



Universität Regensburg

Evaluierung der Akzeptanz-, Leistungs- und Beanspruchungsmanipulation in der Mensch-Maschine-Interaktion kognitiv-immersiver Assistenzsysteme durch Fehlereinflüsse

Inaugural-Dissertation zur Erlangung der Doktorwürde (Dr. phil.)
der Fakultät für Informatik und Data Science
der Universität Regensburg

vorgelegt von
Maximilian Rosilius
aus Bergtheinfeld

Die Arbeit entstand in gemeinsamer Betreuung durch die Fakultät für Informatik und Data Science der Universität Regensburg und dem Institut für Digital Engineering der Technischen Hochschule Würzburg-Schweinfurt

Regensburg 2026

Tag der Abgabe:	27.01.2026
Tag der Disputation:	03.08.2026
Erstgutachter/Betreuer:	Prof. Dr. Bernd Ludwig
Zweitgutachter/Betreuer:	Prof. Dr.-Ing. Volker Bräutigam

Widmung



Für meine Frau Nadja – danke für Deine unendliche Geduld, Deinen Rückhalt und Deine Liebe in all den Jahren.

Vorwort

Im Sinne einer besseren Lesbarkeit und Verständlichkeit wird in dieser Arbeit auf die Nennung mehrerer Geschlechterformen verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter, ausdrücklich wertfrei und ohne Gewichtung. Mit diesem Hinweis im Vorwort wird explizit auf die Gleichstellung aller Geschlechter hingewiesen.

Die Entwicklung und Durchführung der Studien 1 bis 3 wurden teilweise während der COVID-19-Pandemie durchgeführt. Daher ist nicht auszuschließen, dass die Ergebnisse dieser Studien durch Hygieneschutzmaßnahmen sowie den Einsatz von Desinfektionsmitteln an dem verwendeten Equipment beeinträchtigt wurden.

Der Quellcode der im Rahmen dieser Dissertation eingesetzten Software-Forschungsprototypen unterliegt den jeweiligen Forschungsprojekt-Kooperationsverträgen und ist durch Urheber-, Lizenz- und Nutzungsrechte geschützt. Aus diesem Grund kann dieser nicht zur Verfügung gestellt werden.

Die Forschungsmittel sowie die notwendige Infrastruktur wurden dankenswerterweise durch die Technische Hochschule Würzburg-Schweinfurt (THWS) im Rahmen des Forschungsprojekts „PlanAR“ bereitgestellt. Die Förderung erfolgte durch die Bayerische Forschungsförderung (BayVFP), Förderlinie Digitalisierung / FuE-Programm „Informationstechnik und Kommunikationstechnik“ des Freistaates Bayern (Fördernummer: IUK-1811-0022, PlanAR, o. D.). Darüber hinaus wurde der eingesetzte VR-Forschungsprototyp „Bauhausmädels“ dankenswerterweise von Frau Prof. Dr. Yvonne Brandenburger (Fachhochschule Erfurt) aus dem Forschungsprojekt „Virtual Reality für Thüringer Museen: Entwicklung, Leitfaden und Implementierung von Pilotanwendungen“ zur Verfügung gestellt (Universität Erfurt, o. D.-b). Die wissenschaftlichen Methoden, Vorgehensweisen, Forschungsinhalte und Ergebnisse wurden durch die Förderung und die Bereitstellung der Forschungsprototypen nicht beeinflusst und entsprechen ausdrücklich den Grundsätzen guter wissenschaftlicher Praxis. Sie unterlagen ausschließlich den unabhängigen Forschungsinteressen der betreuenden Professoren sowie des Promovenden. Ein besonderer Dank gilt auch dem Bayerischen Wissenschaftsforum (BayWISS), welches u. a. die Mittel für die Probandengelder bereitgestellt hat.

Als Ouvertüre in das Forschungsthema soll ein Gedankenspiel anhand eines Navigationssystems einleiten. Es gilt zu eruieren, inwiefern typische Fehlertypen die Qualität und Effektivität der Mensch-Maschine-Interaktion beeinträchtigen können. Im Fokus stehen dabei Sichtbarkeitsfehler, Datenfehler, Interaktionsfehler sowie Trainingsfehler. Analog gilt es in dem Gedankenspiel zu betrachten, welche Konsequenzen solche Fehler auf die Ankunftszeit am Zielort, auf die subjektive Beanspruchung sowie auf die Akzeptanz der Nutzenden haben. Welche Auswirkungen ergeben sich, wenn Informationen fehlerhaft erkannt oder Inhalte inkorrekt dargestellt werden? Wie beeinflussen beeinträchtigte Interaktionsfunktionen oder Missverständnisse der Bedienlogik die Nutzungserfahrung? In der vorliegenden Dissertation werden diese grundlegenden Fragestellungen – übertragen auf immersive Assistenzsysteme im industriellen Umfeld – systematisch untersucht.

Danksagung

Die vorliegende Dissertation entstand in kooperativer Betreuung zwischen der Fakultät Data Science der Universität Regensburg und meiner wissenschaftlichen Tätigkeit am Institut Digitales Engineering (IDEE) der Technischen Hochschule Würzburg-Schweinfurt. Diese Zusammenarbeit wurde durch das BayWISS-Verbundkolleg ermöglicht und gefördert. Mein herzlicher Dank gilt allen, die mich auf diesem Weg begleitet und unterstützt haben.

Mein besonderer Dank gilt meinem Erstbetreuer Prof. Dr. Bernd Ludwig für die ausgezeichnete wissenschaftliche Betreuung dieser Dissertation. Seine wertvollen Anregungen, sein Vertrauen und die konstruktiven Diskussionen haben diese Arbeit maßgeblich geprägt und entscheidend zu ihrem Gelingen beigetragen.

Ebenso danke ich meinem Zweitbetreuer Prof. Dr. Volker Bräutigam für die Übernahme des Zweitgutachtens. Besonders dankbar bin ich ihm für sein stetes Mentoring in den vergangenen Jahren. Seine unermüdliche Förderung und Motivation haben meine wissenschaftliche sowie persönliche Entwicklung nachhaltig geprägt und meinen Werdegang maßgeblich beeinflusst.

Mein herzlicher Dank gilt meinen Kollegen und Kolleginnen am IDEE für das offene, kollegiale und inspirierende Arbeitsumfeld. Besonders bedanken möchte ich mich bei Dr. Ingo von Eitzen und Steffen Decker für die zahlreichen fachlichen Diskussionen und wertvollen Impulse. Ein besonderer Dank gilt Benedikt Wirsing, Lukas Hügel und Noah Meißner für ihre Geduld, ihr kritisches Feedback und ihre organisatorische Unterstützung bei der Planung und Durchführung komplexer Studien.

Mein persönlicher Dank gilt schließlich meiner Frau und meiner Tochter, meinen Eltern sowie meiner Familie und meinen Freunden. Ihre Unterstützung, ihr Verständnis und ihr Zuspruch haben mich während der gesamten Promotionszeit begleitet und mir auch in anspruchsvollen Phasen den nötigen Rückhalt gegeben. Ohne sie wäre diese Arbeit nicht möglich gewesen.

Zusammenfassung

Infolge der fortschreitenden Digitalisierung wird die Kollaboration von Mensch und Maschine unter anderem durch den Einsatz immersiver Assistenzsysteme (IAS) weiter intensiviert. Für eine erfolgreiche Zusammenarbeit gewinnt die Qualität der Mensch-Maschine-Interaktion (MMI) entscheidend an Bedeutung. Vor diesem Hintergrund untersucht die vorliegende Dissertation systematisch, wie unterschiedliche Fehlertypen in der MMI bei IAS die Leistung, Beanspruchung und Akzeptanz beeinflussen. Ausgehend von Defiziten im bisherigen Forschungsstand – insbesondere der fragmentarischen Evidenz zu Fehlerwirkungen und der fehlenden Differenzierung von Fehlertypen – wird ein integriertes Forschungs- und Evaluationsmodell entwickelt. Dieses verknüpft Leistungsbewertung, Beanspruchung (NASA-TLX und physiologische Parameter) und Akzeptanz (UTAUT2) in einem konsistenten experimentellen Design für industrielle Anwendungsszenarien. Zusätzlich wird eine Konfidenzanzeige als experimentelle Manipulation implementiert, um zu prüfen, inwieweit Systemtransparenz negative Effekte auf die Akzeptanz abschwächen oder kompensieren kann. Im Fokus stehen vier zentrale Fehlerdimensionen: Datenfehler, Visualisierungsfehler, Interaktionsfehler (Sprache und Gesten) sowie Trainingsfehler. Die empirische Basis bilden vier kontrollierte Laborstudien mit insgesamt über 150 Teilnehmenden (S1: $N = 18$; S2: $N = 22$; S3: $N = 52$; S4: $N = 64$). Die statistische Auswertung erfolgt inferenzstatistisch abgesichert durch a-priori-Stichprobenumfangsplanungen sowie Post-Hoc-Poweranalysen, robuste Tests und Change-Score-Verfahren zur Kontrolle interindividueller Unterschiede. Im Sinne des Parsimonie-Prinzips setzt die Arbeit überwiegend robuste, gut interpretierbare Hypothesentests und Varianzanalysen ein. Neben der Entwicklung des übergeordneten Forschungsdesigns leistet die Arbeit methodisch mehrere Beiträge, unter anderem die Einführung eines aggregierten Leistungsindex zur Abbildung des Speed-Accuracy-Tradeoffs mit eindeutiger Bewertung von Bearbeitungszeit und Richtigkeit, die Adaption der Information-Transfer-Rate (ITR) auf Sprachinteraktionen in IAS sowie ein Cut-off-Verfahren zur Bewertung von Trainingsfehlern. Die Ergebnisse belegen signifikante, fehler-typabhängige Effekte auf Leistung, Akzeptanz und Beanspruchung: Datenfehler führen zu einer signifikanten Verschlechterung des Leistungsindex ($p = 0,002$; $Co'd = 0,69$) und gehen zugleich mit einer signifikanten Abnahme der Akzeptanz einher ($Co'd = 0,53$),

während Beanspruchung und physiologische Stressindikatoren keine signifikanten Veränderungen zeigen. Visualisierungsfehler wirken sich primär auf die Leistung aus, wobei sich deskriptiv höhere Effektstärken in VR ($Co'd = 1,15$) als in AR ($Co'd = 0,72$) zeigen; Akzeptanz und Beanspruchung bleiben insgesamt stabil, mit Ausnahme einer signifikant erhöhten Frustration in VR ($Co'd = 1,12$). Interaktionsfehler erweisen sich als besonders kritisch: In der Sprachinteraktion (ASR) verschlechtert sich die Wortfehlerrate signifikant ($p = 0,001$; $Co'd = 0,88$), begleitet von einer signifikanten Reduktion der Interaktionsleistung bei Sprachbefehlen (ITR; $p = 0,001$; $Co'd = 0,88$) und sehr starken Akzeptanzverlusten ($Co'd = 1,55$). Auch bei der Gesteninteraktion zeigen sich signifikante Effekte auf Leistung ($p = 0,006$; $Co'd = 1,08$), Akzeptanz ($Co'd = 0,53$) und Beanspruchung ($Co'd = 0,77$). Trainingsfehler haben durchgängig sehr große, signifikante Effekte auf alle drei Primärfaktoren ($Co'd \geq 1,17$; Power = 1,00). In der vergleichenden Analyse unterscheiden sich insbesondere die Akzeptanz- und Beanspruchungseffekte signifikant zwischen den Fehlertypen, während sich für Leistungs- und physiologische Parameter keine signifikanten Unterschiede zwischen den Fehlertypen nachweisen lassen. Schließlich zeigt sich ein signifikanter positiver Effekt der UV Konfidenzanzeige auf die Akzeptanz bei Interaktionsfehlern ($p = 0,026$; $Co'd = 1,12$; Power = 0,74).

Abstract

As a consequence of progressive digitalisation, collaboration between humans and machines is increasingly intensified through the use of immersive assistance systems (IAS). For successful cooperation, the quality of human-machine-interaction (HMI) becomes a decisive factor. Given this context, the present dissertation systematically investigates how different types of errors in the HMI of IAS affect performance, workload, and acceptance. Motivated by deficiencies in the existing body of research – particularly the fragmented evidence on error effects and the insufficiently differentiated treatment of error types – this work develops an integrated research and evaluation framework. This framework consistently links performance evaluation, workload (NASA-TLX and physiological measures), and acceptance (UTAUT2) within a coherent experimental design tailored to industrial application contexts. In addition, a system confidence indicator is implemented as an experimental manipulation to examine whether system transparency can mitigate or compensate for negative effects on acceptance. The analysis focuses on four central error dimensions: data errors, visualisation errors, interaction errors (speech and gestures), and training errors. The empirical basis comprises four controlled laboratory studies with more than 150 participants in total (S1: $N = 18$, S2: $N = 22$, S3: $N = 52$, S4: $N = 64$). Statistical inference is ensured through a-priori sample size planning and post-hoc power analyses, robust statistical tests, and change-score methods to control for interindividual baseline differences. In line with the principle of parsimony, the study predominantly relies on robust and easily interpretable hypothesis tests and analyses of variance. Beyond the development of the overarching research design, the dissertation makes several methodological contributions, including the introduction of an aggregated performance index to capture the speed-accuracy-trade-off with an unambiguous consideration of execution time and correctness, the adaptation of the Information Transfer Rate (ITR) to speech-based interactions in IAS, and a cut-off procedure for the evaluation of training errors. The results demonstrate significant, error-type-specific effects on performance, acceptance, and workload. Data errors lead to a significant deterioration of the performance index ($p = 0,002$; $Co'd = 0,69$) and are associated with a significant reduction in acceptance ($Co'd = 0,53$), while workload and physiological stress indicators do not show significant changes. Visibility errors primarily impair

performance, with stronger effects in VR ($Co'd = 1,15$) than in AR ($Co'd = 0,72$); acceptance and workload remain largely stable, with the exception of a significant increase in frustration in VR ($Co'd = 1,12$). Interaction errors are particularly critical: in speech interaction (ASR), the word error rate increases significantly ($p = 0,001$; $Co'd = 0,88$), accompanied by a significant decline in interaction performance measured by ITR ($p = 0,001$; $Co'd = 0,88$) and very large losses in acceptance ($Co'd = 1,55$). Gesture-based interactions also show significant effects on performance ($p = 0,006$; $Co'd = 1,08$), acceptance ($Co'd = 0,53$), and workload ($Co'd = 0,77$). Training errors exhibit consistently very large and significant effects across all three primary outcome variables ($Co'd \geq 1,17$; Power = 1,00). Comparative analyses reveal that acceptance and workload effects differ significantly between error types, whereas no significant differences between error types are observed for performance and physiological parameters. Finally, a significant effect of the system confidence indicator on acceptance under interaction errors is observed ($p = 0,026$; $Co'd = 1,12$; Power = 0,74).

Abkürzungsverzeichnis

AF	Akzeptanzfaktoren
ANOVA	Analysis of Variance (Varianzanalyse)
AP	Arbeitsphase
AR	Augmented Reality (erweiterte Realität)
DSS	Decision Support System (Augmented-Reality-Entscheidungsunterstützungssystem)
AS	Assistenzsystem
ASR	Automatic Speech Recognition (automatische Spracherkennung)
AV	Abhängige Variable
BB	Bounding Box (virtuelle Drahtgitterbox)
BCI	Brain-Computer Interface (Gehirn-Computer-Schnittstelle)
BD	Balkendiagramm
BF	Beanspruchungsfaktoren
BI	Behavioral Intention (Nutzungs- bzw. Verhaltensabsicht; Akzeptanz)
BPM	Beats per Minute (Schläge pro Minute)
BRK	Biserialer Rangkorrelationskoeffizient
BSD	Between-Subject Design (Versuchsplan mit unabhängigen Gruppen)
Co'd	Cohen's d (standardisierte Effektstärke)
CWSD	Counterbalanced Within-Subject Design (gegenbalanciertes Versuchsdesign mit Messwiederholung)
DD	Deskriptive Daten
Diff	Differenzwert
DLZ	Durchlaufzeit
DLZS	Durchlaufzeit mit Strafzeiten
DOE	Design of Experiments (Versuchsplanung)
DQ	Datenqualität
DSCF	Dwass-Steel-Critchlow-Fligner-Test
DT	Detailliertes Training
E	Effort (Aufwand/NASA-TLX)
ED	Eigene Darstellung

EE	Effort Expectancy (Aufwandserwartung/UTAUT2)
EER	End-Error (Versuchsreihenfolge Endfehler)
F	Frustration (NASA-TLX)
FC	Facilitating Conditions (unterstützende Bedingungen/UTAUT2)
ff	Fehlerfreie Versuchsbedingung
FF	Forschungsfrage
fh	Fehlerhafte Versuchsbedingung
FT	Fehlertyp
GLBD	Geführtes Learning by Doing
HCI	Human-Computer-Interaction (Mensch-Computer-Interaktion)
HCPS	Human Cyber-Physical System (menschzentrierte cyber-physische Systeme)
HL	Microsoft HoloLens
HM	Hedonic Motivation (hedonische Motivation/UTAUT2)
HMD	Head-Mounted-Display (am Kopf getragenes Anzeigegerät)
HMI	Human-Machine-Interface (Mensch-Maschine-Schnittstelle)
HR	Heart Rate (Herzfrequenz)
HRV	Heart Rate Variability (Herzfrequenzvariabilität)
HT	Habit (Gewohnheit/UTAUT2)
IAS	Immersive Assistenzsysteme
IDEE	Institut Digitales Engineering
IKT	Informations- und Kommunikationstechnologien
IL	Index Leistung
IQ	Interaktionsqualität
IS	Informationssystem
ITR	Information Transfer Rate (Informationsübertragungsrate)
KI	Künstliche Intelligenz
MD	Mental Demand (Mentale Beanspruchung/NASA-TLX)
Mean_	Durchschnittlicher NASA-TLX-Gesamtwert/durchschnittliche Beanspru-
NASA	chung
MMD	Mixed-Method Design (gemischtes Methodendesign)

MMI	Mensch-Maschine-Interaktion
MR	Mixed Reality
MRTK	Mixed-Reality-Toolkit
MT	Minimales Training (Faktorstufe)
MTO	Mensch-Technik-Organisation
ND	Netzdiagramm
NTAI	Nicht trainiert, aber instruiert
OT	Optimiertes Training (Faktorstufe)
P	Performance (Leistung/NASA-TLX)
PE	Performance Expectancy (Leistungserwartung/UTAUT2)
PEU	Perceived Ease of Use (wahrgenommene Benutzerfreundlichkeit/TAM)
PF	Primärfaktoren
PHT	Post-hoc-Test
PP	Physiologische Parameter
PT	Persönliches Training (Faktorstufe)
PU	Perceived Usefulness (wahrgenommener Nutzen/TAM)
RF	Reihenfolge
RP	Ruhephase
S	Studie
SCL	Skin Conductance Level (Hautleitfähigkeitsniveau)
SD	Standardabweichung
SE	Standardfehler
SER	Start-Error (Versuchsreihenfolge Startfehler)
SERA	Systematic Error and Risk Analysis (systematische Fehler- und Risikoanalyse)
SI	Social Influence (sozialer Einfluss/UTAUT2)
SPC	Speech Command (Sprachbefehl)
SPL	Sound Pressure Level (Schalldruckpegel)
SUS	System Usability Scale (Skala zur Bewertung der Gebrauchstauglichkeit eines Systems)
TAM	Technology Acceptance Model (Technologieakzeptanzmodell)

TB	Text und Bilder
TD	Temporal Demand (zeitliche Beanspruchung/NASA-TLX)
TQ	Trainingsqualität
UTAUT2	Unified Theory of Acceptance and Use of Technology 2
UV	Unabhängige Variable
VQ	Visualisierungsqualität (visuelle Qualität)
VR	Virtual Reality (virtuelle Realität)
WER	Word Error Rate (Wortfehlerrate)
XR	Extended Reality (erweiterte Realität)
ZEF	Zielerfüllungsfaktor

Inhaltsverzeichnis

1	Einleitung.....	24
1.1	Motivation und Problemstellung	25
1.2	Themenabgrenzung	28
1.3	Forschungsbedarf und -frage	38
1.4	Aufbau der Arbeit	46
2	Stand der Forschung	48
3	Übergeordnetes Forschungsdesign.....	74
3.1	Hypothesen	74
3.2	Methodik	76
4	Empirische Studien.....	92
4.1	Studie 1 – Sichtbarkeitsfehler I	94
4.1.1	Methodik.....	95
4.1.2	Ergebnisse.....	102
4.1.3	Interpretation	108
4.1.4	Diskussion	109
4.2	Studie 2 – Interaktionsfehler I.....	113
4.2.1	Methodik.....	114
4.2.2	Ergebnisse.....	120
4.2.3	Interpretation	125
4.2.4	Diskussion	126
4.3	Studie 3 – Trainingsfehler	128
4.3.1	Methodik.....	129
4.3.2	Ergebnisse.....	135
4.3.3	Interpretation	142
4.3.4	Diskussion	145
4.4	Studie 4 – Verbundstudie	149
4.4.1	Methodik.....	150
4.4.2	Ergebnisse.....	157
4.4.2.1	<i>Teilstudie 4.1 Fehlerauswirkungen</i>	<i>165</i>
4.4.2.2	<i>Teilstudie 4.2 Fehlervergleich</i>	<i>172</i>
4.4.2.3	<i>Teilstudie 4.3 Akzeptanzmanipulation.....</i>	<i>176</i>
4.4.3	Interpretation	177

4.4.3.1	<i>Teilstudie 4.1 Fehlerauswirkungen</i>	177
4.4.3.2	<i>Teilstudie 4.2 Fehlervergleich</i>	180
4.4.3.3	<i>Teilstudie 4.3 Akzeptanzmanipulation</i>	182
4.4.4	Diskussion	182
4.4.4.1	<i>Teilstudie 4.1 Fehlerauswirkungen</i>	182
4.4.4.2	<i>Teilstudie 4.2 Fehlervergleich</i>	194
4.4.4.3	<i>Teilstudie 4.3 Akzeptanzmanipulation</i>	196
5	Zusammenfassung	199
5.1	Rekurrieren der Erkenntnisse	199
5.2	Implikationen	215
5.3	Limitationen und weiterer Forschungsbedarf	224
5.4	Fazit	232
	Literaturverzeichnis	236
	Anhang A	267
	A1 – Beitrag des Stands der Forschung zur Erklärung der UVs und AVs	267
	A2 – Nutzerbefragungen & Einverständniserklärung	273
	A3 – Studie 1 Sichtbarkeitsfehler I	278
	A4 – Studie 2 Interaktionsfehler I	280
	A5 – Studie 3 Trainingsfehler	282
	A6.1 – Studie 4 Fehlervergleich und -manipulation	287
	A6.2 – Teilstudie 4.2 Fehlervergleich	295

Abbildungsverzeichnis

Abbildung 1. Automatisierungsszenarien im Spannungsfeld MTO (ED).....	25
Abbildung 2. Karikatur zur ambivalenten Wahrnehmung von IAS	27
Abbildung 3: Blockdiagramm des übergeordneten Forschungsdesigns	78
Abbildung 4. VR-Startperspektive (ED) – S1	97
Abbildung 5. VR-Detailansichten versus VQ (ED) – S1	99
Abbildung 6. Versuchsablaufplan S1 (ED).....	100
Abbildung 7. Boxplot DLZ [s] versus der VQ – S1.....	104
Abbildung 8. Boxplot-Richtigkeit versus VQ – S1	105
Abbildung 9. Boxplot-Leistung versus VQ – S1	105
Abbildung 10. Barplots UTAUT2 versus VQ – S1	106
Abbildung 11. Barplots der NASA-TLX versus VQ – S1.....	107
Abbildung 12. Boxplots (a) HR_Diff & (b) SCL_Diff versus VQ – S1.....	107
Abbildung 13. Schematische Darstellung des Versuchsaufbaus – S2	114
Abbildung 14. Amplitude (oben) und Spektrogramm (unten) – S2	115
Abbildung 15. Versuchsablaufplan – S2	117
Abbildung 16. SPC-Dialog des Forschungsprototyps – S2	118
Abbildung 17. Verteilung der Sprachinteraktionserfahrung – S2.....	120
Abbildung 18. Boxplots WER versus UV SPL	121
Abbildung 19. Boxplot ITR versus SPL – S2.....	122
Abbildung 20. Barplots UTAUT2 versus RF – S2	123
Abbildung 21. Barplots NASA-TLX versus RF (ED) – S2.....	124
Abbildung 22. schematische Darstellung des Versuchsaufbaus – S3.....	129
Abbildung 23. Versuchsablaufplan versus TQ – S3	134
Abbildung 24. Erfahrungsverteilung – S3	136
Abbildung 25. Boxplot-Leistung [%/min] versus TQ – S3	137
Abbildung 26. Boxplots ZEF versus TQ – S3	137
Abbildung 27. Boxplots DLZS [s] versus TQ – S3	138
Abbildung 28. Balkendiagramm UTAUT2 versus TQ – S3.....	140
Abbildung 29. Balkendiagramm NASA-TLX versus TQ – S3	141
Abbildung 30. Boxplots (a) HR_Diff & (b) SCL_Diff versus TQ – S3	142

Abbildung 31. Schematische ED des Versuchsaufbaus Verbundstudie 4	153
Abbildung 32. Versuchsablaufplan (ED) – S4.....	155
Abbildung 33. Richtigkeit versus Fehlerfall & Fehlertyp – S4.....	159
Abbildung 34. DLZ versus Fehlerfall & Fehlertyp – S4.....	159
Abbildung 35. Index Leistung versus Fehlerfall & Fehlertyp – S4	160
Abbildung 36. BI versus Fehlerfall & Fehlertyp – S4	161
Abbildung 37. Mean_NASA versus Fehlerfall & Fehlertyp – S4	162
Abbildung 38. HR-Differenz versus Fehlerfall & Fehlertyp – S4.....	163
Abbildung 39. SCL-Differenz versus Fehlerfall & Fehlertyp – S4.....	163
Abbildung 40. HR versus Reihenfolge – S4.....	165
Abbildung 41. Barplots UTAUT versus DQ – S4.1.....	166
Abbildung 42. Barplots NASA-TLX versus DQ – S4.1	167
Abbildung 43. Boxplots der (a) HR_Diff & (b) SCL_Diff versus DQ – S4.1.....	167
Abbildung 44. Barplots UTAUT2 versus UV IQ – S4.1	168
Abbildung 45. Barplots NASA-TLX versus IQ – S4.1	169
Abbildung 46. Boxplots (a) HR_Diff & (b) SCL_Diff versus IQ – S4.1	170
Abbildung 47. Barplots UTAUT2 versus VQ – S4.1	170
Abbildung 48. Barplots NASA-TLX versus VQ – S4.1.....	171
Abbildung 49. Boxplots (a) HR_Diff & (b) SCL_Diff versus VQ – S4.1.....	172
Abbildung 50: Boxplot Leistungs_Diff [%/(10*s)] versus UV Fehlertyp – S4.2.....	173
Abbildung 51. Boxplot Richtigkeit_Diff versus UV Fehlertyp – S4.2	173
Abbildung 52. Boxplot DLZ_Diff [s] versus UV Fehlertyp S4.2.....	174
Abbildung 53. Barplots UTAUT2 versus UV Fehlertyp – S4.2.....	175
Abbildung 54. Barplots NASA-TLX versus Fehlertypen – S4.2	175
Abbildung 55. Boxplots HR_Diff_Diff & SCL_Diff_Diff versus FT – S4.2.....	176
Abbildung 56. Boxplot BI_Diff versus Konfidenzanzeige – S4.3	177
Abbildung 57. ND-Effektstärken des Datenfehlers auf die PF	201
Abbildung 58. ND-Effektstärken des Datenfehlers auf die AF	201
Abbildung 59. ND-Effektstärken des Datenfehlers auf die BF	202
Abbildung 60. BD-Effektstärken des Datenfehlers auf die PP	202
Abbildung 61. ND-Effektstärken des Sichtbarkeitsfehlers auf die PF.....	203

Abbildung 62. ND-Effektstärken des Sichtbarkeitsfehlers auf die AF	204
Abbildung 63. ND-Effektstärken des Sichtbarkeitsfehler auf die BF	205
Abbildung 64. ND-Effektstärken des Sichtbarkeitsfehlers auf die PP	205
Abbildung 65. ND-Effektstärken des Interaktionsfehlers (Gestik) auf die PF	206
Abbildung 66. ND-Effektstärken des Interaktionsfehlers (ASR) auf die PF.....	206
Abbildung 67. ND-Effektstärken des Interaktionsfehlers auf die AF	207
Abbildung 68. ND-Effektstärken des Interaktionsfehlers auf die BF	208
Abbildung 69. BD-Effektstärken des Interaktionsfehlers (Gestik) auf die PP.....	208
Abbildung 70. ND-Effektstärken des Trainingsfehlers auf die PF	209
Abbildung 71. ND-Effektstärken des Trainingsfehlers auf die AF	209
Abbildung 72. ND-Effektstärken des Trainingsfehlers auf die BF	210
Abbildung 73. BD-Effektstärken des Trainingsfehlers auf die PP	210
Abbildung 74. ND-Effektstärken des Fehlervergleichs auf die PF	212
Abbildung 75. ND-Effektstärken des Fehlervergleichs auf die AF	213
Abbildung 76. ND-Effektstärken des Fehlervergleichs auf die BF	214
Abbildung 77. BD-Effektstärken des Fehlervergleichs auf die PP	214
Abbildung 78. Einverständniserklärung zur Erhebung	278

Tabellenverzeichnis

Tabelle 1. Einordnung ausgewählter Fehlertypen in Fehlertaxonomien.....	35
Tabelle 2. Gesamtübersicht des Studiendesigns.....	92
Tabelle 3. Stichprobenumfangsplan – S1	101
Tabelle 4. DD-Leistungsfaktoren versus VQ – S1.....	102
Tabelle 5. DD UTAUT2 versus VQ – S1.....	103
Tabelle 6. DD NASA-TLX versus VQ – S1.....	103
Tabelle 7. DD HR [BPM] & SCL [μ S] versus VQ – S1	104
Tabelle 8. Hypothesentest UTAUT2 versus VQ – S1	106
Tabelle 9. Hypothesentest NASA-TLX versus VQ – S1	106
Tabelle 10. HR [BPM] und SCL [μ S] Test versus VQ – S1.....	107
Tabelle 11. Stichprobenumfangsplanung – S2	118
Tabelle 12. WER [%] – PHT versus SPL – S2.....	120
Tabelle 13. nicht-param. Test WER [%] versus SPL – S2	121
Tabelle 14. DD der WER [%] versus SPL – S2	121
Tabelle 15. ITR [bit/min] PHT versus SPL – S2	121
Tabelle 16. <i>t</i> -Test ITR [bit/min] versus SPL – S2	122
Tabelle 17. DD ITR [bit/min] versus SPL – S2	122
Tabelle 18. Hypothesentest UTAUT2 versus RF – S2	123
Tabelle 19. DD UTAUT2 versus RF – S2.....	123
Tabelle 20. Hypothesentest NASA-TLX versus RF – S2	124
Tabelle 21. DD NASA-TLX versus RF – S2.....	124
Tabelle 22. Überblick UV TQ & Teilaufgaben – S3.....	132
Tabelle 23. Stichprobenumfangsplanung – S3	135
Tabelle 24. Leistung [%/min] PHT versus TQ – S3.....	136
Tabelle 25. DD-Leistung versus TQ – S3.....	137
Tabelle 26. ZEF [%] PHT versus TQ – S3	137
Tabelle 27. DD ZEF [%] versus TQ – S3	138
Tabelle 28. DLZS [s] PHT versus TQ – S3.....	138
Tabelle 29. DD DLZS [s] versus TQ – S3.....	138
Tabelle 30. ANOVA UTAUT2 versus TQ – S3.....	139

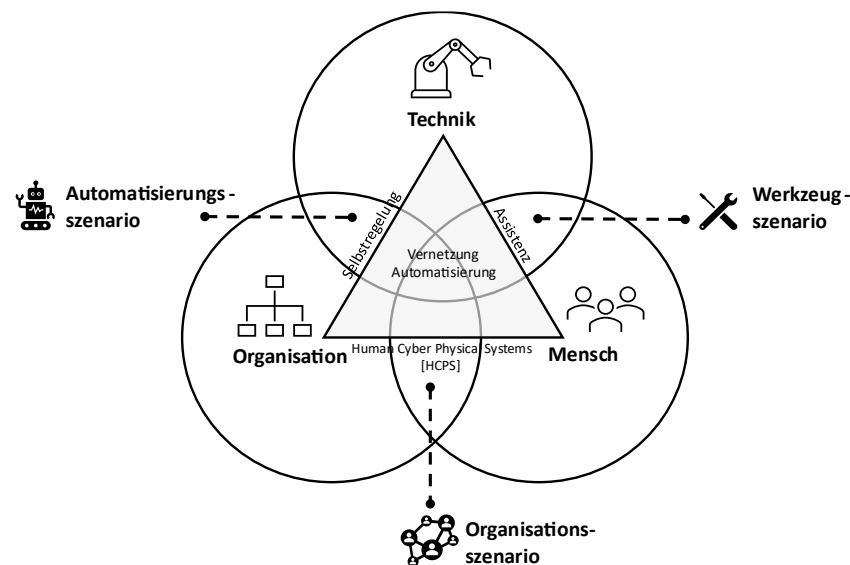
Tabelle 31. UTAUT2 PHT versus TQ – S3	139
Tabelle 32. ANOVA NASA-TLX versus TQ – S3.....	140
Tabelle 33. NASA-TLX PHT versus TQ – S3	141
Tabelle 34. HR [BPM] & SCL [μ S] PHT versus TQ – S3	142
Tabelle 35. DD zu HR [BPM] & SCL [μ S] versus TQ – S3.....	142
Tabelle 36. Stichprobenumfangsplanung – S4	155
Tabelle 37. DD Leistungs-AV versus Fehlerfall & Fehlertyp – S4.....	158
Tabelle 38. DD BI versus Fehlerfall & Fehlertyp – S4	160
Tabelle 39. DD Mean_NASA versus Fehlerfall & Fehlertyp – S4.....	161
Tabelle 40. DD HR [BPM] & SCL [μ S] versus Fehlerfall & Fehlertyp – S4	162
Tabelle 41. Verteilung der Fehlerwahrnehmung – S4.....	164
Tabelle 42. Haupt- & Wechselwirkungseffekt RF & FT * RF – S4	164
Tabelle 43. UTAUT2 PHT versus DQ – S4.1.....	166
Tabelle 44. NASA-TLX PHT versus DQ – S4.1	166
Tabelle 45. HR [BPM] & SCL [μ S] PHT versus DQ – S4.1.....	167
Tabelle 46. UTAUT2 PHT gegen UV IQ – S4.1	168
Tabelle 47. NASA-TLX versus UV IQ – S4.1	169
Tabelle 48. HR [BPM] und SCL [μ S] PHT versus IQ – S4.1	169
Tabelle 49. UTAUT2 versus VQ – S4.1	170
Tabelle 50. NASA-TLX PHT versus VQ – S4.1.....	171
Tabelle 51. HR [BPM] & SCL [μ S] PHT versus VQ – S4.1	171
Tabelle 52. Nicht-param. ANOVA LeistungsAV versus Fehlertyp – S4.2.....	172
Tabelle 53. ANOVA NASA-TLX Differenzen versus Fehlertyp – 4.2	175
Tabelle 54. Gesamtüberblick der Hypothesenbewertungen.....	199
Tabelle 55. Verknüpfung Vorstudien mit Variablen des Forschungsdesign.....	273
Tabelle 56. deutsche Fassung des UTAUT2 angepasst je Studie.....	274
Tabelle 57. deutsche Fassung des NASA-TLX angepasst je Studie	274
Tabelle 58. Soziodemografischer Fragebogen.....	275
Tabelle 59. Spezifische Erhebungen – S1	276
Tabelle 60. Spezifische Erhebungen – S2	277
Tabelle 61. Spezifische Erhebungen – S3	277

Tabelle 62. Spezifische Erhebungen – S4	277
Tabelle 63. Datensatz DLZ, Richtigkeit und Leistung versus VQ – S1.....	279
Tabelle 64. Datensatz NASA-TLX versus VQ – S1	279
Tabelle 65. Datensatz UTAUT2 versus VQ – S1	280
Tabelle 66. Datensatz SCL und HR versus VQ – S1	280
Tabelle 67. Datensatz WER und ITR versus SPL – S2	281
Tabelle 68. Datensatz NASA-TLX versus RF – S2	282
Tabelle 69. Datensatz UTAUT2 versus RF – S2.....	282
Tabelle 70. Datensatz ZEF, DLZ, DLZS versus TQ – S3	283
Tabelle 71. Datensatz NASA-TLX versus TQ – S3.....	284
Tabelle 72. Datensatz UTAUT2 versus TQ – S3.....	285
Tabelle 73. Datensatz SCL und HR versus TQ – S3.....	286
Tabelle 74. Datensatz DLZ, Richtigkeit und Leistung versus Fehlertyp – S4.....	289
Tabelle 75. Datensatz NASA-TLX versus Fehlertyp – S4	291
Tabelle 76. Datensatz UTAUT2 versus Fehlertyp – S4	292
Tabelle 77. Datensatz physiologische Parameter – S4.....	293
Tabelle 78. DD UTAUT2 versus Fehlerfall und Fehlertyp – S4.....	294
Tabelle 79. DD NASA-TLX versus Fehlerfall und Fehlertyp – S4.....	295
Tabelle 80. DD Leistung-, Richtigkeit- und DLZ-Diff versus Fehlertyp – S4.2	295
Tabelle 81. DSCF-Test Leistungsfaktoren versus UV Fehlertyp – S4.2	296
Tabelle 82. PHT UTAUT2 versus UV Fehlertyp – S4.2	296
Tabelle 83. DD Differenzen UTAUT2 versus Fehlertyp	297
Tabelle 84. Studie 4.2 Post-Hoc Test NASA-TLX versus Fehlertyp – S4.2.....	297
Tabelle 85. DD Differenzen NASA-TLX versus Fehlertyp – S4.2.....	297
Tabelle 86. HR_Diff_Diff & SCL_Diff_Diff versus Fehlertyp –S4.2	298
Tabelle 87. DD HR_Diff_Diff & SCL_Diff_Diff versus Fehlertyp – S4.2	298

1 Einleitung

Die fortschreitende Forschung und Entwicklung immersiver Assistenzsysteme (IAS) eröffnet stets neue Anwendungsmöglichkeiten u. a. im industriellen Kontext. Mit weiteren Forschungsaktivitäten sowie erfolgreichen Prototypentests wächst ebenfalls die Technologiereife. Die hohe Komplexität der Systeme stellt nach wie vor ein wesentliches Hindernis für den breiten operativen Einsatz, deren Akzeptanz und damit das Erreichen der Produktivitätsschwelle dar (Mark et al., 2021, S. 23). Das Hauptpotenzial von IAS liegt in der kognitiven Unterstützung der Anwender (Mark et al., 2021, S. 23), die richtigen Informationen zur richtigen Zeit am richtigen Ort zu präsentieren. Somit wird den Nutzern ermöglicht, den gleichen Arbeitsumfang in kürzerer Zeit bei geringerer wahrgenommener Arbeitsbelastung zu erledigen (Fink, 2020; Mark et al., 2021; Ludwig, 2015). Organisatorisch betrachtet zeichnen sich zukünftig drei dominierende Szenarien ab. Gemäß Abbildung 1 wird entsprechend dem Automatisierungsszenario nach Hirsch-Kreinsen (2014) menschliche Arbeitsleistung durch intelligente Systeme substituiert. Die hierdurch freigewordenen menschlichen Ressourcen können durch Erhöhung ihres Kompetenzlevels für neue Aufgabenbereiche genutzt werden. Hingegen wird durch das Werkzeugenszenario nach Windelband und Spöttl (2012) sowie Dombrowski et al. (2014) der Mensch durch intelligente Assistenten befähigt, simple Tätigkeiten um Optimierungstätigkeiten zu erweitern und somit die Vernetzung von Mensch und Maschine zu intensivieren. Das dritte Szenario geht aus der konsequenten Verschmelzung von Mensch und Technik hervor, dem sog. Human Cyber Physical Systems (HCPS) (Hadorn et al., 2016; B. Wang et al., 2020; J. Zhou et al., 2019). Hierbei bedingen „Actio und Reactio“ von Mensch und Maschine einander und führen zu einem hybriden organisatorischen Level. Solche Szenarien bergen auch Risiken und Gefahren, gerade im Hinblick auf die menschlichen Faktoren sowie die Gesamtsystemstabilität respektive Produktivität (Gegoff et al., 2023). Dabei können z. B. durch den sogenannten „Technostress“ (Fuglseth & Sørenbø, 2014; Funk et al., 2019; Ragu-Nathan et al., 2008) Frust, Unbehagen, Misstrauen oder Ängste gegenüber solchen Systemen ausgelöst werden. In der Arbeitswelt von Morgen wandeln sich die Aufgabenstellungen (Hirsch-Kreinsen et al., 2018). Mitarbeiter müssen Systeme bedienen können, die richtigen Informationen erhalten, diese verstehen, Handlungsempfehlungen richtig interpretieren und

sich auf die bereitgestellten Informationen verlassen können (vgl. Ludwig, 2015). Die Qualität und Effektivität menschlicher Arbeit hängen nicht mehr nur von Ausbildung, Erfahrung, Motivation und Führung ab, sondern auch von der Frage nach der Qualität und den Eigenschaften der Informationsübertragung und Interaktion zwischen Mensch und Maschine respektive nach der hybriden Effektivleistung des sog. Human Centered Cyber Physical System (J. Zhou et al., 2019).



Fußnote. ED des Mensch-Technik-Organisationsmodells (MTO) in Anlehnung an Hirsch-Kreinsen et al. (2018) und Strohm (2019)

Abbildung 1. Automatisierungsszenarien im Spannungsfeld MTO (ED)

1.1 Motivation und Problemstellung

Entsprechend meiner autoethnografischen Motivation (Kaltenhauser et al., 2024) zeigt sich die praktische Relevanz der mangelnden Akzeptanz aufgrund von Systemfehlern – insbesondere der organisationalen Ablehnung – von Assistenzsystemen (AS) aus den Erfahrungen, die ich im Rahmen meiner beruflichen Tätigkeit als IoT-Applikationsingenieur sammeln durfte. Ich war mit der Aufgabe betraut, ein digitales Werkzeug zu entwickeln, welches Maschinenbediener beim Einrichten bzw. Einstellen unterstützt. Mit unserem Team bestehend aus Maschineneinrichtern, Applikationsentwicklern des Maschinenherstellers sowie Auszubildenden war es 2018 gelungen, ein prototypisches Werkerassistenzsystem (genannt „Autotuner“) für den CNC-Mehrspindeldrehautomaten Schütte SC9-46 in Betrieb zu nehmen. Dieser Prototyp ermöglichte es

Maschinenbedienern, die manuell erhobenen Messwerte aus der stündlichen statistischen Prozesskontrolle zur automatisierten Korrektur der Führungswerte so zu verrechnen, dass die fertigungsbedingten Maße dieser Wertschöpfungsstufe gen Toleranzmitte lagen. Zusammenfassend konnten durch diesen Piloten drei Primäreffekte erreicht werden: a) die Gesamtanlagenproduktivität konnte rechnerisch um ca. 7,01 % (386 h/a) gesteigert werden; b) der Gesamtausschuss des Prozesses konnte um 0,44 % reduziert werden; c) die Kompetenzanforderungen und der mentale Bedarf bei der Durchführung der Tätigkeit wurden signifikant reduziert, sodass der Prozess zuverlässig von ungelernten Mitarbeitenden zusammen mit diesem Werkerassistenzsystem beherrscht werden konnte (Pfennig, 2018). Trotz dieser Unterstützungsfunktionen der operativen Arbeitsaufgaben zur Nutzung der wirtschaftlichen und organisatorischen Potentiale scheiterte dieses prototypische AS und stieß weder bei den Maschinenbedienern oder -einrichtern noch beim Management auf Akzeptanz. Formal wurden diverse Fehlergründe genannt, die dazu führten, dass das AS abgelehnt wurde. Als primäre Fehlerursachen wurden u. a. nicht angepasste organisatorische Unterstützungsprozesse (u. a. Fehler bei der Messmittelwartung, unzureichende Netzwerkinfrastruktur) sowie der Ausfall der Human-Machine-Interface (HMI) (Anzeige der Messdaten) aufgrund der Vibration der Maschine identifiziert. Sekundär wurde von unterschiedlichen Personas am Shopfloor die Pilotierung aufgrund von Vorbehalten erschwert. Beispielsweise wurden Unsicherheiten durch mangelndes Vertrauen in die Systemlogik (Komplexität), Existenzängste aufgrund der Kompetenzabbildung durch das AS, unzureichende Einweisungen usw. genannt. Dieses Beispiel zeigt, dass Systemfehler, unzureichende Maßnahmen bei der Pilotierung sowie persönliche Ressentiments dazu führten, dass AS von Organisationen trotz offensichtlicher Potentiale nicht angenommen werden.

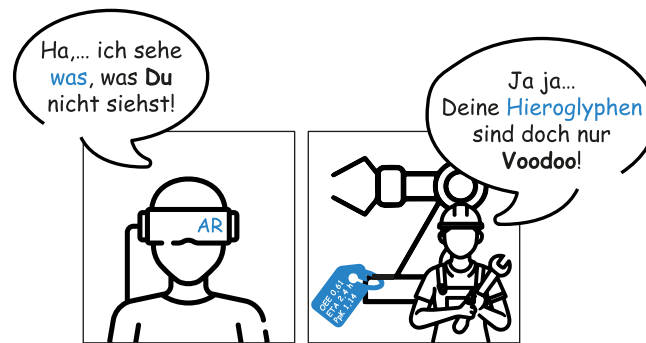


Abbildung 2. Karikatur zur ambivalenten Wahrnehmung von IAS

Quelle: ED unter Verwendung adaptierter Elemente des Icons „Virtuelle Realität“ von Magnific, Flaticon (o. J.), https://www.flaticon.com/de/kostenloses-icon/virtuelle-realitat_1377790, sowie des Icons „Roboter“ von Icons8 (o. J.), <https://icons8.de/icon/933/roboter> (Monthly Icons Subscription erworben). Die Darstellung des Werkers wurde mithilfe von Adobe Illustrator erstellt.

Losgelöst von den Multimomentaufnahmen persönlicher Erfahrungen können solche Effekte unter dem Produktivitätsparadox (Schweikl & Obermaier, 2020) subsumiert werden, dass die Potentiale moderner Information und Kommunikationstechnologien (IKT) angewandt in AS nicht ausgeschöpft werden. Dieses Paradox besagt, dass dank moderner Unterstützung durch Automatisierung und Digitalisierung eine höhere Gesamtproduktivitätsleistung zu erwarten wäre, als sich aktuell bemessen lässt. Nach Schweikl und Obermaier (2020) sind die Gründe hierfür mehrschichtig und umfassen u. a. überzogene Erwartungen, zeitliche Anpassungsverzögerungen, Fehlbewertungen der Produktivität und fehlendes Technologiemanagement. Jedoch sind neben der Produktivität zwei primäre Schlüsselfaktoren im wissenschaftlichen Diskurs zum industriellen Produktivitätsparadox deutlich unterrepräsentiert, denen die Erkenntnisse dieser Dissertation Rechnung tragen sollen: die Systemstabilität respektive Fehlerauswirkungen und konsekutive menschliche Faktoren wie Nutzerakzeptanz und Beanspruchungen innerhalb der Mensch-Maschine-Interaktion (MMI). Auch wenn Kosten und Ressourcen zur Entwicklung, Beschaffung und Implementierung von AS bzw. IKT aufgewendet werden, amortisieren sie sich bzw. bewirken sie nur dann eine Produktivitätssteigerung, wenn diese von den Nutzergruppen auch angenommen werden (Ali & Younes, 2013). „In der Vergangenheit wurden viele wichtige Technologien entwickelt, die jedoch vor allem rein die Produktivität der Unternehmen verbesserten und meist firmenseitig

orientiert waren. Dies ist bei AS anders: Sie verfolgen den Ansatz, den Mitarbeitenden in den Mittelpunkt zu stellen und das Unternehmen insgesamt zu verbessern, indem der Bediener mit adäquater Unterstützung ausgestattet wird“ (Mark et al., 2021, S. 25). Studien zeigen diesbezüglich, dass Nutzer AS u. a. aufgrund mangelnder Akzeptanz und negativer Vorerfahrungen ablehnen (Sochor et al., 2019). In Bezug auf automatisierte Systeme kann eine geringe Systemstabilität zu einer geringeren Akzeptanz führen, da Benutzer möglicherweise das Vertrauen in das System verlieren, wenn es häufig Ausfälle oder Fehler aufweist (Brauner et al., 2019; Gegoff et al., 2023; Konopka et al., 2024a, 2024b). Dies kann zu Frustration (F), Unsicherheit und einer Abnahme der Bereitschaft führen, das System zu verwenden oder auf seine Handlungsempfehlungen zu vertrauen (Gegoff et al., 2023; Weidemann & Rußwinkel, 2021). In diesem Sinne zeigt die Karikatur in Abbildung 2 links einen Werker, der mit einem Augmented-Reality-Headset, ausgestattet durch digitale Augmentierungen (in blau dargestellt), zusätzliche Informationen zu seinem Aufgabenkontext erhält. Er interagiert mit einem IAS und fühlt sich durch diese erweiterten Fähigkeiten seinen analogen Kollegen gegenüber überlegen. Rechts hingegen ist ein traditionell arbeitender Werker mit Helm und Maulschlüssel dargestellt, der die digitalen Einblendungen abwertend als unverständliche „Hieroglyphen“ bezeichnet. Mit dem Begriff „Voodoo“ unterstellt er, die digitalen Informationen seien unklar, fehleranfällig oder sogar kontraproduktiv für den Arbeitsablauf. Die Karikatur verdeutlicht auf überzeichnete Weise das Spannungsfeld zwischen digitaler und anachronistischer Arbeitsweltanschauung in der industriellen Praxis.

Rekurriert führt dies zu der übergeordneten, forschungsleitenden Frage:

Welche Auswirkungen haben Fehler bei IAS innerhalb der MMI auf die Leistung, Beanspruchung und Akzeptanzfaktoren und können adverse Effekte reduziert werden?

1.2 Themenabgrenzung

Dieses Kapitel dient der thematischen Einordnung und Abgrenzung der vorliegenden Monografie innerhalb der bestehenden Forschungslandschaft. Ziel ist es, den thematischen Rahmen einzugrenzen und zugleich eine verständliche Orientierung auch für fachfremde Leser zu schaffen. Dadurch sollen sich die zugrunde liegenden

Fragestellungen, die eingesetzten Methoden sowie die angestrebten Erkenntnisse transparenter und nachvollziehbarer erörtern lassen.

Grundlegend lassen der Titel sowie die forschungsleitende Fragestellung (vgl. Kapitel 1.1) dieser Arbeit zu, sich letzterer aus mehreren Disziplinen zu nähern, u. a. aus den Blickwinkeln der Ingenieur-, Arbeits-, Sozial- und Wirtschaftswissenschaft, der Informatik sowie der angewandten Psychologie. Diese Arbeit ordnet sich methodisch in die interdisziplinären Human-Computer-Interaction (HCI) Forschungsaktivitäten eines praxiserfahrenen Ingenieurs ein.

Besonderheiten und begriffliche Erläuterungen zu immersiven Systemen: Zur begrifflichen Einordnung immersiver Systeme bietet das von Milgram und Kishino (1994) entwickelte Mixed-Reality-Kontinuum eine zentrale Grundlage. Es beschreibt einen fließenden Übergang zwischen realen und virtuellen Umgebungen. Am einen Ende des Spektrums steht die reale Welt, am anderen eine vollständig virtuelle Realität. Dazwischen liegt der Bereich der Mixed Reality (MR), in dem reale und virtuelle Inhalte miteinander kombiniert werden. Innerhalb dieses Zwischenbereichs lassen sich zwei wesentliche Konzepte differenzieren: Bei der Augmented Reality (AR) wird die reale Umgebung durch virtuelle Informationen ergänzt, während bei der Augmented Virtuality eine überwiegend virtuelle Welt durch reale Elemente angereichert wird. Das Modell verdeutlicht, dass Realität und Virtualität nicht als Gegensätze, sondern als fließend ineinander übergehende Zustände zu verstehen sind (Milgram & Kishino, 1994). Ergänzend dazu beschreibt Azuma (1997) drei charakteristische Merkmale immersiver Systeme: die Kombination von realen und virtuellen Inhalten, die Interaktion in Echtzeit sowie die präzise räumliche Verankerung (als sog. Registrierung) virtueller Objekte in der realen Welt. Neben diesen technischen Kriterien spielen auch nutzerzentrierte Dimensionen für die Wirksamkeit und Akzeptanz immersiver Anwendungen eine zentrale Rolle. Laut von Eitzen (2024) beeinflussen insbesondere folgende Faktoren die Akzeptanz maßgeblich. Hedonismus beschreibt das wahrgenommene Vergnügen und die Freude an der Nutzung – unabhängig vom funktionalen Nutzen. Immersion bezeichnet die technische und sensorische Einbettung in eine virtuelle Umgebung, etwa durch realistische 3D-Grafik, räumlichen Klang oder hohe Bildraten, und erzeugt ein intensives Gefühl des Eintauchens. Präsenz meint das subjektive Erleben, sich tatsächlich innerhalb der virtuellen

Welt zu befinden; sie wird durch Immersion gefördert und verstärkt wiederum das hedonistische Nutzungserlebnis. Verkörperung beschreibt die Wahrnehmung, einen eigenen Körper oder Avatar in der virtuellen Welt zu besitzen und mit diesem interagieren zu können, was Präsenz und Vergnügen zusätzlich steigert. Interaktivität ist die Fähigkeit, aktiv mit der virtuellen Umgebung zu agieren, was als Voraussetzung für Präsenz und das Erleben eines Flow-Zustands gilt. Letzterer beschreibt eine Phase vollständiger Konzentration und intrinsischer Motivation. Weitere relevante Dimensionen sind Bedienbarkeit (Ease of Use), also die Benutzerfreundlichkeit des Systems, sowie die daraus resultierende Zufriedenheit, verstanden als positiver emotionaler Zustand, der entsteht, wenn Erwartungen erfüllt und Nutzungserlebnisse als angenehm empfunden werden. Ein hoher Realismus steigert die Authentizität und Glaubwürdigkeit der Virtual Reality (VR) Erfahrung und unterstützt somit sowohl Präsenz als auch Immersion. Schließlich beschreibt Involvement den psychologischen Zustand der starken Fokussierung und Konzentration auf die virtuelle Umgebung, der die Aufmerksamkeit bündelt und das Eintauchen in die VR-Erfahrung intensiviert. Involvement steht in engem Zusammenhang mit Präsenz und verstärkt das Nutzungserlebnis (von Eitzen, 2024).

Differenzierung von IAS: IAS bezeichnen technologische Lösungen, die den Menschen in industriellen Kontexten bei Wahrnehmung, Informationsverarbeitung, Entscheidung und Handlungsausführung unterstützen. IAS zeichnen sich durch eine starke Integration kognitiver und immersiver Komponenten aus. Sie kombinieren sensorische Erfassung, situationsadaptive Reaktionen und multimodale Schnittstellen wie Augmented Reality bis hin zu VR (abhängig vom Grad der Immersion) oder sprachbasierte Interaktion, um nicht nur zu informieren, sondern menschliches Handeln aktiv zu strukturieren und zu erweitern (Fellmann et al., 2017, S. 60; Mark et al., 2021, S. 140). Laut Definition nach Kraus et al. (2021) sind AS softwarebasierte Agenten, die Nutzer bei der Bearbeitung komplexer Aufgaben unterstützen, indem sie auf einen geäußerten oder erkannten Assistenzbedarf reagieren. Sie zeichnen sich durch eine interaktive und kooperative Zusammenarbeit mit dem Nutzer, durch Problemlösungskompetenz sowie häufig durch Lernfähigkeit aus. Ziel eines AS ist es, gemeinsam mit dem Nutzer zielorientiert Aufgaben zu bewältigen, wobei das System sowohl eigene als auch nutzerseitige Aktivitäten beobachten, bewerten und situationsangepasst darauf reagieren kann. Im

Klassifikationsschema von Mark et al. (2021) werden kognitive Systeme verstanden, die Denk- und Entscheidungsprozesse durch intelligente Datenverarbeitung und Echtzeit-feedback begleiten. Sie unterscheiden sich deutlich von sensorischen oder physischen AS, die vorrangig Informationen erfassen oder körperlich entlasten. Der Fokus kognitiver Systeme liegt auf mentaler Entlastung, Fehlervermeidung und Prozesssicherheit in komplexen Produktionsumgebungen. Nach Mark et al. (2021) stellen IAS keine eigenständige Abkehr vom Assistenzsystembegriff dar, sondern sind als systemische Erweiterung klassischer AS um immersive Funktionalitäten zu verstehen, die die Mensch-System-Interaktion räumlich, multimodal und kontextsensitiv erweitern, sog. projektionsbasierte AS. Die theoretische Fundierung der Interaktion zwischen Mensch und IAS lässt sich anhand des kognitiven Handlungsmodells von Wandke (2005) strukturieren. IAS greifen insbesondere in die Phasen der Informationsintegration, Entscheidungsfindung und Rückmeldung ein. Dabei fungieren sie als kognitive Erweiterung, indem sie Wahrnehmungslücken schließen, Alternativen aufzeigen oder korrekatives Feedback geben. Unterstützt wird dies durch ein hohes Maß an Multimodalität, etwa durch visuelle, auditive und taktile Kanäle, die simultan agieren (Wandke, 2005). Fellmann et al. (2017) beschreiben IAS im Kontext smarter Fabriken als hochgradig kontextadaptiv, individualisiert, proaktiv und interaktiv. Sie stehen damit am oberen Ende eines morphologischen Rasters, das zwischen statisch und dynamisch, generisch und spezifisch sowie mono- und multimodal unterscheidet. Ihre Flexibilität erlaubt den Einsatz in variantenreichen, wissensintensiven und nicht standardisierbaren Arbeitssituationen. Im Rahmen des „Operator-4.0“-Konzepts (Romero et al., 2016; zitiert in Mark et al., 2021) werden IAS nicht nur als Werkzeuge, sondern als Erweiterung menschlicher Handlungskompetenz interpretiert. Der Mensch agiert hier als kognitiv augmentierter Akteur in symbiotischer Interaktion mit digitalen Systemen. Ziel ist nicht die Automatisierung, sondern die informierte, selbstbestimmte Entscheidungsfindung – unterstützt durch adaptive Maschinenintelligenz. In der vorliegenden Dissertation liegt der Fokus nicht auf der kontinuierlichen Erfassung des Arbeitsfortschritts oder einer situativen Begleitung der Nutzer. Die Assistenzfunktionalitäten folgen vielmehr einer vereinfachten, starren Logik, um eine hohe Erklärbarkeit und Nachvollziehbarkeit der Systemreaktionen für die

Probanden sicherzustellen. Adaptive oder lernbasierte Mechanismen treten dabei bewusst in den Hintergrund.

Differenzierung des Forschungskontextes: Angelehnt an „Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009-2019)“ von Merino et al. (2020) wird im Folgenden der Forschungskontext der vorliegenden Dissertation systematisch in die Klassifikationskategorien Typ der Forschungsarbeit, Forschungsthema, Evaluationszenario, kognitive Aspekte sowie Studienkonfiguration eingeordnet. Im Rahmen dieser Arbeit wurde ein übergeordnetes Evaluationsdesign entwickelt, mit dem verschiedene Auswirkungen von Fehlern innerhalb der MMI im Kontext fachlicher Aufgabenstellungen untersucht wurden, die an industrielle Anwendungsfelder angelehnt sind. Grundsätzlich handelt es sich dabei um eine nutzerzentrierte Evaluationsstudie, in der Probanden in mehreren Studien unterschiedliche fachliche Aufgaben in einer mobilen Laborumgebung ausführten. Das Spektrum der Aufgaben reichte dabei von der reinen Informationsaufnahme in einer virtuell abgebildeten Museumsumgebung über AR-gestützte Spracheingaben unter industriellen Lärmbedingungen bis hin zu verschiedenen AR-basierten Messaufgaben. In Bezug auf das Forschungsthema lassen sich die Studien den Rubriken Design, menschliche Faktoren, HCI-Technologien, multimodale Schnittstellen sowie Applikationen zuordnen. Als Evaluationsszenarien wurden thematisch übergeordnet sowohl Nutzerleistung als auch User Experience betrachtet. Die kognitiven Aspekte umfassten dabei Wahrnehmung, Usability, Emotionen, kognitive Beanspruchung sowie Lernfähigkeit. Mit dieser Einordnung positioniert sich die vorliegende Arbeit innerhalb der von Merino et al. (2020) identifizierten humanzentrierten Evaluationsansätze und berücksichtigt zudem die Bedeutung von mobilen und realitätsnahen Studienkonfigurationen für die ökologische Validität der Ergebnisse.

Differenzierung der betrachteten Fehlerdimensionen: Innerhalb IAS ist eine differenzierte Analyse von Fehlern auf Interaktionsebene zentral, da Wahrnehmung, Informationsverarbeitung, Bedienung und mentales Modell eng miteinander verflochten sind. Bahaei et al. (2019), Bahaei und Gallina (2019) und Palanque et al. (2020) bieten hierzu komplementäre Perspektiven und erweitern etablierte Mensch-Fehler-Modelle. Dabei greifen sie auf bewährte Klassifikationen zurück wie das „Human Factors Analysis and Classification System“ (Shappell & Wiegmann, 2000), das „Systematic Error and Risk

Analysis“ (kurz SERA, vgl. Hendy, 2003), die „Standardized Plant Analysis Risk – Human Reliability Analysis“-Methode (Gertman et al., 2005), das „Skill-Rule-Knowledge Framework“ (Rasmussen, 1983) sowie die „Fault-Error-Failure-Taxonomie“ zur Systemzuverlässigkeit nach Avizienis et al. (2004). Bahaei und Gallina (2019) schlagen eine featurebasierte Erweiterung bestehender Fehlerklassifikationen vor, insbesondere zur Erfassung visueller Fehler, die spezifisch für AR-Systeme sind. Bahaei et al. (2019) integrieren AR-induzierte Fehlerursachen in ein umfassendes Modell für soziotechnische Systeme und berücksichtigen dabei technische, organisationale und menschliche Einflussfaktoren mit dem Schwerpunkt auf fehlerhaften Informationsflüssen. Palanque et al. (2020) setzen am HCI-Loop an und ordnen Fehler systematisch den perzeptiven, kognitiven und motorischen Subsystemen des Nutzers auf Basis von Echtzeitstörungen in der Interaktion zu. Aus diesen drei Vorarbeiten gehen differenzierte Betrachtungen der aus Nutzersicht zentralen Fehlertypen im Kontext immersiver Mensch-Maschine-Interaktion hervor (vgl. Tabelle 1):

- *Visualisierungsfehler* treten auf, wenn AR-Inhalte nicht korrekt dargestellt werden oder durch Umgebungsbedingungen wie Lichtreflexionen, schlechte Kontraste oder fehlerhafte Platzierung unlesbar sind. Bahaei und Gallina (2019) führen sie als eigene, AR-spezifische Fehlerkategorie ein und erweitern bestehende Fehlerklassifikationen um sogenannte image-centric failure modes. Sie verorten diese Fehler klar auf Systemseite, insbesondere bei nicht optimal gestalteten Displays oder Head-Mounted-Displays (HMD). Auch Bahaei et al. (2019) klassifizieren Visualisierungsfehler als technisch bedingte Ursachen, etwa bei fixen Fokalebene, die zu Wahrnehmungseinbußen führen. Palanque et al. (2020) ordnen diese Fehler der perzeptiven Dimension zu und verstehen sie als Störungen, die aus dem Zusammenspiel von Mensch, Interface und Umgebung resultieren. Alle drei Studien bewerten Visualisierungsfehler als besonders kritisch, da sie das Situationsbewusstsein direkt beeinträchtigen.
- *Informationsfehler* beziehen sich auf falsche, unvollständige oder nicht interpretierbare Systeminformationen. Bahaei et al. (2019) strukturieren diese entlang interner (z. B. kognitive Überforderung) und externer Ursachen (z. B.

fehlerhafte Sensorik) und ordnen sie als Hauptursache für kognitive Fehlentscheidungen ein. Auch Bahaei und Gallina (2019) thematisieren Informationsfehler, insbesondere wenn überlagerte Inhalte widersprüchlich oder irreführend sind. Palanque et al. (2020) klassifizieren sie systematisch als kognitive Fehlerzustände, die durch eine fehlerhafte oder überladene Informationsverarbeitung im Interaktionskontext entstehen. In allen Arbeiten gelten sie als hochrelevant für die Integrität von Entscheidungsprozessen in datenbasierten AR-Systemen.

- *Interaktionsfehler* entstehen, wenn Nutzereingaben – etwa durch Gesten, Sprache oder Touch – vom System nicht korrekt erkannt oder verarbeitet werden. Palanque et al. (2020) nehmen sie als zentrales Element ihrer Fehlerklassifikation auf und unterscheiden zwischen motorischen, sensorischen und kognitiven Interaktionsstörungen, die durch Systemdesign, Umwelt oder Nutzerzustand bedingt sind. Auch Bahaei et al. (2019) betonen externe Ursachen, etwa wenn die Darstellung auf dem HMD die Koordination stört. Bahaei und Gallina (2019) ordnen Interaktionsfehler im Rahmen ihrer featurbasierten Modellierung ebenfalls als Systemschwächen ein, die zu Ausführungsfehlern führen können. Alle drei Arbeiten erkennen diese Fehlerform als besonders relevant für die Bedienbarkeit und Nutzerakzeptanz an.
- *Trainingsfehler* betreffen unzureichende Vorbereitung oder Einweisung in die Nutzung immersiver Systeme. Bahaei et al. (2019) integrieren sie explizit in ihre SERA-basierte Struktur als personelle Einflussgröße. Palanque et al. (2020) behandeln Trainingsfehler indirekt, indem sie auf fehlende mentale Modelle oder kognitive Kapazitäten hinweisen, die zu Fehlhandlungen führen. Auch Bahaei und Gallina (2019) thematisieren das Fehlen notwendiger Strategien als systeminduzierte Fehlerursache. Trotz unterschiedlicher Gewichtung wird deutlich, dass Trainingsfehler besonders relevant sind, wenn neue, nicht-intuitive Interaktionsformen in AR-Systemen genutzt werden.

Fehlertyp	Bahaei & Gallina, 2019	Bahaei et al., 2019	Palanque et al., 2020
Visualisierung	Einführung als neue, AR-spezifische Fehlerkategorie; Fokus auf visuelle Ablenkung und	Technisch-externe Fehlerquelle durch fehlerhafte Darstellung oder	Perzeptive Fehler infolge gestörter visueller Wahrnehmung durch Umwelt oder Display

Fehlertyp	Bahaei & Gallina, 2019	Bahaei et al., 2019	Palanque et al., 2020
Information	Überlagerung durch Systemdesign Indirekt adressiert durch kognitive Verarbeitungsprobleme bei widersprüchlichen oder nicht interpretierbaren AR-Inhalten	unpassende Fokusebene Klar differenziert nach internen (kognitiv) und externen (sensorisch-technisch) Ursachen	Kognitive Fehler durch fehlerhafte, überladene oder unvollständige Informationen
Interaktion	Aufgeführt als Systemverhalten mit negativen Auswirkungen auf die Nutzeraktion (z. B. nicht erkannte Eingabe, Ausführungsfehler)	Extern induzierte Störung im Handlungsvollzug durch Beeinträchtigung der Koordination (z. B. HMD)	Zentrale Kategorie, differenziert nach motorischen, sensorischen und kognitiven Störungen
Training	Betrachtet als systeminduzierte Folge mangelnder Einweisung in AR-spezifische Handlungsstrategien	Explizit modelliert als personelle Fehlerquelle bei fehlender Vorbereitung oder Ausbildung	Indirekt über unzureichende kognitive Ressourcen und fehlende intuitive Systemgestaltung beschrieben

Tabelle 1. Einordnung ausgewählter Fehlertypen in Fehlertaxonomien

Differenzierung der Evaluationsmethoden: In diesem Absatz werden einschlägige Vorarbeiten kurz erörtert, von denen das übergeordnete Forschungsdesign ausgeht, das sich dann weiter differenziert. Das Modell von Parasuraman et al. (2000) bewertet AS entlang vier Funktionsarten – von Informationsgewinnung bis Handlungsausführung – jeweils nach ihrem Automationsgrad und ihren Auswirkungen auf mentale Belastung, Situationsbewusstsein, Übervertrauen und Fähigkeitsabbau aufgrund von Automationen. Ziel ist die Bestimmung eines funktional optimalen Automatisierungsniveaus, das Nutzen maximiert und Risiken begrenzt. Die Studie (S) von Onnasch et al. (2014) erweitert diesen Ansatz empirisch durch die Bewertung von Routineleistung, Störfallleistung, Arbeitsbelastung und Situationsbewusstsein anhand objektiver und subjektiver Messgrößen. Sie zeigt, dass höhere Automatisierung zwar die Effizienz im Normalbetrieb steigert und Belastung reduziert, jedoch gleichzeitig die Reaktionsfähigkeit und das Situationsbewusstsein im Störfall deutlich senkt – besonders ab einem kritischen Automatisierungsgrad. Auch wenn der daraus resultierende Kompromiss zwischen Effizienz und Resilienz Grundlage für eine differenzierte Bewertung zukünftiger AS bildet, wird in der vorliegenden Arbeit der Automationsgrad nicht berücksichtigt. Ausgehend von zwei zentralen Kritikpunkten am bestehenden Forschungsstand – der unzureichenden Berücksichtigung von Akzeptanzfaktoren (AF) und der fehlenden differenzierten

Analyse spezifischer Fehlertypen – rückt diese Dissertation die systematische Verbindung von Leistungsbewertung (im Regel- und Störfall), Beanspruchungs- sowie Akzeptanzfaktoren in den Mittelpunkt. Obwohl das Konstrukt Vertrauen ein zentraler Untersuchungsgegenstand zahlreicher Studien bei AS ist, stellt es in der vorliegenden Arbeit keinen eigenen Forschungsschwerpunkt dar – auch wenn klar anerkannt wird, dass Vertrauen eine notwendige Grundvoraussetzung für die zukünftige Nutzungsabsicht (Akzeptanz) darstellt und diese auch maßgeblich beeinflusst (Nasirian et al., 2017). Mehrere einschlägige Arbeiten bieten hierzu wertvolle theoretische Grundlagen: Konopka et al. (2024a) sowie Konopka et al. (2024b) analysieren, wie Nutzer ihr Vertrauen in AR-basierte Entscheidungsunterstützungssysteme anpassen (Vertrauenskalibrierung) und wie sich dies auf Systemnutzung und Nutzungsintention auswirkt. McGuirl und Sarter (2006) untersuchen Vertrauensfehlkalibrierung in zeitkritischen, automatisierten Entscheidungssituationen, etwa in der Luftfahrt und diskutieren potenzielle Gegenmaßnahmen durch verbessertes Feedbackdesign. Madhavan et al. (2006) zeigen, dass Fehler automatisierter Systeme – insbesondere bei vermeintlich einfachen Aufgaben – das Vertrauen besonders stark beeinträchtigen können.

Differenzierung des Einsatzkontextes: Obwohl zahlreiche Vorstudien zu AS in unterschiedlichsten Einsatzbereichen wertvolle wissenschaftliche Grundlagen liefern, unterscheidet sich der Fokus dieser Dissertation klar durch ihre Ausrichtung auf den industriellen Einsatz von IAS. Im weiteren HCI-Umfang wurden Evaluationsstudien (vgl. Metastudie von Stefanidi et al. (2023)) u. a. in Lernumgebungen bzw. Education (Lin & Yu, 2023; Chang et al., 2022; Zhang et al., 2022), in musealen Kontexten (Y. Zhou et al., 2022), in der medizinischen Versorgung (Sadeghi Milani et al., 2024), in Games (Mekler et al., 2014) sowie für Menschen mit Beeinträchtigungen (Beneteau, 2020) etc. durchgeführt. In der vorliegenden Arbeit gilt es komplementär zu diesen Kontexten methodische und praktikable Beiträge zu leisten.

Differenzierung von kognitiven Wahrnehmungsprozessen: Sadovitch (2020, S. 35) veranschaulicht in seiner Abbildung das erweiterte Modell des Situationsbewusstseins nach Endsley (1995), ergänzt durch systematische Einflussfaktoren gemäß Vollrath und Krems (2011). Das Modell versucht die kognitiven Prozesse, die einem Menschen ermöglichen, in dynamischen und informationsintensiven Umgebungen relevante

Informationen aufzunehmen, zu interpretieren und daraus zielgerichtete Handlungen abzuleiten. Zentral ist die Untergliederung in drei aufeinander aufbauende Ebenen: Auf Level 1 erfolgt die Wahrnehmung relevanter Umweltelemente, etwa sensorischer oder systemgenerierter Reize. Level 2 umfasst das situative Verständnis, also die Integration dieser Elemente in ein sinnvolles mentales Modell. Level 3 beschreibt die Antizipation zukünftiger Entwicklungen auf Grundlage des situativen Verständnisses. Diese drei Ebenen stehen in direkter Verbindung zu den nachgelagerten Prozessen der Handlungsplanung und -ausführung. Sie sind wiederum in einen zyklischen Rückkopplungsprozess eingebettet, da das Ergebnis einer Handlung rückwirkend Einfluss auf das Situationsbewusstsein nimmt. Darüber hinaus benennt das Modell mehrere Dimensionen, die die Ausprägung und Qualität von Situationsbewusstsein beeinflussen. Hierzu zählen individuelle Faktoren wie Ziele, Erfahrung, Erwartungen oder Training ebenso wie kognitive Ressourcen – darunter Arbeits- und Langzeitgedächtnis sowie automatisierte Handlungsroutinen. Ergänzt wird dies durch system- und aufgabenbezogene Einflussgrößen, etwa Komplexität, Schnittstellengestaltung, Informationsdichte oder Stressbelastung. Sadovitch (2020) überträgt dieses Modell auf den Kontext von Augmented-Reality-basierten Head-up-Displays in Fahrerassistenzsystemen. Obwohl das Modell ursprünglich auf den automobilen Kontext angewendet wurde, könnten die darin beschriebenen kognitiven Prinzipien auf IAS im industriellen Bereich übertragen werden. Auch hier müssen Nutzende in komplexen, zeitkritischen Arbeitsumgebungen Informationen effizient aufnehmen, verarbeiten und auf deren Grundlage korrekt handeln. Die Gestaltung dieser Systeme erfordert daher ein tiefes Verständnis der kognitiven Voraussetzungen, um Fehlerquellen zu minimieren und Leistung sowie Sicherheit zu optimieren. In der vorliegenden Dissertation wird jedoch nicht vertiefend auf kognitionspsychologische Informationsverarbeitungsprozesse eingegangen. Das Modell des Situationsbewusstseins dient vielmehr als theoretischer Referenzrahmen, um den menschlichen Wahrnehmungs- und Handlungsprozess kontextuell einzuordnen. Thematisch grenzt sich diese Arbeit davon ab, indem der Fokus auf die Analyse von Fehlertypen hinsichtlich Leistung, Akzeptanz- und Beanspruchungsfaktoren (BF) von IAS in industriellen Anwendungsszenarien gelegt wird.

1.3 Forschungsbedarf und -frage

Wie in Kapitel 1.1 dargestellt, besteht ein begründetes Interesse an einer vertieften Erörterung dieses Themas. Die kritische Literaturrecherche hat einen konkreten Forschungsbedarf identifiziert. Aus diesem wurden die zentralen Forschungsfragen (*FF*) und Hypothesen (*H*) abgeleitet, die dieser Arbeit zugrunde liegen.

Allgemeiner Forschungsbedarf zur Evaluation der Mensch-Maschine-Interaktion: Die Metastudien von Dey et al. (2018) und Merino et al. (2020) verdeutlichen einen klaren Forschungsbedarf im Bereich der Evaluationspraxis von MR-Anwendungen in unterschiedlichen Kontexten. Insbesondere fehlt es an integrativen Evaluationsansätzen, die sowohl technologische als auch nutzerzentrierte Aspekte systematisch und gleichzeitig erfassen. In vergangenen Studien wurden kognitive Dimensionen meist isoliert betrachtet. Zukünftige Forschung sollte daher Methoden entwickeln, die mehrere dieser kognitiven Aspekte simultan abbilden, um ein ganzheitlicheres Verständnis der Nutzererfahrung zu ermöglichen (Merino et al., 2020, S. 9–10). Das Framework von Stowers et al. (2017) stellt ein theoretisches Modell zur umfassenden Bewertung in Mensch-Maschine-Systemen dar, indem es systematisch Inputs, menschliche Prozesse und resultierende Outcomes integriert. Die zentrale Forschungsfrage richtet sich auf die präzise Erfassung dieser Dimensionen unter Berücksichtigung vielfältiger Einflussfaktoren. Daraus ergibt sich ein konkreter Forschungsbedarf hinsichtlich der Operationalisierung des Frameworks: Es bedarf der Entwicklung und Validierung geeigneter Messmetriken sowie anwendbarer methodischer Ansätze, um die identifizierten Faktoren empirisch zu erfassen und zu analysieren. Darüber hinaus ist eine kontextspezifische Adaption erforderlich, um das Framework in Hinblick auf unterschiedliche Anwendungsfelder praxisgerecht anzupassen. Zusätzlich müssen situationsabhängige Einflussfaktoren und neu entstehende Variablen identifiziert und integriert werden, die für spezielle Ausprägungen der Mensch-Maschine-Interaktion von Relevanz sind (Stowers et al., 2017). Die Arbeit von Moencks et al. (2022) befasst sich mit der systematischen Integration von Augmentationstechnologien in industriellen Arbeitsumgebungen mit dem Ziel, menschliche Arbeit technologiegestützt zu optimieren. Hierfür wurde der „Augmented Workforce Canvas“ entwickelt – ein Referenzmodell, das Unternehmen bei der Planung, Implementierung und Bewertung von Projekten zur Mensch-Technologie-Integration unterstützen soll. Aus

der Studie ergeben sich eindeutige Forschungsbedarfe, sodass eine vertiefte Untersuchung der Technologieintegration speziell im industriellen Kontext notwendig ist und bestehende industrielle Rahmenbedingungen detaillierter modelliert werden sollten, um der Komplexität realer Einsatzszenarien gerecht zu werden. Die Forschungsgemeinschaft ist sich darin einig, den Operator als unmittelbaren Nutzer von Augmentations-technologien stärker in den Mittelpunkt von Industrie-4.0-Aktivitäten zu stellen. Dies erfordert vertiefte Untersuchungen und Entwicklungsansätze, die den Menschen konsequent in den Fokus rücken, um die Effektivität und Akzeptanz technologischer Unterstützungssysteme im industriellen Kontext zu erhöhen (Cimini et al., 2020; Fantini et al., 2020; Longo et al., 2017; Moencks et al., 2022; Pacaux-Lemoine et al., 2017). Außerdem braucht es mehr quantitative und statistische Daten für die Charakterisierung menschlicher Tätigkeiten, sodass menschliche Arbeit unter verschiedenen Bedingungen analysiert wird (Fantini et al., 2020).

Forschungsbedarf zu Qualitätsauswirkungen in der Mensch-Maschine-Interaktion: In der Studie von Nasirian et al. (2017) wurden die Auswirkungen von Informationsqualität, Systemqualität und Interaktionsqualität (IQ) auf die Nutzungsintention respektive die Akzeptanz von AS untersucht. Hieraus wird der Forschungsbedarf abgeleitet, dass zukünftige Studien sich auf spezifische Dimensionen der IQ fokussieren, wobei der Fokus nicht nur auf Sprachassistenten gelegt werden soll (Nasirian et al., 2017). Die Taxonomie nach Palanque et al. (2020) erweitert bestehende Fehlerklassifikationen im sog. HCI-Loop, wobei eine Differenzierung nach internen und externen Ursachen sowie den zugrundeliegenden menschlichen Fähigkeiten vorgenommen wird. Im Rahmen dieser Arbeit wird ein Forschungsbedarf formuliert, der die Entwicklung formaler Modelle zur Beschreibung der Operatorhandlungen im HCI-Loop fordert. Darüber hinaus wird die Notwendigkeit betont, das Konzept der MMI in praktischen Anwendungen verschiedener fachlicher Kontexte – beispielsweise in geräuschintensiven Umgebungen – weitergehend zu erforschen und zu validieren. Suh und Prophet (2018) haben zum einen eine umfangreiche Literaturrecherche durchgeführt und zeigen darin klar die Forschungslücke bezüglich der negativen Auswirkungen von immersiven Technologien. Zukünftige Forschung sollte technische Probleme und unerwünschte Effekte stärker einbeziehen. Außerdem sollten zukünftige Untersuchungen vielfältigere Nutzergruppen und

Anwendungskontexte einschließen. Auf der anderen Seite haben Suh und Prophet (2018) mit ihrer grundlegenden Arbeit zur Entwicklung eines konzeptionellen Frameworks basierend auf dem Stimulus-Organismus-Reaktions-Modell bei immersiven Systemen wesentliche theoretische Impulse gesetzt. Aus diesem Ansatz ergibt sich der zentrale Forschungsbedarf der vorliegenden Arbeit.

Forschende könnten bedeutende Beiträge leisten, indem sie die Auswirkungen der Passung bzw. Nicht-Passung zwischen verschiedenen Stimulusarten auf die Nutzererfahrung untersuchen. So könnte empirische Forschung beispielsweise aufzeigen: (1) wie die Passung zwischen technologischen und inhaltlichen Stimuli eine positive Nutzererfahrung hervorruft; (2) wie die Nicht-Passung zwischen sensorischen und perzeptuellen Stimuli zu einer negativen Nutzererfahrung führt; und (3) wie technologische Stimuli den Einfluss inhaltlicher Stimuli auf die Nutzererfahrung abschwächen oder verstärken. Solche Untersuchungen können zum besseren Verständnis der Wirkmechanismen von Nutzererfahrungen in immersiven Technologien beitragen. (Suh & Prophet, 2018, S. 11)

In der Studie von Pourmemar (2024) wurde nachgewiesen, dass Fehler bei der Interaktion mit immersiven AR-Menüs sowohl auf benutzerbezogene Faktoren – wie kognitive Belastung und physische Ermüdung – als auch auf system- und menübezogene Einflussgrößen zurückzuführen sind. Um ein tiefergehendes Verständnis für die Wirkung dieser Einflussfaktoren zu gewinnen und darauf aufbauend evidenzbasierte Designrichtlinien zur Fehlerreduktion in AR-Interfaces zu entwickeln, besteht weiterer quantitativer Forschungsbedarf (Pourmemar, 2024).

Forschungsbedarf zu Auswirkungen mangelhafter Visualisierungsqualität: Die Studie von Alessa et al. (2023) untersuchte die Nutzerbelastung bei der Ausführung von Aufgaben mit unterschiedlichem Schwierigkeitsgrad unter variierenden Informationsdarstellungen, darunter auch AR-basierte Arbeitsanweisungen. Die Ergebnisse verweisen auf einen erheblichen Forschungsbedarf in Bezug auf die Auswirkungen unterschiedlicher Visualisierungsformen und -qualitäten auf die kognitive und physische Beanspruchung der Nutzer. Insbesondere in komplexeren Anwendungskontexten und erweiterten Studiendesigns erscheint eine weiterführende Differenzierung dieser Einflussfaktoren erforderlich (Alessa et al., 2023). Durch die Studie von Sadovitch (2020) konnte nachgewiesen werden, dass bei Augmented-Reality-Head-up-Displays die subjektive Akzeptanz

mit zunehmender Größe bestimmter Fehlerarten, insbesondere von Registrierungsfehlern, signifikant abnimmt. Verallgemeinernd wurde in dieser grundlegenden Untersuchung der Einfluss von Anzeigefehlern auf die Nutzerakzeptanz in Fahrerassistenzsystemen analysiert. Daraus ergibt sich ein weiterführender Forschungsbedarf hinsichtlich der Übertragbarkeit dieser Ergebnisse auf andere Arten von Visualisierungsfehlern sowie auf unterschiedliche Einsatzkontexte von AR-Systemen (Sadovitch, 2020). In der Metaanalyse (Ötting et al., 2022) wird der Einfluss von Designmerkmalen an der Mensch-Maschine-Schnittstelle bei Robotern auf u. a. Leistung, Akzeptanz sowie mentale Belastung untersucht. Die Ergebnisse bestätigen u. a., dass die Sichtbarkeit an den Systemen einen signifikanten Einfluss auf Leistung und Akzeptanz hat. Allerdings basiert diese Bestätigung auf einer begrenzten Anzahl von Studien und Effektgrößen, so dass ein weiterer Forschungsbedarf besteht. Insbesondere ist eine Ausweitung der Untersuchungen erforderlich, um die Generalisierbarkeit der Befunde auf andere Arten von AS sowie unterschiedliche Einsatzkontexte fundiert zu prüfen (Ötting et al., 2022).

Forschungsbedarf zur Interaktionsqualität innerhalb der Mensch-Maschine-Interaktion: Die Metastudie von Al-kfairy et al. (2024) weist auf einen Forschungsbedarf hinsichtlich der IQ und deren Einfluss auf die Akzeptanz hin: „Die von den Nutzern wahrgenommene Komplexität kann sowohl als Barriere als auch als Förderfaktor für die Technologieadoption wirken. Vereinfachte und intuitive Metaverse-Schnittstellen werden bevorzugt, da sie die kognitive Belastung reduzieren und die Akzeptanz erhöhen“ (Al-kfairy et al., 2024). Es besteht Forschungsbedarf hinsichtlich der Untersuchung, wie unterschiedliche Interaktionstypen und Stimuli (technologisch und inhaltlich) zusammenwirken und wie Fehlanpassungen (Misfits) zwischen ihnen negative Nutzererfahrungen auslösen können. Um die Funktionalitäten der automatischen Spracherkennung (ASR) erfolgreich innerhalb der MMI bei IAS einsetzen zu können, müssen die Anforderungen der aktuellen Technologiereife angemessen hinsichtlich der Produktivität berücksichtigt werden (Vajpai & Bora, 2016). Im Bereich der Sprachschnittstellentechnologie gilt es, die entscheidenden nutzerzentrierten Faktoren zu verstehen, die die Produktnutzung und -anwendung beeinflussen (Vajpai & Bora, 2016), woraus sich weiterer Forschungsbedarf. Die Studie von Bräuer und Mazarakis (2018) zeigt, dass externe Störeinflüsse, insbesondere Umgebungsgeräusche, zu erheblichen Funktionsstörungen bei der sprachbasierten

Mensch-Maschine-Interaktion (bei ASR) führten, sodass die Untersuchung zunächst abgebrochen werden musste. Nach dem Wechsel zu einer alternativen Eingabemethode (Bluetooth-Klicker) konnte die Studie fortgeführt werden. Dennoch besteht der Verdacht, dass die Auswirkungen der fehlerhaften ASR und der resultierenden Verzerrungen im Studiendesign auf die Untersuchungsergebnisse unterschätzt wurden. Es ist notwendig, die Fähigkeiten und Auswirkungen der ASR auf der HoloLens (HL) 2 in einer industriellen Umgebung zu evaluieren (Derby et al., 2020). Im Rahmen der Studie von Graichen und Graichen (2023) wurde untersucht, inwiefern beobachtbare Fehler bei der Nutzung gestenbasierter Interaktionen in Virtual-Reality-Umgebungen die subjektive Akzeptanz der Nutzer beeinflussen. Die Ergebnisse deuten darauf hin, dass Nutzer zwar Fehler wahrnehmen, diese jedoch nicht zwangsläufig zu einer Akzeptanzreaktion der Technologie führen. Diese Befunde widersprechen bisherigen Forschungsergebnissen, die einen negativen Zusammenhang zwischen Systemfehlern und Nutzerakzeptanz nahelegen. Die daraus resultierende Diskrepanz unterstreicht den Bedarf an weiterführender Forschung zur Rolle von Fehlern und deren Reaktion in immersiven Interaktionsszenarien (Cockburn et al., 2019).

Forschungsbedarf zu unzureichender Trainingsqualität: Day et al. (2019) legt nahe, dass organisatorische Unterstützung, insbesondere in Form eines systematisch gestalteten Ressourcenmanagements, eine zentrale Rolle für die erfolgreiche Schulung neuer Technologien spielt. Dabei wird betont, dass Trainingsmaßnahmen die individuellen Bedürfnisse der Arbeitnehmer berücksichtigen müssen, um die Technologieakzeptanz zu fördern und Widerstände zu minimieren (Day et al., 2019). Marshall et al. (2008, S. 5) weisen explizit darauf hin, dass die Endbenutzerschulung ein bislang wenig erforschter, jedoch bedeutender Faktor bei der Einführung neuer Technologien ist. Fehlende oder unzureichende Schulungen können maßgeblich zur Ablehnung (mangelnde Akzeptanz) neuer Systeme beitragen. Hier besteht ein wesentlicher Forschungsbedarf hinsichtlich der Frage, wie spezifische Trainingskonzepte entwickelt werden können, die sich an den Anforderungen der Endanwender orientieren und damit die Akzeptanzquote steigern (Marshall et al., 2008). In der Untersuchung von Palanque et al. (2020) wird deutlich, dass Fehler in der Mensch-Computer-Interaktion häufig auf mangelhafte Schulungen zurückzuführen sind. Eine geeignete Auswahl und Gestaltung der Schulungsinhalte

können helfen, diese Fehlerquellen signifikant zu reduzieren. Forschung wird daher benötigt, um Methoden zu identifizieren, wie Trainingsprogramme systematisch zur Fehlerprävention beitragen können. Im industriellen Kontext, insbesondere bei der Implementierung von Augmented-Reality-Anwendungen, wurde ein spezifischer Forschungsbedarf festgestellt (Quandt & Freitag, 2021). Autoren „(...) äußern die folgenden zusätzlichen Bedenken der Nutzer hinsichtlich der Einführung von AR-basierten Systemen in einer industriellen Arbeitsumgebung: Einstellung gegenüber neuer Technologie, fehlende Kenntnisse zur Nutzung des Systems, fehlender Mehrwert, (...) Systemkomplexität (...)“ etc. (Quandt & Freitag, 2021, S. 13). Graser und Böhm (2022) machen zudem darauf aufmerksam, dass bestehende Trainingsmodelle häufig auf schulische oder akademische Lernumgebungen zugeschnitten sind und betriebliche Faktoren nur unzureichend abbilden.

Forschungsbedarf zu Auswirkungen von Automatisierungsfehlern bzw. mangelhafter Datenqualität: In der Studie von Brauner et al. (2019) wurde untersucht, wie Benutzer mit fehlerhaften Entscheidungsunterstützungssystemen interagieren und welche Auswirkungen solche Fehler u. a. auf die Benutzerakzeptanz und die Leistung haben. Aus der Studie ergibt sich eine Empfehlung für zukünftige Forschungsaktivitäten: Zum einen sollen unterschiedliche Komplexitätsgrade sowie verschiedene Einsatzkontexte berücksichtigt werden. Zum anderen wird gefordert, weitere Nutzerperspektiven systematisch einzubeziehen. Darüber hinaus wird die Notwendigkeit betont, ein grundlegendes Verständnis für das Verhalten von Menschen im Umgang mit fehlerhaften AS insbesondere in hochintegrierten Produktionsumgebungen zu entwickeln. Die Studie von Azeroual et al. (2020) weist nach, dass mangelnde Datenqualität (DQ) die Akzeptanz von Forschungssystemen negativ beeinflusst. Die Studie von Grabowski (2015) untersucht den Beitrag und die Auswirkungen von Wearable-Immersive-Augmented-Reality-Technologien auf die nautische Navigation und Sicherheit im maritimen Bereich, wobei sie die Herausforderungen betont, die mit der Einführung neuartiger Technologien in sicherheitskritischen maritimen AR-Systemen verbunden sind. Der Forschungsbedarf besteht darin, systematisch zu untersuchen, wie fehlerhafte oder unvollständige Darstellungen bei AS HCI-Faktoren u. a. Nutzerleistung, mentale Beanspruchung (z. B. Stress, Ermüdung) und Technologieakzeptanz beeinflussen (Grabowski, 2015). Die Studie von

Bahaei und Gallina (2019) untersucht menschliche Fehler im Kontext AR-erweiterter Fähigkeiten und betont die Notwendigkeit, die bestehende Fehlerklassifikation weiterzuentwickeln, empirisch zu validieren und in praxistaugliche Analysewerkzeuge zu überführen. Zudem fordert sie eine vertiefte Forschung zu den Verhaltensauswirkungen durch AR-Technologien (Bahaei & Gallina, 2019). Übergreifend lässt sich im Rahmen dieser Literaturrecherche feststellen, dass Untersuchungen zu den Auswirkungen von Fehlern – insbesondere im Zusammenhang mit fehlerhaften Daten und Automatisierungsfehlern in AS – sowie speziell bei immersiven Systemen bislang nur unzureichend adressiert wurden. Dieses Defizit weist auf eine bestehende Forschungslücke hin, die eine systematische wissenschaftliche Aufarbeitung erfordert.

Forschungsbedarf zu Auswirkungen der Mensch-Maschine-Interaktion auf die Belastung: Die Arbeit von Tomori & Ogunseiju (2025) verdeutlicht grundlegend, in welchem Maße verschiedene Faktoren innerhalb von MR-Umgebungen die kognitive Belastung beeinflussen. Sie unterstreicht insbesondere die Bedeutung eines durchdachten Designs, einer stabilen Systemarchitektur sowie von gezielten Trainingsmaßnahmen zur Reduktion dieser Belastung. Aus den Ergebnissen der Studie ergibt sich zudem ein klarer Forschungsbedarf hinsichtlich der Untersuchung von MR-Lernumgebungen unter realen, dynamischen Bedingungen außerhalb kontrollierter Laborumgebungen (Tomori & Ogunseiju, 2025). Die Studie von Bläsing und Bornewasser (2021) zeigt, dass die Art des eingesetzten AS einen maßgeblichen Einfluss auf die subjektive Wahrnehmung kognitiver Belastung hat. Zwar führen AR-Brillen zu einer leicht erhöhten mentalen Beanspruchung, jedoch wird dies nicht zwangsläufig als negativ empfunden. Zukünftige Untersuchungen sollten verstärkt berücksichtigen, dass wahrgenommene Komplexität und mentale Belastung nicht nur auf die Aufgabe, sondern auch wesentlich auf die spezifischen Eigenschaften und das Design des Systems zurückzuführen sein können (Bläsing & Bornewasser, 2021). Auffällig ist in diesem Zusammenhang die Studie von Alessa et al. (2023). Sie zeigt, dass AR-Arbeitsanweisungen die kognitive und physische Belastung senken können im Vergleich zu einer Arbeitsanweisung aus Papier. Es bedarf vergleichender Untersuchungen zu unterschiedlichen AR-Technologien, Anzeigearten und Interaktionsmodalitäten (visuell, auditiv, haptisch), um deren Einfluss auf die Belastung

und Akzeptanz zu verstehen. Zudem sollten Studien realitätsnahe, komplexe industrielle Aufgaben berücksichtigen (Alessa et al., 2023).

Forschungsbedarf zu Auswirkungen der Mensch-Maschine-Interaktion auf die Akzeptanz(-manipulation): Gegoff et al. (2023) und Hecht et al. (2020) zeigen, dass Transparenzmaßnahmen in AS kontextabhängig wirken: Während Hecht et al. (2020) keinen signifikanten Einfluss prädiktiver Statusanzeigen auf das Vertrauen feststellen konnten, ergaben die Befunde von Gegoff et al. (2023), dass Transparenz bei niedriger Automationszuverlässigkeit „Disuse“ reduziert, die Vertrauenskalibrierung verbessert und die subjektive kognitive Belastung senkt. Disuse bezeichnet dabei die unbegründete Ablehnung korrekter Systemempfehlungen, Misuse hingegen die unkritische Übernahme fehlerhafter Empfehlungen. Bei hoher Systemzuverlässigkeit konnte kein Rückgang von Misuse beobachtet werden, da Nutzer fehlerhafte Ratschläge weiterhin akzeptierten. Diese Ergebnisse verdeutlichen nicht nur die differenzielle Wirkung von Transparenz in Abhängigkeit von der Systemzuverlässigkeit, sondern auch einen weiterführenden Forschungsbedarf hinsichtlich ihrer Generalisierbarkeit. Insbesondere ist zu untersuchen, wie sich Transparenzinformationen in anderen Anwendungs- und Domänenkontexten auswirken. Aus der Studie von McGuirl und Sarter (2006) ergibt sich ein klarer Forschungsbedarf hinsichtlich der Übertragbarkeit von vertrauensbasierten Informationsanzeigen auf andere Automationssysteme. Die Arbeit von Wanner et al. (2022) liefert u. a. empirische Evidenz für die Wirksamkeit eines integrierten Modells, das Transparenz und Vertrauen als zentrale Einflussfaktoren auf die Akzeptanz intelligenter Systeme berücksichtigt. Daraus ergibt sich ein konkreter Forschungsbedarf: Insbesondere im industriellen Kontext bedarf es vertiefter Untersuchungen zur Akzeptanz von AR-Systemen sowie zur Wechselwirkung zwischen Transparenz, Vertrauen und der Überwindung bestehender Akzeptanzbarrieren (Wanner et al., 2022, S. 4). Aus der Studie von Quandt und Freitag (2021) über die systematische Überprüfung der Benutzerakzeptanz in der industriellen Augmented Reality geht die Notwendigkeit für Forschungsaktivitäten hervor, „(...) die die Kombination von Faktoren untersuchen, die eine ganzheitliche Bewertung des Einsatzes von industriellen AR-Systemen mit einem vertretbaren Aufwand ermöglichen“ (Quandt & Freitag, 2021, S. 15).

Bei der kritischen Literaturrecherche zu dieser Forschungsarbeit zeichnete sich ab, dass Studien zur Untersuchung der Fehlerauswirkungen innerhalb der Themeneingrenzung (Abschnitt 1.2) im Hinblick auf die Leistung, Akzeptanz- und Beanspruchungsfaktoren in der aktuellen Forschung allgemein deutlich unterrepräsentiert sind.

Forschungsfragen

Entsprechend des vorgestellten Forschungsbedarfs wurden folgende Forschungsfragen abgeleitet, deren Beantwortung einen relevanten Forschungsbeitrag leisten sollen:

FF_{1.A-D}: Welchen Einfluss hat bzw. haben

A. fehlerhafte **Daten**,

B. eine fehlerhafte **Visualisierung**,

C. eine fehlerhafte **Interaktion**,

D. ein unzureichendes **Trainingsszenario** innerhalb der MMI auf die Leistung, Akzeptanz sowie die Beanspruchung bei IAS?

FF₂: Manipulieren **unterschiedliche Fehlerarten** bei IAS innerhalb der MMI die Leistungen, Akzeptanz und Belastungen unterschiedlich?

FF₃: Welchen Einfluss hat die **Anzeige der Systemkonfidenz** innerhalb der MMI auf die Akzeptanz bei IAS?

1.4 Aufbau der Arbeit

Dieser Abschnitt gibt einen Überblick über die übergeordnete Struktur der Arbeit und dient als Leitfaden zur Orientierung in den einzelnen Kapiteln:

In *Kapitel 1*, der Einleitung, wird grundlegend in den Forschungskontext der MMI eingeleitet und zum Thema fehlerhafter, immersiver AS hingeführt. Hierbei wird vor allem auf die Relevanz des Themas eingegangen. Weiterführend wird in *Abschnitt 1.1* auf die Motivation zur Problemstellung sowie auf den wissenschaftlichen Diskurs des Produktivitätsparadoxons eingegangen. Als nächstes wird in *Abschnitt 1.2* das Themenfeld eingegrenzt und das Forschungsthema präzisiert. In *Abschnitt 1.3* werden der inhärente Forschungsbedarf sowie die Forschungsfragen erläutert. Mit der strukturierten Beschreibung des Aufbaus der Arbeit (*Abschnitt 1.4*) wird die Einleitung abgeschlossen.

In *Kapitel 2* wird der Stand der Forschung mit den theoretischen Grundlagen und Vorarbeiten des Themengebiets abgebildet.

In *Kapitel 3* wird dem Leser das übergeordnete Forschungsdesign präsentiert. Hierbei werden in *Abschnitt 3.1* die dedizierten Forschungsthese und in *Abschnitt 3.2* die Methodik eingehend erläutert.

In *Kapitel 4* werden die empirischen Studien beschrieben, die zur Überprüfung der Thesen und Beantwortung der Forschungsfragen entwickelt wurden: In *Abschnitt 4.1* wird die *Studie 1* zur Untersuchung des Sichtbarkeitsfehlers anhand eines VR-Forschungsprototyps dargestellt. Hierzu wird in *Abschnitt 4.1.1* die spezifische Methodik eingehend beschrieben. Nachfolgend werden in *Abschnitt 4.1.2* die Ergebnisse sachlich sowohl deskriptiv als auch inferenzstatistisch ausgewertet und so präsentiert, dass der Leser ein objektives Bild erhält. Ebenfalls werden eingehend die formalen, soziodemografischen Daten zusammengefasst. In *Abschnitt 4.1.3* erfolgt die Interpretation der Ergebnisse anhand der Überprüfung der Forschungsthese. Nachfolgend werden in *Abschnitt 4.1.4* die Ergebnisse sowohl in Hinblick auf die Limitationen der gewählten Methodik als auch auf die Vorarbeiten diskutiert. Die *Abschnitte 4.1.1-4.1.4* werden als Blaupause bestehend aus Methodik, Ergebnisdarstellung, Interpretation der Ergebnisse sowie der Diskussion ebenfalls auf die weiteren Studien in den *Kapiteln 4.2-4.4* dieser Arbeit abgebildet. In *Abschnitt 4.2* wird die *Studie 2* zur Untersuchung des Interaktionsfehlers I hinsichtlich der Spracheingabe als eines der immersiven Interaktionsparadigmen präsentiert. In *Abschnitt 4.3* wird die *Studie 3* zur Untersuchung des Trainingsfehlers in Bezug auf den Forschungsprototyp einer immersiven Messanwendung dargestellt. In *Abschnitt 4.4* wird die *Studie 4* als Verbundstudie zur Untersuchung unterschiedlicher Fehlertypen sowie deren Vergleich anhand eines gemeinsamen Forschungsprototyps dargelegt.

In *Kapitel 5* werden die Ergebnisse der Arbeit zusammengefasst und systematisch auf die Forschungsfrage rekurriert (*Kapitel 5.1*). Das *Kapitel 5.2* leitet die wissenschaftlichen sowie praktischen Implikationen der Arbeit ab. In *Kapitel 5.3* werden die methodischen Limitationen sowie der daraus abgeleitete Bedarf weiterer Forschungsansätze präsentiert. Abschließend wird in *Kapitel 5.4* ein Resümee über die vorliegende Dissertationsschrift gezogen.

2 Stand der Forschung

Dieses Kapitel fasst die theoretischen Grundlagen und empirischen Erkenntnisse relevanter Vorarbeiten zum Forschungsthema zusammen. Auf Grundlage des identifizierten Forschungsbedarfs gilt es, an die Methoden, die empirischen Studien sowie die Erkenntnisse des Forschungsstands anzuknüpfen. Zur besseren Einordnung und Verständlichkeit werden die Ergebnisse einschlägiger Studien überwiegend konsistent im jeweiligen Studienkontext präsentiert.

Als erweiterten Forschungsbedarf hinsichtlich dieser Arbeit beschreibt Wandke (2005) in seiner Assistenz-Taxonomie sechs Phasen menschlichen Handelns – Motivation, Wahrnehmung, Integration, Entscheidung, Ausführung und Feedback – und formuliert für jede Phase spezifische Anforderungen an die Gestaltung technischer Systeme. Das konzeptionelle Modell verdeutlicht, dass eine unzureichende Berücksichtigung dieser Anforderungen die Nutzererfahrung und Systemnutzung beeinträchtigen kann. Fehlt in der Motivationsphase eine klare, aktivierende Unterstützung, können Orientierungslosigkeit und Handlungsaufschub entstehen. In der Wahrnehmungsphase führen unübersichtliche oder schlecht dargestellte Inhalte zu kognitiver Überlastung, die das Erfassen relevanter Inhalte behindert. Dies setzt sich in der Integrationsphase fort: Fehlende Kontextualisierung oder unzureichende Inhalte können zu Missverständnissen und Fehlentscheidungen führen. Die Entscheidungsphase ist besonders anfällig für Misstrauen und Ablehnung, wenn Empfehlungen und Hinweise nicht nachvollziehbar oder zu dominant sind – Nutzer verlieren ihre Entscheidungshoheit oder folgen Systemvorgaben unreflektiert. In der Handlungsausführung können komplexe oder fehleranfällige Bedienprozesse zu Fehlhandlungen, Systemabbrüchen und Frustration führen. Bleibt in der Feedbackphase die Rückmeldung unklar oder verspätet oder fehlt sie ganz, entsteht Unsicherheit über die eigene Leistung, was das Systemvertrauen schwächen und Lernprozesse hemmen kann. Insgesamt leitet Wandke (2005) konzeptionell ab, dass Überforderung, Fehlverhalten, Frustration, reduzierte Akzeptanz und langfristiges Misstrauen gegenüber dem System die Folge sind, wenn die phasenspezifischen Anforderungen an Darstellungs-, Informations- und Interaktionsqualität vernachlässigt werden. Das Modell von Wandke (2005) zeigt damit deutlich, dass Assistenz nur dann

wirksam ist, wenn die Qualitäten präzise auf den menschlichen Handlungsverlauf abgestimmt sind – andernfalls gefährden sie nicht nur die Effizienz, sondern auch die Akzeptanz technischer Systeme (Wandke, 2005).

Konzeptuelle Untersuchungen in der Mensch-Maschine-Interaktion: Das von Suh & Prophet (2018) entwickelte Framework zur Nutzung immersiver Technologien basiert auf dem Stimulus-Organismus-Response-Modell und bietet einen strukturierten theoretischen Ansatz zur Analyse der Wirkung technologischer Systeme auf Nutzererleben und Nutzungsergebnisse. Es unterteilt den Prozess der Technologieanwendung in drei zentrale Komponenten: Stimulus, Organismus und Response. Die Stimulus-Komponente beschreibt die externen Reize, die von der immersiven Technologie ausgehen. Dazu zählen sowohl technische Merkmale wie Immersionsgrad, Interaktivität, Systemlatenz oder Displayqualität als auch inhaltliche Aspekte wie Relevanz, Qualität und Narrative der virtuellen Inhalte. Diese Merkmale werden im Framework als Einflussgrößen auf die Organismus-Komponente, also die internen kognitiven und affektiven Reaktionen der Nutzer, eingeordnet. Hierzu zählen beispielsweise Aufmerksamkeit, wahrgenommene Kontrolle, mentale Belastung, emotionale Zustände, Nutzerzufriedenheit oder das Erleben eines Flow-Zustands. Die daraus resultierenden Response-Reaktionen umfassen die Nutzungsergebnisse, also etwa die Qualität der User Experience, die wahrgenommene Leistung, das Engagement sowie Verhaltensabsichten wie Akzeptanz oder zukünftige Nutzung. Das Modell verdeutlicht, dass zwischen den Eigenschaften immersiver Systeme, der subjektiven Verarbeitung durch den Nutzer und den daraus abgeleiteten Ergebnissen theoretisch systematische Zusammenhänge angenommen werden. Zudem berücksichtigt das Framework individuelle Unterschiede wie Alter, Vorerfahrung, technologische Affinität oder körperliche Voraussetzungen. Diese Merkmale wirken als Moderatoren, die beeinflussen, wie stark technologische und inhaltliche Stimuli auf die kognitiv-affektiven Nutzerreaktionen wirken und inwieweit diese Reaktionen zu bestimmten Nutzungsergebnissen führen. Das Modell bietet somit eine integrative Perspektive auf die Wirkmechanismen immersiver Technologien. Es verbindet technikorientierte und nutzerzentrierte Forschungsansätze und schafft eine theoretische Grundlage für die Erklärung sowohl positiver Effekte wie gesteigerter Lernleistung, Engagement oder Nutzerbindung als auch potenzieller negativer Effekte wie kognitiver Überlastung oder

Motion Sickness etc. Das Framework stellt jedoch keine eigenständige experimentelle Untersuchung konkreter Fehlerwirkungen dar (Suh & Prophet, 2018).

Die „Augmented Workforce Canvas“ von Moencks et al. (2022) ist ein strategisches Managementtool, das Unternehmen systematisch bei der menschenorientierten und wertgetriebenen Integration von Operator Assistenzsystemen unterstützt. Es gliedert den Prozess in vier Bausteine: Analyse des aktuellen Arbeitsprofils, Erfassung der Wertperspektiven verschiedener Stakeholder, Durchführung von Pilotversuchen und Tests sowie abschließende Bewertung und Anpassung. Dadurch wird eine nachhaltige, anwendungsorientierte und partizipative Einführung neuer Technologien in der Produktion ermöglicht. Im Rahmen dieses konzeptionellen Managementansatzes (Moencks et al., 2022) werden mögliche Folgen technischer Fehler flankierend beschrieben, jedoch nicht experimentell untersucht. Zu den genannten Auswirkungen zählen Zeitverluste durch Unterbrechungen, Qualitätsminderungen bei Produkten oder Dienstleistungen, erhöhte Kosten durch Nacharbeit oder Ausschuss, negative Einflüsse auf die Arbeitsmoral der Mitarbeiter sowie eine Beeinträchtigung der Kundenzufriedenheit und des Unternehmensrufs. Daher sind präventive Maßnahmen und eine sorgfältige Fehlerbewältigung zentral für den Erfolg der Technologieintegration (Moencks et al., 2022).

Auswirkungen der Interaktionsqualität bzw. -fehler: Die Studie von Cockburn et al. (2019) zeigt deutlich, dass Vibrationen die Interaktion mit Touchscreens in sicherheitskritischen Umgebungen wie Cockpits erheblich beeinträchtigen. Unter Vibrationsbedingungen stieg die False-negative-Fehlerquote bei der Single-Finger-Postur auf bis zu 50 %, konnte jedoch durch die stabilisierende Bracing-Haltung signifikant auf 22 % reduziert werden ($F(1;1897) = 216,8; p < 0,001$). Auch die Auswahlzeiten verlängerten sich unter Vibration deutlich, wobei Bracing diese im Vergleich zu Single-Finger signifikant verkürzte (z. B. Task 1: 1517 ms vs. 1280 ms; $F(1;1849) = 7,09; p = 0,01$). Fehlereingaben wie „wrong target“ oder „false-negative“ traten insbesondere bei der Force-Methode und der Single-Finger-Postur gehäuft auf (Fehlerraten bis zu 56 %) und können in sicherheitsrelevanten Kontexten kritische Systemzustände auslösen. Subjektive Rückmeldungen wiesen zudem auf Unterschiede in der wahrgenommenen Ermüdung und im Bedienaufwand hin: Während Single-Finger unter Vibration als ermüdend wahrgenommen wurde, bewerteten Teilnehmende Bracing als ergonomischer. Methoden wie

Dwell und Force erhöhten dabei zusätzlich den Kraftaufwand. Im freien Auswahl-Task bevorzugten Nutzer unter Vibration zu 75 % Bracing, was eine deutliche Nutzerpräferenz für die stabilisierte Interaktion unter Vibration erkennen lässt. Insgesamt belegt die Studie, dass die gezielte Gestaltung von Eingabemethoden und Haltungen entscheidend ist, um Genauigkeit, Reaktionszeit und Nutzerbelastung unter erschwerten Bedingungen zu optimieren (Cockburn et al., 2019).

Die Studie von Pourmemar (2024) zeigt, dass handgestenbasierte Eingabemethoden in immersiven AR-Umgebungen mit hohen Fehler- und Belastungswerten verbunden sind. Drop-down-Menüs mit Handgesten wiesen eine Auswahlfehlerrate von 10 % und eine Erkennungsfehlerrate von 31 % auf; bei radialen Menüs stieg die Auswahlfehlerrate auf 25 %, während die Erkennungsfehlerrate auf 21 % sank. Deutlich besser schnitt die sprachbasierte Eingabe bei Drop-down-Menüs ab: 0 % Auswahlfehler, 13 % Erkennungsfehler. Am leistungsfähigsten war jedoch die Kombination radiales Menü + Head Pointing mit nur 10 % Auswahlfehlern und 0 % Erkennungsfehlern. Die durchschnittliche Zeigezeit bei radialen Menüs mit Handgesten betrug 7,06 s ($SD = 2,20$ s). Negative Korrelationen zeigten sich zwischen körperlicher Belastung und kognitiver Leistung: $r = -0,324$ (gesamt), $r = -0,443$ (Symbolsuche). Auch die Zeigezeit korrelierte negativ mit der kognitiven Leistung ($r = -0,284$). Subjektiv wurden Handgesten als am belastendsten ($M = 4/5$) und frustrierendsten bewertet, während Head Pointing eine geringere Belastung ($M = 2/5$) und Sprache die niedrigste Belastung ($M = 1-1,5$) aufwies. ANOVA -Tests bestätigten signifikante Unterschiede hinsichtlich der körperlichen Belastung ($F(3;72) = 32,471$; $p < 0,05$) sowie hinsichtlich der Frustration ($F(3;72) = 70,795$; $p < 0,05$). Insgesamt wurde Head Pointing im untersuchten Vergleich subjektiv am günstigsten bewertet bei gleichzeitig niedriger Fehlerquote und Belastung (Pourmemar, 2024).

Die Studie von Graichen und Graichen (2023) untersuchte, inwieweit beobachtete Fehler in einer gestenbasierten Virtual-Reality-Anwendung das Vertrauen und die Akzeptanz der Nutzer beeinflussen. Im technologischen Kontext einer VR-Interaktion mittels Mid-Air-Gesten – realisiert durch ein Leap-Motion-System – wurde analysiert, wie sich Fehlfunktionen auf subjektive Bewertungen auswirken. Die Ergebnisse zeigen, dass zwischen der Anzahl der beobachteten Fehler und der allgemeinen Akzeptanz der Anwendung keine signifikante Korrelation bestand ($q = -0,19$; $p = 0,253$). Hingegen

korrelierte die Fehlerhäufigkeit signifikant negativ mit dem Nutzervertrauen ($\rho = -0,35$; $p = 0,041$). Damit kann auf Grundlage der Studie kein signifikanter direkter Akzeptanzverlust durch die beobachteten Fehler nachgewiesen werden. Nachweisbar ist jedoch eine Abnahme des Vertrauens bei zunehmender Fehlerhäufigkeit. Die Autoren betonen, dass Vertrauen in technische Systeme zwar typischerweise mit zunehmender Nutzung wächst, aber durch beobachtete Fehlfunktionen erodieren kann. Insgesamt verdeutlicht die Studie, dass Nutzer zwar grundsätzlich offen für innovative Interaktionsformen wie gestenbasierte VR-Anwendungen sind, aber empfindlich auf erlebte Systeminstabilität im Sinne eines Vertrauensverlusts reagieren.

Die Studie von Caputo et al. (2024) zeigt, dass die Art der gestenbasierten Interaktionsmethoden in immersiven Umgebungen einen wesentlichen Einfluss auf die Nutzerleistung und -wahrnehmung hat. Im Vergleich dreier Interaktionsmethoden – der Standardsteuerung des Mixed-Reality-Toolkits (MRTK), der Zwei-Punkt-Technik mit direkionaler Steuerung sowie einer symmetrischen beidhändigen Steuerung mit integrierter Kontrolle über alle sieben Freiheitsgrade (sogenannte „Handlebar“-Methode) – zeigte die Handlebar-Methode insbesondere bei objektiven Leistungsmaßen Vorteile. In Bezug auf die Effizienz war die symmetrische Methode der Konkurrenz deutlich überlegen: Die Aufgabenbearbeitung verlief signifikant schneller, und es wurden weniger Handbewegungen benötigt ($p < 10^{-32}$ bzw. $p < 10^{-76}$). Auch bei der Genauigkeit – insbesondere beim präzisen Ausrichten von Objekten – zeigte sich, dass erfahrene Nutzer bei den beiden anderen Methoden bessere Ergebnisse erzielten, während die symmetrische Methode unabhängig vom Erfahrungsstand geringere Fehler ermöglichte (Medianfehler: $0,17^\circ \pm 0,53$). Bei der groben Ausrichtungsaufgabe zeigten sich dagegen keine statistisch signifikanten Unterschiede hinsichtlich der allgemein wahrgenommenen Schwierigkeit und Ermüdung. Die Handlebar-Methode erreichte deskriptiv den höchsten mittleren System-Usability-Scale -Wert (SUS) von 72,8 Punkten, verglichen mit 67,5 Punkten bei der Standardsteuerung und 56,3 Punkten bei der Zwei-Punkt-Technik. Die Unterschiede der Gesamtwerte der System Usability Scale waren jedoch nicht signifikant ($p = 0,47$). Darüber hinaus zeigte sich, dass die symmetrische Methode signifikant leichter erlernbar war ($p = 0,01$) und deutlich weniger dominante Handbewegungen erforderte ($p < 10^{-27}$), was auf eine geringere physische Belastung und bessere Ergonomie hinweist.

Insgesamt zeigen die Ergebnisse, dass die symmetrische, visuell integrierte Steuerungsmethode insbesondere hinsichtlich Effizienz, Genauigkeit und Erlernbarkeit Vorteile aufweist. Subjektive Unterschiede bei Schwierigkeit, Ermüdung und der gesamten Usability-Bewertung ließen sich hingegen nicht durchgängig statistisch absichern. Die Studie unterstreicht damit die zentrale Bedeutung nutzerzentrierter Gestaltung und ergonomischer Steuerungskonzepte für die Entwicklung leistungsfähiger, intuitiv bedienbarer MR-Systeme ohne externe Eingabegeräte.

Einflüsse auf kognitive Belastungen: Die Studie von Tomori und Ogunseiju (2025) analysierte Situationsbewusstsein und kognitive Belastung bei 19 Studierenden in einer MR-Lernumgebung. Die Studie untersuchte keine gezielt implementierten Fehler, sondern explorativ die Zusammenhänge zwischen dem Situationsbewusstsein und verschiedenen Dimensionen der kognitiven Belastung. Die Vertrautheit mit der Situation ($M = 7,58$; $SD = 2,36$) korrelierte negativ mit mentaler Belastung ($r \approx -0,31$), zeitlicher Belastung ($r \approx -0,28$) und Anstrengung ($r \approx -0,26$) sowie positiv mit Entspannung ($r \approx 0,39$) und der subjektiven Bewertung der Leistung. Informationsqualität ($M = 7,68$; $SD = 1,67$) und Informationsmenge ($M = 7,58$; $SD = 1,83$) zeigten negative Zusammenhänge mit Frustration ($r \approx -0,36$ bzw. $-0,39$) und Anstrengung ($r \approx -0,23$). Aufmerksamkeit ($M = 7,79$; $SD = 1,75$) korrelierte positiv mit subjektiver Leistungsbewertung ($r \approx 0,39$). Diese Zusammenhänge waren überwiegend nicht statistisch signifikant. In der Korrelationsmatrix ist lediglich der positive Zusammenhang zwischen verfügbarer mentaler Kapazität und physischer Anforderung mit $r = 0,56$ als signifikant gekennzeichnet. Die Ergebnisse liefern somit explorative Hinweise darauf, dass die kognitive Belastung in MR-Umgebungen mit der Gestaltung der Informationspräsentation, der Nutzervertrautheit, den Anforderungen an die Aufmerksamkeit sowie der situativen Komplexität und Stabilität zusammenhängen kann. Die Autoren leiten daraus ab, dass effektives Training, gut gestaltete Informationsflüsse und Strategien zur Stress- und Frustrationsreduktion zur Optimierung des Situationsbewusstseins und der kognitiven Belastung beitragen können. Ein kausaler Einfluss von Informations- oder Trainingsfehlern auf Leistung oder Beanspruchung kann aus dem korrelativen Studiendesign jedoch nicht abgeleitet werden (Tomori & Ogunseiju, 2025).

Einfluss von Darstellungsqualitäten bzw. Visualisierungsfehlern: In der Studie von Alessa et al. (2023) mit 28 gesunden, männlichen Probanden, die in zwei Gruppen unterteilt wurden (AR-basierte vs. papierbasierte Instruktion), wurde ein $2 \times (2 \times 14)$ gemischtes Design angewandt. Die Teilnehmer führten einfache und komplexe Wartungsaufgaben in kontrollierter, balancierter Reihenfolge aus. Ziel war es, die Auswirkungen der Instruktionsmethode auf Leistung und Belastung zu untersuchen. Die Ergebnisse zeigen, dass die AR-basierte Anleitung bei komplexen Aufgaben zu einer signifikanten Reduktion der Aufgabendauer um 14,94 % führte (24,48 vs. 28,76 Minuten; $p < 0,01$), während bei einfachen Aufgaben kein signifikanter Unterschied festgestellt wurde. Auch hinsichtlich der subjektiven Arbeitsbelastung – gemessen mit dem NASA-TLX – zeigten sich teilweise signifikante Vorteile zugunsten der AR-Bedingung. Bei komplexen Aufgaben reduzierte AR die physische Belastung um etwa 12 % und die Frustration um rund 60 % und verbesserte gleichzeitig die Leistungsbewertung (p -Werte zwischen $< 0,05$ und $< 0,01$). Signifikante Interaktionen zwischen Instruktionsmethode und Aufgabenschwierigkeit wurden für physische Belastung und Frustration festgestellt ($p < 0,05$). Die physiologische Messung mittels Herzfrequenzvariabilität unterstreicht diese Ergebnisse: Unter papierbasierter Anleitung führten komplexe Aufgaben zu einer erhöhten mittleren Herzfrequenz (HR) und einer Reduktion von Herzfrequenzvariabilität (HRV) Maßen – ein Hinweis auf eine gesteigerte sympathische Aktivierung ($p < 0,05$). Im Gegensatz dazu blieben bei AR-Instruktion sowohl Herzfrequenz als auch HRV bei einfachen und komplexen Aufgaben stabil, was auf eine geringere physiologische Belastung hinweist. Zusammenfassend zeigen die statistischen Analysen, dass AR-basierte Instruktionen insbesondere bei komplexen Wartungsaufgaben zu einer kürzeren Bearbeitungszeit sowie zu einer Reduktion bestimmter physischer und subjektiver Belastungen führen. Die Studie untersucht allerdings keinen gezielt implementierten Visualisierungs- oder Sichtbarkeitsfehler, sondern vergleicht zwei Instruktionsmedien. Sie liefert daher nur eine indirekte Grundlage für die Bedeutung der Darstellungs- und Instruktionsqualität. Aufgrund ihres Fokus auf die Vermittlung von Handlungsinformationen lässt sie sich zugleich der Trainingsqualität zuordnen (Alessa et al., 2023).

Die empirische Studie von Sadovitch (2020) zur subjektiven Fehlerbewertung von AR-HMD untersucht den Einfluss visueller Fehlervariablen, insbesondere Positions-

und Orientierungsfehler, Latenz und Jitter, auf die subjektive Systembewertung. Die Untersuchung zeigte, dass eine Erhöhung der Fehlerausprägung bei den überwiegenden Fehlerarten zu einer signifikanten Reduktion der subjektiven Systemakzeptanz führte. Die Wirkung ist jedoch differenziert nach Fehlerart und Fehlerausprägung zu betrachten. Jitter hatte den geringsten Einfluss auf die Bewertung; zwischen den beiden untersuchten Ausprägungsstufen zeigten sich weder für die Nützlichkeit ($p = 1,000$) noch für die Zufriedenheit ($p = 0,180$) signifikante Unterschiede. Latenz reduzierte vor allem die hedonische Qualität signifikant, während laterale Orientierungsfehler auch die pragmatische Qualität und Nützlichkeit stark beeinträchtigten. Orientierungsfehler wurden signifikant störender bewertet als Positionsfehler ($p < 0,05$). Insgesamt belegen die Daten, dass höhere Fehlergrade die Akzeptanz und Wahrnehmung von AR-HMDs beeinträchtigen können, mit differenzierten Effekten je Fehlerart (Sadovitch, 2020).

Die Studie von Rosilius et al. (2022) zeigt, dass die Sichtbarkeit optotypischer Zeichen (Landolt-C) stark größenabhängig ist und im AR-System früher abnimmt als im realen Raum. Bereits bei Größen unter 10 mm sank die Sichtbarkeit in AR signifikant, während im realen System dieser Effekt erst unter 4 mm auftrat. So lag der Mittelwert des Sichtbarkeitsindex für 6 mm im realen System bei 21,24 %/s ($SD = 3,85$), verglichen mit 18,05 %/s ($SD = 5,74$) für 16 mm im AR-System. Die Ergebnisse belegen eine systematische Sichtbarkeitsminderung in AR, weshalb visuelle Elemente dort größer dargestellt werden müssen, um eine vergleichbare Sichtbarkeit wie in realen Umgebungen zu gewährleisten. Die Untersuchung bezieht sich damit unmittelbar auf die Sichtbarkeit, weist jedoch keine eigenständigen Effekte auf Akzeptanz oder Beanspruchung nach (Rosilius et al., 2022).

Die Studie von Merenda et al. (2019) untersucht den Einfluss visueller Fidelity – also der grafischen Detailgenauigkeit – auf Fahrleistung und Blickverhalten in simulierten Fahrscenarien mit AR-Head-Up-Displays. Ziel war es, herauszufinden, ob eine höhere visuelle Qualität (VQ) die Reaktionszeit, die Aufmerksamkeit und das Situationsbewusstsein beeinflusst. Die Ergebnisse zeigen, dass hohe visuelle Fidelity zu signifikant längeren Bremszeiten führte ($M = 1,402$ s vs. 1,321 s; $t(356) = 2,368$; $p = 0,018$), was eine Beeinträchtigung der objektiven Reaktionsleistung erkennen lässt. Auch das Blickverhalten veränderte sich deutlich: Fahrer benötigten bei hoher Fidelity längere Blickzeiten

auf Fußgänger ($t = 2,774$; $p = 0,006$) und richteten seltener den Blick auf die Straße ($t = -5,848$; $p < 0,001$). In den subjektiven Bewertungen des Situationsbewusstseins zeigten sich hingegen keine signifikanten Unterschiede ($p > 0,05$). Die Ergebnisse legen nahe, dass ein komplexeres, visuelles Umfeld die objektive Leistung und visuelle Aufmerksamkeit beeinflusst, ohne dass sich das subjektiv eingeschätzte Situationsbewusstsein entsprechend verändert (Merenda et al., 2019).

Die Studie von Arning et al. (2012) untersucht die Leistung, Nutzererfahrung und Akzeptanz zweier mobiler Indoor-Navigationssysteme – eines klassischen Bildschirms (z. B. Smartphone) und eines mobilen Projektors – unter besonderer Berücksichtigung von Faktoren wie technischer Selbstwirksamkeit, Sichtbarkeitsproblemen, Vertrauen und kognitiver Belastung. Die Navigationsleistung war beim mobilen Bildschirm signifikant höher als beim mobilen Projektor (M Bildschirm = 0,91; $SD = 0,09$ vs. M Projektor = 0,85; $SD = 0,10$; $F(1;23) = 12,5$; $p < 0,01$). Das bedeutet, dass die Nutzer beim Navigieren mit dem Projektor eine schlechtere Leistung zeigten. Ein weiteres, zentrales Ergebnis ist, dass die Akzeptanz für die Projektorbedingung insgesamt geringer ausfiel als für den Bildschirm. Konkret wurde in der Strukturgleichungsmodellierung für die Akzeptanz des Projektors ein Erklärungsanteil R^2 von 0,61 ermittelt, während er beim Bildschirm bei 0,50 lag. Dies deutet auf eine insgesamt höhere Varianzaufklärung durch die zugrunde liegenden Einflussfaktoren in Projektbedingungen hin, bei zugleich geringerer mittlerer Akzeptanz ($M = 5,25$; $SD = 0,71$) gegenüber dem Bildschirm ($M = 5,39$; $SD = 0,54$). Eine zentrale Einflussgröße für die Akzeptanz in der Projektorbedingung war das Vertrauen in das System. Dieses wurde durch die Sichtbarkeitsprobleme signifikant negativ beeinflusst ($\beta = -0,43$; $p < 0,01$), die im Mittel bei $M = 4,64$ ($SD \approx 1,00$) lagen – deutlich höher als beim Bildschirm ($M = 1,38$; $SD \approx 0,48$). Das Vertrauen wiederum zeigte einen klar positiven Effekt auf die Akzeptanz ($\beta = 0,49$; $p < 0,01$). Auch die Vertrautheit mit dem Navigationssystem hatte einen positiven, wenn auch indirekten Effekt auf das Vertrauen und darüber hinaus auf die Akzeptanz. Im Gegensatz dazu hatte die Sichtbarkeit keinen direkten Einfluss auf die wahrgenommene Leistung ($\beta = -0,08$; $p = 0,24$), die unter der Projektorbedingung bei $M = 0,85$ ($SD = 0,10$) und beim der Bildschirmbedingung bei $M = 0,91$ ($SD = 0,09$) lag. Zusätzlich ergaben sich Unterschiede in der kognitiven Belastung: Die durchschnittliche Arbeitsbelastung lag beim Projektor bei $M = 13,14$ ($SD = 9,45$)

und war damit höher als beim Bildschirm ($M = 10,0$; $SD = 9,81$). Dennoch zeigte sich, dass der Workload keinen direkten Einfluss auf die Akzeptanz hatte. Vielmehr wurde die Akzeptanz in der Projektorbedingung primär durch Wahrnehmungsfaktoren wie Vertrauen, Sichtbarkeitsprobleme und subjektive Nutzeneinschätzung bestimmt (Arning et al., 2012).

Einfluss von Automatisierungsfehlern sowie Transparenz- bzw. Konfidenzanzeigen: In der Studie von Konopka et al. (2024b) wurde ein AR-basiertes Entscheidungsunterstützungssystem (DSS) im Kontext eines Blackjack-Spiels eingesetzt, um die Auswirkungen gezielt implementierter Systemfehler und Gestaltungsfaktoren auf Vertrauen, Nutzerleistung und Akzeptanz zu untersuchen. Das AR-DSS war so konzipiert, dass es unterschiedliche Fehlfunktionen simulierte, darunter variierende Erkennungsgenauigkeit bei der Kartenerfassung (60 % vs. 90 %) sowie daraus resultierende fehlerhafte Handlungsempfehlungen. Ergänzend wurden unterschiedliche Stufen systemischer Transparenz hinsichtlich der Empfehlungserklärungen (L1–L3) sowie die Schwere der Entscheidungsauswirkungen (hohe vs. niedrige Einsätze) experimentell variiert. Das Design folgte einem dreifaktoriellen Between-Subjects-Ansatz mit zwölf experimentellen Bedingungen. Die Ergebnisse zeigten, dass eine höhere Systemperformance signifikant mit einem höheren Vertrauen in das AR-DSS einherging. Ebenso erhöhte eine höhere Transparenzstufe die Vertrauenswerte der Nutzer signifikant. Das Systemvertrauen hatte jedoch keinen signifikanten Einfluss auf die Aufgabenleistung, da vor allem individuelle Expertise der Teilnehmenden im Blackjack einen stärkeren Effekt auf die Performance zeigte. Entscheidend für die Akzeptanz des Systems war eine realistische Kalibration des Vertrauens: Nutzer, die die Leistungsgrenzen des Systems angemessen einschätzten, zeigten eine signifikant höhere Bereitschaft, das AR-DSS zukünftig weiterzuverwenden. Die Ergebnisse unterstreichen die Bedeutung transparenter Gestaltung und gezielter Vertrauenskalibrierung in AR-Systemen, insbesondere im Umgang mit potenziellen Systemfehlern (Konopka et al., 2024b). Die variierten Transparenzstufen sind dabei als unterschiedliche Erklärungstiefen und nicht als reine Konfidenzanzeige einzuordnen.

Die Studie von Konopka et al. (2024a) untersucht, inwieweit absichtlich integrierte Fehlfunktionen in einem AR-DSS das Vertrauen der Nutzer, ihre emotionalen

Reaktionen sowie ihre Nutzungsabsichten beeinflussen. Die Fehlfunktionen bezogen sich dabei konkret auf gezielt fehlerhafte Handlungsempfehlungen, die das AR-DSS den Teilnehmenden bereitstellte. Im Rahmen eines experimentellen Designs wurden 79 Probanden mit einem solchen System konfrontiert, wobei die Fehlfunktionen in ihrer Schwere (gering vs. gravierend) variierten. Die Ergebnisse zeigen, dass mit zunehmender Fehlerschwere das Vertrauen signifikant abnimmt. Zudem führen die fehlerhaften Anweisungen zu einem wahrgenommenen Vertrauensbruch und verstärken emotionale Reaktionen, die sich ebenfalls negativ auf das Vertrauen auswirken. Besonders bei schwerwiegenden Fehlern zeigte sich ein direkter negativer Einfluss auf das Vertrauen. Auch die Nutzungsabsicht sank tendenziell mit zunehmender Fehleranzahl, wenngleich nicht alle Effekte statistisch signifikant ausfielen. Auffällig war, dass viele Teilnehmende die fehlerhaften Anweisungen nicht explizit als Systemversagen erkannten oder die Verantwortung bei sich selbst suchten – ein möglicher Hinweis auf Übervertrauen in AR-Technologien (Konopka et al., 2024a).

Die Studie von Gegoff et al. (2023) zeigt, dass sowohl Automationszuverlässigkeit als auch Transparenz signifikante Haupteffekte auf Vertrauen, die Nutzung automatisierter Systeme und Usability haben. Hohe Zuverlässigkeit führte zu signifikant höheren Trefferquoten ($M = 0,93$ vs. $0,88$; $F = 26,4$; $p < 0,001$), stärkerem Vertrauen ($M = 3,14$ vs. $2,30$; $F = 65,5$; $p < 0,001$) und besseren Usability-Bewertungen ($M = 68,7$ vs. $56,3$; $F = 33,1$; $p < 0,001$). Erhöhte Transparenz verkürzte signifikant die Entscheidungszeit ($M = 13,5$ s vs. $18,0$ s; $F = 33,1$; $p < 0,001$), senkte die subjektive Arbeitsbelastung ($M = 53,0$ vs. $57,6$; $F = 4,74$; $p = 0,03$) und förderte die angemessene Annahme beziehungsweise Zurückweisung automatisierter Empfehlungen. Eine statistisch signifikante Interaktion zwischen Transparenz und Zuverlässigkeit wurde jedoch nicht gefunden. Innerhalb der Stichprobe ließ sich aber beobachten, dass höhere Transparenz bei niedriger Zuverlässigkeit mit einer erhöhten Annahme korrekter automatisierter Ratschläge einherging, was auf eine Minderung des typischen Disuse-Effekts hindeutet. Der Disuse-Effekt bedeutet, dass Bediener automatisierte Systeme oder deren Ratschläge zunehmend ablehnen oder ignorieren, wenn sie deren Zuverlässigkeit als niedrig wahrnehmen (Gegoff et al., 2023).

Die Studie von Hecht et al. (2020) untersuchte im Fahrsimulator den Einfluss unterschiedlicher Häufigkeiten notwendiger Übernahmeaufforderungen, beispielsweise

aufgrund von Automatisierungsfehlern beim automatisierten Fahren, sowie eines prädiktiven HMI auf Arbeitsbelastung, Akzeptanz, Benutzerfreundlichkeit und Vertrauen. Die Ergebnisse zeigen, dass eine höhere Häufigkeit an Übernahmeaufforderungen (höhere Fehlerhäufigkeit) die Arbeitsbelastung signifikant erhöhte ($F(1;33) = 46,85; p < 0,001; \eta^2 = 0,602$) und die Akzeptanz deutlich verringerte ($\chi^2(2) = 49,115; p < 0,001$). Auch die Benutzerfreundlichkeit nahm mit steigender Frequenz ab. Das Vertrauen in das System blieb davon unbeeinflusst ($\chi^2(2) = 4,160; p = 0,125$). Der Einsatz des prädiktiven HMI hatte jedoch keinen signifikanten Einfluss auf die Arbeitsbelastung ($F(1;31) = 0,81; p = 0,809$), Akzeptanz oder Usability. Auch eine signifikante Verbesserung der Leistung oder Reduktion der Frustration wurde nicht festgestellt. Zusammenfassend erhöhen häufige Übernahmeaufforderungen die Belastung und verringern die Akzeptanz, während das prädiktive HMI keine nachweisbare Entlastung oder Leistungssteigerung bewirkt. Das prädiktive HMI ist allerdings nicht mit einer klassischen Anzeige der numerischen Systemkonfidenz gleichzusetzen.

Die Studie von McGuirl und Sarter (2006) belegt, dass Piloten die Zuverlässigkeit automatisierter Systeme signifikant besser einschätzen können, wenn sie kontinuierlich aktualisierte Informationen zur Konfidenz bzw. Zuverlässigkeit des Systems erhalten. In einem simulierten vereisungsbedingten Notfallszenario zeigte sich, dass Teilnehmer mit dynamischen Vertrauensanzeigen im Mittel nur 1,2 Flugzeugvereisungen pro 28 Flüge erlebten, verglichen mit 3,4 Flugzeugvereisungen in der Kontrollgruppe mit statischer Anzeige ($p < 0,01$; Odds Ratio = 2,01; 95 % CI = 1,36–2,95). Zudem waren Piloten in der dynamischen Bedingung signifikant eher bereit, eine fehlerhafte Erstreaktion zu revidieren: 45 % passten ihr Verhalten bei ineffektiven Reaktionen an, während dies in der Kontrollgruppe nur in 20 % der Fälle geschah (χ^2 -Test; $p < 0,05$). Auch die subjektive Einschätzung der Systemgenauigkeit war in der dynamischen Gruppe deutlich realistischer: Die Abweichung vom tatsächlichen Wert lag bei ± 5 %, gegenüber ± 15 % in der statischen Gruppe (t -Test; $p < 0,01$). Diese Ergebnisse verdeutlichen, dass dynamisch vermittelte Vertrauensinformationen Piloten dabei unterstützen, ihr Vertrauen adaptiv an die tatsächliche Systemleistung anzupassen. Sie unterstützen eine angemessene Kalibrierung des Vertrauens und fördern gleichzeitig kritisches Verhalten bei sinkender Systemgenauigkeit. Insgesamt trägt dies zu einer besseren Mensch-Automation-

Koordination bei und reduziert das Risiko fehlangepasster Automationsnutzung (McGuirl & Sarter, 2006). Die Studie untersucht damit primär Vertrauenskalibrierung und angemessene Systemnutzung, nicht unmittelbar Akzeptanz im Sinne eines Technologieakzeptanzmodells.

Die Studie von Richardson et al. (2018) zeigt, dass die Integration zusätzlicher Informationen im HMI hochautomatisierter LKW-Systeme signifikant positive Effekte auf die Fahrer-Automation-Interaktion hat. Die subjektive Akzeptanz erhöhte sich deutlich, insbesondere hinsichtlich der Zufriedenheit mit dem System, wobei Spearman-Korrelationen zwischen $r = 0,579$ und $0,800$ auf einen starken Zusammenhang zwischen Informationsangebot und Bewertung hinweisen. Auch die wahrgenommene Arbeitsbelastung (erhoben mit DALI) konnte signifikant reduziert werden – insbesondere in der Subskala von situativer Stress ($p = 0,025$), was auf eine geringere Beanspruchung bei Aufgabeninterferenzen und Stresssituationen schließen lässt (Richardson et al., 2018). Die Untersuchung stellt dabei eine Studie zur Informationsdarstellung dar und keine unmittelbare Manipulation eines Systemfehlers.

Die Studie von Brauner et al. (2019) untersuchte experimentell, wie sich die Korrektheit eines DSS auf Leistung, Vertrauen, Akzeptanz und Compliance der Nutzer auswirkt. Grundlage war ein innerhalb der Personen angelegtes Design. In zwei aufeinanderfolgenden Durchläufen spielten die Teilnehmenden ein identisches Geschäftssimulationsspiel – jeweils einmal mit einem korrekt funktionierenden und einmal mit einem defekten DSS. Die Zuteilung der DSS-Zustände erfolgte randomisiert, sodass Reihenfolgeeffekte kontrolliert und individuelle Unterschiede reduziert werden konnten. Die Teilnehmer wurden nicht darüber informiert, ob das System in einer Runde korrekt arbeitete oder fehlerhafte Vorschläge lieferte. In der ersten Runde führte das defekte DSS zu einer signifikanten Leistungseinbuße: Der Mediangewinn lag bei 10,350 Punkten gegenüber 13,350 Punkten bei korrekt arbeitendem DSS ($p = 0,027$). Die Leistungsdifferenz von rund 38 % wurde auch durch eine ANOVA bestätigt ($F(1;38) = 4,110$; $p = 0,050$). In der zweiten Runde, bei der die DSS-Zustände getauscht wurden, verringerte sich die Differenz (Median: korrekt = 12,30 vs. defekt = 11,650), was auf Lern- oder Kompensationseffekte hindeutet. Der Unterschied war hier nur tendenziell signifikant ($p = 0,056$) und in der Varianzanalyse nicht mehr bedeutsam ($F(1;38) = 1,178$; $p = 0,285$). Die

subjektive Bewertung des DSS war ebenfalls deutlich von dessen Zuverlässigkeit abhängig. Eine multivariate Varianzanalyse zeigte signifikante Effekte auf die wahrgenommene Nützlichkeit, Benutzerfreundlichkeit und Nutzungsintention (Pillai's $V = 0,644$; $F(4;33) = 14,941$; $p < 0,001$). Die Nützlichkeit sank beim defekten DSS im Mittel um 57 % ($F(1;36) = 41,353$; $p < 0,001$), die Nutzungsintention um 62 % ($p < 0,001$), und auch die Benutzerfreundlichkeit wurde signifikant schlechter bewertet ($p < 0,05$). Auch das Verhalten der Nutzer in Bezug auf die Einhaltung der DSS-Empfehlungen (Compliance) war abhängig von der Systemzuverlässigkeit. Während in der ersten Runde keine Unterschiede festgestellt wurden ($F(1;36) = 0,008$; $p = 0,928$), war die berichtete Compliance in der zweiten Runde signifikant höher bei funktionierendem DSS ($F(1;33) = 4,955$; $p = 0,033$). Zudem zeigten objektive Verhaltensdaten, dass Teilnehmende beim fehlerhaften DSS deutlich häufiger von den Vorschlägen abwichen (Median: defekt = 15 vs. korrekt = 8,5; $p = 0,025$). Zwischen Compliance und Leistung unter defekten Bedingungen bestand eine starke negative Korrelation ($\rho = -0,668$; $p < 0,001$), was darauf hinweist, dass die Befolgung fehlerhafter Empfehlungen die Leistung maßgeblich verschlechterte. Das Vertrauen in das DSS war ebenfalls signifikant durch die Korrektheit beeinflusst. In Runde 1 zeigte sich ein tendenzieller Effekt zugunsten des funktionierenden DSS ($F(1;37) = 3,401$; $p = 0,073$), der in Runde 2 statistisch signifikant ausfiel ($F(1;37) = 19,763$; $p < 0,001$; $\eta^2 = 0,348$). Der Vertrauenswert stieg von 28 bei defektem auf 53 bei korrektem DSS. Insgesamt belegt die Studie, dass die Systemzuverlässigkeit einen zentralen Einfluss auf Leistung, Vertrauen, Akzeptanz und Compliance hat. Das eingesetzte Zweirunden-Design ermöglicht eine differenzierte Analyse dieser Effekte bei derselben Person und stärkt somit die Aussagekraft der Ergebnisse. Fehlerhafte DSS führen nicht nur zu messbaren Leistungseinbußen, sondern auch zu einem signifikanten Rückgang der subjektiven Bewertung und des Vertrauens. Diese Ergebnisse unterstreichen die hohe Bedeutung der Systemqualität für die Mensch-Technik-Interaktion (Brauner et al., 2019).

Die Metaanalyse von Schaefer et al. (2016) zeigt, dass Informationen über Systemleistung, Zuverlässigkeit und Status relevante Gestaltungsfaktoren für die Vertrauensbildung in automatisierte Systeme sind. Eine transparente und konsistente Rückmeldung über die Systemleistung – insbesondere über die Genauigkeit und Zuverlässigkeit der Automation – kann zu einer angemesseneren Vertrauensbildung beitragen. Klare

und nachvollziehbare Informationen ermöglichen eine realistische Einschätzung der Systemfähigkeit und reduzieren Unsicherheiten sowie potenzielle Fehlinterpretationen. Darüber hinaus wirken sich konsistente und verständliche Statusanzeigen positiv auf die Wahrnehmung der Systemtransparenz aus, während widersprüchliche oder unklare Darstellungen das Vertrauen verringern können. Die Verfügbarkeit akkurater Feedbackinformationen unterstützt nicht nur die Vertrauensentwicklung, sondern fördert auch die Akzeptanz und Nutzungshäufigkeit der Automationssysteme. Dieser Zusammenhang ist besonders ausgeprägt in komplexen oder sicherheitskritischen Aufgaben, in denen Nutzer stärker auf verlässliche Systemrückmeldungen angewiesen sind. Insgesamt verdeutlichen die Ergebnisse, dass eine gut gestaltete Anzeige der Systemgenauigkeit ein effektives Mittel ist, um Vertrauen in automatisierte Systeme aufzubauen, die Nutzung zu fördern und die Interaktionssicherheit zu verbessern (Schaefer et al., 2016).

Die Studie von Wanner et al. (2022) zeigt, dass die Akzeptanz intelligenter Systeme im realitätsnahen Kontext eines industriellen Wartungsarbeitsplatzes von Leistungserwartung, Transparenz und Vertrauen beeinflusst wird. Im erweiterten UTAUT-Modell zeigte sich, dass Systemtransparenz die wahrgenommene Leistungsfähigkeit ($\beta = 0,346$; $p < 0,001$), die Aufwandserwartung ($\beta = 0,539$; $p < 0,001$) und vor allem das Vertrauen in die Systemfähigkeit ($\beta = 0,610$; $p < 0,001$) positiv beeinflusst. Das Vertrauen in die Systemfähigkeit wirkte sich wiederum positiv auf die Vertrauensneigung gegenüber dem intelligenten System aus ($\beta = 0,645$; $p < 0,001$). Darüber hinaus wurde ein signifikanter indirekter Effekt der Systemtransparenz auf die Nutzungsabsicht nachgewiesen ($\beta = 0,278$; $p < 0,001$). Die Leistungserwartung stellte einen bedeutsamen direkten Prädiktor der Verhaltensabsicht dar ($\beta = 0,313$; $p < 0,001$), während die Aufwandserwartung nur indirekt über die Einstellung wirkte. Die Einstellung gegenüber der Technologie beeinflusste die Verhaltensabsicht direkt ($\beta = 0,348$; $p < 0,001$). Zugleich zeigte sich ein indirekter Effekt der Systemtransparenz auf die Einstellung ($\beta = 0,441$; $p < 0,001$). Insgesamt war die Akzeptanz zunächst stark leistungsgetrieben; Transparenz nahm jedoch eine wichtige ergänzende und überwiegend indirekte Rolle für die Vertrauensbildung, die wahrgenommene Systemleistung und die Nutzungsabsicht ein. Eine nachvollziehbare Systemfunktionalität stellt daher einen relevanten Gestaltungsfaktor für die Einführung intelligenter Assistenzsysteme dar (Wanner et al., 2022).

Die Studie von Lorrig & Daw (2024) untersucht die Potenziale und Herausforderungen der Mensch-KI-Zusammenarbeit (Human-AI Teaming, HAIT) im Kontext zukünftiger Flugdeckoperationen, insbesondere im Hinblick auf Minimal- oder Ein-Pilot-Besatzungen. Ziel ist es, durch KI-gestützte Systeme die Effizienz zu steigern, Kosten zu senken und Piloten bei steigender Arbeitslast – insbesondere in Notfallsituationen – durch schnelle und präzise Informationsverarbeitung zu unterstützen. Die Autoren betonen dabei zentrale Risiken durch Systemfehler: Wiederholte Fehlfunktionen oder unklare Empfehlungen können zu Vertrauensverlust oder – im umgekehrten Fall – zu übermäßiger Abhängigkeit führen, die die Situationswahrnehmung und das manuelle Können der Piloten beeinträchtigen. Um diesen Risiken entgegenzuwirken, fordern sie transparente, interpretierbare und fehlertolerante KI-Systeme, die menschliche Eingriffe jederzeit ermöglichen (Lorrig & Daw, 2024). Die dargestellten Fehlerfolgen sind als konzeptionelle Risikoanalyse und nicht als Ergebnis einer eigenen empirischen Fehlerstudie einzuordnen.

Die Studie von Madhavan et al. (2006) untersucht, wie Fehler automatisierter Hilfsmittel das Vertrauen der Nutzer beeinflussen – insbesondere bei Aufgaben, die Menschen als einfach empfinden. In einem Experiment mit drei Gruppen zeigte sich, dass Teilnehmer deskriptiv weniger Vertrauen in das System hatten, wenn es bei einfachen Aufgaben Fehler machte (Easy-Miss-Gruppe: $M = 2,67$), verglichen mit Fehlern bei schwierigen Aufgaben (Difficult-Miss-Gruppe: $M = 3,87$; $t(28) = 1,71$; $p = 0,07$). Der Unterschied erreichte damit das übliche Signifikanzniveau von $\alpha = 0,05$ nicht. Die Easy-Miss-Gruppe erhielt eine Automationshilfe, die zwar bei komplexen Aufgaben korrekt arbeitete, aber sichtbare Fehler bei leichten Aufgaben produzierte – etwa das Verfehlen eindeutig erkennbarer Ziele. Auch die geschätzte Zuverlässigkeit war in dieser Gruppe signifikant geringer (53,2 % vs. 69,7 %; $t(28) = 2,57$; $p < 0,05$), obwohl die tatsächliche Systemgenauigkeit in schwierigen Aufgaben hoch war. Zudem zeigte die Easy-Miss-Gruppe ein überhöhtes Selbstvertrauen ($M = 3,36$), widersprach der korrekten Automation bei schwierigen Aufgaben häufiger und erzielte dadurch eine geringere Gesamtleistung. Die Studie zeigt, dass nicht allein die Fehlerrate, sondern insbesondere die Wahrnehmung der Fehlerart entscheidend für Vertrauen, Nutzung und Leistung in der Mensch-Automation-Interaktion ist – wobei Fehler bei einfachen Aufgaben das

Vertrauen besonders stark beeinträchtigen können, zu überschätztem Selbstvertrauen führen und die Wirksamkeit automatisierter Systeme deutlich mindern, selbst bei objektiv hoher Systemleistung (Madhavan et al., 2006).

Auswirkungen der Qualitätsfaktoren bzw. Fehler innerhalb der MMI: Die Studie von Nasirian et al. (2017) untersucht, welche Qualitätsfaktoren die Akzeptanz KI-basierter Sprachassistenzsysteme beeinflussen. Das auf dem „Information Systems Success Model“ basierende Strukturmodell weist eine erklärte Varianz von $R^2 = 54,9\%$ für die Nutzungsabsicht auf und berücksichtigt insbesondere Interaktionsqualität, Systemqualität (Stabilität bzw. Systemleistung), Informationsqualität (Richtigkeit der Informationen) sowie Vertrauen. Die Interaktionsqualität – also die Fähigkeit des Systems, als dialogfähiger Kommunikationspartner zu agieren – ist der stärkste Prädiktor für Vertrauen. Der Pfad Interaktionsqualität auf Vertrauen ist signifikant positiv ($p < 0,01$), ebenso der Pfad Vertrauen auf Nutzungsabsicht ($p < 0,01$). Die Interaktionsqualität wirkt somit indirekt (d. h. über Vertrauen) signifikant auf die Nutzungsabsicht (Akzeptanz). Im Gegensatz dazu zeigten weder die Informationsqualität noch die Systemqualität signifikante direkte oder indirekte Effekte: Die Pfade Informationsqualität zu Vertrauen und Systemqualität zu Vertrauen waren nicht signifikant ($p > 0,05$). Entsprechend konnten auch keine signifikanten Einflüsse dieser Faktoren auf die Nutzungsabsicht nachgewiesen werden. Die Autoren interpretieren System- und Informationsqualität als Hygienefaktoren im Sinne der Zwei-Faktoren-Theorie nach Herzberg (2015) – als notwendig, aber nicht vertrauensfördernd. Zusammenfassend lässt sich sagen, dass Vertrauen der zentrale vermittelnde Mechanismus für die Nutzungsabsicht von Sprachassistenten ist. Es wird signifikant durch die Interaktionsqualität gefördert, nicht jedoch durch System- oder Informationsqualität. Die Studie basiert allerdings auf einem beobachtenden Strukturmodell und enthält keine experimentelle Manipulation von Interaktions-, Daten- oder Systemfehlern. Aus den nicht signifikanten Pfaden kann daher nicht geschlossen werden, dass konkrete Daten- oder Systemfehler grundsätzlich keine Akzeptanzwirkungen besitzen.

Auswirkungen der Sprachinteraktion bei immersiven Systemen:

(Teile dieses Abschnitts basieren auf der Veröffentlichung: Rosilius, M., Spiertz, M., Wirsing, B., Geuen, M., Bräutigam, V. & Ludwig, B. (2024). Impact of Industrial Noise on

Speech Interaction Performance and User Acceptance when Using the MS HoloLens 2. *Multimodal Technologies and Interaction*, 8(2), 8. <https://doi.org/10.3390/mti8020008>. Die Inhalte, Abbildungen und Tabellen wurden für die vorliegende Monographie übernommen bzw. teilweise überarbeitet und in adaptierter Form eingebracht. Der Originalartikel wurde als Open-Access-Beitrag unter der Lizenz Creative Commons Attribution 4.0 International (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0/>) veröffentlicht, die die Bearbeitung und Wiederverwendung unter Angabe der ursprünglichen Quelle gestattet.)

Die ASR-Technologie ermöglicht es, die MMI beim industriellen Einsatz von AS zu vereinfachen. Durch die Vermeidung von manuellen Eingaben können beim Einsatz eines AS mehrere Aufgaben parallel ausgeführt werden. Laut Vajpai und Bora (2016) steigt die Zahl der ASR-Implementierungen stetig. Damit gewinnt ASR als alternative Interaktionsform innerhalb AS in verschiedenen Industriebereichen an Bedeutung. Heutzutage wird die ASR-Technologie bereits in verschiedenen Bereichen eingesetzt und ist weithin akzeptiert, z. B. in der Medizin, der Industrie, der Forensik, der Strafverfolgung, der Verteidigung, der Luft- und Raumfahrt, der Telekommunikation, der Heilautomatisierung, der Zugangskontrolle und der Unterhaltungselektronik etc. (Vajpai & Bora, 2016).

Die Spracheingabe kann die Ermüdung der Arbeitenden verringern und die Geschwindigkeit und Flexibilität der Befehlsübermittlung erhöhen, da mit den Händen gleichzeitig andere Aufgaben erledigt werden können (Vajpai & Bora, 2016). Dadurch können manuelle Tätigkeiten parallel ausgeführt werden. Die Sprachinteraktion kann u. a. bei der Überwachung und Bedienung der Maschinen angewendet werden, sodass die zugewiesenen Aufgaben leichter, schneller und mit weniger Aufwand erledigt werden können (Derby et al., 2019). ASR unterstützt auch leistungseingeschränkte Mitarbeiter und kann als hilfreiches Werkzeug zu Inklusionsbemühungen beitragen (Vajpai & Bora, 2016).

Als alternatives Interaktionsparadigma zur Gestensteuerung kann ASR auf der HL2 eingesetzt werden (Eckert et al., 2018; Estrella & Mendoza, 2019). Aufgrund der geringen gestenbasierten Eingabegeschwindigkeit stellt ASR eine potenziell effizientere Alternative für Texteingaben und Sprachbefehle dar. Laut Derby et al. (2019) wurde via

Gestensteuerung auf der HL1 lediglich ein Mittelwert von 5,41 Wörtern pro Minute gemessen. ASR ist nicht nur auf die immersive Anwendung selbst beschränkt, sondern ermöglicht auch die Steuerung von Produktionsmaschinen und Robotern. Sprachbefehle (SPC) können im industriellen Kontext z. B. für "Produktion starten oder stoppen", das Quittieren oder das Navigieren von Menüstrukturen innerhalb von Bedienoberflächen herangezogen werden (Fuhrmann et al., 2021). Diese Quellen belegen primär Anwendungsmöglichkeiten und technische Machbarkeit, nicht jedoch eigenständig die Auswirkungen von ASR-Fehlern auf Leistung, Akzeptanz oder Beanspruchung.

In einschlägigen Studien wurden bereits verschiedene ASR-Implementierungen mittels Audiodateien aus unterschiedlichen Quellen an Teilnehmern anhand der Metrik Word-Error-Rate (WER) getestet (Filippidou & Moussiades, 2020; Képuska, 2017; Liu et al., 2022). Diese Untersuchungen sind vor allem als methodische und technische Benchmarks der Erkennungsleistung einzuordnen. ASR wurde bereits in industrienahen Studien untersucht, jedoch bei Schalldruckpegeln (SPL) größer 81 dB als nicht einsatzfähig eingestuft. Marklin et al. (2022) zeigten, dass bei SPL zwischen 81 dB und 84 dB 90 % der SPC nicht empfangen wurden. Die SPC-Fähigkeiten wurden auch hinsichtlich der Hardware HL2 bewertet, jedoch nur unter abstrahierten Bedingungen. Außerdem wurde in diesen Studien lediglich die Richtigkeit gemessen, die effektive Informationsmenge pro Zeit wurde nicht berücksichtigt (Liu et al., 2022). Auf der Hardware HL1 wurden Untersuchungen an sechzehn englischsprachigen Probanden in Bezug auf die abhängigen Variablen (AV), WER, wahrgenommene Arbeitsbelastung und wahrgenommene Benutzerfreundlichkeit (SUS-Fragebogen) in Abhängigkeit des SPL durchgeführt (Derby et al., 2020). Bei der WER und der wahrgenommenen Benutzerfreundlichkeit wurden im Gegensatz zur wahrgenommenen Belastung keine signifikanten Leistungsunterschiede gemessen. Als Hintergrundgeräusch wurde weißes Rauschen mit 40, 55 und 70 dB eingesetzt, wodurch industrienaher Bedingungen nur eingeschränkt abgebildet wurden. Zwischen den Lärmbedingungen zeigten sich keine signifikanten Unterschiede hinsichtlich der Spracheingabegeschwindigkeit und der wahrgenommenen Arbeitsbelastung. Dagegen verschlechterten sich die Erkennungsgenauigkeit und die wahrgenommene Benutzerfreundlichkeit in der 70-dB-Bedingung. Aufgrund technischer Ausfälle konnten lediglich sechs der 16 Teilnehmer diese Bedingung vollständig absolvieren. Basierend auf

dem Forschungsdesign von Derby et al. (2020) wurde eine Studie (mit $N = 16$ Teilnehmern) auch auf der Hardware HL2 durchgeführt (Sinlapanuntakul et al., 2021). Die mittlere WER wurde bei verschiedenen SPL-Bedingungen mit 3,6 % (40 dB), 6,8 % (55 dB) und 4,8 % (70 dB) gemessen (Sinlapanuntakul et al., 2021). Es wurden jedoch keine statistisch signifikanten Unterschiede ($p > 0,05$) zwischen den SPL Bedingungen nachgewiesen. Außerdem konnten statistisch signifikante Unterschiede hinsichtlich der Spracheingabegeschwindigkeit (Mittelwertdifferenz = 3,89 wpm; $p = 0,038$) sowie der mentalen und zeitlichen Beanspruchung aufgezeigt werden (Sinlapanuntakul et al., 2021). Da die Texteingabegeschwindigkeit auf der (HL2 im Mittel bei 37,44 Wörtern pro Minute lag, zeigte sich deskriptiv eine etwa siebenfach höhere Eingabegeschwindigkeit als bei der zuvor untersuchten gestenbasierten Texteingabe auf der HoloLens 1. Daraus kann jedoch nicht unmittelbar auf eine siebenfach erhöhte Technologiereife geschlossen werden. Für die wahrgenommene Systembenutzbarkeit wurden keine statistisch signifikanten Unterschiede in der gegebenen Versuchsanordnung gefunden (Sinlapanuntakul et al., 2021).

Auswirkungen von Trainingsmethoden bei immersiven Systemen:

(Teile dieses Abschnitts basieren auf der Veröffentlichung: Rosilius, M., Hügel, L., Wirsing, B., Geuen, M., von Eitzen, I., Bräutigam, V. & Ludwig, B. (2024). Development and Evaluation of Training Scenarios for the Use of Immersive Assistance Systems. *Applied System Innovation*, 7(5), 73. <https://doi.org/10.3390/asi7050073>. Die Inhalte, Abbildungen und Tabellen wurden für die vorliegende Monographie übernommen bzw. teilweise überarbeitet und in adaptierter Form eingebracht. Der Originalartikel wurde als Open-Access-Beitrag unter der Lizenz Creative Commons Attribution 4.0 International (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0/>) veröffentlicht, die die Bearbeitung und Wiederverwendung unter Angabe der ursprünglichen Quelle gestattet.)

Trotz erheblicher Investitionen in Informations- und Kommunikationstechnologien haben sich in Gänze die erwarteten Produktivitätssteigerungen, gemessen am Ertrag pro Arbeitskraft oder Ertrag pro Kapitaleinsatz, nicht eingestellt. Entsprechend dem wissenschaftlichen Diskurs eines aktuellen Metareviews (Schweikl & Obermaier, 2020) über Auswirkungen, Ursachen und soziotechnische Zusammenhänge lässt sich die Problematik des Produktivitätsparadoxons grundsätzlich auch auf die Einführung immersiver

Assistenzsysteme übertragen. Die Forschung der letzten drei Jahrzehnte legt nahe, dass "Anpassungsverzögerungen, Messprobleme, überzogene Erwartungen und Missmanagement" (Schweikl & Obermaier, 2020) mögliche Hauptursachen für die jüngste Verlangsamung des Produktivitätswachstums sind. Mit einem kritischen Blick auf das Missmanagement wird erwähnt, dass eine unzureichende Einführung von IKT und ein fehlender organisatorischer Wandel zu einem erheblichen Produktivitätsverlust führen können. Darüber hinaus ist die Ertüchtigung eines organisatorischen Wandels zur grundlegenden Umgestaltung der Geschäftsprozesse eine notwendige Maßnahme, die vom Management ergriffen werden muss. Ein Schlüsselaspekt ist ein unzureichender Umgang mit diesen Technologien und das Fehlen ergänzender Voraussetzungen. Den Anwendern fehlt es an der notwendigen Kompetenz, um die Systeme effektiv nutzen zu können. Demnach werden nicht genügend Ressourcen für die notwendige Ausbildung zur Verfügung gestellt (Schweikl & Obermaier, 2020).

Daher ist es wichtig, dass die Arbeitnehmer ihre Fähigkeiten ständig weiterentwickeln und dass die Unternehmen in Schulungs- und Entwicklungsprogramme investieren, um sicherzustellen, dass ihre Mitarbeiter über die notwendigen Fähigkeiten verfügen, um neue Technologien effektiv zu nutzen (Schweikl & Obermaier, 2020). Es besteht ein Bedarf an Schulungs- und Weiterbildungsprogrammen, die den Arbeitnehmern die notwendigen Fähigkeiten vermitteln, um sich in der sich verändernden Arbeitswelt schnell an die neuen Aufgaben und Technologien anpassen zu können. Für immersive Assistenzsysteme lässt sich daraus die Notwendigkeit geeigneter Qualifizierungs- und Trainingsmaßnahmen ableiten; ein unmittelbarer empirischer Nachweis für immersive Systeme wird durch den Metareview jedoch nicht erbracht (Schweikl & Obermaier, 2020). Die Studie von Zoetbrood (2021) zeigt anhand der Korrelationsmatrix lediglich einen sehr geringen und statistisch nicht signifikanten negativen Zusammenhang zwischen dem Einsatz von Ausbildungspraktiken und der Produktivität ($r = -0,069$; $p > 0,05$). Darüber hinaus konnte die Hypothese, dass ein verstärkter Einsatz von Schulungs- und Entwicklungspraktiken in einer Organisation zu einem größeren Produktivitätsnutzen infolge von IKT-Investitionen führt, basierend auf der multivariaten Analyse nicht bestätigt werden ($\beta = 0,27$; $t(98) = 1,003$; $p = 0,318$). Ebenso zeigte sich kein signifikanter Haupteffekt der Schulungs- und Entwicklungspraktiken auf die Arbeitsproduktivität (β

= 0,034; $t(98) = 1,047$; $p = 0,298$). Die Einbeziehung des Interaktionsterms führte lediglich zu einer geringen und statistisch nicht signifikanten Erhöhung der erklärten Varianz ($\Delta R^2 = 0,015$; $F\text{-Change}(2,98) = 0,980$; $p = 0,379$). Insgesamt liefert die Studie somit keinen empirischen Beleg dafür, dass Schulungs- und Entwicklungspraktiken den Produktivitätsnutzen von IKT-Investitionen signifikant erhöhen oder den Zusammenhang zwischen IKT-Investitionen und Arbeitsproduktivität moderieren.

Die Studie von Korunka und Vitouch (1999) zeigt, dass Schulungsbedarf im Rahmen von IT-Implementierungen einen signifikanten Einfluss auf die Wahrnehmung und Bewertung des Veränderungsprozesses durch die Mitarbeitenden hat. Schulung wurde dabei als Teil der sogenannten Anpassungsanforderungen operationalisiert, gemeinsam mit Qualifikationsveränderungen und Funktionsänderungen. Entgegen der häufig geäußerten Annahme, dass ein hoher Schulungsaufwand als Belastung empfunden wird, ergaben die Ergebnisse der Studie, dass höhere Anpassungsanforderungen, zu denen auch Schulungs- und Qualifikationsbedarfe gehörten, mit einer geringeren subjektiven Belastung und einer höheren Arbeitszufriedenheit einhergingen. Diese positive Wirkung von Schulung lässt sich durch ein erhöhtes Gefühl der Kontrolle und Vorbereitung erklären, das Schulungen den Mitarbeitenden vermitteln. In einem von Unsicherheit geprägten Arbeitsmarkt – wie dieser zur Zeit der Erhebung bestand – wurden Weiterbildungen zudem als investive Maßnahme in die eigene Beschäftigungsfähigkeit verstanden, was die psychologische Beanspruchung zusätzlich reduzierte. Dabei zeigte sich, dass der Schulungsbedarf insbesondere dann positiv wirkte, wenn er mit einer aktiven oder passiven Beteiligung am Veränderungsprozess einherging. Statistisch konnte eine starke Korrelation zwischen Partizipation und Anpassungsanforderungen nachgewiesen werden (standardisierte Residual-Korrelation: $r = 0,57$; Korrelationswert in der Korrelationsmatrix: $r \approx 0,75$), die auf eine enge Verbindung zwischen der Beteiligung der Mitarbeitenden und dem Umfang an Qualifikations- bzw. Schulungsmaßnahmen hinweist. Die subjektive Belastung korrelierte dabei invers mit den Anpassungsanforderungen, sodass höhere Anforderungen, unter den untersuchten Kontextbedingungen mit einer geringeren subjektiven Belastung verbunden waren. Diese Befunde unterstreichen die Bedeutung von Qualifizierung und Partizipation als integraler Bestandteil

erfolgreicher IT-Einführungen, insbesondere wenn sie von transparenter Kommunikation und Einbindung der Mitarbeitenden begleitet werden (Korunka & Vitouch, 1999).

Untersuchungen von Day et al. (2019) und Korunka und Vitouch (1999) zeigen, dass angemessen geschulte Arbeitnehmer, die an der Einführung neuer IKT beteiligt sind, weniger Stress erleben. Nach Wang et al. (2008) kann Technostress durch die Förderung einer lernenden Organisation und ausreichende Schulungen verringert werden. Diese Befunde beziehen sich auf organisationale Einführungs- und Schulungskontexte und erlauben keine unmittelbare Aussage über die Qualität konkreter Trainingsmethoden für immersive Assistenzsysteme.

Die Studie von Bala und Venkatesh (2016) zeigt, dass effektive Schulungen die wahrgenommene Bedrohung durch neue IT-Systeme signifikant reduzieren. Der direkte Effekt der Trainingseffektivität (TRNG) auf die wahrgenommene Bedrohung (PTHRT) beträgt $-0,11$ und ist auf dem Niveau von $p < 0,05$ signifikant. Darüber hinaus wurde ein signifikanter Interaktionseffekt zwischen Trainingseffektivität und Nutzerbeteiligung festgestellt, mit Effektstärken von $-0,14$ bis $-0,15$ ($p < 0,05$). Diese Werte belegen, dass der positive Einfluss von Schulungsmaßnahmen besonders dann wirksam ist, wenn Mitarbeitende aktiv in den Implementierungsprozess eingebunden sind. Insgesamt unterstreichen die Befunde, dass eine Kombination aus wirksamer Schulung, hoher Nutzerbeteiligung und organisatorischer Unterstützung wesentlich dazu beiträgt, Ängste und Widerstände gegenüber IT-Implementierungen abzubauen (Bala & Venkatesh, 2016).

Die Studie von Mariani et al. (2013) untersucht den Einfluss von Trainingsmöglichkeiten auf die Akzeptanz von Informationstechnologien und auf die Arbeitszufriedenheit in italienischen Organisationen. Anhand eines strukturgleichungsmodellbasierten Ansatzes mit einer Stichprobe von 497 Arbeitnehmern wurde analysiert, wie das Angebot von IT-Schulungen die Wahrnehmung von Benutzerfreundlichkeit (Ease of Use) und Nützlichkeit (Perceived Usefulness) sowie die Absicht zur Nutzung von IT (Intention to Use) beeinflusst. Die Ergebnisse zeigen, dass Trainingsmöglichkeiten signifikant mit den zentralen Variablen des Technology Acceptance Model (TAM) zusammenhängen. Konkret konnten standardisierte Effekte von $0,16$ für die wahrgenommene Benutzerfreundlichkeit und von $0,21$ für die wahrgenommene Nützlichkeit nachgewiesen werden, was die Bedeutung von Trainingsangeboten bei der Förderung der Technologieakzeptanz

unterstreicht. Darüber hinaus hebt die Studie hervor, dass Trainingsangebote auch das Vertrauen in die eigenen IT-Kompetenzen (IT-Self-Competence) stärken und die allgemeine Arbeitszufriedenheit positiv beeinflussen können. Die Trainingsmöglichkeiten wirken somit als wichtiger Hebel, um sowohl die subjektiven Erfahrungen mit der Technologie als auch die generelle Zufriedenheit am Arbeitsplatz zu verbessern. Zusammenfassend zeigt die Studie, dass IT-Trainings mit einer höheren wahrgenommenen Benutzerfreundlichkeit und Nützlichkeit sowie darüber mit der Nutzungsabsicht in Zusammenhang stehen. Gleichzeitig tragen sie zur Erhöhung der IT-Kompetenz und zur Verbesserung der Arbeitszufriedenheit bei, was insgesamt zum erfolgreichen Einsatz von Technologien in Organisationen beiträgt (Mariani et al., 2013).

Die Studie von Marshall et al. (2008) untersucht den Einfluss der Endanwenderschulung auf die Akzeptanzfaktoren PE (Leistungserwartung) und EE (Aufwandserwartung) im Bereich von Kieferchirurgiesystemen. Anhand einer empirischen Untersuchung unter oralchirurgischen Fachärzten wurde festgestellt, dass eine gezielte Endanwenderschulung signifikant positiv mit PE (Pearson = 0,769) und EE (Pearson = 0,840) korreliert. Dies bedeutet, dass Endanwenderschulung mit einer höheren wahrgenommenen Leistungsfähigkeit der Technologie sowie mit einer leichteren und weniger aufwendigen Nutzung verbunden war. Die Ergebnisse unterstreichen die zentrale Rolle von Endanwenderschulungen bei der Förderung der Technologieakzeptanz, insbesondere in anspruchsvollen medizinischen Anwendungsbereichen wie der Kieferchirurgie (Marshall et al., 2008).

In der Studie von Alqahtani et al. (2022) wurde gezeigt, dass die wahrgenommene Ausbildungsqualität einen signifikanten Einfluss auf die Akzeptanzfaktoren wahrgenommene Benutzerfreundlichkeit (PEU) ($\beta = 0,132$; $p = 0,005$) und wahrgenommenen Nutzen (PU) ($\beta = 0,096$; $p = 0,015$) von E-Learning-Systemen hat. Allerdings konnte kein direkter signifikanter Einfluss der Ausbildungsqualität auf die Nutzungsabsicht festgestellt werden. Dennoch wurde im Strukturgleichungsmodell eine starke Faktorladung von 0,81 sowie eine Korrelation von 0,461 zwischen Ausbildungsqualität und Nutzungsabsicht ermittelt, was auf einen indirekten Zusammenhang hinweist.

Die Untersuchung von Minow (2021) ergab, dass die Desktoptrainingsgruppe (interaktive und computergestützte Vermittlung der Lerninhalte) die

Gebrauchstauglichkeit des AS mit einem SUS-Mittelwert von $79,9 \pm 12,43$ Punkten tendenziell besser bewertete als die Trainingsfilmgruppe ($73,7 \pm 12,99$; $p = 0,057$). Der Unterschied des SUS-Gesamtwerts war damit nicht statistisch signifikant. Besonders die Verständlichkeit im Umgang wurde signifikant besser eingeschätzt ($p < 0,05$). Zudem zeigte die Desktoptrainingsgruppe eine signifikant höhere Lernförderlichkeit sowie tendenziell bessere Werte bei Selbstbeschreibungsfähigkeit, Fehlertoleranz und Individualisierbarkeit. Objektiv waren die Leistungsdaten vor der Puzzlephase deutlich besser in der Desktoptrainingsgruppe (signifikant nach Mann-Whitney-U-Test), während subjektive visuelle Beanspruchung und geistige Anstrengung keine signifikanten Unterschiede zeigten. Die Ergebnisse weisen darauf hin, dass interaktives Desktoptraining insbesondere die Verständlichkeit, einzelne Gebrauchstauglichkeitsdimensionen und bestimmte objektive Leistungswerte verbessern kann. Für den SUS-Gesamtwert sowie für die visuelle Beanspruchung und die geistige Anstrengung wurden hingegen keine signifikanten Unterschiede nachgewiesen (Minow, 2021).

Qualitative Verknüpfung der Beiträge aus den Vorarbeiten mit Variablen des übergeordneten Forschungsdesigns: Die berücksichtigten Ergebnisse der Literaturrecherche weisen hinsichtlich der untersuchten Variablen, ihrer Operationalisierung, der Methodik, der berichteten Ergebnisse sowie der internen und externen Validität eine erhebliche Heterogenität auf. Ein konsistentes Gesamtbild der zugrunde liegenden Wirkungszusammenhänge lässt sich aus dem derzeitigen Forschungsstand daher nicht ableiten. Vielmehr zeigen die Vorarbeiten überwiegend kontextspezifische Tendenzen, deren Vergleichbarkeit durch unterschiedliche Forschungsdesigns, Stichproben, Messinstrumente und Auswertungsverfahren eingeschränkt ist. Vor diesem Hintergrund verknüpft Tabelle 55 im Anhang A1 die unabhängigen Variablen (UV) DQ, IQ, VQ und TQ mit den primären, abhängigen Variablen Leistung, Akzeptanz und Beanspruchung. Da die einbezogenen Studien teilweise deutlich breitere Variablenspektren und differenziertere Wirkungszusammenhänge untersuchen, werden die zentralen Befunde zur Synthese werden jeweils funktional dem nächstgelegenen Primärfaktor (PF) bzw. der nächstgelegenen UV des übergeordneten Forschungsdesign zugeordnet. Eine ganzheitliche, fachliche Einordnung der Variablen in das vorliegende Forschungsdesign erfolgt daher nur auf dieser Abstraktionsebene. Zugleich verdeutlicht die uneinheitliche und nur

eingeschränkt vergleichbare Befundlage den weiterführenden Forschungsbedarf und begründet die systematische Untersuchung der formulierten Forschungsfragen und Forschungsableitungen.

3 Übergeordnetes Forschungsdesign

Aufbauend auf der Problemstellung (vgl. Kapitel 1.1) und der thematischen Abgrenzung (vgl. Kapitel 1.2) wurde eine kritische Literaturrecherche durchgeführt. Deren Ergebnisse flossen sowohl in die Ableitung des Forschungsbedarfs (vgl. Kapitel 1.3) als auch in die Darstellung des aktuellen Forschungsstands (vgl. Kapitel 2) ein. Auf dieser Grundlage wurden die zentralen Forschungsfragen sowie entsprechende Hypothesen formuliert. Das daraus entwickelte übergeordnete Forschungsdesign bildet die konzeptionelle Grundlage der empirischen Untersuchung und wird im Folgenden näher erläutert.

3.1 Hypothesen

Entsprechend wurden die nachfolgenden Hypothesen zur Beantwortung der Forschungsfragen entwickelt. Hierbei lag der Fokus hinsichtlich des Studiendesigns sowie der Stichprobenumfangsplanung primär auf den Auswirkungen der Leistungsfaktoren (v.a. auf dem Index Leistung) sowie der Akzeptanzfaktoren (v.a. der BI) und sekundär auf den Beanspruchungsfaktoren (v.a. der durchschnittlichen Beanspruchung (Mean_NASA)).

Datenfehler (Bezug zu $FF_{1,A}$)

$H_{A.1.1}$: Fehlerhafte Informationen reduzieren bei IAS die Leistung.

$H_{A.1.0}$: Fehlerhafte Informationen beeinflussen bei IAS die Leistung nicht.

$H_{A.2.1}$: Fehlerhafte Informationen reduzieren bei IAS die Nutzerakzeptanz.

$H_{A.2.0}$: Fehlerhafte Informationen beeinflussen bei IAS die Nutzerakzeptanz nicht.

$H_{A.3.1}$: Fehlerhafte Informationen erhöhen bei IAS die Beanspruchung des Anwenders.

$H_{A.3.0}$: Fehlerhafte Informationen beeinflussen bei IAS die Beanspruchung des Anwenders nicht.

Sichtbarkeitsfehler (Bezug zu $FF_{1.B}$)

- $H_{B.1.1}$: Fehlerhafte Sichtbarkeit reduziert bei IAS die Leistung.
- $H_{B.1.0}$: Fehlerhafte Sichtbarkeit beeinflusst bei IAS die Leistung nicht.
- $H_{B.2.1}$: Fehlerhafte Sichtbarkeit reduziert bei IAS die Nutzerakzeptanz.
- $H_{B.2.0}$: Fehlerhafte Sichtbarkeit beeinflusst bei IAS die Nutzerakzeptanz nicht.
- $H_{B.3.1}$: Fehlerhafte Sichtbarkeit erhöht bei IAS die Beanspruchung des Anwenders.
- $H_{B.3.0}$: Fehlerhafte Sichtbarkeit beeinflusst bei IAS die Beanspruchung des Anwenders nicht.

Interaktionsfehler (Bezug zu $FF_{1.C}$)

- $H_{C.1.1}$: Fehlerhafte Interaktion reduziert bei IAS die Leistung.
- $H_{C.1.0}$: Fehlerhafte Interaktion beeinflusst bei IAS die Leistung nicht.
- $H_{C.2.1}$: Fehlerhafte Interaktion reduziert bei IAS die Nutzerakzeptanz.
- $H_{C.2.0}$: Fehlerhafte Interaktion beeinflusst bei IAS die Nutzerakzeptanz nicht.
- $H_{C.3.1}$: Fehlerhafte Interaktion erhöht bei IAS die Beanspruchung des Anwenders.
- $H_{C.3.0}$: Fehlerhafte Interaktion beeinflusst bei IAS die Beanspruchung des Anwenders nicht.

Trainingsfehler (Bezug zu $FF_{1.D}$)

- $H_{D.1.1}$: Unzureichendes Training reduziert bei IAS die Leistung.
- $H_{D.1.0}$: Unzureichendes Training beeinflusst bei IAS die Leistung nicht.
- $H_{D.2.1}$: Unzureichendes Training reduziert bei IAS die Nutzerakzeptanz.
- $H_{D.2.0}$: Unzureichendes Training beeinflusst bei IAS die Nutzerakzeptanz nicht.
- $H_{D.3.1}$: Unzureichendes Training erhöht bei IAS die Beanspruchung des Anwenders.
- $H_{D.3.0}$: Unzureichendes Training beeinflusst bei IAS die Beanspruchung des Anwenders nicht.

Fehlervergleich (Bezug zu FF_2)

- $H_{1.1}$: Unterschiedliche Fehlertypen manipulieren bei IAS die Leistung unterschiedlich.
- $H_{1.0}$: Der Fehlertyp hat bei IAS keinen Einfluss auf die Leistung.
- $H_{2.1}$: Unterschiedliche Fehlerarten manipulieren bei IAS die Nutzerakzeptanz unterschiedlich.
- $H_{2.0}$: Der Fehlertyp hat bei IAS keinen Einfluss auf die Nutzerakzeptanz.
- $H_{3.1}$: Unterschiedliche Fehlerarten manipulieren bei IAS die Beanspruchung unterschiedlich.
- $H_{3.0}$: Der Fehlertyp hat bei IAS keinen Einfluss auf die Beanspruchung.

Akzeptanzmanipulation (Bezug zu FF_3)

- $H_{4.1}$: Die Anzeige der Systemkonfidenz wirkt bei IAS positiv auf die Nutzerakzeptanz.
- $H_{4.0}$: Die Anzeige der Systemkonfidenz hat bei IAS keinen Einfluss auf die Nutzerakzeptanz.

3.2 Methodik

Methodische Herleitung des übergeordneten Forschungsdesigns

Den theoretischen Ausgangspunkt für die Gestaltung des Forschungsdesigns bildet die Metastudie von Suh und Prophet (2018). Ihre umfassende Literaturlauswertung identifiziert systemische Einflussgrößen immersiver Technologien auf die Nutzererfahrungen und -ergebnisse. Das entwickelte Stimulus–Organismus–Reaktion-Modell beschreibt, wie technologische Reize – etwa visuelle, auditive oder haptische Darstellungen – kognitive und affektive Prozesse wie Aufmerksamkeit, Präsenz oder Immersion auslösen, welche wiederum messbare Effekte wie Nutzungsintention, Leistung, Überforderung etc. hervorrufen. Diese Struktur findet sich im Forschungsdesign in der Auswahl der Reaktionsgrößen wieder: Subjektive Beanspruchung (E_2), Akzeptanz (E_1) und physiologische Parameter (E_3) dienen als Indikatoren der Wirkung immersiver Informationsangebote (U_1). Ergänzend liefern Fantini et al. (2020), Moencks et al. (2022) und Stowers et al. (2017) konzeptionelle Grundlagen zur strukturierten Verknüpfung von

Mensch, Technik und Kontext. Während Stowers et al. (2017) ein systemisches Modell präsentieren, das technische (z. B. Automatisierungsgrad), menschliche (z. B. Einstellungen) und kontextuelle (z. B. Aufgabenanforderung) Faktoren integriert, konkretisieren Moencks et al. (2022) diesen Ansatz organisatorisch durch den „Augmented Workforce Canvas“, in dem der Mensch als funktionaler Wertschöpfungsträger innerhalb betrieblicher Rollen modelliert wird. Beide Konzepte fließen in das Forschungsdesign ein, indem der Mensch als Stellglied agiert und die Gestaltung der Assistenzfunktion auf reale Aufgabenprofile und Kontextbedingungen bezogen wird. Fantini et al. (2020) vertiefen diese Perspektive, indem sie menschliche Arbeit in CPS formal beschreiben und simulationsbasiert bewerten – vorrangig aufgabenorientiert und nicht technikzentriert. Ihre Modellierung liefert das methodische Fundament, um die fachliche Aufgabe $W(t)$ kontextsensitiv in die Gestaltung der Interaktion einzubetten. Ein zentraler theoretischer Impuls zur Ausgestaltung der Mensch-Maschine-Schnittstelle stammt von Müller et al. (2008). Sie zeigen, dass adaptive Systeme nur dann akzeptiert werden, wenn ihre Steuerungslogik für den Nutzer sichtbar bleibt. Im Forschungsdesign wird diese Forderung durch eine Konfidenzanzeige umgesetzt, die systemseitige Unsicherheiten transparent macht und dem Nutzer eine nachvollziehbare Einschätzung über die Vertrauenswürdigkeit der Handlungsempfehlungen (U_1) erlaubt – ein zentrales Mittel zur Sicherung von Vertrauen und Akzeptanz. Die Bedeutung systemischer Fehlerquellen in der MMI wird auf Basis der erweiterten Fehlertaxonomie von Palanque et al. (2020) berücksichtigt. Die vier im Forschungsdesign operationalisierten Fehlergrößen – Datenfehler (F_1), Sichtbarkeitsfehler (F_2), Interaktionsfehler (F_3) und Trainingsdefizite (F_4) – wirken als Störeinflüsse auf Leistungsparameter P_i und menschliche Effekte E_j . Alessa et al. (2023) liefern die empirische Grundlage zur Integration sowohl subjektiver Beanspruchungsskalen als auch physiologischer Stressmarker wie Herzrate oder elektrodermalen Aktivität in das Design, um Belastungsreaktionen aus unterschiedlichen Stimuli innerhalb der MMI vergleichend erfassen zu können. Die strukturelle und funktionale Verankerung des Menschen in der Steuerungslogik wird wesentlich durch die Arbeiten von Cimini et al. (2020), Pacaux-Lemoine et al. (2017) und Reuter und Zacher (2008) getragen. Reuter und Zacher (2008) liefern mit dem Regelkreisparadigma die technische Systemgrundlage: Führungsgrößen $G(t)$, Messgrößen $D(t)$, Störgrößen F_k und Stellgrößen $U(t)$

werden formalisiert und auf kognitive Mensch-Technik-Systeme übertragen. Cimini et al. (2020) erweitern dieses Konzept um eine Industrie-4.0-taugliche Steuerungsarchitektur, in der der Mensch bewusst als Entscheidungsträger erhalten bleibt. Pacaux-Lemoine et al. (2017) ergänzen dies um ein dynamisches Kooperationsmodell, das auf situativ angepasste Interaktion statt starre Automatisierungslogik setzt. Im Forschungsdesign resultiert daraus ein geschlossener Rückkopplungskreis: Das System erfasst kontinuierlich $D(t)$, vergleicht diese mit $G(t)$, generiert daraus adaptive Hinweise $U_1(t)$ und überlässt dem Menschen die Ausführung der Aufgabe $W(t)$ – wobei dieser gleichzeitig Bewertungspartner und aktives Stellglied bleibt.

Entsprechend dieser Grundlagen wurde ein Forschungsmodell entwickelt, dessen Komponenten, Variablen und Wirkzusammenhänge in Abbildung 3 dargestellt werden:

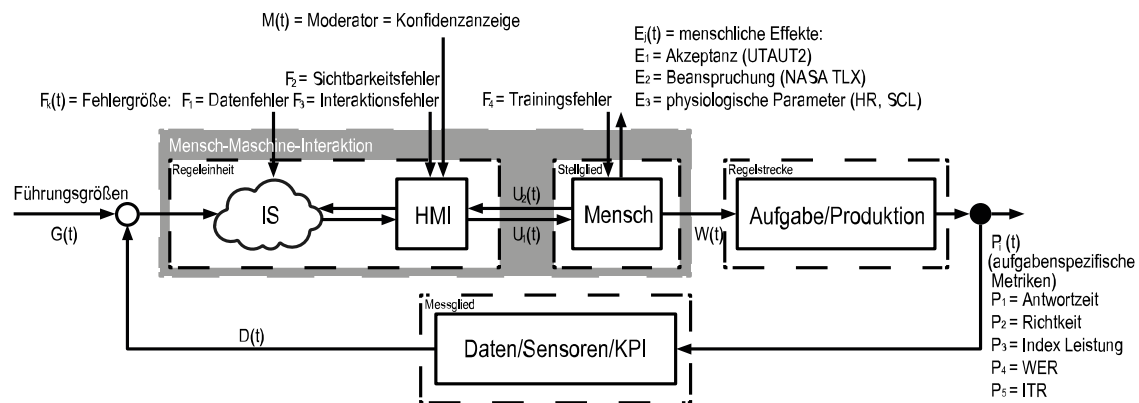


Abbildung 3: Blockdiagramm des übergeordneten Forschungsdesigns

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Abstrahierte Modellbeschreibung im industriellen Kontext

Die Regeleinheit stellt das IAS inkl. der Hardware dar und besteht aus dem Informationssystem (IS) und der HMI, wobei das IS in diesem Modell die immersive Anwendung beinhaltet. Die Regeleinheit hat die Aufgabe, die vom Messglied stammenden Daten D (z. B. KPI, Produktionsdaten etc.) mit den vorgegebenen Führungsgrößen G bzw. aufgabenspezifischen Zielen des jeweiligen Wirkungsbereichs des AS zu verrechnen. Die Regeleinheit (immersives Assistenzsystem) und das Stellglied (Mensch) bilden zusammen die MMI. Der Mensch stellt in diesem Modell das Stellglied dar; er wirkt physisch auf die Regelstrecke ein. Die kontextualen Informationen U_i (z. B. Handlungsempfehlung, in Abhängigkeit des Automatisierungsgrads, Parasuraman et al., 2000) des

Systems werden dem Menschen über das HMI situationsgerecht präsentiert. Wiederum interagiert der Mensch über das HMI mit dem AS, weswegen die Informationsflüsse $U_{1/2}$ bidirektional zwischen dem Menschen und dem AS stattfinden. Als weitere Einschränkung für das Modell wird angenommen, dass es sich um ein kognitives AS handelt (u. a. Mark et al., 2021), bei dem das AS selbst keinen direkten Einfluss auf die Regelstrecke nehmen kann. Somit wird idealisiert angenommen, dass alle Informationen U_1 über das HMI dem Menschen bereitgestellt werden. Die Informationsverarbeitung und Handlungsableitung basieren sowohl auf der Ausgangssituation der fachlichen Aufgabe W sowie auf den bereitgestellten Informationen U_1 durch das AS. Als Handlungsoptionen kann der Mensch zum einen die fachlichen Aufgaben W ausführen, zum anderen mit dem AS U_2 interagieren. Organisatorisch sind Werker entsprechend dem Werkzeugszenario im HCPS-Kontext angehalten, Informationen und Hilfestellungen über digitale Endgeräte zu erhalten sowie mit diesen zu interagieren (z. B. Bedienung von Software oder Dateneingaben). Abstrahiert werden AS so ausgelegt, dass sie kollaborativ operative Aufgaben entsprechend Führungsgrößen erfüllen. Das AS fragt über die IT-Infrastruktur die für die Aufgabe spezifischen Daten D ab und vergleicht diese mit den Zielvorgaben G . Situativ können dem Anwender somit relevante Informationen (Handlungsempfehlungen etc.) U_1 je nach Aufgabe und Automatisierungsunterstützung über die HMI bereitgestellt werden. Der Mensch führt seine Aufgabe W zusammen mit dem AS entsprechend seinen Vorgaben und den Erwartungen an ihn fachlich aus und liefert einen messbaren Produktivitätsbeitrag P_i gemäß der Zielgrößen G (z. B. Stückzahl und Qualität bzw. Richtigkeit seiner Aufgabe etc.). Auf die MMI können unterschiedliche Störeinflüsse einwirken, die in dem Modell als Fehlergrößen F_k dargestellt werden. Zum einen kann ein Datenfehler F_1 auftreten, d. h. dass die bereitgestellten Daten fehlerhaft sind (z. B. Anzeige eines falschen Parameters oder Handlungsempfehlung etc.). Weiter kann ein Sichtbarkeitsfehler F_2 vorliegen, weswegen die Qualität der visuellen Wahrnehmung des Anwenders beeinträchtigt wird (z. B. unzureichende Helligkeit, zu kleine oder unscharfe visuelle Informationen). Ebenfalls kann der Interaktionsfehler F_3 auftreten, bei dem die Nutzereingaben bzw. Interaktion nicht oder fehlerhaft vom System erfasst werden. Darüber hinaus kann ein Trainingsfehler vom Typ F_4 vorliegen, bei dem Anwender entweder die erforderlichen Fähigkeiten oder das spezifische Wissen fehlen, um das

Assistenzsystem korrekt zu bedienen – etwa infolge unzureichender Schulungsqualität oder -quantität, individueller Dispositionen oder ähnlicher Faktoren. Diese Fehlergrößen F_k wirken als Störeinflüsse innerhalb der MMI sowohl auf die Leistung und Qualität der operativen Aufgabe als auch auf die menschlichen Effekte E_j . Hierbei soll der Fokus auf die Akzeptanzfaktoren E_1 des UTAUT2-Fragebogens (Venkatesh et al., 2012), die Beanspruchungsfaktoren E_2 des Fragebogens NASA-TLX (Hart, 2006) aufgrund der Belastung sowie auf die physiologischen Parameter E_3 (Alessa et al., 2023; Giorgi et al., 2021; Kosch et al., 2023; Setz et al., 2010) der Herzrate und elektrodermalen Aktivität zur Erfassung der Stressparameter gelegt werden. Außerdem kann das IAS dem Nutzer über eine Anzeige seine Systemkonfidenz bereitstellen; diese geht als Moderatorvariable $\mathbf{M}$ in das Modell ein.

Entsprechend dieser Modellbeschreibungen im industriellen Kontext wurde ein Design of Experiment (DOE) entwickelt, um mittels Studierhebungen die Auswirkungen der Fehlergrößen auf die menschlichen Effekte sowie die Leistung zu untersuchen.

Beschreibung des übergeordneten Design of Experiments

Das DOE ist für die Erforschung der MMI bei IAS ausgelegt. Es gilt, die Auswirkungen der unabhängigen Variablen der Daten-, Trainings-, Visualisierungs- und Interaktionsqualität mit jeweils unterschiedlichen Faktorstufen auf die abhängigen Variablen Antwortzeit P_1 , Richtigkeit P_2 , Akzeptanz E_1 , Beanspruchung E_2 und physiologischen Parameter E_3 sowie die spezifischen Metriken Wortfehlerrate (WER) P_4 und Informationstransferrate (ITR) P_5 für die Sprachinteraktion zu untersuchen. (Der Index Leistung P_3 als weitere abhängige Variable wird aus den beiden abhängigen Variablen Antwortzeit P_1 und Richtigkeit P_2 verrechnet und erst im weiteren Verlauf introduziert.) Die Faktorstufen niedriger Qualität (Datenfehler F_1 , Sichtbarkeitsfehler F_2 , Interaktionsfehler F_3 , Trainingsfehler F_4) werden je unabhängige Variable entsprechend ausgelegt, sodass sie als fehlerhaft definiert werden. Hierbei sollen primär sowohl die Effekte je Fehlermanipulation als auch der Fehlervergleich untersucht werden. Sekundär wird die Manipulation des stärksten Fehlereffekts aus der Verbundstudie 4 auf die Akzeptanz durch die Konfidenzanzeige als Moderatorvariable $\mathbf{M}$ untersucht.

Datenqualität: Die DQ beschreibt, wie unterstützend und richtig die angezeigten immersiven Informationen U_1 des AS zur fachlichen Aufgabe sind (Nasirian et al., 2017). Bei einer Systemstörung ist es möglich, dass die Daten fehlerhaft sind oder nicht vollständig angezeigt werden, d. h. die Informationen, die der Nutzer vom System erhält, sind u. a. nicht hinreichend genau, vollständig, relevant, widerspruchsfrei oder aktuell (Gegoff et al., 2023; Pal et al., 2019; Wandke, 2005). Zur Untersuchung des Einflusses werden zwei Faktorstufen der DQ gewählt – eine hohe und eine niedrige. Bei der hohen DQ werden dem Nutzer die richtigen Informationen durch das System angezeigt. Hingegen soll bei einer niedrigen DQ davon ausgegangen werden, dass ein Datenfehler F_1 vorliegt, der dem Anwender fehlerhafte Informationen anzeigt.

Visuelle Qualität: VQ ist die Sichtbarkeit augmentierter Annotationen des AS und bezieht sich auf die Fähigkeit, einen gewissen Informationsumfang innerhalb eines Zeitraums zu erkennen, um virtuelle Informationen U_1 eindeutig wahrzunehmen (Nasirian et al., 2017; Rosilius et al., 2021). Zur Untersuchung des Einflusses werden zwei Faktorstufen definiert. Zum einen die native Sichtbarkeit bzw. Repräsentation der Informationen entsprechend einem geeigneten Anwendungsdesign zum normalen Gebrauch. Zum anderen soll eine verminderte Sichtbarkeit als Sichtbarkeitsfehler F_2 entwickelt werden. Hierzu können die Freiheitsgrade der Softwareparameter entsprechend angepasst werden, u. a. durch die Reduzierung der Helligkeit der Szene bzw. des Displays, der Schrift-/Objektgrößen sowie der Opazität oder des Kontrastverhältnisses (Rosilius et al., 2021, 2022).

Interaktionsqualität: Die Interaktionsqualität (IQ) ist ein Maß, das angibt, wie gut die Bedienung bzw. Eingaben des Nutzers via der entsprechenden Interaktionsparadigmen an das System übertragen werden (Chaudary, 2023; Nasirian et al., 2017). In Bezug auf IAS werden im industriellen Kontext die Gestensteuerung und Sprachsteuerung/-eingabe als Interaktionsparadigmen bevorzugt (Rosilius, Spiertz et al., 2024). Zur Untersuchung des Einflusses der Interaktionsqualität werden zwei Faktorstufen definiert. Zum einen werden die nativen Fähigkeiten der Interaktionsparadigmen eingesetzt. Bei der zweiten Faktorstufe wird das Interaktionsvermögen eingeschränkt, welches den Interaktionsfehler F_3 darstellen soll, sodass die Zielerfüllung sowie das Bedienverhalten des Nutzers erschwert wird (Rosilius, Spiertz et al., 2024).

Trainingsqualität: Die Trainingsqualität (TQ) entspricht dem Kompetenzerwerb des Anwenders, der ihm ermöglicht, seine fachliche Aufgabe zusammen mit dem IAS effektiv und effizient auszuführen (Korpelainen & Kira, 2010; Rosilius, Hügel et al., 2024). Die unabhängige Variable TQ wird in drei Faktorstufen dargestellt. Eine Faktorstufe soll dabei mittels eines minimalen Trainingsszenarios den Trainingsfehler F_4 repräsentieren; sie orientiert sich an den nativen Erste-Hand-Lernangeboten von Microsoft. Bei den anderen beiden Faktorstufen werden jeweils eine sehr ausführliche und eine optimierte Trainingsqualität abgebildet (Rosilius, Hügel et al., 2024).

Durchlaufzeit P_1 : Die Durchlaufzeit (DLZ), auch Antwortzeit genannt, ist die Zeit, die der Proband benötigt, um seine operative Aufgabe zusammen mit dem AS auszuführen, oder bis er seine Aufgabe abbricht, falls er nicht in der Lage ist, diese erfolgreich abzuschließen. Nach den umfangreichen Metareviews (Dey et al., 2018; Merino et al., 2020) wurde die Nutzerleistung u. a. mit der Metrik DLZ gemessen.

Richtigkeit P_2 : Die Richtigkeit, auch Genauigkeit genannt, gibt an, mit welcher Qualität bzw. wie fehlerfrei der Proband seine Aufgabe ausgeführt hat. Entsprechend den Metastudien (Dey et al., 2018; Merino et al., 2020) sind Richtigkeit und DLZ die beiden gängigen Metriken, mit denen Aufgaben in Nutzerstudien gemessen und bewertet werden.

Index Leistung P_3 : Im Rahmen dieser Forschungsarbeit sowie innerhalb einer bereits veröffentlichten Studie (Rosilius, Hügel et al., 2024) wurde der Index Leistung (IL) novelliert; er kann je nach Aufgabenstellung kontextual adaptiert werden, vgl. Index Sichtbarkeit (Rosilius et al., 2022) als Quality of Experience (Reichl et al., 2010). Dieser Index erhöht die Präzision und die Aussagefähigkeit von Nutzerstudien gegenüber aktuellen Metareviews (Dey et al., 2018; Merino et al., 2020) aus dem Bereich der Extended-Reality (XR) Technologien. Entwickelt wurde dieses Konzept aus dem sog. Speed-Accuracy-Trade-off (Standage et al., 2015). Dieses besagt, dass es bei (kognitiven) Aufgaben eine Abhängigkeit zwischen der Geschwindigkeit und der Genauigkeit gibt. Dies bedeutet, dass eine erhöhte Geschwindigkeit in der Ausführung einer Aufgabe oft auf Kosten der Genauigkeit geht, während eine höhere Genauigkeit in der Regel zu einer langsameren Ausführung führt. Demnach wird zur Erfassung der Nutzerleistung der Quotient gebildet, wie in der folgenden Gleichung 1 dargestellt:

$$\text{Leistung} = \text{Richtigkeit} / \text{Durchlaufzeit}. \quad (1)$$

Akzeptanz E1: Angesichts der Vielzahl an Modellen und Theorien zur Beschreibung des Konstrukts der (Technologie-) Akzeptanz wird, entsprechend der Studie von Mkhonto und Zuva (2024), in dieser Arbeit nicht der Anspruch auf eine valide Definition erhoben. Stattdessen liegt der Fokus auf dem Anwendungsbezug, der Interpretierbarkeit sowie einer hohen Validität (Mkhonto & Zuva, 2024). Im Rahmen dieser Arbeit wird sich an dem Akzeptanzkonstrukt der Verhaltensabsicht, engl. Behavioral Intention (BI), ein IAS verwenden zu wollen, orientiert. Dieses basiert auf dem multidimensionalen Unified Theory of Acceptance and Use of Technology 2 Modell (UTAUT2) (Venkatesh et al., 2012). Ursprünglich wurde das Unified Theory of Acceptance and Use of Technology (UTAUT)-Modell zur Erklärung und Vorhersage der Akzeptanz neuer Technologien im Bereich der Informationstechnologie entwickelt (Venkatesh et al., 2003). Es kombiniert acht theoretische Ansätze aus Psychologie und Soziologie, um Faktoren zu identifizieren, die das individuelle Verhalten bei der Nutzung neuer Technologien beeinflussen. Die Erweiterung UTAUT2 umfasst insgesamt acht Akzeptanzfaktoren: Leistungserwartung (PE), Anstrengungserwartung (EE), sozialen Einfluss (SI), unterstützende Bedingungen (FC), hedonische Motivation (HM), Preiswert (PV), Gewohnheit (HT) und Verhaltensabsicht (BI). Diese Erweiterung verbessert insbesondere die Erklärungskraft für die Verbraucherakzeptanz von Technologien und ermöglicht eine breitere Anwendbarkeit sowie präzisere Vorhersagen zum Verhalten unterschiedlicher Nutzergruppen in vielfältigen Kontexten (Venkatesh et al., 2012). Die Studie von von Eitzen (2024) zeigt, dass Hedonismus – also das wahrgenommene Vergnügen bei der Nutzung immersiver Technologien – ein zentraler Einflussfaktor für deren Akzeptanz ist. Hedonismus trägt maßgeblich zur Nutzerzufriedenheit bei, die wiederum als entscheidender Prädiktor für Akzeptanz gilt. Gleichzeitig zeigen die Ergebnisse, dass sowohl Präsenz als auch Immersion das hedonistische Erleben fördern, wodurch ein verstärkter Zusammenhang zwischen Nutzungserlebnis und Akzeptanz entsteht. Darüber hinaus beeinflusst auch die Gestaltung der Nutzeravatare (Verkörperung) das wahrgenommene Vergnügen (von Eitzen, 2024). Das UTAUT2-Modell, das auf mehreren etablierten Akzeptanzmodellen wie dem TAM aufbaut, berücksichtigt im Gegensatz zu früheren Ansätzen explizit

hedonistische und emotionale Einflussfaktoren. Aufgrund seiner starken empirischen Validierung eignet es sich besonders gut zur Analyse der Nutzerakzeptanz immersiver Systeme (Kallweit, 2017). Es ermöglicht somit eine differenzierte und realitätsnahe Betrachtung technologischer Nutzungsmotive in immersiven Kontexten. Für die nachfolgenden Studien wurde der Fragebogen manuell aus der Originalfassung (Venkatesh et al., 2012) ins Deutsche übersetzt und auf den Kontext je Studiendesign und Aufgabenstellung passend adaptiert (siehe Anhang A2, Tabelle 56). Das Item Preiswert wurde bewusst nicht abgefragt, da im industriellen Kontext sowie für die Probanden der Studien monetäre Überlegungen keine Rolle spielen. Auf die Ergebnisse der Subskalen HT und SI wird innerhalb dieser Arbeit nicht weiter eingegangen, sie werden jedoch zur Vollständigkeit für weiterführende Forschungsaktivitäten präsentiert. Kausal können ad hoc nach einer Nutzererfahrung (eines IAS) keine Beeinflussung durch das soziale Umfeld oder eine Änderung der Gewohnheiten der Probanden stattfinden. Im Kontext des UTAUT2 gestaltet sich nach Kohl (o. J., S. 39) eine „(...) Änderung von langwierig eingeübten Gewohnheiten (...) schwierig und kann nur über einen längeren Zeitraum erfolgen“. Demnach sind HT und SI trägere Prozesse, die durch Erfahrungen sowie soziale Interaktion sowohl im privaten als auch im beruflichen Umfeld erfolgen und innerhalb der Studien nicht signifikant manipuliert werden können.

Beanspruchung Ez: Entsprechend dem Belastungsbeanspruchungsmodell stellt die Beanspruchung die subjektiven Auswirkungen der Belastungsexposition auf die Arbeitspersonen dar. Es umfasst die subjektiven, psychischen und physischen Reaktionen der Arbeitenden aufgrund der äußeren, objektiven Belastungen (Georg et al., 2013; Rohmert & Rutenfranz, 1975). Vor allem in den deutschsprachigen Arbeitswissenschaften wird diese Unterscheidung herausgestellt. In der englischsprachigen Literatur ist hiermit der „perceived workload“ gemeint. Als gängiges Modell zur Messung des Workloads wurde der NASA-Task-Load-Index (NASA-TLX) als eine Skala eingesetzt, um die Arbeitsbelastung von Bedienern nach der Durchführung einer Aufgabe zu bewerten. Diese multidimensionale Skala besteht aus sechs Subskalen, die mentale (MD), physische (PD) und zeitliche Anforderungen (TD), Frustration (F), Aufwand (E) und Leistung (P) erfassen. Der NASA-TLX wird in verschiedenen Umgebungen eingesetzt, um die Belastung von Bedienern zu bewerten, und hat sich als zuverlässig und einfach

anwendbar erwiesen. Er ist ein wichtiges Instrument für Designer, Hersteller, Manager und Betreiber, um die Leistung von Systemen in Bezug auf die Arbeitsbelastung der Bediener zu verstehen und zu optimieren (Hart, 2006).

Für die nachfolgenden Studien wurde der Fragebogen manuell aus der Originalfassung (Hart, 2006) ins Deutsche übersetzt und auf den Kontext je Studiendesign und Aufgabenstellung passend adaptiert (siehe Anhang A2, Tabelle 57). Zur Erfassung einer Gesamtbeanspruchung wird in den Studien zusätzlich zu den originären Subskalen der Mittelwert der durchschnittlichen Beanspruchung (Mean_NASA) berechnet. Diese Selbsteinschätzung der Arbeitsbelastung wird auch zur Bewertung des empfundenen Stressses herangezogen. MD ist die psychologische Reaktion als ein Bestandteil bei der Messung von Stress (Alsuraykh et al., 2019). Für die strukturierte Erhebung der Fragebögen wurde das Onlinetool SoSci Survey eingesetzt (*SoSci Survey online*, o. D.).

Physiologische Parameter E3: Physiologische Parameter (PP) sind messbare Merkmale oder Eigenschaften des menschlichen Körpers, die Aufschluss über den physiologischen Zustand einer Person geben. Diese Parameter können verschiedene Körperfunktionen und -reaktionen widerspiegeln, wie beispielsweise im vorliegenden Kontext die Herzfrequenz und die elektrodermale Aktivität (Giorgi et al., 2021; Kosch et al., 2023; Setz et al., 2010). Durch die Messung dieser physiologischen Parameter können Rückschlüsse u. a. auf das Stresslevel, die mentale und körperliche Belastung und andere physiologische Reaktionen auf bestimmte Aufgaben oder Stimuli gezogen werden (Alsuraykh et al., 2019). Mit dem tragbaren Messgerät der Empatica E4 können mittels der physiologischen Parameter das Stressniveau als physiologische Reaktion (Alsuraykh et al., 2019) über die Metriken Herzfrequenz (HR) und elektrodermale Aktivität (SCL) gemessen werden. Um die Auswirkungen von Stress durch die Stimuli zu isolieren, werden die HR_Differenz (Diff) und SCL_Diff berechnet, indem die Mittelwerte der physiologischen Parameter der Ruhephasen (RP) (fünf Min. Sitzen) von den Mittelwerten der Aktionsphasen (AP) (Exposition der Stimuli) subtrahiert werden (Giorgi et al., 2021).

Wortfehlerrate P4: Um IAS in operativen Aufgaben effektiv und effizient einzusetzen, bieten sie dem Nutzer die intuitive Möglichkeit der Sprachinteraktion als Alternative zur Gestensteuerung (Rosilius, Spiertz et al., 2024). Die WER ist eine Metrik, die zur Bewertung der Qualität der Sprachinteraktion (vgl. sprecherunabhängige automatische

Spracherkennung), genauer gesagt für das Sprache-zu-Text-Verfahren, herangezogen werden kann, um die Genauigkeit des erkannten Textes im Vergleich zum Referenztext zu bewerten. Die WER misst den prozentualen Anteil der Wörter im erkannten Text, die falsch sind, hinzugefügt, weggelassen oder falsch erkannt wurden, wie Gleichung 2 verdeutlicht (Filippidou & Moussiades, 2020).

$$WER = \frac{\text{Substitutionen} + \text{Auslassungen} + \text{Ergänzungen}}{\text{Gesamtanzahl der Worte}} \quad (2)$$

Informationstransferrate P₅: Eine weitere Teildisziplin der ASR ist der Sprachbefehl als alternatives Steuerungsparadigma von immersiven Systemen (Rosilius, Spiertz et al., 2024). Die Sprachbefehlserkennung bezieht sich auf die Funktion eines Systems, gesprochene Befehle von Benutzern zu verstehen und zu identifizieren. Dieser Prozess umfasst die Umwandlung von gesprochener Sprache in Aktionen, die von IAS verarbeitet werden können (McMahan & Rao, 2018). Im Rahmen dieser Forschungsarbeit sowie der veröffentlichten Studie (Rosilius, Spiertz et al., 2024) wurde die bewährte Metrik der Informationstransferrate aus dem Forschungsfeld der Brain-Computer-Interfaces (BCI) zur Leistungsmessung der Sprachbefehlsinteraktion herangezogen. Die ITR ist eine geeignete Metrik, da sie die Effizienz der Interaktion zwischen dem Benutzer und dem AS quantifiziert. Sie ermöglicht es, die Leistungsfähigkeit zu bewerten und zu vergleichen, insbesondere in Bezug auf die Geschwindigkeit und Genauigkeit der Informationsübertragung (BCI Kleve, o. D.; Dal Seno et al., 2010; Wolpaw et al., 2002).

$$B \left[\frac{\text{Bit}}{\text{Versuch}} \right] = \log_2(N) + P * \log_2(P) + (1 - P) * \log_2 \left(\frac{1 - P}{N - 1} \right), \quad (3)$$

$$Q \left[\frac{\text{Versuche}}{\text{min}} \right] = \frac{Sn}{T}, \quad (4)$$

$$ITR \left[\frac{\text{Bit}}{\text{min}} \right] = B * Q. \quad (5)$$

Die ITR (Gleichung 5) gibt die durchschnittliche Anzahl von übertragenen Bit pro Minute an. Die ITR ist abhängig von folgenden Faktoren und setzt sich entsprechend aus den Gleichungen 3–4 zusammen:

- Anzahl der möglichen Auswahlmöglichkeiten bzw. Ziele (N) pro Versuch,

- Genauigkeit (P),
- Anzahl der Versuche (S_n),
- durchschnittliche Zeit je Versuch (T).

Übergeordnetes, inferenzstatistisches Vorgehen

Hinweis: Dieser Abschnitt gibt einen kurzen Überblick über die eingesetzten statistischen Verfahren sowie deren Randbedingungen. Zur besseren Nachvollziehbarkeit wird das statistische Vorgehen kurz erläutert. Die statistische Auswertung wurde mittels Jamovi 2.3.2 (*jamovi – open statistical software for the desktop and cloud*, o. D.) und für die Studienplanung bzw. Analyse der statistischen Power der Hypothesentests G*Power V3.1.9.7 (*G*Power*, o. D.) durchgeführt. Als Prämisse für die inferenzstatistische Auswertung ist im Rahmen des überordneten Forschungsdesigns die Erklärbarkeit der eingesetzten Verfahren, Modelle sowie die Validität der Ergebnisse von großer Bedeutung. Entsprechend des Parsimony-Prinzips (vgl. Ockhams Rasiermesser) gilt es, unnötige Komplexität zu vermeiden (Sober, 1981). Aus diesem Grund wurden probate Hypothesentests bzw. Varianzanalysen anstelle konfirmatorischer Faktorenanalysen bzw. Strukturgleichungsmodelle bevorzugt. Laut Wirth und Klein (2023, S. 3 ff.) werden Hypothesentests immer dann als Mittel der Wahl eingesetzt, wenn streng kontrollierte Experimente auf Signifikanz von Ergebnissen geprüft werden, z. B. bei Vergleich A oder B. Die Vorteile von Varianzanalysen überwiegen beim angewendeten Forschungsdesign: „Störvariablen werden im Experiment direkt eliminiert“, „Effektstärken können in ihren Real-World-Auswirkungen beurteilt werden“, wenn „(...) konkrete Hypothesen vorliegen, die man prüfen möchte“ (Wirth & Klein, 2023, S. 20). Vor allem stellt dieses Forschungsdesign für praktische Anwendungen ein belastbares Instrument dar (Wirth & Klein, 2023). Dabei wurde besonderer Wert auf die Prüfung der Voraussetzungen und die Wahl robuster Methoden sowie die statistische Validität (Power) durch die formale Stichprobenumfangsplanung gelegt (Döring & Bortz, 2016). In Anlehnung an Döring und Bortz (2016, S. 821 ff.) wird folgendes Vorgehen bei der empirischen Studienplanung angewendet:

Nach Festlegung der Hypothese mit der zugehörigen Nullhypothese wird zuerst versucht, aus der kritischen Literaturrecherche vergleichbare bzw. übertragbare,

standardisierte Effektstärken für das jeweilige DOE zu extrahieren. Wenn keine geeigneten Referenzen verfügbar sind, wird eine Vorstudie zur Abschätzung der zu erwartenden Effektstärke durchgeführt. Im Anschluss wird a priori via G*Power und mit vordefinierten Parametern u. a. einem Signifikanzniveau von $\alpha = 0,05$ und einer Teststärke (Power) von $1 - \beta = 0,8$ die benötigte Stichprobengröße geplant. Nach Erhebung der Studienergebnisse wird eine Post-hoc-Teststärkenanalyse (Power) durchgeführt, welche Aufschluss über die statistische Validität der Signifikanztests und die Ergebnisse bei realen Stichproben gibt. Im Folgenden werden die eingesetzten Verfahren unter Berücksichtigung von Davis (2023) kurz erläutert:

Effektstärke: Cohen's d (Co'd) ist eine standardisierte Effektgröße, die dazu dient, die Größe eines Unterschieds zwischen zwei Gruppen im Verhältnis zur Variabilität der Daten zu quantifizieren. Sie basiert auf der Differenz der Mittelwerte der beiden Gruppen, relativiert durch die Standardabweichung, wodurch Unterschiede unabhängig von der Skala oder Maßeinheit der Messdaten vergleichbar gemacht werden können (Cohen, 1988, S. 37). Bei ungleichen Standardabweichungen zwischen den Gruppen wird üblicherweise die gepoolte Standardabweichung verwendet, die eine gewichtete Schätzung der Varianzen der beiden Gruppen darstellt (Cohen, 1988, S. 84). Ein zentraler Aspekt von Co'd besteht darin, dass die Effektgröße unabhängig von der Stichprobengröße bleibt und somit eine objektive Beurteilung der praktischen Relevanz des Unterschieds ermöglicht. Ein kleiner Effekt (Co'd = 0.2) zeigt, dass der Unterschied zwischen den Gruppen gering ist und oft schwer wahrnehmbar sein kann. Ein moderater Effekt (Co'd = 0.5) deutet auf einen deutlicheren, jedoch nicht dominanten Unterschied hin. Ein großer Effekt (Co'd = 0.8 oder höher) weist auf einen spürbaren Unterschied zwischen den Gruppen hin (Cohen, 1988).

Teststärke (Engl. statistical power): Die Power hängt von der Effektgröße, der Stichprobengröße, dem Signifikanzniveau α , der Variabilität der Daten und der Testart ab. Eine höhere Power reduziert das Risiko, einen vorhandenen Effekt zu übersehen. In der Praxis wird meist eine Power von mindestens 80 % angestrebt, um aussagekräftige Ergebnisse zu gewährleisten (Döring & Bortz, 2016).

Analysis of Variance (ANOVA): Die ANOVA dient der Analyse von Mittelwertunterschieden zwischen mehreren Faktorstufen. Sie setzt voraus, dass die Daten

normalverteilt und die Varianzen zwischen den Gruppen homogen sind. Bei großen Stichproben ($N > 25$) kann die ANOVA robust gegenüber Verletzungen der Normalverteilung angenommen werden (Schmider et al., 2010, S. 2–4).

Repeated Measures ANOVA: Die Repeated Measures ANOVA wird angewandt, wenn abhängige Variablen (AV) mehrfach bei denselben Probanden gemessen werden, z. B. bei unterschiedlichen Bedingungen oder Messzeitpunkten. Diese Methode berücksichtigt die Korrelation zwischen den Messungen bei den Probanden. Sie ist ebenfalls robust gegenüber Verletzungen der Normalverteilung, insbesondere bei wenigen Faktorstufen und großen Stichproben (Blanca et al., 2017, S. 3; Blanca et al., 2023, S. 6; Janczyk & Pfister, 2023). Die Prüfung der Sphärizität ist hierbei wichtig. Bei Verletzungen kann die Greenhouse-Geisser-Korrektur angewendet werden.

t-Tests: Unabhängige *t*-Tests werden eingesetzt, um Mittelwerte zwischen zwei unabhängigen Gruppen zu vergleichen. Die Anwendung dieses Verfahrens setzt ebenfalls voraus, dass die Daten normalverteilt und die Varianzen der Gruppen homogen sind. Im Gegensatz dazu wurden gepaarte *t*-Tests genutzt, um Mittelwerte zwischen zwei abhängigen Gruppen zu vergleichen, beispielsweise bei Messungen derselben Probanden vor und nach einer Intervention. Gepaarte *t*-Tests sind gegenüber Verletzungen der Normalverteilung weniger empfindlich als unabhängige *t*-Tests, da sie die Abhängigkeit der Messungen innerhalb der Gruppen berücksichtigen. Bei Verletzung der Varianzhomogenität wurde der Welch-*t*-Test verwendet, der für den Vergleich von Mittelwerten zwischen zwei Gruppen geeignet ist und robuste Ergebnisse liefert, wenn die Varianzen stark unterschiedlich sind. Falls die Annahme der Normalverteilung verletzt war, wurden nichtparametrische Tests eingesetzt. Der Mann-Whitney-*U*-Test wurde verwendet, um zwei unabhängige Gruppen zu vergleichen, während der Wilcoxon-Rang-Test für den Vergleich von zwei abhängigen Gruppen genutzt wurde (Pagano, 2013). Für den Vergleich von mehr als zwei unabhängigen Gruppen wurde der Kruskal-Wallis-Test als nichtparametrisches Pendant zur ANOVA herangezogen (Institut für Statistik, 2020). Bei nichtparametrischen Tests wird analog zur Effektstärke der biseriale Rangkorrelationskoeffizient (BRK) gemessen. Dieser wird verwendet, um den Zusammenhang zwischen einer dichotomen Variablen (z. B. Ja/Nein, 0/1) und einer ordinalen oder rangskalierten Variablen zu messen. Zur Sicherstellung der Vergleichbarkeit der Ergebnisse können

BRK näherungsweise u. a. auf die standardisierte Effektstärke $Co'd$ umgerechnet werden, wie Gleichung 6 verdeutlicht (Döring & Bortz, 2016, S. 819; Rosenthal & Rosnow, 2008, S. 377).

$$d = \frac{2 * r}{\sqrt{1 - r^2}}, \quad (6)$$

mit r : biserialer Rangkorrelationskoeffizient (BRK)

Post-Hoc-Tests wurden durchgeführt, um signifikante Gruppenunterschiede nach einer ANOVA zu identifizieren. Eine besondere Rolle spielte dabei die Korrektur des Alpha-Fehlers, um die Wahrscheinlichkeit eines Fehlers 1. Art (fälschliche Ablehnung der Nullhypothese) zu minimieren. Der Tukey-Test wird angewandt, wenn die Varianzen zwischen den Gruppen homogen und die Stichprobengrößen gleich sind (Lee & Lee, 2018). Für Szenarien mit Varianzheterogenität oder ungleichen Gruppengrößen kommt der Scheffé-Test zum Einsatz, der als eine konservative Methode zur Kontrolle des Alpha-Fehlers gilt (Staffa & Zurakowski, 2020). In Fällen, in denen die Normalverteilung oder Varianzhomogenität verletzt ist, wird der Dwass-Steel-Critchlow-Fligner-Test (DSCF) verwendet. Dieser nichtparametrische Post-Hoc-Test ist speziell für solche Verletzungen konzipiert und bietet eine robuste Alternative zu parametrischen Post-Hoc-Tests (Lüpsen, 2019).

Fehlstellen- und Ausreißerbereinigung: Um statistische Verzerrungen durch extreme Werte zu minimieren, wird sog. Winsorizing als standardisierte Methode zur Ausreißerbereinigung eingesetzt. Hierbei werden Werte, die über definierten Grenzwerten liegen, durch die nächstgelegenen gültigen Werte ersetzt. Die Berechnung der Grenzwerte basierte auf der Interquartilsdifferenz (IQR), wobei Grenzwerte entsprechend der Gleichungen 7–8 definiert wurden.

$$\text{oberer Grenzwert} = Q3 + 1.5 * IQR, \quad (7)$$

$$\text{unterer Grenzwert} = Q1 - 1.5 * IQR. \quad (8)$$

Besonders wichtig war die Anwendung dieser Methode bei sensitiven, physiologischen Daten wie dem Hautleitwert (SCL) und der Herzfrequenz (HR), da diese Parameter anfälliger für Anomalien und Messartefakte sind. Die Nutzung des Winsorizing stellte somit sicher, dass die Analysen durch einzelne Extremwerte nicht verfälscht bzw. überinterpretiert wurden („Winsorizing“, 2018; Yang et al., 2019). Fehlstellen, sog.

Missing Values, werden stets durch die entsprechenden Gruppenmittelwerte gehoben (Göthlich, 2009, S. 123).

Darstellung der Ergebnisse: Zur Darstellung der Ergebnisse wurden Boxplots, Barplots und Mittelwertdiagramme verwendet. Boxplots zeigen Median, Quartile und die Verteilung der Daten; der Mittelwert wurde zusätzlich markiert. Barplots stellen Gruppenmittelwerte dar und verwenden Standardabweichungen als Fehlerindikatoren zur Darstellung der Streuung innerhalb der Gruppen. Mittelwertdiagramme mit Konfidenzintervallen ermöglichen die differenzierte Darstellung von Gruppenunterschieden und potenziellen Interaktionen. Die Konfidenzintervalle geben dabei die statistische Präzision der Mittelwertschätzung an und erlauben Rückschlüsse auf mögliche Signifikanzen zwischen den Vergleichsgruppen.

Eingesetzte Hardware

Als Hardware für die Studien wurden als Augmented-Reality-HMD die Microsoft HoloLens 2 (*HoloLens 2*, o. D.) (HL2) und als Virtual-Reality-HMD die Oculus Quest 2 von Meta Inc. (*Oculus Meta Quest 2*, o. D.) eingesetzt. Die Messung der physiologischen Parameter wurde mittels E4 Wristband von Empatica (*E4 Wristband Empatica*, o. D.) durchgeführt.

4 Empirische Studien

Abgeleitet aus dem übergeordneten Forschungsdesign wurde ein konsekutiver Studienplan (siehe Tabelle 2) entwickelt. Die Studienfolge ergab sich zum einen aus der aufsteigenden Komplexität des Studiendesigns, zum anderen aus dem Entwicklungsstand der Softwareforschungsprototypen, deren Basis u. a. aus der AR-Plattform des Forschungsprojekts PlanAR (Förderträger: BayVFP Förderlinie Digitalisierung / FuE – Programm „Informationstechnik und Kommunikationstechnik“ des Freistaates Bayern IUK-1811-0022, PlanAR, o. D.) bereitgestellt wurde.

Studie/MR	Design	UV/Fehler/Konf.	AV	FF	Hypothesen erhoben am
1/VR	BSD	VQ/ F_2	P_{1-3} ; E_{1-4}	$FF_{1.B}$	$H_{B.1-3}$ KW 21/2021
2/AR	CWSD	IQ/ F_3	P_{4-5} ; E_{1-4}	$FF_{1.C}$	$H_{C.1-3}$ KW 34/2022
3/AR	BSD	TQ/ F_4	P_{1-3} ; E_{1-4}	$FF_{1.D}$	$H_{D.1-3}$ KW 25–26/2022
4.1/AR	MMD	DQ/ F_1 ; VQ/ F_2 ; IQ/ F_3	P_{1-3} ; E_{1-4}	$FF_{1.A-C}$	H_{A-C} KW 43–49/2023
4.2/AR	BSD	DQ/ F_1 ; VQ/ F_2 ; IQ/ F_3	P_{1-3} ; E_{1-4}	FF_2	H_{1-3} KW 43–49/2023
4.3/AR	BSD	M/Konf.	E_1	FF_3	H_4 KW 49–50/2023

Fußnote. between subject design (BSD); counterbalanced within-subject design (CWSD); Mixed Method Design (MMD)

Tabelle 2. Gesamtübersicht des Studiendesigns

Zur Einordnung der Studienentwicklung in die Methodik werden die Studien nachfolgend kurz skizziert. In den Studien 1–3 sollen die Effekte der Sichtbarkeits- (F_2), Interaktions- (F_2) und Trainingsfehler (F_2) einzeln untersucht werden. Hierbei wurde anhand Studie 1 der empirische Grundansatz des Forschungsdesigns entwickelt und getestet. Aufgrund des umfangreichen, komplexen Studiendesigns hinsichtlich des Aufbaus und der Durchführung zur Bewertung der Sprachinteraktionsqualität sowie der Entwicklung der Faktorstufen der TQ wurden die Studien 2 und 3 separat durchgeführt. Studie 4 wurde als Verbundstudie ausgelegt mit dem Ziel, unterschiedliche Fehlertypen in einem Versuchsaufbau unter vergleichbaren Bedingungen zu untersuchen.

Studie 1 – Sichtbarkeitsfehler F_2 : Bei dieser Studie wurde der Störeinfluss des Sichtbarkeitsfehlers F_2 durch die Manipulation VQ in zwei Faktorstufen im BSD untersucht. Zur Minimierung weiterer Störeinflüsse auf das Forschungsdesign und zur Fokussierung der Probanden auf die immersive Wahrnehmung wurde ein VR-Forschungsprototyp mit einer sehr hohen optischen Repräsentationsgüte gewählt. Als

Basisforschungsprototyp¹ wurde die VR-Rekonstruktion der Ausstellung „Bauhausmädel“ aus dem Anger Museum Erfurt eingesetzt. Von der Museumanwendung wurde eine angepasste Version mit wesentlicher Beeinträchtigung der VQ abgeleitet, welche als Faktorstufe Sichtbarkeitsfehler F_2 eingesetzt wurde.

Studie 2 – Interaktionsfehler F_3 : Wie vorher beschrieben, werden aktuell zwei Interaktionsparadigmen für den industriellen Einsatz favorisiert: die Gestensteuerung und die Sprachsteuerung. Für die Studie 2 wurde eine angepasste Variante des PlanAR-Forschungsprototyps entwickelt, um den Interaktionsfehler F_3 bei der Sprachinteraktion in industriellen Umgebungen zu untersuchen. Hierbei wurde der Störeinfluss nicht artifiziell durch Systemparameter implementiert, sondern durch die limitierende Fähigkeit hinsichtlich des industriellen Umgebungslärms indirekt manipuliert. Im Forschungsprojekt PlanAR wurden die Einflüsse des Schalldruckpegels, der Stationarität sowie der Lärmquelle auf die effektive Leistung und Gebrauchstauglichkeit der Spracheingabe auf der HL2 mittel CWSD untersucht. Aus den Vorstudien des Forschungsverbunds wurden die entsprechenden Faktorstufen der IQ jeweils mit und ohne Interaktionsfehler über die Manipulation des Schalldruckpegels abgeleitet.

Studie 3 – Trainingsfehler F_4 : In dieser Studie wird der Einfluss der TQ in einem BSD untersucht. Hierzu wurden drei Faktorstufen entwickelt. Die erste Stufe stellt den Trainingsfehler F_4 dar und wurde inhaltlich korrekt, jedoch minimal in Umfang und didaktischer Güte ausgeführt. Als zweite Stufe wurde ein persönliches Trainingsszenario eingesetzt, welches sehr ausführlich und umfangreich schult. Als dritte Faktorstufe wurde ein optimiertes Trainingsszenario entwickelt, bei dem ein optimierter Methodenmix angewendet wurde. Dieses optimierte Trainingsszenario wurde in umfangreichen Vorstudien innerhalb des Forschungsprojekts PlanAR entwickelt. Als fachliche Aufgabe wurden die Probanden gebeten, verschiedene Teilaufgaben durchzuführen, u. a. das

¹ Dieser Forschungsprototyp wurde dankenderweise von Prof. Dr. Yvonne Brandenburger aus dem Forschungsprojekt „Virtual Reality für Thüringer Museen: Entwicklung, Leitfaden und Implementierung von Pilotanwendungen“ zur Verfügung gestellt (Universität Erfurt, o. D.-b). Dieses Projekt ist ein interdisziplinäres Kooperationsprojekt der Fachgebiete Angewandte Informatik und Architektur der Fachhochschule Erfurt und der Medien- und Kommunikationswissenschaft der Universität Erfurt, gefördert durch das Landesprogramm „ProDigital“ des Thüringer Ministeriums für Wirtschaft, Wissenschaft und Digitale Gesellschaft (TMWWDG).

Abrufen von Information aus dem Internet sowie die Durchführung mehrerer industrieller Messaufgaben via immersive Anwendungen. Diese AR-Messanwendungen wurden ebenfalls als Softwareforschungsprototypen aus dem PlanAR-Projekt bereitgestellt.

Studie 4 – Fehlervergleich und Akzeptanzmanipulation: Bei der Studie 4 handelt es sich um eine Verbundstudie bestehend aus den drei Teilstudien 4.1, 4.2 und 4.3, bei der die Fehlereinflüsse F_{1-3} in einem MMD (Murrar & Brauer, 2018) hinsichtlich der drei Zielsetzungen untersucht werden. Das erste Ziel der Teilstudie 4.1 ist, die Fehlerauswirkungen an einem Forschungsprototypen sowie einer fachlichen Aufgabe zu bewerten. Aufgrund der Komplexität und des Umfangs der UV TQ wurde der Trainingsfehler hierbei nicht miteinbezogen. Entsprechend dem zweiten Ziel (Teilstudie 4.2) soll geprüft werden, ob Unterschiede zwischen den Fehlerarten hinsichtlich der abhängigen Variablen nachgewiesen werden können. Als drittes Ziel (Teilstudie 4.3) wird entsprechend den Ergebnissen aus Teilstudie 4.1 der Fehler mit der höchsten Effektstärke gewählt und der Einfluss der Konfidenzanzeige als Moderatorvariable $\mathbf{M}$ auf die Akzeptanz untersucht.

Hinweis: Aufgrund der Lesbarkeit und zum besseren Verständnis innerhalb der einzelnen Studien wird im Folgenden auf den Bezug zum übergeordneten Forschungsdesign durch die Nennung der Syntax verzichtet.

4.1 Studie 1 – Sichtbarkeitsfehler I

Das Hauptziel der Studie 1 ist, die Fehlerauswirkungen des Sichtbarkeitsfehlers hinsichtlich der Leistung, Akzeptanz und Beanspruchung zu untersuchen, vgl. $H_{B.1-3}$ bzw. $FF_{1.B}$. Im Rahmen des Forschungskontextes hinsichtlich allgemeiner Sichtbarkeitsfaktoren immersiver Systeme wurden bereits zwei Vorstudien durchgeführt:

1. In der Studie von Rosilius et al. (2021) wurde ein immersiver Landolt-Sehtest auf der HL2 als Forschungsprototyp entwickelt, auf dessen Basis die Sichtbarkeit unter industriellen Anwendungsbedingungen wie Beleuchtungsstärke, Graustufen der Augmentierung, virtueller Hinterlegung und Oberflächenmaterialien untersucht wurde. Neben der Entwicklung eines ersten Forschungsprototyps zur Messung immersiver Sichtbarkeit mittels der Leistungsfaktoren Antwortzeit und Richtigkeit sowie primärer Studienergebnisse wurden flankierend initiale Akzeptanzerhebungen durchgeführt. Innerhalb dieses DOE konnten keine

signifikanten Unterschiede hinsichtlich BI sowie weiterer Dimensionen des UTAUT2 festgestellt werden.

2. Konsekutiv aufbauend darauf wurde die Studie von Rosilius et al. (2022) entwickelt und durchgeführt. Hieran wurde der Index Sichtbarkeit als sogenannter Quality of Service novelliert, dessen adaptierte Metrik als Index Leistung einen zentralen Bestandteil dieser Dissertation darstellt.

4.1.1 Methodik

In dieser Studie wurden die Probanden im Between-Subject-Design gebeten, entweder mit oder ohne Fehlerbedingung hinsichtlich der Sichtbarkeit, sich fachliche Inhalte aus einer Museumsszene (siehe Abbildung 4) so schnell und präzise wie möglich einzuprägen. Im Anschluss beantworteten sie einen Fragebogen als Lernzielkontrolle.

VR-Forschungsprototyp: Als Grundlage des Forschungsprototyps wurde eine VR-Museumsanwendung (virtuelle Rekonstruktion der Ausstellung „Bauhausmädels“ aus dem Anger Museum Erfurt (Universität Erfurt, o. D. -b) entsprechend der UV VQ so adaptiert, dass hieran die Auswirkungen des Sichtbarkeitsfehlers zur Beantwortung der Forschungsfrage $FF_{1.B}$ bzw. die Prüfung der Hypothesen $H_{B.1-3}$ durchgeführt werden konnten. Die VR schirmt die reale Welt vollständig ab, wodurch ein höherer Grad der Immersion erreicht wird. Dies bewirkt eine kontrollierbare und ablenkungsarme Umgebung für das Studiendesign (Slater & Sanchez-Vives, 2016, S. 18–35). Somit können präzise Studienbedingungen realisiert werden. AR-Szenen werden Nachteile durch äußere Störeinflüsse zugesprochen, wie eingeschränktes Sichtfeld, verzerrte räumliche Wahrnehmung, Diskrepanz zwischen visuellem und taktilen Feedback etc. (Caputo et al., 2020, S. 7). An diesen Forschungsprototyp werden die Anforderungen gestellt, ein reproduzierbares, qualitativ hochwertiges, immersives Präsenzerlebnis mit minimalen Störeinflüssen durch die Maskierung der realen Welt sowie einer nativen Fortbewegung zu schaffen. In dieser virtuellen Welt können sich die Probanden auf die visuellen Inhalte konzentrieren, sodass die UV VQ effektiv wahrgenommen werden kann. Diese Museumsausstellung bietet zudem einen hohen Grad an Realismus und optischer Repräsentation mit einer nachvollziehbaren Logik, sodass hierdurch die Präsenzwahrnehmung gefördert wird (von Eitzen, 2024, S. 84), wie in Abbildung 4 dargestellt. Nach von

Eitzen (2024) bedingen die VR-Konstrukte Realismus, Präsenz, Immersion, und Involvement bzw. Aufmerksamkeit einander. Dank dieses qualitativ hochwertigen VR-Forschungsprototyps kann die Aufmerksamkeit fokussiert auf die virtuellen Inhalte gelenkt werden, was das Präsenzerlebnis verstärkt (von Eitzen, 2024, S. 165). Auch wenn es sich bei der betrachteten Museumsanwendung nicht um ein klassisches AS handelt, weist sie aufgrund ihrer strukturellen und funktionalen Merkmale eine hohe Ähnlichkeit zu didaktisch anspruchsvoll gestalteten VR-Lernumgebungen im industriellen Kontext auf. Solche industriellen VR-Schulungsanwendungen können als autodidaktische IAS verstanden werden, die Nutzer bei komplexen Lernprozessen unterstützen. In vergleichbarer Weise erfüllt die Museumsanwendung im nicht-industriellen Kontext die Anforderung, Informationen in immersiver Form zu vermitteln, sodass der Nutzer als Rezipient vollständig in die Inhalte eintauchen kann. Ziel ist es, eine konzentrierte, ablenkungsfreie Auseinandersetzung mit dem präsentierten Wissen zu ermöglichen. Im Rahmen des Forschungsprojekts „Virtual Reality für Thüringer Museen: Entwicklung, Leitfaden und Implementierung von Pilotanwendungen“ (Universität Erfurt, o. D.-b) war eines der zentralen Ziele, das physische Museumserlebnis möglichst authentisch in einer virtuellen Umgebung nachzubilden.



Abbildung 4. VR-Startperspektive (ED) – S1

Design of Experiment

Die Studie wurde in einem BSD durchgeführt. Hierbei wurde innerhalb der UV VQ durch die beiden Faktorstufen fehlerhaft (fh), sprich mit Sichtbarkeitsfehler, und fehlerfrei (ff) zwei Gruppen gebildet. Die Probanden wurden randomisiert diesen beiden Gruppen zugeordnet. Die AVs wurden in Übereinstimmung mit dem übergeordneten Forschungsdesign (vgl. Kapitel 3.2) wie folgt definiert: Erfasst wurden die Durchlaufzeit (DLZ), die Richtigkeit der Aufgabenbearbeitung, der Index Leistung sowie die Akzeptanz (BI) als zentraler Indikator der Systemakzeptanz im Sinne des UTAUT2-Modells. Ergänzend wurden weitere sekundäre Akzeptanzfaktoren aus dem UTAUT2 berücksichtigt. Zur Messung der subjektiven Beanspruchung wurde der Mittelwert der NASA-TLX-Skala (Mean_NASA) einbezogen, einschließlich seiner Subskalen. Darüber hinaus wurden die physiologischen Parameter SCL und HR als objektive Indikatoren kognitiver und emotionaler Belastung erfasst.

Potenzielle Störeinflüsse: Per Design wurde versucht, die potenziellen Störeinflüsse zu kontrollieren bzw. minimieren. Die Probanden sollen sich ausschließlich auf die fachlichen Inhalte der AR-Szene konzentrieren. Um einen höheren Grad an Immersion und Realismus zu erwirken, wurde die Szene akustisch mit der monotonen Soundkulisse einer Museumsumgebung ausgestattet. Da das Virtual-Reality-HMD über mehrere Lautsprecher verfügt, wurde hierdurch versucht, ein Präsenzerlebnis mit einem stärkeren Involvement zu schaffen. Zudem konnten sich die Probanden nativ und ohne Steuerrungskontroller in der VR-Szene bewegen. Dies sollte den Störeinfluss eines ungewohnten Interaktionsparadigmas reduzieren und die Erfahrung eines realen Museumsbesuches suggerieren. Außerdem wurde der virtuelle Begrenzungszaun, der sog. Guardian (Tseng, 2023, S. 1–2), so gesetzt, dass der Proband alle Exponate frei begehen konnte. Insgesamt wurde der Anforderung Rechnung getragen, die reale Welt optisch und akustisch aus der Wahrnehmung des Probanden zu maskieren und die Fokussierung auf die fachlichen Inhalte ohne Beeinträchtigung zu fördern.

Visuelle Qualität: Zwei Faktorstufen wurden definiert: ff mit einer hohen Repräsentationsgüte für eine gute Sichtbarkeit bzw. Lesbarkeit der Inhalte sowie fh mit einer niedrigen Repräsentationsgüte für eine unzureichende Sichtbarkeit (vgl. Abbildung 5). Die Faktorstufen wurden durch die VR-Szeneneinstellungen in Blender angepasst. Hierbei wurden die Qualitäts- bzw. Detailstufen u. a. durch die Manipulation der Schriftgröße, die Darstellungsfläche, das Kontrastverhältnis, die Helligkeit, die Anzahl und Intensität der Lichtquellen entsprechend den beiden Faktorstufen realisiert. Hinsichtlich der Wiedergabeparameter der VR-Brille wurden stets die Standardparameter wie die Standardeinstellung der Texturgröße, 72 Hz Bildwiederholrate, standardmäßiger Fixed-Foveated-Rendering-Wert sowie aktivierte chromatische Aberrationskorrektur gewählt.



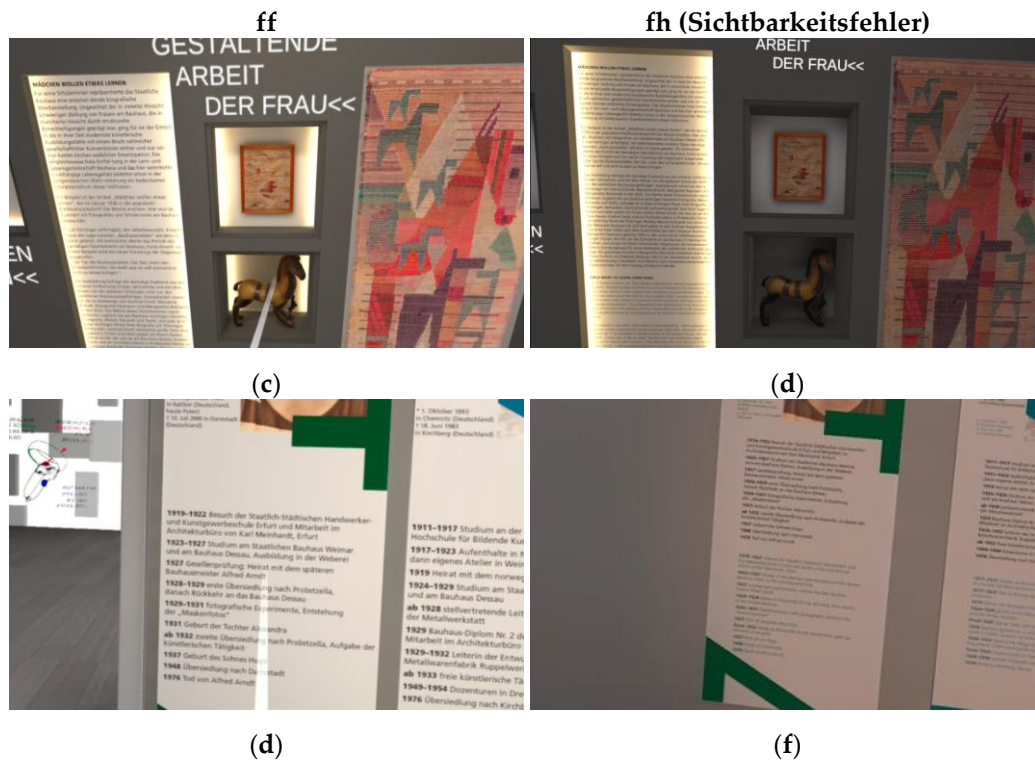


Abbildung 5. VR-Detailansichten versus VQ (ED) – S1

Probandenbefragung: Neben den Umfängen des UTAUT2 (siehe Tabelle 56) und NASA-TLX (siehe Tabelle 57) werden in dieser Studie ebenfalls die soziodemografischen Daten (siehe Tabelle 58), die Sehfähigkeit sowie die Leistungsabfragen (siehe Tabelle 59) erhoben.

Versuchsaufbau und -ablaufplan

Die Studie wurde in Laborumgebungen mit einer freien Fläche von mindestens 10 x 10 Metern durchgeführt. Der detaillierte Ablauf ist in Abbildung 6 dargestellt. Zu Beginn der Studie füllten die Probanden einen initialen Fragebogen am Laptop aus. Anschließend erfolgte eine Ruhephase zur Erfassung der physiologischen Ausgangswerte (Baseline). Der Livestream der Oculus Quest 2 wurde über ein Smartphone übertragen. Im nächsten Schritt erhielten die Teilnehmenden eine Einweisung in den Versuchsablauf. Diese erfolgte sowohl direkt durch den Versuchsleiter als auch über den Livestream, sodass die Probanden den Ablauf gut nachvollziehen konnten. Zur Förderung der extrinsischen Motivation wurde ein kompetitiver Charakter der Aufgabe vermittelt. Die Teilnehmenden wurden dazu angehalten, sich die Museumsinhalte möglichst schnell und präzise einzuprägen. Die individuelle Leistung sollte anschließend im Vergleich zur

Gesamtkohorte bewertet werden. Als Anreiz wurde eine Prämie von zehn Euro für jene Probanden ausgelobt, deren Ergebnisse im oberen Quartil lagen. Unabhängig davon erhielten alle Teilnehmenden eine Aufwandsentschädigung. Im Anschluss wurde die eigentliche Datenerhebung durchgeführt. Während der Durchführung zeichnete die Empatica-E4-Armbanduhr kontinuierlich physiologische Parameter auf. Sobald die Probanden angaben, die virtuelle Museumstour abgeschlossen zu haben, wurde die Zeitmessung gestoppt und der Fragebogen zur Lernzielkontrolle bearbeitet. Abschließend füllten die Teilnehmenden die Fragebögen UTAUT2 und NASA-TLX aus. Mit deren Abschluss war der Versuch beendet.

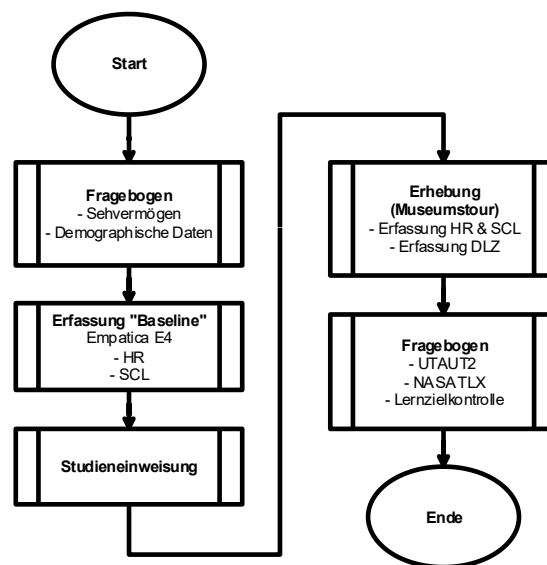


Abbildung 6. Versuchsablaufplan S1 (ED)

Versuchsdurchführung und Stichprobenumfangsplanung

Das Studiendesign wurden unter dem Einfluss der Coronapandemie entwickelt und durchgeführt (Q1 2021). Aufgrund der erschwerten Bedingung (u. a. Akquirierung von Probanden etc.) konnte keine formale Vorstudie durchgeführt werden.

Stichprobenumfangsplanung: Aufgrund des Fehlens geeigneter Effektstärken aus vorherigen Studien wurde zur Schätzung des Stichprobenumfangs auf Studien verwandter Technologien bzw. Forschungsfelder zurückgegriffen. Die Ergebnisse von Arning et al. (2012) zeigen, dass die visuelle Wahrnehmung beim Projektor schlechter ist als beim Monitor. Anhand dieser Daten konnten via G*Power Version 3.1.9.7 Effektstärken hinsichtlich der Leistungsunterschiede von $Co'd = 0,63$ und für die Akzeptanzunterschiede

von $Co'd = 0,5$ geschätzt werden. Diese Effektstärken wurden zur Abschätzung der Stichprobenplanung herangezogen (siehe Tabelle 3).

Statistische Parameter	Leistung	Akzeptanz
$Co'd$	0,63	0,50
Power	0,80	0,80
α	0,05	0,05
Anzahl der Gruppen	2	2
kalkulierter Stichprobenumfang	64	102

Tabelle 3. Stichprobenumfangsplan – S1

Alle Datensätze wurden anonymisiert erfasst. Jeder Teilnehmer wurde über die Richtlinien guter ethischer Forschung informiert und musste seine ausdrückliche Zustimmung zur Teilnahme geben (vgl. Abbildung 78). Die Dauer des Experiments betrug ca. 45 Minuten. Die Geräte wurden nach jedem Teilnehmer vollständig desinfiziert. Die Studie wurde sowohl an der Technischen Hochschule Würzburg-Schweinfurt als auch an der Fachhochschule Erfurt in der KW 21 2021 erhoben. Auch wenn ein Stichprobenumfang nach Vorplanung in Höhe von 64 veranschlagt war, konnte dieser aufgrund fehlender Probanden trotz mehrerer Aufrufe zur Probandengewinnung unter den Coroneinschränkungen nicht realisiert werden. Die Studie wurde beendet und die vorliegenden Ergebnisse wurden ausgewertet. Da die Ergebnisse differenzierte statistische Reaktionen mit aussagekräftigen Effektstärken zeigten und eine gute Vergleichbarkeit zum AR-Sichtbarkeitsfehler aus Studie 4 (vgl. Kapitel 4.4.2.1) vorlag, wurde die Studie nicht zu einem späteren Zeitpunkt fortgeführt.

Statistisches Vorgehen

Zu den Ausführungen in Kapitel 3.2 werden in diesem Kapitel studienspezifische Besonderheiten erwähnt. Zur Studienaushwertung wurde ein unabhängiger t -Test (gerichtet) herangezogen. Zur Überprüfung der Annahmen wurden im Vorfeld Normalität und Homogenität der Varianzen getestet. Wenn die Homogenität der Varianzen verletzt wurde, wurde ein Welch t -Test angewendet. War hingegen die Normalverteilung verletzt, wurde ein Mann-Whitney- U -Test durchgeführt. Dieser wurde für die Auswertung der Leistungsfaktoren, Akzeptanzfaktoren des UTAUT2 sowie Beanspruchungsfaktoren des NASA-TLX herangezogen. Ebenfalls für die Untersuchung der Unterschiede innerhalb der UV VQ wurden für die physiologischen Parameter SCL und HR t -Tests

beziehungsweise Mann-Whitney-*U*-Tests angewendet. Für die nichtparametrischen Tests wurde nach Gleichung 6 der BRK in Co'd umgerechnet.

4.1.2 Ergebnisse

Demografische Daten

Für die Hauptstudie wurden insgesamt $N = 18$ Teilnehmer, 13 männlich und 5 weiblich, über eine regionale Studentenmailingliste rekrutiert. Hiervon hatten 33,33 % eine abgeschlossene Berufsausbildung und waren berufstätig, 66,67 % studierten. Die Teilnehmenden waren zwischen 20 und 39 Jahre alt, ihr Durchschnittsalter lag bei 26,7 Jahren.

Deskriptive Daten (DD)

Die deskriptiven Ergebnisse sind in Tabelle 4 dargestellt. Die DLZ ist beim Fehlerfall ff im Mittel länger ($M = 476$; $SD = 130,41$) als bei fh ($M = 460,2$; $SD = 97,21$). Die Richtigkeit hat im Fehlerfall einen höheren Mittelwert ($M = 0,80$; $SD = 0,152$) als bei fh ($M = 0,529$; $SD = 0,136$). Ebenso zeigt der Index Leistung bei ff einen höheren Mittelwert ($M = 0,19$; $SD = 0,098$) im Vergleich zu fh ($M = 0,12$; $SD = 0,049$).

AV	VQ	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
DLZ [s]	ff	8	476,00	130,41	46,11
	fh	10	460,20	97,21	30,74
Richtigkeit	ff	8	0,80	0,15	0,05
	fh	10	0,53	0,14	0,04
Leistung [%/s]	ff	8	0,19	0,10	0,04
	fh	10	0,12	0,05	0,02

Tabelle 4. DD-Leistungsfaktoren versus VQ – S1

Die deskriptiven Ergebnisse der UTAUT2-Dimensionen in Tabelle 5 zeigen nur geringe Unterschiede zwischen den beiden Bedingungen der VQ ff und fh. Bei der PE liegt der Mittelwert in der Gruppe fh (5,3) leicht über dem von ff (5,04), während bei der EE nahezu identische Mittelwerte festgestellt wurden (ff: 6,25; fh: 6,17). In der Dimension HM weist ff mit 6,58 einen höheren Mittelwert auf als fh (6,2), jedoch mit einer geringeren Streuung. Die BI Mittelwerte sind nahezu identisch (ff: 4,38; fh: 4,3).

Dimension	VQ	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
UTAUT2_PE	ff	8	5,04	1,25	0,44
	fh	10	5,30	1,06	0,34
UTAUT2_EE	ff	8	6,25	0,66	0,23
	fh	10	6,17	0,68	0,21
UTAUT2_SI	ff	8	3,08	1,45	0,51
	fh	10	3,27	1,09	0,34
UTAUT2_FC	ff	8	5,44	1,18	0,42
	fh	10	5,80	0,89	0,28
UTAUT2_HM	ff	8	6,58	0,50	0,18
	fh	10	6,20	1,01	0,32
UTAUT2_HT	ff	8	2,33	1,07	0,38
	fh	10	1,90	1,21	0,38
UTAUT2_BI	ff	8	4,38	1,09	0,39
	fh	10	4,30	1,27	0,40

Tabelle 5. DD UTAUT2 versus VQ – S1

Die deskriptiven Ergebnisse des NASA-TLX in Tabelle 6 zeigen deutliche Unterschiede zwischen den Faktorstufen der VQ fh und ff. Die Faktorstufe ff zeigt höhere Mittelwerte für die Subskalen PD (58,13 vs. 54,5) und P (16,25 vs. 11), während fh bei TD (17 vs. 7,5), E (74,5 vs. 63,75), F (37 vs. 21,25) und der durchschnittlichen Beanspruchung (39,08 vs. 36,98) höhere Mittelwerte aufweist. Besonders bei MD als psychischer Stresskomponente zeigt ff einen höheren Wert (55 vs. 40,5).

Dimension	VQ	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
Mean_NASA	fh	10	39,08	11,00	3,47
	ff	8	36,98	11,40	4,02
NASA_PD	fh	10	54,50	30,60	9,67
	ff	8	58,13	21,50	7,61
NASA_TD	fh	10	17,00	15,30	4,84
	ff	8	7,50	10,00	3,54
NASA_P	fh	10	11,00	21,20	6,70
	ff	8	16,25	21,20	7,49
NASA_E	fh	10	74,50	14,00	4,44
	ff	8	63,75	15,80	5,57
NASA_MD	fh	10	40,50	23,00	7,28
	ff	8	55,00	25,80	9,11
NASA_F	fh	10	37,00	20,60	6,51
	ff	8	21,25	21,70	7,66

Tabelle 6. DD NASA-TLX versus VQ – S1

Tabelle 7 präsentiert die deskriptiven, physiologischen Werte. Es zeigt sich, dass bei HR_Diff die Gruppe mit fehlerhafter visueller Qualität höhere Mittelwerte (12,61 Beats per Minute (BPM) vs. 8,98 BPM) und eine größere Streuung aufweist. Bei SCL_Diff zeigt die fehlerfreie Gruppe leicht höhere Mittelwerte (2,1 vs. 1,88) und eine größere Streuung.

Parameter	VQ	<i>n</i>	<i>M</i>	<i>SE</i>	<i>Median</i>	<i>SD</i>
HR_Diff	fh	10	12,61	13,17	7,07	2,23
	ff	8	8,98	9,57	7,08	2,50
SCL_Diff	fh	10	1,88	1,43	2,26	0,71
	ff	8	2,10	0,41	4,63	1,64

Tabelle 7. DD HR [BPM] & SCL [μ S] versus VQ – S1

Inferenzstatistische Ergebnisse

DLZ: Mittels des ungepaarten *t*-Tests konnte für die DLZ kein statistisch signifikanter Unterschied (gerichtet) zwischen den Faktorstufen der VQ nachgewiesen werden $t(16) = -0,295$; $p = 0,614$; $MD = -15,8$; $Co'd = -0,140$; $Power = 0,09$. Die Voraussetzungen der Normalverteilung und Varianzhomogenität wurden erfüllt. Die Mittelwerte für die ff-Sichtbarkeit lag bei $M = 476s$ ($n = 8$; $SD = 130,41$; $SE = 46,10$) und für die fh-Sichtbarkeit bei $M = 460s$ ($n = 10$; $SD = 97,2$; $SE = 30,7$). Abbildung 7 zeigt den Boxplot der DLZ.

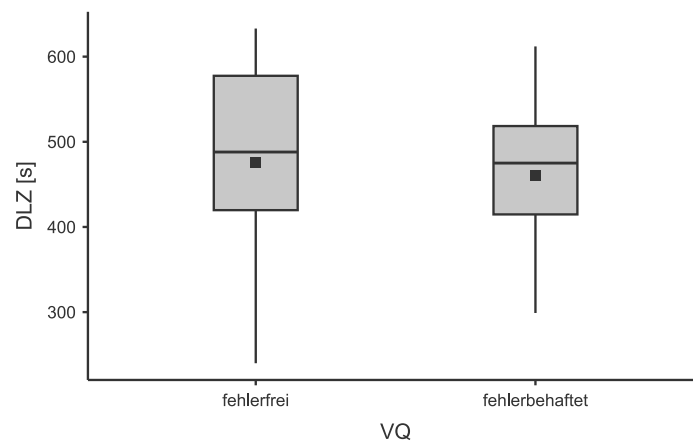


Abbildung 7. Boxplot DLZ [s] versus der VQ – S1

Richtigkeit: Mittels eines ungepaarten *t*-Tests konnte für die Richtigkeit ein statistisch signifikanter Unterschied (gerichtet) zwischen den Faktorstufen der VQ nachgewiesen werden $t(16) = 4,06$; $p < 0,001$; $MD = 0,275$; $Co'd = 1,927$; $Power = 0,99$. Es wurden beide Voraussetzungen, die Normalverteilung und die Varianzhomogenität, erfüllt. Die Mittelwerte für die fehlerfreie Sichtbarkeit lag bei $M = 0,786$ ($n = 8$; $SD = 0,151$; $SE = 0,053$) und für die fehlerbehaftete Sichtbarkeit bei $M = 0,529$ ($n = 10$; $SD = 0,135$; $SE = 0,043$). Abbildung 8 zeigt den Boxplot der Richtigkeit.

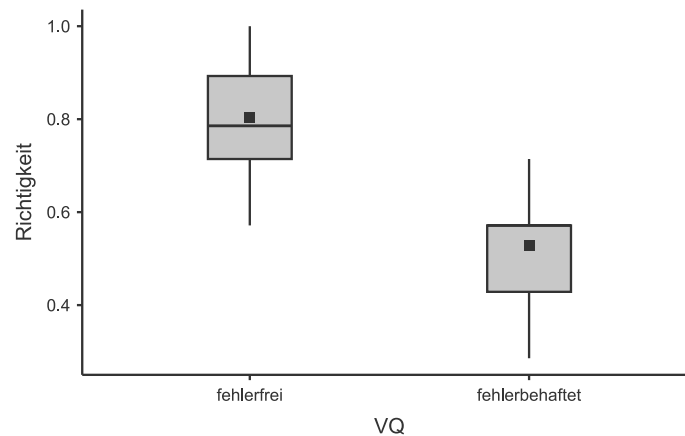


Abbildung 8. Boxplot-Richtigkeit versus VQ – S1

Leistung: Es konnte ein statistisch signifikanter Unterschied (gerichtet) mittels eines unabhängigen Mann-Whitney-*U*-Tests für den Index Leistung zwischen den Faktorstufen der VQ nachgewiesen werden ($U = 20$; $p = 0,042$; $MD = 0,046$ %/s; $BRK = 0,500$; $Co'd = 1,15$; $Power = 0,73$). Im Gegensatz zur Varianzhomogenität wurde die Voraussetzung der Normalverteilung nicht erfüllt. Die Mittelwerte für die ff-Sichtbarkeit lagen bei $M = 0,189\%/s$ ($n = 8$; $SD = 0,163$; $SE = 0,035$) und für die fh-Sichtbarkeit bei $M = 0,123$ %/s ($n = 10$; $SD = 0,049$; $SE = 0,015$). Abbildung 9 zeigt den Boxplot des Index Leistung.

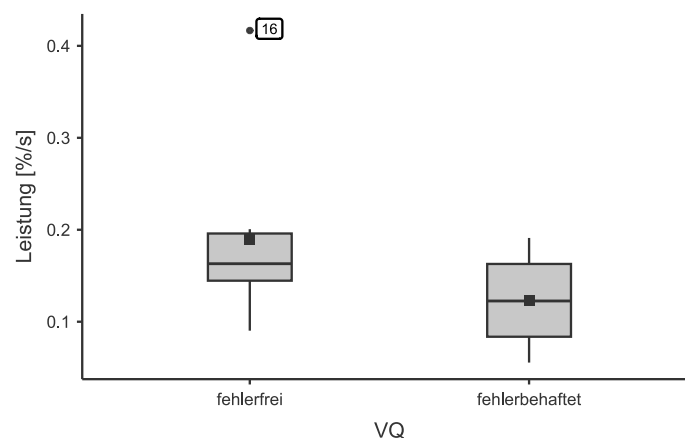


Abbildung 9. Boxplot-Leistung versus VQ – S1

UTAUT2: Alle Subdimensionen des UTAUT2-Fragebogens zeigten keinen statistisch signifikanten Unterschied zwischen den beiden Faktorstufen der VQ. Die Ergebnisse der unabhängigen *t*-Tests (gerichtet) sind in Tabelle 8 dargestellt. Die Homogenität der Varianzen war nicht verletzt, jedoch teilweise die Normalität. Entsprechend der

Verletzung wurde ein Mann-Whitney-*U*-Test durchgeführt. Abbildung 10 stellt die Barplots der UTAUT2-Subdimensionen dar.

Dimension	Tests	<i>t/U</i>	<i>df</i>	<i>p</i>	Co'd/BRK	Power
UTAUT2_PE	Student's <i>t</i>	-0,474	16	0,679	-0,225	0,12
UTAUT2_EE	Student's <i>t</i>	0,237	16	0,408	0,112	0,08
UTAUT2_SI	Student's <i>t</i>	-0,308	16	0,619	-0,146	0,09
UTAUT2_FC	Student's <i>t</i>	-0,745	16	0,767	-0,354	0,18
UTAUT2_HM	Mann-Whitney <i>U</i>	33,000		0,266	0,355/0,175*	0,17
UTAUT2_HT	Student's <i>t</i>	0,795	16	0,219	0,377	0,19
UTAUT2_BI	Student's <i>t</i>	0,132	16	0,448	0,063	0,05

Tabelle 8. Hypothesentest UTAUT2 versus VQ – S1

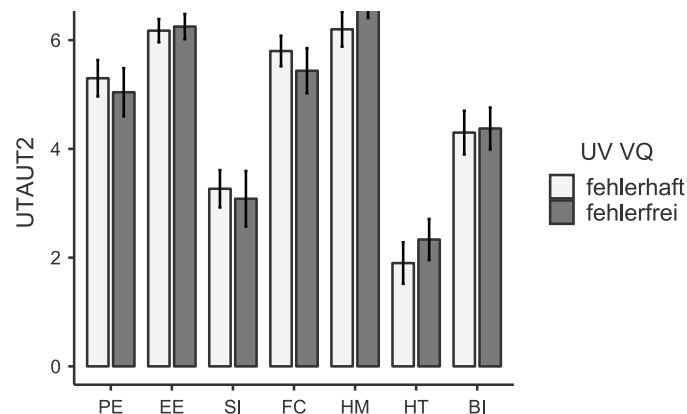


Abbildung 10. Barplots UTAUT2 versus VQ – S1

NASA-TLX: Die durchschnittliche Beanspruchung zeigt keinen signifikanten Unterschied zwischen den Faktorstufen der VQ fh und ff. Alle anderen Subdimensionen des NASA-TLX mit Ausnahme der F zeigten ebenfalls keine signifikanten Unterschiede (siehe Tabelle 9). Abbildung 11 zeigt die Barplots der Subdimensionen. Die Voraussetzung der Varianzhomogenität wurde nicht verletzt, jedoch teilweise die der Normalität. Entsprechend den Verletzungen wurde ein Mann-Whitney-*U*-Test durchgeführt.

Dimension	Tests	<i>t/U</i>	<i>df</i>	<i>p</i>	Co'd/BRK	Power
Mean_NASA	Student's <i>t</i>	0,397	16	0,348	0,189	0,10
NASA_MD	Student's <i>t</i>	-0,283	16	0,610	-0,134	0,08
NASA_PD	Mann-Whitney <i>U</i>	23,500		0,073	0,900/0,412	0,55
NASA_TD	Mann-Whitney <i>U</i>	-25,000		0,926	-0,910/0,375	0,57
NASA_P	Student's <i>t</i>	1,530	16	0,073	0,726	0,43
NASA_E	Mann-Whitney <i>U</i>	-22,500		0,946	-0,974/0,438	0,60
NASA_F	Mann-Whitney <i>U</i>	20,500		0,045	1,118/0,488	0,71

Tabelle 9. Hypothesentest NASA-TLX versus VQ – S1

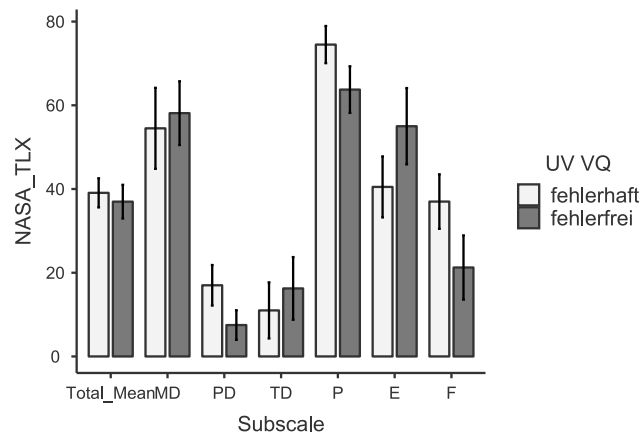


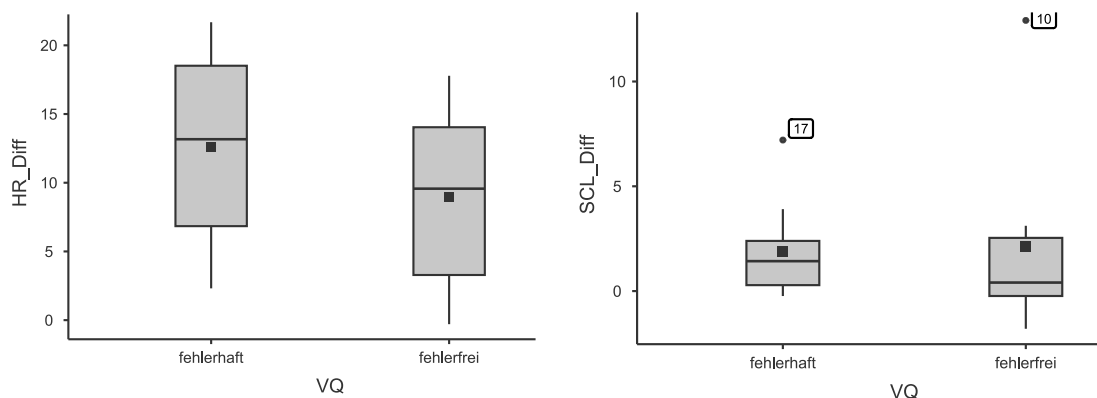
Abbildung 11. Barplots der NASA-TLX versus VQ – S1

Physiologische Parameter: Der statistisch signifikante Einfluss der VQ auf die physiologischen Reaktionen konnte via *t*-Test nicht für die AVs SCL oder HR bestätigt werden. Die Voraussetzungen für die Varianzhomogenität wurden sowohl von der HR als auch von der SCL erfüllt, die Voraussetzung für die Normalverteilung nur bei der HR. Die Abweichung von der Normalverteilung wurde durch den nichtparametrischen Mann-Whitney-*U*-Test kompensiert. Tabelle 10 zeigt die Ergebnisse der *t*-Tests.

Parameter	Tests	<i>t/U</i>	<i>df</i>	<i>p</i>	Co'd/BRK	Power
HR_Diff	Student's <i>t</i>	1,081	16	0,148	0,513	0,27
SCL_Diff	Mann-Whitney <i>U</i>	31,000		0,230	0,461/0,225*	0,23

Tabelle 10. HR [BPM] und SCL [μS] Test versus VQ – S1

Abbildung 12 präsentiert die entsprechenden Boxplots.



(a) HR_Diff [BPM]

(b) SCL_Diff [μS]

Abbildung 12. Boxplots (a) HR_Diff & (b) SCL_Diff versus VQ – S1

Die relevanten Datensätze der Erhebung werden in Anhang A3, Tabelle 63 bis Tabelle 66, bereitgestellt.

4.1.3 Interpretation

Die Ergebnisse der Studie lassen folgende Rückschlüsse hinsichtlich der Hypothesen zu:

$H_{B.1.1}$ Fehlerhafte Sichtbarkeit reduziert bei IAS die Leistung.

Nachweis. Die $H_{B.1.1}$ wird durch die Ergebnisse bestätigt; somit kann die $H_{B.1.0}$ verworfen werden. Durch die UV VQ konnte ein statistisch signifikanter Unterschied sowohl für die Richtigkeit $t(16) = 4,06$; $p < 0,001$; $MD = 0,275$; $Co'd = 1,927$; $Power = 0,99$ als auch für den Index Leistung ($U = 20$; $p = 0,042$; $MD = 0,046$ %/s; $Co'd = 1,15$; $Power = 0,73$) aufgezeigt werden. Aufgrund der fh-Sichtbarkeit verschlechterte sich im Mittel die Richtigkeit der fachlichen Aufgabenausführung um 27,5 %, sodass sich die Gesamtleistung im Mittel um 0,046 % pro Sekunde verschlechterte (vgl. Abbildung 8 und Abbildung 9). Für die DLZ konnte via t -Test $t(16) = -0,295$; $p = 0,614$; $MD = -15,8$; $Co'd = -0,140$; $Power = 0,09$ kein statistisch signifikanter Unterschied nachgewiesen werden (vgl. Abbildung 7). Aufgrund der geringen statistischen Power lassen sich diese Aussagen für den Index Leistung sowie die AV DLZ nur begrenzt verallgemeinern.

$H_{B.2.1}$ Fehlerhafte Sichtbarkeit reduziert bei IAS die Nutzerakzeptanz.

Nachweis. Die $H_{B.2.1}$ wird durch die Ergebnisse nicht bestätigt; somit wird die $H_{B.2.0}$ angenommen, dass eine fehlerhafte Sichtbarkeit keinen Einfluss auf die Akzeptanz bei IAS hat. Keine der Subdimensionen des UTAUT2-Fragebogens zeigte einen statistisch signifikanten Unterschied zwischen den beiden Faktorstufen der VQ (vgl. Tabelle 8). Zum einen ist die statistische Power und Effektstärke sehr gering, u. a. BI ($Co'd = 0,063$; $Power = 0,05$), zum anderen weisen die Mittelwerte nahezu keine Unterschiede und geringe Standardabweichungen innerhalb der Stichprobe auf. Hierdurch kann diese Aussage bestätigt werden (vgl. Tabelle 5).

$H_{B.3.1}$ Fehlerhafte Sichtbarkeit erhöht bei IAS die Beanspruchung.

Nachweis. Die $H_{B.3.1}$ wird durch die Ergebnisse nicht bestätigt; somit wird die $H_{B.3.0}$ angenommen, dass eine fh-Sichtbarkeit keinen Einfluss auf die Beanspruchung bei IAS hat. Die durchschnittliche Beanspruchung zeigt keinen signifikanten Unterschied hinsichtlich der Faktorstufen VQ. Dies wird ebenfalls über alle Subdimensionen des NASA-

TLX-Fragebogens bis auf die F ($U = 20,5$; $p < 0,045$; $Co'd = 1,118$; $Power = 0,71$) bestätigt (vgl. Tabelle 9, Tabelle 6 und Abbildung 11). Ebenfalls wurde für die Ergebnisse der physiologischen Parameter SCL und HR keine signifikanten Unterschiede hinsichtlich der Beanspruchung nachgewiesen. Die vorliegende Stichprobe lässt den Schluss zu, dass eine schlechte Sichtbarkeit keine Stressreaktionen weder bei MD noch physiologisch auslöst.

4.1.4 Diskussion

Die Ergebnisse der Studie zur Untersuchung von Sichtbarkeitsfehlern zeichnen ein differenziertes Bild im Rahmen der Stichprobe. Zwar lässt sich durch die Analyse bestätigen, dass die UV VQ einen signifikanten Einfluss auf die Leistungsfaktoren Richtigkeit und den Index Leistung hat, jedoch nicht für die DLZ. Hier konnte kein signifikanter Unterschied zwischen den Bedingungen ff und fh nachgewiesen werden (ff: $M = 476$; $SD = 130,41$; fh: $M = 460$; $SD = 97,21$; $t(16) = -0,295$, $p = 0,614$). Die geringe Mittelwertsdifferenz ($MD = -15,8$) und die niedrige (negative) Effektstärke ($Co'd = -0,140$) deuten darauf hin, dass die VQ keinen wesentlichen Einfluss auf die DLZ hatte. Im Gegensatz dazu war die Richtigkeit signifikant höher bei ff-Sichtbarkeit (ff: $M = 0,80$; $SD = 0,152$) im Vergleich zu fh-Sichtbarkeit (fh: $M = 0,529$, $SD = 0,136$; $t(16) = 4,06$; $p < 0,001$). Die Mittelwertsdifferenz ($MD = 0,275$) und die hohe Effektstärke ($Co'd = 1,927$) verdeutlichen die Bedeutung einer ff-Darstellung für die Richtigkeit der fachlichen Aufgabe. Zusammengefasst wird durch den Index Leistung bestätigt, dass ein signifikanter Unterschied zugunsten der ff-Bedingung (ff: $M = 0,19$ %/s, $SD = 0,098$; fh: $M = 0,12$ %/s, $SD = 0,049$; $p = 0,042$; $Co'd = 1,15$; $Power = 0,73$) existiert.

Auch wenn die UV VQ einen Einfluss auf die Leistung hat, zeigte die Studie inferenzstatistisch auf, dass eine fehlerhafte Sichtbarkeit keinen signifikanten Einfluss auf die Akzeptanzfaktoren (UTAUT2) sowie auf die Beanspruchungsfaktoren NASA-TLX oder die physiologischen Parameter zu haben scheint, außer auf die F ($U = 20,5$; $p < 0,045$; $Co'd = 1,118$; $Power = 0,71$). Die Werte für PE (fh: $M = 5,3$ vs. ff: $M = 5,04$) und EE (ff: $M = 6,25$ vs. fh: $M = 6,17$) lagen sehr nah beieinander. Auch FC (fh: $M = 5,8$ vs. ff: $M = 5,44$) und HM (ff: $M = 6,58$ vs. fh: $M = 6,2$) wiesen keine relevanten Unterschiede auf. Die NASA-TLX-Werte zeigten uneinheitliche Ergebnisse. Während MD ($M = 55$ vs. $M = 40,5$), PD

($M = 58,13$ vs. $M = 54,5$) und P ($M = 16,25$ vs. $M = 11$) tendenziell höhere Werte bei ff-Sichtbarkeit aufwiesen, waren TD ($M = 17$ vs. $M = 7,5$) und E ($M = 74,5$ vs. $M = 63,75$) bei fh-Sichtbarkeit (fh) höher. Besonders auffällig war die signifikant erhöhte F bei fh ($M = 37$ vs. $M = 21,25$; $p = 0,045$), was auf eine gesteigerte subjektive Belastung bei fh-Sichtbarkeit hinweist. Die physiologischen Parameter wie die HR_Diff und die SCL_Diff wiesen keine signifikanten Unterschiede auf. Dennoch war die HR_Diff bei fh tendenziell höher ($M = 12,61$ BPM, $SD = 13,17$) als bei ff ($M = 8,98$ BPM, $SD = 9,57$), während die SCL_Diff bei ff leicht höhere Mittelwerte zeigte ($M = 2,1$ μ S, $SD = 0,41$) als bei fh ($M = 1,88$ μ S, $SD = 1,43$). Mit Ausnahme der Frustrationsdimension weisen die übrigen NASA-TLX-Dimensionen lediglich geringe Effektstärken ($Co'd < 0,5$) und niedrige Teststärken (Power $< 0,3$) auf, sodass die VQ insgesamt nur einen geringen Einfluss auf die subjektive Beanspruchung erkennen lässt. Um diesen Einfluss valide nachweisen zu können, bedarf es größerer Stichprobenumfänge. Die a-priori ermittelten Effektstärken aus dem Vorversuch und die geschätzten Stichprobenumfänge konnten nicht erreicht werden. Unter der Annahme, dass keiner der Probanden über Vorkenntnisse zu den Inhalten der Museumsanwendung verfügte, kann von vergleichbaren Ausgangsbedingungen ausgegangen werden. Die vermittelten Lerninhalte waren für die Teilnehmenden inhaltlich weitgehend irrelevant, abgesehen von individuellen thematischen Präferenzen. Aufgrund dieser neutralen Ausgangslage sowie der strukturellen Eigenschaften der Lernanwendung erscheint eine Übertragbarkeit der Ergebnisse auf vergleichbare Szenarien im industriellen Kontext plausibel.

Im Vergleich zur vorliegenden Studie verdeutlichen die Ergebnisse von Arning et al. (2012) eine stärker ausgeprägte subjektive und akzeptanzbezogene Wirkung von Sichtbarkeitsproblemen. Zugleich zeigten sich signifikante Unterschiede in der Leistung ($M_{\text{Bildschirm}} = 0,91$; $SD = 0,09$ vs. $M_{\text{Projektor}} = 0,85$; $SD = 0,10$; $F(1,23) = 12,5$; $p < 0,01$). Während in der aktuellen Untersuchung keine signifikanten Effekte auf die Akzeptanzfaktoren oder die subjektiv wahrgenommene Beanspruchung festgestellt wurden, wies die Projektorbedingung bei Arning et al. (2012) signifikant geringere Werte für das Vertrauen in das System auf ($M_{\text{Bildschirm}} = 86,1$; $SD = 11,7$ vs. $M_{\text{Projektor}} = 77,5$; $SD = 14,3$; $F(1,23) = 9,1$; $p < 0,05$) sowie für die Akzeptanz ($M_{\text{Bildschirm}} = 92,7$; $SD = 9,0$ vs. $M_{\text{Projektor}} = 87,5$; $SD = 11,8$; $F(1,23) = 4,9$, $p < 0,05$). Dabei beeinflusste die Sichtbarkeit zunächst das Vertrauen, das

wiederum über die wahrgenommene Desorientierung indirekt auf die Akzeptanz wirkte. Auch die subjektiv wahrgenommene Gesamtbeanspruchung fiel im Projektor-kontext mit $M = 13,1$ ($SD = 9,4$) deskriptiv höher aus als bei der Bildschirmdarstellung mit $M = 10,0$ ($SD = 9,8$), wobei dieser Unterschied nicht statistisch signifikant war. Dieser Befund stimmt insofern mit den Ergebnissen der vorliegenden Studie überein, als auch dort kein signifikanter Einfluss der Sichtbarkeitsqualität auf die subjektiv wahrgenommene Beanspruchung nachgewiesen wurde. Insgesamt legen die Ergebnisse von Arning et al. (2012) nahe, dass Sichtbarkeitsprobleme, die sich aus unterschiedlichen Visualisierungsmodalitäten ergeben, beispielsweise aus der Verwendung eines Projektors anstelle eines Bildschirms, tendenziell stärkere Auswirkungen auf die Akzeptanz entfalten können als graduelle Sichtbarkeitsunterschiede innerhalb einer einheitlichen Darstellungsform. Diese Schlussfolgerung ist jedoch vorsichtig zu interpretieren, da sich die beiden Visualisierungsmodalitäten zugleich hinsichtlich ihrer Bedienung, Vertrautheit und technischen Neuartigkeit unterscheiden.

Die Befunde von Sadovitch (2020) stimmen daher nur teilweise mit den Ergebnissen der vorliegenden Untersuchung überein. Während sich in der vorliegenden Studie keine signifikanten Unterschiede in den Akzeptanzdimensionen PE, EE, FC, HM und BI nachweisen ließen, verdeutlichen die Ergebnisse von Sadovitch, dass visuelle Fehler durchaus zu Akzeptanzverlusten führen können. Entscheidend scheinen dabei jedoch die konkrete Fehlercharakteristik, deren Ausprägung sowie der jeweilige Anwendungskontext zu sein. Eine mögliche Erklärung für die abweichenden Ergebnisse besteht darin, dass der in der vorliegenden Studie untersuchte Sichtbarkeitsfehler subjektiv weniger salient oder weniger störend wahrgenommen wurde als die bei Sadovitch (2020) untersuchten Registrierungsfehler. Darüber hinaus ist der Vergleich nur eingeschränkt möglich, da Sadovitch die Systemakzeptanz anhand der Dimensionen Nützlichkeit und Zufriedenheit erfasste, während in der vorliegenden Untersuchung Konstrukte des UTAUT2-Modells betrachtet wurden. Die fehlenden signifikanten Effekte in PE, EE, FC, HM und BI können daher nicht unmittelbar als Widerspruch zu Sadovitch (2020) interpretiert werden. Vielmehr weisen beide Untersuchungen darauf hin, dass objektiv oder technisch vorhandene visuelle Fehler nicht zwangsläufig in allen subjektiven Bewertungsdimensionen gleichermaßen abgebildet werden. Für zukünftige Untersuchungen erscheint es

deshalb sinnvoll, neben der bloßen Präsenz eines visuellen Fehlers auch dessen Salienz, Dynamik, räumliche Ausprägung und subjektive Wahrnehmbarkeit systematisch zu berücksichtigen.

Bei Alessa et al. (2023) zeigte die Instruktionsform einen signifikanten Einfluss auf den NASA-TLX-Gesamtwert; $F(1,26) = 9,34$; $p = 0,005$; $\eta^2p = 0,26$, sowie insbesondere auf die F $F(1,26) = 70,08$; $p < 0,001$; $\eta^2p = 0,73$, und die subjektiv bewertete Leistung, $F(1,26) = 25,46$; $p < 0,001$; $\eta^2p = 0,50$. Bei der komplexen Aufgabe reduzierte die AR-basierte Instruktion gegenüber der papierbasierten Anleitung den PD und die F. Für MD zeigten sich hingegen weder bei der einfachen noch bei der komplexen Aufgabe signifikante Unterschiede zwischen AR- und papierbasierter Instruktion. Bei der einfachen Aufgabe ergaben sich auch für die PD und die physiologischen Stressparameter keine signifikanten Unterschiede zwischen den Instruktionsformen. Teilweise fiel die AR-basierte Instruktion sogar ungünstiger aus: Gegenüber der papierbasierten Anleitung erhöhte sich TD und P verschlechterte sich, während F geringer ausfiel. Die beanspruchungsreduzierenden Vorteile der AR-Instruktion zeigten sich somit vor allem bei der komplexen Wartungsaufgabe. Die Divergenz der Ergebnisse könnte daher durch unterschiedliche Komplexitätsgrade der Aufgaben erklärt werden. In der Studie von Alessa et al. (2023) wurden komplexe, körperlich anspruchsvolle Wartungstätigkeiten untersucht, während die Aufgaben der vorliegenden Museumsanwendung als kognitiv und körperlich weniger herausfordernd einzustufen sind. Darüber hinaus ist die Vergleichbarkeit der Studien grundsätzlich eingeschränkt, da Alessa et al. (2023) AR-basierte mit papierbasierten Instruktionen verglichen, während in der vorliegenden Untersuchung Sichtbarkeitsunterschiede innerhalb einer VR-Anwendung betrachtet wurden. AR-, VR- und papierbasierte Informationsdarstellungen unterscheiden sich nicht nur hinsichtlich ihrer Sichtbarkeit, sondern auch bezüglich Immersionsgrad, Informationsaufbereitung, Interaktionsform, räumlicher Einbettung sowie körperlicher und kognitiver Anforderungen. Daher kann nicht eindeutig beurteilt werden, ob die Sichtbarkeit bei Alessa et al. (2023) den maßgeblichen Einflussfaktor darstellte oder ob die beobachteten Unterschiede vielmehr auf die Instruktionsform, die Aufgabenkomplexität oder das Zusammenwirken mehrerer Kontextfaktoren zurückzuführen sind.

Auch Merenda et al. (2019) stützen diese Interpretation: Ihre Ergebnisse zeigen, dass ein erhöhtes Maß an visueller Detailfülle (visuelle Fidelity) zwar die objektive Beanspruchung und das Blickverhalten beeinflusst, jedoch nicht zwingend zu einer veränderten subjektiven Belastung führt – ein Effekt, der auch in der aktuellen Studie sichtbar wird. Zusammenfassend unterstreicht die vorliegende Untersuchung somit die Bedeutung differenzierter Betrachtungen bei der Bewertung visueller Fehler: Sichtbarkeitsprobleme scheinen in erster Linie die objektive Leistung zu beeinträchtigen, während Einflüsse auf Akzeptanz und Beanspruchung nur unter bestimmten Bedingungen – etwa bei hoher Fehlerausprägung, komplexen Aufgaben oder affektiv relevanten Kontexten – messbar werden.

4.2 Studie 2 – Interaktionsfehler I

(Teile dieses Kapitels basieren auf der Veröffentlichung: Rosilius, M., Spiertz, M., Wirsing, B., Geuen, M., Bräutigam, V. & Ludwig, B. (2024). Impact of Industrial Noise on Speech Interaction Performance and User Acceptance when Using the MS HoloLens 2. *Multimodal Technologies and Interaction*, 8(2), 8. <https://doi.org/10.3390/mti8020008>. Die Inhalte, Abbildungen und Tabellen wurden für die vorliegende Monographie übernommen bzw. teilweise überarbeitet und in adaptierter Form eingebracht. Der Originalartikel wurde als Open-Access-Beitrag unter der Lizenz Creative Commons Attribution 4.0 International (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0/>) veröffentlicht, die die Bearbeitung und Wiederverwendung unter Angabe der ursprünglichen Quelle gestattet.)

Als fachliche Aufgabe dieser Studie wurden die Probanden gebeten, bei unterschiedlichen Umgebungslautstärken mit dem IAS via Sprache zu interagieren (vgl. Abbildung 13). Zum einen sollten sie einen vorgegebenen Text diktieren, zum anderen Sprachbefehle ausführen. Hierbei wurden die Probanden angehalten, Informationen aus der realen Welt zu erfassen (Lesen des Textes) und diese dem IAS via ASR zu diktieren. Die Ergebnisse der Spracheingabe konnten die Probanden als augmentierten Text live mitverfolgen. Beim Sprachbefehl wurden die Probanden über einen virtuellen Dialog aufgefordert, angezeigte Reglerzustände entsprechend den Instruktionen anzupassen.



Abbildung 13. Schematische Darstellung des Versuchsaufbaus – S2

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

4.2.1 Methodik

Design of Experiment

Als Forschungsdesign wurde ein gegenbalanciertes Within-Subject-Design angewendet. Die Wahl dieses effizienten Designs war notwendig, da der Versuchsaufbau und -ablauf aufwendig und komplex war. Die UV IQ wird latent durch die Manipulation des SPL in vier Faktorstufen (0 dB, 64,2 dB, 84,9 dB und 87 dB) abgebildet. Für die beiden hohen SPL wurde der Fehlerfall definiert. Über eine gegenbalancierte Versuchsreihenfolge (RF) wurden die SPL-Faktorstufen bzw. die Probanden gruppiert. Die eine Gruppe (EER) startete mit den niedrigen beiden SPLs (ff) gefolgt von den hohen SPLs (fh). Bei der anderen Gruppe (SER) war die RF umgekehrt. Durch diese Sequenzierung bzw. Gruppierung kann der sog. „Carry-Over-Effect“ genutzt werden, der innerhalb des Within-Subject-Designs ermöglicht, Unterschiede zwischen der fh-Faktorstufe und der vermeintlichen ff-Faktorstufe zu messen (Brooks, 2012). Die Metriken WER (siehe Gleichung 2) und ITR (siehe Gleichungen 3–5) wurden als AV zur Messung der Sprachinteraktionsleistung herangezogen. Als weitere primäre AV wurde der UTAUT2 erhoben. Zur sekundären Untersuchung der Beanspruchung wurden der NASA-TLX sowie die physiologischen Parameter SCL und HR gemessen.

Definition der SPL: Für die Definition des SPL wurden Messungen (zerspanende Industrie) in einer realen industriellen Produktion durchgeführt. Ebenfalls wurden die für

die Definition der SPL-Faktorstufen gesetzlichen Anforderungen berücksichtigt (Richtlinie 2003/10/EG des Europäischen Parlaments und des Rates, 2003). Der untere Wert von 64,2 dB(A) wurde als durchschnittlicher Mindestwert in der Werkstatt gemessen. 84,9 dB(A) wurde vom oberen Expositionsgrenzwert über acht Stunden pro Tag entsprechend den gesetzlichen Anforderungen an die Sicherheit am Arbeitsplatz abgeleitet. Der SPL-Wert von 87 dB stellt den maximal zulässigen Expositionsgrenzwert dar. Bei der unteren SPL-Faktorstufe 0 dB(A) wird kein Lärm von den Lautsprechern abgestrahlt. Die Studien wurden unter Tragen professioneller Lärmschutzausrüstung bei allen Teilnehmern und Versuchsleitern durchgeführt.

Umgebungsgeräusch: Nachdem aus einer Vorstudie des Forschungsverbunds die These widerlegt werden konnte, dass die Stationarität einen statistisch signifikanten Einfluss auf Spracheingabeleistung hat, wurde im Rahmen dieser Studie als Industriege­räusch eine Tonaufnahme einer Wasserstrahlschneidmaschine verwendet. Grundlage der Vorüberlegungen im Forschungsverbund war, dass ein stationäres Hintergrundgeräusch sowohl für Menschen als auch für Algorithmen leichter auszublen­den sei als ein nicht-stationäres Hintergrundgeräusch (Sainburg & Gentner, 2021). In Abbildung 14 werden die Amplituden und das Spektrogramm des Signals der Wasserstrahlschneiden dargestellt, welches über die Zeit ein nahezu konstantes Spektrum (Stationarität) aufweist.

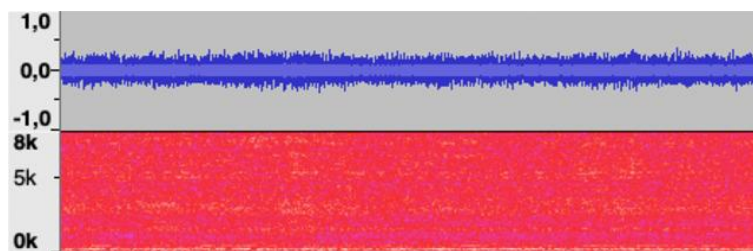


Abbildung 14. Amplitude (oben) und Spektrogramm (unten) – S2

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

Diktierter Text: Die Probanden wurden gebeten, den folgenden Text bestehend aus sieben inhomogenen Satzfragmenten mit insgesamt 39 Wörtern dem IAS zu diktieren:

„Platziere den Gabelstapler im Raum
Fenster im Raum anheften
Ich denke also bin ich
Ich glaube also bin ich

Arbeit besiegt alles unablässiges Mühen bezwingt alles
bringt alles fertig

Gehe zum Ende des Paragraphen

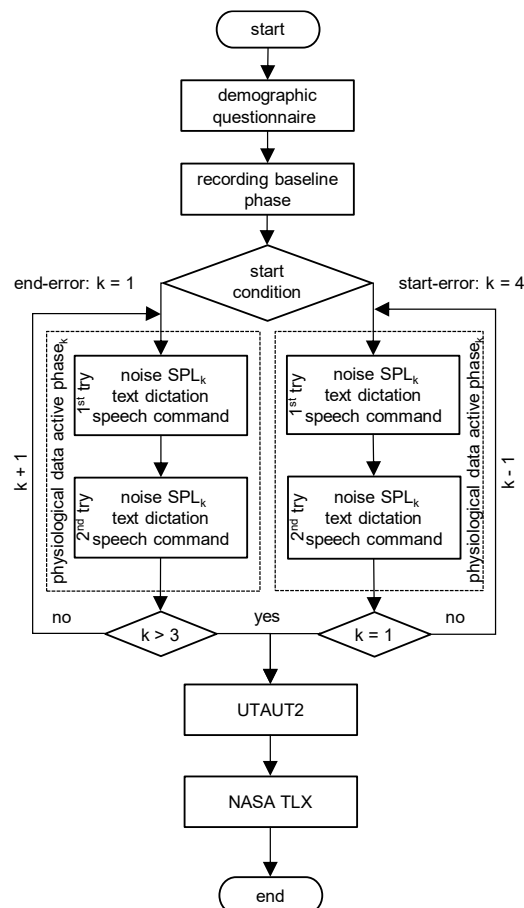
Gehe zum Anfang des Paragraphen`

Auswahl der Umgebungsgeräusche"

Versuchsaufbau und -ablaufplan

Das Experiment wurde in einem Tonaufnahmestudio mit den Abmessungen 10,5 m × 7,3 m durchgeführt. Die Betonwände sowie der Boden waren vollständig mit einer grünen Leinwand verkleidet. Abbildung 13 veranschaulicht den schematischen Versuchsaufbau. Zur Erzeugung der akustischen Reizkulisse kamen Yamaha-MSP5-Aktivlautsprecher zum Einsatz, die in einem Abstand von 0,5 m zu den Wänden positioniert und jeweils in einem Winkel von 45° in Richtung der Raumecken ausgerichtet wurden. Ziel dieser Anordnung war die Erzeugung eines diffusen Schallfelds, wie es typischerweise in industriellen Umgebungen auftritt, in denen Schall durch Reflexionen an zahlreichen Oberflächen eine gleichmäßige Verteilung hinsichtlich Intensität und zeitlicher Verteilung erfährt (Weinzierl, 2008). Die Steuerung der akustischen Stimuli erfolgte über ein externes Mischpult, das von einem der Versuchsleiter bedient wurde. Der andere Versuchsleiter übernahm die Einweisung und Betreuung der Teilnehmenden. Der zu lesende Text wurde auf einem iPad präsentiert, das in etwa 20 cm Entfernung zum Probanden auf einem Ständer befestigt war. Seitlich daneben wurde ein immersives Browserfenster eingeblendet, das die Google Notes Website zeigte und die ASR der HL2 in Echtzeit dokumentierte. Der Versuchsablauf orientierte sich an dem in Abbildung 15 dargestellten experimentellen Ablaufplan. Zu Beginn füllten die Teilnehmenden einen soziodemografischen Fragebogen aus. Es folgte eine fünfminütige Ruhephase (Baseline) zur Erfassung physiologischer Ausgangswerte, während der eine Empatica-E4-Armbanduhr getragen wurde. Anschließend wurden die Probanden mit der HL2 sowie Gehörschutz ausgestattet. Die Zuweisung der Startreihenfolgen erfolgte randomisiert. Während der Aktionsphase, also der eigentlichen Durchführung des Experiments, wurden die physiologischen Parameter kontinuierlich aufgezeichnet. Für jede Faktorstufe des SPL wurde ein standardisiertes Vorgehen angewandt: Abhängig von der jeweiligen Versuchskonfiguration wurde die SPL über die externe Soundkarte Scarlett 2i2 in

Verbindung mit dem Steuergerät Digital Sound 8928 gemäß Versuchsplan kalibriert. Im Anschluss wurde der Proband dem entsprechenden SPL ausgesetzt und las dabei den vorbereiteten Text laut vor. Nach Abschluss dieser Phase wurde der Programmablauf zur Untersuchung des SPC aktiviert. Die Probanden lösten durch das Drücken eines virtuellen türkisfarbenen Schalters (siehe Abbildung 13) die automatische Zeitmessung aus.

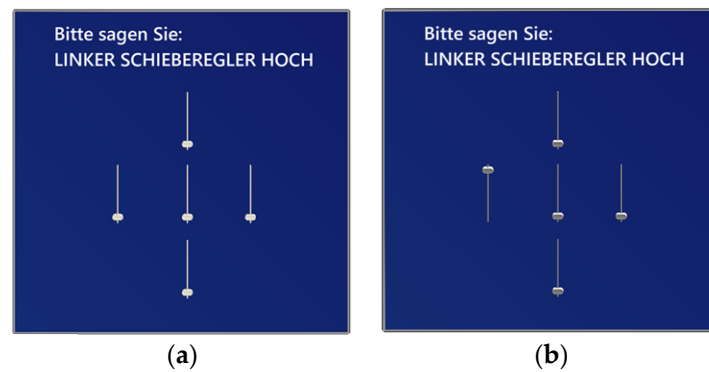


Fußnote. Variable k: 1—0 dB(A); 2—64,2 dB(A); 3—84,9 dB(A) and 4—87 dB(A); wobei jede SPL ≠ 0 dB (A) zweimal nacheinander durchlaufen wurde

Abbildung 15. Versuchsablaufplan – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Die Probanden wurden aufgefordert, den Befehl, der für jeden Schieberegler instruiert wurde, auszusprechen (vgl. Abbildung 16) bspw. „Linker Schieberegler hoch“. Sobald alle fünf entsprechenden Befehle ausgesprochen waren, wurde die Zeitmessung bei Erfolg automatisch beendet.



Fußnote. (a) vor (linker Schieberegler ist unten) und (b) nach dem ersten erfolgreichen SPC (linker Schieberegler ist oben)

Abbildung 16. SPC-Dialog des Forschungsprototyps – S2

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

Die Korrektheit der Erkennung des Systems wurde durch den Studienleiter dokumentiert. Wenn der Teilnehmer das Kommando korrekt artikulierte, der SPC aber nicht erkannte wurde, wurde der Befehl durch den Studienleiter (via Bluetooth-Klicker) ferngesteuert aktiviert und dokumentiert. Dann wurde das Prozedere mit dem nächsten SPL wiederholt. Nach den ersten beiden SPL wurde die Aufzeichnung der physiologischen Parameter gestoppt und mit den letzten beiden SPL neu gestartet. Am Ende des Experiments wurde der Teilnehmer gebeten, den UTAUT2- und den NASA-TLX-Fragebogen auszufüllen.

Stichprobenumfangsplanung und Versuchsdurchführung

Statistische Parameter	Werte
η^2p	0,157
Power	0,80
α	0,05
Anzahl der Gruppen	4
Anzahl Messwiederholungen	2
berechneter Stichprobenumfang	16

Tabelle 11. Stichprobenumfangsplanung – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Um a priori die angemessene Stichprobengröße zu schätzen und das Konzept des Forschungsdesigns zu erproben, wurde eine Vorstudie mit fünf Teilnehmern durchgeführt, aus der die ersten Daten generiert und berechnet wurden (siehe Tabelle 11). Die Hauptstudie wurde in KW 34 2022 durchgeführt. Alle Datensätze wurden anonymisiert erfasst. Die Teilnehmer nahmen freiwillig an der Studie teil und erhielten keine

finanzielle Entschädigung. Jeder Teilnehmer wurde über die Richtlinien für ethische Forschung informiert und musste seine ausdrückliche Zustimmung zur Teilnahme geben. Die Versuchsdauer betrug ca. 50 Minuten. Die Geräte wurden nach jedem Teilnehmer vollständig desinfiziert. Die RF hinsichtlich EER und SER wurde im Vorfeld zufällig festgelegt.

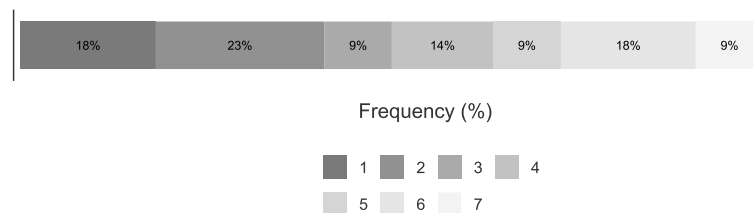
Statistisches Vorgehen

Ergänzend zu den Ausführungen in Kapitel 3.2 werden in diesem Kapitel studien-spezifische Ergänzungen angeführt. Um den Einfluss der unabhängigen Variablen SPL auf die abhängigen Variablen WER und ITR zu analysieren, wurde eine ANOVA mit wiederholten Messungen und Hypothesentests durchgeführt (E. Park et al., 2009). Um die Auswirkungen von undefinierten, sprachlichen, technischen oder organisatorischen Problemen herauszufiltern, wurden sowohl die Antwortzeit (Obergrenze = 27,9 s) als Faktor der ITR als auch der WER (Obergrenze = 77,85%) „winsorisiert“. Aufgrund der logischen Kausalität war 0 die natürliche Untergrenze für beide Variablen. Zur Überprüfung der Annahmen wurde auf Sphärizität getestet. Im Falle signifikanter Unterschiede zwischen den SPL wurde ein Post-Hoc-Vergleich mit Scheffe- α -Fehlerkorrektur durchgeführt. Da die Faktorstufe SPL 0 dB(A) im Vergleich zu den übrigen Faktorstufen nur die Hälfte der Datenpunkte enthielt, wurde aufgrund unterschiedlicher Stichprobenumfänge keine ANOVA durchgeführt. Hierfür wurde ein *t*-Test für gepaarte Stichproben (einseitige Hypothese) für SPL 0 dB(A) gegenüber den anderen SPLs herangezogen (Institut für Statistik, 2020). Zur Überprüfung der Annahmen wurde zunächst die Normalverteilung getestet. Wurde diese verletzt, musste ein nichtparametrischer Wilcoxon-Rang durchgeführt werden. Die Auswirkungen des Interaktionsfehlers wurden sowohl für den NASA-TLX als auch für den UTAUT2 als unabhängiger *t*-Test mit einseitiger Hypothese bewertet. Zur Überprüfung der Annahmen wurden im Vorfeld Normalität und Homogenität der Varianzen getestet. Wenn die Homogenität der Varianzen verletzt wurde, wurde ein Welch-*t*-Test angewendet. Wurde hingegen die Normalverteilung verletzt, wurde ein Mann-Whitney-*U*-Test durchgeführt. Aus den gemessenen SCL wurden Mittelwerte für die RF (SER versus EER) berechnet.

4.2.2 Ergebnisse

Demografische Daten

Für die Hauptstudie wurden insgesamt $N = 22$ Teilnehmer über eine regionale Studentenmailingliste rekrutiert. Folgende demografische Daten charakterisierten die Teilnehmerkohorte: 55 % männlich und 45 % weiblich; 9 % der Teilnehmer hatten eine abgeschlossene Berufsausbildung, 55 % hatten die Hochschulreife und 36 % einen Hochschulabschluss. Die Teilnehmer waren zwischen 20 und 41 Jahre alt, ihr Durchschnittsalter lag bei 25,7 Jahren. Abbildung 17 zeigt die Ergebnisse des Erfahrungsniveaus im Umgang mit Sprachinteraktion in Bezug auf die Selbsteinschätzung der Teilnehmer.



Fußnote. Stufe 1 (niedrig) bis Stufe 7 (hoch)

Abbildung 17. Verteilung der Sprachinteraktionserfahrung – S2

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

Inferenzstatistische Ergebnisse

WER: Hinsichtlich des WER wurde die statistische Signifikanz der unabhängigen Variable SPL $F(2;40) = 14,81$; $p < 0,001$; $\eta^2p = 0,425$; Power = 1,00 mittels ANOVA mit wiederholten Messungen nachgewiesen. Die Sphärizität wurde nicht verletzt. Tabelle 12 zeigt den Post-Hoc-Vergleich.

SPL [dB]		MD	SE	df	t	p_{scheffe}	Co'd	Power
64,2	84,9	12,57	3,97	20	3,173	0,017	1,352	0,99
64,2	87,0	19,74	4,03	20	4,900	0,001	0,712	0,61
84,9	87,0	7,18	2,92	20	2,463	0,071	0,254	0,12

Tabelle 12. WER [%] – PHT versus SPL – S2.

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Tabelle 13 zeigt die Ergebnisse der WER mit einem t -Test mit gepaarten Stichproben für die SPL-Faktorstufe 0 dB(A) im Vergleich zu den anderen. Da die Normalität verletzt wurde, wurde der nichtparametrische Wilcoxon-Rang für gepaarte Stichproben durchgeführt.

SPL [dB]	<i>df</i>	<i>t</i>	<i>p</i>	<i>MD</i>	<i>SE</i>	<i>Co'd</i>	Power
0 64,2	13	3,383	0,005	8,06	2,38	0,715	0,95
0 84,9	13	5,093	0,001	20,61	4,05	0,884	0,99
0 87	13	6,572	0,001	28,85	4,39	0,990	1,00

Tabelle 13. nicht-param. Test WER [%] versus SPL – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Tabelle 14 zeigt die zugehörigen deskriptiven Daten.

SPL [dB]	<i>n</i>	<i>M</i>	<i>Median</i>	<i>SD</i>	<i>SE</i>
0	22	3,03	1,28	3,93	0,84
64,2	44	15,09	9,62	16,20	3,45
84,9	44	27,97	22,44	23,05	4,91
87	44	36,77	29,49	28,28	6,03

Tabelle 14. DD der WER [%] versus SPL – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Abbildung 18 zeigt die entsprechenden Boxplots.

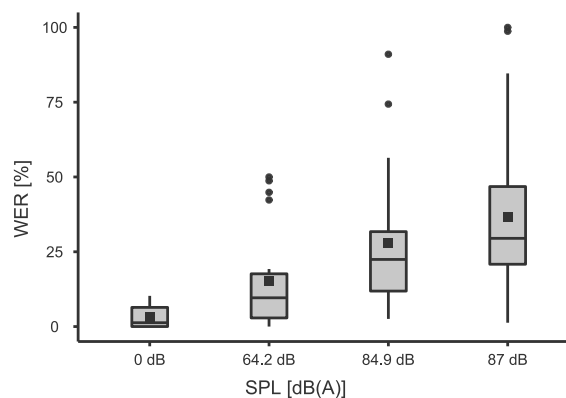


Abbildung 18. Boxplots WER versus UV SPL

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

ITR: Für die ITR wurde die statistische Signifikanz der unabhängigen Variable SPL $F(2;40) = 13,208$; $p < 0,001$; $\eta^2 p = 0,398$; Power = 1,00 mittels ANOVA mit wiederholten Messungen nachgewiesen. Die Sphärizität wurde nicht verletzt. Tabelle 15 zeigt die Post-Hoc-Tests.

SPL [dB]	<i>MD</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i> _{scheffe}	<i>Co'd</i>	Power
64,2 84,9	-11,62	1,88	20	-6,190	0,001	1,350	0,99
64,2 87,0	-8,93	2,75	20	-3,240	0,015	0,713	0,61
84,9 87,0	2,69	2,39	20	1,123	0,542	0,254	0,12

Tabelle 15. ITR [bit/min] PHT versus SPL – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Tabelle 16 zeigt die Ergebnisse der ITR mit *t*-Tests für gepaarte Stichproben der SPL-Faktorstufe 0 dB(A) im Vergleich zu den anderen. Die Normalität wurde nicht verletzt.

SPL [dB]	<i>df</i>	<i>t</i>	<i>p</i>	<i>MD</i>	<i>SE</i>	<i>Co'd</i>	<i>Power</i>
0	64,2	21	-0,937	0,180	-2,72	2,90	0,200
0	84,9	21	-4,121	0,001	-14,53	3,53	0,879
0	87,0	21	-3,461	0,001	-11,79	3,41	0,738

Tabelle 16. *t*-Test ITR [bit/min] versus SPL – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Tabelle 17 enthält die deskriptiven Werte und Abbildung 19 die zugehörigen Boxplots.

SPL [dB]	<i>n</i>	<i>M</i>	<i>Median</i>	<i>SD</i>	<i>SE</i>
0	22	38,51	40,94	13,70	2,93
64,2	44	35,83	40,84	12,61	2,69
84,9	44	24,02	23,43	12,30	2,62
87	44	26,84	28,60	13,23	2,82

Tabelle 17. DD ITR [bit/min] versus SPL – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

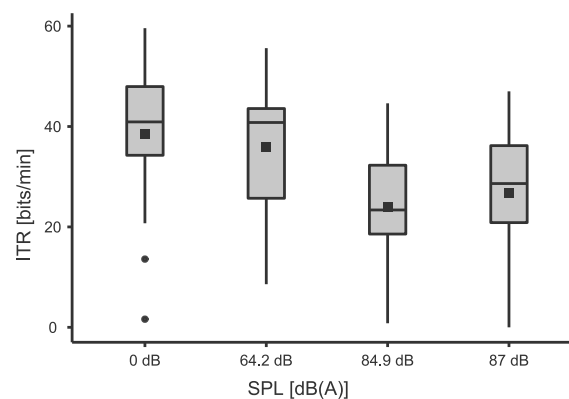


Abbildung 19. Boxplot ITR versus SPL – S2

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

UTAUT2: Alle Subdimensionen des *UTAUT2*-Fragebogens zeigten einen statistisch signifikanten Unterschied zwischen den Faktorstufen SER und EER. Die Ergebnisse der unabhängigen *t*-Tests sind in Tabelle 18 dargestellt. Die Homogenität der Varianzen war nicht verletzt, jedoch teilweise die Normalität. Entsprechend der Verletzung wurde daraufhin ein Mann-Whitney-*U*-Test durchgeführt. Tabelle 19 zeigt die Ergebnisse der deskriptiven Analyse und Abbildung 20 die Barplots der *UTAUT2*-Subdimensionen.

Dimension	Tests	<i>t/U</i>	<i>df</i>	<i>p</i>	<i>Co'd/BRK</i>	<i>Power</i>
UTAUT2_PE	Student's <i>t</i>	2,170	20	0,021	0,926	0,67
UTAUT2_EE	Mann-Whitney <i>U</i>	20,000		0,004	1,800/0,669	0,97

Dimension	Tests	<i>t/U</i>	<i>df</i>	<i>p</i>	Co'd/BRK	Power
UTAUT2_SI	Student's <i>t</i>	2,164	20	0,022	0,921	0,67
UTAUT2_FC	Student's <i>t</i>	1,981	20	0,031	0,845	0,60
UTAUT2_HM	Student's <i>t</i>	3,794	20	0,001	1,614	0,97
UTAUT2_HT	Mann-Whitney <i>U</i>	27,000		0,014	1,330/0,554	0,83
UTAUT2_BI	Student's <i>t</i>	3,651	20	0,001	1,555	0,97

Tabelle 18. Hypothesentest UTAUT2 versus RF – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

Dimension	Group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
UTAUT2_PE	SER	11	4,91	1,30	0,39
	EER	11	3,61	1,51	0,45
UTAUT2_EE	SER	11	6,00	0,87	0,26
	EER	11	4,52	1,54	0,46
UTAUT2_SI	SER	11	4,73	1,58	0,48
	EER	11	3,21	1,71	0,52
UTAUT2_FC	SER	11	5,55	1,01	0,31
	EER	11	4,27	1,88	0,57
UTAUT2_HM	SER	11	6,42	0,80	0,24
	EER	11	4,33	1,65	0,50
UTAUT2_HT	SER	11	3,21	1,89	0,57
	EER	11	1,76	1,22	0,37
UTAUT2_BI	SER	11	5,03	1,36	0,41
	EER	11	2,73	1,59	0,48

Tabelle 19. DD UTAUT2 versus RF – S2

Quelle: Adaptiert nach Rosilius, Spiertz et al. (2024), CC BY 4.0.

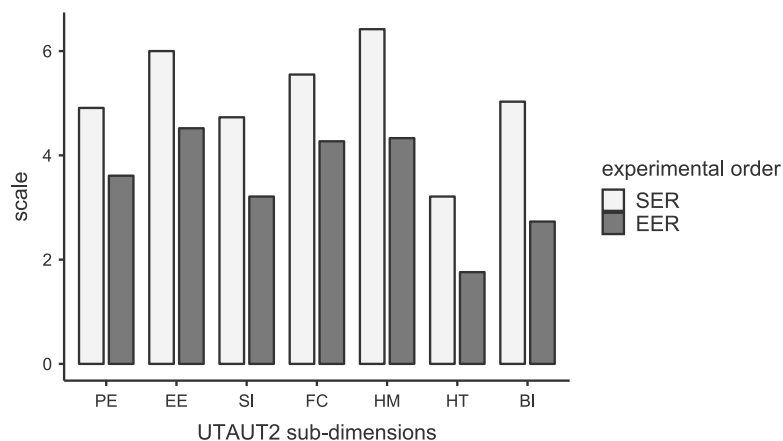


Abbildung 20. Barplots UTAUT2 versus RF – S2

Quelle: Rosilius, Spiertz et al. (2024), CC BY 4.0.

NASA-TLX: Die Skalen der durchschnittlichen Beanspruchung MD und F zeigen einen signifikanten Unterschied zwischen den Faktorstufen der RF via *t*-Tests für unabhängige Stichproben. Alle anderen Subdimensionen des NASA-TLX zeigten keine signifikanten Unterschiede zwischen den Faktorstufen der RF (siehe Tabelle 20). Tabelle 21

zeigt die Ergebnisse der deskriptiven Analyse und Abbildung 21 die Barplots der Subdimensionen.

Dimension	Tests	<i>t/U</i>	<i>df</i>	<i>p</i>	Co'd/BRK	Power
Mean_NASA	Mann-Whitney <i>U</i>	31,000		0,028	1,120/0,490	0,71
NASA_PD	Student's <i>t</i>	1,214	20,00	0,120	0,520	0,32
NASA_TD	Mann-Whitney <i>U</i>	51,502		0,286	0,301/0,149	0,15
NASA_P	Student's <i>t</i>	1,932	20,00	0,055	0,820	0,58
NASA_E	Mann-Whitney <i>U</i>	54,000		0,345	0,201/0,100	0,11
NASA_MD	Mann-Whitney <i>U</i>	35,504		0,050	0,906/0,413	0,55
NASA_F	Welch's	3,074	11,40	0,005	1,310	0,80

Tabelle 20. Hypothesentest NASA-TLX versus RF – S2

Dimension	Group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
Mean_NASA	SER	11	5,80	1,87	0,56
	EER	11	7,53	3,41	1,03
NASA_PD	SER	11	2,36	1,43	0,43
	EER	11	3,18	1,72	0,52
NASA_TD	SER	11	3,73	2,87	0,86
	EER	11	5,64	5,84	1,76
NASA_P	SER	11	4,18	4,00	1,21
	EER	11	8,45	6,15	1,86
NASA_E	SER	11	4,73	3,47	1,05
	EER	11	6,55	6,49	1,96
NASA_MD	SER	11	4,09	4,13	1,25
	EER	11	7,73	6,75	2,04
NASA_F	SER	11	3,09	1,76	0,53
	EER	11	9,55	6,74	2,03

Tabelle 21. DD NASA-TLX versus RF – S2

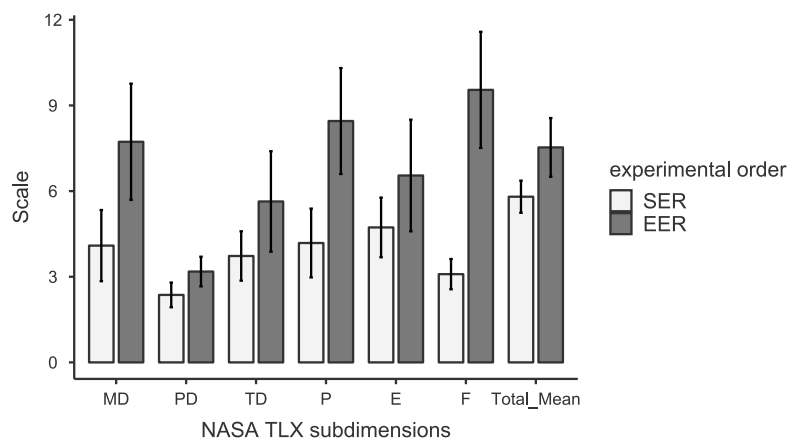


Abbildung 21. Barplots NASA-TLX versus RF (ED) – S2

physiologische Parameter: Die physiologischen Parameter konnten nicht korrekt gemessen werden. Die Probanden sind zu Fuß in den vierten Stock gestiegen und es wurde bei ihnen direkt zu Beginn die Ruhephase gemessen. Aufgrund der körperlichen Anstrengung unmittelbar vor Studienbeginn wurden die Daten verworfen, da die

Differenzrechnung aus Aktionsphase abzüglich der Ruhephase durchgehend negative Werte ergab. Diese sind somit nicht plausibel.

Die relevanten Rohdaten der Erhebung werden in Anhang A4, Tabelle 67 bis Tabelle 69, bereitgestellt.

4.2.3 Interpretation

Die Ergebnisse der Studie lassen folgende Rückschlüsse auf die Hypothesen zu:

$H_{C.1.1}$ Fehlerhafte Interaktion reduziert bei IAS die Leistung.

Nachweis. Die $H_{C.1.1}$ wird durch die Ergebnisse bestätigt; somit kann die $H_{C.1.0}$ verworfen werden. Zum einen zeigten die ANOVAs für die ASR-Leistung einen signifikanten Einfluss der unabhängigen Variable SPL sowohl hinsichtlich der WER $F(2;40) = 14,81$; $p < 0,001$; $\eta^2 p = 0,425$; Power 1,00 als auch auf der ITR $F(2;40) = 13,21$; $p < 0,001$; $\eta^2 = 0,095$; $\eta^2 p = 0,398$; Power 1,00. Zum anderen kann für die beiden AVs WER und ITR über Post-hoc-Tests (vgl. Tabelle 12, Tabelle 13, Tabelle 16, Tabelle 17) nachgewiesen werden, dass es signifikante Unterschiede zwischen den beiden Faktorstufen fh und ff gibt. Als konkrete Referenz kann beispielsweise der signifikante Unterschied zwischen 0 dB und 84,9 dB der WER $t(13) = 5,09$; $MD = 20,61$; $p < 0,001$; $Co'd = 0,884$; Power = 0,99 und ITR $t(21) = -4,12$; $MD = -14,53$; $p < 0,001$; $Co'd = 0,88$; Power = 0,99 angeführt werden. So liegt die Leistung bei Fehlerfreiheit (0 dB(A), 64,2 dB(A)) im Mittelwert bei WER 9,06 % bzw. ITR 37,15 bit/min und im Fehlerfall (84,9 dB(A), 87 dB(A)) im Mittelwert bei WER 32,37 % bzw. ITR 25,4 bit/min.

$H_{C.2.1}$ Fehlerhafte Interaktion reduziert bei IAS die Nutzerakzeptanz.

Nachweis. Die $H_{C.2.1}$ wird durch die Ergebnisse bestätigt, somit kann die $H_{C.2.0}$ verworfen werden. Alle Unterdimensionen des UTAUT2 zeigten einen statistisch signifikanten Unterschied innerhalb der unabhängigen Variable der RF (siehe Tabelle 18). Stellvertretend sind hier die Ergebnisse des t -Tests für BI zu nennen $t(20) = 3,65$; $p < 0,001$; $Co'd = 1,555$; Power = 0,97. Demnach zeigt Tabelle 19, dass alle Mittelwerte des UTAUT2 signifikant niedriger waren, wenn die Teilnehmer die RF EER erlebten.

$H_{C.3.1}$ Fehlerhafte Interaktion erhöht bei IAS die Beanspruchung des Anwenders.

Nachweis. Die $H_{C.3.1}$ wurde durch die Studienergebnisse bestätigt, somit kann die $H_{C.3.0}$ verworfen werden. Die Subskalen des NASA-TLX MD ($U = 35,5$; $p = 0,05$; $Co'd = 0,906$; Power = 0,55) und F ($t(11,4) = 3,07$; $p = 0,005$; $Co'd = 1,310$; Power = 0,8) sowie die durchschnittliche Beanspruchung ($U = 31$; $p = 0,028$; $Co'd = 1,12$; Power = 0,71) zeigen statistisch signifikante Unterschiede innerhalb der UV-RF.

4.2.4 Diskussion

Das Forschungsdesign trug u. a. der Anforderung des Herstellers Rechnung, ASR-Funktionen bis zu einem Schalldruckpegel von 90 dB(A) (bzw. > 70 dB(A)) einsetzen zu können (*Microsoft's HoloLens 2 Team Answers More Questions About Biometric Security, Audio, & Hand Tracking*, 2019). Im Gegensatz zu Derby et al. (2020) wurde der untersuchte SPL-Bereich > 70 dB erweitert und somit näher an reale, industrielle Einsatzszenarien herangeführt. In der vorgestellten Studie wurden die Experimente jedoch auf 87 dB(A) begrenzt, da dies entsprechend der Arbeitsstättenverordnung (Richtlinie 2003/10/EG des Europäischen Parlaments und des Rates vom 6. Februar 2003 über Mindestvorschriften zum Schutz von Sicherheit und Gesundheit der Arbeitnehmer vor der Gefährdung durch physikalische Einwirkungen (Lärm); 17. Einzelrichtlinie im Sinne des Artikels 16 Absatz 1 der Richtlinie 89/391/EWG, 2003) der für die Gesundheit tolerierbare Expositionspegel ist. Um gesundheitsschädliche Auswirkungen abzuschwächen und die Bedingungen in der Industrie widerzuspiegeln, wurden die Experimente mit Gehörschutz durchgeführt.

Zusammenfassend zeigen frühere Ergebnisse, dass die HL1 bei 70 dB(A) eine WER von 55 % aufwies, wobei sich die Fehlerrate zwischen den untersuchten Geräuschbedingungen signifikant unterschied, $F(1,37; 20,47) = 13,56$; $p < 0,05$; $\eta^2p = 0,475$ (Derby et al., 2020). Die HL2 erreichte nach Sinlapanuntakul et al. (2021) über den Bereich von 40 bis 70 dB(A) eine durchschnittliche WER von 5,1 %, wobei zwischen den Geräuschbedingungen keine signifikanten Unterschiede festgestellt wurden, $p > 0,05$. Hierbei sollten weitere Untersuchungen zur besseren Faktorstufendefinition der IQ an dem sog. Quality of Service, sprich der durch den Nutzer empfundenen Qualität, ausgerichtet werden.

Sinlapanuntakul et al. (2021) fanden dagegen signifikante Unterschiede nur für MD und TD. Die MD stieg von $M = 5,31$ bei 40 dB auf $M = 7,54$ bei 70 dB, TD von $M = 5,38$ auf $M = 6,46$. Für die F und den Mean_NASA wurden keine signifikanten Effekte berichtet. Derby et al. (2020) konnten trotz einer erheblichen Verschlechterung der ASR-Leistung keine signifikanten Unterschiede in den NASA-TLX-Subskalen nachweisen. Die F stieg lediglich deskriptiv von $M = 6,75$ bei 40 dB auf $M = 9,13$ bei 70 dB, blieb jedoch nicht signifikant. Auch MD unterschied sich nicht signifikant zwischen den Lärmbedingungen. Damit zeigt die vorliegende Studie einen breiteren und stärkeren Beanspruchungseffekt als die beiden Vergleichsstudien.

Im Rahmen einer kritischen Analyse von Derby et al. (2020) und Sinlapanuntakul et al. (2021) kann jedoch davon ausgegangen werden, dass das Forschungsdesign aufgrund des within-subject-Designs hinsichtlich des Erhebungszeitpunkts und der Vorbedingungen Gewöhnungseffekte („Carry-Over-Effect“, Charness et al., 2012) induziert. In der vorliegenden Studie wurden die Reihenfolgen und der Erhebungszeitpunkt so gewählt, dass die Probanden alle Bedingungen jedoch in gegenbalancierter RF isoliert durchlaufen und anschließend nur einmal in Fragebögen befragt werden. Dies ermöglicht, die Unterschiede des Fehlereffekts zwischen unabhängigen Gruppen (SER versus EER) innerhalb eines Repeated-Measures-Designs zu ermitteln (Brooks, 2012).

Im Gegensatz zu früheren Arbeiten wurde die Studie in deutscher Sprache durchgeführt, was einerseits den Umfang und die Variabilität der Beiträge erweitert, andererseits jedoch die Vergleichbarkeit einschränkt. In dieser Studie wurde ein größerer Stichprobenumfang ($N = 22$) berücksichtigt als Derby et al. (2020) mit $n = 16$ und Sinlapanuntakul et al. (2021) mit $n = 13$, was die statistische Aussagekraft erhöht. Es wurde nicht nur die Diktatfunktion, sondern auch die SPC-Funktion analysiert, welche eine weitere Funktionalität mit spezifischeren Anforderungen darstellt. Der Nachteil der Studie war, dass kein standardisierter Text zur besseren Vergleichbarkeit verwendet wurde. Zur Bewertung der SPC wurde die ITR-Metrik novelliert, deren Vergleichbarkeit und Aussagekraft jedoch aufgrund fehlender Vergleichsstudien bislang nicht nachgewiesen werden konnte.

Bezogen auf die konkrete Forschungsfrage- und -hypothesen kann konstatiert werden, dass der Interaktionsfehler bei ASR sowohl die Leistungsfaktoren WER und ITR als

auch die Akzeptanzfaktoren des UTAUT2 sowie die Beanspruchungsfaktoren des NASA-TLX signifikant beeinflusst. Je stärker das SPL, desto größer ist der Fehlereffekt. In produzierenden Umgebungen werden Nutzer zwangsläufig bei ASR-Funktionalitäten mit diesen Fehlereffekten konfrontiert. Der Fehlereffekt bei den SPCs war insgesamt geringer als bei der Texteingabe. Wenn ein Sprachbefehl ignoriert wird, kann er im Allgemeinen wiederholt werden, da die fehlende Aktion leicht identifiziert werden kann. Im Gegensatz dazu ist beim Diktat eine Wiederholung im Falle eines nichttolerierbaren Fehlereffekts nicht ohne größere Auswirkungen auf die User Journey möglich. Der ASR-Interaktionsfehler wirkte sich signifikant negativ auf alle Dimensionen des UTAUT2 aus. Hinsichtlich der Subskalen des NASA-TLX wirkte sich der Interaktionsfehler lediglich signifikant negativ auf Mean_NASA, MD (psychische Stressreaktion) und F aus. Aus dem Studiendesign kann allerdings nicht abgeleitet werden, ob die Akzeptanz- und Beanspruchungsverluste durch die fehlerhafte ASR oder durch die Exposition der SPL induziert werden. Jedoch wurde der Versuch über alle SPL-Bedingungen hinweg unter Tragen von akustischer Schutzausrüstung durchgeführt. Daher kann vermutet werden, dass die subjektiven Effekte der Lärmwahrnehmung durch die Schutzausrüstung hinreichend egalisiert wurden. Wie eingehend erwähnt, konnten die Daten der physiologischen Parameter aufgrund unerwünschter Störeinflüsse (unverhältnismäßig hoher körperlicher Anstrengung der Probanden kurz vor Versuchsbeginn) nicht für die Auswertung herangezogen werden.

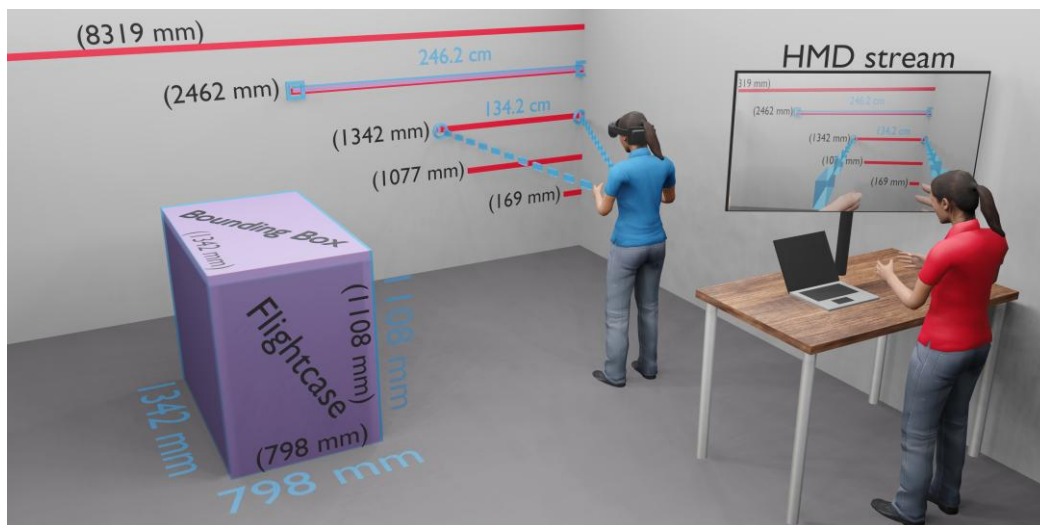
4.3 Studie 3 – Trainingsfehler

(Teile dieses Kapitels basieren auf der Veröffentlichung: Rosilius, M., Hügel, L., Wirsing, B., Geuen, M., von Eitzen, I., Bräutigam, V. & Ludwig, B. (2024). Development and Evaluation of Training Scenarios for the Use of Immersive Assistance Systems. *Applied System Innovation*, 7(5), 73. <https://doi.org/10.3390/asi7050073>. Die Inhalte, Abbildungen und Tabellen wurden für die vorliegende Monographie übernommen bzw. teilweise überarbeitet und in adaptierter Form eingebracht. Der Originalartikel wurde als Open-Access-Beitrag unter der Lizenz Creative Commons Attribution 4.0 International (CC BY 4.0; <https://creativecommons.org/licenses/by/4.0/>) veröffentlicht, die die Bearbeitung und Wiederverwendung unter Angabe der ursprünglichen Quelle gestattet.)

In dieser Studie wurden Probanden gebeten, nach dem Durchlaufen eines Trainings szenarios fachliche Teilaufgaben auszuführen. Ziel dieser Studie ist, die Auswirkungen der Trainingsqualität auf die Aufgabenleistung, Akzeptanz sowie Beanspruchung der Probanden zu untersuchen.

4.3.1 Methodik

Hierbei wurden drei Faktorstufen exponiert – von sehr einfach (unzureichender Trainingsqualität) über persönlich bzw. sehr ausführlich bis hin zu einem optimierten Trainings szenario. Die Teilaufgaben waren konsekutiv, aufeinander aufbauend über die drei Systemebenen Hardware-, Betriebssystem- und AR-Applikation und spezifisch ausgelegt. Kurz zusammengefasst bestand die sog. „User Journey“ aus augmentiertem Surfen im Internet, dem Starten des immersiven Forschungsprototyps sowie der virtuellen Vermessung von physischen Referenzobjekten im realen Versuchsaufbau (vgl. Abbildung 22).



Fußnote. Rote Linien sind physikalische Referenzen; violetter physikalischer Flight Case; schwarze Informationen dienen nur der Beschreibung für den Leser; alle blauen Anmerkungen wurden über HL2 virtuell augmentiert

Abbildung 22. schematische Darstellung des Versuchsaufbaus – S3

Quelle: Rosilius, Hügel et al. (2024), CC BY 4.0.

Design of Experiment

In dieser Studie wurde ein Between-Subjects-Design angewendet. Als UV wurde die TQ in drei unterschiedlichen Faktorstufen definiert, deren Auswirkungen wie folgt untersucht wurden. Als erste AV wurde die Leistung der Teilaufgaben anhand von zwei Metriken gemessen: dem Zielerfüllungsfaktor (ZEF) des Teilaufgabenfortschritts (analog AV-Richtigkeit) und die dafür benötigte DLZ, welche zur Berechnung des in Kapitel 3.2 beschriebenen Indexes Leistung herangezogen wurde. Zudem wurden die AVs BI (inkl. sekundärer Faktoren des UTAUT2), die durchschnittliche Beanspruchung (inkl. sekundärer Faktoren des NASA-TLX) sowie die physiologischen Parameter SCL und HR definiert.

Trainingsqualität: Die TQ wurde in drei Faktorstufen unterteilt: minimales, detailliertes und optimiertes Trainingsszenario, wobei das minimale Training den Trainingsfehler abbilden soll.

- **Minimales Training (MT):** Bei dem MT handelt es sich um ein Trainingsszenario im E-Learning-Format, bei dem ein Schritt-für-Schritt-Tutorial bestehend aus einer Mischung aus Text- und Bildbeschreibungen mit Beispielvideos angereichert wird. Zusätzlich wurde für das Interaktionstraining die native HL2 „Getting started“-Anwendung „Tipps“ verwendet, um dem Probanden ein interaktives Lernen zu ermöglichen. Jeder Teilnehmer hatte die Möglichkeit, das Tutorial so oft und so lange wie nötig zu verinnerlichen. Tabelle 22 zeigt die Trainingsinhalte für die Teilaufgaben und die jeweiligen Systemebenen. Dieses Szenario wurde so knapp wie möglich entwickelt, um ein herausforderndes, unzureichendes Trainingsszenario nicht von den Inhalten her, jedoch hinsichtlich des Trainingsumfangs abzubilden.
- **Detailliertes Training (DT):** Bei dieser Art von Training hat der Versuchsleiter dem Probanden die Inhalte (siehe Tabelle 22) persönlich vorgestellt und ausführlich erklärt. Der Teilnehmer konnte den Livestream der HL2 auf einem Monitor verfolgen. Er hatte auch die Möglichkeit, direkt Fragen zu stellen. Bei dieser Schulungsart wurde ebenfalls versucht, individuell auf die Probanden einzugehen.

- Optimiertes Training (OT): In diesem Szenario wurden verschiedene Methoden auf vorteilhafte Weise kombiniert. Das Interaktionsparadigma wurde dem Probanden wiederum interaktiv über die Anwendung „Tipps“ vermittelt. In Bezug auf Tabelle 22 wurden die grundlegenden Hardwarefunktionen über Videotutorials erklärt. Die Teilaufgaben des Betriebssystems wurden über kommentierte Videotutorials trainiert. Darüber hinaus wurde dem Probanden die Möglichkeit gegeben, die Inhalte durch angeleitetes Learning by Doing noch einmal zu üben. Die Teilaufgaben der AR-Anwendung wurden zunächst in einer persönlichen Vorführung geschult. Anschließend wurde die HL2 an den Teilnehmer übergeben und die Inhalte durch angeleitetes Learning by Doing vertieft. Der Studienleiter begleitete den Probanden währenddessen auf dem AR-Stream und unterstützte ihn bei Rückfragen. Die Faktorstufe OT wurde im Verlauf des PlanAR-Forschungsverbunds entwickelt und zur weiteren Untersuchung für diese Studie herangezogen.

Zielerfüllungsfaktor (ZEF): Der Quotient errechnet sich aus der Anzahl der Teilaufgaben, die der Proband bis zum Abbruchkriterium erfolgreich erledigen konnte (vgl. Gleichung 9). Als Abbruchkriterium wurde die Aufgabe des Probanden definiert (vgl. Genauigkeit des Systems, Dey et al., 2018).

$$\text{ZEF} = \frac{\text{Anzahl erfolgreich abgeschlossener Teilaufgaben}}{\text{Gesamtanzahl der Teilaufgaben}}, \quad (9)$$

DLZ: Beschreibt die Zeitdauer, die ein Proband für seine Teilaufgaben benötigt (vgl. Gleichung 10).

$$\text{DLZ} = \text{Summe der Zeitdauer für die erledigten einzelnen Teilaufgaben}_k, \quad (10)$$

DLZ mit Strafzeiten (DLZS): Die Gesamtzeit, die der Proband bis zur letzten erfolgreichen Teilaufgabe a benötigt hat, zuzüglich der Strafzeiten für jede nicht erfolgreiche Teilaufgabe $(15-a)$ (vgl. Gleichung 11). Die Strafzeit pro Teilaufgabe k ist definiert als das 75. Perzentil, das über alle Versuchspersonen für jede Teilaufgabe berechnet wurde (vgl. Gleichung 12).

DLZS = Gesamtzeit bis zur letzten erfolgreich abgeschlossenen Teilaufgabe_a + Summe an Strafzeiten für die nicht erfolgreich beendeten Teilaufgaben_(15-a) (11)

Strafzeit_k = Perzentil (0.75) der Durchlaufzeit für Teilaufgabe_k über alle Probanden (12)

Die Leistung je Proband wird aus ZEF dividiert durch DLZS analog Gleichung 1 berechnet.

ZEF	System Level Teilaufgabe	MT	DT	OT
01/15 Hardware	korrekte Platzierung des HMD auf dem Kopf, Starten und Herunterfahren des Betriebssystems	TB	PT	AV
02/15 Betriebssystem	Systemanmeldung über Benutzeranmeldedaten	TB	PT	AV+GLBD
03/15	Bildschirmfreigabe über Miracast mit Monitor	TB	PT	AV+GLBD
04/15	Verbindung mit WLAN	TB	PT	AV+GLBD
05/15	Starten der Browseranwendung MS Edge	NTAI	NTAI	NTAI
06/15	Eingabe der URL zur Website und des 'Promo-Codes'	NTAI	NTAI	NTAI
07/15	Schließen der MS-Edge-Browseranwendung	NTAI	NTAI	NTAI
08/15	Start des Forschungsprototyps "PlanAR"-Anwendung	AV	PT	AV+GLBD
09/15	Benutzeranmeldung über Anmeldedaten	AV	PT	AV+GLBD
10/15 AR Prototyp App	Öffnen des Hand-Menü-Dialogs und Aktivieren Zwei-Finger-Messfunktion	AV	PT	P+GLBD
11/15	Messung der ersten Referenz mittels Zwei-Finger-Messfunktion	AV	PT	PT+GLBD
12/15	Messung der zweiten Referenz über Zwei-Finger-Messfunktion	AV	PT	PT+GLBD
13/15	Wechsel zur 'Bounding-Box'-Messfunktion und Messung des physischen Referenzobjekts	AV	PT	PT+GLBD
14/15	Wechsel zur „Punkt-zu-Punkt“-Messung und Messung dedizierter Referenzen	AV	PT	PT+GLBD
15/15	Beenden aller Anwendungen, Herunterfahren des Betriebssystems und Abnehmen des HMD	NTAI	NTAI	NTAI

Fußnote. TB = Text und Bilder; AV = annotiertes Video; PT = persönliches Training; GLBD = geführtes Learning by Doing; NTAI = nicht trainiert, aber instruiert.

Tabelle 22. Überblick UV TQ & Teilaufgaben – S3

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Versuchsaufbau und -ablaufplan

Das Experiment wurde in einem Labor (8,32 m x 5,5 m) durchgeführt. Abbildung 22 zeigt den Aufbau schematisch. Zu Beginn der Studie wurden die Probanden ins Labor gebeten und setzten sich an den Tisch (unten rechts), an dem sie die Fragebögen auf einem Laptop beantworteten. In einer anschließenden Ruhephase wurde die physiologische Baseline aufgezeichnet. Der Livestream der HL2 wurde über den MS Miracast HDMI-Stick auf den Monitor übertragen. Die theoretische Trainingseinheit fand auf der linken Seite des Tisches statt. An dem Tisch konnten die Probanden sowohl die Handlung des Versuchsleiters direkt als auch den AR-Stream auf dem Bildschirm verfolgen. Fragen des Teilnehmers wurden direkt vom Versuchsleiter beantwortet. Für die Teilaufgabe 11/15 mussten die Probanden zwei Referenzmaße 169 mm und 1077 mm mittels der Zwei-Finger-Messfunktion messen. Diese Messfunktion ermöglichte dem Probanden, direkt mit beiden Händen zu interagieren. Als Verlängerung seiner beiden Unterarme wurde ein Raycast von jeder seiner Hände immersiv gegen die Wand (blau) projiziert. Die Kollision des Handstrahls mit der Wand wurde mit Hilfe eines 2D-Rings als Cursor ergänzt. Die nächstgrößeren Referenzen (1342 mm und 2462 mm) wurden für die Teilaufgabe 12/15 gemessen. Für die Teilaufgabe 14/15 wurden alle Zweipunktmaße einschließlich 8319 mm (von Wand zu Wand) mit Hilfe der Punkt-zu-Punkt-Messanwendung gemessen. Mit diesem Feature können Messpunkte als virtuelle Assets (kleine Würfel) nach dem MRTK-Interaktionsparadigma (Semple et al., 2022) beliebig translatiert und positioniert werden. Zwischen den kleinen Würfeln ist ein virtuelles Maßband eingezeichnet (siehe Maßband 2462 mm in Abbildung 22). Teilaufgabe 13/15 beinhaltet die Vermessung des Flightcases durch exakte Skalierung und Positionierung der virtuellen blauen "Bounding Box" (BB) über dem Flightcase als einhüllendes Objekt unter Verwendung des Interaktionsparadigmas des MRTK (Semple et al., 2022). Die kartesischen Dimensionen werden an der blauen BB dargestellt.

Versuchsablaufplan

Die Studie wurde gemäß des Versuchsablaufplans durchgeführt (siehe Abbildung 23). Zu Beginn des Experiments wurden die Probanden gebeten, einen soziodemografischen Fragebogen zu beantworten. TQ wurde jedem Probanden randomisiert

zugewiesen. Der Proband wurde vom Versuchsleiter entsprechend dem sich daraus ergebenden Trainingsszenario ausgebildet. Der Proband wurde in die jeweilige Teilaufgabe eingewiesen und aufgefordert, die Teilaufgabe entsprechend auszuführen. Sowohl die DLZ als auch der ZEF der Teilaufgabe wurden vom Betreuer manuell dokumentiert. Sobald der Teilnehmer die Teilaufgabe erfolgreich abgeschlossen hatte, wurde die nächste Teilaufgabe analog zum Vorgehen instruiert. Dieser Ablauf wurde so lange fortgesetzt, bis entweder alle fünfzehn Teilaufgaben erfolgreich bearbeitet wurden oder bis eine Teilaufgabe nicht erfolgreich bearbeitet wurde. Anschließend wurde die sogenannte Ruhephase (Basislinie) der physiologischen Parameter für fünf Minuten mit der tragbaren EmpathicaE4-Uhr gemessen. Am Ende des Experiments wurde der Teilnehmer gebeten, den UTAUT2- und den NASA-TLX-Fragebogen auszufüllen.

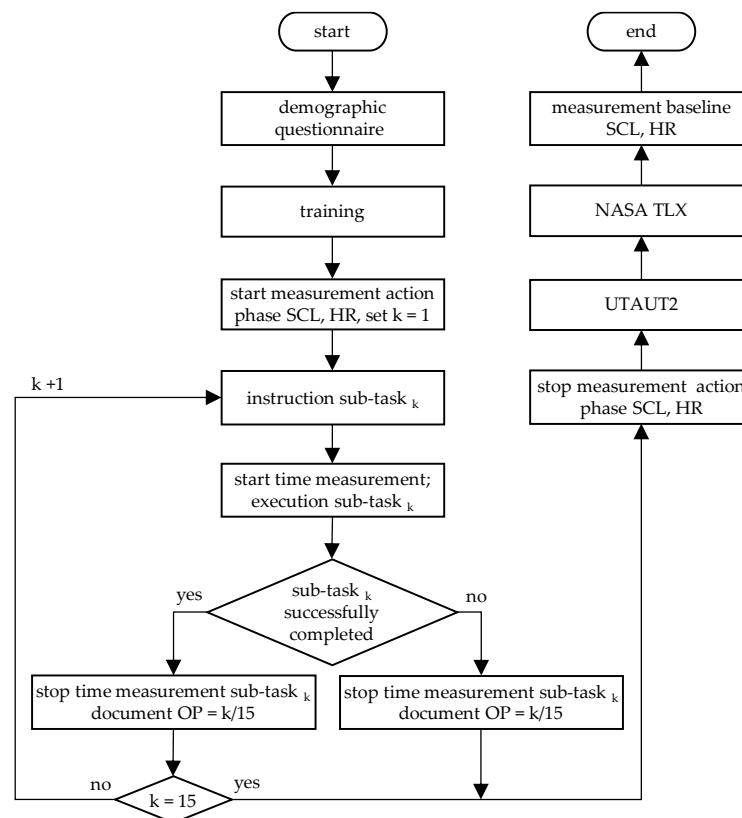


Abbildung 23. Versuchsablaufplan versus TQ – S3

Quelle: Rosilius, Hügel et al. (2024), CC BY 4.0.

Stichprobenumfangsplanung und Versuchsdurchführung

Da aus den einschlägigen Publikationen keine Schätzung des Stichprobenumfangs abgeleitet werden konnte, wurde aus den Daten der explorativen Vorstudie (in KW 43

des Jahres 2021 mit $N = 20$ Probanden) eine A-priori-Berechnung mittels G*Power V3.1.9.7 durchgeführt (vgl. Tabelle 23).

Statistische Parameter	Leistung	BI (UTAUT2)	
η^2p		0,30	0,12
Power		0,80	0,80
α		0,05	0,05
Anzahl der Gruppen		2	2
kalkulierter Stichprobenumfang		22	58

Tabelle 23. Stichprobenumfangsplanung – S3

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Die Studie wurde in den KW 25 und 26 2022 durchgeführt. Alle Datensätze wurden anonymisiert erfasst. Die Teilnehmer nahmen freiwillig an der Studie teil und erhielten eine finanzielle Aufwandsentschädigung von zwölf Euro. Jeder Teilnehmer wurde über die Richtlinien für ethische Forschung informiert und musste seine ausdrückliche Zustimmung zur Teilnahme geben. Die Dauer des Experiments betrug ca. 45 Minuten. Die Geräte wurden nach jedem Teilnehmer vollständig desinfiziert.

Statistisches Vorgehen

Ergänzend zu den Ausführungen in Kapitel 3.2 werden in diesem Kapitel studien-spezifische Vorgehensweisen vorgestellt. Um den Einfluss von TQ auf die abhängigen Variablen (ZEF, DLZS, Leistung, Akzeptanz und physiologische Parameter) zu analysieren, wurden einzelne ANOVAs mit anschließenden Post-hoc-Tests durchgeführt (E. Park et al., 2009). Zur Überprüfung der Annahmen wurden Homogenität und Normalität getestet. Bei einer Verletzung der Homogenität wurde eine Welch-ANOVA durchgeführt. Sollten die Vorbedingungen der Normalverteilung nicht gegeben sein, ist die ANOVA ausreichend robust, da ein Stichprobenumfang von mehr als 25 ($N = 52$ Probanden) erreicht wurde (Blanca et al., 2017; Schmider et al., 2010).

4.3.2 Ergebnisse

Demografische Daten

Insgesamt nahmen $N = 52$ Teilnehmer an der Studie teil, 90 % davon waren Männer und 10 % Frauen. Alle Teilnehmer waren als Studenten an der Technischen Hochschule Würzburg-Schweinfurt eingeschrieben. Die Teilnehmer waren zwischen 19 und 39 Jahre

alt, ihr Durchschnittsalter lag bei 25,8 Jahren. Abbildung 24 zeigt die Ergebnisse des Erfahrungsniveaus im Umgang mit den MR-Technologien hinsichtlich der Selbsteinschätzung der Teilnehmer.

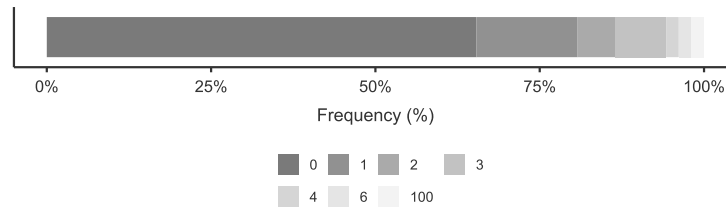


Abbildung 24. Erfahrungsverteilung – S3

Quelle: Rosilius, Hügel et al. (2024), CC BY 4.0.

Inferenzstatistische Ergebnisse

TQ: Jede Bedingung der unabhängigen Variable TQ führte zu statistisch signifikanten Veränderungen bei den abhängigen Variablen ZEF, DLZS, Leistung, Akzeptanz und mittlerer Gesamtbeanspruchung, die im Folgenden detailliert dargestellt werden.

Leistung: Hinsichtlich der Leistung zeigten sich statistisch signifikante Unterschiede für die UV TQ $F(2;29,6) = 12,83$; $p < 0,001$; $\eta^2p = 0,40$; Power = 0,99. Die Voraussetzungen der Normalverteilung und der Varianzhomogenität waren nicht erfüllt. Zur Kompensation wurde eine Welch-ANOVA berechnet, die für $N > 30$ bei fehlender Normalverteilung hinreichend robust ist. Tabelle 24 zeigt den Post-Hoc-Vergleich und Abbildung 25 die entsprechenden Box-Plots.

UV TQ		MD	df	t	p _{Scheffe}	Co'd	Power
MT	DT	-0,59	49,00	-1,274	0,451	-0,436	0,59
MT	OT	-2,47	49,00	-5,410	0,001	-1,828	1,00
DT	OT	-1,88	49,00	-4,121	0,001	-1,392	1,00

Tabelle 24. Leistung [%/min] PHT versus TQ – S3

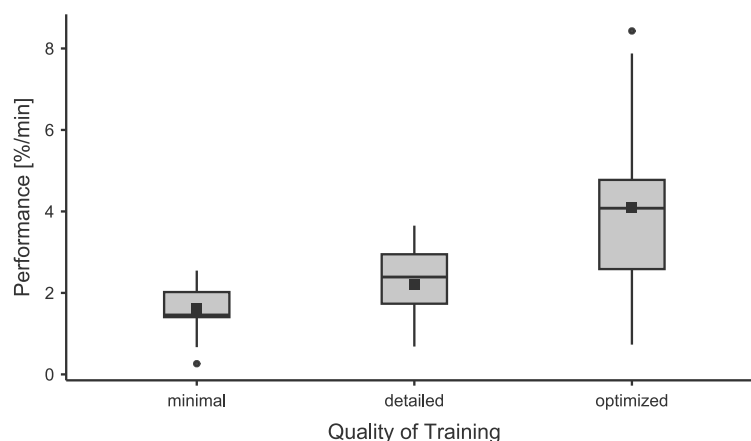


Abbildung 25. Boxplot-Leistung [%/min] versus TQ – S3

Quelle: Rosilius, Hügel et al. (2024), CC BY 4.0.

Tabelle 25 zeigt die deskriptiven Werte.

TQ	<i>n</i>	<i>M</i>	<i>Median</i>	<i>SD</i>	<i>SE</i>
MT	17	1,61	1,46	0,58	0,14
DT	17	2,20	2,39	0,79	0,19
OT	18	4,09	4,08	2,09	0,49

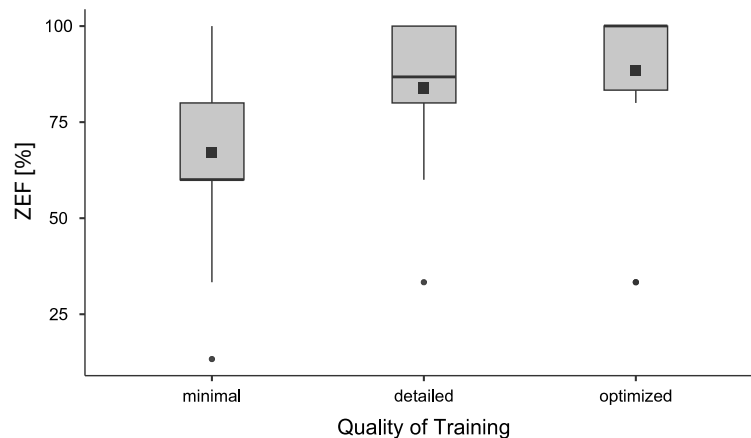
Tabelle 25. DD-Leistung versus TQ – S3

ZEF: Für den ZEF konnte mittels ANOVA ein statistisch signifikanter Unterschied innerhalb der unabhängigen Variable TQ $F(2;49) = 5,28$; $p = 0,008$; $\eta^2p = 0,177$; Power = 0,84 nachgewiesen werden. Im Gegensatz zur Varianzhomogenität ist die Voraussetzung der Normalverteilung verletzt. Aufgrund des großen Stichprobenumfangs kann die Verletzung kompensiert werden. Ein Post-Hoc-Vergleich wurde entsprechend diesen Voraussetzungen durchgeführt (vgl. Tabelle 26).

TQ	<i>MD</i>	<i>df</i>	<i>t</i>	<i>p_{Tukey}</i>	<i>Co'd</i>	<i>Power</i>
MT DT	-16,86	49,00	-2,405	0,051	-0,825	0,98
MT OT	-21,46	49,00	-3,103	0,009	-1,049	0,99
DT OT	-4,59	49,00	-0,664	0,785	-0,225	0,20

Tabelle 26. ZEF [%] PHT versus TQ – S3

Abbildung 26 stellt die entsprechenden Boxplots dar.

**Abbildung 26. Boxplots ZEF versus TQ – S3**

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Tabelle 27. DD ZEF [%] versus TQ – S3 zeigt die deskriptiven Werte.

TQ	<i>n</i>	<i>M</i>	<i>Median</i>	<i>SD</i>	<i>SE</i>
MT	17	67,12	60,00	21,10	5,13
DT	17	83,94	86,70	18,62	4,50
OT	18	88,50	100,00	21,40	5,05

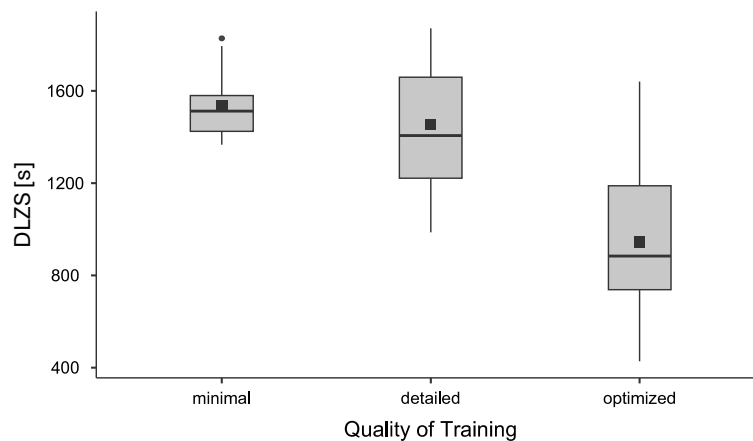
Tabelle 27. DD ZEF [%] versus TQ – S3

DLZS: Auch für *DLZS* konnten statistisch signifikante Unterschiede zwischen den TQ nachgewiesen werden $F(2;28,8) = 21,67$; $p < 0,001$; $\eta^2p = 0,51$; Power = 0,99. Da die Voraussetzung der Homogenität verletzt wurde, wurde eine Welch-ANOVA durchgeführt. Außerdem war die Normalitätsverteilung gegeben. Tabelle 28 zeigt die Ergebnisse des entsprechenden Post-Hoc-Vergleichs.

	TQ	MD	df	t	p_{Scheffe}	Co'd	Power
MT	DT	81,4	49,00	0,892	0,674	0,306	0,33
MT	OT	593,3	49,00	6,592	0,001	2,230	1,00
DT	OT	511,9	49,00	5,688	0,001	1,924	1,00

Tabelle 28. DLZS [s] PHT versus TQ – S3

Abbildung 27 zeigt die entsprechenden Boxplots.

**Abbildung 27. Boxplots DLZS [s] versus TQ – S3**

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Tabelle 29 enthält die deskriptiven Werte.

	TQ	n	M	Median	SD	SE
MT		17	1536	1512	139	33,6
DT		17	1454	1406	260	63,0
OT		18	943	883	350	82,5

Tabelle 29. DD DLZS [s] versus TQ – S3

UTAUT2: Die Subskalen des UTAUT2 wurden mittels ANOVA auf Unterschiede zwischen den drei Stufen der unabhängigen Variable TQ getestet (siehe Tabelle 30). Die Bedingung der Varianzhomogenität ist für alle Subskalen erfüllt. Die Normalitätsverteilung ist nur für die Skalen SI, FC und HM verletzt. Die ANOVA ist jedoch robust gegenüber dieser Verletzung, da die Stichprobengröße mit $N = 52$ ausreichend groß ist (siehe oben). Das Hauptaugenmerk liegt auf der BI. Die Ergebnisse zeigen einen statistisch

signifikanten Unterschied zwischen den Trainingsszenarien. Außerdem beeinflusste das Training auch die Akzeptanzfaktoren PE, EE, FC und HM.

Subskalen	<i>df1, df2</i>	<i>F</i>	<i>p</i>	η^2p	Power
PE	2,48	3,530	0,037	0,128	0,66
EE	2,48	3,410	0,041	0,124	0,65
SI	2,48	0,445	0,643	0,018	0,12
FC	2,48	4,384	0,007	0,188	0,86
HM	2,48	4,960	0,011	0,171	0,81
HT	2,48	1,310	0,280	0,052	0,29
BI	2,48	5,990	0,005	0,200	0,89

Tabelle 30. ANOVA UTAUT2 versus TQ – S3

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Tabelle 31 zeigt die Ergebnisse des entsprechenden Post-Hoc-Vergleichs.

Subskalen	TQ		<i>MD</i>	<i>t</i>	<i>p</i> _{Tukey}	<i>Co'd</i>	Power
PE	MT	DT	-0,34	-1,040	0,554	-0,363	0,44
	MT	OT	-0,85	-2,630	0,030	-0,903	0,99
	DT	OT	-0,51	-1,600	0,257	-0,540	0,76
EE	MT	DT	-0,41	-1,340	0,380	-0,467	0,64
	MT	OT	-0,78	-2,610	0,032	-0,898	0,99
	DT	OT	-0,37	-1,270	0,417	-0,431	0,57
SI	MT	DT	-0,04	-0,094	0,995	-0,033	0,05
	MT	OT	0,30	0,751	0,734	0,258	0,25
	DT	OT	0,34	0,861	0,668	0,291	0,30
FC	MT	DT	-0,46	-1,500	0,300	-0,522	0,73
	MT	OT	-1,01	-3,330	0,005	-1,143	1,00
	DT	OT	-0,55	-1,840	0,169	-0,621	0,87
HM	MT	DT	-0,35	-1,190	0,466	-0,414	0,54
	MT	OT	-0,90	-3,110	0,009	-1,068	1,00
	DT	OT	-0,55	-1,930	0,140	-0,654	0,90
HT	MT	DT	-0,75	-1,594	0,258	-0,555	0,78
	MT	OT	-0,49	-1,057	0,545	-0,363	0,44
	DT	OT	0,26	0,568	0,838	0,192	0,16
BI	MT	DT	-0,93	-2,260	0,071	-0,788	0,97
	MT	OT	-1,38	-3,420	0,004	-1,174	1,00
	DT	OT	-0,46	-1,140	0,494	-0,386	0,48

Tabelle 31. UTAUT2 PHT versus TQ – S3

Abbildung 28 zeigt das entsprechende Balkendiagramm (BD).

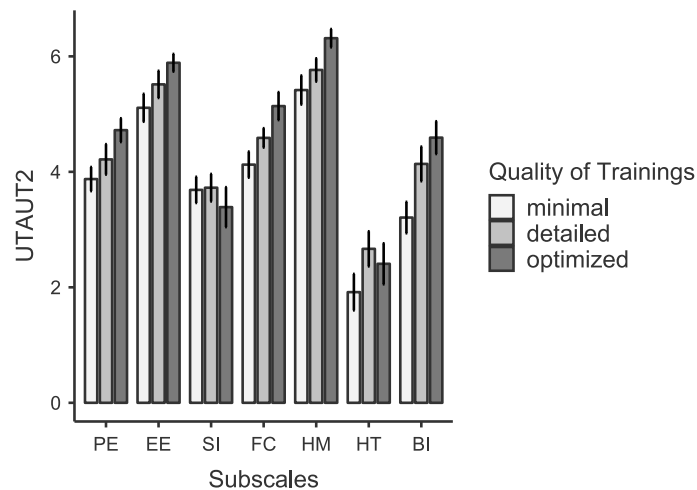


Abbildung 28. Balkendiagramm UTAUT2 versus TQ – S3

Quelle: Rosilius, Hügel et al. (2024), CC BY 4.0.

NASA-TLX: Zur Bewertung der wahrgenommenen Gesamtbelastung wurde der Mittelwert über alle Subskalen des NASA-TLX berechnet. Die ANOVA $F(2;48) = 8,15$; $p < 0,001$; $\eta^2p = 0,254$; Power = 0,96 zeigt statistisch signifikante Unterschiede in der Beanspruchung durch die TQ. Darüber hinaus wurden entsprechende ANOVAs über alle Subskalen berechnet (siehe Tabelle 32). Die Voraussetzungen für die Normalverteilung der Daten sind für alle Subskalen mit Ausnahme der körperlichen Anstrengung und Leistung erfüllt. Bei der Prüfung auf Varianzhomogenität wurden die Voraussetzungen für alle Subskalen außer für die Leistung erfüllt. Wie zuvor wurde eine Welch-ANOVA für die Analyse bei Verletzung der Voraussetzungen verwendet.

Subskalen	df1; df2	F	p	η^2p	Power
MD	2; 48	4,840	0,012	0,167	0,81
PD	2; 48	0,369	0,693	0,015	0,11
TD	2; 48	0,879	0,422	0,035	0,20
P	2; 30	5,130	0,001	0,282	0,98
E	2; 48	4,960	0,011	0,171	0,82
F	2; 48	5,790	0,006	0,194	0,87

Tabelle 32. ANOVA NASA-TLX versus TQ – S3

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Tabelle 33 zeigt die Ergebnisse des entsprechenden Post-Hoc-Vergleichs.

Subskalen	α -Korrektur	QoT	MD	t	p	Co'd	Power
MD	Tukey	MT DT	-2,37	-0,335	0,940	-0,117	0,09
		MT OT	17,19	2,464	0,045	0,846	0,99
		DT OT	19,56	2,848	0,017	0,963	1,00
PD	Tukey	MT DT	2,79	0,518	0,863	0,180	0,14

Subskalen α -Korrektur		QoT		MD	t	p	Co'd	Power
TD	Tukey	MT	OT	4,56	0,856	0,671	0,294	0,31
		DT	OT	1,76	0,336	0,940	0,114	0,09
		MT	DT	-7,52	-1,035	0,559	-0,361	0,43
		MT	OT	1,19	0,166	0,985	0,0570	0,06
P	Scheffe	DT	OT	8,71	1,235	0,439	0,4175	0,54
		MT	DT	16,10	1,960	0,159	0,681	0,92
		MT	OT	35,30	4,330	0,001	1,488	1,00
		DT	OT	19,10	2,390	0,068	0,807	0,98
E	Tukey	MT	DT	-5,29	-0,774	0,721	-0,270	0,27
		MT	OT	15,17	2,248	0,073	0,772	0,97
		DT	OT	20,46	3,081	0,009	1,042	1,00
		MT	DT	5,46	0,579	0,832	0,202	0,17
F	Tukey	MT	OT	29,42	3,162	0,008	1,086	1,00
		DT	OT	23,96	2,616	0,031	0,885	0,99
		MT	DT	3,35	0,667	0,784	0,232	0,21
		MT	OT	20,32	4,106	0,001	1,411	1,00
Mean_	Tukey	DT	OT	16,97	3,484	0,003	1,178	1,00
NASA								

Tabelle 33. NASA-TLX PHT versus TQ – S3

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

In Abbildung 29 ist das entsprechende Balkendiagramm dargestellt.

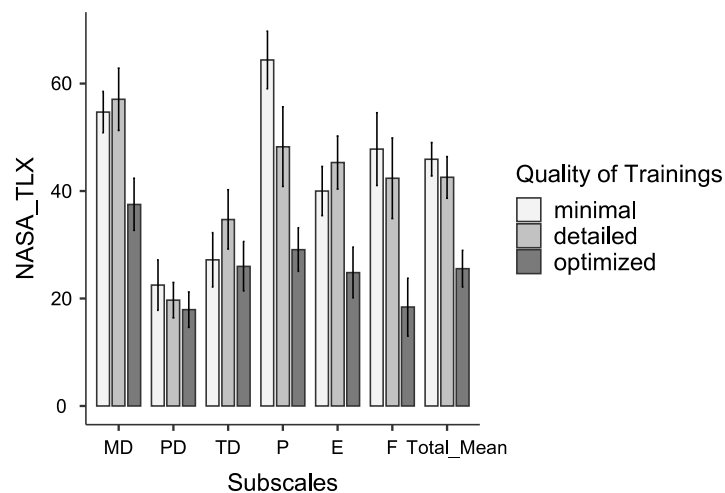


Abbildung 29. Balkendiagramm NASA-TLX versus TQ – S3

Quelle: Rosilius, Hügel et al. (2024), CC BY 4.0.

Physiologische Parameter: Der statistisch signifikante Einfluss der Trainingsqualität auf die physiologischen Reaktionen wurde durch ANOVAs für die HR_Diff $F(5;45) = 12,2$; $p < 0,001$; $\eta^2 p = 0,352$; Power = 1,00 sowie die SCL_Diff $F(2;45) = 5,20$; $p = 0,009$; $\eta^2 p = 0,188$; Power = 0,83 nachgewiesen. Die Voraussetzungen für die Varianzhomogenität wurden sowohl von der HR als auch von der SCL erfüllt, die Voraussetzung für die Normalverteilung nur von der SCL. Die Abweichung von der Normalverteilung kann

jedoch durch den hohen Stichprobenumfang von $N = 48$ kompensiert werden. Tabelle 34 zeigt die Ergebnisse der Post-hoc-Tests.

Parameter	TQ		MD	t	p	Co'd	Power
HR_Diff	MT	DT	-3,20	-1,910	0,940	-0,676	0,92
	MT	OT	4,91	2,880	0,045	1,036	1,00
	DT	OT	8,11	4,920	0,017	1,712	1,00
SCL_Diff	MT	DT	-0,35	-0,815	0,696	-0,289	0,30
	MT	OT	0,96	2,231	0,077	0,802	0,98
	DT	OT	1,31	3,130	0,008	1,090	1,00

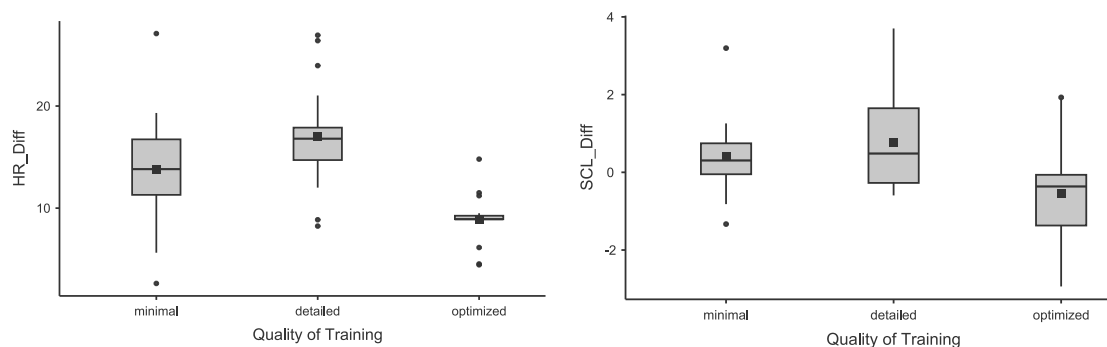
Tabelle 34. HR [BPM] & SCL [μ S] PHT versus TQ – S3

Tabelle 35 präsentiert die deskriptiven physiologischen Werte.

Parameter	TQ	n	M	SE	Median	SD
HR_Diff	MT	15	13,82	1,50	13,82	5,81
	DT	17	17,28	1,28	16,80	5,26
	OT	16	8,91	0,63	8,91	2,50
SCL_Diff	MT	15	0,41	0,26	0,31	1,02
	DT	17	0,76	0,30	0,48	1,24
	OT	16	-0,55	0,33	-0,36	1,31

Tabelle 35. DD zu HR [BPM] & SCL [μ S] versus TQ – S3

Abbildung 30 präsentiert die entsprechenden Boxplots.



(a) HR_Diff [BPM]

(b) SCL_Diff [μ S]

Abbildung 30. Boxplots (a) HR_Diff & (b) SCL_Diff versus TQ – S3

Quelle: Adaptiert nach Rosilius, Hügel et al. (2024), CC BY 4.0.

Die relevanten Rohdaten der Erhebung werden in Anhang A5 bereitgestellt.

4.3.3 Interpretation

$H_{D.1.1}$ Unzureichendes Training reduziert bei IAS die Leistung.

Nachweis. Die $H_{D.1.1}$ wird durch die Ergebnisse bestätigt, somit kann die $H_{D.1.0}$ verworfen werden. Im Einzelnen lässt sich mit statistischer Signifikanz zeigen, dass der mittlere ZEF zwischen OT und MT um 21,46 % ($p = 0,009$; Co'd = -1,049) und von DT

nach MT um 16,86 % ($p = 0,050$; $Co'd = -0,825$) abnimmt (vgl. Tabelle 26, Tabelle 27 und Abbildung 26). Aus der Perspektive der DLZS kann der Mittelwert von OT zu MT um 593,3s ($p < 0,001$; $Co'd = 2,230$) und von OT zu DT um 511,9s ($p < 0,001$; $Co'd = 1,924$) signifikant gesteigert werden (vgl. Tabelle 28, Tabelle 29 und Abbildung 27). Der Mittelwert des Index Leistung wird hinsichtlich der TQ von OT zu MT um 2,47 %/min ($p < 0,001$; $Co'd = -1,828$) und von OT zu DT um 1,88 %/min ($p < 0,001$; $Co'd = -1,392$) signifikant reduziert (vgl. Tabelle 24, Tabelle 25 und Abbildung 25). Unter der Annahme, dass das MT ein fehlerhaftes Trainingsszenario aufweist, konnte insgesamt eine statistisch signifikante Verschlechterung zwischen dem optimierten und dem fehlerhaften Trainingsszenario nachgewiesen werden. Aufgrund der hohen statistischen Power und der hohen Effektstärken kann die $H_{D.1.1}$ bestätigt werden.

$H_{D.2.1}$: Unzureichendes Training reduziert bei IAS die Nutzerakzeptanz.

Nachweis. Die $H_{D.2.1}$ wird durch die Ergebnisse bestätigt, somit kann die $H_{D.2.0}$ verworfen werden. Die zentrale Aussagekraft hinsichtlich der Nutzerakzeptanz liegt in der Dimension BI, dessen ANOVA $F(2;48) = 5,99$; $p = 0,005$; $\eta^2p = 0,200$; Power = 0,89 einen signifikanten Unterschied innerhalb der unabhängigen Variable TQ zeigt. Im Einzelnen konnte mittels Post-hoc-Tests eine signifikante Reduktion des Mittelwertes um 1.38 Einheiten (basierend auf einer 7-Punkte-Likert-Skala) von der Faktorstufe OT zu MT nachgewiesen werden. Diese Aussage wird auch durch signifikante Unterschiede in weiteren Subskalen des UTAUT2 gestützt (vgl. Tabelle 30, Tabelle 31 und Abbildung 28). Das fehlerhafte Trainingsszenario reduziert ebenfalls die Subskalen (im Vergleich zu OT) der PE (Mean_Diff = -0,847; $p = 0,030$; $Co'd = -0,903$), EE (Mean_Diff = -0,780; $p = 0,032$; $Co'd = -0,898$), FC (Mean_Diff = -1,014; $p = 0,005$; $Co'd = -1,143$) und HM (Mean_Diff = -0,898; $p = 0,009$; $Co'd = -1,068$). Diese Ergebnisse lassen den Schluss zu, dass der Nutzer bei einem unzureichenden Trainingsszenario eine geringere Leistung (Arbeit pro Zeit sowie Qualität) und weniger hedonische Motivation (Freude) bei gleichzeitig höherem Aufwand erwartet. Die Ergebnisse der Dimension FC lassen den Schluss zu, dass bei unzureichenden Trainings die Voraussetzungen wie Ressourcen, Fachwissen, Vorerfahrungen und Unterstützungen geringer eingeschätzt werden. Die Subskalen SI und HT zeigen erwartungsgemäß keine statistisch signifikanten Unterschiede, was auch logisch

nachvollziehbar ist. Unmittelbar nach einer schlechten Trainings- bzw. Arbeitserfahrung kann keine Beeinflussung auf das soziale Umfeld oder die Gewohnheiten der Probanden stattgefunden haben. HT und SI sind trägere Prozesse, die durch Erfahrungen sowie soziale Interaktion sowohl im privaten als auch im beruflichen Umfeld einhergehen.

H_{D.3.1}: Unzureichendes Training erhöht bei IAS die Beanspruchung des Anwenders.

Nachweis. Die *H_{D.3.1}* wird durch die Ergebnisse bestätigt; somit kann die *H_{D.3.0}* verworfen werden. Die ANOVA des Mean_NASA (empfundene Gesamtbeanspruchung) $F(2;48) = 8,15$; $p < 0,001$; $\eta^2p = 0,254$; Power = 0,96 zeigt über alle Subskalen des NASA-TLX einen signifikanten Unterschied innerhalb der unabhängigen Variable TQ. Der Mittelwert wurde um 20,32 Einheiten (Skala 0–100) von OT zu MT und um 16,97 Einheiten von OT und DT erhöht. Diese Aussage wird auch durch signifikante Unterschiede in weiteren Subskalen des NASA-TLX gestützt (vgl. Tabelle 32, Tabelle 33 und Abbildung 29). Die mittlere MD stieg signifikant von OT zu MT um 17,19 Einheiten ($p = 0,045$; $Co'd = 0,846$) und von OT zu DT um 19,56 Einheiten ($p = 0,017$; $Co'd = 0,963$). Die invertierte mittlere Leistung stieg von OT zu MT um 35,3 Einheiten ($p < 0,001$; $Co'd = 1,488$). Die mittlere F stieg statistisch signifikant von OT zu MT um 29,42 Einheiten ($p = 0,008$; $Co'd = 1,086$). Die mittlere E stieg signifikant von OT zu DT um 20,46 Einheiten ($p = 0,009$; $Co'd = 1,042$) und von OT zu MT um 15,17 Einheiten, was jedoch statistisch nicht nachgewiesen werden konnte ($p = 0,073$; $Co'd = 0,772$). Erwartungsgemäß konnten keine statistisch signifikanten Auswirkungen aufgrund fehlerhafter Trainings auf PD und TD festgestellt werden. Nach Alsuraykh et al. (2019) ist MD als psychologische Reaktion ein Hinweis auf Stress, der sich in dieser Studie als statistisch signifikant erwiesen hat. Die Vermutung, dass das Trainingsszenario einen Einfluss auf Stress hat, wird auch durch die physiologische Reaktion der Mittelwerte HR_Diff $F(5;45) = 12,2$; $p < 0,001$; $\eta^2p = 0,352$; Power = 1,00 und SCL_Diff $F(2;45) = 5,20$; $p = 0,009$; $\eta^2p = 0,188$; Power = 0,83 über ANOVAs bestätigt. Im Einzelnen ist die HR_Diff von OT zu MT um 4,91 BPM ($p = 0,045$; $Co'd = 1,036$) und von OT zu DT um 8,11 ($p = 0,017$; $Co'd = 1,712$) signifikant erhöht worden. Darüber hinaus zeigt die SCL_Diff auch einen signifikanten Anstieg von DT zu OT um 1,309 μS ($p = 0,008$; $Co'd = 1,090$).

4.3.4 Diskussion

Im Gegensatz zu den meisten AR-Studien (De Giorgio et al., 2023; Garcia Fracaro et al., 2022; Martins et al., 2023) konzentriert sich die vorliegende Studie nicht auf das Lernen mit AR-Systemen, sondern auf Trainingsszenarien für immersive Systeme. Im Rahmen der kritischen Literaturrecherche konnten bislang keine Beiträge gefunden werden, die sich mit den Auswirkungen von Trainingsszenarien bzw. mit unzureichender Qualität sowie deren Evaluationsmethoden für das Erlernen des Einsatzes immersiver Systeme überhaupt (geschweige denn im industriellen Kontext) beschäftigen. Dörner und Horst (2021) haben einen AR-Parkour für das universitäre Lernumfeld entwickelt, dessen Implikationen nur sehr eingeschränkt auf den industriellen Einsatz übertragbar sind. Außerdem hatten die Autoren weder eine Evaluation durchgeführt noch Auswirkungen der Trainingsszenarien untersucht. Diese Studie adressiert die bestehende wissenschaftliche Unsicherheit darüber, wie und in welchem Ausmaß die – insbesondere unzureichende – TQ die Leistung, Akzeptanz und Beanspruchung beeinflusst. Die Ergebnisse leisten einen Beitrag hinsichtlich der Forschungslücke: Auch wenn hierüber ein breiter Konsens herrscht, dass Trainingsszenarien an sich ein wichtiger Faktor sind (Day et al., 2019; Marshall et al., 2008; Quandt & Freitag, 2021), ist in diesem Forschungskontext ein Bedarf nach der Untersuchung konkreter Auswirkungen schlechter TQ implizit gegeben (Graser & Böhm, 2022; Palanque et al., 2020; Rosilius, Spiertz et al., 2024).

Auswirkung der TQ auf die Leistung: Im Gegensatz zu den Ergebnissen von Zoetbrood (2021) zeigt die vorliegende Studie, dass die Trainingsqualität tatsächlich einen signifikanten Einfluss auf den Index Leistung hat ($p = 0,009$; $\text{Co'd} = -1,049$). Die Korrelationsmatrix von Zoetbrood (2021) weist nur einen sehr geringen, nicht statistisch signifikanten negativen Zusammenhang (Pearson = $-0,069$; $p > 0,05$) zwischen dem Einsatz von Ausbildungspraktiken und der Produktivität auf. Weiter konnte die Hypothese „Je mehr Schulungs- und Entwicklungspraktiken in einer Organisation eingesetzt werden, desto größer ist der Produktivitätsnutzen aufgrund von IKT-Investitionen“ auf der Grundlage der multivariaten Analyse nicht bestätigt werden ($\beta = 0,27$, $t(98) = 1,003$, $p = 0,318$; Zoetbrood, 2021). Ebenso zeigte sich bei Zoetbrood (2021) kein signifikanter Haupteffekt der Schulungs- und Entwicklungspraktiken auf die Arbeitsproduktivität ($\beta = 0,034$; $t(98) = 1,047$; $p = 0,298$). Zudem führte die Berücksichtigung des

Interaktionsterms lediglich zu einer geringen und statistisch nicht signifikanten Erhöhung der erklärten Varianz ($\Delta R^2 = 0,015$; $F\text{-Change}(2,98) = 0,980$; $p = 0,379$). Insbesondere die Entwicklung der Hypothese von Zoetbrood (2021) basiert nur auf indirekten und schwachen Annahmen, sodass hier weiterer Forschungsbedarf besteht ist. Darüber hinaus lassen das Forschungsdesign und die darin abgebildeten Vorarbeiten keine Rückschlüsse auf eine differenzierte Analyse der Qualität und Quantität der inhärenten Trainingsszenarien zu (Zoetbrood, 2021), was wiederum einen relevanten Unterschied zu dem präsentierten DOE darstellt. Gleichzeitig weist Zoetbrood (2021) auf weitere periphere, organisationale Interaktionseffekte hin, wonach die Wirkung von Trainingsszenarien nur dann effektiv genutzt werden kann, wenn auch andere förderliche Bedingungen, wie z. B. die Integration von Bildungsmaßnahmen in Prozessstrukturen, verbunden mit organisatorischen Veränderungen und der notwendigen Kontinuität umgesetzt werden. Die präsentierten Ergebnisse hinsichtlich der Sensitivität der Leistung in Bezug auf fehlerbehaftete TQ unterstützen das postulierte Missmanagement als eine Ursache des Produktivitätsparadoxons (Schweikl & Obermaier, 2020), was darauf hindeutet, dass eine unzureichende Schulung der IKT zu einem erheblichen Produktivitätsverlust führen kann. Auch die Aussage von Schweikl und Obermaier [3], dass Unternehmen in Aus- und Weiterbildungsprogramme investieren sollten, um den richtigen Umgang mit neuen Technologien zu gewährleisten und ihre Belegschaft zu stärken, wird durch die vorliegenden Ergebnisse gestützt. Jedoch lassen sich aufgrund der Unterschiede in den verwendeten Technologien, DOEs und Hypothesen die Ergebnisse der Vorarbeiten nur eingeschränkt mit den präsentierten Ergebnissen vergleichen.

Auswirkung der TQ auf die Akzeptanz: In Übereinstimmung mit der Studie von Marshall et al. (2008) aus dem Bereich Kieferchirurgiesysteme über den Einfluss der Endanwenderschulung auf die Akzeptanzfaktoren PE (Pearson = 0,769 mit Signifikanz) und EE (Pearson = 0,840 mit Signifikanz) konnten die Ergebnisse analog bei IAS mit PE ($p = 0,030$; $\text{Co'd} = -0,903$) und EE ($p = 0,032$; $\text{Co'd} = -0,898$) bestätigt werden. Jedoch gilt es, die DOEs und Fragestellungen hierbei differenziert zu betrachten. Marshall et al. (2008) berichten starke Korrelationen, während sich in der vorliegenden Studie große standardisierte Gruppenunterschiede zeigen. Jedoch lassen Marshall et al. (2008) keine Rückschlüsse auf den Umfang und die Qualität der Trainingsszenarien zu. In ähnlicher Weise

zeigten Alqahtani et al. (2022), dass die wahrgenommene Ausbildungsqualität einen signifikanten Einfluss auf die beiden Akzeptanzfaktoren wahrgenommene Benutzerfreundlichkeit PEU (Ausbildungsqualität auf PUE $\beta = 0,132$; $p = 0,005$) und wahrgenommenen Nutzen PU (Ausbildungsqualität auf PU $\beta = 0,096$; $p = 0,015$) für E-Learning-Systeme hat. Jedoch wurde durch Alqahtani et al. (2022) kein direkter, signifikanter Einfluss der Trainingsqualität auf die Nutzungsabsicht untersucht. Lediglich die Beziehung von Trainingsqualität auf die Nutzungsabsicht wurde korrelativ mit $r = 0,461$ im Strukturgleichungsmodell beziffert (Alqahtani et al., 2022). Die vorliegende Studie weist ebenfalls einen signifikanten Einfluss der Trainingsqualität auf die konzeptionell verwandten UTAUT2-Konstrukte PE und EE nach. Für PE zeigte sich ein signifikanter Unterschied mit großer Effektstärke ($p = 0,030$; $d = -0,90$), ebenso für EE ($p = 0,032$; $d = -0,898$). Darüber hinaus konnte – im Unterschied zu Alqahtani et al. (2022) – ein statistisch signifikanter Einfluss der Trainingsqualität auf die Verhaltensabsicht (BI) und damit auf die Akzeptanz immersiver Assistenzsysteme nachgewiesen werden. Wie bereits erwähnt, wurde im Gegensatz zu Alqahtani et al. (2022) und Mariani et al. (2013) keine konfirmatorische Faktorenanalyse auf der Basis des TAM-Modells durchgeführt. Mariani et al. (2013) wiesen signifikante direkte Effekte der wahrgenommenen Trainingsmöglichkeiten auf die wahrgenommene Benutzerfreundlichkeit ($\beta = 0,16$; $p < 0,01$) und die wahrgenommene Nützlichkeit ($\beta = 0,21$; $p < 0,01$) nach.

Auswirkung der TQ auf die Beanspruchung: In Übereinstimmung mit den Schlussfolgerungen von Korunka und Vitouch (1999) und K. Wang et al. (2008) sind die Ergebnisse der vorliegenden Studie mit der Annahme vereinbar, dass eine angemessene TQ technikbezogene Beanspruchung reduzieren kann. Diese Vorarbeiten deuten auf die Notwendigkeit geeigneter Trainingsszenarien hin, lieferten jedoch bislang keinen Nachweis für diesen Zusammenhang. Bala und Venkatesh (2016) untersuchten die Auswirkungen der Trainingseffektivität auf die wahrgenommene Bedrohung durch IKT, wobei sich hochsignifikante Effekte zeigten. In der vorliegenden Arbeit stieg die wahrgenommene Gesamtbelastung (Mean_NASA) über verschiedene Subskalen des NASA-TLX um 20,32 Einheiten von OT zu MT und um 16,97 Einheiten von OT zu DT. Andere Subskalen des NASA-TLX wiesen ebenfalls signifikante Unterschiede auf (siehe Tabelle 33). Der mittlere MD als psychologische Reaktion auf Stress stieg signifikant von OT und MT um

17,19 Einheiten ($p = 0,045$; $Co'd = 0,846$) und zwischen OT und DT um 19,56 Einheiten ($p = 0,017$; $Co'd = 0,963$). Die physiologischen Reaktionen bestätigten ebenfalls die Auswirkungen eines unzureichenden Trainings. HR_Diff steigt signifikant von OT zu MT um 4,91 BPM ($p = 0,045$; $Co'd = 1,036$) und von OT zu DT um 8,11 BPM ($p = 0,017$; $Co'd = 1,712$). SCL_Diff zeigt einen signifikanten Anstieg von DT zu MT um $1,309 \mu S$ ($p = 0,008$; $Co'd = 1,09$). Die Ergebnisse des SCL_Diff zeigen jedoch einen auffälligen Unterschied im Detail, da DT im Vergleich zu MT ($p = 0,008$) einen statistisch signifikanten Unterschied aufweist, MT im Vergleich zu OT jedoch nicht ($p = 0,077$). Die Analysen zeigen negative Werte für Mean SCL_Diff bei OT, d. h., dass die SCL in der Baseline (Ruhephase) höher war als in der Aktionsphase. Dies könnte auf die hohe Störanfälligkeit des Messsystems sowie externe Störfaktoren (Kosch et al., 2023) zurückzuführen sein, z. B. durch Abnutzung der Elektroden (Setz et al., 2010), Hautpflegeprodukte, Messfehler durch Bewegung, Desinfektion der Hände oder falsche Positionierung.

Nach Barnard et al. (2013) und Mariani et al. (2013) spielen Alter, Bildungsstand, Technikaffinität, Lernbereitschaft und organisatorische Bedingungen sowie die Erwartung von produktivem Erfolg eine zentrale Rolle bei der Differenzierung der Faktoren. Hierdurch könnte die große Standardabweichung der mittleren Leistung im OT-Szenario von 2,09 %/min erklärt werden. Diese Erkenntnisse decken sich ebenfalls mit den Erfahrungen aus dem Forschungsverbundprojekt PlanAR.

Eine kritische Betrachtung dieser Studie lässt die folgende Behauptung zu: Mit zusätzlichem Entwicklungsaufwand hätten die Faktorstufen der unabhängigen Variable TQ wirkungsvoller gestaltet werden können. Um realistische Unternehmensbedingungen abzubilden, wurden im MT-Szenario die Lerninhalte korrekt bereitgestellt, jedoch mit nur vereinfachten und didaktisch unzureichend gestalteten Trainingsmethoden. Selbstredend gibt es keine Grenze, wie unzureichend (fehlerhaft) dieses Szenario gestaltet werden kann. Die mittlere DT-Stufe wurde so konzipiert, dass ein Trainer versucht, die Trainingsinhalte so effektiv wie möglich persönlich in einer Trainingssitzung zu vermitteln. Hierbei wären auch alternative Szenarien anwendbar, z. B., dass ein Trainer eine GLBD-Sitzung durchgehend begleitet, was theoretisch vielversprechende Ergebnisse liefern könnte. Wie in Abschnitt 4.3.1 beschrieben, wurde ein optimiertes Szenario mit

gemischten Methoden gewählt. Für das OT-Szenario, wie Trainingsszenarien optimal gestaltet werden können, herrscht umfangreiches Forschungspotenzial.

Im Gegensatz zu den bisherigen Usability-Studien von Cao et al. (2020) und Dey et al. (2018), in denen die Leistungen stets unabhängig voneinander über die Dauer und die Richtigkeit bzw. Zielerfüllung bewertet wurden, wurde in diesem DOE der Index Leistung adaptiert und angewendet. Dieser hat den Vorteil, dass eindeutige und aussagekräftige Ergebnisse ohne nachträgliches Interpretieren der beiden abhängigen, konkurrierenden Variablen Dauer und Richtigkeit (vgl. richtige, lange Durchführung versus falsche, aber kurze Durchführung) notwendig sind, um die Nutzerleistung zu bewerten. Die Leistung wird anhand des sog. „Speed-Accuracy-Tradeoffs“ (Standage et al., 2015) gemäß Gleichung 1 ermittelt. Dieser Index ist analog zu den von Rosilius et al. (2022) vorgeschlagenen Sichtbarkeitsindex. In Bezug auf den übergreifenden Forschungsansatz von Rosilius, Spiertz et al. (2024) hat der Beitrag gezeigt, dass ein mangelhaftes Trainingsszenario einen negativen Einfluss auf Leistung, Akzeptanz und Beanspruchung hat. Damit wurde ein Beitrag in diesem Forschungskontext geleistet und ein eigenes Forschungsdesign für die UV TQ entwickelt, das einen Ansatz für weitere Forschungsaktivitäten bietet. Quantitativ konnte die BI durch ein fehlerhaftes Trainingsszenario um durchschnittlich 1,4 Einheiten reduziert werden. Im Vergleich zu Rosilius, Spiertz et al. (2024) wurde die BI durch eine fehlerhafte Sprachinteraktion im Mittel um 2,3 Einheiten verschlechtert. Aufgrund der unterschiedlichen Versuchsdesigns ist dieser Vergleich lediglich deskriptiv zu interpretieren.

4.4 Studie 4 – Verbundstudie

Das Hauptziel der Verbundstudie ist, die Fehlerauswirkungen unterschiedlicher Fehlertypen störeinflussarm in einem einheitlichen Versuchsaufbau bzw. an einem Forschungsprototypen zu untersuchen und zu vergleichen. Entsprechend den Hypothesen wurde hierbei je Teilstudie einem dedizierten Forschungsziel nachgegangen:

Erstens wird in Teilstudie 4.1 (vgl. $H_{A-C.1-3}$ bzw. $FF_{1.A-C}$) der Einfluss unterschiedlicher Fehlertypen (Datenfehler, Sichtbarkeitsfehler und Interaktionsfehler) unabhängig analysiert. Dies wird u. a. als Erweiterung zu den Studien 1 und 2 durchgeführt. Zweitens untersucht Teilstudie 4.2 (vgl. $H_{1-3.1}$ bzw. $FF_{1.A-C}$) die Unterschiede zwischen den

Fehlertypen. Drittens wird in Teilstudie 4.3 (vgl. *H_{4.1} bzw. FF₃*) der Einfluss der Moderatorvariable Systemkonfidenzanzeige als mögliche Manipulation auf die Akzeptanz (BI) hin analysiert.

4.4.1 Methodik

In dieser Studie wurden die Probanden im Rahmen einer fachlichen Aufgabe in zwei Versuchsbedingungen – mit und ohne Fehlereinfluss – dazu aufgefordert, die Maße eines Tisches (Länge, Tiefe und Höhe) so schnell und präzise wie möglich zu ermitteln. Der Tisch lag dabei umgedreht mit der Tischfläche nach unten auf dem Boden (vgl. Abbildung 30). Als digitales Messwerkzeug nutzten die Probanden die BB-Messfunktion des IAS. Die grundsätzliche Funktionsweise dieser fachlichen Messaufgabe wurde bereits in Kapitel 4.3 erläutert.

Design of Experiment

Die Studie wurde in einem MMD (Murrar & Brauer, 2018) durchgeführt. Hierbei wurde ein randomisiertes, paritätisch-gegenbalanciertes Design zur Kompensation von potenziellen Störeffekten angewendet, die sich aus RF ergeben würden. Es wurden zwei UV definiert: Zum einen wurde die UV Fehlertyp (FT) mit den drei Between-Subject-Faktorstufen Datenfehler, Sichtbarkeitsfehler und Interaktionsfehler eingesetzt. Zum anderen wurde die UV Fehlerfall mit den beiden Within-Subject-Faktorstufen ff und fh analog den entsprechenden UVs DQ (des Datenfehlers), VQ (Sichtbarkeitsfehler) und IQ (Interaktionsfehler) herangezogen. Als AVs wurden entsprechend dem übergeordneten Forschungsdesign (vgl. Kapitel 3.2) DLZ, Richtigkeit, IL, BI inkl. sekundärer Akzeptanzfaktoren des UTAUT2, die durchschnittliche Beanspruchung inkl. der sekundären NASA-TLX-Faktoren sowie die physiologischen Parameter SCL und HR definiert. Jeder Proband durchlief die Studie zweimal hintereinander, einmal in der ff- und einmal in der fh-Bedingung in einer Faktorstufe der UV Fehlertyp. Die Reihenfolge, in der ein Proband startete (ob mit fh oder ff), wurde per Zufallsprinzip aus einer festgelegten Gesamtheit gemäß dem Versuchsplan bestimmt. A priori wurde definiert, dass über das DOE hinweg die Häufigkeiten der Startbedingung paritätisch vorliegen. Somit konnten per se Reihenfolgeeffekte minimiert werden. Weiter wurden die Reihenfolgen der UV

Fehlertyp innerhalb des Versuchsplans ebenfalls randomisiert. Als Moderatorvariable wurde die Anzeige der Systemkonfidenz definiert.

Fehlerfall des Datenfehlers: In dieser Versuchsbedingung wurden die annotierten, numerischen Kantenlängen der BB um den Faktor 2 vergrößert. Den Probanden wurden somit idealtypisch anstelle der tatsächlichen Maße $T0,74 \text{ m} \times L1,81 \text{ m} \times H0,75 \text{ m}$ die fehlerbehafteten Dimensionen $T1,48 \text{ m} \times L3,62 \text{ m} \times H1,50 \text{ m}$ angezeigt. Die Seitenverhältnisse der BB blieben dabei unverändert.

Fehlerfall des Sichtbarkeitsfehlers: Im Fehlerfall des Sichtbarkeitsfehlers wurde die BB in der Farbe RGB 21/59/132 mit 50 % Opazität und 30 % Displayhelligkeit dargestellt – anstelle der nativen Darstellung mit RGB 21/59/132 bei 100 % Opazität und 100 % Helligkeit.

Fehlerfall des Interaktionsfehlers: In dieser Versuchsbedingung wurde anstelle der nativen Steuerung mit drei Freiheitsgraden (Translation entlang der Achsen X, Y, Z) die vollständige Manipulation über alle sechs Freiheitsgrade ermöglicht. Dabei wurden sowohl Rotationen um die Achsen (XX, YY, ZZ) als auch Translationen (X, Y, Z) beim Greifen der BB (virtuelle Drahtgitterbox) als Steuerungsbefehl übergeben. Das bedeutet, dass die Nutzer sowohl in der direkten („Near-Interaction“) als auch in der indirekten („Far-Interaction“) Manipulation gemäß Iolambean (2023) die BB nicht nur verschieben, sondern auch rotieren konnten. Die Pose der BB reagierte mit hoher Sensitivität auf die Handbewegungen, was das präzise Positionieren deutlich erschwerte.

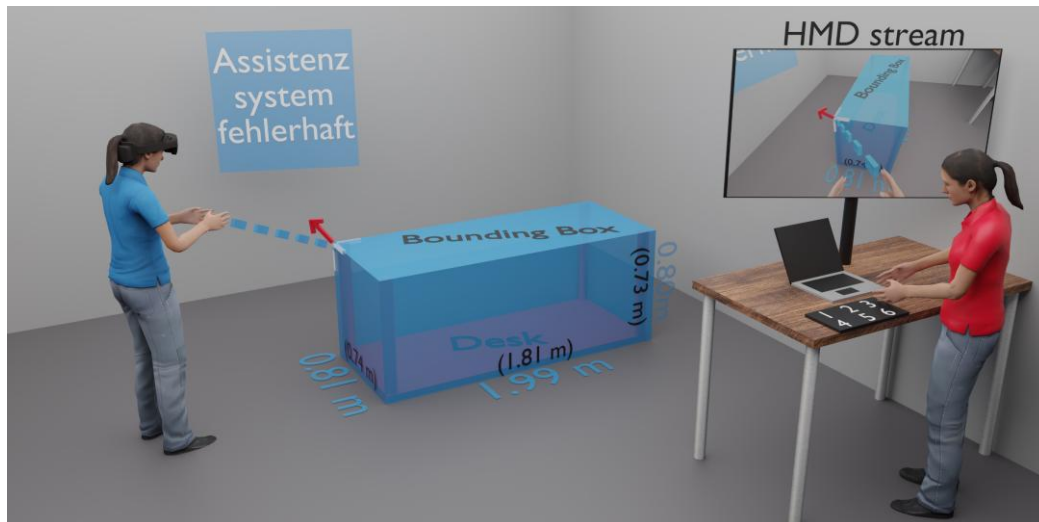
Anzeige der Systemkonfidenz: Die Moderatorvariable Systemkonfidenz diente als visuelle Rückmeldung und zeigte den Probanden die jeweils gültige Faktorstufe der UV Fehlerfall an. Entsprechend Hypothese $H_{4.1}$ wird angenommen, dass diese Anzeige die BI positiv beeinflusst. Sie wurde den Probanden jedoch ausschließlich im fh-Durchlauf des jeweiligen Fehlertyps eingeblendet und im ff-Durchlauf nicht präsentiert.

Bounding-Box: Entsprechend dem MRTK von Microsoft stellt die BB ein interaktives Begrenzungsrahmenobjekt dar, mit dem über das MRTK-Interaktionsparadigma (Semple et al., 2022) durch Skalieren, Rotieren und Translatieren interagiert werden kann. Der Benutzer kann an den Eckgriffen (sog. Corner Handles) diagonal ziehen, um die Größe des Objekts zu verändern. Mithilfe der Drehgriffe (sog. Rotation Handles), die mittig auf den jeweiligen Kanten des Rahmens angeordnet sind, kann die BB um die

entsprechende Achse durch den Objektschwerpunkt rotiert werden. Das Objekt kann innerhalb des Rahmens gegriffen und in der Ebene translatiert werden (radicalad, 2021). Im Rahmen dieser Studie konnte die BB ausschließlich proportional skaliert werden.

Probandenbefragung: Neben den oben erwähnten Umfängen des UTAUT2 und NASA-TLX werden in der Verbundstudie ebenfalls die soziodemografischen Daten (vgl. Tabelle 58) sowie der studienspezifische Fragebogen erhoben (vgl. Tabelle 62). Zum einen werden die Vorerfahrung und das Sehvermögen erhoben, zum anderen wird zum Ende des Versuchs abgefragt, ob und welche Fehlertypen die Probanden subjektiv wahrgenommen haben.

Teilstudien: Für die Teilstudien 4.1 und 4.2 wurden die Daten gemeinsam erhoben; die Differenzierung der Studien wurde erst im Rahmen der Auswertung relevant. Das Ziel der Teilstudie 4.1 bestand darin, die Auswirkungen verschiedener Fehlerarten im Kontext eines Forschungsprototyps sowie einer fachlichen Aufgabe zu untersuchen. Die Teilstudie 4.2 hatte zum Ziel, mögliche Unterschiede zwischen den Fehlerarten in Bezug auf die AVs zu identifizieren. Zur Minimierung von Störeinflüssen durch individuelle Dispositionen der Probanden – etwa Begabung oder Vorerfahrung – wurde ein Auswertungsverfahren gewählt, das statistische Verzerrungen durch solche Unterschiede reduziert. Hierzu kam die sogenannte Change-Score-Metrik zum Einsatz (Becker, 1988). Laut Becker (1988) bieten Change-Score-Analysen den Vorteil, Veränderungen direkt zu messen, individuelle Unterschiede in Ausgangswerten zu kontrollieren und die statistische Power zu erhöhen – insbesondere bei hoher Korrelation zwischen Vorher- und Nachher-Werten. Diese Methode stellt eine etablierte Praxis in empirischen AR-Studien dar (Garzón & Acevedo, 2019). Konkret wurden für die deskriptiven und inferenzstatistischen Verfahren nicht die absoluten AVs der fh-Bedingungen herangezogen, sondern die individuellen Differenzwerte berechnet (fh abzüglich ff je Proband). Die Werte und Vorzeichen wurden dabei so definiert, dass ein erwarteter Fehlereffekt stets als positiver Wert dargestellt wird. In der Teilstudie 4.3 wurde auf Basis der Ergebnisse aus Teilstudie 4.1 die Fehlerbedingung mit der höchsten Effektstärke ausgewählt, um anschließend den Einfluss der Moderatorvariable Systemkonfidenzanzeige auf die Akzeptanz (BI) zu analysieren.



Fußnote. Blaue Elemente stellen augmentierte Inhalte dar; schwarze Annotationen sind lediglich Anmerkungen für den Leser; die Kantenlängen wurden an der virtuellen BB in Echtzeit präsentiert

Abbildung 31. Schematische ED des Versuchsaufbaus Verbundstudie 4

Versuchsaufbau und -ablaufplan

Das Experiment wurde durchgehend in einem Labor mit matten, grauen Fußböden durchgeführt, um optische Irritationen zu vermeiden. Die Abbildung 31 präsentiert den schematischen Aufbau der Studie und Abbildung 32 den entsprechenden Versuchsablaufplan. Zu Beginn der Studie wurden die Probanden gebeten, sich an den Tisch (vorne rechts) zusetzen, um am Laptop die Fragebögen zu beantworten. In einer anschließenden Ruhephase wurde die Baseline der physiologischen Parameter aufgezeichnet. Der Livestream der HL2 wurde über den MS-Miracast-HDMI-Stick auf den Monitor übertragen. Im Anschluss wurden die Probanden in das Vorgehen der Studie sowohl allgemein als auch in den Umgang mit der HL2, der Interaktion sowie in die Messanweisung eingewiesen. Die Probanden konnten die Einweisung teilweise direkt als auch über den Livestream verfolgen, während der Versuchsleiter sie unterwies. Zum Maßnehmen sollten die Probanden den Tisch (vgl. roten Desk in Abbildung 31) via der virtuellen BB passend einhüllen. Um die Komplexität der fachlichen Aufgabenstellung zu reduzieren und die Vergleichbarkeit der Ergebnisse bzw. der Fehlereffekte zu erhöhen, wurde den Probanden ein gewisser Formalismus zum Messen vorgeschlagen: Zuerst sollten die Probanden die BB kongruent entlang der Tischkanten an einer Ecke positionieren und ausrichten. Da die BB lediglich proportional skalierbar ist, kann sie im Anschluss über die

obere Ecke zur Tischmitte hin auf die korrekte Tischgröße skaliert werden. Als nächstes wurden die Probanden gebeten, unter Anleitung praktische Erfahrungen mit der Handhabung und vor allem mit dem Interaktionsprinzip des Forschungsprototyps zu sammeln (durchschnittlich ca. fünf Minuten), um sich mit der immersiven Messanwendung vertraut zu machen. Ziel dieser Übung war es, dass die Probanden die Interaktion bewusst und reproduzierbar bedienen können. Danach bereitete der Versuchsleiter die Ausgangsanordnung der immersiven Szene standardisiert vor: Die BB wurde initial mit einem Skalierungsfaktor von 0,5 instanziiert und im lichten Abstand von 1 m über dem Tisch mittig platziert. Je nach Bedingung wurde die Systemkonfidenzanzeige den Probanden als Dialogfeld „Assistenzsystem fehlerhaft“ links neben der BB (relativ 1 m/1 m/1,5 m bzgl. linker oberer Ecke) mit weißer Schrift (RGB 245/245/245) auf blauer Hinterlegung (RGB 21/59/132 mit 100 % Opazität) visualisiert und dank der Objekteigenschaft des sog. „Billboard-Effekts“ stets dem Betrachter zugewandt (Anandapadmanaban, 2020, S. 63). Über die Bluetooth-Tastatur konnte der Versuchsleiter die unterschiedlichen Versuchsbedingungen über die Tasten F1 bis F6 konfigurieren. Als nächstes wurde der erste Versuchsdurchlauf unter Zeitmessung durchgeführt, währenddessen die physiologischen Parameter via der EmpaticaE4-Uhr aufgezeichnet wurden. Sobald die Probanden rückmeldeten, dass sie den Tisch mittels der BB erfolgreich abgebildet haben, wurde die Zeitmessung gestoppt, und die virtuellen Dimensionen der BB wurden dem Versuchsleiter mitgeteilt. Im Anschluss wurden die Probanden gebeten, die Fragebögen (UTAUT2 und NASA-TLX) auszufüllen. Der Versuchsleiter bereitete die immersive Szene für den nächsten Durchlauf entsprechend dem gegenläufigen Fehlerfall im Vergleich zum ersten Durchlauf vor. Danach führte der Proband analog zum ersten Durchlauf den Versuch durch. Es folgte die finale Bearbeitung des Fragebogens inkl. der subjektiven Fehlerbeurteilung durch den Probanden.

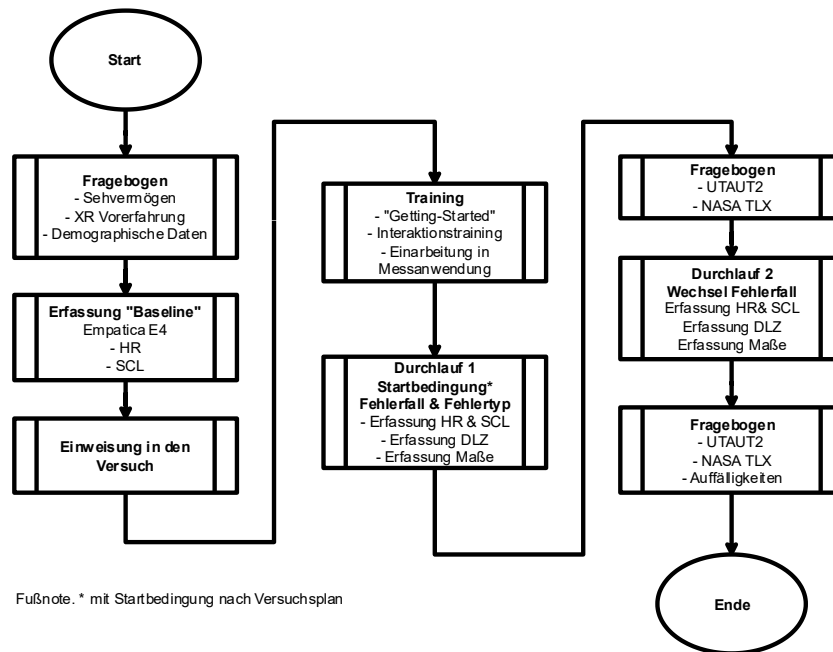


Abbildung 32. Versuchsablaufplan (ED) – S4

Versuchsdurchführung und Stichprobenumfangsplanung

Stichprobenumfangsplanung: Aufgrund des Fehlens geeigneter Daten aus vorherigen Studien zur Schätzung des Stichprobenumfangs für die Effekte der primären AVs wurde eine A-priori-Berechnung des Stichprobenumfangs basierend auf den Ergebnissen einer explorativen Vorstudie (KW 43 2023 mit $N = 12$ Teilnehmenden) mithilfe der Software G*Power Version 3.1.9.7 durchgeführt. Die Stichprobenumfangsplanung wurde primär auf die Ergebnisse der Teilstudie 4.2 ausgerichtet (siehe Abbildung 36).

Statistische Parameter	Leistung_Diff	BI_Diff	Mean_NASA_Diff
η^2p	0,155	0,513	0,136
Power	0,80	0,80	0,80
α	0,05	0,05	0,05
Anzahl der Gruppen	3	3	3
kalkulierter Stichprobenumfang	54	15	63

Tabelle 36. Stichprobenumfangsplanung – S4

Die Studie wurde von KW 43 bis KW 49 2023 durchgeführt. Alle Datensätze wurden anonymisiert erfasst. Jeder Teilnehmer wurde über die Richtlinien ethischer Forschung informiert und musste seine ausdrückliche Zustimmung zur Teilnahme geben (vgl. Abbildung 77). Die Dauer des Experiments betrug ca. 55 Minuten. Die Geräte wurden nach jedem Teilnehmer vollständig desinfiziert.

Probandengelder als Incentivierung: Zu Beginn der Studie wurden die Probanden mittels eines kompetitiven Studiencharakters inkl. eines finanziellen Anreizes extrinsisch motiviert, sich bei der Studiendurchführung ernsthaft zu bemühen (Bershadsky et al., 2022). Sie wurden gebeten, die immersive Vermessung des realen Tisches so schnell und gleichzeitig richtig wie möglich durchzuführen. Die Studienleistung der Probanden würde am Ende gegen die Ergebnisse der Kohorte gemessen werden. Sollten die Probanden mit ihren Leistungen im oberen Quartil im Vergleich zur Kohorte liegen, würden sie zwanzig Euro als Gratifikation erhalten. Selbstverständlich haben alle Teilnehmer zwanzig Euro als Probandengeld erhalten. Für diese Studie wurden diese Finanzmittel vom Bayerischen Wissenschaftsforum (BayWISS) dankenderweise bereitgestellt. Organisatorisch wurde die Auszahlung von der Technischen Hochschule Würzburg-Schweinfurt rechtskonform verwaltet und ausgehändigt.

Statistisches Vorgehen

Ergänzend zu den Ausführungen in Kapitel 3.2 wurde für die drei Teilstudien jeweils das folgende statistische Verfahren angewendet:

Teilstudie 4.1: Das Ziel ist es, an einem Forschungsprototypen unterschiedliche Fehlertypen sowie den Fehlerfall (fh versus ff) zu untersuchen, wobei jede Faktorstufe der UV Fehlertyp einzeln bezüglich ihrer UV Fehlerfall betrachtet werden soll. Ungeachtet der Voraussetzung der Normalverteilung bietet sich hierfür eine ANOVA mit Wiederholmessungen mit der Variablen des Fehlerfalls als Within-Faktor und der Fehlertypen im Between-Subjekt-Design an. Generell gilt die Repeated Measures ANOVA als sehr robust gegenüber einer Verletzung der Normalverteilungsvoraussetzung (Blanca et al., 2023). Ebenfalls können hieran die Alphafehlerkorrekturen probat angewendet werden. Neben der Voraussetzung der Normalverteilung werden zusätzlich die Varianzhomogenität und die Sphärität geprüft. Entsprechend der Homogenität der Varianzen werden die spezifischen Post-hoc-Tests nach Fischer bzw. Welch zur Kompensation bei Verletzung angewendet. Die Erfüllung der Sphärität stellt keine notwendige Bedingung dar, da es sich hierbei um eine Repeated Measures ANOVA mit lediglich zwei Faktorstufen handelt (Janczyk & Pfister, 2023).

Teilstudie 4.2: Das Ziel ist es, an einem Forschungsprototypen unterschiedliche Fehlertypen zu untersuchen, wobei die Störeinflüsse der menschlichen Varianzen aufgrund bekannter Moderatoren (vgl. UTAUT2) wie z. B. Vorerfahrungen, Alter und Technologieaffinität egalisiert und die Fehlereinflüsse isoliert werden sollten. Aus diesem Grund wurden die abhängigen Variablen der beiden Durchläufe je Proband subtrahiert und die Differenzen via ANOVA im Between-Subjekt-Design hinsichtlich der UV Fehlertypen untersucht. Die relevanten Hypothesenentscheidungen werden zusätzlich im Falle einer Verletzung der Normalverteilungsvoraussetzung durch den nichtparametrischen Kruskal-Wallis-Test validiert (Davis, 2023, S. 64). Außerdem wird im Falle einer fehlenden Normalverteilung der Dwass-Steel-Critchlow-Fligner (DSCF) bei dem Post-hoc-Test angewendet (Davis, 2023, S. 64; Lüpsen, 2019, S. 59).

Teilstudie 4.3: Das Ziel ist es, den Einfluss der Konfidenzanzeige auf die Akzeptanz hinsichtlich des stärksten Fehlereinflusses der Studie zu untersuchen. Hierzu werden analog zu Teilstudie 4.2 die Differenzen der beiden Durchläufe je Proband herangezogen. Mittels des unabhängigen *t*-Tests (Student) unter Berücksichtigung der Voraussetzungen der Normalverteilung und Varianzhomogenität wurden die Unterschiede der AV BI_Diff durch die Moderatorvariable **M** (Systemkonfidenzanzeige) untersucht.

4.4.2 Ergebnisse

Demografische Daten

Für die Hauptstudie wurden insgesamt 70 Probanden sowohl an der Universität Regensburg als auch an der Technischen Hochschule Würzburg-Schweinfurt über einen Studentenmailverteiler rekrutiert. $N = 64$ valide Datensätze konnten zur Auswertung herangezogen werden, bei den übrigen sechs Datensätzen sind irreversible, ablauftechnische Störungen aufgetreten, sodass diese Daten verworfen wurden. Von den $N = 64$ Teilnehmenden waren 47 männlich und 17 weiblich. 77,65 % der Teilnehmenden waren Studierende und 22,35 % berufstätig (überwiegend im wissenschaftlichen Umfeld). Die Teilnehmenden waren zwischen 18 und 37 Jahre alt, ihr Durchschnittsalter lag bei 23,5 Jahren. Die Gesamtkohorte der Probanden wurde in drei Between-Faktor-Bedingungen (Faktorstufen) der UV Fehlertyp unterteilt: Sichtbarkeitsfehler ($n = 19$), Datenfehler ($n = 21$) und Interaktionsfehler ($n = 18$, ohne Manipulation). Zusätzlich wurde sechs

Probanden beim Fehlertyp Interaktionsfehler die Systemkonfidenz als Moderatorvariable angezeigt.

Deskriptive Analyse

Zur detaillierten Vergleichbarkeit der umfangreichen Ergebnisse dieser Verbundstudie werden die deskriptiven Daten wie folgt dargestellt. In Tabelle 37 werden die Leistungsdaten entsprechend der AVs Richtigkeit (siehe Abbildung 33), DLZ (siehe Abbildung 34) und des Index Leistung (Abbildung 35) abgebildet.

AV/Fehlerfall	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
DLZ_Fehler [s]	Datenfehler	21	405,91	201,25	43,92
	Sichtbarkeitsfehler	19	343,68	212,79	48,82
	Interaktionsfehler	18	468,28	196,60	46,34
DLZ_ohne [s]	Datenfehler	21	246,14	131,59	28,72
	Sichtbarkeitsfehler	19	227,79	119,15	27,34
	Interaktionsfehler	18	259,17	112,37	26,49
Richtigkeit_Fehler	Datenfehler	21	0,95	0,07	0,02
	Sichtbarkeitsfehler	19	0,95	0,09	0,02
	Interaktionsfehler	18	0,97	0,04	0,01
Richtigkeit_ohne	Datenfehler	21	0,96	0,04	0,01
	Sichtbarkeitsfehler	19	0,96	0,06	0,01
	Interaktionsfehler	18	0,98	0,02	0,01
Leistung_Fehler [%/(s*10)]	Datenfehler	21	3,10	2,00	0,44
	Sichtbarkeitsfehler	19	3,88	2,34	0,54
	Interaktionsfehler	18	2,38	0,85	0,20
Leistung_ohne [%/(s*10)]	Datenfehler	21	5,07	3,08	0,67
	Sichtbarkeitsfehler	19	5,41	2,80	0,64
	Interaktionsfehler	18	4,52	2,04	0,48

Tabelle 37. DD Leistungs-AV versus Fehlerfall & Fehlertyp – S4

Speziell auf die Stichproben bezogen fällt auf, dass der Mittelwert der Richtigkeit beim Datenfehler im Fehlerfall (siehe Abbildung 33) am niedrigsten ist, gefolgt von dem Sichtbarkeitsfehler, und für den Interaktionsfehler am höchsten ist (vgl. Tabelle 37). Die Ergebnisse zeigen weiter, dass für den Sichtbarkeitsfehler die Mittelwertdifferenz zwischen den Fehlerfällen am kleinsten ausfällt.

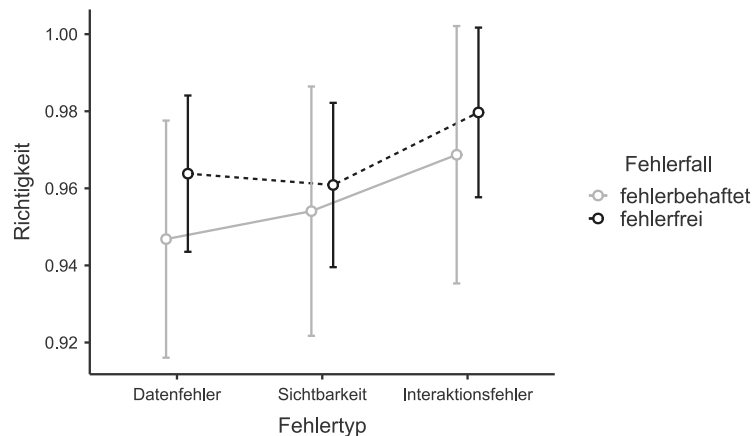


Abbildung 33. Richtigkeit versus Fehlerfall & Fehlertyp – S4

Die Stichproben (vgl. Tabelle 37 und Abbildung 34) zeigen hinsichtlich der AV DLZ, dass die Mittelwerte für den Fehlertyp Sichtbarkeit kürzer gemessen wurden als für den Datenfehler bzw. den Interaktionsfehler. Ebenfalls ergibt sich für den Sichtbarkeitsfehler eine geringere Mittelwertdifferenz zwischen den Fehlerfällen. Für den Interaktionsfehler fällt sowohl der Mittelwert als auch die Mittelwertdifferenz zwischen den Fehlerfällen am höchsten aus.

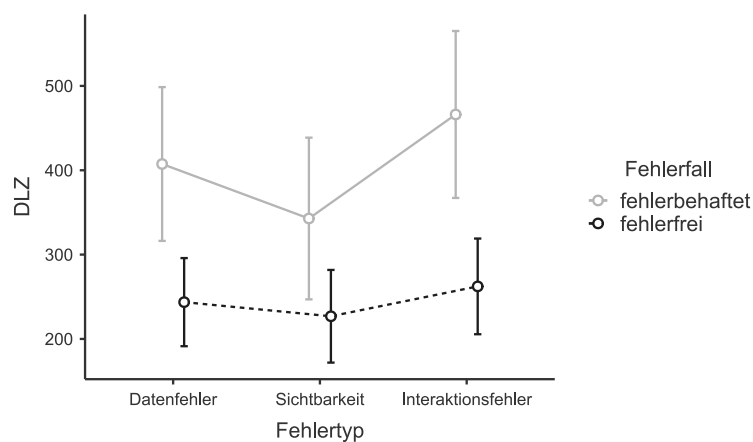


Abbildung 34. DLZ versus Fehlerfall & Fehlertyp – S4

Die Stichproben (vgl. Tabelle 37 und Abbildung 35) zeigen für die Mittelwerte des Index Leistung, dass sie für den Fehlertyp Sichtbarkeitsfehler höher liegen als bei den anderen beiden und für den Interaktionsfehler am niedrigsten sind.

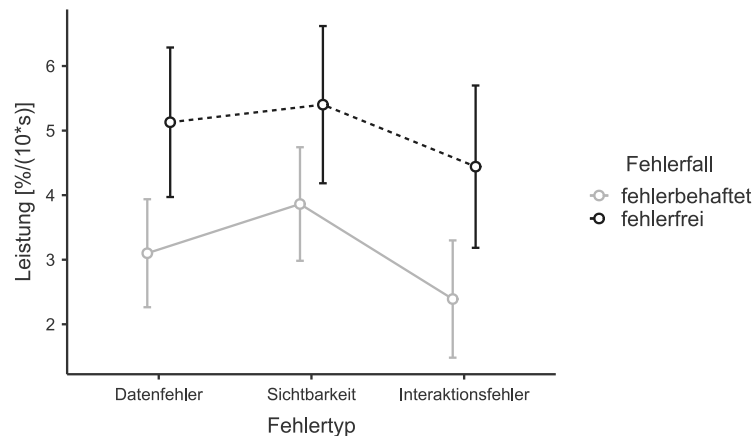


Abbildung 35. Index Leistung versus Fehlerfall & Fehlertyp – S4

Insgesamt zeigt sich, dass die Mittelwertdifferenzen für die AV Richtigkeit ($MD = 0,01$; $Max = 1$) durch die UV Fehlerfall über alle Fehlertypen hinweg am kleinsten ausfallen im Gegensatz zu AV DLZ ($MD = 161s$; $Max = 930s$) sowie AV Index Leistung ($MD = 1,88 \text{ } \%/ (s \cdot 10)$; $Max = 14,4 \text{ } \%/ (s \cdot 10)$).

Abbildung 36 zeigt die Mittelwerte der BI (vgl. Tabelle 38). Weitere Akzeptanzfaktoren des UTAUT2 sind der Tabelle 78 in Anhang 6.1 zu entnehmen.

AV/Fehlerfall	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
BI_Fehler	Interaktionsfehler	18	3,17	1,48	0,35
	Stichtbarkeitsfehler	19	4,40	1,55	0,36
	Datenfehler	21	3,45	1,22	0,27
BI_ohne	Interaktionsfehler	18	3,56	1,70	0,40
	Stichtbarkeitsfehler	19	4,12	1,75	0,40
	Datenfehler	21	3,91	1,27	0,28

Tabelle 38. DD BI versus Fehlerfall & Fehlertyp – S4

Die Stichproben zeigen, dass beim Sichtbarkeitsfehler die BI-Mittelwerte der Faktorstufe fh einen höheren Skalenwert als die der Faktorstufe ff aufweisen. Im Gegensatz dazu verhalten sich die Mittelwerte der BI bei den anderen beiden Fehlertypen umgekehrt. Für den Interaktionsfehler fällt der Mittelwert der Faktorstufe fh am niedrigsten aus.

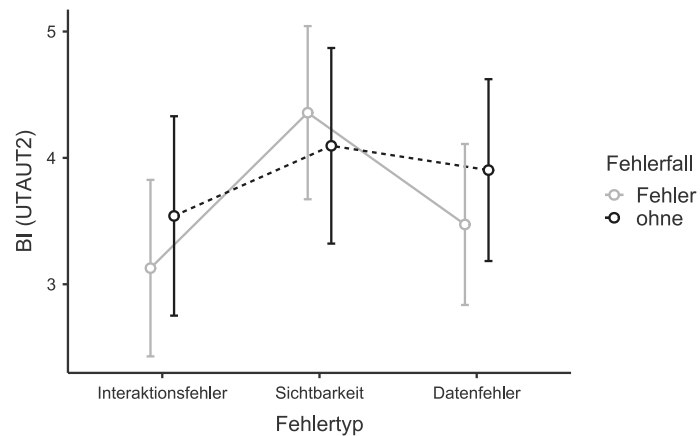


Abbildung 36. BI versus Fehlerfall & Fehlertyp – S4

Die durchschnittliche Gesamtbeanspruchung des NASA-TLX wird in Abbildung 37 sowie in Tabelle 39 dargestellt. Die deskriptiven Daten der übrigen Dimensionen des NASA-TLX können entsprechend der Tabelle 79 (Anhang 6.1) entnommen werden.

AV/Fehlerfall	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
Mean_NASA_Fehler	Interaktionsfehler	18	56,10	17,30	4,08
	Sichtbarkeitsfehler	19	45,70	18,50	4,25
	Datenfehler	21	46,80	20,50	4,48
Mean_NASA_ohne	Interaktionsfehler	18	42,70	17,30	4,08
	Sichtbarkeitsfehler	19	47,80	15,60	3,58
	Datenfehler	21	47,40	22,40	4,88

Tabelle 39. DD Mean_NASA versus Fehlerfall & Fehlertyp – S4

Der Mittelwert der durchschnittlichen Beanspruchung liegt beim Interaktionsfehler im Fehlerfall um 10,4 Skalenwerte höher als beim Sichtbarkeitsfehler bzw. um 9,3 Skalenwerte höher als beim Datenfehler. Ebenfalls ist die Mittelwertdifferenz der durchschnittlichen Beanspruchung zwischen den beiden Fehlerfällen beim Interaktionsfehler um Faktor 6,3 größer als bei den anderen beiden Fehlertypen.

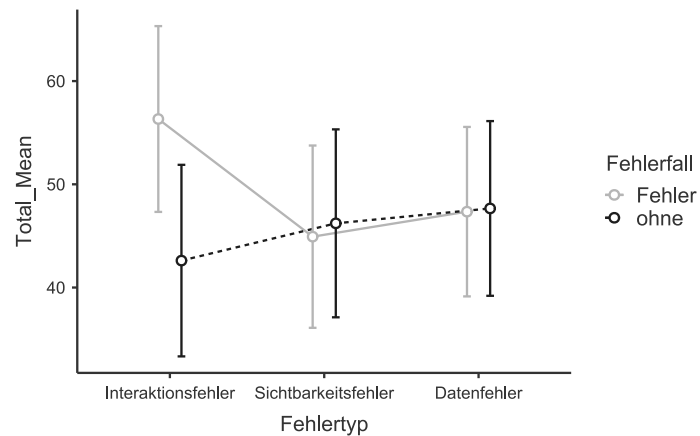


Abbildung 37. Mean_NASA versus Fehlerfall & Fehlertyp – S4

Abbildung 38 und Abbildung 39 zeigen die physiologischen Parameter HR und SCL (siehe Tabelle 40).

Physiologische Parameter	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
HR_Diff_Fehler	Datenfehler	21	10,09	11,01	2,40
	Interaktionsfehler	19	6,08	12,23	2,88
	Sichtbarkeitsfehler	18	8,34	12,58	2,96
HR_Diff_ohne	Datenfehler	21	10,16	11,92	2,60
	Interaktionsfehler	19	4,61	15,24	3,59
	Sichtbarkeitsfehler	18	10,39	15,03	3,54
SCL_Diff_Fehler	Datenfehler	21	1,23	2,34	0,51
	Interaktionsfehler	19	1,72	2,84	0,67
	Sichtbarkeitsfehler	18	1,47	2,60	0,61
SCL_Diff_ohne	Datenfehler	21	1,18	2,26	0,49
	Interaktionsfehler	19	1,53	2,26	0,53
	Sichtbarkeitsfehler	18	1,25	2,00	0,47

Tabelle 40. DD HR [BPM] & SCL [μ S] versus Fehlerfall & Fehlertyp – S4

In Bezug auf die Mittelwerte der HR-Differenz zeigt sich, dass diese im Fehlerfall beim Interaktionsfehler tendenziell am niedrigsten ausfallen, während sie beim Datenfehler am höchsten sind. Auch beim Sichtbarkeitsfehler liegen die Mittelwerte im Fehlerfall unter denen der fehlerfreien Bedingung, jedoch über denen des Interaktionsfehlers (vgl. Tabelle 40).

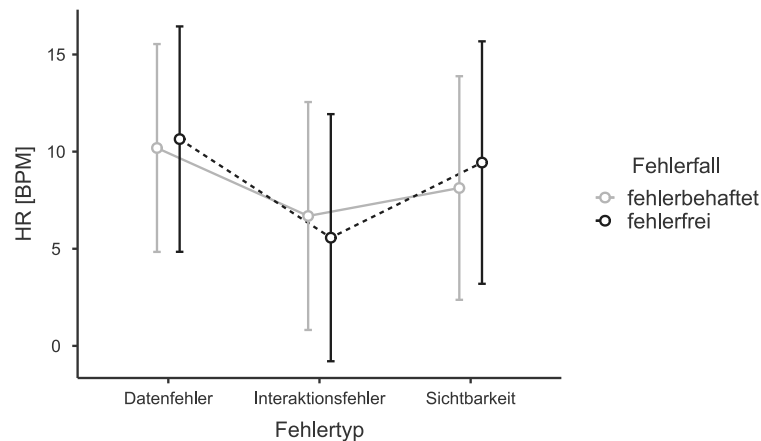


Abbildung 38. HR-Differenz versus Fehlerfall & Fehlertyp – S4

Bei den Mittelwerten der SCL-Differenz verhält es sich genau gegenläufig zur HR-Differenz.

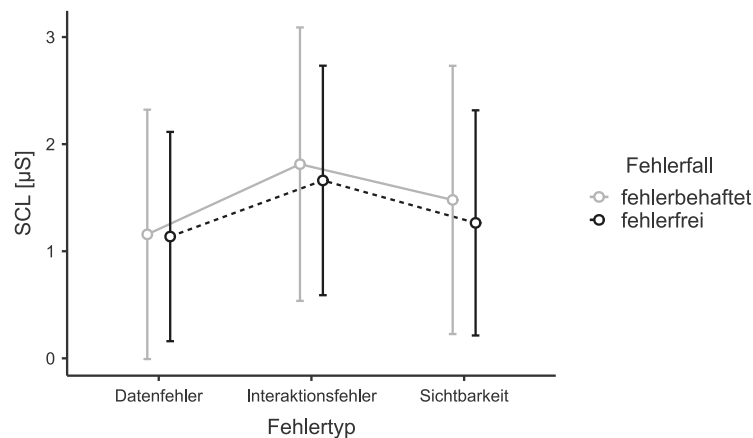


Abbildung 39. SCL-Differenz versus Fehlerfall & Fehlertyp – S4

Diese Abbildungen und Beschreibungen sollen einen ersten, qualitativen Gesamtüberblick über die Ergebnisse der Verbundstudie geben und werden im Folgenden entsprechend den Hypothesen und Forschungsfragen weiter konsekutiv ausgewertet. Zum besseren Verständnis werden die Ergebnisse der Teilstudie 4.3 zusammenhängend inkl. der deskriptiven Daten in Abschnitt 4.4.2.3 präsentiert.

Subjektive Fehlerwahrnehmung

Die Probanden wurden nach beiden Durchläufen entsprechend dem Fragebogen gemäß Tabelle 62 befragt, welche Qualitätsverluste sie bei welchem der beiden Durchläufe wahrgenommen haben. Hierbei war eine Mehrfachauswahl möglich. Die prozentuale Verteilung der Ergebnisse ist der nachfolgenden Tabelle 41 zu entnehmen. Bei den

Ergebnissen wurde differenziert zwischen der UV Fehlertyp sowie der eindeutig wahr, wahr-positiven (wp) und falsch-positiven (fp) Fehlerwahrnehmungen.

wahrgenommen/ wahr	eindeutig wahr	Sichtbarkeitsfehler	Interaktionsfehler	Datenfehler
Datenfehler	52,4 %	14,3 % (fp)	38,1 % (fp)	81,0 % (wp)
Interaktionsfehler	55,6 %	11,1 % (fp)	77,8 % (wp)	22,2 % (fp)
Sichtbarkeitsfehler	52,6 %	73,7 % (wp)	36,8 % (fp)	10,5 % (fp)
Indikator (Interakt.)	83,3 %	0,0 % (fp)	100,0 % (wp)	16,7 % (fp)

Tabelle 41. Verteilung der Fehlerwahrnehmung – S4

Analyse der Fehlerreihenfolge als Störeinfluss

Das Forschungsdesign wurde so ausgelegt, dass die potenziellen Störeinflüsse der Reihenfolge durch ein paritätisch, gegenbalanciertes DOE minimiert wurden. Flankierend wurde der Störeffekt der RF mittels einer ANOVA mit Wiederholmessungen für alle AVs ausgewertet, wobei die RF und der Fehlertyp als Between-Subject-Faktoren und der Fehlerfall (fh und ff) als Within-Subject-Faktor eingingen. Hinsichtlich der RF wurde außer für die physische Stressreaktion HR bei keiner weiteren AV ein signifikanter Unterschied (siehe Abbildung 40) nachgewiesen, auch nicht für die Wechselwirkungseffekte Fehlertyp x Reihenfolge. Tabelle 42 zeigt die Ergebnisse der Between-Subject-Effekte der relevanten AVs auf.

AV	Between Subject	F	p	η^2p
DLZ	RF	0,301	0,586	0,006
DLZ	FT * RF	0,033	0,586	0,001
Richtigkeit	RF	0,234	0,631	0,004
Richtigkeit	FT * RF	0,314	0,732	0,012
Leistung	RF	0,951	0,334	0,018
Leistung	FT * RF	0,811	0,450	0,030
BI	RF	0,332	0,567	0,006
BI	FT * RF	0,495	0,613	0,019
MD	RF	0,094	0,760	0,002
MD	FT * RF	0,574	0,567	0,022
F	RF	0,107	0,745	0,002
F	FT * RF	0,049	0,953	0,002
Mean_NASA	RF	0,112	0,739	0,002
Mean_NASA	FT * RF	0,329	0,721	0,013
PE	RF	0,293	0,591	0,006
PE	FT * RF	1,229	0,301	0,046
HR	RF	5,718	0,021	0,101
HR	FT * RF	0,399	0,673	0,015
SCL	RF	0,001	0,972	0,000
SCL	FT * RF	0,732	0,732	0,028

Tabelle 42. Haupt- & Wechselwirkungseffekt RF & FT * RF – S4

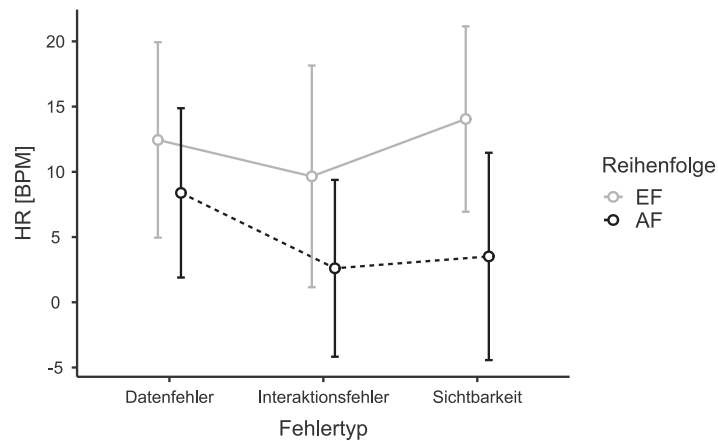


Abbildung 40. HR versus Reihenfolge – S4

Die Datensätze der Studie 4 (Teilstudien 4.1–4.3) sind in Tabelle 74 bis Tabelle 77 in Anhang A6.1 bereitgestellt.

4.4.2.1 Teilstudie 4.1 Fehlerauswirkungen

Datenfehler

Leistungsfaktoren: Die Post-hoc-Tests Datenfehler * Fehlertyp der Repeated Measures ANOVAs zeigen im Gegensatz zur Richtigkeit $t(52) = 1,570$; $MD = 0,017$; $p_{\text{Tukey}} = 0,622$; $Co'd = 0,277$; Power = 0,34 nur für die AV DLZ $t(52) = 4,56$; $MD = 163,8$ s; $p_{\text{Tukey}} < 0,001$; $Co'd = 0,831$; Power = 0,98 und den Index Leistung $t(52) = 4,026$; $MD = 2,029$ %/(10*s); $p_{\text{Tukey}} = 0,002$; $Co'd = 0,686$; Power = 0,92 einen signifikanten Unterschied zwischen den Faktorstufen fh und ff hinsichtlich der UV DQ. Für alle drei Variablen ist die Voraussetzung der Varianzhomogenität erfüllt. Die zugehörigen Mittelwerte sind grafisch in Abbildung 33 bis Abbildung 35 einsehbar.

UTAUT2: Bis auf BI zeigten die Post-hoc-Tests der Subdimensionen des UTAUT2-Fragebogens keinen statistisch signifikanten Unterschied zwischen den beiden Faktorstufen der DQ. Die Ergebnisse wurden in Tabelle 43 dargestellt (vgl. Abbildung 41). Die Varianzhomogenität war lediglich bei SI nicht gegeben.

Dimension	MD	t	df	p_{Tukey}	Co'd	Power
UTAUT2_PE	0,46	2,040	51	0,335	0,480	0,68
UTAUT2_EE	0,24	1,219	51	0,825	0,378	0,51
UTAUT2_SI	0,26	1,315	51	0,775	0,195	0,21
UTAUT2_FC	0,15	0,823	51	0,914	0,165	0,18
UTAUT2_HM	0,11	0,628	51	0,988	0,279	0,34
UTAUT2_HT	0,34	2,030	51	0,340	0,408	0,56

Dimension	<i>MD</i>	<i>t</i>	<i>df</i>	<i>p</i> _{Tukey}	<i>Co'd</i>	Power
UTAUT2_BI	0,43	2,238	51	0,030	0,527	0,75

Tabelle 43. UTAUT2 PHT versus DQ – S4.1

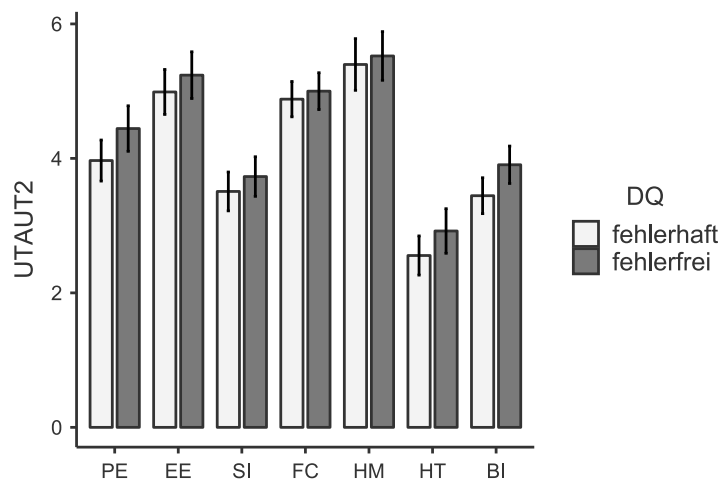


Abbildung 41. Barplots UTAUT versus DQ – S4.1

NASA-TLX: Im verbundenen Design zeigten beim Post-hoc-Test weder die durchschnittliche Beanspruchung noch die einzelnen Subskalen des NASA-TLX signifikante Unterschiede zwischen den Faktorstufen der DQ fh und ff (siehe Tabelle 44 und Abbildung 42). Die Voraussetzungen der Varianzhomogenität sind erfüllt.

Dimension	<i>MD</i>	<i>t</i>	<i>df</i>	<i>p</i> _{Tukey}	<i>Co'd</i>	Power
NASA_MD	-5,49	-0,927	51	0,938	-0,125	0,14
NASA_PD	2,29	0,408	51	0,998	0,015	0,05
NASA_TD	-4,31	-0,731	51	0,977	-0,167	0,18
NASA_P	10,07	1,712	51	0,531	0,257	0,31
NASA_E	-10,00	-1,588	51	0,610	-0,307	0,39
NASA_F	5,56	0,764	51	0,972	0,160	0,17
Mean_NASA	-0,31	-0,072	51	1,000	-0,026	0,06

Tabelle 44. NASA-TLX PHT versus DQ – S4.1

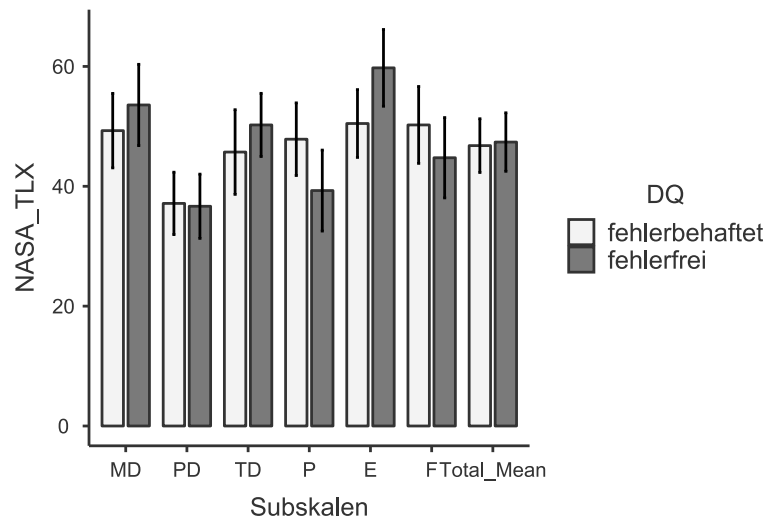


Abbildung 42. Barplots NASA-TLX versus DQ – S4.1

Physiologische Parameter: Es konnte kein statistisch signifikanter Einfluss der DQ auf die physiologischen Reaktionen via des Post-hoc-Tests auf die AVs SCL (siehe Abbildung 43 b) oder HR (siehe Abbildung 43 a) bestätigt werden (vgl. Tabelle 45). Die Voraussetzung der Varianzhomogenität wurde sowohl für die HR_Diff als auch für die SCL_Diff erfüllt.

Parameter	MD	t	df	p_{Tukey}	Co'd	Power
HR_Diff	-0,46	-0,179	51	1,000	-0,012	0,06
SCL_Diff	0,02	0,073	51	1,000	0,047	0,08

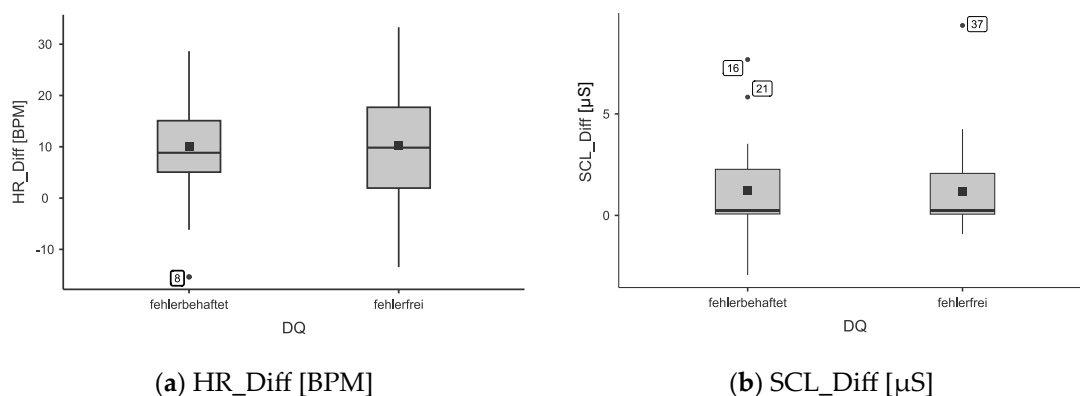
Tabelle 45. HR [BPM] & SCL [μ S] PHT versus DQ – S4.1

Abbildung 43. Boxplots der (a) HR_Diff & (b) SCL_Diff versus DQ – S4.1

Interaktionsfehler II

Leistungsfaktoren: Die Post-hoc-Tests Interaktionsfehler * Fehlertyp der Repeated Measures ANOVAs zeigen im Gegensatz zur Richtigkeit $t(52) = 0,935$; $MD = 0,011$; $p_{\text{Tukey}} = 0,936$; $Co'd = 0,322$; Power 0,36 nur für die AV DLZ $t(52) = 5,225$; $MD = 203,9s$; $p_{\text{Tukey}} <$

0,001; $Co'd = 1,23$; Power = 1 und den Index Leistung $t(52) = 3,748$; $MD = 2,05 \text{ \%}/(10*s)$; $p_{\text{Tukey}} = 0,006$; $Co'd = 1,08$; Power = 1 einen signifikanten Unterschied zwischen den Faktorstufen fh und ff hinsichtlich der UV IQ. Für alle drei Variablen ist die Voraussetzung der Varianzhomogenität nicht verletzt. Die zugehörigen Mittelwerte sind grafisch in Abbildung 33 bis Abbildung 35 einsehbar.

UTAUT2: Bis auf BI zeigten alle Subdimensionen des UTAUT2-Fragebogens keinen statistisch signifikanten Unterschied zwischen den beiden Faktorstufen der IQ. Die Ergebnisse der Post-hoc-Tests innerhalb der UV IQ zwischen den beiden Faktorstufen ff und fh wurden in Tabelle 46 dargestellt (vgl. Abbildung 44). Die Voraussetzung der Varianzhomogenität ist nur für SI und HM verletzt.

Dimension	MD	t	df	p_{Tukey}	Co'd	Power
UTAUT2_PE	0,49	1,981	51	0,255	0,244	0,29
UTAUT2_EE	0,25	1,151	51	0,857	0,203	0,21
UTAUT2_SI	-0,16	-0,721	51	0,978	-0,215	0,22
UTAUT2_FC	0,01	0,063	51	1,000	0,000	0,05
UTAUT2_HM	0,08	0,405	51	0,999	0,183	0,18
UTAUT2_HT	0,09	0,505	51	0,996	0,248	0,26
UTAUT2_BI	0,41	2,240	51	0,045	0,529	0,69

Tabelle 46. UTAUT2 PHT gegen UV IQ – S4.1

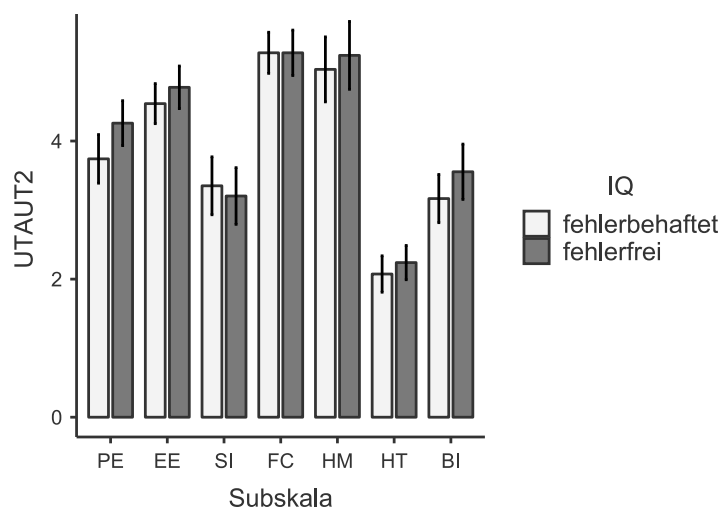


Abbildung 44. Barplots UTAUT2 versus UV IQ – S4.1

NASA-TLX: Im verbundenen Design zeigten die durchschnittliche Beanspruchung sowie MD, P und F des NASA-TLX signifikante Unterschiede zwischen den Faktorstufen der UV IQ fh und ff (siehe Tabelle 47). Die Tabelle 79 in Anhang 6.1 zeigt die Ergebnisse der deskriptiven Analyse über alle Subskalen. Außer bei TD und P waren die Voraussetzungen der Varianzhomogenität erfüllt.

Dimension	MD	t	df	p _{Tukey}	Co'd	Power
NASA_MD	19,94	3,073	51	0,038	0,926	0,98
NASA_PD	6,69	1,087	51	0,382	0,094	0,10
NASA_TD	6,46	1,003	51	0,422	0,243	0,25
NASA_P	13,80	2,140	51	0,047	0,703	0,88
NASA_E	10,16	1,472	51	0,247	0,321	0,36
NASA_F	25,26	3,168	51	0,029	0,823	0,95
Mean_NASA	13,72	2,875	51	0,042	0,770	0,93

Tabelle 47. NASA-TLX versus UV IQ – S4.1

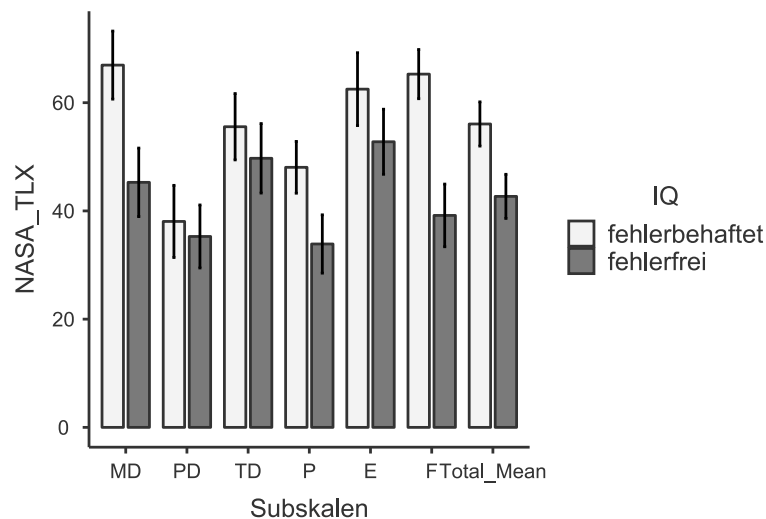


Abbildung 45. Barplots NASA-TLX versus IQ – S4.1

Physiologische Parameter: Es konnte kein statistisch signifikanter Einfluss der IQ auf die physiologischen Reaktionen via des Post-hoc-Tests der AVs SCL (siehe Abbildung 46 b) oder HR (siehe Abbildung 46 a) bestätigt werden (vgl. Tabelle 48). Die Voraussetzung der Varianzhomogenität wurde für beide physiologischen AVs erfüllt.

Parameter	MD	t	df	p _{Tukey}	Co'd	Power
HR_Diff	-1,12	-0,399	51	0,999	-0,144	0,14
SCL_Diff	0,15	0,500	51	0,996	0,136	0,13

Tabelle 48. HR [BPM] und SCL [μS] PHT versus IQ – S4.1

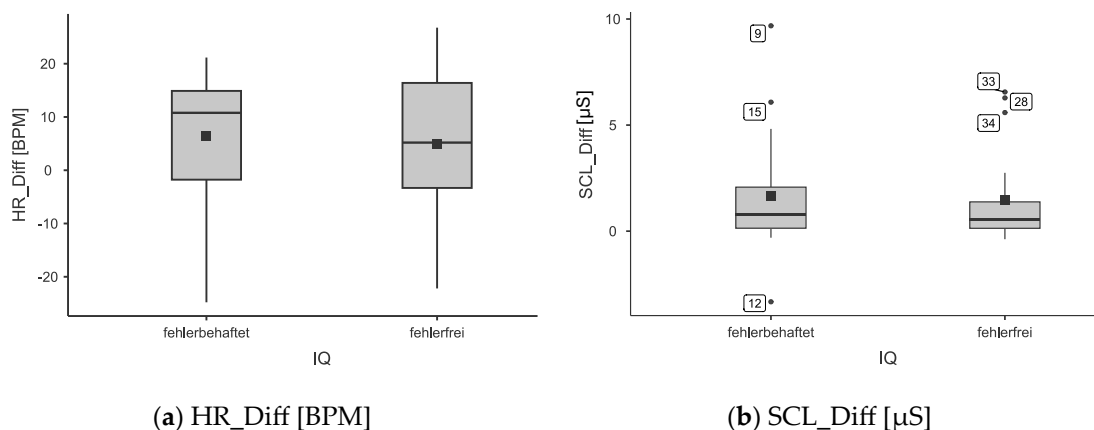


Abbildung 46. Boxplots (a) HR_Diff & (b) SCL_Diff versus IQ – S4.1

Sichtbarkeitsfehler II

Leistungsfaktoren: Die Post-hoc-Tests Sichtbarkeitsfehler * Fehlertyp der Repeated Measures ANOVAs zeigen im Gegensatz zur Richtigkeit $t(52) = 1,570$; $MD = 0,0170$; $p_{\text{Tukey}} = 0,622$; $Co'd = 0,150$; Power = 0,16 nur für die AV DLZ $t(52) = 4,56$; $MD = 163,8$ s; $p_{\text{Tukey}} < 0,001$; $Co'd = 0,813$; Power = 0,97 und den Index Leistung $t(52) = 4,026$; $MD = 2,029$ %/(10*s); $p_{\text{Tukey}} = 0,002$; $Co'd = 0,721$; Power = 0,94 einen signifikanten Unterschied zwischen den Faktorstufen fh und ff hinsichtlich der UV VQ. Für alle drei Variablen ist die Voraussetzung der Varianzhomogenität erfüllt. Die zugehörigen Mittelwerte sind grafisch in Abbildung 33 bis Abbildung 35 einsehbar.

UTAUT2: Keine Subdimensionen des UTAUT2-Fragebogens zeigte im Post-hoc-Test einen statistisch signifikanten Unterschied zwischen den beiden Faktorstufen der VQ. Die Ergebnisse der Post-hoc-Tests innerhalb der UV VQ wurden in Tabelle 49 dargestellt (vgl. Abbildung 47). Die Homogenität der Varianzen wurde nicht verletzt.

Dimension	MD	t	df	p_{Tukey}	Co'd	Power
UTAUT2_PE	-0,26	-1,015	51	0,872	-0,233	0,25
UTAUT2_EE	-0,16	-0,865	51	0,970	-0,198	0,21
UTAUT2_SI	0,02	0,111	51	1,000	0,026	0,06
UTAUT2_FC	0,13	0,664	51	0,982	0,152	0,16
UTAUT2_HM	-0,09	-0,570	51	0,996	-0,131	0,14
UTAUT2_HT	-0,11	-0,666	51	0,991	-0,153	0,16
UTAUT2_BI	-0,28	-1,303	51	0,710	-0,299	0,35

Tabelle 49. UTAUT2 versus VQ – S4.1

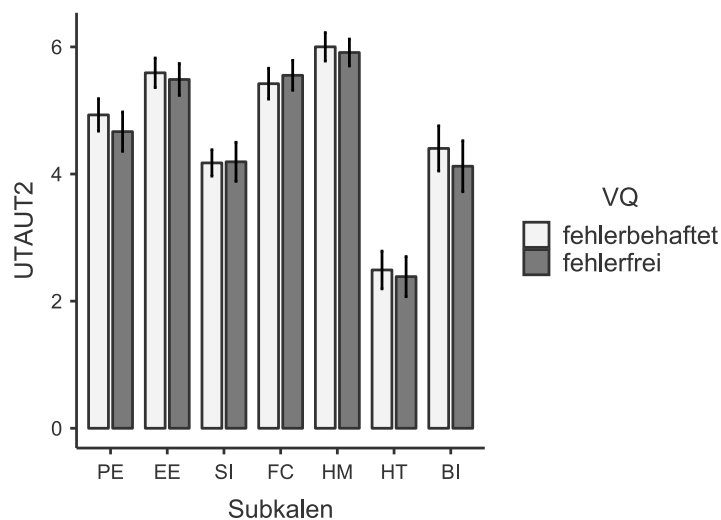


Abbildung 47. Barplots UTAUT2 versus VQ – S4.1

NASA-TLX: Im verbundenen Design zeigten die Post-hoc-Tests weder die durchschnittliche Beanspruchung noch die einzelnen Subskalen des NASA-TLX signifikante Unterschiede zwischen den Faktorstufen der VQ fh und ff (siehe Tabelle 50 und Abbildung 48). Die Voraussetzung der Varianzhomogenität war für alle Subskalen erfüllt.

Dimension	MD	t	df	p_{Tukey}	Co'd	Power
NASA_MD	8,16	1,061	52	0,640	0,244	0,27
NASA_PD	-6,05	-1,492	52	0,900	-0,342	0,41
NASA_TD	0,79	0,130	52	1,000	0,030	0,06
NASA_P	-10,00	-1,563	52	0,662	-0,359	0,44
NASA_E	1,32	0,216	52	1,000	0,050	0,08
NASA_F	-6,32	-0,839	52	0,976	-0,192	0,20
Mean_NASA	-2,02	-0,450	52	1,000	-0,103	0,11

Tabelle 50. NASA-TLX PHT versus VQ – S4.1

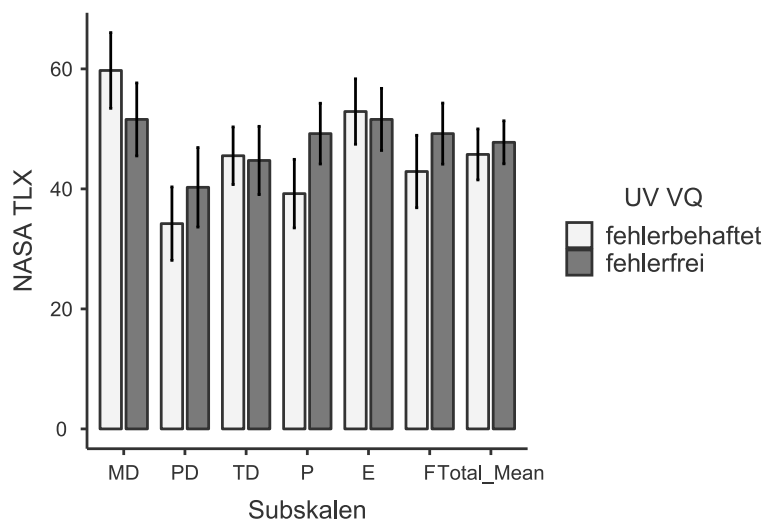


Abbildung 48. Barplots NASA-TLX versus VQ – S4.1

Physiologische Parameter: Es konnte kein statistisch signifikanter Einfluss der VQ auf die physiologischen Reaktionen via des Post-hoc-Tests der AVs SCL (siehe Abbildung 49 b) und HR (siehe Abbildung 49 a) bestätigt werden (vgl. Tabelle 51). Die Voraussetzungen für die Varianzhomogenität wurden erfüllt.

Parameter	MD	t	df	p_{Tukey}	Co'd	Power
HR_Diff	-1,31	-0,477	51	0,997	-0,120	0,12
SCL_Diff	0,21	0,719	51	0,979	0,177	0,18

Tabelle 51. HR [BPM] & SCL [μ S] PHT versus VQ – S4.1

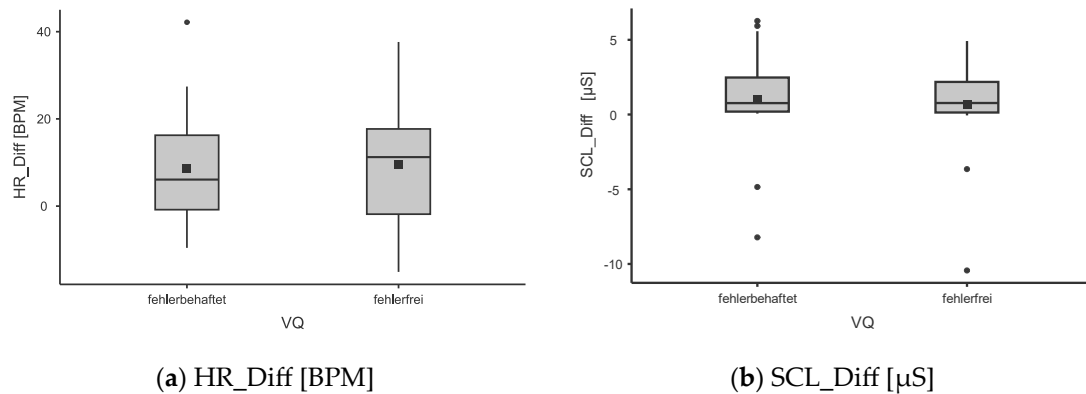


Abbildung 49. Boxplots (a) HR_Diff & (b) SCL_Diff versus VQ – S4.1

4.4.2.2 Teilstudie 4.2 Fehlervergleich

Leistungsfaktoren: Die einfaktoriellen ANOVAs (nach Fischer) zeigen für alle drei AVs Leistungsdifferenz $F(2;54) = 0,222$; $p = 0,802$; $\eta^2p = 0,008$; Power = 0,08, vgl. Abbildung 50, Richtigkeitsdifferenz $F(2;54) = 0,236$; $p = 0,791$; $\eta^2p = 0,009$; Power = 0,09, vgl. Abbildung 51 und DLZ-Differenz $F(2;54) = 1,58$; $p = 0,216$; $\eta^2p = 0,055$, Power = 0,32, vgl. Abbildung 52, dass hinsichtlich der Faktorstufen der UV Fehlertyp keine signifikanten Unterschiede nachgewiesen werden können. Im Gegensatz zur Varianzhomogenität wurden die Voraussetzungen für alle drei AVs bei der Normalverteilung verletzt. Auch wenn ANOVAS robust gegenüber nicht normalverteilten Daten sind, und diese Verletzung ebenfalls durch große Stichprobenumfänge (Blanca et al., 2017; Schmider et al., 2010) kompensiert werden könnten, werden die Aussagen nochmal anhand einer nichtparametrischen, einfaktoriellen ANOVA (Kruskal-Wallis) validiert (vgl. Tabelle 52). Die Q-Q-Plots zeigen für alle drei Leistungsfaktoren hinsichtlich ihrer Verteilung relevante Abweichungen von der Diagonalen.

AV	χ^2	p
Leistungs_Diff	1.137	0,566
Richtigkeits_Diff	0.392	0,822
DLZ_Diff	5.022	0,081

Tabelle 52. Nicht-param. ANOVA LeistungsAV versus Fehlertyp – S4.2

Deswegen wurde als Post-hoc-Test ein nichtparametrischer Dwass-Steel-Critchlow-Fligner (DSCF) konsekutiv angewendet. Aus diesem gingen keine statistisch signifikanten Unterschiede innerhalb der UV Fehlertyp hervor (siehe Tabelle 81 in Anhang 6.2). Die entsprechenden deskriptiven Werte können der Tabelle 80 in Anhang 6.2 entnommen werden. Bei den Differenzbetrachtungen der AVs zwischen den beiden

Durchlaufen je Proband ist zu beachten, dass die Subtrahenden so gewählt wurden, dass die zu erwartende Fehlerwirkung (Leistungs-/Akzeptanzverlust etc.) mit positivem Betrag gewertet wird. Jedoch gehen auch aus der deskriptiven Analyse keine nennenswerten Auffälligkeiten hervor.

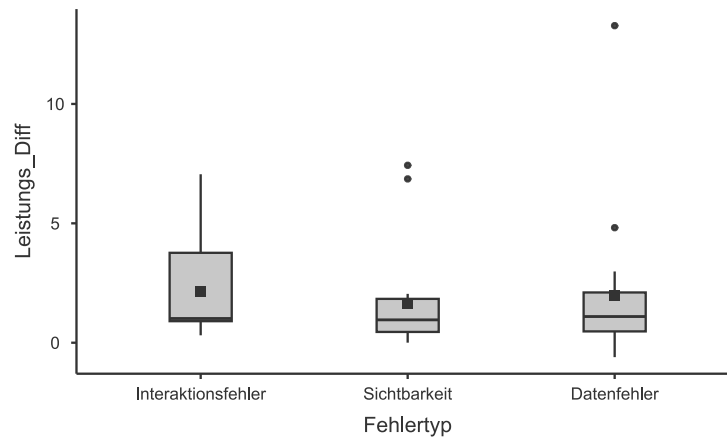


Abbildung 50: Boxplot Leistungs_Diff [%/(10*s)] versus UV Fehlertyp – S4.2

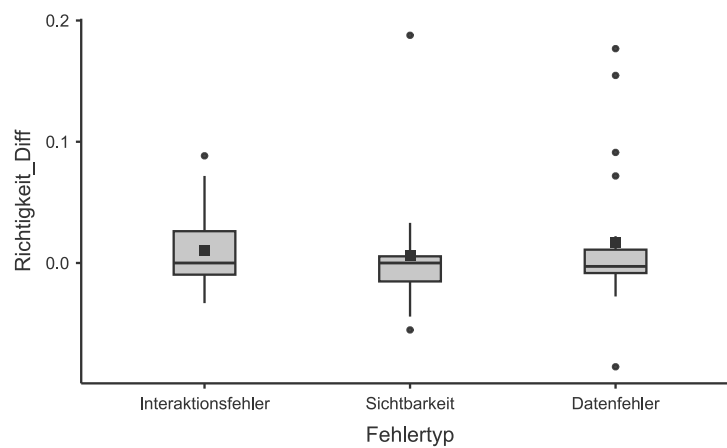


Abbildung 51. Boxplot Richtigkeit_Diff versus UV Fehlertyp – S4.2

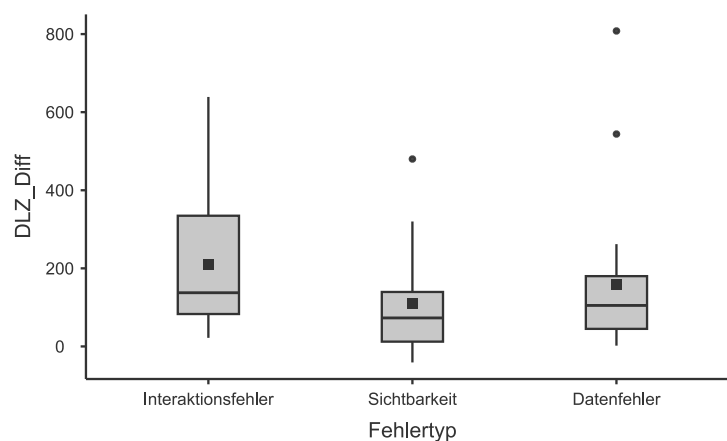


Abbildung 52. Boxplot DLZ_Diff [s] versus UV Fehlertyp S4.2

UTAUT2: Lediglich die Differenzen BI_Diff $F(2;54) = 4,14$; $p = 0,021$; $\eta^2 p = 0,133$; Power = 0,74 und PE_Diff $F(2;54) = 3,28$; $p = 0,045$; $\eta^2 p = 0,108$; Power = 0,63 der Subdimensionen des UTAUT2-Fragebogens zeigen via der ANOVA (Fischer) einen statistisch signifikanten Unterschied zwischen den Fehlertypen. Die Voraussetzungen für die Varianzhomogenität ist für alle Diskriminanten erfüllt, hingegen bei der Normalverteilung nur für EE_Diff, FC_Diff und BI_Diff. Wie in Abschnitt 3.2 beschrieben, sind die Dimensionen SI und HT nur fakultativ und werden nicht weiter berücksichtigt. Die Analyse des Q-Q-Plots der PE_Diff zeigt hingegen, dass die Verteilung der Werte im zentralen Bereich eng an der Diagonalen liegen und nur an den beiden Randbereichen abweichen. Zusammen mit der hohen Stichprobengröße ($N = 58$) kann der Post-hoc-Test ohne Einschränkung durchgeführt werden. Er zeigt für BI_Diff zwischen den Faktorstufen Sichtbarkeit- und Datenfehler ($t(54) = -2,66$; $MD = -0,737$; $p_{\text{Tukey}} = 0,027$; $Co'd = -0,854$) und Interaktion- und Sichtbarkeitsfehler ($t(54) = 2,317$; $MD = 0,666$ $p_{\text{Tukey}} = 0,049$; $Co'd = 0,772$; Power = 0,97) sowie für PE_Diff zwischen Sichtbarkeit- und Datenfehler ($t(54) = -2,21$; $MD = -0,736$; $p_{\text{Tukey}} = 0,048$; $Co'd = -0,710$; Power = 0,94) und zwischen Sichtbarkeitsfehler und Interaktionsfehler ($t(54) = 2,252$; $MD = 0,778$; $p_{\text{Tukey}} = 0,050$; $Co'd = 0,751$; Power = 0,96) einen signifikanten Unterschied. Die übrigen Ergebnisse des Post-hoc-Tests können aus Tabelle 82 in Anhang 6.2 entnommen werden. Die deskriptiven Daten werden in Tabelle 83 des Anhangs A6.2 beigefügt. Abbildung 53 präsentiert die entsprechenden Barplots des UTAUT2.

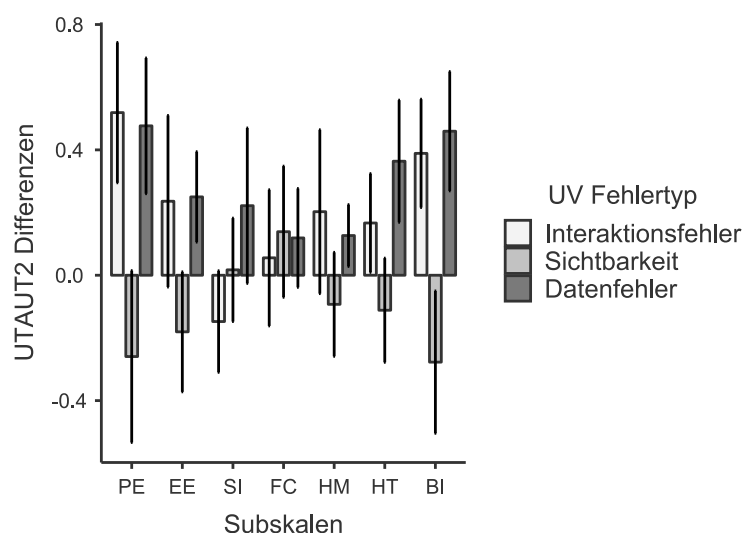
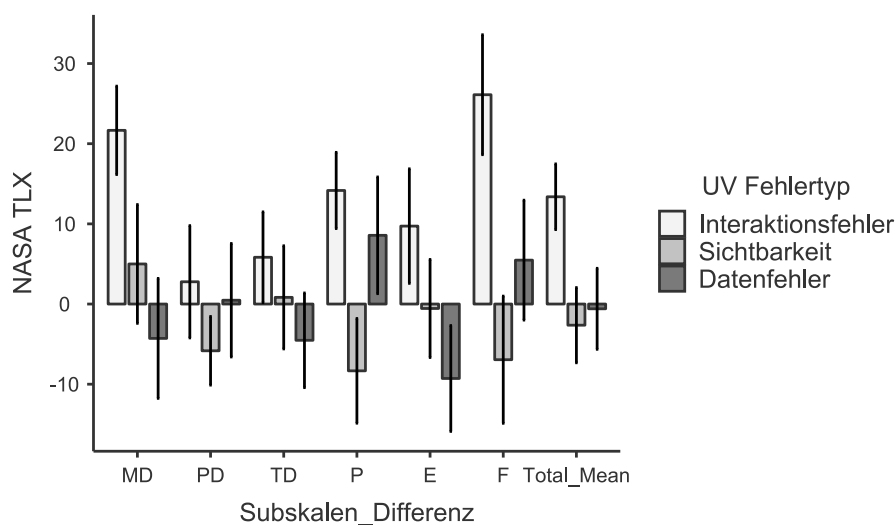


Abbildung 53. Barplots UTAUT2 versus UV Fehlertyp – S4.2

NASA-TLX: Die ANOVA (Fischer) des NASA-TLX zeigt statistisch signifikante Unterschiede zwischen den Faktorstufen Fehlertypen für MD_Diff, P_Diff, F_Diff und der durchschnittlichen Beanspruchung (siehe Tabelle 53). Die Tabelle 85 des Anhangs 6.2 zeigt die Ergebnisse der deskriptiven Analyse über alle Subdimensionen (vgl. Abbildung 54). Die Voraussetzungen der Normalverteilung wurden bis auf PD und P sowie hinsichtlich der Varianzhomogenität für alle Subskalen erfüllt. In der Tabelle 84 des Anhangs 6.2 werden die Ergebnisse der Post-hoc-Tests der Subskalen des NASA-TLX aufgezeigt.

Subskalen	<i>df1; df2</i>	<i>F</i>	<i>p</i>	η^2p	Power
MD_Diff	2; 54	3,601	0,034	0,118	0,68
PD_Diff	2; 54	0,467	0,629	0,017	0,13
TD_Diff	2; 54	0,759	0,473	0,027	0,18
P_Diff	2; 54	3,185	0,049	0,106	0,62
E_Diff	2; 54	2,083	0,134	0,072	0,44
F_Diff	2; 54	4,535	0,015	0,144	0,78
Mean_NASA_Diff	2; 54	3,324	0,044	0,110	0,64

Tabelle 53. ANOVA NASA-TLX Differenzen versus Fehlertyp – 4.2**Abbildung 54. Barplots NASA-TLX versus Fehlertypen – S4.2**

Die Ergebnisse der Post-hoc-Tests zeigen, dass für MD_Diff ($t(54) = 2,66$; $MD = 25,95$; $p_{\text{Tukey}} = 0,027$; $Co'd = -0,856$; $Power = 0,74$) ein signifikanter Unterschied zwischen dem Daten- und dem Interaktionsfehler besteht. Ebenfalls konnten für P_Diff ($t(54) = 2,414$; $MD = 22,5$; $p = 0,05$; $Co'd = -0,805$; $Power = 0,66$), F_Diff ($t(54) = 2,98$; $MD = 33,1$; $p_{\text{Tukey}} = 0,012$; $Co'd = 0,993$; $Power = 0,84$) und die durchschnittliche Beanspruchung ($t(54) = 2,349$;

$MD = 16.02$; $p_{\text{Tukey}} = 0.048$; $Co'd = 0,783$; $Power = 0,64$) signifikante Unterschiede zwischen dem Interaktions- und Sichtbarkeitsfehler nachgewiesen werden.

Physiologische Parameter: Via der einfaktoriellen ANOVA nach Fischer konnte kein statistisch signifikanter Unterschied der AVs SCL_Diff_Diff $F(2;55) = 0,270$; $p = 0,765$; $\eta^2p = 0,010$; $Power = 0,09$, siehe Abbildung 55 b), und HR_Diff_Diff $F(2;55) = 0,211$; $p = 0,811$; $\eta^2p = 0,008$; $Power = 0,08$, siehe Abbildung 55 a), zwischen den Fehlertypen festgestellt werden. Analog zeigen die Ergebnisse der zugehörigen Post-hoc-Tests (vgl. Tabelle 86 in Anhang 6.2) keine Unterschiede. Die Voraussetzungen für die Varianzhomogenität wurden im Gegensatz zur Normalverteilung erfüllt. Aufgrund der Verletzung der Normalverteilungsvoraussetzung wurden beide Aussagen nochmal via der nichtparametrischen Kruskal-Wallis ANOVA mit demselben Ergebnis gegengeprüft. Die entsprechenden deskriptiven Daten sind der Tabelle 87 in Anhang 6.2 zu entnehmen.

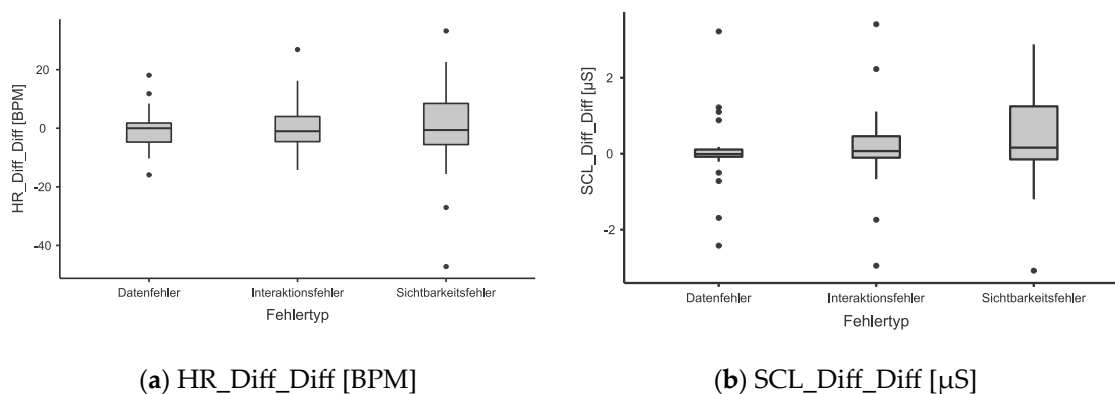


Abbildung 55. Boxplots HR_Diff_Diff & SCL_Diff_Diff versus FT – S4.2

4.4.2.3 Teilstudie 4.3 Akzeptanzmanipulation

Aufgrund des signifikanten Unterschieds in der BI mit der höchsten Effektstärke zwischen den Faktorstufen der UV Fehlerfall aus der Teilstudie 4.1 wurde für Teilstudie 4.3 der Fehlertyp Interaktionsfehler ausgewählt ($p = 0,045$; $Co'd = 0,529$).

Für die BI_Diff zeigt der unabhängige t -Test $t(22) = 2,38$; $p = 0,026$; $MD = 0,889$; $Co'd = 1,12$; $Power = 0,74$; $N_I = 6$; $N_{th} = 18$ innerhalb des Interaktionsfehlers einen signifikanten Unterschied zwischen den Faktorstufen der Moderatorvariable ohne und mit Indikatoranzeige der Systemkonfidenz (vgl. Abbildung 56).

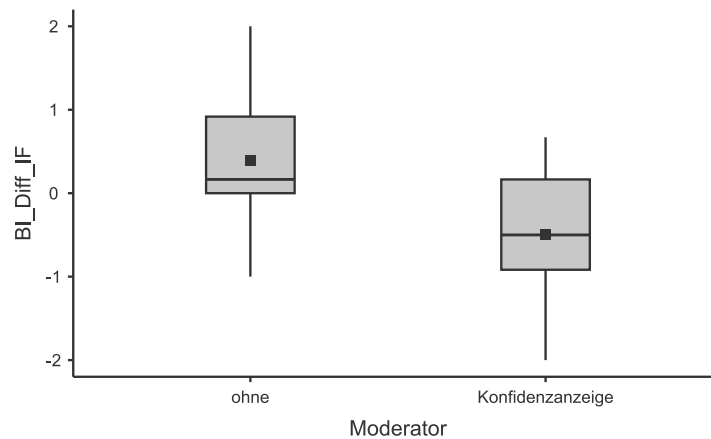


Abbildung 56. Boxplot BI_Diff versus Konfidenzanzeige – S4.3

Die beiden Voraussetzungen der Normalverteilung und der Varianzhomogenität sind erfüllt. Die BI-Mittelwertdifferenz liegt ohne Konfidenzanzeige ($n = 18$) bei 0,389 ($SD = 0,734$), hingegen mit Konfidenzanzeige ($n = 6$) bei -0,500 ($SD 0,961$). Das negative Vorzeichen zeigt an, dass die absolute BI ohne Interaktionsfehler niedriger war als mit Interaktionsfehler und Konfidenzanzeige.

4.4.3 Interpretation

Die Ergebnisse der Verbundstudie zeigen ganzheitlich einen komplexen Sachverhalt, den es gegen die dedizierten Hypothesen zu differenzieren gilt.

4.4.3.1 Teilstudie 4.1 Fehlerauswirkungen

Die erste Teilstudie 4.1 untersucht die Fehlerauswirkungen der Fehlerfälle (fh versus ff) unterschiedlicher Fehlertypen an einem Forschungsprototyp und beantwortet die folgenden Hypothesen:

Datenfehler (Bezug zu $FF_{1.A}$):

$H_{A.1.1}$: Fehlerhafte Informationen reduzieren bei IAS die Leistung.

Nachweis. Die $H_{A.1.1}$ wird durch die Ergebnisse bestätigt, somit kann die $H_{A.1.0}$ verworfen werden. Im Einzelnen zeigen die Ergebnisse (vgl. Abbildung 33 bis Abbildung 35), dass die DLZ bei fh-Daten im Vergleich zu ff-Informationen signifikant länger ist $t(52) = 4,56$; $MD = 163,8$ s; $p_{Tukey} < 0,001$; $Co'd = 0,831$; Power = 0,74. Ebenso ist der Index Leistung bei fehlerhaften Informationen signifikant niedriger $t(52) = 4,026$; $MD = 2,029$

$\%/(10*s)$; $p_{\text{Tukey}} = 0,002$; $Co'd = 0,92$. Hinsichtlich der AV Richtigkeit konnte kein statistisch signifikanter Unterschied nachgewiesen werden.

$H_{A.2.1}$: Fehlerhafte Informationen reduzieren bei IAS die Nutzerakzeptanz.

Nachweis. Die $H_{A.2.1}$ wird durch die Ergebnisse bestätigt, somit kann die $H_{A.2.0}$ verworfen werden. Die Ergebnisse zeigen, dass nur die Subskala BI einen signifikanten Unterschied zwischen den Faktorstufen der UV Fehlerfall aufweist $t(51) = -2,238$; $MD = -0,430$; $p_{\text{Tukey}} = 0,030$; $Co'd = 0,527$; $Power = 0,75$. Alle anderen Subdimensionen des UTAUT2-Fragebogens zeigen hierbei jedoch keinen statistisch signifikanten Unterschied (vgl. Abbildung 41).

$H_{A.3.1}$: Fehlerhafte Informationen erhöhen bei IAS die Beanspruchung des Anwenders.

Nachweis. Die $H_{A.3.1}$ wird durch die Ergebnisse nicht gestützt, somit kann die $H_{A.3.0}$ angenommen werden. Es konnten keine statistisch signifikanten Unterschiede bei der durchschnittlichen Beanspruchung sowie in den einzelnen Subskalen des NASA-TLX zwischen den Bedingungen fh- und ff-Informationen nachgewiesen werden. Dieses Ergebnis wird ebenfalls durch die fehlenden signifikanten Unterschiede der physiologischen Reaktionen (Stress) der AVs SCL und HR (vgl. Abbildung 43) gestützt.

Interaktionsfehler II (Bezug zu $FF_{1.c}$):

$H_{C.1.1}$: Fehlerhafte Interaktion reduziert bei IAS die Leistung.

Nachweis. Die $H_{C.1.1}$ wird durch die Ergebnisse gestützt, somit kann die $H_{C.1.0}$ verworfen werden. Die Post-hoc-Tests der Repeated Measures ANOVAs zeigen signifikante Unterschiede zwischen den Faktorstufen der UV IQ fh- und ff-Interaktion für die AV DLZ $t(52) = 5,225$; $MD = 203,9$ s; $p_{\text{Tukey}} < 0,001$; $Co'd = 1,23$; $Power = 1$ sowie den Index Leistung $t(52) = 3,748$; $MD = 2,05$ $\%/(10*s)$; $p_{\text{Tukey}} = 0,006$; $Co'd = 1,08$; $Power = 1$, vgl. Abbildung 33 und Abbildung 35. Für die UV Richtigkeit konnte kein statistisch signifikanter Unterschied nachgewiesen werden.

H_{C.2.1}: Fehlerhafte Interaktion reduziert bei IAS die Nutzerakzeptanz.

Nachweis. Die *H_{C.2.1}* wird durch die Ergebnisse gestützt, somit kann die *H_{C.2.0}* verworfen werden. Lediglich die Subdimension BI des UTAUT2-Fragebogens zeigt einen statistisch signifikanten Unterschied zwischen den Faktorstufen fehlerfreie und fehlerbehaftete Interaktion der UV IQ $t(51) = 2,24$; $MD = 0,413$; $p_{\text{Tukey}} = 0,045$; $Co'd = 0,529$; Power = 0,69. Alle anderen Subdimensionen zeigen keine signifikanten Unterschiede (vgl. Abbildung 44).

H_{C.3.1}: Fehlerhafte Interaktion erhöht bei IAS die Beanspruchung des Anwenders.

Nachweis. Die *H_{C.3.1}* wird durch die Ergebnisse gestützt, somit kann die *H_{C.3.0}* verworfen werden. Die durchschnittliche Beanspruchung $t(51) = 2,875$; $MD = 13,72$; $p_{\text{Tukey}} = 0,042$; $Co'd = 0,770$; Power = 0,93 sowie die inhärenten Dimensionen MD als psychologische Stressreaktion $t(51) = 3,073$; $MD = 19,94$; $p = 0,038$; $Co'd = 0,926$; Power = 0,98, $P t(51) = 2,139$; $MD = 13,80$; $p_{\text{Tukey}} = 0,047$; $Co'd = 0,703$; Power = 0,88 und $F t(51) = 3,168$; $MD = 25,26$; $p_{\text{Tukey}} = 0,029$; $Co'd = 0,823$; Power = 0,95 des NASA-TLX zeigen signifikante Unterschiede zwischen den Faktorstufen fh- und ff-Interaktion (vgl. Abbildung 45). Jedoch konnte die Beanspruchung des Nutzers nicht durch signifikante Unterschiede der physiologischen Reaktionen (Stress) der AVs SCL und HR nachgewiesen werden (vgl. Abbildung 45).

Sichtbarkeitsfehler II (Bezug zu FF_{1.B}):

H_{B.1.1}: Fehlerhafte Sichtbarkeit reduziert bei IAS die Leistung.

Nachweis. Die *H_{B.1.1}* wird durch die Ergebnisse gestützt, somit kann die *H_{B.1.0}* verworfen werden. Die Post-hoc-Tests der Repeated Measures ANOVAs zeigen signifikante Unterschiede zwischen den Faktorstufen fh- (F_2) und ff-Sichtbarkeit der UV VQ für die AV DLZ $t(52) = 4,56$; $MD = 163,8$ s; $p_{\text{Tukey}} < 0,001$; $Co'd = 0,813$; Power = 0,97 und den Index Leistung $t(52) = 4,026$; $MD = 2,029$ %/(10*s); $p_{\text{Tukey}} = 0,002$; $Co'd = 0,721$; Power = 0,94, vgl. Abbildung 33 bis Abbildung 35. Für die AV Richtigkeit konnte kein signifikanter Unterschied nachgewiesen werden.

$H_{B.2.1}$: Fehlerhafte Sichtbarkeit reduziert bei IAS die Nutzerakzeptanz.

Nachweis. Die $H_{B.2.1}$ wird durch die Ergebnisse nicht gestützt, somit wird die $H_{B.2.0}$ angenommen. Keine der Subdimensionen des UTAUT2-Fragebogens zeigte im Post-hoc-Test einen statistisch signifikanten Unterschied zwischen den beiden Faktorstufen fh- und ff-Sichtbarkeit der UV VQ (vgl. Abbildung 47).

$H_{B.3.1}$: Fehlerhafte Sichtbarkeit erhöht bei IAS die Beanspruchung des Anwenders.

Nachweis. Die $H_{B.3.1}$ wird durch die Ergebnisse nicht gestützt, somit wird die $H_{B.3.0}$ angenommen. Die Post-hoc-Tests zeigen weder für die durchschnittliche Beanspruchung noch für die einzelnen Subskalen des NASA-TLX signifikante Unterschiede zwischen den Faktorstufen fh und ff der UV VQ (vgl. Abbildung 48). Dies wird ebenfalls durch die Ergebnisse der physiologischen Reaktionen (Stress) via der AVs SCL und HR bestätigt (vgl. Abbildung 49).

4.4.3.2 Teilstudie 4.2 Fehlervergleich

Die zweite Teilstudie 4.2 vergleicht die Fehlerauswirkungen zwischen den Fehlertypen (Datenfehler, Interaktionsfehler II und Sichtbarkeitsfehler II) an einem Forschungsprototyp und beantwortet die folgenden Hypothesen (Bezug zu FF_2):

$H_{1.1}$: Unterschiedliche Fehlertypen manipulieren bei IAS die Leistung unterschiedlich

Nachweis. Die $H_{1.1}$ wird durch die Ergebnisse nicht gestützt, somit wird die $H_{1.0}$ angenommen. Die einfaktoriellen ANOVAs zeigen für die Leistungsdifferenz $F(2;54) = 0,222$; $p = 0,802$; $\eta^2p = 0,008$; Power = 0,08, die Richtigkeitsdifferenz $F(2;54) = 0,236$; $p = 0,791$; $\eta^2p = 0,009$; Power = 0,09 sowie die DLZ-Differenz $F(2;54) = 1,58$; $p = 0,216$; $\eta^2p = 0,055$; Power = 0,32 keine signifikanten Unterschiede zwischen den Fehlertypen (vgl. Abbildung 50 bis Abbildung 52). Ebenfalls zeigen die Post-hoc-Tests keine statistisch signifikanten Unterschiede zwischen den Fehlertypen.

H_{2.1}: Unterschiedliche Fehlertypen manipulieren bei IAS die Akzeptanz unterschiedlich

Nachweis. Die **H_{2.1}** wird durch die Ergebnisse gestützt, somit kann die **H_{2.0}** verworfen werden. Die ANOVA-Ergebnisse zeigen signifikante Unterschiede bei den Differenzen BI_Diff $F(2;54) = 4,14$; $p = 0,021$; $\eta^2p = 0,133$; Power = 0,74 und PE_Diff $F(2;54) = 3,28$; $p = 0,045$; $\eta^2p = 0,108$; Power = 0,63 zwischen den Fehlertypen (vgl. Abbildung 53). Die Post-hoc-Tests zeigen, dass sich BI_Diff signifikant zwischen Sichtbarkeits- und Datenfehlern $t(54) = -2,66$; $MD = -0,737$; $p_{\text{Tukey}} = 0,027$; $\text{Co'd} = -0,854$; Power = 0,99 sowie zwischen Interaktions- und Sichtbarkeitsfehlern $t(54) = 2,317$; $MD = 0,666$; $p_{\text{Tukey}} = 0,049$; $\text{Co'd} = 0,772$; Power = 0,97 unterscheidet. Für PE_Diff zeigen sich signifikante Unterschiede zwischen Sichtbarkeits- und Datenfehlern $t(54) = -2,21$; $MD = -0,736$; $p_{\text{Tukey}} = 0,048$; $\text{Co'd} = -0,710$; Power = 0,94 sowie zwischen Sichtbarkeits- und Interaktionsfehlern $t(54) = 2,252$; $MD = 0,778$; $p_{\text{Tukey}} = 0,050$; $\text{Co'd} = 0,751$; Power = 0,96.

H_{3.1}: Unterschiedliche Fehlertypen manipulieren bei IAS die Beanspruchung unterschiedlich

Nachweis. Die **H_{3.1}** wird durch die Ergebnisse gestützt, somit kann die **H_{3.0}** verworfen werden. Die ANOVA-Ergebnisse zeigen signifikante Unterschiede bei MD_Diff als empfundener Stress $F(2, 54) = 3,601$; $p = 0,034$; $\eta^2p = 0,118$; Power = 0,68, P_Diff $F(2;54) = 3,185$; $p = 0,049$; $\eta^2p = 0,106$; Power = 0,62, F_Diff $F(2;54) = 4,535$; $p = 0,015$; $\eta^2p = 0,144$; Power = 0,78 und der durchschnittlichen Beanspruchung $F(2;54) = 3,324$; $p = 0,044$; $\eta^2p = 0,110$; Power = 0,64. Die Post-hoc-Tests zeigen signifikante Unterschiede für MD_Diff $t(54) = 2,66$; $MD = 25,95$; $p_{\text{Tukey}} = 0,027$; $\text{Co'd} = -0,856$; Power = 0,74 zwischen Daten- und Interaktionsfehlern sowie bei P_Diff $t(54) = 2,414$; $MD = 22,5$; $p_{\text{Tukey}} = 0,05$; $\text{Co'd} = -0,805$; Power = 0,66, F_Diff $t(54) = 2,98$; $MD = 33,1$; $p_{\text{Tukey}} = 0,012$; $\text{Co'd} = 0,993$; Power = 0,84 und der durchschnittlichen Beanspruchung $t(54) = 2,349$; $MD = 16,02$; $p_{\text{Tukey}} = 0,048$; $\text{Co'd} = 0,783$; Power = 0,64 zwischen Interaktions- und Sichtbarkeitsfehlern. Diese Unterschiede spiegeln sich jedoch nicht in den physiologischen Reaktionen wider. Für die AVs SCL und HR konnten zwischen den Fehlertypen keine statistisch signifikanten Unterschiede nachgewiesen werden.

4.4.3.3 Teilstudie 4.3 Akzeptanzmanipulation

Die dritte Teilstudie 4.3 untersucht den Effekt der UV Systemkonfidenzanzeige mit den Faktorstufen „ohne“ und „mit Konfidenzanzeige“ auf die Akzeptanz (AV BI) beim Fehlertyp Interaktionsfehler und beantwortet die folgenden Hypothesen (Bezug zu FF_3):

$H_{4.1}$: Die Anzeige der Systemkonfidenz wirkt bei IAS positiv auf die Akzeptanz

Nachweis. Die $H_{4.1}$ wird durch die Ergebnisse gestützt, somit kann die $H_{4.0}$ verworfen werden. Der unabhängige t -Test zeigt einen signifikanten Unterschied bei BI_Diff innerhalb des Interaktionsfehlers $t(22) = 2,38$; $p = 0,026$; $MD = 0,889$; $Co'd = 1,12$. Der Mittelwert der BI-Differenz ohne Konfidenzanzeige liegt bei 0,389 ($SD = 0,734$), während er mit Konfidenzanzeige bei -0,500 ($SD = 0,961$) liegt. Damit schwächte die Konfidenzanzeige den negativen Einfluss des Interaktionsfehlers auf die BI ab und führte deskriptiv sogar zu einer höheren BI als im fehlerfreien Zustand.

4.4.4 Diskussion

Nachdem die Ergebnisse in Abschnitt 4.4.2 sowohl deskriptiv als auch inferenzstatistisch präsentiert und ausgewertet wurden, wurden die Implikationen der Ergebnisse auf die einzelnen Hypothesen in Abschnitt 4.4.3 bereits interpretiert. Im Folgenden gilt es, diese Erkenntnisse zueinander sowie gegen wissenschaftliche Vorarbeiten zu diskutieren.

4.4.4.1 Teilstudie 4.1 Fehlerauswirkungen

Datenfehler

In der vorliegenden Stichprobe wirkt sich der Datenfehler hinsichtlich der Leistungsfaktoren signifikant auf die DLZ ($MD = 163,8$ s) sowie auf den Index Leistung ($MD = -2,03 \text{ } \%/ (10 \cdot s)$) aus, jedoch nicht auf die AV Richtigkeit. Dies beeinflusst in Bezug auf die Akzeptanzfaktoren des UTAUT2 lediglich die BI ($MD = -0,430$) signifikant, während alle übrigen Dimensionen keine statistische Signifikanz erreichten. Dennoch zeigen sämtliche UTAUT2-Dimensionen Mittelwertsdifferenzen in der erwarteten Effektrichtung, wenngleich diese – mit Ausnahme der BI – statistisch nicht signifikant ausfallen.

Im Gegensatz dazu weisen die Subskalen des NASA-TLX kein konsistentes Reaktionsmuster auf. Zwar konnten keine statistisch signifikanten Unterschiede nachgewiesen werden, jedoch zeigen die Effektvorzeichen sowohl positive (PD, P, F) als auch negative (MD, TD, E, Mean_NASA) Richtungen innerhalb der Stichprobe. Die Ergebnisse deuten darauf hin, dass der Datenfehler primär die objektive Leistung und die Akzeptanz beeinflusst, ohne gleichzeitig konsistente subjektive oder physiologische Beanspruchungsreaktionen auszulösen. Der objektive Leistungsverlust spiegelt sich somit weder in den subjektiven Bewertungen des NASA-TLX noch in den meisten UTAUT2-Diskriminanten widers. Es kann vermutet werden, dass ein Teil der Nutzer im Laufe der Nutzungsdauer eine Unstimmigkeit zwischen den fehlerhaft angezeigten Maßen (Faktor 2) und der subjektiven Tischgröße wahrnimmt. Diese Vermutung wird durch die Erhebung der wahr-positiven Fehlerwahrnehmung des Datenfehlers mit 81 % (52,4 % eindeutig) gestützt. Es ist anzunehmen, dass ein Teil der Nutzer versuchen, diesen Sachverhalt zu plausibilisieren, was zu einer Erhöhung der DLZ bzw. einer Reduzierung der Leistung beigetragen haben könnte. Dass sich zugleich nur für die BI ein signifikanter Unterschied zeigte, könnte darauf hindeuten, dass der Fehler zwar die Nutzungsintention beeinträchtigte, jedoch nicht ausreichte, um die übrigen Akzeptanzdimensionen oder die wahrgenommene Beanspruchung messbar zu verändern. Entsprechend konnten auch für die physiologischen Parameter HR und SCL keine signifikanten Effekte nachgewiesen werden. Außerdem zeigen sie auch keine konsistenten Reaktionsmuster, da HR_Diff und SCL_Diff trotz fehlender Signifikanz entgegengesetzte Effektrichtungen aufweisen. Insgesamt können durch den Datenfehler weder die psychischen noch die physiologischen Stressreaktionen mit statistischer Signifikanz nachgewiesen werden. Bei dem Versuch, die BB weiter zu adjustieren, stellten die Probanden vermutlich fest, dass sie diesen Fehler (Messfehler um Faktor 2) nicht beeinflussen konnten. Möglicherweise orientierten sie sich deshalb stärker an der sichtbaren Übereinstimmung der BB mit der physischen Tischkante als an den fehlerhaften Maßangaben. Obwohl der Datenfehler von einem Großteil der Teilnehmenden erkannt wurde, scheint die maßliche Diskrepanz für die eigentliche Aufgabenbearbeitung von untergeordneter Bedeutung gewesen zu sein. Es bleibt offen, inwieweit ungünstige Darstellungsbedingungen das Situationsbewusstsein und den Fokus auf die immersive Informationsdarstellung zur kontextualen

Aufgabenstellung beeinflussen – beispielsweise durch perspektivisch bedingte Überlagerungen, Informationsüberlastung innerhalb des eingeschränkten Sichtfelds oder eine geringere Aufmerksamkeit für peripher dargestellte Informationen. (He & Zhou, 2024; Woodward & Ruiz, 2023). In der Diskussion bleibt der Widerspruch ungeklärt, warum die Probanden den Fehler bemerkten, jedoch dessen Auswirkungen nicht signifikant auf PE bzw. P rekurrierten. Ein Erklärungsversuch wäre, dass die Probanden mit ihrem Arbeitsergebnis, der Kongruenz der virtuellen BB bezüglich der physischen Tischkante, zufrieden waren, und der maßlichen Diskrepanz keine weitere Beachtung schenkten bzw. diese nicht weiter plausibilisierten. Dies würde sich anhand der Forschungsergebnisse der Vorarbeiten hinsichtlich Situationsbewusstsein und Wahrnehmung von virtuellen Informationen erklären lassen. Motivieren sollte die initiale Aufgabenstellung, dass sie ein Probandengeld in Höhe von zwanzig Euro nur erhielten, wenn sie richtig und schnell im Verhältnis zur Vergleichsgruppe messen würden. Gegebenenfalls sind diese Effekte bzw. Fragestellungen, ob Datenfehler Auswirkungen auf die Akzeptanz- sowie Belastungsfaktoren als auch auf die vermuteten Inhibitoren haben, in weiterführenden Forschungsaktivitäten eingehender zu erforschen.

Die Ergebnisse erweitern die Befunde von Nasirian et al. (2017), die Vertrauen als wichtigen Prädiktor der Nutzungsintention identifizierten. In deren Untersuchung beeinflusste insbesondere die wahrgenommene Interaktionsqualität das Vertrauen signifikant, während sich für die Informations- und Systemqualität keine signifikanten direkten Effekte auf das Vertrauen zeigten. Die vorliegende Untersuchung ergänzt diese wahrnehmungsbezogene Betrachtung um einen experimentell erzeugten Datenfehler, der sowohl die objektive Leistung als auch die BI signifikant beeinträchtigte. Da Vertrauen nicht erhoben wurde, lässt sich jedoch nicht bestimmen, ob der Effekt auf die BI unmittelbar oder über eine veränderte Vertrauensbewertung vermittelt wurde. Die Befunde stehen somit nicht im Widerspruch zu Nasirian et al. (2017). Vielmehr bleibt offen, ob der festgestellte Unterschied der BI einen unmittelbaren Akzeptanzeffekt des Datenfehlers darstellt oder ebenfalls durch einen nicht erhobenen Vertrauensverlust vermittelt wurde.

Im Vergleich zur vorliegenden Studie, in der ein technischer Datenfehler die Leistung sowie die Nutzungsintention signifikant beeinträchtigte – jedoch ohne

Auswirkungen auf Beanspruchung oder physiologische Parameter –, berichteten Konopka et al. (2024a), dass gezielt fehlerhafte Handlungsempfehlungen in einem AR-DSS das Vertrauen signifikant reduzieren ($F = 8,45$; $p < 0,01$). Ein wesentlicher Unterschied liegt darin, dass in der vorliegenden Studie Vertrauen nicht direkt gemessen wurde; die signifikante Reduktion der BI ist jedoch mit der Annahme vereinbar, dass der erkannte Fehler (Fehlerwahrnehmung 81 %) von den Nutzenden als relevantes Systemdefizit bewertet wurde, das sich negativ auf die Nutzungsabsicht auswirkt. Ein daraus resultierender Vertrauensverlust kann jedoch nicht nachgewiesen werden.

Während Konopka et al. (2024b) neben Vertrauensverlusten auch emotionale Reaktionen auf technologische Fehlfunktionen nachweisen, zeigten sich in der vorliegenden Studie keine signifikanten subjektiven oder physiologischen Beanspruchungsreaktionen. Dies könnte darauf hindeuten, dass der Datenfehler eher die Bewertung der zukünftigen Systemnutzung als das Beanspruchungserleben beeinflusste.

Gegoff et al. (2023) zeigten, dass eine hohe Automationszuverlässigkeit Vertrauen, Genauigkeit und Entscheidungsgeschwindigkeit verbessert, während höhere Transparenz zusätzlich die subjektive Arbeitsbelastung reduziert. Die vorliegenden Ergebnisse weisen in eine ähnliche Richtung, da der Datenfehler die DLZ, den Index Leistung und die BI signifikant beeinträchtigte. Im Unterschied zu Gegoff et al. (2023) zeigte sich jedoch keine signifikante Veränderung der subjektiven oder physiologischen Beanspruchung. Dies deutet darauf hin, dass der Datenfehler in der vorliegenden Aufgabe vor allem die Leistung und Akzeptanz beeinflusste, ohne gleichzeitig eine messbare Belastungsreaktion auszulösen.

Ein Vergleich mit der Studie von Brauner et al. (2019) zeigt deutliche Parallelen hinsichtlich der Auswirkungen fehlerhafter Systemempfehlungen. Beide Arbeiten belegen signifikante Leistungseinbußen bei Datenfehlern: Brauner et al. (2019) berichteten eine durchschnittliche Reduktion des Mediangewinns um 38 % ($p = 0,027$); auch in der vorliegenden Studie führte ein fehlerhaftes System zu signifikant schlechterer Leistung. Zudem untersuchten beide Studien die Systemakzeptanz, die sich unter fehlerhaften Bedingungen in beiden Studien verringerte. Während Brauner et al. (2019) zusätzlich den Einfluss auf Vertrauen und Compliance analysierten, konzentriert sich die vorliegende Untersuchung stärker auf Leistungskennzahlen, Beanspruchung und Akzeptanz.

Madhavan et al. (2006) berichteten bei Fehlern in einfach lösbaren Aufgaben ein niedrigeres Vertrauen in die Automation ($M = 2,67$ gegenüber $M = 3,87$). Der Unterschied verfehlte mit $t(28) = 1,71$ und $p = 0,07$ jedoch das übliche Signifikanzniveau. Die Ergebnisse weisen dennoch darauf hin, dass Fehler bei Aufgaben, deren richtige Lösung Nutzer selbst sicher beurteilen können, für die Vertrauenskalibrierung besonders relevant sein können.

Interaktionsfehler II

In der vorliegenden Stichprobe wirkt sich der Interaktionsfehler bei den Leistungsfaktoren signifikant auf die DLZ ($MD = 203,9$ s) sowie auf den Index Leistung ($MD = 2,05$ %/(10*s) aus, jedoch nicht auf die AV Richtigkeit. Dies beeinflusst hinsichtlich der Akzeptanzfaktoren des UTAUT2 lediglich die BI ($MD = 0,41$) und nicht die Dimension der PE. Weiter zeigen die Daten, dass der Interaktionsfehler signifikante Auswirkungen sowohl auf Subskalen MD ($MD = 19,94$), P ($MD = 13,80$), F ($MD = 25,26$) als auch auf die durchschnittliche Gesamtbeanspruchung ($MD = 13,72$) des NASA-TLX hat. Das bedeutet, dass der objektive Leistungsverlust durch den Interaktionsfehler subjektiv nicht eindeutig als Leistungsverlust (bei P, aber nicht PE) wahrgenommen wird. Außerdem empfinden die Probanden psychischen Stress in Form einer erhöhten MD sowie F. Insgesamt kann durch den Interaktionsfehler in Summe ein signifikanter Akzeptanzverlust beobachtet werden. Jedoch können hierdurch keine physischen Stressreaktionen via Pulsrate oder Hautleitfähigkeit nachgewiesen werden. 77,8 % (wahr-positiv) der Probanden haben im Fehlerfall den Interaktionsfehler als solchen wahrgenommen (55,6 % eindeutig wahr). Als mögliche kausale Begründung der Effekte wird vermutet, dass die Probanden sich stärker mental (MD) anstrengen mussten, die BB exakt zu positionieren, da die Manipulation aller sechs Freiheitsgrade anstatt der nativen drei Freiheitsgrade (X, Y, Z) die Handrotation (XX, YY, ZZ) und Translation (X, Y, Z) an der BB beim Greifen direkt ermöglichte. Dies führte außerdem zu einer höheren Gesamtbeanspruchung (Mean_NASA) und längeren DLZ, um genauso exakt zu positionieren wie ohne Fehler. Dies spiegelt sich in höheren Frustrationswerten wider.

Der Interaktionsfehler II (Gestensteuerung) lässt sich aufgrund interaktionsspezifischer Metriken nicht direkt mit dem Interaktionsfehler I (vgl. Kapitel 4.2, Studie 2)

vergleichen. Näherungsweise können hier hinsichtlich der Richtigkeit die AV WER bei der ASR und die AV Richtigkeit bei der Gesteninteraktion für einen Vergleich herangezogen werden. Bei dem Vergleich zeigt sich hierbei, dass eine fehlerhafte ASR-Interaktion eine signifikante Auswirkung auf die Richtigkeit (WER, SPL 0 dB – 87 dB ($p = 0,001$; $Co'd = 0,99$) im Gegensatz zur Gestensteuerung (Richtigkeit UV Fehlerfall ($p = 0,936$; $Co'd = -0,32$) hat. Analog können für den Vergleich der Leistung die Metriken der AVs ITR bei der ASR und der Index Leistung bei der Gestensteuerung herangezogen werden. Vergleichend zeigt sich, dass die UV Fehlerfall für beide Interaktionsparadigmen signifikante Unterschiede aufweist. Für den Interaktionsfehler der Gestensteuerung bemisst sich der Unterschied des Index Leistung auf $t(52) = 3,748$; $MD = 2.05 \text{ } \%/ (10*s)$; $p_{\text{Tukey}} = 0,006$; $Co'd = 1,08$; $\text{Power} = 1,00$ und für die ASR analog der ITR für die extremen Faktorstufen 0 dB – 87 dB auf $t(21) = 3,461$; $MD = 11,79$; $p_{\text{Tukey}} = 0,001$; $Co'd = 0,738$; $\text{Power} = 0,96$. Da es sich hierbei um unterschiedliche Metriken handelt, kann der Vergleich nur unter Vorbehalt numerisch und nicht entsprechend des „Quality of Service“ geführt werden. In dieser Arbeit tendiert der Fehlerfall bei der Gestensteuerung ($Co'd = 1,08$; relativ 45 % Leistungsverlust bezogen auf den Mittelwert ohne Fehler) zu stärkeren Effekten auf die Leistung als bei der ASR ($Co'd = 0,738$; relativ 31 % Leistungsverlust bezogen auf den Mittelwert ohne Fehler – 0 dB). Hinsichtlich der Akzeptanz zeigt sich ein klareres Bild. Bei der ASR konnte durch die Fehlerreihenfolge die BI signifikant relativ um 46 % manipuliert werden, bezogen auf den Mittelwert der SER-Reihenfolge ($Co'd = 1,56$). Hingegen konnte bei der Gestensteuerung im Fehlerfall nur ein Akzeptanzverlust von 12 % ($Co'd = 0,53$) nachgewiesen werden. Bei der ASR sind im Gegensatz zur Gestensteuerung durch die UV Fehlerfall alle Unterschiede der Akzeptanzfaktoren signifikant mit gleichem Vorzeichen beeinflusst (vgl. den Abschnitt 4.2.2). Hinsichtlich der Beanspruchung konnte bei der ASR mittels der unterschiedlichen Fehlerreihenfolge beim Mean_NASA eine signifikante relative Manipulation um 29,8 % festgestellt werden, bezogen auf den Mittelwert ($Co'd = 0,49$). Im Gegensatz dazu konnte bei der Gestensteuerung im Fehlerfall lediglich ein mittlerer Beanspruchungsanstieg von relativ 32 % ($Co'd = 0,77$) nachgewiesen werden. Jedoch ist darauf hinzuweisen, dass eine direkte Vergleichbarkeit der Fehlereffekte zwischen der Gestensteuerung und der ASR im Fehlerfall nur eingeschränkt möglich ist. Bei der Gestensteuerung wurde der Akzeptanzverlust bzw. der

Beanspruchungsanstieg in einem gepaarten Versuchsdesign mit einem Within-Subject-Faktor erhoben. Hingegen wurde beim ASR mit einem Between-Subject-Faktor innerhalb eines gegenbalancierten Wiederholmessungsdesigns verfahren. Diese Ansätze waren aus ganzheitlicher Betrachtung der DOEs entsprechend gegeben.

Die vorliegenden Ergebnisse und Erkenntnisse stehen der AR-Interaktionsfehlerstudie von Graichen & Graichen (2023) entgegen. In der Studie mit 35 Teilnehmern wurde keine signifikante Korrelation zwischen der Anzahl der beobachteten Fehler und der allgemeinen Akzeptanz der gestenbasierten VR-Anwendung festgestellt ($\rho = -0,19$, $p = 0,253$). Allerdings zeigte sich eine signifikante negative Korrelation zwischen dem Nutzervertrauen und den Fehlern ($\rho = -0,35$, $p = 0,041$). Dies weist darauf hin, dass Nutzer, obwohl sich ihre Einstellung gegenüber der Gesteninteraktion nicht änderte, ihr Vertrauen in das System verloren, wenn sie mehr Fehler beobachteten.

Dementsprechend sind in der vorliegenden Studie die Aussagen von Suh und Prophet (2018) vereinbar, dass aufgrund von Interaktionsfehlern eine schlechte Passung zwischen motorischen, sensorischen und wahrnehmungsbezogenen Stimuli zu einer kognitiven Überlastung führt und Unbehagen u. a. Frustration fördert.

Ebenfalls konnte somit die Aussage von Moencks et al. (2022) gestützt werden, dass technische Fehler zu Zeitverlust führen.

Bei Cockburn et al. (2019) wurde festgestellt, dass Vibration bei der Toucheingabe die Fehlerquote um ca. 50 % erhöht. Abgesehen von exakten Einheiten konnte in der vorliegenden Studie nachgewiesen werden, dass eine fehlerhafte Interaktion Auswirkungen auf die effektive Leistung (Ausführungszeit) hat. Die Ergebnisse sind mit der vorliegenden Studie insofern vergleichbar, als erschwerte Interaktionsbedingungen die objektive Leistung beeinträchtigen können und eine angepasste Interaktionsgestaltung diese Auswirkungen reduzieren kann.

Die Ergebnisse von Pourmemar (2024) stehen im Einklang mit der vorliegenden Studie, da beide Untersuchungen zeigen, dass eine ungünstig gestaltete Gesteninteraktion sowohl die objektive Leistung als auch die subjektive Beanspruchung beeinträchtigen kann. Bei Pourmemar (2024) unterschieden sich die Auswahlzeiten signifikant zwischen den Interaktionsformen, $F(3,297) = 13,298$; $p < 0,05$. Auch für die körperliche Beanspruchung, $F(3,72) = 32,471$; $p < 0,05$, und die F $F(3,72) = 70,795$; $p < 0,05$, zeigten sich

signifikante Unterschiede. Allerdings verglich Pourmemar (2024) verschiedene regulär funktionierende Eingabemodalitäten, während in der vorliegenden Studie die Gestensteuerung gezielt fehlerhaft beziehungsweise erschwert gestaltet wurde.

Caputo et al. (2024) zeigten, dass eine symmetrische und integrierte Gestensteuerung die Leistung signifikant verbessern kann. Die Handlebar-Steuerung führte gegenüber der MRTK-Standardsteuerung und der an klassischen Desktop-Anwendungen angelehnten DTP-Steuerung, bei der Translation, Rotation und Skalierung über separate virtuelle Bedienelemente nacheinander gesteuert werden, zu signifikant kürzeren Bearbeitungszeiten, signifikant weniger erforderlichen Bewegungen und einer signifikant höheren Rotationsgenauigkeit. Im Unterschied dazu wurde in der vorliegenden Studie keine alternative Interaktionsmethode verglichen, sondern ein gezielter Interaktionsfehler untersucht. Dieser beeinträchtigte vor allem die Geschwindigkeit und Effizienz, während sich für die Richtigkeit kein signifikanter Effekt zeigte. Darüber hinaus erhöhte der Interaktionsfehler die mentale Beanspruchung und Frustration signifikant, wohingegen Caputo et al. (2024) keine signifikanten Unterschiede hinsichtlich der wahrgenommenen Schwierigkeit und Ermüdung feststellten. Die vorliegenden Ergebnisse erweitern ebenfalls Erkenntnisse der Vorarbeit von Nasirian et al. (2017). Diese Studie untersucht, welche Qualitätsfaktoren die Akzeptanz KI-basierter Sprachassistenzsysteme beeinflussen. Dabei erweist sich die Interaktionsqualität – verstanden als die Fähigkeit des Systems, als dialogfähiger Kommunikationspartner zu agieren – als stärkster Prädiktor für das Vertrauen der Nutzer. Der Pfad von der Interaktionsqualität zum Vertrauen ist signifikant positiv ($p < 0,01$), ebenso der Pfad vom Vertrauen zur Nutzungsintention ($p < 0,01$). Somit entfaltet die Interaktionsqualität einen indirekten, signifikant positiven Einfluss auf die Nutzungsabsicht und damit auf die Akzeptanz der Systeme.

Sichtbarkeitsfehler II

In der vorliegenden Stichprobe wirkt sich der Sichtbarkeitsfehler bei den Leistungsfaktoren signifikant auf die DLZ ($MD = 163,8s$) sowie auf den Index Leistung ($MD = 2,03 \%/ (10*s)$) aus, jedoch nicht auf die AV Richtigkeit. Demgegenüber beeinflusst der Sichtbarkeitsfehler weder die Akzeptanzfaktoren des UTAUT2 noch die Beanspruchungsfaktoren des NASA-TLX signifikant. Ebenfalls kann durch den Sichtbarkeitsfehler keine

physische Stressreaktion der Pulsrate bzw. Hautleitfähigkeit nachgewiesen werden. Diese Ergebnisse lassen vermuten, dass der objektive Leistungsverlust durch den Sichtbarkeitsfehler subjektiv nicht in vergleichbarem Ausmaß wahrgenommen wird. Außerdem wird die Akzeptanz (BI) durch den Leistungsverlust des Sichtbarkeitsfehlers nicht beeinflusst. 73,7 % (wahr-positiven) der Probanden haben im Fehlerfall den Sichtbarkeitsfehler als solchen wahrgenommen (52,6 % eindeutig wahr). Bemerkenswert ist, dass der Sichtbarkeitsfehler zwar die beiden Leistungsfaktoren DLZ und Index Leistung negativ beeinflusst, jedoch lässt sich, wie vorher beschrieben, kein signifikanter Verlust bei den Akzeptanzfaktoren des UTAUT2 sowie der Beanspruchungsskalen des NASA-TLX feststellen. Auffällig ist, dass in dieser Stichprobe die Mittelwertdifferenzen ($MD_{PE} = 0,263$; $MD_{EE} = 0,158$; $MD_{HM} = 0,090$; $MD_{HT} = 0,105$; $MD_{BI} = 0,280$) für fast alle Akzeptanzfaktoren im Fehlerfall tendenziell höher als ohne Fehler liegen, mit Ausnahme der Dimensionen SI und FC. Auch beim NASA-TLX zeigt sich überwiegend ein analoges Verhalten. Für die Skalen PD ($MD = -6,053$), P ($MD = -10,001$), F ($MD = -6,316$) und Mean_NASA ($MD = -2,018$) liegt die entsprechende Dimension der Beanspruchung im Fehlerfall niedriger als ohne Fehler. Eine mögliche Erklärung für diese unerwarteten Ergebnisse ist, dass die reduzierte Überblendung der realen Welt durch die verringerten Helligkeitswerte (30 %) des virtuellen Objekts im Fehlerfall objektiv die Leistungsfaktoren beeinflusst. Gleichzeitig führt die geringere Opazität bzw. höhere Transparenz beim Positionieren der BB zu einem stärkeren Kontrast der Umrandungen des weißen Tisches gegenüber dem Boden, wodurch das zu vermessende Objekt weniger maskiert wird. Dieser optische Effekt könnte subjektiv diese widersprüchlichen Ergebnisse von der Kausalität her begründen. Beim Vergleich mit Studie 1 (Sichtbarkeitsfehler I – VR) zeigen beide Studien, dass fehlerhafte Sichtbarkeit die Leistung bei der Nutzung interaktiver AS beeinträchtigt, jedoch in unterschiedlichen Ausprägungen. Beim Sichtbarkeitsfehler I (VR) war insbesondere die Richtigkeit der Aufgabenausführung betroffen ($t(16) = 4,06$; $p < 0,001$; $MD = 0,275$; $Co'd = 1,927$; Power = 0,99, während die Durchlaufzeit (DLZ) keinen signifikanten Unterschied aufwies ($t(16) = -0,295$; $p = 0,614$; $MD = -15,8$ s). Im Gegensatz dazu zeigte Sichtbarkeitsfehler II (AR) signifikante Leistungseinbußen bei der DLZ ($t(52) = 4,56$; $MD = 163,8$ s; $p < 0,001$; $Co'd = 0,813$; Power = 0,97 sowie beim Leistungsindex ($t(52) = 4,026$; $MD = 2,029$ %/(10 s); $p = 0,002$; $Co'd = 0,721$; Power = 0,94, jedoch

ohne signifikanten Unterschied in der Richtigkeit. Auch in Sichtbarkeitsfehler II wiesen weder die UTAUT2-Dimensionen noch die Beanspruchungsmaße (NASA-TLX sowie SCL/HR) signifikante Effekte auf. Bezüglich Akzeptanz und Beanspruchung zeigen beide Studien konsistente Ergebnisse: Es wurden keine signifikanten Unterschiede zwischen ff- und fh-Sichtbarkeit festgestellt. Bei Sichtbarkeitsfehler I waren beispielsweise die Unterschiede der Akzeptanzdimension BI minimal ($Co'd = 0,063$; $Power = 0,05$), ebenso zeigten sich keine signifikanten Unterschiede bei den NASA-TLX-Skalen mit Ausnahme einer einzelnen Subdimension. Zusammenfassend lässt sich festhalten, dass Sichtbarkeitsfehler in beiden Studien die objektive Leistung beeinträchtigen, jedoch auf unterschiedliche Leistungsaspekte wirken, während Akzeptanz und subjektive Beanspruchung unbeeinflusst bleiben.

Die von Wandke (2005) postulierten Effekte kognitiver Überlastung infolge unübersichtlicher Inhalte – die im weiteren Sinne mit Sichtbarkeitsfehlern vergleichbar sind – konnten in der vorliegenden Studie nicht bestätigt werden.

Wie bereits im Rahmen des Sichtbarkeitsfehlers I werden auch die Ergebnisse des Sichtbarkeitsfehlers II den Befunden von Arning et al. (2012) gegenübergestellt. Im Vergleich zur vorliegenden Untersuchung zeigt die Studie von Arning et al. (2012) eine stärkere subjektive und akzeptanzbezogene Wirkung von Sichtbarkeitsproblemen. Zugleich ergaben sich signifikante Unterschiede in der Leistung zugunsten der Bildschirmdarstellung ($M_{\text{Bildschirm}} = 0,91$; $SD = 0,09$ vs. $M_{\text{Projektor}} = 0,85$; $SD = 0,10$; $F(1,23) = 12,5$, $p < 0,01$). Während der Sichtbarkeitsfehler in der aktuellen Studie keine signifikanten Effekte auf die Akzeptanzfaktoren oder die wahrgenommene Beanspruchung hatte, wurden bei Arning et al. die Sichtbarkeit ($M_{\text{Bildschirm}} = 93,8$; $SD = 7,9$ vs. $M_{\text{Projektor}} = 77,3$; $SD = 14,4$; $F(1,23) = 22,9$; $p < 0,001$), das Vertrauen ($M_{\text{Bildschirm}} = 86,1$; $SD = 11,7$ vs. $M_{\text{Projektor}} = 77,5$; $SD = 14,3$; $F(1,23) = 9,1$; $p < 0,05$) und die Akzeptanz ($M_{\text{Bildschirm}} = 92,7$; $SD = 9,0$ vs. $M_{\text{Projektor}} = 87,5$; $SD = 11,8$; $F(1,23) = 4,9$; $p < 0,05$) signifikant schlechter bewertet. Dabei wirkte die Sichtbarkeit positiv auf das Vertrauen ($\beta = 0,545$; $p < 0,001$), welches wiederum die Desorientierung verringerte ($\beta = -0,805$; $p < 0,001$). Eine stärkere Desorientierung führte schließlich zu einer geringeren Akzeptanz ($\beta = -0,632$; $p < 0,001$). Die subjektiv wahrgenommene Gesamtbeanspruchung fiel bei der Projektornutzung mit $M = 13,1$ ($SD = 9,4$) zwar deskriptiv höher aus als bei der Bildschirmdarstellung mit $M = 10,0$ ($SD = 9,8$), unterschied sich

jedoch nicht signifikant. Insgesamt deuten die Befunde darauf hin, dass Sichtbarkeitsprobleme zwischen unterschiedlichen Visualisierungsmodalitäten stärkere indirekte Auswirkungen auf Vertrauen, Desorientierung und Akzeptanz entfalten können als Sichtbarkeitsunterschiede innerhalb einer einheitlichen Visualisierungsart.

Nachdem Alessa et al. (2023) bereits beim ersten Sichtbarkeitsfehler (vgl. Abschnitt 4.1.4) diskutiert wurden, ist der Vergleich mit dem AR-Sichtbarkeitsfehler II ergänzend zu verstehen. Besonders relevant ist, dass sich die Vorteile der AR-basierten Instruktion bei Alessa et al. (2023) überwiegend bei komplexen Wartungsaufgaben zeigten. Bei einfachen Aufgaben ergaben sich dagegen weder für die physische und mentale Anforderung noch für die untersuchten HRV-Parameter signifikante Unterschiede zwischen AR- und papierbasierter Instruktion. Deskriptiv fiel die AR-Bedingung teilweise sogar ungünstiger aus, da die zeitliche Anforderung höher und die subjektiv bewertete Leistung schlechter war; lediglich die Frustration war geringer. Die Vergleichbarkeit mit dem AR-Sichtbarkeitsfehler II ist jedoch eingeschränkt. Alessa et al. (2023) untersuchten keinen isolierten Sichtbarkeitsfehler, sondern verglichen AR- und papierbasierte Instruktionen, die sich zugleich hinsichtlich Informationsdarstellung, Interaktion, räumlicher Einbettung und körperlicher Handhabung unterschieden. Zudem bestand die Stichprobe ausschließlich aus 28 männlichen Teilnehmern, was die Vergleichbarkeit deutlich aufgrund der Selektivität der Stichprobe weiter einschränkt. Daher lässt sich nicht eindeutig bestimmen, welchen Anteil die Sichtbarkeit selbst an den beobachteten Effekten hatte. Insgesamt sprechen die Befunde eher dafür, dass subjektive und physiologische Entlastungseffekte insbesondere bei hoher Aufgabenkomplexität auftreten, während sie bei einfacheren Aufgaben deutlich geringer ausfallen.

Wie bereits bei der Diskussion des Sichtbarkeitsfehlers I (vgl. Abschnitt 4.1.4) herausgearbeitet, stimmen die Befunde von Sadovitch (2020) nur teilweise mit den Ergebnissen der vorliegenden Untersuchungen überein. Auch beim AR-Sichtbarkeitsfehler zeigten sich keine signifikanten Veränderungen der Akzeptanzdimensionen PE, EE, FC, HM und BI. Gleichzeitig führte der Fehler jedoch zu einer signifikanten Verlängerung der DLZ um $MD = 163,8 \text{ s}$; $t(52) = 4,56$; $p < 0,001$; $Co'd = 0,813$ sowie zu einer signifikanten Verringerung des Index Leistung um $2,03 \text{ \%}/(10\text{-s})$: $t(52) = 4,03$; $p = 0,002$; $Co'd = 0,721$. Für die Richtigkeit zeigte sich dagegen kein signifikanter Effekt. Dieses Muster spricht dafür,

dass die Teilnehmenden die visuelle Beeinträchtigung teilweise durch eine langsamere und kontrolliertere Aufgabenbearbeitung kompensierten. Der Fehler wirkte sich somit objektiv auf Geschwindigkeit und Effizienz aus, ohne die korrekte Aufgabenlösung oder die subjektive Systemakzeptanz wesentlich zu beeinträchtigen. Im Unterschied zu den bei Sadovitch untersuchten Registrierungsfehlern war der Sichtbarkeitsfehler möglicherweise weniger salient, weniger störend oder für die grundsätzliche Bewertung des Systems weniger relevant. Die Ergebnisse verdeutlichen daher, dass objektive Leistungseinbußen durch visuelle Fehler nicht zwangsläufig mit Akzeptanzverlusten einhergehen, sondern von Fehlerart, Salienz, funktionaler Bedeutung und Kompensierbarkeit abhängen.

Folgend der Diskussion des Sichtbarkeitsfehlers I, stehen auch die Befunde von Merenda et al. (2019) grundsätzlich im Einklang mit den Ergebnissen des Sichtbarkeitsfehlers II. In beiden Untersuchungen beeinflussten visuelle Bedingungen die objektive Leistung, ohne dass sich entsprechende Veränderungen in den subjektiven Bewertungen nachweisen ließen. Während die erhöhte visuelle Fidelity bei Merenda et al. (2019) zu längeren Brems- und Blickzeiten sowie zu einer geringeren Ausrichtung des Blicks auf die Straße führte, zeigte sich beim Sichtbarkeitsfehler II eine signifikante Beeinträchtigung der Bearbeitungsgeschwindigkeit und Effizienz, jedoch kein signifikanter Effekt auf die Richtigkeit, die Akzeptanz oder die wahrgenommene Beanspruchung. Damit bestätigt sich das bereits für den Sichtbarkeitsfehler I diskutierte Muster, wonach visuelle Beeinträchtigungen objektive Leistungsparameter verändern können, ohne von den Nutzenden in gleicher Weise subjektiv abgebildet zu werden. Inhaltlich unterscheiden sich die Studien jedoch hinsichtlich der visuellen Manipulation: Merenda et al. (2019) untersuchten keine eingeschränkte Sichtbarkeit, sondern eine erhöhte grafische Detailfülle, die vermutlich zusätzliche visuelle Aufmerksamkeit band und dadurch die Reaktionsleistung beeinträchtigte. Der Sichtbarkeitsfehler der vorliegenden Studie erschwerte dagegen unmittelbar die Wahrnehmung aufgabenrelevanter Informationen. Die übereinstimmenden Befunde legen somit nahe, dass sowohl eine zu hohe visuelle Komplexität als auch eine unzureichende Sichtbarkeit die effiziente Informationsverarbeitung beeinträchtigen können. Dass sich diese objektiven Nachteile weder bei Merenda et al. (2019) im subjektiven Situationsbewusstsein noch in der vorliegenden Studie

in der Akzeptanz oder Beanspruchung widerspiegeln, spricht dafür, dass Nutzende visuelle Leistungseinbußen teilweise kompensieren oder nicht bewusst als systembedingte Beeinträchtigung wahrnehmen.

4.4.4.2 Teilstudie 4.2 Fehlervergleich

Die Ergebnisse der Teilstudie 4.2 zum Fehlervergleich zeigen ein differenziertes Bild der Auswirkungen verschiedener Fehlertypen. Eine stringente Aussage, die eine klare Unterscheidung zwischen den Fehlertypen erlaubt, lässt sich aufgrund der komplexen Befunde nicht ableiten und erfordert eine detaillierte Diskussion.

Obwohl die einfaktorielle ANOVA und die Post-hoc-Tests keine signifikanten Unterschiede der Leistungsfaktoren zwischen den Fehlertypen zeigten, liefern die deskriptiven Ergebnisse wichtige Hinweise. Interaktionsfehler erwiesen sich durch hohe Mittelwerte und geringe Streuung als besonders kritisch. Sichtbarkeitsfehler hatten den geringsten Einfluss bei gleichzeitig sehr stabilen Werten. Der Datenfehler zeigte zwar den größten zeitlichen Einfluss, war jedoch durch eine hohe Streuung gekennzeichnet. Diese lässt sich darauf zurückführen, dass ein Teil der Probanden den Fehler nicht wahrnahm, während andere aktiv versuchten, die fehlerhaften Maßangaben zu kompensieren – was sich negativ auf DLZ und Richtigkeit auswirkte. Diese Unterschiede sind lediglich im Rahmen dieser Studie interpretierbar und nicht generalisierbar. Die geringe Teststärke (Power = 0,08 bis 0,32) weist darauf hin, dass die Wahrscheinlichkeit zum Nachweis tatsächlich vorhandener Unterschiede begrenzt war. Für zukünftige Untersuchungen erscheint daher ein größerer Stichprobenumfang sinnvoll.

Bei den Akzeptanzfaktoren konnten nur für BI_Diff ($p = 0,021$; $\eta^2p = 0,133$) und PE_Diff ($p = 0,045$; $\eta^2p = 0,108$) innerhalb der Faktorstufen der UV Fehlertyp signifikante Unterschiede nachgewiesen werden. Die konsekutiven Post-hoc-Tests zeigen für BI_Diff und PE_Diff, dass die Fehlerauswirkungen des Interaktions- und Datenfehlers auf einem Niveau liegen. Hingegen liegt der Sichtbarkeitsfehler auf einem signifikant anderen Niveau. Es lassen sich sowohl bei BI_Diff zwischen den Faktorstufen Sichtbarkeits- und Datenfehler ($MD = -0,737$; $p_{\text{Tukey}} = 0,027$; $Co'd = -0,854$) als auch zwischen Interaktions- und Sichtbarkeitsfehler ($MD = 0,666$ $p_{\text{Tukey}} = 0,049$; $Co'd = 0,772$) signifikante Unterschiede nachweisen. Gleiches gilt bei PE_Diff zwischen Sichtbarkeits- und Datenfehler ($MD =$

-0,736; $p_{\text{Tukey}} = 0,048$; $\text{Co'd} = -0,710$) und zwischen Sichtbarkeits- und Interaktionsfehler ($\text{MD} = 0,778$; $p_{\text{Tukey}} = 0,050$; $\text{Co'd} = 0,751$). Für den Sichtbarkeitsfehler weisen die Differenzwerte negative Vorzeichen auf, die besagen, dass BI bzw. PE im Fehlerfall höher liegen als ohne Fehler. Ein analoger Effekt zeigt sich deskriptiv für die Dimensionen EE_Diff, HM_Diff und HT_Diff (vgl. Abbildung 53). Basierend darauf lässt sich vermuten, dass sich die vorgenannten Akzeptanzfaktoren durch den Sichtbarkeitsfehler im Fehlerfall steigern lassen. Wie in Abschnitt 4.4.4.1 (Absatz Sichtbarkeitsfehler) bereits erwähnt, wird vermutet, dass im Fehlerfall die tatsächliche Sichtbarkeit bzw. das Situationsbewusstsein der Probanden für diese spezifische Messaufgabe höher ist als ohne Fehler. Im Gegensatz dazu lassen die Ergebnisse der vorliegenden Studie den Schluss zu, dass die anderen beiden Fehlertypen BI und PE auf vergleichbarem Niveau liegen. Die inferenzstatistischen Ergebnisse des UTAUT2 decken sich mit den deskriptiven Ergebnissen der Leistungsfaktoren, konnten jedoch innerhalb dieser Studie nicht mit statistischer Signifikanz belegt werden. Dies stützt weiterhin die These, dass differenziertere Forschungsaktivitäten notwendig sind. Eine mögliche Erklärung besteht darin, dass die experimentelle Umsetzung des Sichtbarkeitsfehlers im konkreten Anwendungskontext unbeabsichtigt zu einer verbesserten, visuellen Situationswahrnehmung führte.

Bei den Beanspruchungsfaktoren des NASA-TLX zeigen sich für die UV Fehlertyp bei MD_Diff ($p = 0,034$; $\eta^2p = 0,118$), P_Diff ($p = 0,049$; $\eta^2p = 0,106$), F_Diff ($p = 0,015$; $\eta^2p = 0,144$) sowie der durchschnittlichen Beanspruchung ($p = 0,044$; $\eta^2p = 0,110$) signifikante Unterschiede. Entsprechend den Ergebnissen der Post-hoc-Tests zeigt sich, dass die MD_Diff ($\text{MD} = 25,95$; $p_{\text{Tukey}} = 0,027$; $\text{Co'd} = -0,856$) als eine psychologische Stressreaktion beim Interaktionsfehler signifikant höher ausfällt als beim Datenfehler. Ebenfalls konnten für P_Diff ($\text{MD} = 22,5$; $p = 0,05$), F_Diff ($\text{MD} = 33,1$; $p_{\text{Tukey}} = 0,012$) und für die durchschnittliche Beanspruchung ($\text{MD} = 16,02$; $p_{\text{Tukey}} = 0,048$) signifikante Unterschiede zwischen dem Interaktions- und Sichtbarkeitsfehler nachgewiesen werden. Das bedeutet, dass die wahrgenommene Leistung beim Sichtbarkeitsfehler größer ist als beim Interaktionsfehler, obwohl sich objektiv keine Unterschiede der Leistungsparameter nachweisen lassen. Außerdem sind die F und die durchschnittliche Beanspruchung beim Sichtbarkeitsfehler geringer als beim Interaktionsfehler. Auch diese Ergebnisse sind analog

zu den zuvor diskutierten AVs und stützen die vorherige Diskussion bezüglich des fehlerhaften Versuchsdesigns.

Für die physiologische Stressreaktion HR_Diff und SCL_Diff konnten keine statistisch signifikanten Unterschiede für die UV Fehlertyp nachgewiesen werden. Aufgrund der geringen statistischen Power 0,09 (SCL_Diff) bzw. 0,08 (HR_Diff) ist für diese Aussage jedoch keine statistische Validität gegeben. Ebenfalls lässt sich aus dem deskriptiven Vergleich keine Tendenz ableiten.

Insgesamt lassen die Ergebnisse vermuten, dass der Stichprobenumfang zu klein ausgelegt wurde, da für einige AVs die statistische Power $\ll 0,8$ ausfällt. Aus den deskriptiven Analysen lassen sich wie vorher diskutiert weitere Tendenzen ableiten, dass der Einfluss des Interaktionsfehlers größer ist als des Datenfehlers. Der Einfluss vom Sichtbarkeitsfehler wird am kleinsten vermutet. Es gilt, diesen in weiterführenden Forschungsaktivitäten zu untersuchen. Außerdem zeigen die Ergebnisse Unterschiede zwischen dem Sichtbarkeitsfehler und den anderen beiden Fehlertypen auf, obwohl objektiv kein Unterschied beim Leistungsverlust inferenzstatistisch nachgewiesen werden konnte. Diese Diskrepanz könnte mit dem fachlichen Anwendungskontext sowie gegebenenfalls mit einem Designfehler begründet werden (siehe oben).

4.4.4.3 Teilstudie 4.3 Akzeptanzmanipulation

Anknüpfend an den methodischen Aufbau des Forschungsdesigns sowie die Ergebnisse aus Teilstudie 4.1 wurde der Interaktionsfehler als größter, signifikanter Fehlereffekt mit einer Stärke von $Co'd = 0,529$ auf BI im Vergleich zum Daten- und Sichtbarkeitsfehler identifiziert. Die Ergebnisse des unabhängigen t -Tests zeigen einen signifikanten Unterschied ($t(22) = 2,38$; $p = 0,026$) zwischen den Bedingungen mit und ohne Konfidenzanzeige der Systemleistung. Mit einer mittleren Differenz von 0,889 und einer Effektstärke ($Co'd$) von 1,12 deutet die Analyse auf eine substantielle Beeinflussung der BI_Diff durch die Konfidenzanzeige hin. Die Ergebnisse legen nahe, dass die Konfidenzanzeige ein wirksames Mittel zur Beeinflussung der Akzeptanz sein kann. Der Vergleich der Mittelwerte zwischen den beiden Faktorstufen zeigt, dass die BI-Differenz in der Bedingung ohne Konfidenzanzeige ($M = 0,389$; $SD = 0,734$) höher ausfiel als in der Bedingung mit Konfidenzanzeige ($M = -0,500$; $SD = 0,961$). Dies kann als Hinweis

interpretiert werden, dass die Konfidenzanzeige die Wahrnehmung der Interaktion beeinflusst, und möglicherweise zu einer höheren Akzeptanz führt. Ein möglicher Erklärungsansatz hierfür könnte sein, dass Nutzer aufgrund der höheren Transparenz durch die Konfidenzanzeige dem IAS sowie dessen Systemleistung stärker vertrauen (Konopka et al., 2024b).

Die Erfüllung der Voraussetzungen der Normalverteilung und Varianzhomogenität verleiht den Ergebnissen methodische Robustheit. Gleichzeitig sollte die ungleiche Gruppengröße zwischen den Bedingungen ($n = 18$ vs. $n = 6$) kritisch betrachtet werden, da dies die statistische Power und die Generalisierbarkeit der Ergebnisse beeinflussen könnte. Die berechnete Power von 0,74 liegt zwar im akzeptablen Bereich, lässt jedoch Raum für Verbesserungen durch größere Stichproben (Döring & Bortz, 2016). Die Ergebnisse legen nahe, dass die Systemkonfidenzanzeige eine wertvolle Komponente für die Gestaltung von Systemen sein kann, um Akzeptanzprobleme aufgrund von Systemfehlern zu adressieren. Jedoch sollte die Implementierung solcher Anzeigen sorgfältig getestet werden, um unerwünschte Nebeneffekte wie die potenzielle Verstärkung von F bei einer unzutreffenden Konfidenzanzeige der Systemvalidität zu vermeiden. Die Ergebnisse stützen die Aussagen der Publikation von Richardson et al. (2018), in der die Akzeptanz der HMI-Konzepte, analog gemessen durch die Dimensionen Nützlichkeit und Zufriedenheit, durch die Bereitstellung zusätzlicher Informationen signifikant positiv beeinflusst wurde.

Im Gegensatz dazu konnten in der Studie von Hecht et al. (2020) keine signifikanten Effekte aufgrund zusätzlicher HMI-Informationen auf die Akzeptanz nachgewiesen werden ($F(1;31) = 0,16$, $p = 0,695$). Die vorliegende Studie liefert neue Erkenntnisse zum Einfluss transparenter Systemrückmeldungen auf die Akzeptanz intelligenter Systeme, insbesondere durch die Einführung von Konfidenzanzeigen. Zentrales Ergebnis ist, dass Transparenz in Form expliziter Unsicherheitskommunikation nicht nur bei fehlerhaften Daten oder unzutreffenden Empfehlungen positiv wirkt, wie frühere Studien nahelegen (z. B. Gegoff et al., 2023; Konopka et al., 2024b; Schaefer et al., 2016), sondern auch bei Interaktionsfehlern eine signifikant akzeptanzsteigernde Wirkung entfaltet. Dies spricht dafür, dass Nutzer bereit sind, Fehler zu tolerieren, sofern das System transparent über seine Grenzen informiert. Die Ergebnisse bestätigen somit nicht nur die Bedeutung von

Transparenz für Vertrauen und Akzeptanz, wie sie Schaefer et al. (2016) und Wanner et al. (2022) betonen, sondern erweitern diese Perspektive auf bislang weniger betrachtete Fehlertypen in der Mensch-Technik-Interaktion.

Aus dieser Diskussion lassen sich zwei relevante Forschungsfragen hinsichtlich der Systemkonfidenzanzeige ableiten. Erstens: Wie muss eine solche Anzeige innerhalb des Nutzungskontexts IAS gestaltet sein, um einen positiven Beitrag zur BI zu gewährleisten? Zweitens: Für welchen Fehlertyp ist die Anzeige der Systemkonfidenz wirksam bzw. wie groß ist ihre Effektstärke im industriellen Umfeld?

Eine wesentliche Limitation der Teilstudie ist die kleine Stichprobengröße insbesondere in der Bedingung mit Konfidenzanzeige ($n = 6$). Dies schränkt die Aussagekraft der Ergebnisse ein und sollte in zukünftigen Studien adressiert werden. Da vor Beginn der Studie nicht vorhersehbar war, welcher Fehlertyp den stärksten Einfluss auf die BI ausüben würde, erfolgte zunächst eine explorative Auswertung der Teilstudie 4.1. Auf Grundlage dieser Ergebnisse wurde der Interaktionsfehler für die nachfolgende Untersuchung der Systemkonfidenzanzeige ausgewählt. Im Weiteren wurde keine a-priori Stichprobenumfangsplanung durchgeführt. Stattdessen erfolgte eine studienbegleitende Auswertung mit dem Ziel, eine statistische Power von 0,8 zu erreichen. Die tatsächlich erreichte Teststärke lag jedoch nur bei 0,74. Aufgrund von Verständnisproblemen bei der Versuchsdurchführung hinsichtlich der Konfidenzanzeige musste ein Versuchsdurchlauf verworfen werden.

5 Zusammenfassung

Nachdem in Kapitel 4 die empirischen Studien von der Methodik über die Ergebnisdarstellung und -interpretation bis hin zu deren Diskussion umfangreich präsentiert wurden, gilt es in den folgenden Unterkapiteln, die gewonnenen Erkenntnisse der einzelnen Hypothesenbewertungen zusammenzufassen und gegen die Forschungsfragen zu rekurrieren. Weiter gilt es, die konkreten Implikationen sowie Limitationen dieser Arbeit herauszustellen und den weiteren Forschungsbedarf abzuleiten.

5.1 Rekurreren der Erkenntnisse

In Tabelle 54 werden die durchgeführten Hypothesenbewertungen nachfolgend zusammengetragen und im Kontext des übergeordneten Forschungsdesigns erörtert.

Hypothese	Fehler	UV	Prüfung	Studie	AV	Co'd/BRK	Power	n
<i>H_{B.1.1}</i>	<i>F₂</i>	VQ	ja	1	Leistung	1,15/0,500	0,73	18
<i>H_{B.2.1}</i>			nein		BI	0,063	0,05	
<i>H_{B.3.1}</i>			nein		Mean_NASA	0,189	0,10	
<i>H_{C.1.1}</i>	<i>F₃</i>	IQ	ja	2	WER	0,884	0,99	22
<i>H_{C.1.1}</i>			a		ITR	0,880	0,99	
<i>H_{C.2.1}</i>			ja		BI	1,550	0,97	
<i>H_{C.3.1}</i>			Ja		Mean_NASA	1,12/0,490	0,71	
<i>H_{D.1.1}</i>	<i>F₄</i>	TQ	ja	3	Leistung	1,828	1,00	52
<i>H_{D.2.1}</i>			ja		BI	1,174	1,00	
<i>H_{D.3.1}</i>			ja		Mean_NASA	1,411	1,00	
<i>H_{A.1.1}</i>	<i>F₁</i>	DQ	ja	4.1	Leistung	0,686	0,92	21
<i>H_{A.2.1}</i>			ja		BI	0,527	0,75	
<i>H_{A.3.1}</i>			nein		Mean_NASA	-0,026	0,06	
<i>H_{C.1.1}</i>	<i>F₃</i>	IQ	ja		Leistung	1,080	1,00	19
<i>H_{C.2.1}</i>			ja		BI	0,529	0,69	
<i>H_{C.3.1}</i>			ja		Mean_NASA	0,770	0,93	
<i>H_{B.1.1}</i>	<i>F₂</i>	VQ	ja		Leistung	0,721	0,94	18
<i>H_{B.2.1}</i>			nein		BI	-0,299	0,35	
<i>H_{B.3.1}</i>			nein		Mean_NASA	-0,103	0,11	
<i>H_{1.1}</i>	<i>F₁₋₃</i>	FT	nein	4.2	Leistung_Diff	0,216	0,19	58
<i>H_{2.1}</i>			ja		BI_Diff	0,854	0,99	
<i>H_{3.1}</i>			ja		Mean_NASA_Diff	0,783	0,74	
<i>H_{4.1}</i>	<i>F₃</i>	M/IQ	ja	4.3	BI_Diff	1,120	0,74	24

Fußnote. Mean_NASA = Mittelwert der NASA-TLX-Dimensionen; * BRK = Biseraler

Rangkorrelationskoeffizient für nichtparametrische Hypothesentests aufgrund Verletzung der Normalverteilungsannahme; bei mehreren Faktorstufen wird der stärkste Effekt angegeben; FT = Fehlertyp

Tabelle 54. Gesamtüberblick der Hypothesenbewertungen

Zur Beantwortung der Forschungsfragen werden die einzelnen Hypothesen sowie deren Bewertungen hinsichtlich der Ergebnisse der einzelnen Studien herangezogen. Aussagekräftig lassen sich die Bewertungen über die jeweilige standardisierten Effektstärken nach $Co'd$, die statistische Teststärke (Power) als auch über die statistische Signifikanz je Primärfaktor des Index Leistung, BI und des Mean_NASA-TLX zusammenfassen. Flankierend werden die Ergebnisse der weiteren sekundären Faktoren des UTAUT2, NASA-TLX sowie der physiologischen Parameter präsentiert. Sofern nicht explizit angegeben, impliziert ein positiver Effekt eine erwartete negative Fehlerauswirkung. Negative Effekte deuten auf atypische Ergebnisse hin, was bedeutet, dass ein Fehler die AV oder den jeweiligen Faktor entgegen der Logik beeinflusst. Subsumierend wird für jede UV der paarweise Vergleich der Post-hoc-Tests mit der höchsten Effektstärke präsentiert. Verallgemeinerungen sind nur bei statistisch signifikanten Effekten möglich und können in Abhängigkeit der Teststärke als valide angenommen werden. Teilergebnisse ohne statistische Signifikanz weisen darauf hin, dass die Effekte auf die konkrete Stichprobe beschränkt sein und zufällig auftreten können, wodurch kein valider Einfluss der jeweiligen UV auf die AV nachgewiesen werden konnte. Dies sollte bei der nachfolgenden Beantwortung der Forschungsfragen stets berücksichtigt werden:

FF_{1.A}: Welchen Einfluss haben fehlerhafte Daten innerhalb der MMI auf die Leistung, Akzeptanz und die Beanspruchung bei IAS?

Antwort: In Studie 4.1. wurde die UV DQ hinsichtlich der beiden Faktorstufen ff und fh untersucht. Aus dem Netzdiagramm (ND) Abbildung 57 geht hervor, dass der Datenfehler einen statistisch signifikanten Einfluss auf die Leistung (mittlere Effektstärke $Co'd = 0,69$) sowie auf die Akzeptanz mit mittlerer Effektstärke ($Co'd = 0,53$) und hohen Teststärken (zwischen 92 % bzw. 75 %) hat, jedoch keinen signifikanten Einfluss auf die Beanspruchung bei sehr kleiner Effekt- und Teststärke ($Co'd = -0,03$; Power = 0,06).

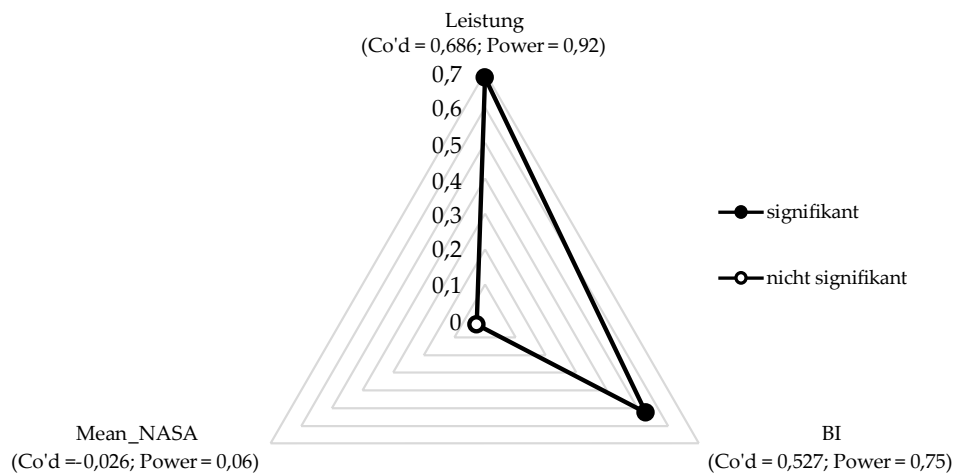


Abbildung 57. ND-Effektstärken des Datenfehlers auf die PF

Begleitend werden die sekundären Einflüsse des Datenfehlers auf die Akzeptanzfaktoren in Abbildung 58, auf die Beanspruchungsfaktoren des NASA-TLX in Abbildung 59 und auf die physiologischen Parameter in Abbildung 60 dargestellt.

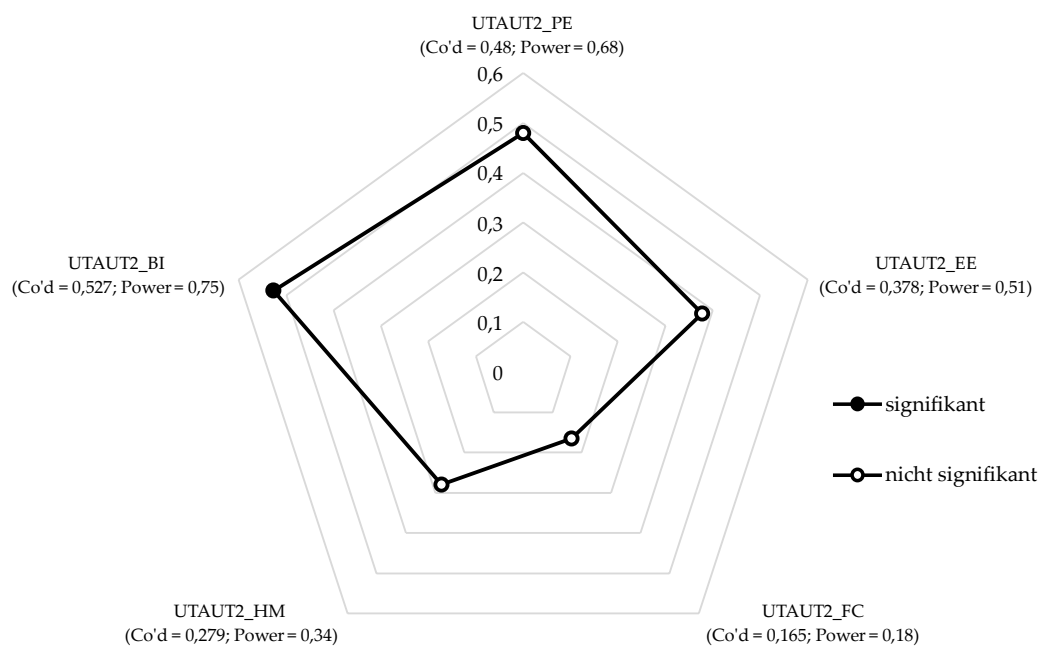


Abbildung 58. ND-Effektstärken des Datenfehlers auf die AF

Außer bei BI zeigt der Datenfehler keinen statistisch signifikanten Einfluss auf weitere Akzeptanzfaktoren. Für PE ergibt die Stichprobe die Tendenz einer mittleren Effektstärke und für EE eine kleine Effektstärke; beide mit einer mittleren Teststärke von 68 % respektive von 51 %. Für HM und FC konnten nur kleine Effekte mit geringer statistischer Aussagekraft (Power < 0,39) nachgewiesen werden.

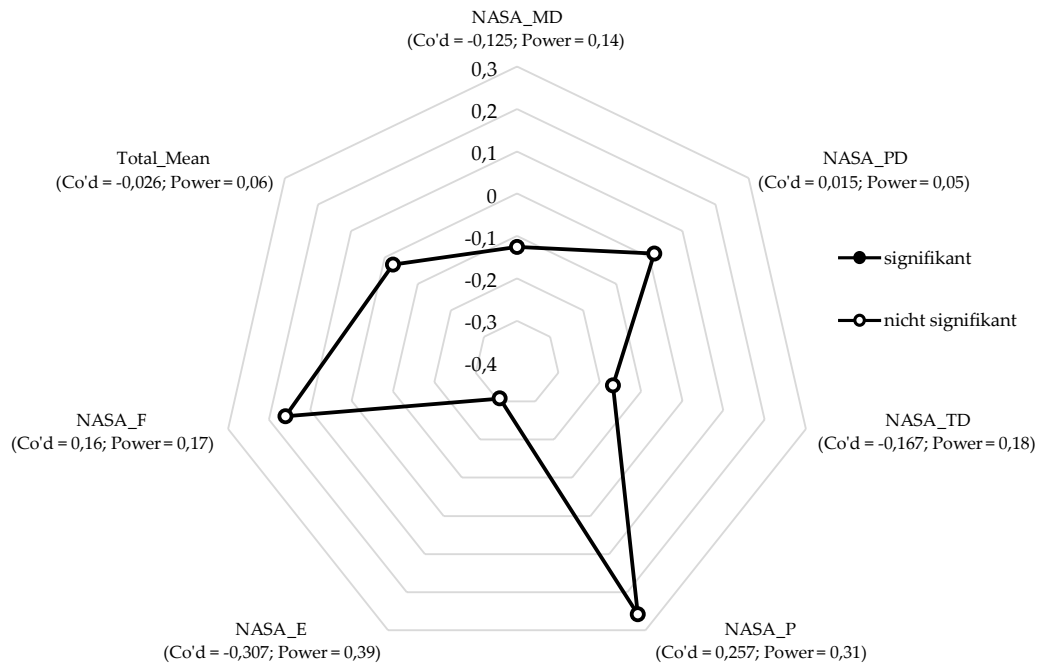


Abbildung 59. ND-Effektstärken des Datenfehlers auf die BF

Entsprechend konnten keine statistisch signifikanten Effekte durch den Datenfehler auf die Beanspruchungsfaktoren nachgewiesen werden. Die Effektstärken liegen hierbei zwischen 0,26 und -0,30, jedoch bei sehr geringen Teststärken (Power < 0,39).

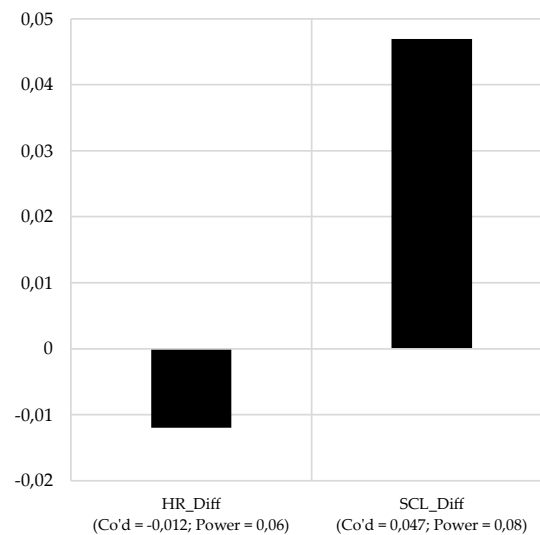


Abbildung 60. BD-Effektstärken des Datenfehlers auf die PP

Ebenfalls konnten nur sehr kleine Effekte ohne statistische Signifikanz des Datenfehlers auf die physiologischen Parameter nachgewiesen werden. Die Teststärken belaufen sich hierbei auf 6 % beziehungsweise 8 %.

FF_{1.B}: Welchen Einfluss hat eine fehlerhafte Visualisierung innerhalb der MMI auf die Leistung, Akzeptanz und die Beanspruchung bei IAS?

Antwort: In den Studien 1 (VR) und 4.1 (AR) wurde die UV VQ in Bezug auf die beiden Faktorstufen ff und fh untersucht. Abbildung 61 zeigt, dass der Sichtbarkeitsfehler statistisch signifikante Einflüsse auf die Leistung mit hoher (VR) bzw. mittlerer Effektstärke (AR) bei hoher Teststärke hat, jedoch keine statistisch signifikanten Einflüsse auf die Akzeptanz und die Beanspruchung. Die Ergebnisse der beiden Studien zeigen außerdem für die Leistung, dass die Effektstärke aufgrund der VQ für VR bei $Co'd = 1,15$ liegt, im Gegensatz zu AR mit $Co'd = 0,72$. Für die Akzeptanz ist der Vergleich der Effektstärken sehr auffällig, da die $Co'd$ bei AR negativ wird: $Co'd$ beträgt 0,06 für VR und $-0,30$ für AR.

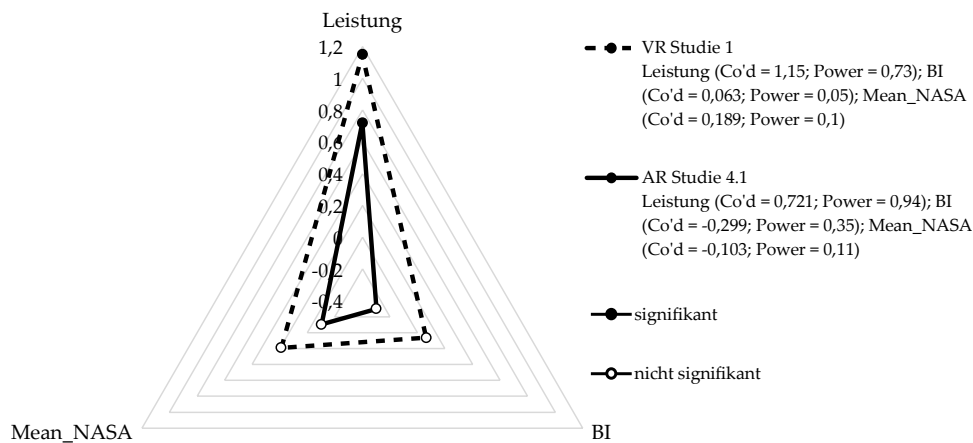


Abbildung 61. ND-Effektstärken des Sichtbarkeitsfehlers auf die PF

Die Ergebnisse der mittleren Beanspruchung weisen nur sehr kleine, aber für AR eine negative Effektstärke (VR $Co'd = 0,19$ und AR $Co'd = -0,10$) auf. Die Teststärken ($\approx 10\%$) fallen innerhalb der Stichproben ebenfalls sehr gering aus. Es werden die sekundären Einflüsse des Sichtbarkeitsfehlers auf die Akzeptanzfaktoren (siehe Abbildung 62), auf die Beanspruchungsfaktoren des NASA-TLX (siehe Abbildung 63) sowie auf die physiologischen Parameter (siehe Abbildung 64) präsentiert.

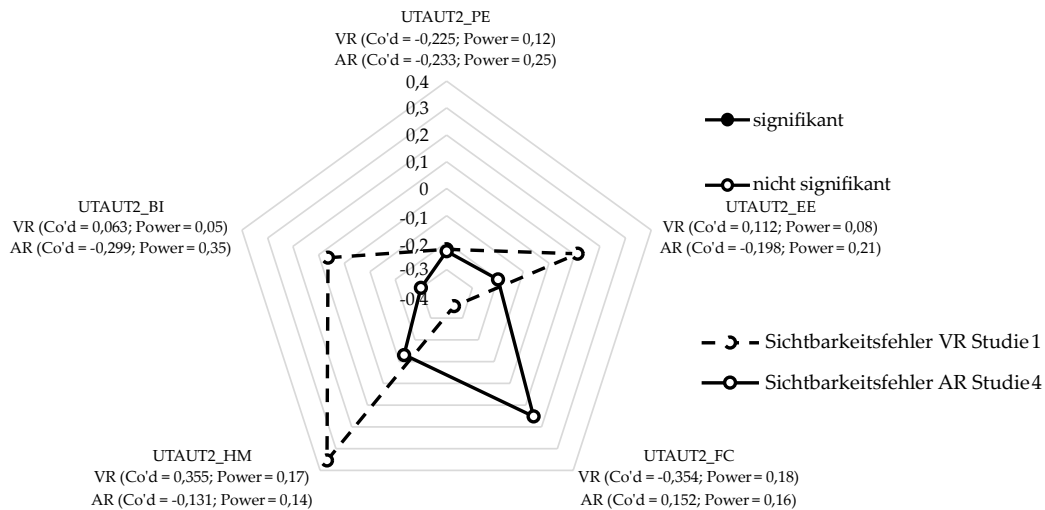


Abbildung 62. ND-Effektstärken des Sichtbarkeitsfehlers auf die AF

Der Sichtbarkeitsfehler zeigt für keine Dimension der Akzeptanzfaktoren einen statistisch signifikanten Einfluss. Betrachtet man zunächst die Beträge der Effektstärken, ergeben sich in der VR-Studie für die Dimensionen HM (Co'd = 0,35) und FC (Co'd = 0,36) die größten Effekte, während in der AR-Studie alle Effektstärken mit maximal Co'd = 0,30 (BI) gering ausfallen. Insgesamt zeigen beide Studien überwiegend kleine Effektstärken und eine geringe Teststärke (Power = 5–35 %), sodass die Ergebnisse nur eingeschränkt belastbar sind. Hinsichtlich der Effektrichtung ergibt sich kein konsistentes Muster zwischen VR und AR. Die Dimensionen PE weisen in beiden Studien negative Effektstärken auf. Dagegen ändern EE, FC, HM und BI ihre Effektrichtung zwischen den Studien: Während EE, HM und BI in der VR-Studie positive Effektstärken zeigen, fallen diese in der AR-Studie negativ aus. Umgekehrt weist FC in der VR-Studie eine negative und in der AR-Studie eine positive Effektstärke auf.

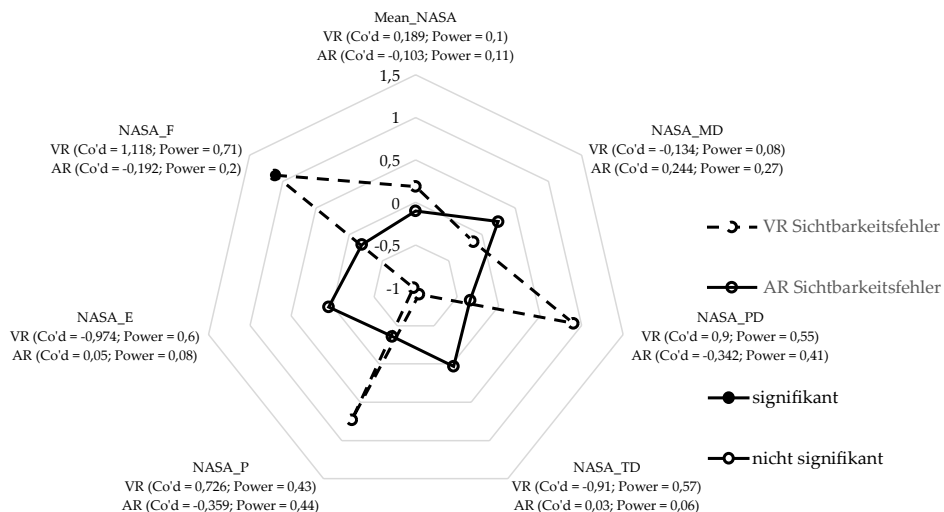
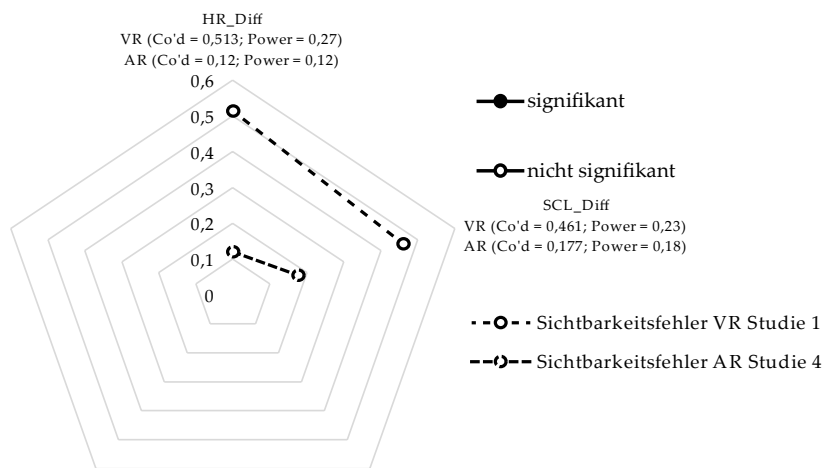


Abbildung 63. ND-Effektstärken des Sichtbarkeitsfehlers auf die BF

Hinsichtlich der Beanspruchungsfaktoren konnte für den Sichtbarkeitsfehler lediglich für die Dimension F in VR ein signifikanter Einfluss mit hoher Effektstärke ($Co'd = 1,12$) und ausreichender Teststärke (Power = 0,71) nachgewiesen werden. Für alle übrigen NASA-TLX-Dimensionen zeigten sich in VR und AR lediglich kleine Effektstärken bei geringer Teststärke (Power = 6–44 %). Auffällig sind die teilweise gegenläufigen Effektrichtungen zwischen VR und AR, beispielsweise für MD (VR: $Co'd = -0,13$; AR: $Co'd = 0,24$) und PD (VR: $Co'd = 0,90$; AR: $Co'd = -0,34$). Analog zu den Akzeptanzfaktoren ergeben sich damit keine konsistenten Reaktionsmuster der Beanspruchung infolge des Sichtbarkeitsfehlers

**Abbildung 64. ND-Effektstärken des Sichtbarkeitsfehlers auf die PP**

Hinsichtlich der physiologischen Parameter konnten ohne statistische Signifikanz mittlere Effekte bei VR und kleine Effekte bei AR nachgewiesen werden. Bei diesen Aussagen liegt hierbei die Teststärke jedoch nur zwischen 12 % und 27 %.

FF_{1.c}: Welchen Einfluss hat eine fehlerhafte Interaktion innerhalb der MMI auf die Leistung, Akzeptanz und die Beanspruchung bei IAS?

Antwort: In den Studien 2 (Spracheingabe) und 4.1 (Gestik) wurde die UV IQ in Bezug auf die beiden Faktorstufen ff und fh untersucht. Abbildung 65 zeigt, dass der Interaktionsfehler bei der Gesteninteraktion statistisch signifikante Einflüsse auf alle drei Primärfaktoren hat. Auf den Index Leistung wurde eine hohe Effektstärke ($Co'd = 1,08$) und Teststärke (Power = 1) nachgewiesen; wobei sich für die durchschnittliche

Beanspruchung ($Co'd = 0,77$) und die Akzeptanz ($Co'd = 0,53$) nur mittlere Effekte mit hohen Teststärken mit 93 % und 69 % ergeben.

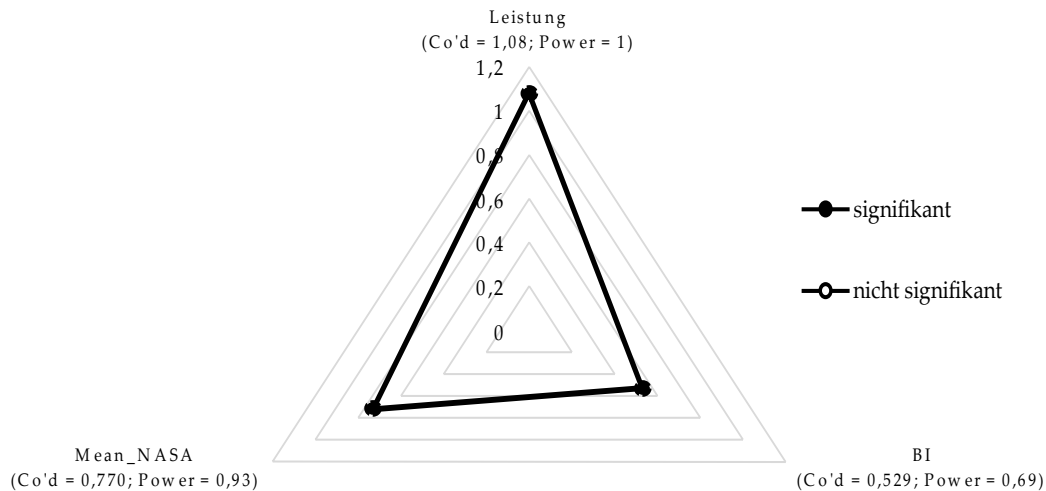


Abbildung 65. ND-Effektstärken des Interaktionsfehlers (Gestik) auf die PF

Abbildung 66 zeigt, dass der Interaktionsfehler bei der Sprachinteraktion ebenfalls statistisch signifikante Einflüsse auf die Primärfaktoren ausprägt. Für die spezifischen Leistungsfaktoren WER ($Co'd = 0,88$) und ITR ($Co'd = 0,88$) sowie für die durchschnittliche Beanspruchung ($Co'd = 1,12$) und Akzeptanz ($Co'd = 1,55$) zeigen sich hohe Effekt- und Teststärken (Power $\geq 0,71$).

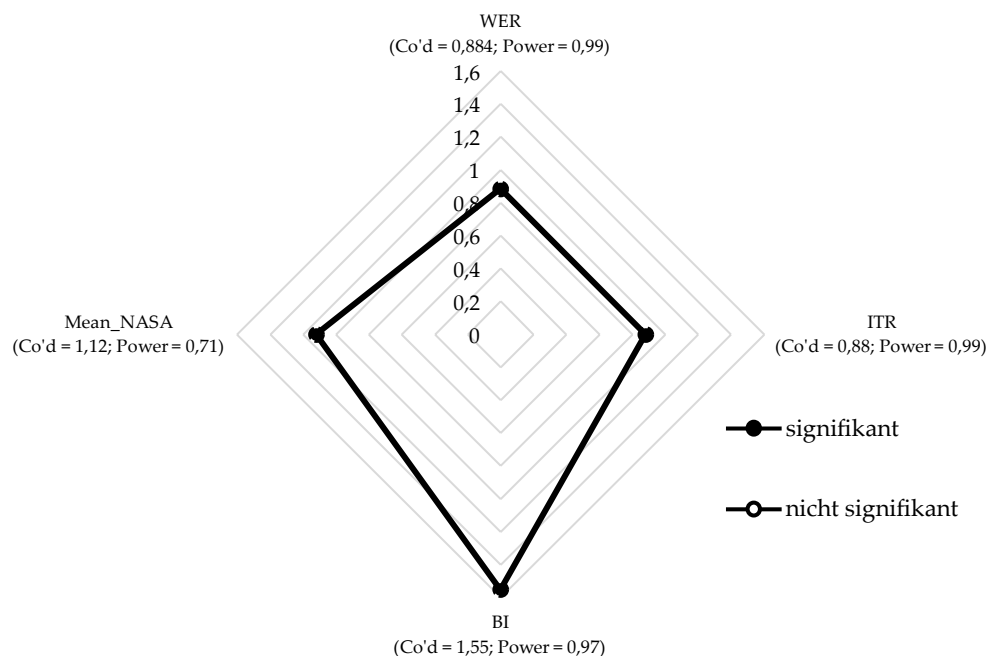


Abbildung 66. ND-Effektstärken des Interaktionsfehlers (ASR) auf die PF

Die Ergebnisse der Studie 2 zeigen, dass der Interaktionsfehler bei der ASR für alle Akzeptanzfaktoren des UTAUT2 (siehe Abbildung 67) bis auf FC und PE signifikante Einflüsse mit durchweg hohen Effektstärken hoher statistischer Validität aufweist. Hinsichtlich der Gesteninteraktion konnte durch den Interaktionsfehler lediglich für BI ein mittlerer Effekt ($Co'd = 0,53$) mit hoher Power nachgewiesen werden. Für alle anderen Akzeptanzfaktoren ergaben sich innerhalb der Studie geringe Effekt- und Teststärken (durchschnittlich $Co'd = 0,16$ bei $Power \leq 0,29$).

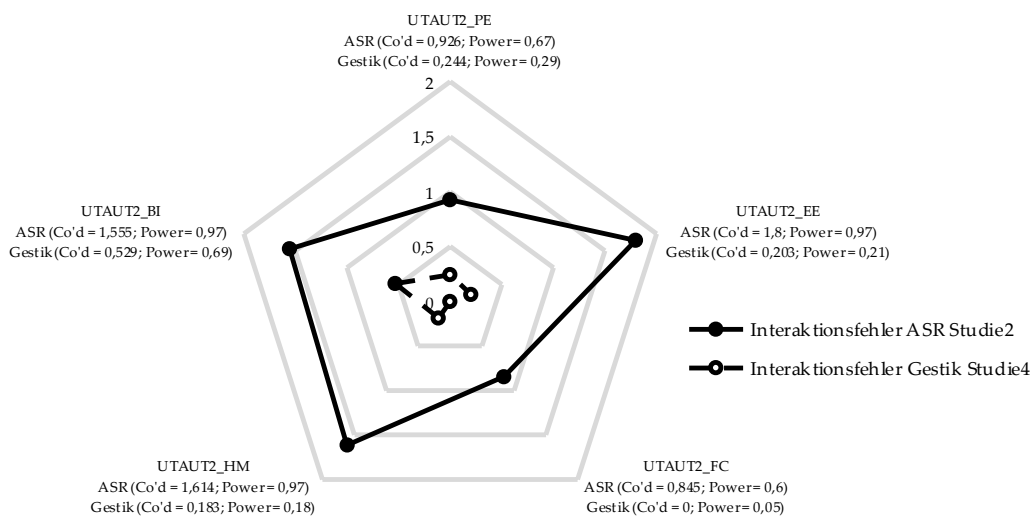


Abbildung 67. ND-Effektstärken des Interaktionsfehlers auf die AF

Die Ergebnisse (vgl. Abbildung 68) zeigen, dass Interaktionsfehler bei der ASR und der Gestensteuerung ähnliche Einflüsse auf die Beanspruchungsfaktoren haben. Für beide Interaktionskonzepte weisen die durchschnittliche Beanspruchung und F signifikante Effekte mit hohen Effektstärken und Teststärken ($Co'd \geq 0,8$; $Power \geq 0,71$) auf. Für die MD zeigten beide Interaktionsarten ebenfalls signifikante Effekte mit hohen Effektstärken, jedoch bei ASR nur mit mittlerer Teststärke ($Power = 0,55$). Die übrigen Dimensionen wie PD und TD wiesen in beiden Bedingungen keine signifikanten Ergebnisse mit nur geringen Effekt- und Teststärken auf. Besonders auffällig ist, dass die Dimension E in beiden Interaktionsarten keine signifikanten Einflüsse mit nur sehr kleinen Effekt- und Teststärken zeigten.

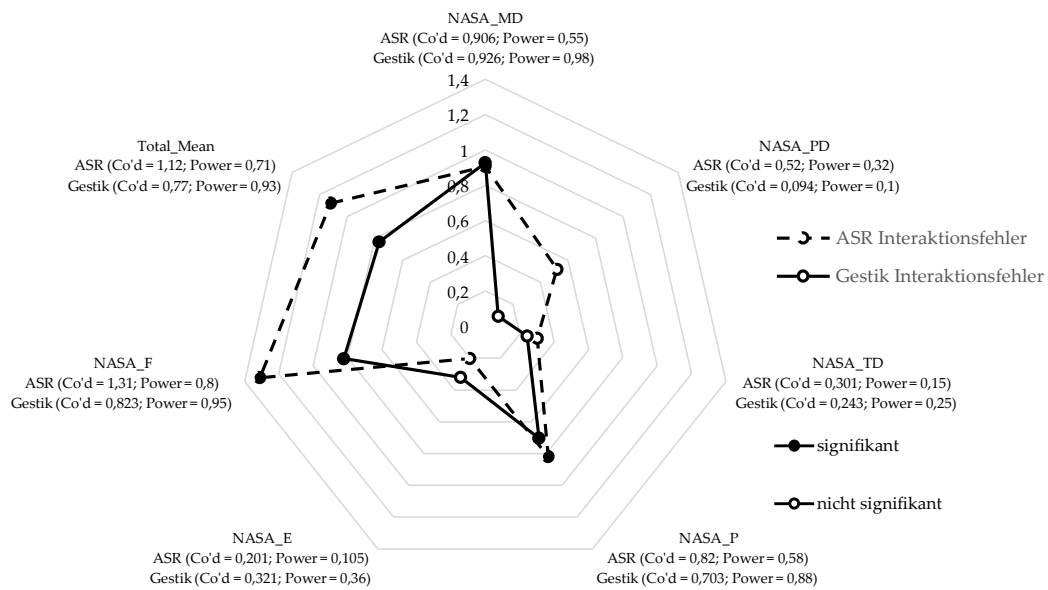


Abbildung 68. ND-Effektstärken des Interaktionsfehlers auf die BF

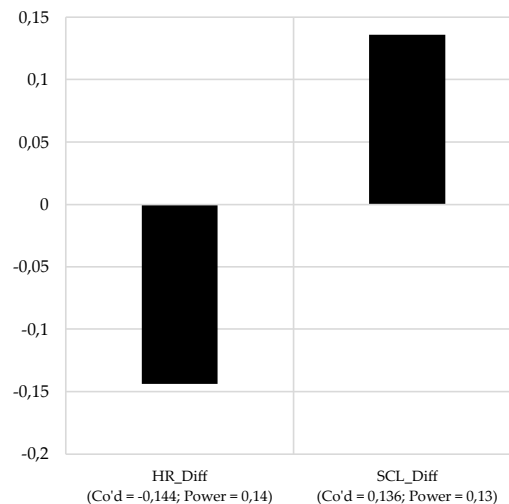


Abbildung 69. BD-Effektstärken des Interaktionsfehlers (Gestik) auf die PP

Hinsichtlich der physiologischen Parameter (vgl. Abbildung 69) konnten nur für den Gesteninteraktionsfehler kleine Effekte ohne statistische Signifikanz nachgewiesen werden. Bei diesen Aussagen liegt er bei sehr kleinen Teststärken ($\text{Power} \leq 0,14$). Aufgrund eines Fehlers im DOE war für ASR die Erhebung der physiologischen Parameter ungültig, sodass die Werte verworfen wurden.

FF_{1.D}: Welchen Einfluss hat ein unzureichendes Trainingsszenario innerhalb der MMI auf die Leistung, Akzeptanz und die Beanspruchung bei IAS?

Antwort: In Studie 3 wurde die UV TQ hinsichtlich der Faktorstufe MT im Vergleich zu den anderen beiden (ff) untersucht. Aus Abbildung 70 geht hervor, dass der

Trainingsfehler einen statistisch signifikanten Einfluss auf alle drei Primärfaktoren hat. Aufgrund des Trainingsfehlers konnten für Leistung, Akzeptanz und die durchschnittliche Beanspruchung sehr hohe Effektstärken ($Co'd \geq 1,17$) bei 100 % Teststärke nachgewiesen werden.

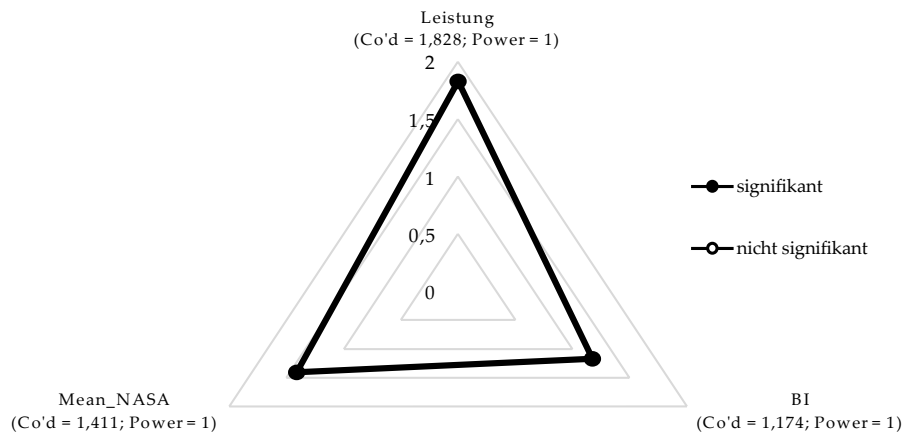


Abbildung 70. ND-Effektstärken des Trainingsfehlers auf die PF

Die Ergebnisse (siehe Abbildung 71) zeigen, dass der Trainingsfehler die Akzeptanzdimensionen PE, FC, HM und BI des UTAUT2 durchweg signifikant mit hohen Effektstärken und maximaler statistischer Power $\geq 0,99$ beeinflusst. Es wurden Effektstärken über alle Akzeptanzfaktoren hinweg mit $Co'd$ zwischen 0,90 und 1,17 berechnet.

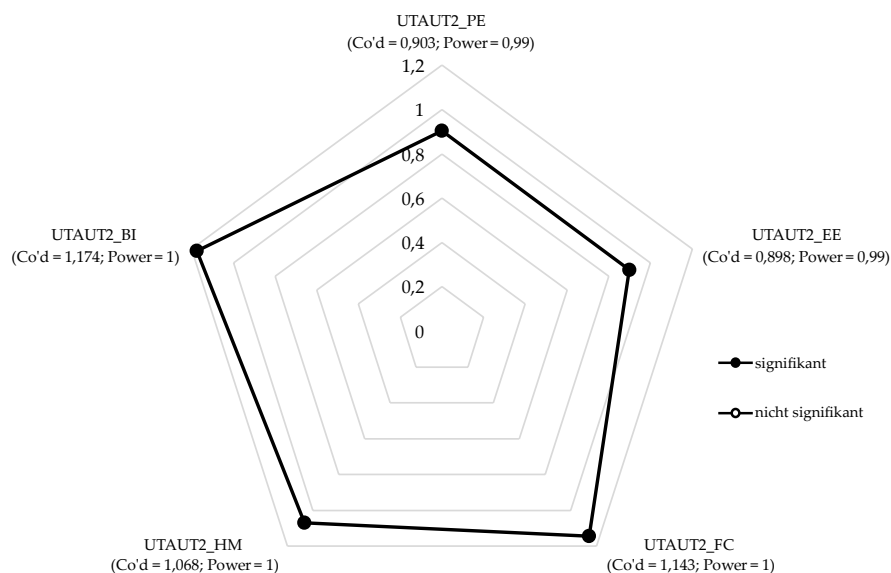


Abbildung 71. ND-Effektstärken des Trainingsfehlers auf die AF

Die Ergebnisse in Abbildung 72 zeigen, dass der Trainingsfehler die Dimensionen durchschnittliche Beanspruchung, F, P und MD des NASA-TLX signifikant und mit hohen Effektstärken ($Co'd \geq 0,85$) und Teststärken ≥ 97 % beaufschlagt. Für die

Dimensionen PD und TD wurden hingegen geringere Effektstärken ($Co'd \leq 0,3$) ohne statistische Signifikanz und mit einer Power von $\leq 31\%$ aufgezeigt. Für E konnte zwar ein beinahe hoher Effekt mit $Co'd = 0,77$ bei einer Teststärke von 97% nachgewiesen werden, jedoch erreichte der Post-hoc-Test keine statistische Signifikanz.

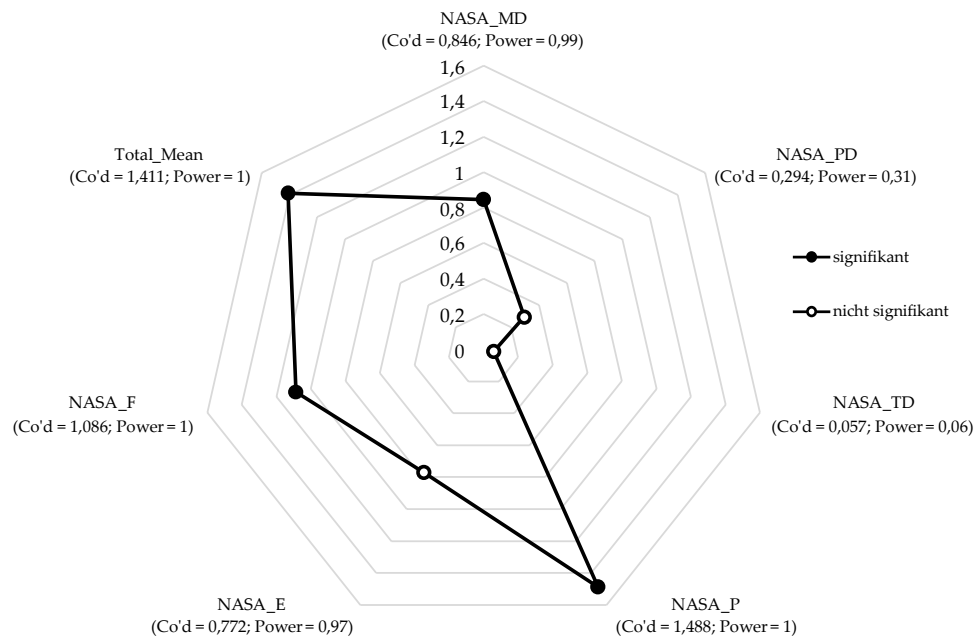


Abbildung 72. ND-Effektstärken des Trainingsfehlers auf die BF

Hinsichtlich der physiologischen Parameter (vgl. Abbildung 73) konnten für den Trainingsfehler relativ große Effektstärken ($Co'd \geq 0,80$) mit hoher statistischer Validität ($Power \geq 0,98$) nachgewiesen werden. Ein statistisch signifikanter Unterschied wurde jedoch ausschließlich für HR_Diff festgestellt, während SCL_Diff trotz hoher Effektstärke im Post-hoc-Test nicht signifikant wurde.

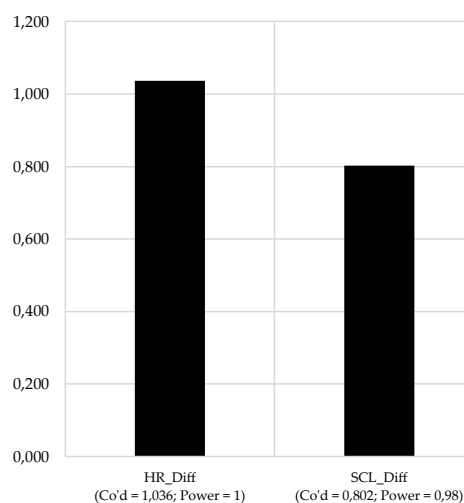


Abbildung 73. BD-Effektstärken des Trainingsfehlers auf die PP

FF₂: Manipulieren unterschiedliche Fehlerarten bei IAS innerhalb der MMI die Leistungen, Akzeptanz und Belastungen unterschiedlich?

Antwort: In Studie 4.2 wurde der Einfluss der UV Fehlertyp hinsichtlich der drei Faktorstufen Daten-, Sichtbarkeits- und Interaktionsfehler untersucht. Aus Abbildung 74 geht hervor, dass nur für die durchschnittliche Beanspruchung und zwischen Sichtbarkeitsfehler und Interaktionsfehler ein signifikanter Unterschied mit einer mittleren Effektstärke von $Co'd = 0,78$ bei einer Teststärke von 64 % nachgewiesen wurde. Für die anderen beiden Vergleiche zwischen Sichtbarkeitsfehler und Datenfehler ($Co'd = 0,68$; Power = 0,54) sowie zwischen Interaktionsfehler und Datenfehler ($Co'd = -0,10$; Power = 0,06) wurden trotz mittlerer bzw. geringer Effektstärken keine statistisch signifikanten Unterschiede festgestellt. Für die AV Leistung_Diff konnten keine signifikanten Unterschiede zwischen den Bedingungen der UV Fehlertyp festgestellt werden. Die Effektstärken und Teststärken fielen insgesamt niedrig aus, wobei die höchste Differenz zwischen Interaktionsfehler und Datenfehler ($Co'd = 0,49$; Power = 0,41) zu beobachten war. Zwischen Sichtbarkeitsfehler und Interaktionsfehler ($Co'd = -0,24$; Power = 0,17) sowie zwischen Sichtbarkeitsfehler und Datenfehler ($Co'd = 0,12$; Power = 0,1) waren die Effektstärken noch geringer und wiesen eine sehr niedrige Teststärke auf. Für die Akzeptanz (BI_Diff) zeigt sich ein signifikanter Unterschied zwischen Interaktionsfehler und Sichtbarkeitsfehler mit einer hohen Effektstärke sowie Teststärke ($Co'd = 0,77$; Power = 0,97). Zwischen Sichtbarkeitsfehler und Datenfehler konnte ebenfalls ein signifikanter Unterschied mit einer hohen Effektstärke und hoher statistischer Validität nachgewiesen werden ($Co'd = -0,85$; Power = 0,99). Für den Vergleich zwischen Interaktionsfehler und Datenfehler wurde hingegen kein signifikanter Unterschied festgestellt ($Co'd = -0,08$; Power = 0,07). Die Effektstärke als auch die Teststärke fallen hierbei sehr niedrig aus.

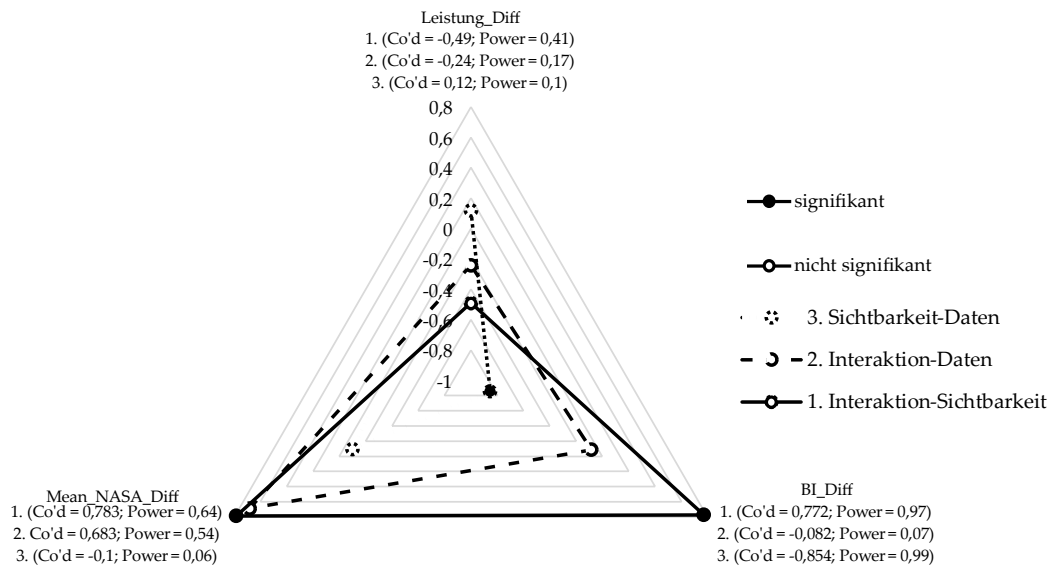


Abbildung 74. ND-Effektstärken des Fehlervergleichs auf die PF

Die Ergebnisse der Akzeptanzfaktoren aus Abbildung 75 zeigen, dass nur für die BI und die PE teilweise signifikante Unterschiede zwischen den Fehlertypen vorliegen. Zwischen Interaktionsfehler und Sichtbarkeitsfehler sowie zwischen Sichtbarkeitsfehler und Datenfehler traten bei beiden Dimensionen signifikante Unterschiede mit hohen Effektstärken und hoher Teststärke auf, was auf deutliche Unterschiede zwischen diesen Bedingungen hinweist. Der Vergleich zwischen Interaktionsfehler und Datenfehler sowohl bei BI als auch bei PE zeigte hingegen eine geringe Effektstärke mit niedriger Teststärke und war nicht signifikant. Für die EE wurde zwischen Interaktionsfehler und Sichtbarkeitsfehler eine mittlere Effektstärke mit mittlerer Teststärke ohne statistische Signifikanz festgestellt. Zwischen Interaktionsfehler und Datenfehler hingegen zeigte sich eine geringe Effektstärke mit niedriger Teststärke, während zwischen Sichtbarkeitsfehler und Datenfehler eine mittlere Effektstärke mit mittlerer Teststärke vorlag. Beide Vergleiche waren nicht signifikant. Für HM und FC waren die Vergleiche zwischen Interaktionsfehler und Sichtbarkeitsfehler, Interaktionsfehler und Datenfehler sowie Sichtbarkeitsfehler und Datenfehler nicht signifikant. Außerdem blieben alle Effektstärken gering und es wurde keine ausreichend hohe Teststärke erreicht.

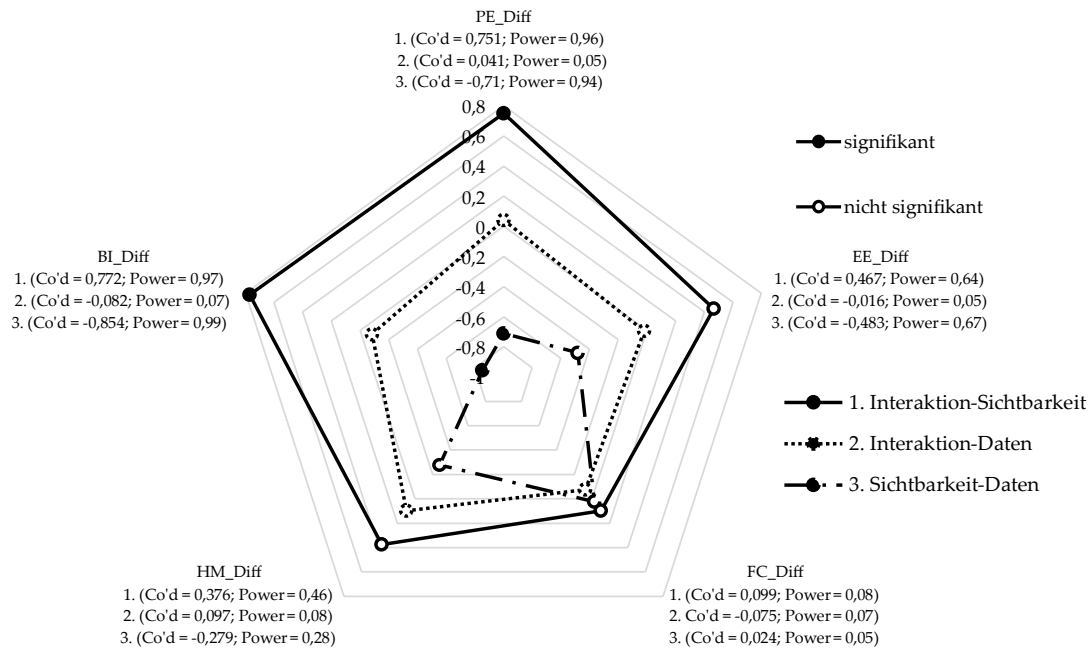


Abbildung 75. ND-Effektstärken des Fehlervergleichs auf die AF

Die Ergebnisse der Beanspruchungsfaktoren aus Abbildung 76 zeigen, dass für die durchschnittliche Beanspruchung sowie für die Dimensionen MD, F und P teilweise signifikante Unterschiede zwischen den Fehlertypen vorliegen. Für die durchschnittliche Beanspruchung sowie die Dimensionen F und P traten signifikante Unterschiede zwischen Interaktions- und Sichtbarkeitsfehlern auf. Für die Dimension MD zeigte sich hingegen ein signifikanter Unterschied zwischen Interaktions- und Datenfehlern. Für die übrigen Vergleiche konnten keine signifikanten Unterschiede nachgewiesen werden. Insgesamt wiesen alle Vergleiche geringe bis moderate Effektstärken bei gleichzeitig geringer Teststärke auf.

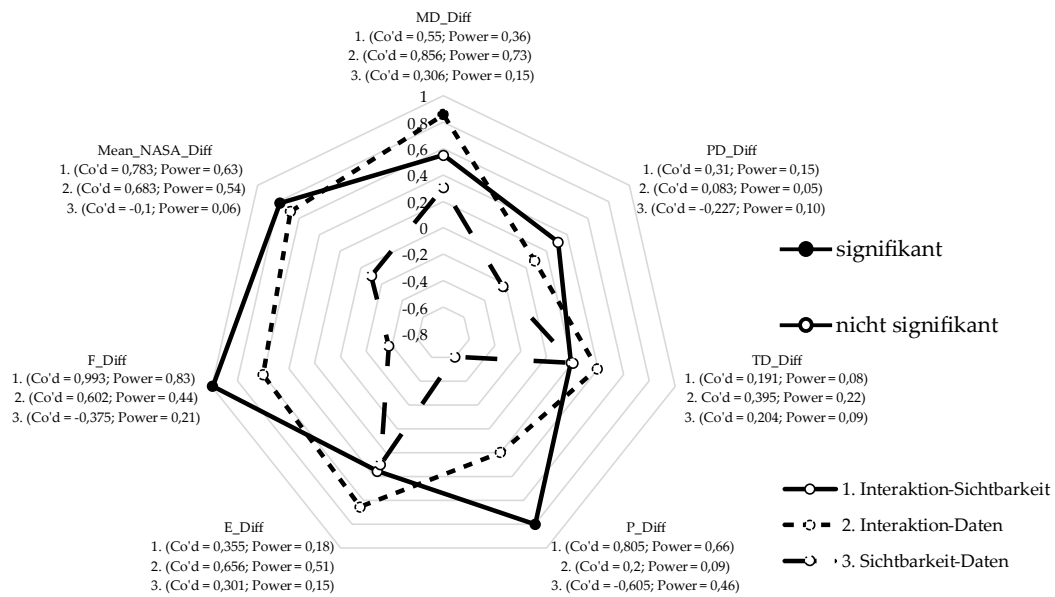


Abbildung 76. ND-Effektstärken des Fehlervergleichs auf die BF

Die Ergebnisse der physiologischen Parameter SCL und HR aus Abbildung 77 zeigen, dass zwischen den Fehlertypen keine signifikanten Unterschiede festgestellt wurden. Für SCL wiesen alle Vergleiche zwischen Interaktionsfehler und Sichtbarkeitsfehler, Interaktionsfehler und Datenfehler sowie Sichtbarkeitsfehler und Datenfehler nur geringe Effektstärken mit niedriger Teststärke auf, sodass keine bedeutsamen Unterschiede zwischen den Faktorstufen der UV Fehlertyp nachgewiesen werden konnten. Ein vergleichbares Bild zeigte sich bei der HR, wo ebenfalls in allen Vergleichen nur geringe Effektstärken mit niedriger Teststärke nachgewiesen wurden.

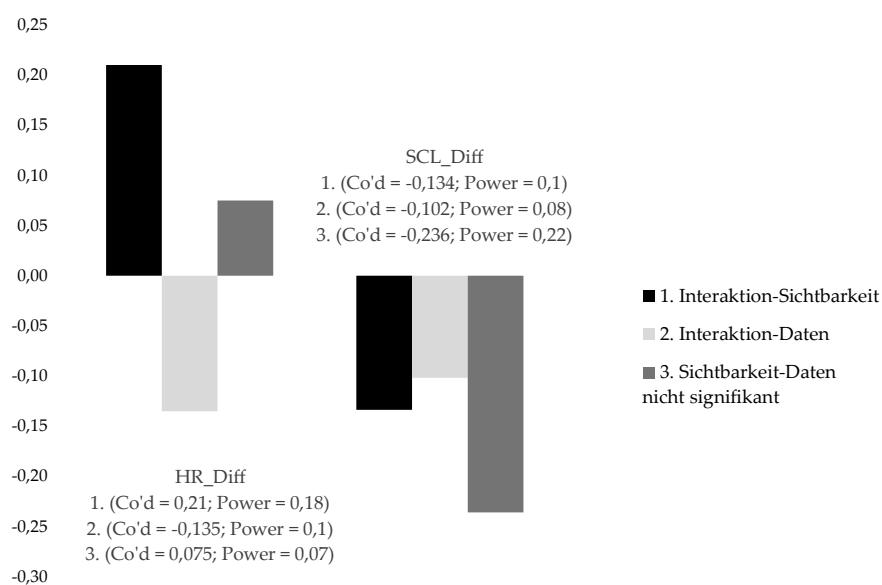


Abbildung 77. BD-Effektstärken des Fehlervergleichs auf die PP

FF₃: Welchen Einfluss hat die Anzeige der Systemkonfidenz innerhalb der MMI auf die Akzeptanz bei IAS?

Antwort: In Studie 4.3 wurde der Einfluss der Systemkonfidenzanzeige auf Akzeptanz (BI) beim Interaktionsfehler (Gestik) untersucht. Die Ergebnisse zeigen, dass ein signifikanter Unterschied mit hoher Effektstärke ($Co'd = 1,12$) bei einer Teststärke von 74 % nachgewiesen werden konnte. Außerdem zeigen die Ergebnisse, dass die absolut gemessene Akzeptanz während des Interaktionsfehlers bei Anzeige der Systemkonfidenz höher ausfiel als in der fehlerfreien Bedingung.

5.2 Implikationen

Die Metastudie von Onnasch et al. (2014) hat gezeigt, welche kausalen Fehlerfolgen hinsichtlich des Situationsbewusstseins, der Arbeitsbelastung sowie der Gesamtleistung bei automatisierten Entscheidungssystemen (kognitiven AS) über Automatisierungsgradienten auftreten. Jedoch wurde bei der Modellierung dieser Basisstudie weder zwischen den Fehlertypen an sich unterschieden, noch wurden die Fehlerauswirkungen auf die Akzeptanz untersucht. Subsumierend wurden Fehler nur als Datenfehler bzw. Systemfehler interpretiert. Überwiegend wurden hierbei Studien aus der Luft- und Raumfahrtpsychologie herangezogen. Durch die Eingrenzung des Forschungsthemas (siehe Abschnitt 1.2) u. a. durch die umfänglichen Fehlertaxonomien von Bahaei et al. (2019), Bahaei und Gallina (2019) und Palanque et al. (2020) ist ersichtlich, dass erweiterte Anforderungen an IAS innerhalb der MMI gestellt werden als bei reinen Entscheidungssystemen (kognitiv). Hieraus geht relevanter Forschungsbedarf hervor (vgl. Kapitel 1.3). Neben den Erkenntnissen aus unterschiedlichen Vorstudien ließen sich bislang nur sehr fragmentierte Aussagen über Fehler bei IAS ableiten (vgl. Kapitel 2). Entsprechend positioniert sich die vorliegende Dissertation und beansprucht aus dem Mangel an relevanten Ergebnissen und Spezifika die forschungsleitende Fragestellung: *Welche Auswirkungen haben Fehler bei IAS innerhalb der MMI auf die Leistung, Beanspruchung und Akzeptanzfaktoren und können adverse Effekte reduziert werden?*

Methodische Implikationen des übergeordneten Forschungsdesigns: Hinsichtlich der Methodik (Abschnitt 3.2) wurde ein spezifischer Ansatz als Forschungsmodell innerhalb der Bilanzhülle der MMI im industriellen Kontext entwickelt (vgl. Abbildung 3;

novelliert in Rosilius, Spiertz et al., 2024). Abgeleitet wurde das Forschungsmodell von dem „Framework to Guide Assessment of Human-Machine Systems“ (Stowers et al., 2017) und entsprechend der Kontextualisierung für industrielle IAS weiterentwickelt. Dieses Modell konkretisiert die Wirkbeziehungen zwischen den Determinanten Leistung, Akzeptanz und Beanspruchung aufgrund ausgewählter Fehlereinflüsse (Datenfehler, Sichtbarkeitsfehler, Interaktionsfehler, Trainingsfehler) bei IAS. Durch die Berücksichtigung der Systemkonfidenzanzeige als Moderatorvariable bietet das Modell einen Ansatz zur Manipulation adverser Effekte auf die Akzeptanz. Dies liefert insgesamt einen relevanten Beitrag zu den einschlägigen Arbeiten wie u. a. Alessa et al. (2023), Cimini et al. (2020), Fantini et al. (2020), Hadorn et al. (2016), Moencks et al. (2022), Pacaux-Lemoine et al. (2017), Palanque et al. (2020), Reuter & Zacher (2008), Stowers et al. (2017), Suh & Prophet (2018), Wandke (2005), B. Wang et al. (2020) und J. Zhou et al. (2019). Diese Arbeit trägt dazu bei, dass die theoretischen Grundlagen zur Abbildung und Evolution von IAS innerhalb produktiver Prozesse weiter operationalisiert werden. Des Weiteren liefert diese Arbeit einen nennenswerten Beitrag zum aktuellen Forschungsstand hinsichtlich der Bewertung der Gesamtleistung im Verbund aus Nutzer und IAS während einer konkreten, fachlichen Arbeitsaufgabe. Als Lösungsansatz des Zielkonflikts bzw. Kompromisses zwischen Richtigkeit und Bearbeitungsdauer bei der Aufgabenbewertung wurde der Index Leistung entwickelt (vgl. „Speed-Accuracy-Tradeoff“ in Standage et al., 2015). Dieser ist definiert entsprechend Gleichung 1 (siehe Definition Index Leistung in Abschnitt 3.2) als Quotient aus Zielerfüllungsgrad bzw. Richtigkeit des Arbeitsinhalts und Bearbeitungsdauer und wurde novelliert in Rosilius, Hügel, et al. (2024). Dieser Index adressiert die Herausforderung, eine eindeutige Bewertung anhand der beiden unabhängigen Metriken Richtigkeit und Bearbeitungsdauer zu treffen. Wie der Tenor der Metastudien von Dey et al. (2018) beschreibt, wird die „(...) Nutzerleistung typischerweise durch die genutzte Zeit und die Richtigkeit in Experimenten gemessen, die der Nutzer bis zur Aufgabenerfüllung benötigt (...)“ (Merino et al., 2020). Mittels der Verrechnung erlaubt der Index Leistung dementsprechend eine quantifizierte Aussage, selbst wenn die Zeitdauer und die Richtigkeit der Aufgabenerfüllung unterschiedliche Einzelbewertungen ergeben.

Methodische Implikationen hinsichtlich der visuellen Qualität: Bei der Untersuchung des Sichtbarkeitsfehlers in Studie 1 (Abschnitt 4.1) wurde der Fehler mittels reduzierter Darstellungsparameter der virtuellen Szene abgebildet. Diese sichtbarkeitsmanipulierten Parameter sind u. a. die Schriftgröße, die Darstellungsfläche, das Kontrastverhältnis, die Helligkeit, die Anzahl und Intensität der Lichtquellen und wurden aus den einschlägigen Vorarbeiten heraus abgeleitet (Gabbard et al., 2007; Rosilius et al., 2021, 2022; Zhao et al., 2017). Für die Untersuchung der Sichtbarkeitsqualität wurde gezielt ein VR-Forschungsprototyp gewählt, da Virtual Reality – im Gegensatz zu Augmented Reality – eine vollständig abgeschirmte, immersive und störungsfreie Umgebung bietet (Caputo et al., 2020, S. 7). So konnten externe Einflüsse wie Licht, Bewegung oder Ablenkungen ausgeschlossen und reproduzierbare Studienbedingungen geschaffen werden (Slater & Sanchez-Vives, 2016, S. 18–35). Die Wahl fiel auf die VR-Museumsanwendung „Bauhausmädels“ (Universität Erfurt, o. D.-b, 10.03.2024). Aufgrund seiner hohen inhaltlichen und visuellen Qualität fördert der VR-Forschungsprototyp durch eine realistische und logisch aufgebaute Darstellung die Präsenzwahrnehmung, indem er die Aufmerksamkeit gezielt auf die virtuellen Inhalte lenkt (von Eitzen, 2024). Die native, gesten- und controllerfreie Interaktion minimierte Bedienhürden und lenkte die Aufmerksamkeit gezielt auf die visuelle Wahrnehmung, wodurch der Effekt des Sichtbarkeitsfehlers präzise messbar wurde.

Methodische Implikationen hinsichtlich der Sprachinteraktionsqualität: Bei der Untersuchung des Sprachinteraktionsfehlers in Studie 2 (Abschnitt 4.2) wurde der Fehler mittels eines reproduzierbaren, diffusen Schallfelds (Wasserstrahlschneide) erzeugt. Hierzu wurden realistische, industrielle Bedingungen in einem kontrollierbaren Versuchsaufbau abgebildet. Das Forschungsdesign berücksichtigte die Herstelleranforderungen der HL2, dass ASR-Funktionen bis zu einem Schalldruckpegel (SPL) von 90 dB bzw. >70 dB(A) einsetzbar sein sollten (Microsoft's HoloLens 2 Team Answers More Questions About Biometric Security, Audio, & Hand Tracking, 2019). Im Gegensatz zu früheren Arbeiten u. a. (Derby et al., 2020), wurde der untersuchte SPL-Bereich erweitert, um realistischere industrielle Einsatzbedingungen abzubilden. Die Experimente wurden jedoch auf einen maximalen Pegel von 87 dB(A) begrenzt, da dies gemäß der Arbeitsstättenverordnung noch als gesundheitlich tolerierbare Exposition gilt (Richtlinie 2003/10/EG des

Europäischen Parlaments und des Rates, 2003). Um realistische Bedingungen zu schaffen und negative Auswirkungen zu minimieren, wurde entsprechend der Arbeitsstättenverordnung während der Experimente Gehörschutz getragen. Im Gegensatz zu den einschlägigen Vorstudien (Derby et al., 2020; Sinlapanuntakul et al., 2021) wurde eine gegenbalancierte RF im Within-Subject-Design gewählt, um vermutete Carry-Over-Effekte der Vorstudien zu vermeiden. Ebenfalls wurde die statistische Validität zur Bewertung des Schalldruckpegels bzw. Fehlereinflusses durch die Erweiterung des Probandenumfangs von $N = 13$ bzw. $N = 16$ auf $N = 22$ sowie durch die Absicherung der effektiven Teststärke erhöht. Zur Bewertung der Interaktionsleistung bei der ASR wurde neben der WER-Metrik (siehe Gleichung 2) zusätzlich die ITR (siehe Gleichung 5) herangezogen; diese wurde in Rosilius, Spiertz et al. (2024) novelliert. Ursprünglich wurde die ITR für die Bewertung von Brain-Computer-Interfaces entwickelt (Dal Seno et al., 2010; Wolpaw et al., 2002). Innerhalb der vorliegenden Arbeit wurde validiert, dass sich diese Metrik ebenfalls zur Bewertung der ASR-Sprachbefehl-Interaktionen eignet. Hierdurch wird nicht nur die Dimension der Richtigkeit der Sprachbefehle berücksichtigt, sondern auch die zeitliche Komponente sowie die Gesamtanzahl der Sprachbefehle bzw. Versuche. Diese Metrik schafft vor allem für den professionellen Einsatz eine stärkere Differenzierung hinsichtlich der Gebrauchstauglichkeit und Vergleichbarkeit der Ergebnisse. Die Adaption und der Einsatz dieser spezifischen Metrik liefern einen relevanten Beitrag zu den einschlägigen Vorstudien hinsichtlich der Bewertung der ASR-Leistungsfähigkeit bei immersiven Systemen (Derby et al., 2020; Liu et al., 2022; Sinlapanuntakul et al., 2021).

Methodische Implikationen hinsichtlich der Trainingsqualität: Bei der Untersuchung des Trainingsfehlers in Studie 3 (Abschnitt 4.3 ff.) waren Probanden teilweise nicht in der Lage, die an sie gestellten Aufgaben vollständig zu erledigen. Somit war der Zielerfüllungsfaktor als UV ein berücksichtigter Faktor im DOE. Zur Bewertung von unvollendeten Teilaufgaben wurde eine von diesen abhängige Strafzeit post-hoc ermittelt und als Cut-off-Wert angewendet. Die Strafzeit pro Teilaufgabe k wurde als das 75. Perzentil über alle Versuchspersonen hinweg für jede Teilaufgabe definiert (novelliert in Rosilius, Hügel et al., 2024). Auch wenn dieses Verfahren vom sog. Cut-off-Verfahren abgeleitet wurde (Osawa & Ren, 2004), stellt dies bei der Bewertung von Systemleistungen

aufgrund von Trainingsqualitäten einen klaren Mehrwert, auch außerhalb immersiver Systeme, dar. Flankierend kann aus dem Forschungsumfang bei IAS festgehalten werden, dass nicht das ausführliche, detaillierte Trainingsszenario (Leistung = 2,20 %/min) die höchste Nutzerleistung bewirkt, sondern ein optimiertes Szenario (4,09 %/min). Die vorliegende Studie 3 unterscheidet sich von bisherigen AR-Studien (De Giorgio et al., 2023; Garcia Fracaro et al., 2022; Martins et al., 2023), da sie nicht das Lernen mit AR-Systemen fokussiert, sondern sich auf Trainingsszenarien für immersive Systeme konzentriert. Außerdem wurde ein Forschungsdesign zur Untersuchung der Wirksamkeit und Auswirkungen von Trainingsszenarien entwickelt.

Methodische Implikationen der Verbundstudie: In der Verbundstudie 4 wurden die Probanden mithilfe eines Forschungsprototyps an einer fachlichen Aufgabe gezielt drei Fehlertypen (Daten-, Interaktions- und Sichtbarkeitsfehler) exponiert. Das Hauptziel dieser Studie ist, zum einen durch ein störeinflussarmes Versuchsdesign einen aussagekräftigen Fehlervergleich zu gewährleisten, zum anderen den Einfluss der Moderatorvariable Systemkonfidenzanzeige auf die Akzeptanz hin zu untersuchen. Flankierend wurden die Fehlerauswirkungen unabhängig voneinander analysiert, u. a. der Datenfehler, der ausschließlich in dieser Studie manipuliert wurde. Die Wahl des MMD (Murrar & Brauer, 2018) in Kombination mit einem randomisierten, paritätisch-gegenbalancierten Versuchsaufbau trägt wesentlich zur internen Validität der Ergebnisse bei. Zum einen konnten durch die systematische Kontrolle potenzieller Reihenfolgeeffekte und zum anderen durch die Change-Score-Methode, sprich das Referenzieren des jeweiligen Fehlertyps auf den fehlerfreien Fall, Störeinflüsse wie individuelle Trainingseffekte, Technologieaffinität oder Vorerfahrungen der Probanden reduziert werden (Becker, 1988; Jennings & Cribbie, 2016). Im Zentrum der Untersuchung stand der direkte Vergleich unterschiedlicher Fehlertypen – Daten-, Sichtbarkeits- und Interaktionsfehler. Die gezielte Variation zweier unabhängiger Variablen – des Fehlertyps Between-Subjects und des Fehlerfalls Within-Subjects – ermöglichte eine differenzierte Betrachtung sowohl interindividueller als auch intraindividueller Reaktionsmuster. Das methodische Design bildet somit eine belastbare Grundlage für die empirische Analyse der Auswirkungen spezifischer Fehlerarten. Diese Verbundstudie stellt in vielerlei Hinsicht ein Novum dar. Es konnte keine einschlägige Studie recherchiert werden, bei der mehrere Fehlertypen

an einem immersiven Forschungsprototypen untersucht wurden. Außerdem leistet das Versuchsdesign dank seines validen Forschungsdesigns, u. a. der „Change-Score“-Methode, einen relevanten Beitrag als veritabler Ansatz für den Vergleich von Fehlertypen bei zukünftigen Forschungsbemühungen im IAS-Kontext. Die Wahl des zugrunde liegenden Forschungsdesigns hat unmittelbare Auswirkungen auf die Interpretierbarkeit und praktische Relevanz der Ergebnisse. Durch die bewusste Orientierung am Parsimony-Prinzip (Sober, 1981) wurde auf übermäßige Modellkomplexität verzichtet, was sowohl die Transparenz der inferenzstatistischen Auswertung als auch die Nachvollziehbarkeit für praktische Anwendungen stärkt. Anstelle komplexer Modellierungsverfahren wie Strukturgleichungsmodelle kamen Hypothesentests und Varianzanalysen zum Einsatz, da diese – gerade bei streng kontrollierten experimentellen Bedingungen – als effizient, robust und aussagekräftig gelten (Wirth & Klein, 2023). Besonders hervorzuheben ist, dass durch die gewählte Analysemethode Störvariablen minimiert, Effektstärken interpretierbar gemacht und konkrete Hypothesen stringent überprüft werden konnten. Diese methodische Klarheit unterstützt die Anwendungsorientierung der Befunde und erhöht deren Übertragbarkeit auf vergleichbare Kontexte. Darüber hinaus wurde die statistische Aussagekraft durch eine formal geplante Stichprobengröße (Poweranalyse) abgesichert, was die Validität der getesteten Effekte zusätzlich stärkt (Döring & Bortz, 2016). Nachdem die vorhergehenden Absätze dieses Kapitels die Implikationen sowohl des relevanten Forschungsbedarfs als auch der methodischen Ansätze herausgestellt haben, wird im Folgenden konkret auf die wissenschaftlichen Ergebnisse bezüglich der Forschungsfragen eingegangen. Als Vorwort zu den Ergebnissen lässt sich betonen, dass sich in der einschlägigen Literatur keine vergleichbare Forschungsarbeit finden lässt, die zusammenhängend verschiedene Fehlerarten sowie deren Auswirkungen innerhalb der MMI im IAS-Kontext untersucht.

Implikationen hinsichtlich der DQ: Die Ergebnisse der Studie 4.1 zeigen, dass fehlerhafte Daten innerhalb der MMI bei IAS einen signifikanten Einfluss auf Leistung und Akzeptanz haben, jedoch keine relevanten Auswirkungen auf die Beanspruchung hervorrufen. Konkret wurde festgestellt, dass fehlerhafte Daten die Leistung der Nutzenden mit mittlerer Effektstärke ($Co'd = 0,69$) und hoher statistischer Aussagekraft (Power = 92 %) negativ beeinflussen. Ebenso besteht ein signifikanter Einfluss auf die Akzeptanz

mit mittlerer Effektstärke ($Co'd = 0,53$) und ebenfalls hoher Teststärke (Power = 75 %). Die detaillierte Betrachtung der sekundären Akzeptanzfaktoren ergibt, dass der Einfluss des Datenfehlers nur bei der BI signifikant ist, während für die übrigen Akzeptanzfaktoren PE, EE, HM sowie FC keine signifikanten, sondern lediglich geringe bis moderate Effekte mit konsistenten Vorzeichen und niedriger bis mittlerer Teststärke nachweisbar waren. Die Beanspruchung der Nutzenden sowie die physiologischen Parameter zeigen insgesamt keine statistisch signifikanten Effekte, wobei sich für die Beanspruchungsfaktoren und die physiologischen Parameter keine konsistenten Effektrichtungen innerhalb der Stichprobe erkennen lassen. Effektstärken für die Beanspruchungsfaktoren bewegen sich zwischen $-0,31$ und $0,26$ mit einer sehr niedrigen statistischen Aussagekraft (Power $< 0,39$). Ebenso konnten für die physiologischen Parameter nur sehr kleine, nicht signifikante Effekte bei äußerst geringer Teststärke (6–8 %) festgestellt werden.

Implikationen hinsichtlich der visuellen Qualität: Die Ergebnisse aus den Studien 1 (VR) und 4.1 (AR) zeigen, dass eine fehlerhafte Visualisierung durch den Sichtbarkeitsfehler in der MMI bei IAS signifikante, negative Auswirkungen auf die Nutzerleistung hat. Die festgestellten Effektstärken für die Leistung unterscheiden sich zwischen den Technologien, wobei in der VR-Umgebung mit $Co'd = 1,15$ eine hohe Effektstärke und in der AR-Umgebung mit $Co'd = 0,72$ eine mittlere Effektstärke nachgewiesen wurde. Im Gegensatz zur Leistung wurden für keine Dimension des UTAUT2 ein statistisch signifikanter Einfluss der fehlerhaften Visualisierung festgestellt. Jedoch besitzen die Ergebnisse des UTAUT2 aufgrund der niedrigen Teststärken (5–35 %) eine nur sehr eingeschränkte statistische Aussagekraft. Auch im Bereich der Nutzerbeanspruchung ergab sich insgesamt kein systematisch signifikanter Effekt des Sichtbarkeitsfehlers. Lediglich im VR-Kontext wurde für die Frustration ein statistisch signifikanter Einfluss mit hoher Effektstärke ($Co'd = 1,12$) und hinreichender Teststärke (Power = 71 %) festgestellt. Für die übrigen NASA-TLX-Dimensionen zeigten sich sowohl in VR als auch in AR lediglich kleine Effektstärken bei überwiegend geringer Teststärke (Power zwischen 6 % und 44 %). Zudem traten zwischen beiden Technologien teilweise gegenläufige Effektrichtungen auf, sodass sich keine konsistenten Auswirkungen des Sichtbarkeitsfehlers auf die subjektive Beanspruchung ableiten lassen. Die physiologischen Parameter ergaben aufgrund fehlerhafter Visualisierungen keine statistisch signifikanten Effekte bei insgesamt niedriger

Teststärke (12–27 %), was auf vernachlässigbare physiologische Auswirkungen innerhalb der MMI schließen lässt.

Implikationen hinsichtlich der Interaktionsqualität: Die Ergebnisse der Studien 2 (ASR) und 4.1 (Gestik) zeigen, dass ein Interaktionsfehler innerhalb der MMI von IAS signifikante Auswirkungen auf alle Primärfaktoren (Leistung, Akzeptanz, Beanspruchung) hat. Für die Leistung zeigt sich in beiden Interaktionsformen ein signifikanter Effekt mit hoher Effekt- und Teststärke ($Co'd \geq 0,74$; $Power \geq 0,96$). Für die Akzeptanz ergaben sich für ASR sehr hohe signifikante Effekte ($Co'd = 1,55$), während sich bei der Gestik lediglich für BI ein signifikanter mittlerer Effekt zeigte ($Co'd = 0,53$). Im Detail wurden für ASR nahezu alle Akzeptanzfaktoren signifikant und hoch negativ beeinflusst, außer FC und PE. Für alle anderen Akzeptanzfaktoren ergaben sich lediglich geringe Effekte mit niedrigen Teststärken ($Co'd \leq 0,25$; $Power \leq 0,29$). Bezüglich der Beanspruchung konnten sowohl bei ASR als auch bei Gestik signifikante Effekte auf die Dimensionen F ($Co'd \geq 0,8$; $Power \geq 0,71$) und MD festgestellt werden. Zusätzlich zeigten sich bei der Gestik signifikante Effekte für P sowie die durchschnittliche Gesamtbeanspruchung. Hingegen ergaben sich für PD, TD und E keine signifikanten Effekte bei nur sehr geringen Effektstärken. Für physiologische Parameter wurde lediglich bei der Gestik ein sehr geringer, nicht signifikanter Effekt mit niedriger Teststärke ($Power \leq 0,14$) beobachtet. Die Daten zu physiologischen Parametern bei ASR mussten aufgrund eines Erhebungsfehlers ausgeschlossen werden.

Implikationen hinsichtlich der Trainingsqualität: Die Ergebnisse der Studie 3 zeigen, dass ein Trainingsfehler innerhalb der MMI von IAS signifikante und sehr starke negative Einflüsse auf alle Primärfaktoren hat. Für die Primärfaktoren Leistung, Akzeptanz sowie die durchschnittliche Beanspruchung wurden aufgrund des Trainingsfehlers durchgehend sehr hohe Effektstärken ($Co'd \geq 1,17$) mit hoher Teststärke ($Power = 100\%$) nachgewiesen. Bezüglich der Akzeptanz zeigt der Trainingsfehler signifikante und hohe negative Einflüsse für sämtliche untersuchten Akzeptanzfaktoren des UTAUT2 mit durchgehend hohen Effektstärken ($Co'd$ zwischen 0,89 und 1,17) und hoher Teststärke ($Power \geq 0,99$). Auch hinsichtlich der Beanspruchung ergaben sich deutliche negative Effekte des Trainingsfehlers auf die Dimensionen F, P, MD sowie die durchschnittliche Beanspruchung mit signifikanten und überwiegend hohen Effektstärken ($Co'd \geq 0,8$) bei

hoher Teststärke (Power $\geq 97\%$). Die Dimensionen PD und TD zeigten hingegen geringe, nicht signifikante Effekte ($Co'd \leq 0,3$) mit niedriger Teststärke (Power $\leq 31\%$). Die Dimension E weist trotz hoher Effektstärke ($Co'd = 0,77$) und Teststärke (Power = 97%) keine statistische Signifikanz auf. Die physiologischen Parameter wurden durch den Trainingsfehler stark beeinflusst. Es konnten große und für HR sogar signifikante Effekte ($Co'd \geq 0,80$) mit hoher Teststärke (Power $\geq 0,98$) nachgewiesen werden.

Implikationen hinsichtlich des Fehlervergleichs: Die Ergebnisse der Studie 4.2 zeigen, dass unterschiedliche Fehlertypen (Datenfehler, Sichtbarkeitsfehler, Interaktionsfehler) innerhalb der MMI von IAS vor allem in Bezug auf Akzeptanz und Beanspruchung signifikant differenzieren, während für die Leistung und physiologischen Parameter keine bedeutsamen Unterschiede festgestellt wurden. Für die Akzeptanz (gemessen über BI) zeigten sich signifikante Unterschiede zwischen Interaktionsfehler und Sichtbarkeitsfehler ($Co'd = 0,77$; Power = $0,97$) sowie zwischen Sichtbarkeitsfehler und Datenfehler ($Co'd = -0,85$; Power = $0,99$). Kein signifikanter Unterschied ergab sich zwischen Interaktionsfehler und Datenfehler ($Co'd = -0,08$; Power = $0,07$). Auch bei den Akzeptanzfaktoren BI und PE zeigten diese Fehlertypen hohe Effektstärken, während für EE, HM und FC keine signifikanten Unterschiede festgestellt wurden. In Bezug auf die Beanspruchung ergab sich ein signifikanter Unterschied bei der durchschnittlichen Beanspruchung zwischen Sichtbarkeitsfehler und Interaktionsfehler ($Co'd = 0,78$; Power = $0,64$). Auch MD, F und die wahrgenommene Leistung zeigten signifikante Unterschiede zwischen denselben Fehlertypen sowie zwischen Sichtbarkeitsfehler und Datenfehler. Keine signifikanten Unterschiede wurden zwischen Interaktionsfehler und Datenfehler beobachtet (z. B. bei durchschnittlicher Beanspruchung $Co'd = -0,1$; Power = $0,06$). Für die Leistung konnten keine signifikanten Unterschiede festgestellt werden. Die höchsten Effektstärken traten zwischen Interaktionsfehler und Datenfehler auf ($Co'd = 0,49$; Power = $0,41$), jedoch ohne Signifikanz. Die übrigen Vergleiche lagen im Bereich $Co'd \leq 0,24$ mit sehr geringer Teststärke. Auch bei den physiologischen Parametern (SCL, HR) zeigten sich in keinem Vergleich signifikante Unterschiede. Alle Effektstärken lagen im niedrigen Bereich ($Co'd \leq 0,2$) mit geringer Teststärke (Power $\leq 0,3$).

Implikationen hinsichtlich der Akzeptanzmanipulation: Die Ergebnisse der Studie 4.3 zeigen, dass die Anzeige der Systemkonfidenz innerhalb der MMI einen signifikanten

positiven Einfluss auf die Akzeptanz (BI) beim Interaktionsfehler (Gestik) hat. Es wurde ein signifikanter Unterschied mit hoher Effektstärke ($Co'd = 1,12$) bei einer Teststärke von 74 % nachgewiesen. Bemerkenswert ist außerdem, dass die gemessene BI bei Vorliegen eines Interaktionsfehlers in Kombination mit einer Systemkonfidenzanzeige sogar höher ausfiel als im fehlerfreien Zustand. Dies deutet darauf hin, dass Transparenz über die Systemzuverlässigkeit die negativen Auswirkungen von Interaktionsfehlern auf die Akzeptanz abschwächen kann und damit einen wichtigen Beitrag zur Gestaltung akzeptanzfördernder Mensch-Maschine-Interaktionen leistet.

5.3 Limitationen und weiterer Forschungsbedarf

Obgleich die vorliegenden Studien zentrale Erkenntnisse zu den Auswirkungen von Fehlern innerhalb der MMI bei IAS liefern, unterliegen sie – wie jede empirische Forschung – bestimmten methodischen, technischen und theoretischen Begrenzungen. Die nachfolgende Betrachtung beleuchtet zentrale Limitationen der durchgeführten Einzelstudien sowie des übergeordneten Forschungsdesigns. Aus diesen Einschränkungen lassen sich zugleich wertvolle Ansatzpunkte für zukünftige Forschungsarbeiten ableiten, um die gewonnenen Befunde zu vertiefen, zu erweitern oder unter realitätsnäheren Bedingungen zu überprüfen. Außerdem gilt es, die methodischen Novellierungen kritisch zu hinterfragen.

Limitationen des übergeordneten Forschungsdesigns: Auch wenn sich bei der Versuchsdurchführung sowie Auswertung eng an den Standard wissenschaftlicher Praxis hinsichtlich Stichprobenumfangsplanung sowie Post-Hoc-Betrachtung der statistischen Teststärke gehalten wurde (Döring & Bortz, 2016), lag die effektive Teststärke in einer Vielzahl von Variablen unter 80 %. Hierbei ist die statistische Validität aufgrund unzureichender Stichprobengröße teilweise stark eingeschränkt. Es ist zu empfehlen, weiterführende Studien mit optimierten Stichprobenumfängen durchzuführen, um die vorliegenden Ergebnisse zu validieren bzw. die statistische Aussagefähigkeit zu erhöhen. Das übergeordnete Forschungsdesign basiert auf einem MMD (Murrar & Brauer, 2018), das vier methodisch und konzeptionell heterogene Einzelexperimente umfasst. Diese Modularität ermöglicht zwar die differenzierte Analyse verschiedener Fehlertypen – darunter Sichtbarkeits-, Interaktions-, Trainings- und Datenfehler –, sie geht jedoch mit

erheblichen Einschränkungen hinsichtlich der internen Kohärenz einher. Die deutlichen Unterschiede in den eingesetzten Forschungsprototypen, Versuchsbedingungen, Erhebungsinstrumenten (z. B. Fragebögen, Metriken) sowie in der Operationalisierung der unabhängigen Variablen erschweren eine konsistente Vergleichbarkeit der Ergebnisse. Diese Form methodischer Fragmentierung wird in der aktuellen Forschung zunehmend kritisch diskutiert, da sie die konzeptuelle Anschlussfähigkeit zwischen den Studien beeinträchtigen und die Gültigkeit übergreifender Schlussfolgerungen unterminieren kann (Kaltenhauser et al., 2024, S. 14). Ähnlich wie Pettersson et al. (2018) im Kontext der UX-Forschung kritisieren, erfolgt zwar eine methodische Diversifikation, jedoch ohne kohärente Zusammenführung der Ergebnisse im Sinne einer analytischen Triangulation. Dadurch bleiben potenzielle Zusammenhänge zwischen den erhobenen Indikatoren unerkannt, was die Erklärungskraft und interne Validität des Gesamtmodells erheblich einschränkt. Insbesondere mangelt es an einem übergreifenden statistischen Modell, welches gemeinsame Wirkmechanismen identifiziert oder konsistente Muster theoretisch begründen kann. Weiterführenden Forschungsaktivitäten wird empfohlen, mittels analytischer Triangulation eine systematische Verknüpfung und integrierte Auswertung dieser Datenquellen über die Einzelexperimente hinweg durchzuführen. Die Generalisierbarkeit der Studienergebnisse von IAS ist im Kontext der vorliegenden Dissertation aufgrund mehrerer forschungspraktischer Rahmenbedingungen eingeschränkt. So erfolgte die Evaluation primär anhand artifizieller, prototypischer Laboranwendungen innerhalb kontrollierter Umgebungen mit begrenzter Realitätsnähe und nur bedingt zeitkritischem Anforderungsprofil. Obwohl diese Studien die isolierte Untersuchung spezifischer Fehlerwirkungen erlaubten, bilden sie nicht das volatile Umfeld realer industrieller Hochlastanwendungen ab, dessen Nutzungskontext (vgl. Kap. 1.1–1.2) adressiert werden soll. Die Übertragbarkeit der Ergebnisse in reale Umgebungen ist an die technologischen Limitationen per se gebunden, u. a. das Trackingvermögen bzw. die Spatial Awareness des Systems im industriellen Umfeld, was bei IAS ein zentrales Argument darstellt (Merino et al., 2020). Zudem erfolgte die Rekrutierung der Versuchspersonen größtenteils aus hochschulnahen Gruppen (z. B. Studierende oder Hochschulpersonal), was zu einer eingeschränkten Repräsentativität führt. Diese Probanden verfügen typischerweise über eine höhere Technologieaffinität sowie ein gesteigertes

Explorationsverhalten. Das sind Eigenschaften, die sich signifikant von industriellen Fachanwendern unterscheiden können (z. B. Maschinenführer, Wartungspersonal). In der Folge ist davon auszugehen, dass insbesondere die subjektiven Bewertungen (z. B. UTAUT2, NASA-TLX) sowie die Leistungsmaße unter industriellen Bedingungen deutlich abweichende Ergebnisse zeigen könnten. Für eine belastbare Extrapolation der Ergebnisse sind deshalb weiterführende Feldstudien im realen Anwendungskontext sowie mit praxisnahen Zielgruppen erforderlich (H. Park et al., 2024). Das vorliegende Forschungsdesign beschränkt sich lediglich auf Momentaufnahmen, wodurch zeitliche Entwicklungen von Akzeptanz, Beanspruchung und Leistung unberücksichtigt bleiben. U.a. nach Kim & Malhotra (2005) wäre ein longitudinales Design bei IAS erforderlich, um Veränderungsprozesse und habituelle Aneignung systematisch abzubilden. Zudem wurde die technologische Vorerfahrung der Teilnehmenden zwar erhoben, jedoch nicht als moderierende Variable in die Analyse integriert, obwohl diese das Erleben und Bewerten interaktiver Systeme laut Bekier & Molesworth (2017) signifikant beeinflussen kann. Bei der Manipulation der jeweiligen unabhängigen Variablen wurde kein Gütekriterium hinsichtlich der Fehlermerkmalsausprägung definiert, welches die interne Validität der Studien stützt. Somit ist unklar, inwieweit die Effektstärken von der Merkmalsausprägung der jeweiligen Fehlerintensität abhängen (Chester & Lasko, 2021). Über alle Einzelstudien hinweg wurden die physiologischen Parameter HR und SCL erfasst. Allerdings konnten signifikante Unterschiede lediglich in Teilstudie 3 im Zusammenhang mit dem Trainingsfehler nachgewiesen werden. Angesichts des Mangels an belastbaren Ergebnissen erscheint es notwendig, die verwendeten Messmethoden, Analyseverfahren sowie die eingesetzte Hardware zur Erfassung physiologischer Stressreaktionen im Kontext von IAS kritisch zu hinterfragen. Diese Aspekte sollten in weiterführenden Studien sowohl aus medizinisch-messtechnischer als auch aus psychologischer Perspektive eingehend untersucht werden. Darüber hinaus bedarf es einer vertieften Auseinandersetzung mit den grundsätzlichen Einschränkungen der Aussagekraft von SCL- und HR-Messungen in diesem Anwendungsfeld. Wie auch in einschlägigen Studien (Alessa et al., 2023; Giorgi et al., 2021; Kosch et al., 2023; Setz et al., 2010) hervorgehoben wird, unterliegen diese physiologischen Metriken verschiedenen Limitationen. Dazu zählen u. a. die Abnutzung der Elektroden, individuelle Unterschiede wie

Hauttrockenheit, die Verwendung von Pflegeprodukten oder ausgeprägtes Schwitzen sowie externe Einflussfaktoren wie Temperaturveränderungen und Bewegungsartefakte. Diese Aspekte sollten in zukünftigen Arbeiten systematisch berücksichtigt werden, um die Validität physiologischer Stressmessungen im Rahmen von IAS-Evaluationen zu verbessern. In der vorliegenden Dissertation wurden für die Erhebung ausschließlich validierte und standardisierte Fragebogenkonstrukte verwendet. Ziel war es nicht, neue theoretische Modelle zu entwickeln oder bestehende Modelle umfassend zu evaluieren, sondern konkrete Forschungsfragen auf Grundlage valider Vorarbeiten zu erörtern. Dennoch unterliegt die Operationalisierung der eingesetzten Fragebögen gewissen Einschränkungen. Zum einen wurden die nativen Items und Konstrukte bestmöglich an den spezifischen Anwendungskontext angepasst. Zum anderen erfolgte eine Übersetzung der Fragebögen ins Deutsche, wodurch potenziell Verzerrungen und Vergleichbarkeit der zugrunde liegenden Modelle entstanden sein könnten (Tractinsky, 2018; Vagnetti et al., 2024).

Limitationen hinsichtlich der visuellen Qualität (VR): Bezogen auf den Sichtbarkeitsfehler in Studie 1 bedeutet dies, dass die Sichtbarkeit bzw. Lesbarkeit mit einer statischen Manipulationsstärke, u. a. Textgröße, Kontrast und Helligkeit, definiert wurde. Jedoch ist eine Verallgemeinerung der Ergebnisse dahingehend eingeschränkt, dass nicht abzuleiten ist, inwieweit Reaktionsmuster der abhängigen Variablen mit der Fehlerintensität korrelieren. Ebenfalls würde sich diese Fragestellung für weiterführende Studien anbieten.

Limitationen hinsichtlich der Interaktionsqualität (Sprache): Im Kontext des Interaktionsfehlers I (Sprachsteuerung – Studie 2) wurde der Sprachinteraktionsfehler über die Exposition gegenüber Umgebungslärm induziert. Diese Manipulation spiegelt einerseits ein realistisches Fehlerbild wider, wie es im industriellen Anwendungskontext nativ auftreten kann. Andererseits birgt sie das Risiko überlagerter Effekte (Superpositionen), die im Rahmen des Forschungsdesigns nicht differenziert kontrolliert wurden und somit potenziell zu Verzerrungen in der Ergebnisinterpretation führen können. So ist bekannt, dass Sprecher in geräuschintensiven Umgebungen ihre Stimme unbewusst anpassen – etwa durch erhöhte Lautstärke, veränderte Artikulation oder Modulation der Tonhöhe (vgl. Lombard-Effekt; Junqua, 1993). Auch das Tragen von Gehörschutz kann diese

stimmliche Selbstregulation beeinflussen und führt nachweislich zu veränderter Sprachproduktion (Tufts & Frank, 2003). Es kann daher nicht ausgeschlossen werden, dass beobachtete Akzeptanzverluste auch teilweise durch Störfaktoren wie Lärmbelastung oder das Tragen des Gehörschutzes bedingt sind. Diese Wechselwirkungen sollten in zukünftigen Untersuchungen explizit berücksichtigt werden. Zudem wäre eine methodische Alternative, den Sprachinteraktionsfehler softwarebasiert generisch zu implementieren, um dessen Intensität gezielter operationalisieren und systematisch variieren zu können.

Limitationen hinsichtlich der Trainingsqualität: Der in Studie 3 untersuchte Trainingsfehler wurde über ein Trainingsszenario mittels unzureichender Trainingsqualität operationalisiert und mit Faktorstufen verglichen, die selektiv verbesserte Trainingsbedingungen boten. Zwar basiert die Ausgestaltung dieser Faktorstufen auf empirischen Erkenntnissen des Forschungsprojekts (PlanAR, o. D.), jedoch bedarf die konzeptionelle und empirische Fundierung der jeweiligen TQ-Stufen weiterer systematischer Untersuchung. Eine generalisierbare Aussagekraft der Ergebnisse ist derzeit nur eingeschränkt gegeben, da unklar bleibt, in welchem Ausmaß Trainingsszenarien für IAS kontext- bzw. anwendungsspezifisch gestaltet sein müssen. Darüber hinaus fehlt eine tragfähige empirische Studienbasis, welche bestehende Methoden und Ergebnisrahmen für Trainingsszenarien in IAS umfassend darlegt. Auch spezifische Evaluationsverfahren zur Qualität von Trainingsszenarien lassen sich bislang nicht aus der einschlägigen Fachliteratur ableiten. Aus diesen Gründen erscheint eine vertiefte Weiterentwicklung dieser Forschungsperspektive sowie der Validierung der Ergebnisse unter Berücksichtigung ihrer Kontextsensitivität dringend geboten.

Limitationen hinsichtlich der visuellen Qualität (AR): In Teilstudie 4.1 wurden die Auswirkungen der Fehlertypen – Sichtbarkeits-, Daten- und Interaktionsfehler – auf Basis eines AR-Forschungsprototyps (vgl. Kapitel 4.4.1) systematisch im Rahmen einer fachlich anspruchsvollen Messaufgabe manipuliert. Die Aussagekraft der Ergebnisse ist jedoch aufgrund der selektiven Definition der Fehlerintensität begrenzt. Besonders auffällig ist das atypische Reaktionsmuster des Sichtbarkeitsfehlers II (AR), welches sich in der Datenanalyse zeigt: Trotz signifikanter Leistungseinbußen ließen sich keine signifikanten Effekte auf die Beanspruchung oder Akzeptanz feststellen. Im Gegenteil, es wurden negative Mittelwertdifferenzen in Bezug auf die Akzeptanz und mittlere

Beanspruchung gemessen, was darauf hindeutet, dass im Fehlerfall die subjektive Bewertung des AR-Systems teilweise positiver ausfiel als in der fehlerfreien Bedingung. Diese Beobachtung lässt die Annahme zu, dass trotz objektiv reduzierter Systemleistung durch den Sichtbarkeitsfehler die subjektive Akzeptanz und wahrgenommene Beanspruchung durch fachliche Kontextfaktoren kompensiert wurden. Ein möglicher Erklärungsansatz ergibt sich aus der spezifischen Gestaltung der Fehlerinduktion: Die künstliche Reduktion der Augmentierungsopazität um 50 % erschwerte zwar die Sichtbarkeit bzw. Lesbarkeit virtueller Inhalte, ermöglichte aber gleichzeitig eine bessere Sicht auf reale Referenzpunkte (die Tischunterkante), was die Genauigkeit der Platzierung relativ zur BB subjektiv erleichtert haben könnte. Diese widersprüchlichen Effekte zwischen objektiver Leistungseinbuße und subjektiver Bewertungsvorteile im Kontext eingeschränkter Sichtbarkeit sollten in nachfolgenden Studien vertieft untersucht werden – insbesondere im Hinblick auf kontextuelle Einflussfaktoren der fachlichen Aufgabe sowie auf die notwendigen Wechselwirkungen von Realweltwahrnehmung und virtueller Überlagerung bei IAS.

Limitationen hinsichtlich der DQ: Eine Generalisierbarkeit der Ergebnisse im Zusammenhang mit dem Datenfehler ist sowohl durch das Studiendesign per se – insbesondere der gewählten Fehlerschwere –, als auch durch die Motivation und Ernsthaftigkeit der Probanden an der Messaufgabe eingeschränkt. Um die Beteiligung der Teilnehmenden am Aufgabenerfolg zu fördern, wurde ein kompetitiver Studiencharakter mit monetärer Incentivierung als extrinsischer Motivationsfaktor eingesetzt. Dennoch bleibt unklar, ob die Probanden die Unplausibilität der systematisch überdimensionierten Objektgröße – im Fehlerfall doppelt so groß wie das reale Referenzobjekt (Tisch) – tatsächlich wahrgenommen und die Kausalität des entstandenen Messfehlers dem System oder sich selbst zugeschrieben haben. Dementsprechend konnte im Studiendesign weder die tatsächliche Einstellung der Probanden zur Aufgabe noch deren effektive Wahrnehmung des Fehlers als potenzielle Störvariablen ausgeschlossen oder deren Einfluss zuverlässig quantifiziert werden. In diesem Zusammenhang betonen Brauner et al. (2019), dass nicht alle Versuchspersonen das notwendige Maß an Ernsthaftigkeit und Engagement aufbringen, was die Ergebnisse hinsichtlich der Leistung bzw. der Akzeptanz beeinflussen kann. Im Hinblick auf dateninduzierte Fehler in IAS besteht somit ein fortbestehender

Forschungsbedarf – insbesondere hinsichtlich der Abgrenzung technikinduzierter Fehlerwirkungen von wahrnehmungspsychologischen und motivatorischen Einflussfaktoren.

Limitationen hinsichtlich der Interaktionsqualität (Gestik): Im Hinblick auf den Interaktionsfehler II (Gestik) aus Teilstudie 4.1 sind die Ergebnisse nicht frei von Einschränkungen. Der Fehler wurde experimentell erzeugt, indem die Rotationsfreiheitsgrade gezielt an die Translationsinteraktion gekoppelt wurden (vgl. integrierte Steuerung der Freiheitsgrade; Caputo et al., 2024). Diese Kopplung erhöhte die Steuerungssensitivität – sowohl im Nahbereich als auch bei der Ferninteraktion bei der virtuellen BB – und führte dadurch zu einer erhöhten Komplexität beim Positionieren für die Versuchspersonen. Caputo et al. (2024) benennen darüber hinaus weitere Limitationen, die das Interaktionsverhalten sowie die Messergebnisse wesentlich beeinflussen können. Diese wurden in der vorliegenden Studie jedoch nur ansatzweise berücksichtigt. So bestehen insbesondere Forschungsbedarfe im Hinblick auf potenziell adverse Effekte gestenbasierter Steuerung. Hierzu zählen u. a. eine unzureichende Gestenerkennung sowie ein mangelhaftes visuelles Feedback, die zu Irritationen oder Fehlinterpretationen bei der Interaktion führen können. Darüber hinaus kann bei längerer Nutzung eine körperliche Ermüdung auftreten, während gleichzeitig die Steuerungseffizienz mit zunehmender Distanz abnimmt. Auch die gleichzeitige Steuerung mehrerer Freiheitsgrade kann als kognitiv überfordernd wahrgenommen werden und die wahrgenommene Intuitivität der Benutzungsschnittstelle negativ beeinflussen. Schließlich wirken sich unterschiedliche Erfahrungsniveaus der Anwender auf die Interaktionsqualität aus und können die individuelle Anpassungsfähigkeit an neue Steuerungskonzepte begrenzen. Eine weitere Limitation liegt darin, dass aus alternativen Metriken wie dem sog. Durchsatz (vgl. Engl. Throughput) oder dem Fitts-Gesetz spezifische Ansätze zur Bewertung von Interaktionsfehlern für IAS abgeleitet werden können (Amini et al., 2025; Nielsen & Levy, 1994). Während der Throughput die Effizienz und Geschwindigkeit einer Interaktion unter Berücksichtigung der Präzision bewertet, ermöglicht das Fitts-Gesetz eine Vorhersage der benötigten Bewegungszeit in Abhängigkeit von Distanz und Zielgröße. Beide Metriken könnten daher zusätzliche Perspektiven auf Interaktionsfehler bieten, indem sie auch Effizienz und Genauigkeit im Kontext der Nutzerleistung berücksichtigen. Ihre

Nichtberücksichtigung könnte zu einer eingeschränkten Bewertung der tatsächlichen Interaktionsqualität führen, da dadurch wichtige Aspekte der Benutzererfahrung außer Acht gelassen werden.

Limitationen hinsichtlich des Fehlervergleichs: In Teilstudie 4.2 wurden die verschiedenen Fehlerarten mithilfe eines Forschungsprototyps innerhalb eines systematischen Versuchsaufbaus miteinander verglichen. Auch wenn die Ergebnisse den Sichtbarkeitsfehler im Vergleich zu den anderen Fehlertypen deutlich exponieren, sind diese Resultate nur eingeschränkt zu verallgemeinern. Vor allem die Metastudie von Merino et al. (2020) betont, dass Ergebnisse in XR-Studien stark von der Umgebung und dem Nutzungskontext abhängig sind und nicht-artifizielle Evaluationsbedingungen notwendig sind, um valide und realitätsnahe Resultate zu erhalten. Insbesondere die Manipulation des Sichtbarkeitsfehlers wirft hinsichtlich seiner spezifischen Bedingungen und der atypischen Präferenz – ausgedrückt durch eine negative Effektstärke (Sichtbarkeitsfehler verbessert teilweise die Faktoren innerhalb der Stichprobe) – im direkten Vergleich zu anderen Fehlertypen berechtigte Zweifel an seiner Validität auf. Aus diesem Grund werden vertiefte Forschungsaktivitäten empfohlen, um die Kontextabhängigkeit und differenzielle Wirkung verschiedener Fehlertypen in realitätsnahen XR-Szenarien systematisch zu untersuchen.

Limitationen hinsichtlich der Akzeptanzmanipulation mittels Konfidenzanzeige: Obwohl in Teilstudie 4.3 ein signifikanter Einfluss einer Konfidenzanzeige auf die Akzeptanz nachgewiesen werden konnte, unterliegen die Ergebnisse mehreren Einschränkungen. Zum einen war die Stichprobenverteilung zwischen den Versuchsgruppen erheblich un- ausgeglichen (z. B. $n = 6$ vs. $n = 18$), was zu einer asymmetrischen Häufigkeitsverteilung innerhalb der Stichprobe führte und trotz ausreichender Teststärke die Aussagekraft der Befunde einschränkt. Darüber hinaus wurde die Manipulation der Konfidenzanzeige im Rahmen des kompromissbehafteten Designs der Verbundstudie ausschließlich für Interaktionsfehler implementiert. Eine systematische Übertragung auf andere Fehlertypen, wie etwa Sichtbarkeits- oder Datenfehler, erfolgte nicht. Somit bleibt offen, ob die beobachteten Manipulationseffekte tatsächlich generalisierbar sind oder spezifisch auf Interaktionsfehler beschränkt bleiben. Zudem basierte die Ausgestaltung der Konfidenzanzeige nicht auf einer systematischen Herleitung aus Vorstudien oder etablierten

Modellen. Die tatsächliche Wahrnehmung und Verarbeitung der Konfidenzanzeige durch die Teilnehmenden wurden nicht weiterführend erfasst. Es bleibt daher unklar, in welchem Umfang die Anzeige aktiv überwacht, verstanden oder lediglich oberflächlich wahrgenommen wurde (Gegoff et al., 2023). Darüber hinaus erfolgte aufgrund des explorativen Charakters der Untersuchung keine a-priori-Stichprobenumfangsplanung. Die Stichprobengröße wurde studienbegleitend angepasst, wobei eine statistische Teststärke von 80 % angestrebt wurde. Mit der vorliegenden Stichprobe konnte jedoch lediglich eine Teststärke von 74 % erreicht werden. Objektive Maße wie Eye-Tracking, wie sie beispielsweise von Gegoff et al. (2023, S. 14) vorgeschlagen werden, könnten in zukünftigen Studien wichtige Erkenntnisse zur Wahrnehmung und Nutzung der Konfidenzanzeige liefern.

5.4 Fazit

Im Rahmen dieser Dissertation wurde entsprechend der forschungsleitenden Fragestellung untersucht, welche Auswirkungen Fehler innerhalb der Mensch-Maschine-Interaktion bei immersiven Assistenzsystemen auf Leistung, Beanspruchung und Akzeptanz haben und inwiefern adverse Effekte durch gezielte Maßnahmen reduziert werden können. Motiviert durch praktische Erfahrungen aus der eigenen industriellen Anwendung und durch den Forschungsbedarf der kritischen Literaturrecherche, wurden systematische Untersuchungen zu den Fehlerwirkungen durchgeführt, um wissenschaftlich belastbare Erkenntnisse für den Einsatz und die Gestaltung künftiger Systeme zu gewinnen.

Die empirischen Ergebnisse zeigen, dass Fehler in IAS signifikante Einflüsse auf die untersuchten Primärfaktoren haben, wobei sich die Auswirkungen je nach Fehlerart unterschiedlich manifestieren. Fehlerhafte Daten beeinflussen insbesondere die Leistung und Akzeptanz negativ. Konkret konnte festgestellt werden, dass fehlerhafte Daten zu einer Reduktion der Nutzerleistung mit mittlerer Effektstärke ($Co'd = 0,69$) führen, wobei ebenfalls ein signifikanter Einfluss auf die Akzeptanz ($Co'd = 0,53$) nachweisbar war. Die Beanspruchung sowie die physiologischen Stressparameter hingegen zeigten keine signifikanten Effekte, was darauf hinweist, dass fehlerhafte Daten primär kognitive und

einstellungsbezogene und weniger beanspruchungsbezogene Reaktionsmuster hervorgerufen.

Ebenfalls zeigen fehlerhafte Visualisierungen einen maßgeblichen Einfluss, primär auf die Leistung der Nutzer. Deskriptiv zeigten sich höhere Effektstärken in VR-Umgebungen ($Co'd = 1,15$) als in AR-Umgebungen ($Co'd = 0,72$). Auf die Akzeptanz und Beanspruchung hatten Visualisierungsfehler hingegen insgesamt keine signifikant negativen Auswirkungen, mit Ausnahme der Frustrationsdimension in VR, bei der eine hohe Effektstärke ($Co'd = 1,12$) festgestellt wurde. Diese Ergebnisse unterstreichen, dass eine mangelhafte Visualisierung vor allem die Aufgabenbewältigung behindert, während sich in AR-Settings keine signifikante Erhöhung der Beanspruchung zeigte.

Besonders prominent zeigen sich die Effekte bei fehlerhafter Interaktion. Sowohl bei Sprachinteraktionen (ASR) als auch bei Gesteninteraktionen wurden signifikant hohe Effektstärken auf Leistung, Akzeptanz und Beanspruchung nachgewiesen. Die Auswirkungen bei der Sprachinteraktion waren hierbei besonders ausgeprägt: Akzeptanz ($Co'd = 1,55$) und Leistung ($Co'd = 0,74$) wurden stark negativ beeinflusst. Bei der Gestensteuerung konnten ebenfalls mittlere bis hohe Effektstärken auf die Leistung ($Co'd = 1,08$), die Akzeptanz ($Co'd = 0,529$) und die Beanspruchung ($Co'd = 0,770$) nachgewiesen werden. Diese Ergebnisse verdeutlichen, dass Interaktionsfehler besonders kritisch sind, da sie sowohl die performative als auch die nutzerbezogene Dimension erheblich beeinträchtigen.

Ein maßgebliches Ergebnis dieser Arbeit betrifft den Einfluss unzureichender Trainingsqualität. Die Ergebnisse zeigen, dass unzureichende Trainingsszenarien starke negative Auswirkungen auf die betrachteten Primärfaktoren hatten. Für Leistung, Akzeptanz und Beanspruchung konnten sehr hohe Effektstärken ($Co'd \geq 1,17$) bei maximaler Teststärke (Power = 1,00) nachgewiesen werden. Die Erkenntnis, dass nicht nur Inhalte und Umfang, sondern auch die Qualität des Trainingsszenarios entscheidend für die spätere Nutzungseffizienz und -akzeptanz ist, liefert einen wichtigen Impuls für die Gestaltung künftiger Einführungs- und Schulungskonzepte von IAS. Flankierend zur Untersuchung wurde innerhalb des Forschungsverbunds ein optimiertes Trainingsszenario entwickelt, das sich in wesentlichen Aspekten von der bisherigen Herangehensweise im Stand der Forschung unterscheidet. Während bestehende Studien überwiegend den

Fokus auf Lernprozesse mithilfe immersiver Systeme legen, zielt das entwickelte Konzept darauf ab, Nutzer gezielt für den Einsatz immersiver Assistenzsysteme auszubilden. Das Trainingsszenario orientiert sich dabei an praxisrelevanten Anforderungen industrieller Anwendungsfälle. Durch diese gezielte Vorbereitung konnte demonstriert werden, dass nicht die bloße Nutzung immersiver Medien, sondern die methodisch fundierte Befähigung zum Umgang mit IAS entscheidend für die spätere Leistung, Akzeptanz und Systemintegration ist.

Die differenzierte Betrachtung der Fehlertypen zeigte weiterhin, dass sich die Auswirkungen auf Akzeptanz und Beanspruchung signifikant zwischen den Fehlerarten unterscheiden. Für die Leistung sowie die physiologischen Parameter konnten dagegen keine signifikanten Unterschiede zwischen den Fehlertypen festgestellt werden. Diese Befunde verdeutlichen, dass Fehler nicht homogen wirken, sondern spezifische, fehlerartenspezifische Mechanismen entfalten, die bei der Systemgestaltung berücksichtigt werden müssen.

Ein weiterer zentraler Befund betrifft die positive Wirkung der Anzeige der Systemkonfidenz auf die Akzeptanz. Durch die transparente Kommunikation der Systemzuverlässigkeit konnte die Akzeptanz auch bei Vorliegen des Interaktionsfehlers signifikant gesteigert werden ($\text{Co'd} = 1,12$; $\text{Power} = 0,74$). Besonders bemerkenswert ist hierbei, dass die Akzeptanz im fehlerhaften Zustand mit Konfidenzanzeige sogar höher lag als im fehlerfreien Zustand ohne Konfidenzanzeige. Dieses Erkenntnis unterstreicht die Bedeutung von Transparenz und Erwartungsmanagement für die Akzeptanz von IAS.

Methodisch leistet die Arbeit einen relevanten Beitrag innerhalb ihres Forschungsfelds. Zum einen wurde mittels des übergeordneten Forschungsdesigns ein kontextualisiertes Forschungsmodell entwickelt, welches aufbauend auf einschlägigen Vorarbeiten spezifisch auf den industriellen Kontext immersiver Assistenzsysteme operationalisiert wurde. Zum anderen wurde mit dem Index Leistung eine Metrik eingeführt, die den Zielkonflikt zwischen Bearbeitungsdauer und Richtigkeit von Aufgaben adressiert. Ferner wurde die ITR-Metrik, ursprünglich für Brain-Computer-Interfaces entwickelt, erfolgreich auf die Bewertung von Sprachinteraktionen bei IAS adaptiert. Zur Bewertung von Trainingsfehlern wurde ein Cut-off-Verfahren entwickelt, das teilaufgabenbezogene Strafzeiten auf Grundlage des 75. Perzentils der Bearbeitungsdauer abbildet und

somit eine realitätsnahe Leistungsbewertung ermöglicht – selbst wenn Probanden fachliche Teilaufgaben zusammen mit dem IAS nicht erfolgreich beenden. Darüber hinaus wurde durch die differenzierte Betrachtung von vier unterschiedlichen Einzelstudien eine methodisch valide Bewertung der Fehlerwirkungen durchgeführt. Besonders hervorzuheben ist der Einsatz der Change-Score-Methode, die intraindividuelle Faktoren minimiert und damit zur Erhöhung der internen Validität der Ergebnisse beiträgt.

Insgesamt leistet diese Dissertation einen substanziellen Beitrag zur wissenschaftlichen Fundierung der Entwicklung und Bewertung von IAS. Durch die erstmalige, umfassende und systematische Untersuchung der Auswirkungen unterschiedlicher Fehlerarten innerhalb der MMI, die Analyse der Trainingsqualität sowie die Implementierung eines Mechanismus zur Reduzierung adverser Akzeptanzeffekte wird der Forschungsstand erweitert. Die gewonnenen Erkenntnisse zeigen, dass die erfolgreiche Implementierung von IAS nicht allein von der Frage der technischen Leistungsfähigkeit abhängt, sondern maßgeblich von der Qualität der Mensch-System-Interaktion, der Resilienz gegenüber Fehlern sowie der Gestaltung nutzerzentrierter Implementierungsstrategien abhängt. Die Arbeit liefert damit wertvolle Implikationen sowohl für die wissenschaftliche Weiterentwicklung als auch für die praktische Umsetzung immersiver Assistenzsysteme in industriellen Kontexten.

Literaturverzeichnis

- Alessa, F. M., Alhaag, M. H., Al-harkan, I. M., Nasr, M. M., Kaid, H. & Hammami, N. (2023). Evaluating Physical Stress across Task Difficulty Levels in Augmented Reality-Assisted Industrial Maintenance. *Applied Sciences*, 14(1), 363. <https://doi.org/10.3390/app14010363>
- Ali, B. M. & Younes, B. (2013). The impact of information systems on user performance: An exploratory study. *Journal of Knowledge Management, Economics and Information Technology*, 3(2), 128–154.
- Al-kfairy, M., Ahmed, S. & Khalil, A. (2024). Factors impacting users' willingness to adopt and utilize the metaverse in education: A systematic review. *Computers in Human Behavior Reports*, 15, 100459. <https://doi.org/10.1016/j.chbr.2024.100459>
- Alqahtani, M. A., Alamri, M. M., Sayaf, A. M. & Al-Rahmi, W. M. (2022). Exploring student satisfaction and acceptance of e-learning technologies in Saudi higher education. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.939336>
- Alsuraykh, N. H., Wilson, M. L., Tennent, P. & Sharples, S. (2019). How Stress and Mental Workload are Connected. *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 371–376. <https://doi.org/10.1145/3329189.3329235>
- Amini, M., Stuerzlinger, W., Teather, R. J. & Batmaz, A. U. (2025). A Systematic Review of Fitts' Law in 3D Extended Reality. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–25. <https://doi.org/10.1145/3706598.3713623>

- Anandapadmanaban, E. (2020). *vMCC: A Virtual Reality Framework for Augmenting Mission Control Operations* [Master Thesis, Massachusetts Institute of Technology].
<https://hdl.handle.net/1721.1/129128>
- Arning, K., Ziefle, M., Li, M. & Kobbelt, L. (2012). Insights into user experiences and acceptance of mobile indoor navigation devices. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, 1–10.
<https://doi.org/10.1145/2406367.2406417>
- Avizienis, A., Laprie, J.-C., Randell, B. & Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, 1(1), 11–33. <https://doi.org/10.1109/TDSC.2004.2>
- Azeroual, O., Saake, G., Abuosba, M. & Schöpfel, J. (2020). Data Quality as a Critical Success Factor for User Acceptance of Research Information Systems. *Data*, 5(2), 35. <https://doi.org/10.3390/data5020035>
- Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4), 355–385. <https://doi.org/10.1162/pres.1997.6.4.355>
- Bahaei, S. S. & Gallina, B. (2019). Augmented Reality-extended Humans: Towards a Taxonomy of Failures – Focus on Visual Technologies. *Proceedings of the 29th European Safety and Reliability Conference (ESREL)*, 207–214.
https://doi.org/10.3850/978-981-11-2724-3_0922-cd
- Bahaei, S. S., Gallina, B., Laumann, K. & Skogstad, M. R. (2019). Effect of Augmented Reality on Faults Leading to Human Failures in Socio-technical Systems. *2019 4th International Conference on System Reliability and Safety (ICSRS)*, 236–245.
<https://doi.org/10.1109/ICSRS48664.2019.8987586>

- Bala, H. & Venkatesh, V. (2016). Adaptation to Information Technology: A Holistic Nomological Network from Implementation to Job Outcomes. *Management Science*, 62(1), 156–179. <https://doi.org/10.1287/mnsc.2014.2111>
- Barnard, Y., Bradley, M. D., Hodgson, F. & Lloyd, A. D. (2013). Learning to use new technologies by older adults: Perceived difficulties, experimentation behaviour and usability. *Computers in Human Behavior*, 29(4), 1715–1724. <https://doi.org/10.1016/j.chb.2013.02.006>
- BCI Kleve. (o. D.). *BCI ITR Kalkulator*. Abgerufen 10. März 2024, von <https://bci-lab.hochschule-rhein-waal.de/de/itr.html>
- Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, 41(2), 257–278. <https://doi.org/10.1111/j.2044-8317.1988.tb00901.x>
- Bekier, M. & Molesworth, B. R. C. (2017). Altering users' acceptance of automation through prior automation exposure. *Ergonomics*, 60(6), 745–753. <https://doi.org/10.1080/00140139.2016.1216610>
- Beneteau, E. (2020). Who Are You Asking?: Qualitative Methods for Involving AAC Users as Primary Research Participants. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3313831.3376883>
- Bershadskyy, D., Krüger, J., Çalık, G., Otto, S., Zabel, S., Greif, J. & Heyer, R. (2022). A Laboratory Experiment on Using Different Financial-Incentivization Schemes in Software-Engineering Experimentation (Version 9). arXiv. <https://doi.org/10.48550/ARXIV.2202.10985>

- Blaine, B. E. (2018). Winsorizing. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vol. 4, pp. 1817–1818). SAGE Publications.
<https://doi.org/10.4135/9781506326139.n747>
- Blanca, M. J., Alarcón, R. & Arnau, J. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29(4), 552–557. <https://doi.org/10.7334/psicothema2016.383>
- Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R. & Bono, R. (2023). Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power. *Psicothema*, 35(1), 21–29. <https://doi.org/10.7334/psicothema2022.292>
- Bläsing, D. & Bornewasser, M. (2021). Influence of Increasing Task Complexity and Use of Informational Assistance Systems on Mental Workload. *Brain Sciences*, 11(1), 102. <https://doi.org/10.3390/brainsci11010102>
- Bräuer, P. & Mazarakis, A. (2018). AR in order-picking – experimental evidence with Microsoft HoloLens. *Mensch und Computer 2018, Workshop-band*. <https://doi.org/10.18420/MUC2018-WS07-0463>
- Brauner, P., Philipsen, R., Calero Valdez, A. & Ziefle, M. (2019). What happens when decision support systems fail? — The importance of usability on performance in erroneous systems. *Behaviour & Information Technology*, 38(12), 1225–1242. <https://doi.org/10.1080/0144929X.2019.1581258>
- Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, 17(4), 600.
- Cao, Y., Qian, X., Wang, T., Lee, R., Huo, K. & Ramani, K. (2020). An Exploratory Study of Augmented Reality Presence for Tutoring Machine Tasks. *Proceedings of the*

2020 CHI Conference on Human Factors in Computing Systems, 1–13.

<https://doi.org/10.1145/3313831.3376688>

Caputo, A., Bartolomioli, R., Orso, V., Mingardi, M., Da Granaiola, L., Gamberini, L. &

Giachetti, A. (2024). Comparison of deviceless methods for distant object manipulation in mixed reality. *Computers & Graphics*, 122, 103959.

<https://doi.org/10.1016/j.cag.2024.103959>

Caputo, A., Jacota, S., Krayevskyy, S., Pesavento, M., Pellacini, F. & Giachetti, A. (2020).

XR-Cockpit: A comparison of VR and AR solutions on an interactive training station. 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), 1, 603–610. <https://ieeexplore.ieee.org/abstract/document/9212043/>

Chang, H.-Y., Binali, T., Liang, J.-C., Chiou, G.-L., Cheng, K.-H., Lee, S. W.-Y. & Tsai, C.-

C. (2022). Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact. *Computers & Education*, 191, 104641. <https://doi.org/10.1016/j.compedu.2022.104641>

Charness, G., Gneezy, U. & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8.

Chaudary, A. (2023). The role of interaction quality in use of AI-based voice-assistant systems: Evidence from Finland [Master Thesis, Jyväskylä University School of Business and Economics]. <https://jyx.jyu.fi/handle/123456789/85968>

Chester, D. S. & Lasko, E. N. (2021). Construct Validation of Experimental Manipulations in Social Psychology: Current Practices and Recommendations for the Future.

Perspectives on Psychological Science, 16(2), 377–395.

<https://doi.org/10.1177/1745691620950684>

Cimini, C., Pirola, F., Pinto, R. & Cavalieri, S. (2020). A human-in-the-loop manufacturing control architecture for the next generation of production systems. *Journal of Manufacturing Systems*, 54, 258–271. <https://doi.org/10.1016/j.jmsy.2020.01.002>

Cockburn, A., Masson, D., Gutwin, C., Palanque, P., Goguey, A., Yung, M., Gris, C. & Trask, C. (2019). Design and evaluation of braced touch for touchscreen input stabilisation. *International Journal of Human-Computer Studies*, 122, 21–37. <https://doi.org/10.1016/j.ijhcs.2018.08.005>

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Dal Seno, B., Matteucci, M. & Mainardi, L. T. (2010). The Utility Metric: A Novel Method to Assess the Overall Performance of Discrete Brain–Computer Interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(1), 20–28. <https://doi.org/10.1109/TNSRE.2009.2032642>

Davis, C. (Hrsg.). (2023). *Statistical testing with jamovi psychology: Second Edition*. VOR Press.

Day, A., Barber, L. K. & Tonet, J. (2019). Information Communication Technology and Employee Well-Being: Understanding the “iParadox Triad” at Work. In R. N. Landers (Hrsg.), *The Cambridge Handbook of Technology and Employee Behavior* (S. 580–607). Cambridge University Press. <https://doi.org/10.1017/9781108649636.022>

- De Giorgio, A., Monetti, F. M., Maffei, A., Romero, M. & Wang, L. (2023). Adopting extended reality? A systematic review of manufacturing training and teaching applications. *Journal of Manufacturing Systems*, 71, 645–663. <https://doi.org/10.1016/j.jmsy.2023.10.016>
- Derby, J. L., Rarick, C. T. & Chaparro, B. S. (2019). Text Input Performance with a Mixed Reality Head-Mounted Display (HMD). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 1476–1480. <https://doi.org/10.1177/1071181319631279>
- Derby, J. L., Rickel, E. A., Harris, K. J., Lovell, J. A. & Chaparro, B. S. (2020). “We Didn’t Catch That!” Using Voice Text Input on a Mixed Reality Headset in Noisy Environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1), 2102–2106. <https://doi.org/10.1177/1071181320641509>
- Dey, A., Billingham, M., Lindeman, R. W. & Swan, J. E. (2018). A Systematic Review of 10 Years of Augmented Reality Usability Studies: 2005 to 2014. *Frontiers in Robotics and AI*, 5. <https://doi.org/10.3389/frobt.2018.00037>
- Dombrowski, U., Riechel, C. & Evers, M. (2014). Industrie 4.0 – Die Rolle des Menschen in der vierten industriellen Revolution. *Industrie*, 4, 129–153.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer-Verlag. <https://doi.org/10.1007/978-3-642-41089-5>
- Dörner, R. & Horst, R. (2021). Conveying Firsthand Experience: The Circuit Parcours Technique for Efficient and Engaging Teaching in Courses about Virtual Reality and Augmented Reality. *Eurographics 2021 - Education Papers*. <https://doi.org/10.2312/EGED.20211002>

- Eckert, M., Blex, M. & Friedrich, C. M. (2018). Object Detection Featuring 3D Audio Localization for Microsoft HoloLens— A Deep Learning based Sensor Substitution Approach for the Blind: *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, 555–561. <https://doi.org/10.5220/0006655605550561>
- Empatica. (o. D.). E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors. Abgerufen 9. Januar 2025, von <https://www.empatica.com/research/e4>
- Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 32–64. <https://doi.org/10.1518/001872095779049543>
- Estrella, J. & Mendoza, B. (2019). Voice Controlled Augmented Reality: A Comparison of Speech Recognition Tools for AR Applications. *Proceedings of the 2019 Fall ASEE Midatlantic Conference, New York, NY, USA*, 1–2. <https://openlab.citytech.cuny.edu/bmendoza/files/2020/05/Estrella-MendozaMidAtlantic-ASEE-Fall2019.pdf>
- Fantini, P., Pinzone, M. & Taisch, M. (2020). Placing the operator at the centre of Industry 4.0 design: Modelling and assessing human activities within cyber-physical systems. *Computers & Industrial Engineering*, 139, 105058. <https://doi.org/10.1016/j.cie.2018.01.025>
- Fellmann, M., Robert, S., Büttner, S., Mucha, H. & Röcker, C. (2017). Towards a Framework for Assistance Systems to Support Work Processes in Smart Factories. In A. Holzinger, P. Kieseberg, A. M. Tjoa & E. Weippl (Hrsg.), *Machine Learning and*

Knowledge Extraction (S. 59–68). Springer-Verlag. https://doi.org/10.1007/978-3-319-66808-6_5

Filippidou, F. & Moussiades, L. (2020). A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In I. Maglogiannis, L. Iliadis & E. Pimenidis (Hrsg.), *Artificial Intelligence Applications and Innovations* (S. 73–82). Springer-Verlag. https://doi.org/10.1007/978-3-030-49161-1_7

Fink, K. (2020). Cognitive assistance systems for manual assembly throughout the German manufacturing industry. *Journal of Applied Leadership and Management*, 8, 38–53.

Fuglseth, A. M. & Sørenbø, Ø. (2014). The effects of technostress within the context of employee use of ICT. *Computers in Human Behavior*, 40, 161–170.

Fuhrmann, F., Weber, A., Ladstätter, S., Dietrich, S. & Rella, J. (2021). Multimodal Interaction in the Production Line—An OPC UA-based Framework for Injection Molding Machinery. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 837–838. <https://doi.org/10.1145/3462244.3481300>

Funk, M., Hartwig, M. & Wischniewski, S. (2019). Evaluation of assistance systems for manual assembly work. *2019 IEEE/SICE International Symposium on System Integration (SII)*, 794–798. <https://ieeexplore.ieee.org/abstract/document/8700420/>

Gabbard, J. L., Swan, J. E., Hix, D., Kim, S.-J. & Fitch, G. (2007). Active Text Drawing Styles for Outdoor Augmented Reality: A User-Based Study and Design Implications. *2007 IEEE Virtual Reality Conference*, 35–42. <https://doi.org/10.1109/VR.2007.352461>

- Garcia Fracaro, S., Glassey, J., Bernaerts, K. & Wilk, M. (2022). Immersive technologies for the training of operators in the process industry: A Systematic Literature Review. *Computers & Chemical Engineering*, 160, 107691. <https://doi.org/10.1016/j.compchemeng.2022.107691>
- Garzón, J. & Acevedo, J. (2019). Meta-analysis of the impact of Augmented Reality on students' learning gains. *Educational Research Review*, 27, 244–260. <https://doi.org/10.1016/j.edurev.2019.04.001>
- Gegoff, I., Tatasciore, M., Bowden, V., McCarley, J., & Loft, S. (2023). Transparent Automated Advice to Mitigate the Impact of Variation in Automation Reliability. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 66(8), 2008–2024. <https://doi.org/10.1177/00187208231196738>
- Georg, A., Meyn, C., & Peter, G. (2013). Belastung und Beanspruchung. In *Nomos Verlagsgesellschaft mbH & Co. KG eBooks* (S. 113–118). <https://doi.org/10.5771/9783845273099-113>
- Gertman, D., Blackman, H., Marble, J., Byers, J., & Smith, C. (2005). The SPAR-H human reliability analysis method. *US Nuclear Regulatory Commission*, 230(4), 35.
- Giorgi, A., Ronca, V., Vozzi, A., Sciaraffa, N., Di Florio, A., Tamborra, L., Simonetti, I., Aricò, P., Di Flumeri, G., Rossi, D., & Borghini, G. (2021). Wearable Technologies for Mental Workload, Stress, and Emotional State Assessment during Working-Like Tasks: A Comparison with Laboratory Technologies. *Sensors*, 21(7), 2332. <https://doi.org/10.3390/s21072332>

- Göthlich, S. E. (2009). Zum Umgang mit fehlenden Daten in großzahligen empirischen Erhebungen. In *Gabler Verlag eBooks* (S. 119–135). https://doi.org/10.1007/978-3-322-96406-9_9
- G*Power. (2024, 11. November). <https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower>
- Grabowski, M. (2015). Research on Wearable, Immersive Augmented Reality (WIAR) Adoption in Maritime Navigation. *Journal of Navigation*, 68(3), 453–464. <https://doi.org/10.1017/S0373463314000873>
- Graichen, L., & Graichen, M. (2023). Do Users Tolerate Errors? Effects of Observed Failures on the Subjective Evaluation of a Gesture-Based Virtual Reality Application. In H. P. Da Silva & P. Cipresso (Hrsg.), *Computer-Human Interaction Research and Applications* (S. 232–243). Springer-Verlag. https://doi.org/10.1007/978-3-031-49425-3_14
- Graser, S., & Böhm, S. (2022). *A Systematic Literature Review on Technology Acceptance Research on Augmented Reality in the Field of Training and Education*. In *CENTRIC 2022: The Fifteenth International Conference on Advances in Human Oriented and Personalized Mechanisms, Technologies, and Services* (Vol. 12). Lisbon
- Hadorn, B., Courant, M., & Hirsbrunner, B. (2016). *Towards human-centered cyber-physical systems: A modeling approach*. https://www.academia.edu/download/107389062/IWP16_01_diuf.pdf
- Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904–908. <https://doi.org/10.1177/154193120605000909>

- He, X., & Zhou, K. (2024). Display in the Air: Balancing Distraction and Workload in AR Glasses Interfaces for Driving Navigation. *arXiv (Cornell University)*.
<https://doi.org/10.48550/ARXIV.2404.18357>
- Hecht, T., Kratzert, S., & Bengler, K. (2020). The Effects of a Predictive HMI and Different Transition Frequencies on Acceptance, Workload, Usability, and Gaze Behavior during Urban Automated Driving. *Information*, 11(2), 73.
<https://doi.org/10.3390/info11020073>
- Hendy, K. C. (2003). *A Tool for Human Factors Accident Investigation, Classification and Risk Management*. <https://apps.dtic.mil/sti/citations/ADA623696>
- Herzberg, F. (2015). Motivation-hygiene theory. *Organizational behavior* 1, 61–74.
- Hirsch-Kreinsen, H. (2014). Wandel von Produktionsarbeit– „Industrie 4.0 “. *WSI-Mitteilungen*, 67(6), 421–429.
- Hirsch-Kreinsen, H., Ittermann, P., & Niehaus, J. (2018). Digitalisierung industrieller Arbeit: Die Vision Industrie 4.0 und ihre sozialen Herausforderungen. Nomos Verlag.
- HoloLens 2: Technische Daten und Funktionen – Microsoft HoloLens 2*. (o. D.). Microsoft Store. Abgerufen 9. Januar 2025, von <https://www.microsoft.com/de-de/d/holo-lens-2/91pnzzznzwc>
- Institut für Statistik, S. (2020). Der t-, Welch- und U-Test im psychotherapiewissenschaftlichen Forschungskontext. *SFU Forschungsbulletin*, 87–105.
<https://doi.org/10.15135/2020.8.1.87-105>
- jamovi—Open statistical software for the desktop and cloud. (o. D.).
<https://www.jamovi.org/>

- Janczyk, M., & Pfister, R. (2023). Repeated-Measures Analysis of Variance (ANOVA). In *Springer eBooks* (S. 145–155). https://doi.org/10.1007/978-3-662-66786-6_10
- Jennings, M. A., & Cribbie, R. A. (2016). Comparing pre-post change across groups: Guidelines for choosing between difference scores, ANCOVA, and residual change scores. *Journal of Data Science*, 14(2), 205–229.
- Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1), 510–524.
- Kallweit, K. (2019). *RETAIL INNOVATIONS–THE USER ACCEPTANCE OF MOBILE-SERVICE* [kumulative Dissertation, Georg-August-Universität Göttingen]. <https://dx.doi.org/10.53846/goediss-7462>
- Kaltenhauser, A., Stefanidi, E., & Schöning, J. (2024). Playing with Perspectives and Unveiling the Autoethnographic Kaleidoscope in HCI – A Literature Review of Autoethnographies. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20. <https://doi.org/10.1145/3613904.3642355>
- Kępuska, V. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). *International Journal of Engineering Research and Applications*, 07(03), 20–24. <https://doi.org/10.9790/9622-0703022024>
- Kim, S. S., & Malhotra, N. K. (2005). A Longitudinal Model of Continued IS Use: An Integrative View of Four Mechanisms Underlying Postadoption Phenomena. *Management Science*, 51(5), 741–755. <https://doi.org/10.1287/mnsc.1040.0326>
- Kohl, C. O. (2021). *Prognose der Akzeptanz von Innovationen am Beispiel autonomer Fahrzeuge* [Dissertation, Technische Universität München]. <https://nbn-resolving.org/urn:nbn:de:bvb:91-diss-20211115-1609670-1-7>

- Konopka, B., Hönemann, K., & Wiesche, M. (2024a). Augmented Reality, Trust and Technology Malfunctions: An Initial Theory and Empirical Study. *Proceedings of the 2024 Computers and People Research Conference*, 1–8. <https://doi.org/10.1145/3632634.3655884>
- Konopka, B., Hönemann, K., & Wiesche, M. (2024b). Trust Calibration in Augmented Reality-based Decision Support Systems. *2024 26th International Conference on Business Informatics (CBI)*, 168–177. <https://doi.org/10.1109/CBI62504.2024.00028>
- Korpelainen, E., & Kira, M. (2010). Employees' choices in learning how to use information and communication technology systems at work: Strategies and approaches. *International Journal of Training and Development*, 14(1), 32–53. <https://doi.org/10.1111/j.1468-2419.2009.00339.x>
- Korunka, C., & Vitouch, O. (1999). Effects of the implementation of information technology on employees' strain and job satisfaction: A context-dependent approach. *Work & Stress*, 13(4), 341–363. <https://doi.org/10.1080/02678379950019798>
- Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., & Woźniak, P. W. (2023). A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Computing Surveys*, 55(13s), 1–39. <https://doi.org/10.1145/3582272>
- Kraus, M., Ludwig, B., Minker, W., & Wagner, N. (2021). Assistenzsysteme. In *Handbuch der Künstlichen Intelligenz* (S. 859–906). De Gruyter Oldenbourg. <https://doi.org/10.1515/9783110659948-020>
- Lee, S., & Lee, D. K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology*, 71(5), 353–360. <https://doi.org/10.4097/kja.d.18.00242>

- Lin, Y., & Yu, Z. (2023). A meta-analysis of the effects of augmented reality technologies in interactive learning environments (2012–2022). *Computer Applications in Engineering Education*, 31(4), 1111–1131. <https://doi.org/10.1002/cae.22628>
- Liu, Y., Dong, H., Zhang, L., & Saddik, A. E. (2022). *Technical Evaluation of HoloLens for Multimedia: A First Look*. <https://doi.org/10.48550/ARXIV.2212.12907>
- Lolambean. (o. D.). *Interactable object—Mixed Reality*. Abgerufen 23. Januar 2023, von <https://learn.microsoft.com/en-us/windows/mixed-reality/design/interactable-object>
- Longo, F., Nicoletti, L., & Padovano, A. (2017). Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context. *Computers & Industrial Engineering*, 113, 144–159. <https://doi.org/10.1016/j.cie.2017.09.016>
- Lorrig, P., & Daw, Z. (2024). Advances and Challenges Towards Enabling Human-AI-Teaming Applications for Flight Deck Operations. *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, 1–10. <https://doi.org/10.1109/DASC62030.2024.10749425>
- Ludwig, B. (2015). *Planbasierte Mensch-Maschine-Interaktion in multimodalen Assistenzsystemen*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-44819-9>
- Lüpsen, H. (2019). *Multiple Mittelwertvergleiche—Parametrisch und nichtparametrisch—Sowie α -Adjustierungen mit praktischen Anwendungen mit R und SPSS. Version 2.0*. Universitäts- und Stadtbibliothek Köln.
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human*

Factors: The Journal of the Human Factors and Ergonomics Society, 48(2), 241–256.

<https://doi.org/10.1518/001872006777724408>

Mariani, M. G., Curcuruto, M., & Gaetani, I. (2013). Training opportunities, technology acceptance and job satisfaction: A study of Italian organizations. *Journal of Workplace Learning*, 25(7), 455–475. <https://doi.org/10.1108/JWL-12-2011-0071>

Mark, B. G., Rauch, E., & Matt, D. T. (2021). Industrial Assistance Systems to Enhance Human–Machine Interaction and Operator’s Capabilities in Assembly. In D. T. Matt, V. Modrák & H. Zsifkovits (Hrsg.), *Implementing Industry 4.0 in SMEs* (S. 129–161). Springer International Publishing. https://doi.org/10.1007/978-3-030-70516-9_4

Marklin, R. W., Toll, A. M., Bauman, E. H., Simmins, J. J., LaDisa, J. F., & Cooper, R. (2022). Do Head-Mounted Augmented Reality Devices Affect Muscle Activity and Eye Strain of Utility Workers Who Do Procedural Work? Studies of Operators and Manhole Workers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 64(2), 305–323. <https://doi.org/10.1177/0018720820943710>

Marshall, B., Mills, R., & Olsen, D. (2008). The Role Of End-User Training In Technology Acceptance. *Review of Business Information Systems (RBIS)*, 12(2), 1–8. <https://doi.org/10.19030/rbis.v12i2.4384>

Martins, B. R., Jorge, J. A., & Zorzal, E. R. (2023). Towards augmented reality for corporate training. *Interactive Learning Environments*, 31(4), 2305–2323. <https://doi.org/10.1080/10494820.2021.1879872>

McGuirl, J. M., & Sarter, N. B. (2006). Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human*

Factors: The Journal of the Human Factors and Ergonomics Society, 48(4), 656–665.

<https://doi.org/10.1518/001872006779166334>

McMahan, B., & Rao, D. (2018). Listening to the World Improves Speech Command Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

<https://doi.org/10.1609/aaai.v32i1.11284>

Mekler, E. D., Bopp, J. A., Tuch, A. N., & Opwis, K. (2014). A systematic review of quantitative studies on the enjoyment of digital entertainment games. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 927–936.

<https://doi.org/10.1145/2556288.2557078>

Merenda, C., Suga, C., Gabbard, J., & Misu, T. (2019). Effects of Vehicle Simulation Visual Fidelity on Assessing Driver Performance and Behavior. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1679–1686. <https://doi.org/10.1109/IVS.2019.8813863>

Merino, L., Schwarzl, M., Kraus, M., Sedlmair, M., Schmalstieg, D., & Weiskopf, D. (2020). Evaluating mixed and augmented reality: A systematic literature review (2009-2019). *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 438–451. <https://ieeexplore.ieee.org/abstract/document/9284762/>

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12), 1321–1329.

Minow, A. (2021). Arbeitsphysiologische Untersuchungen beim Einsatz digitaler Assistenzsysteme für variantenreiche Montageprozesse in der Arbeitswelt 4.0 [kumulative Dissertation, Otto-von-Guericke-Universität Magdeburg]. <https://doi.org/10.25673/39299>

- Mkhonto, M., & Zuva, T. (2024). Technology Acceptance: A Critical Review of Technology Adoption Theories and Models. In R. Silhavy & P. Silhavy (Hrsg.), *Software Engineering Methods in Systems and Network Systems* (S. 414–428). Springer-Verlag. https://doi.org/10.1007/978-3-031-54813-0_38
- Moencks, M., Roth, E., Bohné, T., Romero, D., & Stahre, J. (2022). Augmented Workforce Canvas: A management tool for guiding human-centric, value-driven human-technology integration in industry. *Computers & Industrial Engineering*, 163, 107803. <https://doi.org/10.1016/j.cie.2021.107803>
- Müller, H., Pezzè, M., & Shaw, M. (2008). Visibility of control in adaptive systems. *Proceedings of the 2nd International Workshop on Ultra-Large-Scale Software-Intensive Systems*, 23–26. <https://doi.org/10.1145/1370700.1370707>
- Murrar, S., & Brauer, M. (2018). Mixed model analysis of variance. *The SAGE encyclopedia of educational research, measurement, and evaluation*, 1, 1075–1078.
- Nasirian, F., Ahmadian, M., & Lee, O.-K. D. (2017). *AI-based voice assistant systems: Evaluating from the interaction and trust perspectives*. https://www.researchgate.net/profile/Mohsen-Ahmadian/publication/322665841_AI-Based_Voice_Assistant_Systems_Evaluating_from_the_Interaction_and_Trust_Perspectives/links/5a67992b0f7e9b76ea8f054f/AI-Based-Voice-Assistant-Systems-Evaluating-from-the-Interaction-and-Trust-Perspectives.pdf
- Nielsen, J., & Levy, J. (1994). Measuring usability: Preference vs. performance. *Communications of the ACM*, 37(4), 66–75. <https://doi.org/10.1145/175276.175282>

Oculus Meta Quest 2. (o. D.). Abgerufen 9. Januar 2025, von <https://www.conrad.de/de/p/oculus-meta-quest-2-virtual-reality-system-usb-c-6gb-ram-128gb-qualcom-2445443.html>

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation: An Integrated Meta-Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(3), 476–488. <https://doi.org/10.1177/0018720813501549>

Osawa, N., & Ren, X. (2004). Virtual 3D gearbox widget technique for precise adjustment by hand motion in immersive VR. *IEICE TRANSACTIONS on Information and Systems*, 87(10), 2408–2414.

Ötting, S. K., Masjutin, L., Steil, J. J., & Maier, G. W. (2022). Let's Work Together: A Meta-Analysis on Robot Design Features That Enable Successful Human–Robot Interaction at Work. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 64(6), 1027–1050. <https://doi.org/10.1177/0018720820966433>

Pacaux-Lemoine, M.-P., Trentesaux, D., Zambrano Rey, G., & Millot, P. (2017). Designing intelligent manufacturing systems through Human-Machine Cooperation principles: A human-centered approach. *Computers & Industrial Engineering*, 111, 581–595. <https://doi.org/10.1016/j.cie.2017.05.014>

Pagano, R. R. (2013). *Understanding statistics in the behavioral sciences* (10th ed.). Wadsworth/Cengage Learning.

Pal, D., Arpnikanondt, C., Funilkul, S., & Varadarajan, V. (2019). User Experience with Smart Voice Assistants: The Accent Perspective. *2019 10th International Conference*

on Computing, Communication and Networking Technologies (ICCCNT), 1–6.

<https://doi.org/10.1109/ICCCNT45670.2019.8944754>

Palanque, P., Cockburn, A., & Gutwin, C. (2020). A Classification of Faults Covering the Human-Computer Interaction Loop. In *Lecture notes in computer science* (S. 434–448). https://doi.org/10.1007/978-3-030-54549-9_29

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

Park, E., Cho, M., & Ki, C.-S. (2009). Correct Use of Repeated Measures Analysis of Variance. *Annals of Laboratory Medicine*, 29(1), 1–9. <https://doi.org/10.3343/kjlm.2009.29.1.1>

Park, H., Ahn, D., & Lee, J. (2024). Lessons From Working in the Metaverse: Challenges, Choices, and Implications from a Case Study. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3613904.3641972>

Pettersson, I., Lachner, F., Frison, A.-K., Riener, A., & Butz, A. (2018). A Bermuda Triangle?: A Review of Method Application and Triangulation in User Experience Evaluation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3173574.3174035>

Pfennig, M. (2018). Prozessoptimierung mit digitaler Übergabe der Messwerte an einen Mehrspindeldrehautomaten via OPC-UA zur teilautomatisierten Werkzeugverschleißkorrektur und Anbindung der Messmittel an das CAQ-System [Bachelorarbeit, Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt].

- PlanAR. (o. D.). Technische Hochschule Würzburg-Schweinfurt. Abgerufen 9. Januar 2025, von <https://www.thws.de/forschung/institute/idee/abgeschlossene-idee-projekte/planar/>
- Pourmemar, M. (2024). *Predicting User Performance for the Evaluation of User Interfaces in Immersive Augmented Reality* [Dissertation, Concordia Universität Montréal]. <https://spectrum.library.concordia.ca/id/eprint/993383/>
- Quandt, M., & Freitag, M. (2021). A Systematic Review of User Acceptance in Industrial Augmented Reality. *Frontiers in Education*, 6. <https://doi.org/10.3389/feduc.2021.700760>
- Radicalad. (2021, 15. Juni). *Bounding box and App bar—Mixed Reality*. Microsoft Learn. <https://learn.microsoft.com/en-us/windows/mixed-reality/design/app-bar-and-bounding-box>
- Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008). The Consequences of Technostress for End Users in Organizations: Conceptual Development and Empirical Validation. *Information Systems Research*, 19(4), 417–433. <https://doi.org/10.1287/isre.1070.0165>
- 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*, 3, 257–266.
- Reichl, P., Egger, S., Schatz, R., & D’Alconzo, A. (2010). The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment. *2010 IEEE International Conference on Communications*, 1–5. <https://doi.org/10.1109/ICC.2010.5501894>

- Reuter, M., & Zacher, S. (2008). Regelungstechnik für Ingenieure: Analyse, Simulation und Entwurf von Regelkreisen; mit 77 Beispielen und 34 Aufgaben (12 Aufl.). Vieweg + Teubner.
- Richardson, N. T., Flohr, L., & Michel, B. (2018). Takeover Requests in Highly Automated Truck Driving: How Do the Amount and Type of Additional Information Influence the Driver–Automation Interaction? *Multimodal Technologies and Interaction*, 2(4), 68. <https://doi.org/10.3390/mti2040068>
- Richtlinie 2003/10/EG des Europäischen Parlaments und des Rates vom 6. Februar 2003 über Mindestvorschriften zum Schutz von Sicherheit und Gesundheit der Arbeitnehmer vor der Gefährdung durch physikalische Einwirkungen (Lärm) (17. Einzelrichtlinie im Sinne des Artikels 16 Absatz 1 der Richtlinie 89/391/EWG), EP, CONSIL, 042 OJ L (2003). <http://data.europa.eu/eli/dir/2003/10/oj/deu>
- Rohmert, W., & Rutenfranz, J. (1975). Arbeitswissenschaftliche Beurteilung der Belastung und Beanspruchung an unterschiedlichen industriellen Arbeitsplätzen. Der Bundesminister für Arbeit und Sozialordnung.
- Romero, D., Stahre, J., Wuest, T., Noran, O., Bernus, P., Fast-Berglund, Å., & Gorecky, D. (2016). Towards an operator 4.0 typology: A human-centric perspective on the fourth industrial revolution technologies. *Proceedings of the international conference on computers and industrial engineering (CIE46), Tianjin, China*, 29–31. https://www.researchgate.net/profile/David-Romero-32/publication/309609488_Towards_an_Operator_40_Typology_A_Human-

Centric_Perspective_on_the_Fourth_Industrial_Revolution_Technologies/links/58e435e7a6fdccc85bdf36fa/Towards-an-Operator-40-Typology-A-Human-Centric-Perspective-on-the-Fourth-Industrial-Revolution-Technologies.pdf

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3. Ed.). McGraw-Hill.

Rosilius, M., Hügél, L., Wirsing, B., Geuen, M., von Eitzen, I., Bräutigam, V., & Ludwig, B. (2024). Development and Evaluation of Training Scenarios for the Use of Immersive Assistance Systems. *Applied System Innovation*, 7(5), 73. <https://doi.org/10.3390/asi7050073>

Rosilius, M., Spiertz, M., Wirsing, B., Geuen, M., Bräutigam, V., & Ludwig, B. (2024). Impact of Industrial Noise on Speech Interaction Performance and User Acceptance when Using the MS HoloLens 2. *Multimodal Technologies and Interaction*, 8(2), 8. <https://doi.org/10.3390/mti8020008>

Rosilius, M., Wilhelm, M., Seitz, P., von Eitzen, I., Wirsing, B., Rabenstein, M., Decker, S., & Bräutigam, V. (2022). Equalization of the Visibility Loss between AR and Real Stimuli Sizes. *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 821–826. <https://ieeexplore.ieee.org/abstract/document/9974190/>

Rosilius, M., Wirsing, B., von Eitzen, I., Wilhelm, M., Schmitt, J., Engelmann, B., & Bräutigam, V. (2021). Evaluation of Visual Requirements and Software-Design for Immersive Visibility in Industrial Applications. *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 234–239. <https://ieeexplore.ieee.org/abstract/document/9585845/>

- Sadeghi Milani, A., Cecil-Xavier, A., Gupta, A., Cecil, J., & Kennison, S. (2024). A Systematic Review of Human–Computer Interaction (HCI) Research in Medical and Other Engineering Fields. *International Journal of Human–Computer Interaction*, 40(3), 515–536. <https://doi.org/10.1080/10447318.2022.2116530>
- Sadovitch, V. (2020). Fehlertolerante Anzeigengestaltung für Augmented Reality Head-up-Displays: Bewertung und Kompensation von Registrierungsfehlern im automobilen Kontext. Springer-Verlag.
- Sainburg, T., & Gentner, T. Q. (2021). Toward a Computational Neuroethology of Vocal Communication: From Bioacoustics to Neurophysiology, Emerging Tools and Future Directions. *Frontiers in Behavioral Neuroscience*, 15. <https://doi.org/10.3389/fnbeh.2021.811737>
- Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(3), 377–400. <https://doi.org/10.1177/0018720816634228>
- Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is It Really Robust?: Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption. *Methodology*, 6(4), 147–151. <https://doi.org/10.1027/1614-2241/a000016>
- Schweikl, S., & Obermaier, R. (2020). Lessons from three decades of IT productivity research: Towards a better understanding of IT-induced productivity effects.

Management Review Quarterly, 70(4), 461–507. <https://doi.org/10.1007/s11301-019-00173-6>

Sample, K., Olminkhof, A., Patel, M., Tieto, V., & Wen, Q. (2022, 16. Dezember). *MRTK2-Unity Developer Documentation—MRTK 2*. <https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/?view=mrtkunity-2022-05>

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G., & Ehlert, U. (2010). Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2), 410–417. <https://doi.org/10.1109/TITB.2009.2036164>

Shappell, S. A., & Wiegmann, D. A. (2000). *The human factors analysis and classification system—HFACS*. Scholarly Commons. <https://commons.erau.edu/publication/737/>

Sinlapanuntakul, W., Skilton, K., Mathew, J. N., Collard, A., & Chaparro, B. S. (2021). *Assessing Mixed Reality Voice Dictation with Background Noise*. Scholarly Commons. <https://commons.erau.edu/db-srs/2021/poster-session-two/4/>

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3. <https://doi.org/10.3389/frobt.2016.00074>

Sober, E. (1981). The Principle of Parsimony. *The British Journal for the Philosophy of Science*, 32(2), 145–156. <https://doi.org/10.1093/bjps/32.2.145>

Sochor, R., Kraus, L., Merkel, L., Braunreuther, S., & Reinhart, G. (2019). Approach to Increase Worker Acceptance of Cognitive Assistance Systems in Manual Assembly. *Procedia CIRP*, 81, 926–931. <https://doi.org/10.1016/j.procir.2019.03.229>

SoSci Survey ▶ Professional Online Surveys, European Product, GDPR compliant. (o. D.).

<https://www.soscisurvey.de/>

Staffa, S. J., & Zurakowski, D. (2020). Strategies in adjusting for multiple comparisons: A primer for pediatric surgeons. *Journal of Pediatric Surgery*, 55(9), 1699–1705.

<https://doi.org/10.1016/j.jpedsurg.2020.01.003>

Standage, D., Wang, D., Heitz, R. P., & Simen, P. (2016). *Toward a Unified View of the Speed-Accuracy Trade-Off: Behaviour, Neurophysiology and Modelling*. Frontiers Media SA.

<https://doi.org/10.3389/fnins.2015.00139>

Stefanidi, E., Bentvelzen, M., Woźniak, P. W., Kosch, T., Woźniak, M. P., Mildner, T., Schneegass, S., Müller, H., & Niess, J. (2023). Literature Reviews in HCI: A Review of Reviews. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–24. <https://doi.org/10.1145/3544548.3581332>

Stowers, K., Oglesby, J., Sonesh, S., Leyva, K., Iwig, C., & Salas, E. (2017). A Framework to Guide the Assessment of Human–Machine Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 59(2), 172–188.

<https://doi.org/10.1177/0018720817695077>

Strange, A. (2019, 12. März). *Microsoft's HoloLens 2 team answers more questions about biometric security, audio & hand Tracking*. Next Reality. <https://hololens.reality.news/news/microsofts-hololens-2-team-answers-more-questions-about-biometric-security-audio-hand-tracking-0194712/>

Strohm, O., & Ulich, E. (Hrsg.). (2019). *Unternehmen arbeitspsychologisch bewerten: Ein Mehr-Ebenen-Ansatz unter besonderer Berücksichtigung von Mensch, Technik und Organisation*. vdf Hochschulverlag AG an der ETH Zürich.

- Suh, A., & Prophet, J. (2018). The state of immersive technology research: A literature analysis. *Computers in Human Behavior*, 86, 77–90.
<https://doi.org/10.1016/j.chb.2018.04.019>
- Tomori, M., & Ogunseiju, O. (2025). Assessment of Situational Awareness in the Implementation of Sensing Technologies on a Virtual Construction Site within a Mixed Reality Environment. In A. GhaffarianHoseini, A. Ghaffarianhoseini, F. Rahimian & M. Babu Purushothaman (Hrsg.), *Proceedings of the International Conference on Smart and Sustainable Built Environment (SASBE 2024)* (S. 349–358). Springer Nature Singapore.
- Tractinsky, N. (2018). The Usability Construct: A Dead End? *Human–Computer Interaction*, 33(2), 131–177. <https://doi.org/10.1080/07370024.2017.1298038>
- Tseng, W.-J. (2023). Understanding Physical Breakdowns in Virtual Reality. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–5.
<https://doi.org/10.1145/3544549.3577064>
- Tufts, J. B., & Frank, T. (2003). Speech production in noise with and without hearing protection. *The Journal of the Acoustical Society of America*, 114(2), 1069–1080.
<https://doi.org/10.1121/1.1592165>
- Universität Erfurt. (o. D.-a). *Anger-Museum Erfurt*. Abgerufen 4. Januar 2025, von <https://www.uni-erfurt.de/philosophische-fakultaet/seminare-professuren/medien-und-kommunikationswissenschaft/professuren/forschung-und-publicationen/projekt-vr-in-thueringer-museen/anger-museum-erfurt>
- Universität Erfurt. (o. D.-b). *Projekt VR in Thüringer Museen*. Abgerufen 10. März 2024, von <https://www.uni-erfurt.de/philosophische-fakultaet/seminare->

professuren/medien-und-kommunikationswissenschaft/professuren/forschung-und-publikationen/projekt-vr-in-thueringer-museen

Vagnetti, R., Camp, N., Story, M., Ait-Belaid, K., Mitra, S., Zecca, M., Di Nuovo, A., & Magistro, D. (2024). Instruments for Measuring Psychological Dimensions in Human-Robot Interaction: Systematic Review of Psychometric Properties. *Journal of Medical Internet Research*, 26, e55597. <https://doi.org/10.2196/55597>

Vajpai, J., & Bora, A. (2016). Industrial applications of automatic speech recognition systems. *International Journal of Engineering Research and Applications*, 6(3), 88–95.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3), 425. <https://doi.org/10.2307/30036540>

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157–178. <https://doi.org/10.2307/41410412>

Vollrath, M., & Krems, J. F. (2011). *Verkehrspsychologie: Ein Lehrbuch für Psychologen, Ingenieure und Informatiker*. Kohlhammer Verlag. <https://books.google.com/books?hl=de&lr=&id=qnUiE-AAAQBAJ&oi=fnd&pg=PP1&dq=Vollrath,+M.,+%26+Krems,+&ots=mX5ehi2VMg&sig=kqL9i9Nw8RdTWxmGIv-FEyXvPNCc>

von Eitzen, I. M. (2024). *Faktoren zur Akzeptanz von Virtual Reality Anwendungen* [Dissertation, Universität Würzburg]. <https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/34632>

- Wandke, H. (2005). Assistance in human-machine interaction: a conceptual framework and a proposal for a taxonomy. *Theoretical Issues in Ergonomics Science*, 6(2), 129–155. <https://doi.org/10.1080/1463922042000295669>
- Wang, B., Li, X., Freiheit, T., & Epureanu, B. I. (2020). Learning and intelligence in human-cyber-physical systems: Framework and perspective. *2020 Second International Conference on Transdisciplinary AI (TransAI)*, 142–145. <https://ieeexplore.ieee.org/abstract/document/9253155/>
- Wang, K., Shu, Q., & Tu, Q. (2008). Technostress under different organizational environments: An empirical investigation. *Computers in Human Behavior*, 24(6), 3002–3013. <https://doi.org/10.1016/j.chb.2008.05.007>
- Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets*, 32(4), 2079–2102. <https://doi.org/10.1007/s12525-022-00593-5>
- Weidemann, A., & Rußwinkel, N. (2021). The Role of Frustration in Human-Robot Interaction – What Is Needed for a Successful Collaboration? *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.640186>
- Weinzierl, S. (2008). *Handbuch der Audiotechnik*. Springer-Verlag. <https://books.google.com/books?hl=de&lr=&id=OLgY0QpXD0YC&oi=fnd&pg=PA2&dq=Handbuch+der+Audiotechnik&ots=CDKb7e03nA&sig=dF0FOMpv9Ia1wBkgbw7NI0ov9n4>
- Windelband, L., & Spöttl, G. (2012). Diffusion von Technologien in die Facharbeit und deren Konsequenzen für die Qualifizierung am Beispiel des "Internet der Dinge". https://www.pedocs.de/frontdoor.php?source_opus=7118

- Wirth, T., & Klein, H.-W. (2023). Varianzanalyse versus Strukturgleichungsmodell – ein Vergleich aus der Praxis. In *Springer eBooks* (S. 119–140).
https://doi.org/10.1007/978-3-658-39625-1_5
- Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6), 767–791.
- Woodward, J., & Ruiz, J. (2023). Analytic Review of Using Augmented Reality for Situational Awareness. *IEEE Transactions on Visualization and Computer Graphics*, 29(4), 2166–2183. <https://doi.org/10.1109/TVCG.2022.3141585>
- Yang, J., Rahardja, S., & Fränti, P. (2019). Outlier detection: How to threshold outlier scores? *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, 1–6. <https://doi.org/10.1145/3371425.3371427>
- Zhang, J., Li, G., Huang, Q., Feng, Q., & Luo, H. (2022). Augmented Reality in K–12 Education: A Systematic Review and Meta-Analysis of the Literature from 2000 to 2020. *Sustainability*, 14(15), 9725. <https://doi.org/10.3390/su14159725>
- Zhao, Y., Hu, M., Hashash, S., & Azenkot, S. (2017). Understanding Low Vision People’s Visual Perception on Commercial Augmented Reality Glasses. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 4170–4181.
<https://doi.org/10.1145/3025453.3025949>
- Zhou, J., Zhou, Y., Wang, B., & Zang, J. (2019). Human–cyber–physical systems (HCPs) in the context of new-generation intelligent manufacturing. *Engineering*, 5(4), 624–636.

Zhou, Y., Chen, J., & Wang, M. (2022). A meta-analytic review on incorporating virtual and augmented reality in museum learning. *Educational Research Review*, 36, 100454. <https://doi.org/10.1016/j.edurev.2022.100454>

Zoetbrood, M. (2021). Do training and development practices help in overcoming the “IT productivity paradox”? ICT investments and Labor Productivity. <https://theses.ubn.ru.nl/items/55e3d066-7c8b-4a41-ba59-a5cfc1fef6d4>

Anhang A

A1 – Beitrag des Stands der Forschung zur Erklärung der UVs und AVs

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
Wandke (2005)	DQ; VQ; IQ	Leistung; Akzeptanz; Beanspruchung	Die Assistenz-Taxonomie verknüpft u. a. DQ, VQ und IQ mit den Phasen menschlichen Handelns. Defizite können Überlastung, Missverständnisse, Fehlentscheidungen und Fehlhandlungen verursachen und dadurch Leistung und Akzeptanz beeinträchtigen. Die Zusammenhänge sind konzeptionell abgeleitet, nicht experimentell geprüft.
Suh & Prophet (2018)	DQ; VQ; IQ	Leistung; Akzeptanz; Beanspruchung	U.a. DQ, IQ, und VQ werden als technologische Stimuli eingeordnet. Sie können Aufmerksamkeit, mentale Belastung, wahrgenommene Kontrolle und Zufriedenheit verändern und darüber Leistung und Nutzungsabsicht beeinflussen. Der Beitrag liefert einen theoretischen Rahmen, jedoch keinen experimentellen Fehlervergleich.
Moencks et al. (2022)	allgemeine Leistung Fehler		Technische Fehler werden als Auslöser von Unterbrechungen, Zeitverlusten, Qualitätsminderungen, Nacharbeit und Ausschuss beschrieben. Damit wird vor allem ein möglicher negativer Einfluss auf die Leistung beziehungsweise Prozessqualität aufgezeigt.
Konopka et al. (2024a)	DQ	Akzeptanz	Fehlerhafte Handlungsempfehlungen unterschiedlicher Schwere reduzierten die Datenqualität. Mit zunehmender Fehlerschwere sanken Vertrauen und teilweise Nutzungsabsicht; Vertrauensbruch und negative Emotionen verstärkten diesen Effekt. Die Wirkung auf Akzeptanz verlief damit insbesondere über Vertrauen und Fehlerwahrnehmung.
Tomori & Ogunseiju (2025)	DQ	Beanspruchung	Informationsqualität und Informationsmenge wurden korrelativ mit Dimensionen der kognitiven Belastung betrachtet. Die meisten Zusammenhänge waren nicht signifikant; lediglich verfügbare mentale Kapazität und physische Anforderung korrelierten signifikant. Die Quelle liefert daher nur explorative Hinweise zu DQ.
Brauner et al. (2019)	DQ	Leistung; Akzeptanz	DQ wurde durch ein korrektes beziehungsweise defektes Entscheidungsunterstützungssystem variiert. Das defekte System verringerte Leistung, Nützlichkeit, Benutzerfreundlichkeit und Nutzungsabsicht. Zudem verschlechterte die Befolgung fehlerhafter Empfehlungen die Leistung zusätzlich, wodurch geringe DQ objektive und subjektive Zielgrößen beeinflusste.

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
Madhavan et al. (2006)	DQ	Leistung; Akzeptanz	Fehler bei subjektiv einfachen Aufgaben senkten die geschätzte Zuverlässigkeit und führten zu überhöhtem Selbstvertrauen, häufigerem Widerspruch gegen korrekte Automation und geringerer Gesamtleistung. Damit zeigte sich ein Beitrag der DQ zur Leistung und zum Vertrauen.
Konopka et al. (2024b)	DQ; Konfidenzanzeige	Leistung; Akzeptanz	DQ wurde über die Systemperformance und die Konfidenzanzeige über unterschiedliche Transparenz- beziehungsweise Erklärungstiefen variiert. Höhere Performance und mehr Systeminformation stärkten das Vertrauen. Vertrauen wirkte nicht signifikant auf die Leistung, eine realistische Vertrauenskalibrierung erhöhte jedoch die zukünftige Nutzungsabsicht.
Wanner et al. (2022)	DQ; Konfidenzanzeige	Leistung; Akzeptanz	Systemtransparenz und bereitgestellte Systeminformationen können als Konfidenzanzeige verstanden werden. Sie erhöhten Leistungserwartung, Aufwandserwartung und Vertrauen. Leistungserwartung und Einstellung beeinflussten die Nutzungsabsicht direkt, während die Konfidenzanzeige überwiegend indirekt über Vertrauen, wahrgenommene Leistung und Einstellung wirkte.
Gegoff et al. (2023)	DQ; Konfidenzanzeige	Leistung; Akzeptanz; Beanspruchung	DQ wurde durch die Zuverlässigkeit der Empfehlungen und die Konfidenzanzeige durch unterschiedliche Transparenzstufen bestimmt. Hohe Zuverlässigkeit verbesserte Trefferquote, Vertrauen und Usability. Mehr Systeminformation verkürzte Entscheidungszeiten, senkte Workload und unterstützte eine angemessene Nutzung der Handlungsempfehlungen.
Lorrig & Daw (2024)	DQ; Konfidenzanzeige	Leistung; Akzeptanz	Wiederholte Fehlfunktionen oder unklare Empfehlungen können Vertrauen reduzieren oder Übervertrauen fördern und dadurch Situationswahrnehmung und manuelle Fähigkeiten beeinträchtigen. Die Zielgrößen wurden nicht empirisch erhoben.
Schaefer et al. (2016)	Konfidenzanzeige	Akzeptanz	Die Konfidenzanzeige umfasst Informationen über Systemleistung, Zuverlässigkeit und Status. Konsistente Systeminformationen unterstützen eine realistische Einschätzung der Systemfähigkeit und eine angemessene Vertrauenskalibrierung. Akzeptanz wurde in der Metaanalyse nicht als eigenständige Zielgröße geprüft, kann aber über Vertrauen und Nutzung mittelbar/indirekt beeinflusst sein.
Hecht et al. (2020)	Konfidenzanzeige	Akzeptanz; Beanspruchung	Die Konfidenzanzeige wurde durch ein prädiktives HMI umgesetzt, das Informationen über bevorstehende Systemgrenzen bereitstellte.

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
			Häufigere Übernahmeaufforderungen erhöhten Workload und reduzierten Akzeptanz und Usability. Das prädiktive HMI konnte diese negativen Effekte jedoch nicht signifikant kompensieren.
McGuirl & Sarter (2006)	Konfidenzanzeige	Leistung; Akzeptanz	Dynamische Konfidenzinformationen ermöglichen eine realistischere Einschätzung der Systemzuverlässigkeit. Piloten passten ihr Vertrauen besser an, korrigierten Fehlreaktionen häufiger und nutzten das System angemessener. Die Quelle belegt damit vor allem einen Beitrag zur Vertrauenskalibrierung und Leistung, nicht unmittelbar zur Technologieakzeptanz.
Estrella & Mendoza (2019)	IQ	/	Die Quelle beschreibt sprachbasierte Interaktion als technische Alternative zur Gestensteuerung und leistet damit einen konzeptionellen Beitrag zur IQ.
Fuhrmann et al. (2021)	IQ	/	Das multimodale Konzept kombiniert Gesten- und Sprachsteuerung zur Bedienung einer Spritzgussmaschine. Es zeigt Anwendungsmöglichkeiten unterschiedlicher IQ-Ausprägungen für Maschinensteuerung und Menünavigation.
Graichen & Graichen (2023)	IQ	Akzeptanz	Beobachtete Gestenfehler bildeten eine verminderte Interaktionsqualität ab. Mit zunehmender Fehlerzahl sank das Vertrauen signifikant, während kein signifikanter Zusammenhang mit der allgemeinen Akzeptanz bestand. IQ beeinflusst die Akzeptanz somit nicht unmittelbar, sondern zunächst vor allem das Vertrauen.
Derby et al. (2019)	IQ	Leistung	Die Studie liefert einen Leistungsbenchmark für gestenbasierte Texteingabe auf der HL1. Die geringe Eingabegeschwindigkeit zeigt, dass die gewählte Interaktionsmethode die Eingabeeffizienz begrenzt und alternative IQ-Ausprägungen wie Spracheingabe sinnvoll sein können.
Filippidou & Moussides (2020)	IQ	Leistung	Die WER dient als objektives Maß der Erkennungsleistung und damit als methodische Grundlage zur Bewertung der IQ. Eine geringere WER weist auf eine bessere Spracherkennung hin.
Képuska (2017)	IQ	Leistung	Der Vergleich mehrerer ASR-Systeme (Microsoft API, Google API And CMU Sphinx) anhand der WER liefert einen technischen Benchmark der IQ. Damit wird die Erkennungsleistung objektiv vergleichbar.
Liu et al. (2022)	IQ	Leistung	Die ASR der HL2 wurde unter abstrahierten Bedingungen vor allem hinsichtlich der Erkennungsrichtigkeit bewertet. Damit liefert die Quelle einen Beitrag der IQ zur Leistung.

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
			Informationsmenge pro Zeit, Akzeptanz und Beanspruchung wurden nicht umfassend betrachtet.
Cockburn et al. (2019)	IQ	Leistung; Beanspruchung	Vibration verschlechterte die IQ und erhöhte Fehlerraten sowie Auswahlzeiten. Die „Bracing“-Haltung stabilisierte die Eingabe, reduzierte Fehler und verkürzte die Bearbeitungszeit. Zugleich wurde sie als ergonomischer bewertet, sodass IQ sowohl die objektive Leistung als auch die körperliche Beanspruchung beeinflusste.
Pourmemar (2024)	IQ	Leistung; Beanspruchung	Die Interaktionsqualität wurde über Handgesten, Sprache und Head Pointing variiert. Handgesten führten zu höheren Auswahl- und Erkennungsfehlern, längeren Zeigezeiten und höherer körperlicher Belastung sowie Frustration. Sprache und insbesondere Head Pointing zeigten günstigere Leistungs- und Beanspruchungswerte.
(Caputo et al. (2024)	IQ	Leistung; Beanspruchung	Die drei Interaktionsmethoden führten zu deutlichen Unterschieden bei Bearbeitungszeit, Genauigkeit, Handbewegungen und Erlernbarkeit. Die Handlebar-Methode schnitt objektiv am günstigsten ab. Schwierigkeit, Ermüdung und SUS-Gesamtwerte unterschieden sich dagegen nicht durchgängig signifikant, sodass IQ vor allem die Leistung beeinflusste.
Marklin et al. (2022)	IQ	Leistung; Beanspruchung	Unter realitätsnahen industriellen Lärmbedingungen war die sprachbasierte IQ auf der HL2 stark eingeschränkt. Bei 81 bis 84 dB wurden 90 % der Befehle nicht erfolgreich übertragen. Umgebungslärm beeinträchtigte damit die Leistung deutlich; zusätzlich wurden körperliche Beanspruchungsaspekte der HMD-Nutzung erfasst.
Derby et al. (2020)	IQ	Leistung; Beanspruchung	Die Spracheingabe auf der HL1 wurde bei 40, 55 und 70 dB untersucht. Eingabegeschwindigkeit und wahrgenommene Arbeitsbelastung unterschieden sich nicht signifikant; bei 70 dB verschlechterten sich jedoch Erkennungsgenauigkeit und Usability. IQ wirkte somit vor allem auf Leistung und Systembenutzbarkeit.
Sinlapanuntakul et al. (2021)	IQ	Leistung; Beanspruchung	Die Spracheingabe auf der HL2 wurde bei verschiedenen Lärmbedingungen geprüft. Die WER unterschied sich nicht signifikant, wohl aber Eingabegeschwindigkeit sowie mentale und zeitliche Beanspruchung. Die Systembenutzbarkeit blieb unverändert, sodass IQ selektiv auf einzelne Leistungs- und Belastungsmaße wirkte.

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
Nasirian et al. (2017)	IQ; DQ	Akzeptanz	IQ, DQ und Systemqualität wurden als wahrgenommene Qualitätsfaktoren modelliert. Nur die IQ erhöhte Vertrauen signifikant; Vertrauen steigerte wiederum die Nutzungsabsicht. DQ und Systemqualität zeigten keine signifikanten Effekte, woraus jedoch keine generelle Wirkungslosigkeit konkreter Datenfehler folgt.
Mariani et al. (2013)	TQ	Akzeptanz	Trainingsmöglichkeiten stehen signifikant mit wahrgenommener Benutzerfreundlichkeit und Nützlichkeit in Zusammenhang, die wiederum die Nutzungsabsicht beeinflussen. TQ trägt damit indirekt/mittelbar (ohne Nachweis von Kausalität) zur Akzeptanz bei.
Alqahtani et al. (2022)	TQ	Akzeptanz	Ausbildungsqualität erhöhte wahrgenommene Benutzerfreundlichkeit und wahrgenommenen Nutzen signifikant. Ein direkter Effekt auf die Nutzungsabsicht war nicht signifikant. TQ beeinflusste somit vorgelagerte Akzeptanzfaktoren; ein indirekter Effekt wurde jedoch nicht statistisch nachgewiesen.
Bala & Venkatesh (2016)	TQ	Akzeptanz; Beanspruchung (vgl. Bedrohung)	TQ wurde über die wahrgenommene Effektivität der Schulung operationalisiert. Effektive Schulung reduzierte die wahrgenommene Bedrohung durch neue IT-Systeme, insbesondere bei hoher Nutzerbeteiligung. Damit kann TQ Akzeptanzbarrieren senken.
Korunka & Vitouch (1999)	TQ	Beanspruchung	Schulung wurde als Bestandteil umfassender Anpassungsanforderungen betrachtet. Umfangreiche organisatorische Änderungen waren unter bestimmten Kontextbedingungen mit geringerer Belastung und höherer Arbeitszufriedenheit verbunden, insbesondere bei Partizipation.
Day et al. (2019)	TQ	Beanspruchung	Training und Beteiligung an der IKT-Einführung wurden als organisationale Ressourcen betrachtet. Angemessen geschulte und einbezogene Beschäftigte berichteten weniger Stress. Der Befund stützt einen Beitrag von TQ zur Beanspruchung.
Wang et al. (2008)	TQ	Beanspruchung	Ausreichende Schulungen und eine lernende Organisation können Technostress reduzieren. TQ wird damit als organisationale Ressource zur Verringerung der Beanspruchung eingeordnet.
Schweikl & Obermaier (2020)	TQ	Leistung	Das Metareview ordnet TQ als organisatorische Voraussetzung für IKT-Produktivität ein. Fehlende Kompetenzen, geringe Trainingsressourcen und unzureichende Einführung können verhindern, dass Produktivitätspotenziale realisiert werden.

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
Zoetbrood (2021)	TQ	Leistung	TQ wurde über organisationale Schulungs- und Entwicklungspraktiken abgebildet. Zwischen diesen Praktiken und Produktivität bestand nur ein sehr kleiner, nicht signifikanter Zusammenhang. Auch Haupteffekt und moderierende Wirkung auf IKT-bedingte Produktivität waren nicht signifikant.
Marshall et al. (2008)	TQ	Leistung; Akzeptanz; Beanspruchung	Endanwenderschulung korrelierte stark positiv mit Leistungserwartung und Aufwandserwartung. Höhere TQ war damit mit höherer wahrgenommener Leistungsfähigkeit und einem geringeren Nutzungsaufwand verbunden
Minow (2021)	TQ	Leistung; Beanspruchung	Das Trainingsformat wurde als Desktoptraining gegenüber einem Trainingsfilm variiert. Desktoptraining verbesserte Verständlichkeit, Lernförderlichkeit und einzelne objektive Leistungswerte. SUS-Gesamtwert, visuelle Beanspruchung und geistige Anstrengung unterschieden sich dagegen nicht signifikant.
Sadovitch (2020)	VQ	Akzeptanz	Die Sichtbarkeitsqualität wurde durch Art und Ausprägung visueller Registrierungsfehler variiert. Höhere Fehlergrade verschlechterten Nützlichkeit, Zufriedenheit sowie pragmatische und hedonische Bewertungen, wobei die Wirkung vom Fehlertyp abhing.
Merenda et al. (2019)	VQ	Leistung	VQ wurde über die grafische Detailtreue variiert. Höhere Fidelity verlängerte Bremszeiten, erhöhte Blickzeiten auf Fußgänger und reduzierte den Blickanteil auf die Straße. Damit verschlechterte sich die objektive Leistung.
Arning et al. (2012)	VQ	Leistung; Akzeptanz; Beanspruchung	Beim mobilen Projektor waren Navigationsleistung und Akzeptanz geringer und die Arbeitsbelastung höher als beim Bildschirm. Sichtbarkeitsprobleme reduzierten das Vertrauen, welches wiederum die Akzeptanz beeinflusste. VQ wirkte damit sowohl auf Leistung und Beanspruchung als auch indirekt über Vertrauen auf Akzeptanz.
Richardson et al. (2018)	VQ; Konfidenzanzeige	Akzeptanz; Beanspruchung	Die Sichtbarkeitsqualität beziehungsweise Konfidenzanzeige wurde durch zusätzliche Systeminformationen im HMI erweitert. Das größere Informationsangebot erhöhte Zufriedenheit und subjektive Akzeptanz und reduzierte die wahrgenommene Arbeitsbelastung, insbesondere situativen Stress. Eine direkte Fehlermanipulation erfolgte nicht.
Alessa et al. (2023)	VQ; TQ	Leistung; Beanspruchung	AR- und papierbasierte Instruktion unterschieden sich hinsichtlich Darstellungs- und Trainingsform. Bei komplexen Aufgaben verkürzte

Referenz	zugeordnete UVs	zugeordnete Primär-AVs	Erkenntnisse/Beitrag
			AR die Bearbeitungszeit und reduzierte ausgewählte Belastungsmaße; physiologische Werte blieben vergleichsweise stabiler. Der Beitrag zu VQ und TQ ist indirekt, da kein gezielter Sichtbarkeits- oder Trainingsfehler manipuliert wurde.

Fußnote. Die zentralen Befunde zur Synthese werden jeweils funktional dem nächstgelegenen PF bzw. Der nächstgelegenen UV des übergeordneten Forschungsdesign zugeordnet.

Tabelle 55. Verknüpfung Vorstudien mit Variablen des Forschungsdesign

A2 – Nutzerbefragungen & Einverständniserklärung

Variable	Item/Antwort
UTAUT2_PE1	Ich finde * nützlich im täglichen Leben.
UTAUT2_PE3	* hilft mir dabei Museumsinhalte besser zu verstehen.
UTAUT2_PE4	* steigert mein Interesse.
UTAUT2_EE1	Den Umgang mit * zu erlernen, fällt mir leicht.
UTAUT2_EE2	Die Interaktion in * ist klar und verständlich.
UTAUT2_EE3	* ist einfach zu nutzen.
UTAUT2_EE4	Ich finde, dass es einfach ist, geübt im Umgang mit * zu werden.
UTAUT2_SI1	Andere Personen, die mir nahestehen, sind der Meinung, dass ich * nutzen sollte.
UTAUT2_SI2	Andere Personen, die mein Verhalten beeinflussen, sind der Meinung, dass ich * nutzen sollte.
UTAUT2_SI3	Personen, deren Meinung ich schätze, würden es begrüßen, wenn ich * nutze.
UTAUT2_FC2	Ich verfüge über das notwendige Wissen, um * zu nutzen.
UTAUT2_FC4	Ich kann Hilfe von anderen bekommen, falls ich Probleme mit der Nutzung (von) * habe.
UTAUT2_HM1	* macht Spaß.
UTAUT2_HM2	* ist amüsant.
UTAUT2_HM3	* ist sehr unterhaltsam.
UTAUT2_HT1	* zu nutzen, ist für mich zu einer Gewohnheit geworden.
UTAUT2_HT2	Ich bin süchtig nach *.
UTAUT2_HT3	Ich muss * nutzen.
UTAUT2_BI1	Ich beabsichtige * auch in der Zukunft zu nutzen.
UTAUT2_BI2	Ich werde immer versuchen * in meinem täglichen Leben zu nutzen.
UTAUT2_BI3	Ich plane * weiterhin häufig zu nutzen.
Rückgabewerte	1 = stimme überhaupt nicht zu 2 = stimme nicht zu 3 = stimme eher nicht zu 4 = weder noch 5 = stimme eher zu 6 = stimme zu 7 = stimme völlig zu -9 = nicht beantwortet

Fußnote. * wurde je Studie mit dem entsprechenden Artikel wie folgt angepasst: Studie 1 = VR-Museumsanwendung; Studie 2 = AR-Sprachinteraktion; Studie 3 = Augmented Reality; Studie 4 = Augmented-Reality-Messanwendung

Tabelle 56. deutsche Fassung des UTAUT2 angepasst je Studie

Variable	Item/Antwort
NASA-TLX_MD	Wie viel geistige Anstrengung war bei der Informationsaufnahme und -verarbeitung erforderlich? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erforderte sie hohe Genauigkeit oder war sie fehlertolerant?
NASA-TLX_PD	Wie viel körperliche Aktivität war erforderlich? War die Aufgabe leicht oder schwer, einfach oder anstrengend, erhol-sam oder mühselig?
NASA-TLX_TD	Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder des Taktes, mit dem Aufgaben oder Aufgabenelemente auftraten? War die Abfolge langsam und geruhsam oder schnell und hektisch?
NASA-TLX_P	Wie erfolgreich haben Sie Ihrer Meinung nach die gestellte Aufgabe be-wältigt? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?
NASA-TLX_E	Wie hart mussten Sie arbeiten, um Ihre Aufgabenerfüllung zu errei-chen?
NASA-TLX_F	Wie unsicher, entmutigt, irritiert, gestresst und verärgert fühlten Sie sich während der Aufgabe?
Rückgabewerte	1 = gering bis 21 = hoch -9 = nicht beantwortet

Tabelle 57. deutsche Fassung des NASA-TLX angepasst je Studie

Variable	ITEM/Antwort
Geschlecht	Welches Geschlecht haben Sie?
Rückgabewerte	1 = weiblich 2 = männlich 3 = divers -9 = nicht beantwortet
Alter	Wie alt sind Sie?
Rückgabewerte	Ganze Zahl
Beschäftigung	Was machen Sie beruflich?
Rückgabewerte	1 = Schüler/in 2 = In Ausbildung 3 = Student/in 4 = Angestellte/r 5 = Beamte/r 6 = Selbstständig 7 = Arbeitslos/Arbeit suchend 8 = Sonstiges: -9 = nicht beantwortet
Formale Bildung	Welchen Bildungsabschluss haben Sie
Rückgabewerte	11 = Volks- oder Hauptschulabschluss, Quali 4 = Mittlere Reife, Realschul- oder gleichwertiger Abschluss 5 = Abgeschlossene Lehre 6 = Fachabitur, Fachhochschulreife 7 = Abitur, Hochschulreife 8 = Fachhochschul-/Hochschulabschluss 10 = Anderer Abschluss, und zwar: -9 = nicht beantwortet

Tabelle 58. Soziodemografischer Fragebogen

Variable	ITEM/Antwort
Sehhilfe	Benötigen Sie eine Sehhilfe?
Rückgabewerte	1 = Brille 2 = Kontaktlinsen 3 = keine -9 = nicht beantwortet
Sehschwäche	Haben Sie eine Sehschwäche
Rückgabewerte	1 = ja 2 = ja, wird aber durch meine Sehhilfe vollständig korrigiert 3 = nein -9 = nicht beantwortet
Abfrage 1	Wie heißt der Mädchenname von Marianne Brandt? Nach dem Eingang auf der linken Seite die mittlere Dame?
Rückgabewerte	1 = Liebe* 2 = Wagner 3 = Rosenberg -9 = nicht beantwortet
Abfrage 2	Wie heißt die Ausstellung
Rückgabewerte	1 = Bauhausmädels* 2 = Bauhaus Mädchen 3 = Bauhaus Frauen

Variable	ITEM/Antwort
	4 = Mädels des Bauhaus -9 = nicht beantwortet
Abfrage 3	Aus welchem Jahrzehnt stammt die Verkörperung der „Neuen Frau“ als Vorbildcharakter ihrer Generation?
Rückgabewerte	1 = 1910er 2 = 1920er* 3 = 1930er -9 = nicht beantwortet
Abfrage 4	Wie heißt das Untermotto der Ausstellung „Mädchen wollen etwas lernen“?
Rückgabewerte	1 = Frauen wollen etwas gestalten* 2 = Frauen wollen etwas arbeiten 3 = Frauen wollen etwas erleben -9 = nicht beantwortet
Abfrage 5	Welche Figur steht im Wandregal gegenüber der Tür?
Rückgabewerte	1 = Fisch 2 = Pferd* 3 = Katze 4 = Rind -9 = nicht beantwortet
Abfrage 6	Wie hieß die Sportlehrerin, deren Portrait als Ikone/Prototyp des neuen Frauentypus diente?
Rückgabewerte	1 = Petra Groß 2 = Karla Müller 3 = Karla Grosch* -9 = nicht beantwortet
Abfrage 7	In welcher Zeitschrift wurde der Artikel „Mädchen wollen etwas lernen“ veröffentlicht?
Rückgabewerte	1 = Die Zeit 2 = Der Spiegel 3 = Die Woche* 4 = Der Markt -9 = nicht beantwortet

Fußnote. * markiert die richtige Antwort.

Tabelle 59. Spezifische Erhebungen – S1

Variable	ITEM/Antwort
Erfahrung	Haben Sie Erfahrung mit Spracheingaben?
Rückgabewerte	1 = gering bis 7 = hoch -9 = nicht beantwortet
Muttersprache	Ist Deutsch Ihre Muttersprache?
Rückgabewerte	1 = ja 2 = nein -9 = nicht beantwortet
Artikulation	Waren Sie jemals in logopädischer Behandlung?
Rückgabewerte	1 = ja 2 = nein -9 = nicht beantwortet

Tabelle 60. Spezifische Erhebungen – S2

Variable	ITEM/Antwort
Erfahrung HL2	Wie viele Stunden haben Sie die HoloLens2 bereits genutzt?
Rückgabewerte	Ganze Zahl
Erfahrung MR	Haben Sie bereits Mixed-Reality-Erfahrung mit Augmented oder Virtual Reality gemacht?
	1 = gering
	bis
Rückgabewerte	7 = hoch
	-9 = nicht beantwortet

Tabelle 61. Spezifische Erhebungen – S3

Variable	ITEM/Antwort
Sehhilfe	Benötigen Sie eine Sehhilfe?
	1 = Brille
	2 = Kontaktlinsen
Rückgabewerte	3 = keine
	-9 = nicht beantwortet
Sehschwäche	Haben Sie eine Sehschwäche
	1 = ja
	2 = ja, wird aber durch meine Sehhilfe vollständig korrigiert
Rückgabewerte	3 = nein
	-9 = nicht beantwortet
Q-Unterschied	Ist Ihnen ein (Qualitäts-)Unterschied der AR-gestützten Anwendung zwischen den beiden Durchläufen aufgefallen?
	1 = ja
Rückgabewerte	2 = nein
	-9 = nicht beantwortet
Fehlereigenschaft	Mir ist aufgefallen, dass in einem der beiden Durchläufe folgende Anwendungseigenschaft fehlerhaft ist:
	Mehrfachauswahl
	1 = Qualität der Darstellung (Sichtbarkeit/Helligkeit der augmentierten Inhalte)
Rückgabewerte	2 = Interaktionsvermögen/-eigenschaften (Fehler beim Handling)
	3 = Maße/Dimensionen (Fehler in der Skalierung)
	-9 = nicht beantwortet

Tabelle 62. Spezifische Erhebungen – S4



0% ausgefüllt

Willkommen

Vielen Dank, dass Sie sich bereit erklärt haben an dieser Studie teilnehmen. Bitte lesen Sie sich folgende Hinweise sorgfältig durch und wenden Sie sich bei eventuell auftretenden Fragen an den Versuchsleiter bzw. die Versuchsleiterin.

Hinweise zur Teilnahme

- Diese Befragung ist **kein Test**. Es gibt weder richtige, noch falsche Antworten. Es findet ebenfalls keinerlei Bewertung Ihrer Antworten statt. Sie können die Befragung also weder bestehen, noch durchfallen.
- Bitte antworten Sie spontan, wahrheitsgemäß und Ihrer persönlichen Meinung entsprechend.
- Die Teilnahme ist grundsätzlich **freiwillig**. Sie können dementsprechend jederzeit die Teilnahme ohne Konsequenzen abbrechen, indem Sie die Versuchsleitung darüber in Kenntnis setzen.

Hinweise zum Datenschutz

- Ihre Daten werden vollständig anonymisiert erhoben und vertraulich behandelt. Ein **Rückschluss auf Ihre Person** ist im Nachhinein **nicht möglich**, d.h. es können auch keine Daten rückwirkend gelöscht werden.
- Die erhobenen Daten werden in ihrer **angesammelten, anonymisierten Form** nach wissenschaftlichen Standards ausgewertet und die Auswertungen in Abschlussarbeiten, Berichten und/oder Fachzeitschriften veröffentlicht bzw. **verwendet**.
- Es erfolgt **keine Weitergabe der Daten an Dritte**.
- Die **Speicherung der Daten** erfolgt gemäß **rechtlicher Regelungen**, z. B. in Hinsicht auf Aufbewahrungsfristen.

Die Richtlinien guter ethischer Forschung sehen vor, dass sich die Teilnehmer/-innen an empirischen Studien explizit und nachvollziehbar mit der Teilnahme einverstanden erklären.

Für Rückfragen finden Sie im Anschluss oder im Impressum Kontaktdaten des Versuchsleiters bzw. der Versuchsleiterin.

Hiermit bestätige ich, dass ich mindestens 18 Jahre alt bin sowie die Einverständniserklärung gelesen und verstanden habe.

- ☐ Nein (nicht an der Studie teilnehmen)
- ☐ Ja

1. Hier bitte die Versuchsnummer eingeben:

Abbildung 78. Einverständniserklärung zur Erhebung

A3 – Studie 1 Sichtbarkeitsfehler I

Proband Nr.	DLZ [s]	Richtigkeit	Leistung [%/s]	VQ
202		324	0,57	0,18fehlerbehaftet
203		460	0,71	0,16fehlerbehaftet
204		441	0,86	0,19fehlerfrei
205		633	0,57	0,09fehlerfrei
206		490	0,43	0,09fehlerbehaftet
207		356	0,71	0,20fehlerfrei
208		514	0,29	0,06fehlerbehaftet
209		470	0,71	0,15fehlerfrei
210		520	0,43	0,08fehlerbehaftet

211	574	1,00	0,17fehlerfrei
212	612	0,43	0,07fehlerbehaftet
213	506	0,71	0,14fehlerfrei
214	542	0,57	0,11fehlerbehaftet
215	588	0,86	0,15fehlerfrei
216	409	0,57	0,14fehlerbehaftet
217	240	1,00	0,42fehlerfrei
218	432	0,71	0,17fehlerbehaftet
220	299	0,57	0,19fehlerbehaftet

Tabelle 63. Datensatz DLZ, Richtigkeit und Leistung versus VQ – S1

Proband Nr.	VQ	MD	TD	PD	P	F	Mean_NASA
202	fehlerbehaftet	25	20	5	55	35	26,67
203	fehlerbehaftet	70	15	0	65	60	47,50
204	fehlerfrei	50	25	10	50	20	34,17
205	fehlerfrei	80	10	20	60	75	41,67
206	fehlerbehaftet	80	10	40	95	20	45,83
207	fehlerfrei	85	0	20	75	75	50,00
208	fehlerbehaftet	75	5	0	85	70	41,67
209	fehlerfrei	35	0	10	80	65	31,67
210	fehlerbehaftet	100	55	0	70	50	55,83
211	fehlerfrei	30	0	5	60	55	25,00
212	fehlerbehaftet	25	20	5	55	15	23,33
213	fehlerfrei	80	20	0	35	65	35,83
214	fehlerbehaftet	55	20	60	75	40	49,17
215	fehlerfrei	45	0	0	70	10	22,50
216	fehlerbehaftet	70	5	0	70	65	40,00
217	fehlerfrei	60	5	65	80	75	55,00
218	fehlerbehaftet	45	0	0	90	50	34,17
220	fehlerbehaftet	0	20	0	85	0	26,67

Fußnote. transformiert -1*5 & P invertiert

Tabelle 64. Datensatz NASA-TLX versus VQ – S1

Proband Nr.	VQ	PE	EE	SI	FC	HM	HT	BI
202	fehlerbehaftet	3,67	5,75	4,00	6,50	6,00	2,67	4,67
203	fehlerbehaftet	5,33	6,25	4,00	7,00	6,00	3,33	5,33
204	fehlerfrei	3,33	5,75	1,00	6,50	6,00	1,67	3,33
205	fehlerfrei	6,67	7,00	2,33	4,00	7,00	2,67	4,33
206	fehlerbehaftet	4,00	5,00	2,00	4,50	6,33	1,33	4,00
207	fehlerfrei	6,67	7,00	4,67	6,50	7,00	2,33	5,67
208	fehlerbehaftet	5,33	7,00	1,00	5,50	7,00	1,00	2,33
209	fehlerfrei	4,00	5,50	2,00	5,00	7,00	1,00	3,00
210	fehlerbehaftet	4,67	6,00	4,33	5,50	6,00	1,00	2,00
211	fehlerfrei	4,67	7,00	4,00	3,50	6,67	1,00	3,67
212	fehlerbehaftet	4,67	6,25	4,00	5,00	3,67	1,00	3,67
213	fehlerfrei	6,00	6,25	3,67	5,50	7,00	2,67	6,00
214	fehlerbehaftet	6,00	6,00	3,00	5,00	6,00	2,33	5,00
215	fehlerfrei	4,33	5,75	2,00	6,00	6,00	3,33	4,00
216	fehlerbehaftet	6,33	7,00	3,67	5,50	7,00	1,00	5,33
217	fehlerfrei	4,67	5,75	5,00	6,50	6,00	4,00	5,00
218	fehlerbehaftet	6,00	5,50	2,67	6,50	7,00	1,00	5,33

Proband Nr.	VQ	PE	EE	SI	FC	HM	HT	BI
220	fehlerbehaftet	7,00	7,00	4,00	7,00	7,00	4,33	5,33

Tabelle 65. Datensatz UTAUT2 versus VQ – S1

Proband Nr.VQ		SCL_RP	SCL_AP	HR_RP	HR_AP
202	fehlerbehaftet	3,01	4,49	97,98	110,31
203	fehlerbehaftet	3,87	5,24	62,25	80,96
204	fehlerfrei	1,71	4,82	81,98	92,36
205	fehlerfrei	9,44	7,64	73,20	86,02
206	fehlerbehaftet	1,95	2,31	125,44	145,27
207	fehlerfrei	4,36	3,99	67,21	84,99
208	fehlerbehaftet	4,54	4,79	52,32	55,54
209	fehlerfrei	0,46	1,08	77,79	77,49
210	fehlerbehaftet	2,62	4,19	83,68	105,36
211	fehlerfrei	5,84	18,76	67,91	68,42
212	fehlerbehaftet	3,97	7,88	90,03	100,42
213	fehlerfrei	1,09	1,28	109,04	113,25
214	fehlerbehaftet	5,45	8,11	78,82	96,74
215	fehlerfrei	2,00	4,35	72,46	90,13
216	fehlerbehaftet	4,65	4,42	73,62	87,63
217	fehlerfrei	0,98	0,79	76,04	84,80
218	fehlerbehaftet	9,54	16,74	59,02	61,33
220	fehlerbehaftet	1,36	1,57	63,99	69,64

Tabelle 66. Datensatz SCL und HR versus VQ – S1

A4 – Studie 2 Interaktionsfehler I

Proband Nr.	WER 0 dB	WER 64,2 dB	WER 84,9 dB	WER 87 dB	ITR 0 dB	ITR 64,2 dB	ITR 84,9 dB	ITR 87 dB
<i>Durchlauf 1</i>								
1	0,00	7,69	53,85	100,00	48,95	17,18	1,76	0,00
2	2,56	7,69	25,64	17,95	20,73	12,49	18,62	9,52
3	7,69	15,38	58,97	33,33	13,59	0,00	16,19	0,00
4	0,00	15,38	15,38	23,08	41,71	18,98	12,18	51,98
5	0,00	2,56	23,08	15,38	1,62	1,25	1,49	49,26
6	0,00	2,56	20,51	17,95	47,07	46,47	38,85	38,91
7	7,69	2,56	10,26	15,38	49,65	40,59	46,19	32,90
8	0,00	2,56	10,26	7,69	59,59	51,37	14,51	32,14
9	2,56	10,26	35,90	56,41	37,19	35,85	41,46	22,89
10	10,26	10,26	17,95	48,72	26,86	37,92	17,94	39,35
11	10,26	7,69	15,38	28,21	47,48	41,79	36,53	15,38
12	0,00	5,13	10,26	7,69	49,76	43,54	33,01	29,55
13	0,00	100,00	56,41	100,00	35,41	42,92	42,17	37,65
14	2,56	28,21	66,67	87,18	47,71	1,12	1,02	36,86
15	2,56	0,00	2,56	33,33	34,48	40,43	7,47	40,26
16	0,00	12,82	38,46	23,08	29,39	32,60	27,83	14,13
17	0,00	2,56	5,13	0,00	48,04	44,85	37,65	29,23
18	10,26	0,00	15,38	20,51	36,86	35,72	36,66	37,39
19	0,00	25,64	2,56	0,00	44,57	46,13	45,14	40,13
20	7,69	33,33	100,00	100,00	40,13	43,81	43,11	49,44
21	0,00	20,51	30,77	33,33	53,05	47,39	43,19	18,75

Proband Nr.	WER 0 dB	WER 64,2 dB	WER 84,9 dB	WER 87 dB	ITR 0 dB	ITR 64,2 dB	ITR 84,9 dB	ITR 87 dB
22	2,56	2,56	15,38	46,15	34,20	39,35	35,58	35,18
<i>Durchlauf 2</i>								
1		12,82	56,41	35,90		0,00	0,00	0,00
2		76,92	33,33	41,03		26,79	8,04	0,00
3		20,51	77,85	66,67		45,23	9,30	0,00
4		12,82	25,64	5,13		6,40	7,03	26,28
5		10,26	30,77	53,85		34,20	36,26	9,34
6		10,26	17,95	28,21		43,19	6,29	14,84
7		5,13	25,64	25,64		46,75	8,15	36,36
8		0,00	7,69	10,26		59,84	39,44	36,86
9		15,38	43,59	38,46		46,35	1,88	50,73
10		77,85	30,77	41,03		44,94	18,98	16,67
11		10,26	7,69	15,38		43,00	13,35	24,00
12		0,00	5,13	23,08		43,08	15,36	14,01
13		0,00	5,13	77,85		37,31	40,90	43,40
14		10,26	46,15	77,85		41,96	0,00	13,11
15		0,00	5,13	25,64		40,57	32,97	18,15
16		0,00	25,64	41,03		37,33	12,57	5,64
17		2,56	10,26	2,56		41,71	36,66	38,06
18		2,56	10,26	33,33		38,83	28,47	36,04
19		0,00	2,56	2,56		49,44	43,11	43,81
20		56,41	77,85	77,85		45,14	46,13	44,57
21		12,82	23,08	15,38		44,77	24,53	37,39
22		2,56	15,38	17,95		36,99	27,35	4,87

Tabelle 67. Datensatz WER und ITR versus SPL – S2

Proband Nr.	ReihenfolgeMD	TD	PD	P	F	
1	SER	10	5	5	95	5
2	EER	35	20	25	25	55
3	SER	0	0	0	40	55
4	EER	100	100	15	35	100
5	SER	70	40	10	100	15
6	EER	25	25	20	70	75
7	SER	25	10	5	70	25
8	SER	25	25	15	90	80
9	EER	40	30	15	65	40
10	EER	5	5	15	100	0
11	SER	0	0	0	85	25
12	SER	20	35	15	65	5
13	EER	90	5	15	20	85
14	SER	5	20	20	80	10
15	EER	40	5	5	95	0
16	EER	0	0	0	50	0
17	SER	15	5	0	85	5
18	EER	25	20	0	85	50
19	SER	0	0	0	100	0
20	EER	5	45	10	40	55
21	SER	0	10	5	100	20
22	EER	5	0	0	100	10

Proband Nr.	Reihenfolge	MD	TD	PD	P	F
-------------	-------------	----	----	----	---	---

Fußnote. transformiert -1*5 & P invertiert

Tabelle 68. Datensatz NASA-TLX versus RF – S2

Proband Nr.	Reihenfolge	PE	EE	SI	FC	HM	HT	BI
1	SER		5,00	6,75	4,33	4	7	2,33
2	EER		1,00	1,25	1	1	2	1
3	SER		5,33	5	5	5	5	2,67
4	EER		3,67	4,75	4,33	6,00	5,00	4,67
5	SER		6,00	6,50	6,00	6,50	6,00	4,67
6	EER		4,00	5,00	3,33	1,50	4,33	1,00
7	SER		5,33	5,25	6,00	7,00	6,67	4,33
8	SER		5,67	5,50	6,00	6,00	6,00	5,33
9	EER		4,67	5,00	4,00	5,00	4,33	2,33
10	EER		4,67	5,50	2,00	6,00	5,33	1,67
11	SER		4,33	4,75	1,33	4,50	7,00	1,00
12	SER		3,00	5,75	3,33	4,50	7,00	1,33
13	EER		1,00	2,00	1,00	6,00	1,00	1,00
14	SER		2,67	7,00	5,00	5,50	5,00	3,00
15	EER		4,67	4,75	1,00	2,00	5,00	1,00
16	EER		3,00	5,00	6,00	5,00	6,00	1,00
17	SER		4,00	5,50	4,00	5,50	7,00	1,67
18	EER		4,33	5,50	4,00	5,50	4,33	1,33
19	SER		7,00	7,00	7,00	7,00	7,00	7,00
20	EER		2,00	10,00	3,00	13,00	3,67	1,00
21	SER		1,00	3,00	2,00	0,00	7,00	2,00
22	EER		2,00	1,00	1,00	0,00	6,67	3,33

Tabelle 69. Datensatz UTAUT2 versus RF – S2

A5 – Studie 3 Trainingsfehler

Proband Nr.	UV TQ	Vorerfahrung XR [h]	DLZ [s]	AV ZEF [%]	AV DLZS [s]	AV Leistung [%/min]
8	MT		1	218	60	1488
9	DT		1	752	86,7	1249
10	MT		0	548	80	1560
11	DT		0	102	33,3	1755
12	MT		0	633	80	1645
13	DT		3	550	80	1562
14	MT		3	362	80	1374
15	DT		2	1286	93,3	1406
16	MT		2	1413	100	1413
17	DT		4	1221	100	1221
18	MT		1	358	80	1370
19	DT		2	755	80	1767
20	MT		0	235	60	1505
21	DT		0	602	60	1872
22	MT		1	142	33,3	1795
23	DT		0	394	80	1406
24	MT		6	45	13,3	1828
25	DT		0	503	60	1773

Proband Nr.	UV TQ	Vorerfahrung XR [h]	DLZ [s]	AV ZEF [%]	AV DLZS [s]	AV Leistung [%/min]
26	MT		0	242	60	1,43
27	DT		1	1218	100	2,96
28	MT		0	267	60	1,41
29	DT		0	1208	93,3	2,53
30	MT		0	413	80	2,02
31	DT		1	1498	100	2,40
32	MT		0	421	60	1,28
33	DT		1	986	100	3,65
34	MT		3	211	60	1,46
35	DT		100	1201	100	3,00
36	MT		0	310	60	1,37
37	DT		3	648	80	1,73
38	MT		0	1247	93,3	2,46
39	DT		1	1189	100	3,03
40	MT		0	526	80	1,87
41	DT		0	622	80	1,76
48	OT		0	152	33,3	0,73
49	OT		0	884	100	4,07
50	OT		0	427	100	8,43
51	OT		0	881	100	4,09
52	OT		0	577	93,3	5,09
53	OT		0	257	80	2,38
54	OT		0	269	80	2,36
55	OT		0	1126	100	3,20
56	OT		0	1105	100	3,26
57	OT		0	457	100	7,88
58	OT		0	650	93,3	4,58
59	OT		0	763	100	4,72
60	OT		0	765	100	4,71
61	OT		0	79	33,3	0,77
62	OT		0	335	80	2,24
63	OT		0	566	100	6,36
64	OT		0	751	100	4,79
65	OT		0	918	100	3,92

Tabelle 70. Datensatz ZEF, DLZ, DLZS versus TQ – S3

Proband Nr.	UV TQ	MD	TD	PD	P	F	Mean_NASA	
8	MT		60	50	40	55	60	54
9	DT		50	60	60	45	20	39
10	MT		70	15	10	70	45	45
11	DT		50	50	40	30	30	40
12	MT		65	55	5	80	85	51
13	DT		90	10	25	95	90	76
14	MT		65	10	5	95	65	54
15	DT		40	40	5	85	50	41
16	MT		30	5	0	50	15	23
17	DT		40	10	45	60	10	41
18	MT		55	15	15	70	45	47
19	DT		40	20	10	65	50	39

Proband Nr.	UV TQ	MD	TD	PD	P	F	Mean_NASA	
20	MT		40	0	15	90	15	36
21	DT		70	50	50	70	60	64
22	MT		50	15	25	30	30	35
23	DT		60	30	10	40	95	49
24	MT		80	30	20	60	65	57
25	DT		0	25	0	80	0	21
26	MT		65	55	20	40	45	44
27	DT		75	20	20	10	15	30
28	MT		75	10	5	5	75	46
29	DT		60	45	10	80	80	56
30	MT		75	90	20	70	70	60
31	DT		70	30	0	65	45	49
32	MT		80	65	15	90	40	55
33	DT		50	25	45	15	0	29
34	MT		65	25	25	65	55	56
35	DT		30	25	20	15	5	18
36	MT		65	0	25	5	0	23
37	DT		20	10	50	60	75	52
38	MT		80	55	20	65	80	54
39	DT		30	30	10	15	10	15
40	MT		50	50	30	80	60	58
41	DT		40	10	0	50	20	29
48	OT		55	20	20	20	0	20
49	OT		0	5	0	0	0	0
50	OT		60	65	30	45	10	34
51	OT		45	25	15	25	10	22
52	OT		80	25	15	60	80	60
53	OT		45	25	15	70	70	46
54	OT		50	55	50	25	20	38
55	OT		55	50	30	20	25	37
56	OT		0	0	0	25	0	5
57	MT		35	45	35	20	0	24
58	DT		50	25	15	20	0	23
59	MT		15	15	10	20	15	13
60	DT		30	0	0	15	10	15
61	MT		30	40	15	35	30	30
62	DT		20	30	20	30	0	14
63	MT		35	0	35	15	20	22
64	DT		30	33	18	29	21	28
65	MT		60	50	40	55	60	54

Fußnote. transformiert -1*5 & P invertiert

Tabelle 71. Datensatz NASA-TLX versus TQ – S3

Proband Nr.	UV TQ PE	EE	SI	FC	HM	HT	BI	
8	MT	5,00	5,00	4,33	4,00	6,00	1,33	3,33
9	DT	3,00	5,00	3,00	3,00	5,00	1,33	2,67
10	MT	3,67	4,50	4,00	3,50	4,00	3,00	5,00
11	DT	4,00	5,75	1,33	4,00	6,00	1,33	4,67
12	MT	4,67	3,50	4,00	4,00	4,67	1,33	1,33
13	DT	2,33	2,50	4,33	3,50	4,67	1,00	2,33

Proband Nr.	UV TQ PE	EE	SI	FC	HM	HT	BI	
14	MT	3,00	6,25	4,00	4,00	4,00	1,00	2,67
15	DT	5,00	4,75	4,00	3,50	5,00	2,67	4,33
16	MT	3,33	5,75	4,00	4,00	4,00	1,00	3,67
17	DT	4,67	5,25	4,00	6,00	6,00	1,33	4,00
18	MT	5,00	6,00	4,67	4,00	6,00	3,67	3,33
19	DT	3,00	5,50	3,00	5,50	6,33	1,33	3,67
20	MT	4,00	5,75	4,33	5,50	7,00	5,33	2,67
21	DT	4,00	5,00	4,00	5,00	5,00	1,00	1,33
22	MT	3,33	5,75	2,00	3,00	6,00	1,00	2,33
23	DT	4,00	5,50	4,00	3,50	7,00	3,00	4,00
24	MT	5,00	5,00	2,67	4,50	5,00	3,67	4,00
25	DT	2,33	4,75	1,00	5,00	5,33	1,00	1,00
26	MT	4,67	4,25	4,00	5,00	5,67	1,00	4,33
27	DT	5,00	5,75	5,00	5,00	6,00	4,00	4,67
28	MT	5,00	6,75	4,00	4,50	6,00	4,00	4,67
29	DT	4,00	4,75	3,00	4,00	5,67	1,33	3,00
30	MT	5,33	5,25	4,00	5,00	6,00	4,00	5,33
31	DT	5,00	6,50	4,00	3,50	7,00	1,00	5,33
32	MT	4,33	4,00	5,00	4,00	4,00	4,00	4,33
33	DT	2,67	6,25	3,00	6,00	6,00	2,00	4,00
34	MT	5,33	7,00	4,00	3,00	5,67	1,67	4,33
35	DT	5,00	5,75	5,00	5,00	7,00	3,33	6,00
36	MT	4,67	6,50	4,00	5,00	5,33	2,67	3,67
37	DT	3,67	6,25	3,33	4,50	6,00	3,33	4,33
38	MT	4,67	5,00	4,00	4,50	6,00	4,33	5,33
39	DT	3,00	6,00	4,00	5,00	7,00	1,33	4,00
40	MT	2,00	4,00	3,33	4,50	4,33	2,67	2,00
41	DT	5,33	5,50	2,00	4,50	5,00	1,33	5,00
48	OT	3,33	6,50	2,67	5,00	5,33	2,00	3,67
49	OT	5,00	6,25	4,33	7,00	6,33	3,67	4,67
50	OT	5,33	5,50	4,00	5,00	6,33	3,67	5,33
51	OT	4,33	6,00	1,00	4,50	6,33	1,00	3,67
52	OT	6,00	5,25	4,67	4,00	6,33	1,33	6,00
53	OT	4,33	6,00	4,00	5,00	6,00	1,00	3,33
54	OT	6,67	6,50	5,33	5,50	7,00	5,00	6,00
55	OT	4,67	6,75	5,67	6,00	7,00	6,00	6,00
56	OT	4,33	6,75	4,00	7,00	7,00	1,00	2,67
57	MT	4,33	6,00	1,00	5,50	5,00	1,00	2,33
58	DT	4,00	4,50	1,00	4,00	6,00	1,00	5,67
59	MT	4,00	6,25	3,33	3,50	6,67	1,67	4,00
60	DT	3,67	5,25	2,00	5,50	7,00	4,00	5,00
61	MT	4,33	4,75	4,00	5,50	7,00	2,00	6,00
62	DT	5,00	6,25	4,00	4,00	6,33	2,33	4,67
63	MT	6,00	6,25	4,67	6,50	7,00	2,67	5,33
64	DT	4,33	5,75	3,33	4,50	6,00	2,67	3,33
65	MT	5,00	5,00	4,33	4,00	6,00	1,33	3,33

Tabelle 72. Datensatz UTAUT2 versus TQ – S3

Proband Nr.	UV TQ	SCL_RP [μ S]	SCL_AP [μ S]	HR_RP [BPM]	HR_AP [BPM]
9	DT	1,47	5,17	80,44	98,33
10	MT	1,50	2,13	51,81	61,12
11	DT	6,99	7,58	78,76	87,62
12	MT	2,86	6,05	72,39	90,28
13	DT	0,86	1,73	73,53	89,82
14	MT	2,42	2,72	71,37	74,00
15	DT	1,68	1,34	70,34	82,35
17	DT	1,68	2,16	66,27	92,67
18	MT	1,10	1,22	78,39	89,52
19	DT	5,78	7,43	64,82	82,44
20	MT	1,26	1,18	46,52	52,15
21	DT	1,92	1,62	98,64	125,57
22	MT	2,38	3,64	76,98	88,44
23	DT	0,95	1,62		74,33
24	MT	1,33	0,51	68,89	88,21
25	DT	2,94	5,69	66,85	83,66
26	MT	2,78	3,61		90,76
27	DT	0,21	0,15	79,00	94,68
28	MT	2,90	1,56	52,29	79,39
29	DT	2,71	2,79	49,12	64,29
30	MT	12,61	12,60		
31	DT	0,52	0,24	84,10	97,82
32	MT	0,45	0,32	65,59	
33	DT	2,57	2,89	97,84	118,87
34	MT	1,15	1,81	80,15	95,74
35	DT	1,82	4,00	97,85	121,81
36	MT	3,00	3,40	75,38	89,00
37	DT	2,19	1,60		95,24
38	MT	4,02	4,11	76,32	94,66
39	DT	4,01	5,66	68,79	83,50
40	MT	2,25	3,31		79,11
41	DT	1,94	1,46	88,27	96,51
49	OT	6,45	6,09	68,90	73,44
50	OT	6,84	4,61	66,70	75,87
51	OT	7,14	5,12		82,25
52	OT	5,46	3,80		
53	OT	2,47	1,20		
54	OT	0,23	2,10		78,91
55	OT	10,30	9,75		104,60
57	OT	14,38	11,44		94,01
58	OT	24,60	24,23		56,82
59	OT	1,66	3,59		79,97
60	OT	0,15	0,08	71,02	80,52
61	OT	0,53	0,47	72,40	83,61
62	OT	4,19	3,40	75,39	81,54
63	OT	0,30	0,42	76,03	90,83
64	OT	0,43	0,37	63,77	68,21
65	OT	5,77	5,40	69,16	80,66

Tabelle 73. Datensatz SCL und HR versus TQ – S3

A6.1 – Studie 4 Fehlervergleich und -manipulation

Proband Nr./ Reihenfolge	DLZ [s]	Richtigkeit	Leistung [%/(s*10)]	Indikator	Fehlertyp
1_EER	254	0,91	3,57	nein	ohne
1_EER	276	0,88	3,20	nein	Interaktion
2_SER	116	0,98	8,43	nein	Sichtbarkeit
2_SER	114	1,00	8,77	nein	ohne
4_EER	214	0,99	4,62	nein	ohne
4_EER	259	1,00	3,85	nein	Daten
5_SER	417	0,94	2,25	nein	Sichtbarkeit
5_SER	97	0,88	9,11	nein	ohne
7_SER	342	0,80	2,34	nein	Sichtbarkeit
7_SER	269	0,79	2,92	nein	ohne
8_EER	336	0,96	2,85	nein	ohne
8_EER	441	0,78	1,77	nein	Daten
9_SER	324	0,99	3,07	nein	Sichtbarkeit
9_SER	201	1,00	4,98	nein	ohne
10_EER	336	0,99	2,96	nein	ohne
10_EER	444	0,98	2,20	nein	Interaktion
11_SER	319	0,96	3,00	nein	Interaktion
11_SER	254	0,99	3,89	nein	ohne
12_SER	319	1,00	3,14	nein	Sichtbarkeit
12_SER	246	0,99	4,02	nein	ohne
13_EER	402	0,92	2,30	nein	ohne
13_EER	449	0,93	2,06	nein	Daten
15_SER	930	0,91	0,98	nein	Interaktion
15_SER	506	1,00	1,98	nein	ohne
16_SER	217	0,98	4,51	nein	Daten
16_SER	180	0,97	5,40	nein	ohne
18_EER	92	0,96	10,45	nein	ohne
18_EER	105	0,98	9,31	nein	Sichtbarkeit
19_SER	334	1,00	2,99	nein	Interaktion
19_SER	162	1,00	6,17	nein	ohne
20_SER	529	1,00	1,89	nein	Daten
20_SER	349	0,99	2,83	nein	ohne
21_EER	346	0,84	2,43	nein	ohne
21_EER	826	0,65	0,79	nein	Sichtbarkeit
22_EER	267	0,99	3,70	nein	ohne
22_EER	543	0,98	1,80	nein	Sichtbarkeit
23_SER	536	0,98	1,82	nein	Daten
23_SER	274	0,98	3,59	nein	ohne
24_SER	481	0,98	2,05	nein	Interaktion
24_SER	146	0,97	6,66	nein	ohne
25_EER	174	0,99	5,72	nein	ohne
25_EER	570	1,00	1,75	nein	Interaktion
26_EER	187	0,99	5,32	nein	ohne
26_EER	304	0,99	3,27	nein	Sichtbarkeit
27_SER	877	0,99	1,13	nein	Daten
27_SER	69	0,99	14,41	nein	ohne
28_EER	197	1,00	5,08	nein	ohne

Proband Nr./ Reihenfolge	DLZ [s]	Richtigkeit	Leistung [%/(s*10)]	Indikator	Fehlertyp
28_EER	234	0,97	4,16 nein		Interaktion
29_EER	207	0,99	4,78 nein		ohne
29_EER	211	0,98	4,66 nein		Sichtbarkeit
30_SER	706	0,81	1,15 nein		Daten
30_SER	162	0,97	5,97 nein		ohne
31_EER	646	0,99	1,54 nein		ohne
31_EER	788	1,00	1,27 nein		Daten
32_EER	331	0,98	2,97 nein		ohne
32_EER	486	0,99	2,05 nein		Interaktion
33_SER	804	1,00	1,24 nein		Interaktion
33_SER	165	0,99	5,99 nein		ohne
34_EER	167	0,82	4,93 nein		ohne
34_EER	204	0,91	4,46 nein		Daten
35_EER	603	0,99	1,65 nein		ohne
35_EER	808	0,99	1,23 nein		Sichtbarkeit
36_SER	374	1,00	2,67 nein		Interaktion
36_SER	254	1,00	3,94 nein		ohne
37_SER	379	0,98	2,58 nein		Sichtbarkeit
37_SER	271	0,97	3,59 nein		ohne
39_EER	217	0,99	4,58 nein		ohne
39_EER	297	0,99	3,35 nein		Interaktion
40_EER	117	0,98	8,41 nein		ohne
40_EER	182	0,99	5,45 nein		Daten
41_SER	165	0,92	5,56 nein		Sichtbarkeit
41_SER	139	0,95	6,84 nein		ohne
42_SER	519	0,99	1,91 nein		Interaktion
42_SER	154	0,99	6,42 nein		ohne
43_SER	421	0,80	1,90 nein		Daten
43_SER	180	0,87	4,85 nein		ohne
44_EER	357	0,92	2,57 nein		ohne
44_EER	359	0,95	2,63 nein		Daten
45_SER	249	0,99	3,97 nein		Daten
45_SER	222	0,98	4,41 nein		ohne
46_EER	297	0,99	3,86 nein		ohne
46_EER	256	0,98	3,31 nein		Sichtbarkeit
47_EER	464	0,98	2,11 nein		ohne
47_EER	798	0,95	1,19 nein		Interaktion
48_SER	271	0,99	3,65 nein		Sichtbarkeit
48_SER	259	0,99	3,84 nein		ohne
49_EER	389	0,98	1,93 nein		ohne
49_EER	511	0,98	2,53 nein		Daten
50_SER	424	0,99	2,35 nein		Interaktion
50_SER	364	0,97	2,66 nein		ohne
51_SER	97	0,98	10,08 nein		Daten
51_SER	87	0,99	11,37 nein		ohne
52_EER	157	0,96	6,37 nein		ohne
52_EER	150	1,00	6,37 nein		Sichtbarkeit
53_EER	167	0,98	5,86 nein		Sichtbarkeit
53_EER	201	0,99	4,95 nein		ohne

Proband Nr./ Reihenfolge	DLZ [s]	Richtigkeit	Leistung [%/(s*10)]	Indikator	Fehlertyp
54_EER	84	0,99	11,77	nein	ohne
54_EER	229	0,99	4,34	nein	Sichtbarkeit
55_SER	598	0,99	1,66	nein	Sichtbarkeit
55_SER	291	0,99	3,42	nein	ohne
56_EER	277	1,00	3,61	nein	ohne
56_EER	462	0,99	2,14	ja	Interaktion
57_SER	329	0,99	3,02	nein	Interaktion
57_SER	237	0,96	4,06	nein	ohne
58_SER	312	0,98	3,13	ja	Interaktion
58_SER	165	0,98	5,96	nein	ohne
59_SER	663	0,88	1,33	ja	Interaktion
59_SER	269	0,86	3,20	nein	ohne
60_EER	159	0,97	6,12	nein	ohne
60_EER	384	0,99	2,59	ja	Interaktion
61_EER	105	0,99	9,47	nein	ohne
61_EER	137	0,98	7,14	ja	Interaktion
63_EER	194	0,99	5,13	nein	ohne
63_EER	454	0,99	2,18	ja	Interaktion
64_SER	354	0,96	2,72	nein	Daten
64_SER	204	0,98	4,82	nein	ohne
65_EER	254	0,99	3,89	nein	ohne
65_EER	321	0,90	2,80	nein	Daten
66_EER	174	0,99	5,72	nein	ohne
66_EER	252	1,00	3,96	nein	Daten
67_SER	339	0,99	2,93	nein	Interaktion
67_SER	99	0,99	9,99	nein	ohne
68_EER	351	0,94	2,68	nein	ohne
68_EER	471	0,87	1,84	nein	Interaktion
69_EER	204	0,96	4,71	nein	ohne
69_EER	366	0,98	2,66	nein	Daten
70_SER	406	0,99	2,45	nein	Daten
70_SER	182	0,99	5,43	nein	ohne

Fußnote. In der ersten Spalte die RF mit SER = Anfangsfehler und EER = Endfehler codiert

Tabelle 74. Datensatz DLZ, Richtigkeit und Leistung versus Fehlertyp – S4

Proband Nr./ Reihenfolge	MD1	MD2	PD1	PD2	TD1	TD2	P1	P2	E1	E2	F1	F2
1_EER	16	65	10	95	65	75	25	40	35	85	15	80
2_SER	4	75	25	10	75	85	55	25	75	20	80	30
4_EER	7	70	20	70	35	30	15	55	70	70	25	70
5_SER	8	80	20	10	50	10	25	25	45	20	30	20
7_SER	14	75	25	25	25	20	30	40	35	25	25	60
8_SER	17	80	20	35	20	70	20	90	35	65	20	75
9_SER	11	50	35	65	55	65	20	50	25	40	20	45
10_SER	16	90	70	70	70	65	65	55	65	75	80	65
11_SER	15	80	20	55	5	10	55	30	70	85	30	15
12_SER	19	100	75	85	30	55	50	20	95	80	95	75
13_EER	2	35	10	5	45	20	80	65	50	25	20	10

Proband Nr./ Reihenfolge	MD1	MD2	PD1	PD2	TD1	TD2	P1	P2	E1	E2	F1	F2
15_SER	12	75	20	20	35	75	40	25	60	65	50	45
16_SER	2	5	20	15	0	0	5	5	5	15	5	0
18_EER	1	35	5	0	35	15	25	0	40	45	50	20
19_SER	7	65	10	30	35	10	35	10	30	15	75	5
20_SER	15	55	25	30	55	60	35	65	60	75	55	80
21_EER	18	100	50	45	90	50	75	5	100	65	85	45
22_EER	14	75	60	45	40	40	65	65	55	40	45	30
23_SER	18	80	70	80	90	90	30	15	90	90	95	95
24_SER	4	30	20	45	55	60	80	50	60	25	50	10
25_EER	16	35	10	20	35	80	15	65	25	80	0	75
26_EER	13	80	20	10	25	20	65	45	70	35	70	25
27_SER	5	75	25	15	75	60	65	5	80	15	80	15
28_EER	3	5	15	10	5	10	15	25	10	5	35	45
29_EER	5	60	15	10	60	20	65	65	60	20	70	25
30_SER	9	60	20	25	65	60	70	10	55	30	80	40
31_EER	17	95	50	70	70	75	45	55	95	85	70	75
32_EER	21	85	10	65	85	95	75	85	65	90	45	95
33_SER	6	65	55	40	20	25	30	15	90	55	70	55
34_EER	8	70	35	60	10	70	70	70	90	60	95	75
35_EER	13	80	40	60	30	75	85	95	80	80	50	70
36_SER	11	70	30	65	80	55	15	5	55	65	75	40
37_SER	7	75	70	65	55	60	70	45	60	65	45	35
39_EER	15	45	35	65	75	40	35	50	45	75	55	65
40_EER	5	85	35	10	70	20	65	100	75	20	70	15
41_SER	10	75	15	45	75	40	50	45	60	50	50	40
42_SER	5	100	15	20	75	25	40	15	90	45	80	10
43_SER	19	30	5	95	95	80	85	90	90	80	85	75
44_EER	19	95	70	60	55	35	60	95	80	65	60	60
45_SER	3	35	10	30	0	30	35	25	20	15	10	25
46_EER	17	70	20	25	45	50	15	5	50	65	20	0
47_EER	17	75	25	20	70	75	65	75	90	75	65	50
48_SER	4	70	20	25	60	35	35	25	55	35	65	25
49_EER	19	70	5	40	35	15	15	45	100	65	30	50
50_SER	19	85	5	10	70	95	65	30	65	85	70	45
51_SER	5	75	45	35	100	80	25	5	25	95	50	10
52_EER	4	50	55	65	80	55	50	50	25	55	10	60
53_EER	6	15	70	15	10	5	75	20	35	15	70	20
54_EER	9	50	0	0	10	55	65	25	65	35	55	25
55_SER	5	85	90	100	55	55	35	75	100	65	85	80
56_EER	15	55	25	30	50	70	45	75	30	45	30	55
57_SER	5	85	90	100	55	55	35	75	100	65	85	80
58_SER	17	75	75	70	50	35	70	35	60	35	10	10
59_SER	18	80	25	70	35	25	75	80	70	80	75	90
60_EER	14	30	20	45	55	40	25	55	50	75	10	70
61_EER	5	75	20	15	10	20	15	5	45	45	5	100
63_EER	13	75	70	65	20	15	40	35	70	60	100	10
64_SER	13	40	70	45	35	30	40	55	30	60	65	55
65_SER	13	15	15	75	5	55	20	70	30	60	10	60
66_EER	16	40	15	50	45	70	20	65	50	75	25	75

Proband Nr./ Reihenfolge	MD1	MD2	PD1	PD2	TD1	TD2	P1	P2	E1	E2	F1	F2
67_SER	3	20	40	40	70	55	35	15	10	25	75	50
68_EER	7	45	35	35	30	55	55	30	75	20	70	25
69_EER	7	20	20	30	60	65	10	15	25	35	5	25
70_SER	2	30	60	30	20	15	10	10	40	20	45	10

Fußnote. transformiert -1*5 & P invertiert

Tabelle 75. Datensatz NASA-TLX versus Fehlertyp – S4

Proband / Reihenfolge	PE1	PE2	EE1	EE2	SI1	SI2	FC1	FC2	HM1	HM2	HT1	HT2	BI1	BI2
1_EER	3,33	3,00	5,25	2,25	1,00	1,00	7,00	6,00	6,00	2,67	1,00	1,00	3,33	1,33
2_SER	6,00	7,00	6,00	7,00	6,00	7,00	6,00	7,00	6,00	7,00	6,00	7,00	6,00	7,00
4_EER	3,00	3,00	5,50	5,25	4,00	4,00	5,00	5,00	5,00	5,00	3,33	2,00	2,33	2,33
5_SER	4,00	4,67	6,50	5,50	4,33	3,67	6,00	6,00	6,67	6,67	2,33	2,00	3,33	3,00
7_SER	4,67	5,00	5,25	5,00	4,33	4,33	5,00	4,50	4,00	4,00	2,67	2,67	4,00	4,33
8_SER	3,67	3,67	4,50	4,75	4,00	4,00	6,50	6,50	6,33	6,33	3,00	2,67	4,67	4,33
9_SER	4,67	4,67	5,25	4,50	4,00	2,00	4,50	4,00	6,00	5,00	1,33	1,67	3,67	3,33
10_SER	1,33	1,00	3,25	1,50	1,00	1,00	5,50	4,00	2,00	1,00	1,00	1,00	1,00	1,00
11_SER	4,67	4,67	6,25	5,25	6,33	6,00	6,00	6,50	6,67	7,00	2,67	3,33	4,67	5,33
12_SER	2,67	2,67	4,00	4,25	6,00	7,00	7,00	7,00	7,00	6,00	3,67	2,00	7,00	7,00
13_EER	5,33	5,00	6,50	6,25	2,00	2,00	6,50	6,50	7,00	7,00	1,33	1,67	4,33	4,00
15_SER	5,67	5,33	5,00	5,25	6,00	5,00	4,50	6,00	6,67	7,00	3,33	3,00	4,00	5,00
16_SER	4,67	5,33	6,75	7,00	4,00	4,00	7,00	6,50	5,67	6,00	1,33	2,67	2,67	3,00
18_EER	4,00	5,00	6,25	7,00	4,00	4,00	6,50	6,00	5,33	5,67	4,00	4,00	2,67	4,00
19_SER	4,67	4,00	6,00	5,50	2,00	3,33	4,50	5,50	7,00	6,33	2,00	2,33	3,67	3,67
20_SER	5,33	6,00	5,75	5,50	2,00	2,00	6,00	6,00	7,00	7,00	4,33	5,00	3,67	3,67
21_EER	3,33	4,67	3,25	4,00	3,33	4,00	3,00	3,50	5,00	5,67	1,67	3,33	1,67	3,00
22_EER	5,33	4,00	5,25	5,25	5,00	4,33	4,50	4,00	6,33	6,67	3,33	4,00	5,33	5,33
23_SER	5,67	5,67	6,25	6,50	1,67	1,00	4,50	4,50	6,67	7,00	1,67	2,00	3,33	5,67
24_SER	5,33	5,67	6,00	7,00	4,00	4,67	5,50	6,50	7,00	7,00	3,67	4,33	5,67	5,67
25_EER	5,33	5,33	5,50	5,00	4,00	4,00	5,50	4,50	7,00	7,00	2,67	4,33	5,00	4,67
26_EER	5,33	6,00	5,25	5,50	4,00	4,00	6,00	6,50	5,67	6,33	1,00	1,33	5,67	6,33
27_SER	2,67	5,33	4,75	5,75	4,00	4,00	4,50	5,00	4,33	5,00	3,67	4,00	3,00	4,67
28_EER	5,33	5,67	5,50	6,50	4,00	4,00	4,00	4,00	6,00	6,00	1,33	1,67	2,00	2,33
29_EER	5,33	5,33	6,50	5,50	4,33	4,33	6,00	5,50	7,00	7,00	1,33	1,00	3,33	3,00
30_SER	3,33	4,33	3,50	4,25	4,00	3,00	5,00	4,50	3,33	3,67	1,67	2,00	3,67	4,00
31_EER	1,33	1,00	3,00	2,50	4,00	1,00	6,00	5,00	1,00	1,33	1,00	1,00	1,67	2,33
32_EER	4,67	4,33	5,50	5,25	4,00	4,00	5,00	5,50	6,00	6,00	4,00	4,00	4,00	4,00
33_SER	5,00	5,67	4,50	5,50	6,00	6,00	7,00	7,00	5,67	6,00	2,33	2,33	3,00	2,00
34_EER	5,00	5,33	6,50	5,50	4,33	4,00	4,00	4,00	6,33	6,33	3,33	4,00	4,67	4,67
35_EER	3,67	4,33	3,50	5,75	4,67	4,33	6,50	5,00	5,33	5,67	1,33	1,67	5,33	5,33
36_SER	3,00	5,00	4,00	4,75	1,00	1,00	5,50	6,00	6,67	7,00	2,00	3,33	3,67	5,00
37_SER	3,67	2,33	6,50	6,50	4,00	4,00	7,00	5,50	6,00	5,33	1,00	1,00	4,00	1,67
39_EER	5,33	3,67	4,75	3,75	4,00	4,00	5,50	6,00	6,00	5,33	2,00	1,67	5,33	4,67
40_EER	3,00	3,00	3,75	5,25	4,00	4,00	3,50	4,00	6,00	5,67	1,00	1,00	1,67	1,67
41_SER	3,67	3,67	5,00	5,50	2,67	2,00	5,50	6,00	5,67	6,33	1,33	1,33	2,33	2,33
42_SER	3,67	6,00	3,75	5,75	4,00	4,00	7,00	6,00	6,33	7,00	1,00	2,33	5,00	6,00
43_SER	2,33	2,00	2,25	2,25	4,00	4,67	4,00	4,00	6,00	5,00	1,00	2,00	2,00	2,67
44_EER	2,33	3,00	2,25	2,25	4,00	4,00	4,00	3,00	2,33	2,00	2,00	1,67	4,00	2,00
45_SER	5,00	4,67	4,50	5,75	5,00	5,67	6,50	6,00	5,00	5,00	1,33	1,67	4,00	3,33

Proband / Reihenfolge	PE1	PE2	EE1	EE2	SI1	SI2	FC1	FC2	HM1	HM2	HT1	HT2	BI1	BI2
46_EER	6,00	6,33	7,00	7,00	5,33	5,00	5,50	6,50	6,33	7,00	3,00	3,00	6,00	4,67
47_EER	2,67	2,00	2,00	2,75	1,00	2,00	1,00	2,00	1,00	1,00	1,00	1,00	1,00	1,33
48_SER	4,33	5,67	4,00	5,00	2,00	2,33	4,00	6,00	4,33	6,00	1,33	2,33	1,00	1,67
49_EER	4,67	4,67	6,50	6,00	4,00	4,00	3,50	4,50	6,33	6,67	2,33	3,33	4,33	4,67
50_SER	3,67	3,33	4,50	4,00	2,00	2,00	3,50	4,00	5,00	4,00	2,33	2,33	2,33	2,33
51_SER	3,00	5,33	5,50	6,25	1,00	1,00	3,50	3,50	5,67	6,33	2,67	5,00	3,33	4,33
52_EER	6,33	6,67	6,50	6,75	4,33	4,00	5,50	6,00	7,00	6,67	2,33	2,33	4,00	6,00
53_EER	2,67	6,33	4,25	4,25	4,00	4,00	5,00	4,00	4,33	4,00	2,67	2,33	4,00	4,00
54_EER	6,33	6,33	5,75	6,50	4,33	4,00	6,00	6,00	7,00	7,00	2,67	2,67	6,00	6,33
55_SER	5,00	4,67	6,25	6,50	4,00	4,00	5,00	5,00	6,67	6,67	2,00	2,00	4,33	4,00
56_EER	5,67	5,67	5,75	6,25	4,33	5,33	6,00	6,00	7,00	7,00	3,33	3,00	3,00	4,00
57_SER	1,00	3,33	3,75	4,00	1,00	1,00	6,00	4,50	2,67	4,67	1,00	1,67	1,00	1,33
58_SER	3,33	4,00	3,50	5,25	1,00	1,00	3,50	3,50	6,33	6,67	2,00	2,33	2,00	1,33
59_SER	3,00	1,33	5,00	4,50	1,67	1,00	4,50	4,50	7,00	7,00	1,33	1,00	1,33	1,00
60_EER	5,00	4,33	5,75	5,00	5,00	5,00	5,50	6,00	5,33	4,67	2,00	2,33	5,00	4,67
61_EER	3,67	4,00	6,25	6,00	3,00	4,00	7,00	7,00	7,00	7,00	3,67	1,67	4,67	4,00
63_EER	5,33	6,00	5,75	4,25	6,67	6,67	6,50	6,00	7,00	7,00	4,00	3,00	5,00	7,00
64_SER	2,33	2,33	2,50	2,75	2,67	3,00	3,00	2,50	2,33	3,67	2,00	3,33	3,00	3,67
65_SER	5,67	5,00	6,25	5,00	4,67	4,33	5,50	5,00	6,00	6,00	3,33	1,67	1,67	3,00
66_EER	6,00	3,33	6,75	6,00	6,33	3,00	5,50	3,50	7,00	7,00	7,00	6,00	7,00	5,67
67_SER	1,67	2,33	3,75	4,25	4,00	2,00	6,50	6,00	2,33	2,33	1,00	1,00	2,33	2,33
68_EER	3,67	3,67	4,75	5,50	3,67	4,00	6,00	5,50	3,00	4,67	1,33	1,33	3,67	2,33
69_EER	6,33	5,33	7,00	7,00	4,00	6,00	7,00	6,00	7,00	7,00	4,00	3,33	5,00	5,67
70_SER	6,00	6,67	6,25	6,50	4,67	5,00	5,00	6,00	7,00	7,00	3,67	4,00	4,33	4,67

Tabelle 76. Datensatz UTAUT2 versus Fehlertyp – S4

Proband / Reihenfolge	HR RP [BPM]	HR_Fehler AP [BPM]	HR_ohne AP [BPM]	SCL RP [µS]	SCL_Fehler AP [µS]	SCL_ohne AP [µS]
1_EER	63,49	84,65	80,37	0,31	1,70	0,87
2_SER	62,95	64,46	66,91	0,16	0,47	0,93
4_EER	64,08	92,26	90,11	0,07	3,36	2,14
5_SER	95,46	113,62	80,37	0,89	1,72	0,87
7_SER	85,54	83,41	83,04	3,46	9,03	7,28
8_SER	82,09	97,19	95,41	0,53	0,69	0,66
9_SER	65,48	80,99	81,94	0,58	0,66	0,52
10_SER	89,65	83,34	87,98	0,93	3,69	3,68
11_SER	74,56	85,34	86,37	0,42	1,00	1,67
12_SER	86,38	86,19	86,48	0,70	1,41	1,43
13_EER	85,03	91,59	96,74	5,19	2,25	4,66
15_SER	61,45	75,26	75,35	0,49	0,54	0,62
16_SER	67,02	82,01	73,58	0,38	2,65	1,55
18_EER	72,30	75,54	102,59	0,50	1,09	1,28
19_SER	88,92	79,86	68,32	2,01	3,89	1,66
20_SER	49,81	73,09	54,99	1,73	0,89	0,81
21_EER	68,56	110,72	93,37	0,22	2,68	2,42
22_EER	93,64	103,50	105,12	0,46	0,61	1,27
23_SER	74,18	102,82	102,82	0,30	0,43	0,43
24_SER	79,22	82,91	84,41	0,39	0,37	0,50
25_EER	88,15	82,41	86,89	0,24	1,02	0,86

Proband / Reihenfolge	HR RP [BPM]	HR_Fehler AP [BPM]	HR_ohne AP [BPM]	SCL RP [μS]	SCL_Fehler AP [μS]	SCL_ohne AP [μS]
26_EER	80,98	79,95	91,97	0,29	0,74	0,72
27_SER	77,18	85,91	74,10	1,36	2,94	3,44
28_EER	62,91	78,89	86,79	0,43	1,19	0,87
29_EER	66,07	75,25	78,46	0,10	0,15	0,14
30_SER	90,99	75,63	77,54	1,15	3,48	3,68
31_EER	73,24	89,65	94,36	1,69	5,22	5,94
32_EER	113,01	130,56	137,61	0,83	10,51	7,10
33_SER	81,91	93,17	76,94	0,36	0,50	0,58
34_EER	93,79	87,61	88,96	0,65	0,48	0,55
35_EER	79,27	88,20	94,56	0,16	2,79	1,40
36_SER	93,55	88,70	92,49	0,05	0,19	0,39
37_SER	89,55	88,48	79,12	0,36	2,20	5,28
39_EER	92,22	93,53	78,62	3,44	0,11	3,06
40_EER	78,05	84,03	85,13	0,34	0,45	0,48
41_SER	85,19	75,62	122,84	22,80	17,96	19,16
42_SER	96,74	71,95	74,54	1,96	1,65	1,57
43_SER	87,60	87,24	85,66	0,67	0,53	0,54
44_EER	77,51	79,05	79,99	0,41	0,58	0,47
45_SER	68,74	77,57	70,43	0,24	0,30	0,35
46_EER	90,59	86,62	80,63	0,46	1,28	1,09
47_EER	67,36	76,86	83,27	0,25	5,07	6,81
48_SER	72,79	75,32	74,93	2,06	4,54	4,16
49_EER	92,16	104,35	103,80	0,20	0,45	0,45
50_SER	70,93	90,91	88,00	3,33	9,41	8,92
51_SER	75,00	80,06	95,97	3,82	11,50	13,18
52_EER	76,28	93,20	108,79	0,43	6,69	4,94
53_EER	94,91	122,33	113,02	0,67	6,60	3,73
54_EER	85,87	102,38	79,73	14,76	6,54	4,33
55_SER	89,38	106,47	97,16	2,67	4,46	5,05
56_EER	95,45	113,27	124,37	2,45	2,61	2,49
57_SER	67,94	80,46	94,71	0,09	1,75	1,32
58_SER	94,81	101,73	103,06	0,31	1,48	1,41
59_SER	99,68	116,13	89,26	0,58	2,85	1,74
60_EER	75,18	84,09	89,29	8,40	12,93	10,87
61_EER	74,64	84,29	82,31	0,66	2,15	1,93
63_EER	84,06	79,17	86,63	7,10	5,26	4,93
64_SER	72,78	83,87	82,61	0,39	0,29	0,34
65_SER	73,90	85,77	91,60	0,51	1,79	0,91
66_EER	60,57	86,62	93,87	0,16	1,25	1,06
67_SER	76,04	82,22	78,79	2,31	3,91	3,82
68_EER	56,72	69,39	65,62	0,57	0,76	0,71
69_EER	72,15	79,35	89,68	0,19	0,52	0,52
70_SER	104,40	106,44	106,35	4,49	10,31	7,10

Tabelle 77. Datensatz physiologische Parameter – S4

UTAUT2	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
PE_Fehler	Interaktion	18	3,74	1,49	0,35
	Sichtbarkeit	19	4,93	1,11	0,26
	Daten	21	3,97	1,40	0,31

UTAUT2	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
PE_ohne	Interaktion	18	4,26	1,38	0,33
	Sichtbarkeit	19	4,67	1,35	0,31
	Daten	21	4,44	1,55	0,34
EE_Fehler	Interaktion	18	4,54	1,23	0,29
	Sichtbarkeit	19	5,59	1,02	0,23
	Daten	21	4,99	1,53	0,33
EE_ohne	Interaktion	18	4,78	1,31	0,31
	Sichtbarkeit	19	5,43	1,13	0,26
	Daten	21	5,24	1,59	0,35
SI_Fehler	Interaktion	18	3,35	1,78	0,42
	Sichtbarkeit	19	4,17	0,91	0,21
	Daten	21	3,51	1,32	0,29
SI_ohne	Interaktion	18	3,20	1,74	0,41
	Sichtbarkeit	19	4,19	1,34	0,31
	Daten	21	3,73	1,35	0,29
FC_Fehler	Interaktion	18	5,28	1,26	0,30
	Sichtbarkeit	19	5,42	1,06	0,24
	Daten	21	4,88	1,20	0,26
FC_ohne	Interaktion	18	5,28	1,40	0,33
	Sichtbarkeit	19	5,55	1,03	0,24
	Daten	21	5,00	1,26	0,27
HM_Fehler	Interaktion	18	5,04	2,00	0,47
	Sichtbarkeit	19	6,00	0,98	0,22
	Daten	21	5,40	1,76	0,38
HM_ohne	Interaktion	18	5,24	2,09	0,49
	Sichtbarkeit	19	5,91	0,92	0,21
	Daten	21	5,52	1,66	0,36
HT_Fehler	Interaktion	18	2,07	1,12	0,26
	Sichtbarkeit	19	2,49	1,29	0,30
	Daten	21	2,56	1,33	0,29
HT_ohne	Interaktion	18	2,24	1,05	0,25
	Sichtbarkeit	19	2,39	1,38	0,32
	Daten	21	2,92	1,52	0,33
BI_Fehler	Interaktion	18	3,17	1,48	0,35
	Sichtbarkeit	19	4,40	1,55	0,36
	Daten	21	3,45	1,22	0,27
BI_ohne	Interaktion	18	3,56	1,70	0,40
	Sichtbarkeit	19	4,12	1,75	0,40
	Daten	21	3,91	1,27	0,28

Tabelle 78. DD UTAUT2 versus Fehlerfall und Fehlertyp – S4

NASA-TLX	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
MD_Fehler	Interaktion	18	66,90	26,70	6,29
	Sichtbarkeit	19	59,70	27,50	6,31
	Daten	21	49,30	28,50	6,21
MD_ohne	Interaktion	18	45,30	26,90	6,33
	Sichtbarkeit	19	51,60	26,50	6,08
	Daten	21	53,60	31,10	6,78
PD_Fehler	Interaktion	18	38,10	28,30	6,67
	Sichtbarkeit	19	34,20	26,70	6,12

NASA-TLX	Fehlertyp	<i>n</i>	<i>M</i>	<i>SD</i>	<i>SE</i>
PD_ohne	Daten	21	37,10	23,90	5,20
	Interaktion	18	35,30	24,70	5,82
	Sichtbarkeit	19	40,30	28,90	6,63
TD_Fehler	Daten	21	36,70	24,60	5,36
	Interaktion	18	55,60	26,00	6,13
	Sichtbarkeit	19	45,50	20,90	4,80
TD_ohne	Daten	21	45,70	32,40	7,07
	Interaktion	18	49,70	27,30	6,42
	Sichtbarkeit	19	44,70	24,80	5,69
P_Fehler	Daten	21	50,20	24,20	5,27
	Interaktion	18	48,10	20,30	4,79
	Sichtbarkeit	19	39,20	24,90	5,71
P_ohne	Daten	21	47,90	27,80	6,06
	Interaktion	18	33,90	22,90	5,39
	Sichtbarkeit	19	49,20	22,10	5,06
E_Fehler	Daten	21	39,30	31,00	6,75
	Interaktion	18	62,50	28,60	6,73
	Sichtbarkeit	19	52,90	23,80	5,45
E_ohne	Daten	21	50,50	25,90	5,66
	Interaktion	18	52,80	25,50	6,01
	Sichtbarkeit	19	51,60	22,50	5,17
F_Fehler	Daten	21	59,80	29,30	6,40
	Interaktion	18	65,30	19,30	4,55
	Sichtbarkeit	19	42,90	26,30	6,04
F_ohne	Daten	21	50,20	29,40	6,41
	Interaktion	18	39,20	24,60	5,79
	Sichtbarkeit	19	49,20	22,20	5,09
Mean_NASA_Fehler	Daten	21	44,80	30,70	6,70
	Interaktion	18	56,10	17,30	4,08
	Sichtbarkeit	19	45,70	18,50	4,25
Mean_NASA_ohne	Daten	21	46,80	20,50	4,48
	Interaktion	18	42,70	17,30	4,08
	Sichtbarkeit	19	47,80	15,60	3,58
	Daten	21	47,40	22,40	4,88

Tabelle 79. DD NASA-TLX versus Fehlerfall und Fehlertyp – S4

A6.2 – Teilstudie 4.2 Fehlervergleich

Leistungsfaktoren	Fehlertyp	<i>n</i>	<i>MD</i>	<i>SD</i>	<i>SE</i>
Leistung_Diff	Interaktion	18	2,141	1,99	0,47
	Sichtbarkeit	19	1,622	2,11	0,50
	Daten	21	1,975	2,88	0,63
Richtigkeit_Diff	Interaktion	18	0,010	0,03	0,01
	Sichtbarkeit	19	0,006	0,05	0,01
	Daten	21	0,017	0,06	0,01
DLZ_Diff	Interaktion	18	209,111	169,96	40,06
	Sichtbarkeit	19	109,056	135,69	31,98
	Daten	21	159,762	192,29	41,96

Tabelle 80. DD Leistung-, Richtigkeit- und DLZ-Diff versus Fehlertyp – S4.2

Leistung	Fehlertyp	MD	W	p	BRK	Co'd	Power
Leistung_Diff	Interaktion Sichtbarkeit	0,519	-1,43	0,569	-0,24	-0,490	0,41
	Interaktion Daten	0,165	-0,76	0,854	-0,12	-0,240	0,17
	Sichtbarkeit Daten	-0,353	0,92	0,794	0,06	0,120	0,10
DLZ_Diff	Interaktion Sichtbarkeit	100,100	-0,97	0,774	-0,16	-0,320	0,24
	Interaktion Daten	49,300	-0,48	0,939	-0,08	-0,160	0,12
	Sichtbarkeit Daten	-50,700	0,26	0,982	0,04	0,080	0,08
Richtigkeit_Diff	Interaktion Sichtbarkeit	0,004	-3,07	0,077	-0,50	-1,150	0,95
	Interaktion Daten	-0,006	1,81	0,406	0,29	0,600	0,56
	Sichtbarkeit Daten	-0,011	1,61	0,489	0,25	0,520	0,47

Tabelle 81. DSCF-Test Leistungsfaktoren versus UV Fehlertyp – S4.2

UTAUT2	Fehlertyp	MD	t	p _{Tukey}	Co'd	Power
PE_Diff	Interaktion Sichtbarkeit	0,778	2,252	0,050	0,751	0,96
	Interaktion Daten	0,042	0,127	0,991	0,041	0,05
	Sichtbarkeit Daten	-0,736	-2,210	0,048	-0,710	0,94
EE_Diff	Interaktion Sichtbarkeit	0,417	1,402	0,347	0,467	0,64
	Interaktion Daten	-0,014	-0,049	0,999	-0,016	0,05
	Sichtbarkeit Daten	-0,431	-1,504	0,297	-0,483	0,67
FC_Diff	Interaktion Sichtbarkeit	-0,083	-0,297	0,953	0,099	0,08
	Interaktion Daten	-0,064	-0,235	0,970	-0,075	0,07
	Sichtbarkeit Daten	0,020	0,073	0,997	0,024	0,05
HM_Diff	Interaktion Sichtbarkeit	0,296	1,127	0,502	0,376	0,46
	Interaktion Daten	0,076	0,301	0,951	0,097	0,08
	Sichtbarkeit Daten	-0,219	-0,868	0,662	-0,279	0,28
BI_Diff	Interaktion Sichtbarkeit	0,666	2,317	0,049	0,772	0,97
	Interaktion Daten	-0,071	-0,255	0,965	-0,082	0,07
	Sichtbarkeit Daten	-0,737	-2,660	0,027	-0,854	0,99

Tabelle 82. PHT UTAUT2 versus UV Fehlertyp – S4.2

UTAUT2_Diff	Fehlertyp	n	MD	SD	SE
PE_Diff	Interaktion	18	0,519	0,95	0,22
	Sichtbarkeit	19	-0,259	1,16	0,27
	Daten	21	0,477	0,99	0,22
EE_Diff	Interaktion	18	0,236	1,16	0,27
	Sichtbarkeit	19	-0,181	0,81	0,19
	Daten	21	0,250	0,66	0,14
SI_Diff	Interaktion	18	-0,148	0,69	0,16
	Sichtbarkeit	19	0,017	0,70	0,17
	Daten	21	0,222	1,14	0,25
FC_Diff	Interaktion	18	0,056	0,92	0,22
	Sichtbarkeit	19	0,139	0,89	0,21
	Daten	21	0,119	0,72	0,16
HM_Diff	Interaktion	18	0,203	1,11	0,26
	Sichtbarkeit	19	-0,093	0,70	0,17
	Daten	21	0,127	0,45	0,10
HT_Diff	Interaktion	18	0,167	0,67	0,16
	Sichtbarkeit	19	-0,112	0,71	0,17
	Daten	21	0,364	0,89	0,20
BI_Diff	Interaktion	18	0,389	0,73	0,17
	Sichtbarkeit	19	-0,277	0,96	0,23

UTAUT2_Diff	Fehlertyp	<i>n</i>	<i>MD</i>	<i>SD</i>	<i>SE</i>
	Daten	21	0,460	0,87	0,19

Tabelle 83. DD Differenzen UTAUT2 versus Fehlertyp

NASA-TLX	Fehlertyp	<i>MD</i>	<i>t</i>	<i>p</i> _{Tukey}	Co'd	Power
MD_Diff	Interaktion Sichtbarkeit	16,670	1,650	0,234	0,550	0,37
	Interaktion Daten	25,950	2,666	0,027	0,856	0,74
	SichtbarkeitDaten	9,290	0,954	0,609	0,306	0,16
PD_Diff	Interaktion Sichtbarkeit	8,610	0,930	0,624	0,310	0,15
	Interaktion Daten	2,300	0,258	0,964	0,083	0,06
	SichtbarkeitDaten	-6,310	-0,707	0,760	-0,227	0,11
TD_Diff	Interaktion Sichtbarkeit	5,000	0,572	0,051	0,191	0,09
	Interaktion Daten	10,360	1,230	0,441	0,395	0,22
	SichtbarkeitDaten	5,360	0,636	0,801	0,204	0,10
P_Diff	Interaktion Sichtbarkeit	22,500	2,414	0,050	0,805	0,66
	Interaktion Daten	5,600	0,623	0,808	0,200	0,09
	SichtbarkeitDaten	-16,900	-1,882	0,154	-0,605	0,46
E_Diff	Interaktion Sichtbarkeit	10,280	1,064	0,541	0,355	0,18
	Interaktion Daten	19,010	2,041	0,112	0,656	0,51
	SichtbarkeitDaten	8,730	0,938	0,619	0,301	0,15
F_Diff	Interaktion Sichtbarkeit	33,100	2,980	0,012	0,993	0,84
	Interaktion Daten	20,600	1,930	0,140	0,602	0,45
	SichtbarkeitDaten	-12,400	-1,160	0,481	-0,375	0,21
Mean_NASA_Diff	Interaktion Sichtbarkeit	16,020	2,349	0,049	0,783	0,64
	Interaktion Daten	13,970	2,127	0,094	0,683	0,54
	SichtbarkeitDaten	-2,040	-0,311	0,948	-0,100	0,06

Tabelle 84. Studie 4.2 Post-Hoc Test NASA-TLX versus Fehlertyp – S4.2

NASA-TLX	Fehlertyp	<i>n</i>	<i>MD</i>	<i>SD</i>	<i>SE</i>
MD_Diff	Interaktion	18	21,667	23,39	5,51
	Sichtbarkeit	19	5,000	31,44	7,41
	Daten	21	-4,286	34,29	7,48
PD_Diff	Interaktion	18	2,778	29,72	7,00
	Sichtbarkeit	19	-5,833	18,17	4,28
	Daten	21	0,476	32,44	7,08
TD_Diff	Interaktion	18	5,833	24,03	5,66
	Sichtbarkeit	19	0,833	27,29	6,43
	Daten	21	-4,524	27,06	5,91
P_Diff	Interaktion	18	14,167	20,16	4,75
	Sichtbarkeit	19	-8,333	27,71	6,53
	Daten	21	8,571	33,36	7,28
E_Diff	Interaktion	18	9,722	30,32	7,15
	Sichtbarkeit	19	-0,556	25,95	6,12
	Daten	21	-9,286	30,26	6,60
F_Diff	Interaktion	18	26,111	31,74	7,48
	Sichtbarkeit	19	-6,944	33,66	7,93
	Daten	21	5,476	34,27	7,48
Mean_NASA_Diff	Interaktion	18	13,380	17,39	4,10
	Sichtbarkeit	19	-2,639	19,90	4,69
	Daten	21	-0,595	23,15	5,05

Tabelle 85. DD Differenzen NASA-TLX versus Fehlertyp – S4.2

Parameter	Fehlertyp		MD	t	p _{Tukey}	Co'd	Power
[BPM]	Daten	Interkation	-1,665	-0,426	0,905	-0,135	0,10
	Daten	Sichtbarkeit	0,928	0,234	0,970	0,075	0,07
	Interaktion	Sichtbarkeit	2,593	0,638	0,800	0,210	0,18
SCL_Diff_Diff	Daten	Interkation	-0,127	-0,321	0,945	-0,102	0,08
[µS]	Daten	Sichtbarkeit	-0,294	-0,734	0,744	-0,236	0,22
	Interaktion	Sichtbarkeit	-0,168	-0,408	0,912	-0,134	0,10

Tabelle 86. HR_Diff_Diff & SCL_Diff_Diff versus Fehlertyp –S4.2

UTAUT2_Diff	Fehlertyp	n	MD	SD	SE
HR_Diff_Diff	Daten	21	-0,076	7,43	1,62
[BPM]	Interaktion	19	1,589	9,91	2,27
	Sichtbarkeit	18	-1,004	18,02	4,25
SCL_Diff_Diff	Daten	21	0,050	1,08	0,24
[µS]	Interaktion	19	0,177	1,31	0,30
	Sichtbarkeit	18	0,344	1,36	0,32

Tabelle 87. DD HR_Diff_Diff & SCL_Diff_Diff versus Fehlertyp – S4.2