Masterarbeit

Alles eine Frage des persönlichen Optimums – Validierung des DLR-Workload Assessment Tools

---

All a matter of the personal optimum – Validation of the DLR-Workload Assessment Tool

Technische Universität Carolo-Wilhelmina zu Braunschweig

Institut für Psychologie

Abteilung für Ingenieur- und Verkehrspsychologie

Braunschweig, den 30.09.2022

Vorgelegt von: Anne Seiler

Matrikelnummer: 4790153

Erstgutachten: Prof. Dr. Mark Vollrath

Zweitgutachten: Dr. Anja Katharina Huemer

Betreuung: Dr. Jan Grippenkoven

Autorinnenhinweis

Herzlichen Dank an Dr. Mark Vollrath und meinen Betreuer Dr. Jan Grippenkoven für die kritischen Anregungen und Unterstützung über die gesamte Zeit. Ein Dankeschön geht außerdem an meine Freunde und Freundinnen und Familie für ihren Beistand und Teilnahme an meiner Studie.

Kontakt: Anne Seiler, E-Mail: a.seiler@tu-braunschweig.de

Zusammenfassung

In der heutigen Zeit verändert sich die Arbeit des Menschen aufgrund von Digitalisierung und Automatisierung immer mehr in die Richtung von Kontroll- und Überwachungsaufgaben. In solchen Situationen kann Über- oder Unterbeanspruchung entstehen. Diese unmittelbaren Auswirkungen aufgrund von externen Einflüssen können mit einer Vielzahl an Instrumenten erhoben werden. Mittels einer Simulationsstudie im Bahnkontext und einer Studie mit acht Teilexperimenten im Universitätskontext sollte überprüft werden, ob der neu entwickelte Fragebogen, der DLR-WAT, sich ebenfalls zur Erfassung von Beanspruchung eignet. Die Objektivität wurde positiv bewertet, Cronbachs Alpha für die Reliabilität ergab gute bis exzellente Werte (α = .8-.9). Für die Bestimmung der Validität wurden Korrelationen zwischen den DLR-WAT-Skalen und Skalen des bewährten Instrumentes, dem NASA-TLX berechnet. Die Korrelationen der Gesamtskalen belaufen sich auf r = .67-.81. Hinsichtlich von Zusammenhängen mit Leistungsmaßen korrelierten höhere Bewertungen der Beanspruchung meist mit schnelleren zeitlichen Reaktionen, jedoch mit schlechterer sonstiger Leistung. Zur Überprüfung der Sensitivität wurden t-Test bei abhängigen Stichproben gerechnet. In vier von fünf Situationen hat der DLR-WAT empfindlich auf Veränderung der Belastung reagiert. Zur Überprüfung der Spezifität wurden für die zweite Studie deskriptiv Mittelwertdifferenzen analysiert, wobei in drei von vier Fällen der DLR-WAT die schwerpunktartig induzierte Belastung erkannte. Laut zwei konfirmatorischer Faktorenanalysen kann die Faktorenstruktur des DLR-WAT nicht als gutes Modell bestätigt werden, jedoch konnte die hierfür nötige Stichprobengröße nicht erreicht werden. Aufgrund der gewonnenen Erkenntnisse kann der DLR-WAT geprüft, als Instrument zur Erfassung von Beanspruchung, eingesetzt werden. Der DLR-WAT sollte in praktischen Anwendungen untersucht werden, um die Annäherung an ein optimales Beanspruchungsniveau zu testen.

Schlagworte: Beanspruchung; Belastung; Validierung; DLR-WAT; Gütekriterien

Abstract

In the present the human work is changing because of digitalisation and automation towards controlling and monitoring tasks. In such situations the workload can be too high or too low. The imminent consequences on the basis of external influces can be measured by a variety of instruments. Through a train-simulation study and a study with eight partial experiments in the environment of the university the aim was to test if the newly developed questionaire DLR-WAT is also suitable for recording workload. The objectivity was rated positivly, Cronbachs Alpha for the reliability produced good to excellent results (α = .8-.9). For the determination of the validity, correlations were conducted between the scales of the DLR-WAT and scales the proven instrument, the NASA-TLX. The correlations between the whole scales are r = .67-.81. In terms of corrleations with performance measures it showed, that higher workload scores correlated mostly with faster reactions, however with poorer performance otherwise. For the examination of the sensitivity t-tests for paired samples were calculated. In four of five situations the DLR-WAT was sensitive für changes in the taskload. For the examination of the specificity, the descriptive mean differences from the second study were analysed, whereby three of four cases the DLR-WAT identified pointly induced taskload. According to two confirmatory factor analyses the factor structure of the DLR-WAT could not be confirmed, but the necessary sample size was not met. Because of the gained knowledge the DLR-WAT can be used as an examined instrument to measure workload. The DLR-WAT should be analysed in practical applications to test for the approach for an optimal wokload level.

Key words: workload; taskload; validation; DLR-WAT; quality criteria

Abkürzungsverzeichnis

Abkürzung	Bedeutung der Abkürzung
CFA	konfirmatorische Faktorenanalyse
EEG	Elektroenzephalographie
DLR-WAT	Deutsches Zentrum für Luft- und Raumfahrt-Workload Assessment Tool
NASA-TLX	National Aeronautics and Space Administration-Task Load Index
RMSEA	root mean square error of approximation
SRMR	standardized root mean square
SWAT	Subjective Workload Assessment Technique

Gliederung

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Arbeit ist ein zielgerichteter Prozess und verbraucht materielle und menschliche Ressourcen für die Erzeugung von Gütern und Dienstleistungen (Bläsing, 2020; Kauffeld, 2019). Dabei muss der Arbeits- und Gesundheitsschutz berücksichtigt werden und stets an die aktuellen Begebenheiten angepasst werden (Evers, 2009). So sind beispielsweise Automatisierung und Digitalisierung zentrale Themen der heutigen Gesellschaft. Automatisierung wird definiert, als eine Arbeitsausführung durch eine Maschine, welche zuvor von einem Menschen ausgeführt wurde (Parasuraman & Riley, 1997). Hierbei kommt es zu einer teilweisen oder vollständigen Aufgabenübernahme durch das System, wodurch sich die Tätigkeiten der Menschen ändern (Parasuraman, Sheridan & Wickens, 2000). Neue Technologien können mehr Autonomie und Entscheidungskompetenz sowie kommunikative Unterstützung bedeuten. Sie können allerdings auch dafür sorgen, dass dem Menschen vermehrt einförmige, einseitige und einfache Aufgaben überlassen werden, was zu Unterbeanspruchung führen kann (Antoni & Bungard, 1989; Richter, 2000; Wickens et al., 2010). Für den Menschen bedeutet dies eine sich verändernde Rolle. Mittelfristig werden Kontrollaufgaben, fernsteuernde Tätigkeiten sowie Entstörungsaufgaben einen Großteil der Arbeit ausmachen (Frey et al., 1992; Richter, 2000; Grippenkoven et al., 2018). Damit insbesondere die Benutzerorientierung und die Anforderungsvielfalt von den Humankriterien zur Gestaltung von Arbeitsaufgaben (DIN EN ISO 9241-2) dennoch erfüllt werden, ist eine gute Arbeitsplatz- und Arbeitsgestaltung essenziell.

Um Beschäftigte vor Unfällen und Gesundheitsschäden zu schützen, müssen Gefahren vorausschauend entdeckt werden (Nohl, 1989). Angesichts der sich verändernden Arbeitstätigkeiten wird nicht nur darauf zu achten sein, Beschäftigte nicht zu überlasten, sondern auch darauf, sie nicht mit kontinuierlichen visuellen Überwachungstätigkeiten zu unterfordern, da beides zu schlechterer Performanz führen kann (Bainbridge, 1983; Brandenburger & Jipp, 2017; Endsley & Kaber, 1999; Young et al., 2006). Schlussendlich bedeutet dies, Tätigkeiten so zu gestalten, dass der Mensch im Zusammenspiel mit den Maschinen adäquat beansprucht wird. Wo genau diese adäquate Beanspruchung liegt, ist eine zentrale Frage, welcher die Forschung auch in Hinblick auf die Entwicklung und Verwendung geeigneter Methoden nachgehen muss (Grippenkoven et al., 2018). Da bei bisherigen Verfahren keine Aussage über das persönliche Optimum der Personen generiert wird, ist ein neues Instrument nötig. Diese Lücke versucht das DLR-Workload Assessment Tool (DLR-WAT) zu schließen.

Diese Arbeit befasst sich mit der Frage, ob der DLR-WAT empirisch begründet in der Praxis angewandt werden kann, also ob der Fragebogen ein geeignetes Messinstrument zur Erfassung von Beanspruchung darstellt. Zur Beantwortung dieser Forschungsfrage wird zunächst auf die Konstrukte der Belastung und Beanspruchung eingegangen und Messinstrumente werden erläutert. Im Anschluss wird auf die Methode eingegangen und die Ergebnisse werden präsentiert. Schließlich werden die Ergebnisse diskutiert und mit den Ergebnissen anderer Studien verglichen sowie Grenzen der vorliegenden Studie aufgezeigt.

2 Theoretische Grundlagen

In diesem Kapitel wird auf die Definitionen und Grundannahmen von Belastung sowie Beanspruchung eingegangen. Außerdem werden gängige Verfahren zur Erfassung von Beanspruchung und die Fragestellung dieser Arbeit erläutert.

2.1 Belastung und Beanspruchung

In der Arbeits-, Ingenieur- und Verkehrspsychologie sind die Konstrukte Belastung und Beanspruchung von zentraler Bedeutung. Laut der internationalen Norm DIN EN ISO 10075-1 ist psychische Belastung definiert als „Gesamtheit aller erfassbaren Einflüsse, die von außen auf den Menschen zukommen und psychisch auf ihn einwirken“ (DIN EN ISO 10075-1, S. 87). Beanspruchung wird als Konsequenz der Belastung gesehen und ist definiert als „unmittelbare (nicht langfristige) Auswirkung der psychischen Belastung im Individuum in Abhängigkeit von seinen jeweiligen überdauernden und augenblicklichen Voraussetzungen, einschließlich individueller Bewältigungsstrategien“ (DIN EN ISO 10075-1, S. 87). Synonym finden sich in der Literatur die Begriffe „Stressor“, „Stressfaktor“ oder im Englischen „stress“ und „taskload“ für Belastung. Beanspruchung wird auch als „Stressreaktion“ oder im Englischen als „strain“ oder „workload“ bezeichnet (Evers, 2009; Manzey, 1998; Wickens et al., 2021).

Bekannte Ausgangssituationen und routiniertes Handeln benötigen weniger Ressourcen als Handlungen mit unbekannten Reizen. Dies geht allerdings mit einer geringeren Wachsamkeit einher und kann so zu Fehlern führen (Bläsing, 2020; Schlick et al., 2018). Kahnemann (2012) unterscheidet zwischen zwei unterschiedlichen kognitiven Systemen. Das erste System handelt basierend auf Vorerfahrungen oder Heuristiken und erzeugt so schnelle und automatische Antworten. Dies generiert allerdings nicht zwingend die richtige Reaktion. Das zweite System handelt, wenn das erste System seine Grenzen erreicht. Das zweite System ist logisch und rational ausgerichtet, ist dadurch allerdings langsamer und führt zu einer höheren Beanspruchung und

schnelleren mentalen Erschöpfung aufgrund des hohen Ressourcenverbrauchs, was Fehler generieren kann. Dementsprechend ist es wichtig ein gutes Maß zu finden, da sowohl eine zu hohe als auch eine zu geringe Belastung zu einer Fehlbeanspruchung führen können (Bläsing, 2020).

Young et al. (2015) nehmen im Red-Lines-Modell an, dass es in Verbindung mit der Beanspruchung einen symmetrischen Verlauf der Performanz gibt. Die benötigten Ressourcen steigen bis zu dem Zeitpunkt, an dem die Anforderungen die Ressourcen übersteigen, dieser Schnittpunkt wird als Start der Überbeanspruchung definiert. Beide Bereiche, also Unterbeanspruchung und Überbeanspruchung wirken sich negativ auf die Performanz aus (Wilson & Rajan, 1995; de Waard 1996). Eine mittlere Beanspruchung wird als förderlich für die Performanz gesehen, da hier Anforderungen und die Ressourcen in einem günstigen Zusammenhang zueinanderstehen (Johannsen, 1993; Hancock & Warm, 1993). Dabei können bei Individuen mit gleich hoher Performanz, unterschiedliche Beanspruchungsniveaus vorliegen (Yeh & Wickens, 1988). Außerdem müssen unterschiedliche Beanspruchungsniveaus bei verschiedenen Individuen nicht zu unterschiedlichen Leistungen führen (Sperandio, 1971).

Laut Chen et al. (2016) kann eine ansteigende mentale Beanspruchung verhindert werden, indem der Umfang der aufzunehmenden Informationen reduziert wird. Ein Design der Informationsdarbietung, bei dem alle wichtigen Informationen sofort aufgenommen und verarbeitet werden können, kann den Arbeitsfluss unterstützen (Mattsson & Fast-Berglund, 2016). Mentale Beanspruchung kann außerdem durch zu viele Alternativen, generierte Unsicherheit und zu hohe Komplexität entstehen, was sich negativ auf den Prozess der Entscheidungsfindung auswirkt (Bläsing, 2020).

Die Beachtung der Beanspruchung ist in vielen unterschiedlichen Tätigkeitsfeldern von größter Relevanz. Mit Blick auf die Montagearbeit kommt es bei Großserienmontage zu einer geringen Beanspruchung und bei den modernen Mixed-Model-Montagesystemen zu einer höheren Beanspruchung aufgrund der häufigen Wechsel und geringen Losgrößen (Bläsing, 2020; Kahnemann 2012). Laut Dunn und Williamson (2012) führt Monotonie im Bahnkontext zu geringerer Performanz, wobei sich eine leichte Erhöhung in der Beanspruchung positiv auf die Performanz auswirken kann. Aufgaben mit hoher Monotonie kennzeichnen sich durch ein geringes Stresslevel aus, wodurch die Wachsamkeit sinkt (Warm et al., 2008). Dementsprechend ist Forschung nötig, um die Performanz mithilfe von optimalen Beanspruchungsniveaus zu maximieren (Grippenkoven et al., 2018). Die individuelle Beanspruchung kann mit unterschiedlichen Methoden erfasst werden, auf die nachfolgend eingegangen werden soll.

2.2 Instrumente zur Erhebung des Beanspruchungsniveaus

Im Folgenden werden einige Instrumente erläutert, mit denen die Beanspruchung erfasst werden kann. Es wird auf physiologische Methoden, Methoden der Performanz sowie Fragebogen eingegangen.

2.2.1 Physiologische Methoden

Im Folgenden soll kurz auf einige physiologische Methoden eingegangen werden, die Rückschlüsse auf die Beanspruchung ermöglichen. Das Ziel dieser Methoden ist es eine objektivere Möglichkeit zu erhalten, um Beanspruchung zu erfassen, da sie weniger anfällig für Fehler ist. Außerdem ermöglichen es diese Instrumente Beanspruchung in Echtzeit zu erfassen (Bläsing, 2020). Ein Nachteil dieser Methoden ist es, dass es eine Chance gibt, dass bei der Erfassung von mentaler Beanspruchung Artefakte durch körperliche Beanspruchung entstehen. Um dem entgegenzuwirken, sind beispielsweise Ruhephasen für die Versuchspersonen nötig (Mehta & Parasuraman, 2013).

Anhand des Auges lassen sich Beanspruchungszustände erkennen. Möglichkeiten bestehen hier durch die Pupillometrie oder auch durch Blickbewegungen. Der Pupillendurchmesser ist einer der Parameter für die Erfassung von Beanspruchung. Der Pupillendurchmesser ist allerdings nicht nur abhängig von der Beanspruchung, sondern auch vom Lichteinfluss, von der Entfernung zu den zu fixierenden Objekten, vom Alter, von der Gesundheit und von dem aktuellen kognitiven und emotionalen Erleben der Person (Mathôt, 2018). Bei steigender neuronaler Erregung und gesteigerter Wachsamkeit dehnt sich die Pupille aus, was eine relativ schnelle Reaktion ist. Durch Veränderungen in der Pupillengröße können in weniger als einer Sekunde Veränderungen in der Beanspruchung erfassbar gemacht werden (Marquart & de Winter, 2015). Weitere Parameter, die sich aus der Blickbewegung ergeben, sind sehr situativ. In dieser Arbeit soll nicht näher darauf eingegangen werden, zum Vergleich ist es z.B. nachlesbar bei Bläsing (2020).

Elektroenzephalographie (EEG) ist eine weitere physiologische Möglichkeit um Beanspruchung zu erfassen. Aufgrund der Informationsverarbeitung nach einer Stimulusinduktion entstehen Spannungsveränderungen in Hirnarealen, welche an der Kopfoberfläche aufgezeichnet werden können. Diese EEG-Signale können in Frequenzbänder eingeteilt werden, wobei für die Beanspruchung das Theta-, Alpha- und Beta-Band in Beziehung zueinander gestellt werden müssen. Ein Vorteil von der Benutzung von EEG ergibt sich durch die gute zeitliche Auflösung hinsichtlich der der Echtzeiterfassung von der Beanspruchung. Die Benutzung von EEG ist

allerdings stark eingeschränkt, da Bewegungen und Sprechen Artefakte bei den Daten erzeugen können (Pope et al., 1995).

Elektrokardiografie ist eine weitere Methode. Hierbei wird die Erregungsausbreitung der Herzmuskelzellen genutzt. Die Technologien umfassen z.B. Pulsoxymetrie am Handgelenk, wobei bei vermehrter Bewegung die Präzision sinkt. Ein anderes Instrument ist z.B. die Erfassung mittels Elektroden bei klinischen Messgeräten (Georgiou et al., 2018). Bei dieser Methode wird die Herzfrequenz und die Herzfrequenzvariabilität genutzt, um die Beanspruchung zu erfassen. Wobei die Herzfrequenzvariabilität, wie die Intervalle zueinander variieren, ein besseres Maß ist als die Herzfrequenz alleine. Die Herzfrequenz steigt in Zuständen von Beanspruchung an und die Herzfrequenzvariabilität nimmt ab (Sammito et al., 2015). Aufgrund der vermehrten Ausbreitung von Smart-Watches, die den Puls und dessen Veränderung erfassen können, steigt die Akzeptanz bei der Nutzung solcher Geräte bei den Versuchspersonen (Bläsing, 2020).

Weitere Maße sind die Hautleitfähigkeit und die funktionelle Magnetresonanztomografie. Die Hautleitfähigkeit ist abhängig von der Schweißdrüsenproduktion. Der Hautleitwert steigt an, wenn Schweiß produziert wird. Die Schweißproduktion steigt kurzfristig bei emotional-affektiven Reaktionen oder auch bei mentaler Beanspruchung an. Die Hautleitwertreaktion kann sich allerdings in ihrem Grundniveau und in ihrer Variabilität zwischen Individuen unterscheiden. So haben jüngere Männer höhere Hautleitwerte als ältere Männer und generell Männer höhere als Frauen (Greil et al., 2008). Die funktionelle Magnetresonanztomografie ist sehr komplex und kann kaum in praktischen Kontexten genutzt werden, so ist sie z.B. im Verkehrskontext nicht anwendbar und wird in dieser Arbeit nicht näher beleuchtet. Auf dieses Verfahren geht z.B. Groth et al. (2011) näher ein.

2.2.2 Performanz-Methoden

Die Idee, die der Nutzung von Nebenaufgaben zur Beurteilung der Beanspruchung zugrunde liegt, ist die Theorie der multiplen Ressourcen. Demnach bestimmen verschiedene Ressourcen die Leistung. Gleichartige Aufgaben sind schwieriger gleichzeitig auszuführen, als solche, die sich in Aspekten unterscheiden (Wickens, 2008). Um mit dieser Methode Beanspruchung zu erfassen, muss die Nebenaufgabe mit der Hauptaufgabe interferieren. Zudem muss sie eigeständig in ihrer Schwierigkeit veränderbar sein und kontinuierlich messbare Ergebnisse liefern. Die Beanspruchung lässt sich dann berechnen als Leistung in der Nebenaufgabe minus die Leistung bei der Nebenaufgabe bei der Ausführung der Hauptaufgabe. Da durch die Interferenz der Aufgaben ein Leistungsabfall entsteht, kann diese Methode nur in Simulationen genutzt

werden. Die Versuchspersonen werden dabei so instruiert, dass die Ausführung der Hauptaufgabe Priorität im Vergleich zur Ausführung der Nebenaufgabe hat und diese nur ausgeübt werden soll, wenn die Hauptaufgabe dies zulässt (Eggemeier, 1988). Ein Beispiel für eine Interferenz ist das zeitgleiche Ablesen von Informationen bei paralleler Kommunikation mit Kollegen (Bläsing, 2020). Weitere mögliche Nebenaufgaben sind z.B. Rechenaufgaben, die n-back-Methode und die Surrogate reference task (Radlmayr, 2016).

Eine weitere Möglichkeit zur Erfassung von Beanspruchung bilden die Reaktionszeiten und die Fehlerraten der Primäraufgabe. Die Güte der Erfüllung der Primäraufgabe sinkt allerdings erst im Bereich der Überbeanspruchung. Somit ist dieses Maß nicht sehr sensitiv. Ein weiteres Problem ist die schlechte Vergleichbarkeit der Beanspruchung über verschiedene Aufgaben (Eggemeier, 1988).

2.2.3 Fragebogen

Es gibt eine Vielzahl von Fragebogen, mit denen die Beanspruchung subjektiv erfasst werden kann. Fragebogen sind sehr ökonomisch in der Durchführung und Auswertung und sorgen für bessere Vergleichbarkeit zwischen Versuchspersonen (Bläsing, 2020). Ein negativer Punkt ist allerdings, dass aufgrund der verspäteten Beantwortung im Vergleich zur Aufgabe Erinnerungsfehler auftreten können und im Gegensatz zu physiologischen Instrumenten keine kontinuierlichen Daten erfasst werden (Eggemeier, 1988). Fragebogen setzen außerdem die Fähigkeit der Versuchsperson voraus eine Introspektion vorzunehmen und die Erkenntnisse in numerische Werte umzuwandeln (Chen et al., 2012). Ein weiterer Nachteil ist es, dass es zur sozialen Erwünschtheit, also die subjektive Berücksichtigung sozialer Erwartungen, kommen kann (King & Bruner, 2000). Im Anschluss sollen die Fragebogen Rating Scale Mental Effort, NASA-TLX sowie DLR-WAT vorgestellt werden.

2.2.3.1 Rating Scale Mental Effort

Die Rating Scale Mental Effort (Zijlstra & van Doorn, 1985) wurde als eindimensionaler Fragebogen entwickelt, der mithilfe einer Linie, markiert mit neun Ankerpunkten, die Beanspruchung evaluieren soll. Bei jedem Ankerpunkt steht ein Label über das Maß an Anstrengung. Die Theorie dahinter ist, dass mit steigender Beanspruchung auch die mentale Anstrengung steigt. Der Fragebogen ist dabei unabhängig von besonderen Geräten, ist einfach zu verstehen, durchzuführen und auszuwerten. Das Instrument ist günstig, schnell in der Durchführung und kann am Arbeitsplatz ohne Interferenz bei der Arbeit durchgeführt werden (Ghanbary Sartang

et al., 2016). Der Fragbogen ermöglicht allerdings keine differenzierten Rückschlüsse auf die Beanspruchung. Diese Lücke schließt der NASA-TLX.

2.2.3.2 NASA-Task Load Index

Beim NASA-TLX wird die Beanspruchung auf sechs unabhängigen Skalen erfasst: „mental demand“, „physical demand“, „temporal demand“, „effort“, „performance“ und „frustration level“ (Hart & Staveland, 1988). Es gibt unterschiedliche Versionen vom Fragebogen. Im Folgenden wird auf die Version ohne Gewichtung eingegangen (Hart, 2006; NASA TLX Paper and Pencil Version, 2022, s. Anhang B). Die Skala geht vom Wert 0, welcher für geringe Beanspruchung steht, bis Wert 100, welcher für hohe Beanspruchung steht. Bei der Skala Aufgabenbewältigung stehen die Pole für Perfektion und Versagen. Bei der Skala Frustration stehen die Pole für niedrige und hohe Frustration. Die Skalierung ist in Fünferschritten. Der Durchschnitt aller Skalen kann als Gesamtwert der Beanspruchung betrachtet werden (Hart, 2006).

In der Zeit seit der Entwicklung des NASA-TLX hat sich die Arbeit weiterentwickelt und es gibt viele Mensch-Automations-Systeme, wodurch es zu leichten Unterforderungszuständen kommen kann (Grippenkoven, 2018). Beim NASA-TLX stellen die Skalen ein Kontinuum von sehr geringer bis sehr hoher Beanspruchung dar, ohne dabei konkret ein persönliches Optimum oder Unter- und Überbeanspruchungsbereiche zu berücksichtigen, welches bei der Weiterentwicklung des NASA-TLX, dem DLR-WAT ergänzt wurde.

2.2.3.3 DLR-Workload Assessment Tool

Die Skalen des DLR-WAT (s. Anhang A) sind angelehnt an die Skalen des NASA-TLX. Da in den modernen Arbeitskontexten die geistige Beanspruchung im Vergleich zur körperlichen Beanspruchung an Wichtigkeit gewinnt, wurde beim DLR-WAT die geistige Beanspruchung in die drei Unterskalen Informationsaufnahme, Wissensabruf und Entscheidungsfindung aufgespaltet. Dies ermöglicht eine differenziertere Betrachtung. Dies stellt die Schritte der Informationsverarbeitung von Menschen dar (Wickens & Carswell, 2012). Der Durchschnitt dieser Skalen kann dabei als Gesamtwert der geistigen Beanspruchung genutzt werden. Die anderen fünf Skalen lauten: „motorische und körperliche Beanspruchung“, „zeitliche Beanspruchung“, „Anstrengung“, „Frustration“ und „Aufgabenbewältigung“ (Grippenkoven et al., 2018).

Der DLR-WAT wurde mit dem Ziel gestaltet, sowohl Überbeanspruchung als auch Unterbeanspruchung auf mehreren Unterskalen zu erheben, insbesondere in unterschiedlichen stark automatisierten Mensch-Maschine-Systemen. Durch den neuen Maßstab eines persönlichen

Optimums soll so die Aufgabe mit der bestmöglichen Beanspruchung entdeckt werden. Das persönliche Optimum stellt dabei den Wert 100 auf den Skalen dar, welche den Wertebereich von 0 bis 200 umfassen. Der Wert 0 stellt die größtmögliche Unterbeanspruchung und der Wert 200 die größtmögliche Überbeanspruchung dar. Lediglich die beiden Skalen Frustration und Aufgabenbewältigung bilden die Ausnahme. Die Skala Frustration umfasst den Wertebereich 100 bis 200, da ein optimales Frustrationsniveau durch das Fehlen von Frustration gekennzeichnet ist. Die Skala Aufgabenbewältigung umfasst den Wertebereich 0 bis 100, da eine optimale Aufgabenbewältigung durch die maximale Aufgabenbewältigung gekennzeichnet ist. Aufgrund dieser Operationalisierung stellt über alle Skalen hinweg eine Abweichung von dem Optimum, dem Wert 100, eine Unter- bzw. Überbeanspruchung dar. Wie beim NASA-TLX kann der Durchschnitt aller Skalen als Gesamtbeanspruchung gewertet werden (Grippenkoven et al., 2018).

Im Vergleich zum NASA-TLX bietet der DLR-WAT eine explizite Unterscheidung des Beanspruchungsniveaus in einen Unterbeanspruchungs- und einen Überbeanspruchungsbereich, wobei in dessen Mitte das persönliche Optimum liegt. Diese Differenzierung entspringt dem Zieleinsatzgebiet des DLR-WAT, Aufgaben und Tätigkeiten mit möglichen Unter- und Überbeanspruchungen, wie z.B. im stark automatisierten Verkehrskontext. Da der DLR-WAT zudem im Bereich der mentalen Beanspruchung zwischen den unterschiedlichen Verarbeitungsstufen differenziert, können so die Ursachen für eine Fehlbeanspruchung besser festgestellt werden (Grippenkoven et al., 2018).

Zusammenfassend kann gesagt werden, dass es eine Vielzahl an Methoden gibt, um Beanspruchung zu erfassen. Diese haben unterschiedliche Vor- und Nachteile. Besonders im Hinblick auf die praktische Handhabbarkeit ist zu berücksichtigen, dass es in einem praktischen Anwendungsfeld schwierig ist, andere Instrumente als Fragebogen zu benutzen. Insbesondere in Feldern, bei denen Sicherheit eine große Rolle spielt, wie z.B. dem Verkehrsbereich ist dies der Fall. Fragebogen sind zudem sehr ökonomisch, sowohl was die Durchführung, als auch die gewonnenen Daten und deren Auswertung und Interpretation angeht.

2.3 Fragestellung

Diese Arbeit zielt darauf ab, den neu entwickelten Fragebogen DLR-WAT (Grippenkoven et al., 2018) empirisch zu testen. Mit den gewonnenen Ergebnissen sollen Aussagen bezüglich der Eignung des Fragebogens und Erfassung von Beanspruchung abgeleitet werden. Die starken Veränderungen in der Arbeitswelt bestätigen die Relevanz des Themas und die Notwendigkeit

auch Unterbeanspruchung sensitiv erfassen zu können und optimale Beanspruchungsniveaus für die Menschen zu finden (Grippenkoven et al., 2018). Mit dieser Untersuchung soll langfristig die Forschung im Feld der Beanspruchung und Belastung vorangebracht werden, indem Erkenntnisse zur Erfassung von Beanspruchung gewonnen werden. Dazu werden zwei Studien ausgewertet, die den DLR-WAT sowie Leistungsmaße für Beanspruchung genutzt haben. Die Auswertung fokussiert sich auf die Gütekriterien Objektivität, Reliabilität und Validität des Fragebogens. Es wird außerdem die Sensitivität, die Spezifität und die Faktorenstruktur untersucht. Auch im Hinblick auf den Vergleich mit anderen Erhebungsmaßen für Beanspruchung soll geprüft werden, ob sich der DLR-WAT als ein solches Instrument eignet.

3 Methode

Zur Validierung des DLR-WAT wurden zwei Studien durchgeführt. Die erste Studie wurde durch das DLR im Rahmen des Projekts „Next Generation Train“ durchgeführt. Die zweite Studie erfolgte an der Technischen Universität Braunschweig. In diesem Abschnitt werden beide Studien erläutert, wobei auf den Ablauf der Experimente, deren Stichproben, Durchführung sowie die statistische Analyse eingegangen wird. Für die gesamte Analyse, bis auf die konfirmatorische Faktorenanalyse, welche mit R (R Core Team, 2020) durchgeführt wurde, wurde SPSS (IBM Corp., 2021) verwendet.

3.1 Erste Studie

Die erste Studie, die vom DLR durchgeführt wurde, hatte das Ziel, die Leistungsfähigkeit von Remote-Zugoperatoren in Kontrollzentren zu untersuchen. Da in der Zukunft der Schienenverkehr einen hohen Grad an Automatisierung zu verzeichnen haben wird, ist es fraglich, wie sich die Tätigkeit und damit einhergehend das Anforderungsprofil eines Triebfahrzeugführers von dem, in der Zukunft wahrscheinlichem, Remote-Operators unterscheidet (Benderoth, unveröffentlicht).

3.1.1 Versuchsplan und -durchführung

Das within-subjects Experiment wurde an zwei Tagen über den Verlauf eines fiktiven Arbeitstages in einem Labor des Institutes für Luft- und Raumfahrtmedizin (AMSAN) durchgeführt. Die Daten wurden zwischen August 2021 und Dezember 2021 erhoben. Hierbei wurde eine neuentwickelte Simulationsumgebung für das Bahn-Kontrollzentrum genutzt. Die Versuchspersonen durchliefen erst das Setting A und am zweiten Experimentaltag des Setting B. Die Experimentaltage wurden jeweils in vier Sitzungen von jeweils 1,5 Stunden Länge eingeteilt.

Nach jeder Sitzung erfolgt eine Pause von 20 Minuten, in der Mess- und Fragebogendaten erhoben wurden (s. Abbildung 1). Vor der Erhebung fanden eine medizinische Voruntersuchung sowie ein Training statt. Zwischen den Tagen gab es eine Pause von mindestens einer Woche (Benderoth, unveröffentlicht).

Zeit	9:00 – 10:30	10:30 - 10:50	10:50 - 12:20	12:20 - 13:05	13:05 - 14:35	14:35 - 14:55	14:55 - 16:25
Task-Load A (High-Low)	High	Pause	High	Pause + Mittagessen	Low	Pause	Low
Task-Load B (Low-High)	Low	(20 Min.)	Low	(45 Min.)	High	(20 Min.)	High

Abbildung 1. Ablauf eines Versuchstages.

Die primäre Aufgabe war es, eine Simulation eines Bahn-Kontrollzentrums, genauer gesagt ein Konzept eines Fernsteuerarbeitsplatzes für Schienenfahrzeuge, an einem Computerarbeitsplatz zu bearbeiten. Die Züge operierten dabei hauptsächlich automatisiert. Dabei wurde zwischen den Sitzungen die Höhe der Belastung (hoch oder niedrig) variiert. Bei der hohen Belastung waren häufiger Reaktionen nötig, da häufiger Tiere (n = 18 bzw. 19 im Vergleich zu n = 14 bzw. 15) im Gleis auftauchten und Langsamfahrstrecken (n = 9) beachtet werden mussten. Bei beiden Gefahrenstellen mussten die Versuchspersonen möglichst schnell reagieren und den betroffenen Zug abbremsen. Verspätungen sollten dabei geringgehalten werden. Priorität hatte hierbei allerdings die Sicherheit, den Zug abzubremsen. Bei einer Gefahrenstelle sendete ein automatisch fahrender Zug eine Anfrage an die Versuchsperson, welche die Kontrolle über den Zug übernehmen konnte, um das Problem zu lösen. Als Nebenaufgabe sollten die Versuchspersonen zudem Rechenaufgaben lösen (Benderoth, unveröffentlicht).

Der Schlaf-Wach-Rhythmus wurde über die Zeit mit Hilfe von Aktigraphie kontrolliert. Erhoben wurden die psychomotorische Vigilanz mittels dreiminütigem Test, die Müdigkeit mittels der Karolinska Sleepiness Scale (Shahid et al., 2011), die Motivation mittels 5-stufiger Likert-Skala, die Übernahmezeit für die Züge (wann die Versuchspersonen aktiv die Steuerung des Zuges übernahmen), die kumulierte Verspätung sowie die Beanspruchung mittels NASA-TLX sowie DLR-WAT. In dieser Arbeit werden lediglich die Fragebogendaten sowie die Übernahmezeiten und kumulierte Verspätungen analysiert (Benderoth, unveröffentlicht).

3.1.2 Stichprobe

Es wurde elf Versuchspersonen (vier Frauen, sieben Männer, Alter: M = 28.36, SD = 5.08) erhoben. Die Versuchspersonen erhielten eine Vergütung für die Teilnahme. Alle hatten

normale Schlafgewohnheiten und keine physischen oder psychischen Besonderheiten, welche die Performanz negativ beeinflussen können (Benderoth, unveröffentlicht).

3.2 Zweite Studie

Das Ziel der zweiten Studie war weitere Experimente in Verbindung mit dem DLR-WAT durchzuführen, um eine Basis für die Validierung des Fragebogens zu schaffen. Außerdem wurden die Experimente so gewählt, dass die Spezifität von vier Skalen des DLR-WAT überprüft werden konnte.

3.2.1 Versuchsplan

Es gab acht Experimente mit einem 4x2 within-subjects-Design. Dabei gab es jeweils zwei Experimente, die hauptsächlich auf eine Skala des DLR-WAT abzielen sollten. Die Ziel-Skalen waren Informationsaufnahme, Wissensabruf, Entscheidungsfindung und zeitliche Beanspruchung. Dabei wurde jeweils ein Experiment mit der Annahme programmiert, dass damit niedrige (im Folgenden auch „einfache Bedingung“ genannt) bzw. hohe Belastung (im Folgenden auch „schwere Bedingung“ genannt) induziert wird. Für die Programmierung wurde die Seite PsyToolKit (Stoet, 2010; Stoet, 2017) genutzt. Erfasst wurden die Reaktionsgeschwindigkeit sowie die Fehler pro Experiment. Außerdem wurde nach jedem Experiment die Beanspruchung mittels NASA-TLX und DLR-WAT erhoben. Die Reihenfolge der Experimente wurde mit Hilfe von Lateinischen Quadraten festgelegt. Durch die Methode erscheint jede Bedingung gleich häufig an jeder Stelle der Abfolge (Sedlmeier & Renkewitz, 2013). Die Reihenfolge der Fragebogen nach den Experimenten wurde über die Versuchspersonen hinweg randomisiert.

3.2.2 Stichprobe

Die Versuchspersonen wurden im Umfeld der TU Braunschweig angeworben und als Belohnung für die Teilnahme wurde Kuchen ausgegeben. Es wurde 32 Versuchspersonen (22 Frauen, zehn Männer, Alter: M = 26,25, SD = 1,78) erhoben. Davon studierten 17 und 15 gingen vorrangig der Arbeit nach.

3.2.3 Durchführung

Die acht Teilexperimente sowie die Fragebogen, die im Anschluss an jedes Experiment beantwortet wurden, erfolgte online am Computer und wurde in Zeitraum vom 08.09. bis 18.09.2022 durchgeführt. Insgesamt dauerte die Studie ca. 30 Minuten. Die Versuchspersonen wurden zuvor auf Datenschutzrechte hingewiesen und aufgefordert die einzelnen Experimente schnell durchzuführen, da die Reaktionszeit erfasst wurde.

Die zwei Experimente, die besonders einen Einfluss auf die Skala Informationsaufnahme vom DLR-WAT haben sollten, wurden so gestaltet, dass auf dem Bildschirm rote Ts gefunden werden mussten (s. Abbildung 2). Wenn es eins gab, musste „j“ gedrückt werden, wenn es nur rote Ts gab, die falsch herum dargestellt wurden oder blaue Ts zu sehen waren, musste „f“ gedrückt werden. In dem Setting mit niedriger Belastung wurde in allen 50 Durchgängen jeweils nur ein Zielreiz bzw. Distraktor gezeigt. Im Setting mit hoher Belastung wurden jeweils um die 17 Distraktoren gezeigt, unabhängig davon ob der Zielreiz auch vorhanden war oder nicht. Vor dem Versuch wurde die Instruktion gezeigt und es gab eine Trainingsphase, bei der es ein Feedback gab. Die Reaktionszeit sowie die Fehlerrate wurden erfasst.

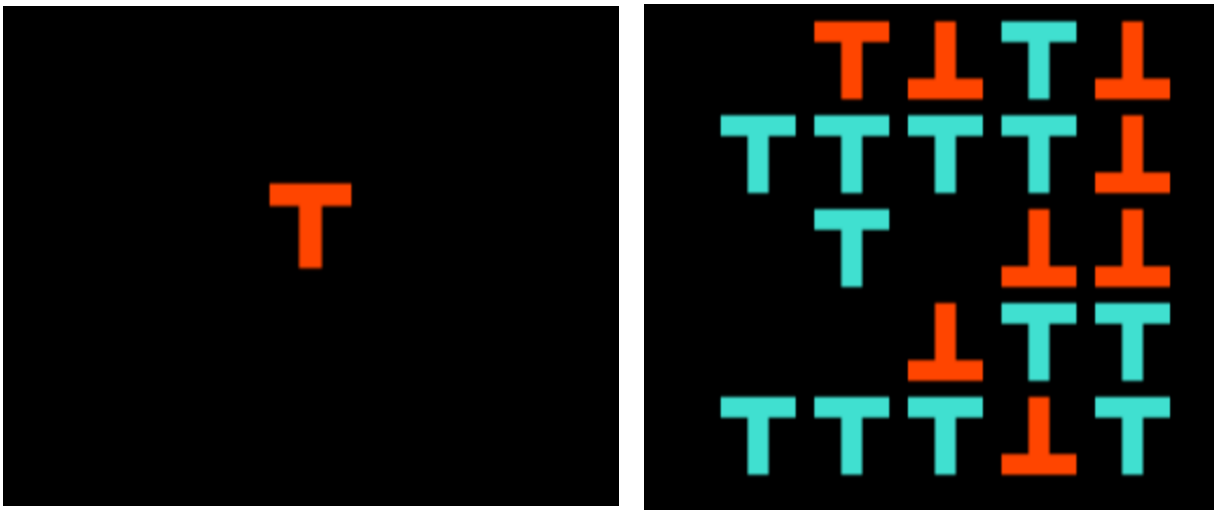


Abbildung 2. Experimente mit Schwerpunkt Informationsaufnahme. Links in einfacher und rechts in schwerer Ausführung.

Bei den zwei Experimenten, die besonders darauf abzielen sollten den Wissensabruf zu beeinflussen, mussten Rechenaufgaben gelöst werden. In beiden Settings waren es Additionsaufgaben im Zehnerbereich. Auf dem Bildschirm wurde die Aufgabe gezeigt und die Versuchspersonen mussten die korrekte Antwort auf ihrer Tastatur eingeben. In dem Setting mit hoher Belastung wurden die Zahlen in Buchstaben umgewandelt. Dies mussten die Versuchspersonen lernen. Bei einer falschen Antwort wurde die Buchstaben-Zahlen-Transformation erneut kurz auf dem Bildschirm eingeblendet (s. Abbildung 3). Vor dem Versuch wurde die Instruktion gezeigt und es gab eine Trainingsphase. Die Reaktionszeit sowie die Fehlerrate wurden erfasst. Es gab 20 Durchgänge.

Abbildung 3. Experimente mit Schwerpunkt Wissensabruf. Links in einfacher und rechts in schwerer Ausführung.

Die zwei Experimente, die besonders einen Einfluss auf die Skala Entscheidungsfindung haben sollte, waren so entworfen, dass eine Auswahl zwischen zwei hypothetischen Hotels getroffen werden musste. Dazu wurden fünf Attribute gewählt, auf die geachtet werden sollte: Zentrumsnähe, Helligkeit, Sauberkeit, Preis und die Inkludierung von Mahlzeiten. Diese Punkte wurden jeweils stichpunktartig für beide Hotel kurz ausgeführt und permanent auf dem Bildschirm eingeblendet. Bei einer Wahl für das Hotel auf der linken Seite des Bildschirms, sollte die Taste „f“ gedrückt werden. Bei der Wahl für das Hotel auf der rechten Seite, sollte die Taste „j“ gedrückt werden. Bei dem Setting mit niedriger Belastung wurden jeweils zwei der Attribute deutlich zum Vorteil eines Hotels verändert, die anderen drei Attribute waren gleich. Bei dem Setting mit hoher Belastung gab es keine objektiv richtige Antwort. Es wurde ein Attribut minimal zum Vorteil des einen Hotels verändert und ein anderes Attribut minimal zum Vorteil des anderen Hotels (s. Abbildung 4). Vor dem Versuch wurde die Instruktion gezeigt und es gab eine Trainingsphase. Die Reaktionszeit wurde erfasst. Es gab zehn Durchgänge.

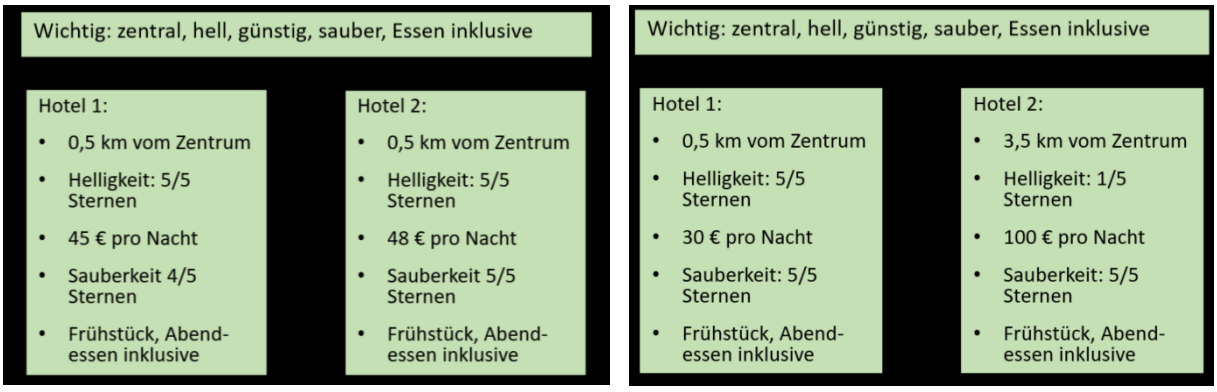


Abbildung 4. Experimente mit Schwerpunkt Entscheidungsfindung. Links in einfacher und rechts in schwerer Ausführung.

Bei den zwei Experimenten, die besonders auf zeitliche Beanspruchung abzielen sollten, mussten auftauchende Quadrate so schnell wie möglich mit dem Mauszeiger erreicht, aber nicht angeklickt werden. Für den Versuchsstart musste in der linken oberen Ecke ein kleines gelbes Quadrat angeklickt werden. Bei dem Experiment mit hoher Belastung konnten die Quadrate überall auf dem Experimentalbildschirm auftauchen. Die Versuchspersonen hatten nur 600 Millisekunden Zeit die Quadrate zu erreichen. Beim Experiment mit niedriger Belastung konnten die Quadrate nur in einem stark begrenzen Raum (siehe grünes Rechteck in Abbildung 5) auftauchen und das Quadrat verschwand nicht direkt, wie im Setting mit hoher Belastung, sobald der Mauszeiger es erreichte, sondern erst nach 500 Millisekunden. Vor dem Versuch wurde die Instruktion gezeigt und es gab eine Trainingsphase. Die Reaktionszeit wurde erfasst. Wenn die Reaktionszeit von 600 ms überschritten wurde, wurde dies als Fehler gewertet und das nächste Quadrat erschien. Es gab 30 Durchgänge.

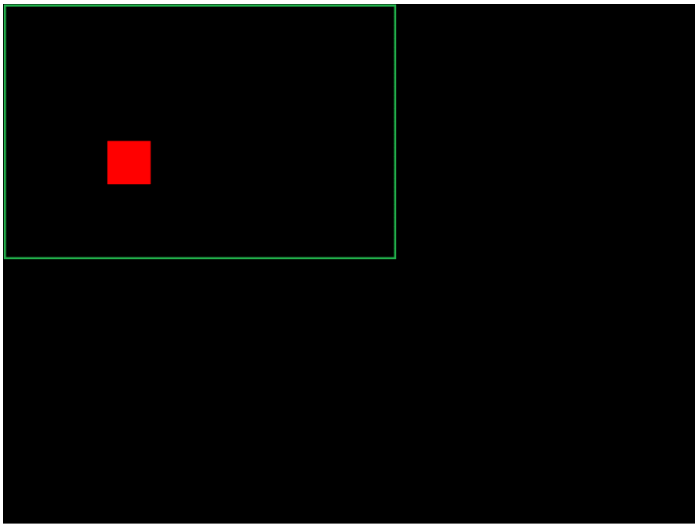


Abbildung 5. Experimente mit Schwerpunkt zeitliche Beanspruchung. Der Bereich für die einfachere Bedingung ist zur Veranschaulichung durch ein grünes Rechteck markiert.

3.3 Statistische Analyse

In diesem Abschnitt wird auf die zu untersuchenden Maße eingegangen. Es werden die Gütekriterien Objektivität, Reliabilität sowie Validität erläutert. Außerdem wird auf die Sensitivität und Spezifität sowie die konfirmatorische Faktorenanalyse eingegangen.

3.3.1 Objektivität

Die Objektivität als Gütekriterium ist wichtig, um die erforderliche Vergleichbarkeit von Untersuchungen von verschiedenen Versuchspersonen sicherzustellen. Dieses Kriterium umfasst die Durchführung, die Auswertung und die Interpretation. Ein Verfahren ist objektiv, wenn unabhängig von der testenden und auswertenden Person sowie von Ort und Zeit eine bestimmte

Versuchsperson dasselbe Ergebnis und Ergebnisinterpretation liefert, da sowohl die Testdarbietung, die Testauswertung und die Interpretationsregeln genau festgelegt sind (Moosbrugger & Kelava, 2012).

Die Durchführungsobjektivität kann als gegeben angesehen werden, da keine Interaktionen mit der testenden Person nötig sind und sowohl die Anweisungen als auch der Fragbogen selbst schriftlich vorlagen. Die Auswertungsobjektivität kann ebenfalls als gegeben angesehen werden. Das Antwortformat ist geschlossen, was die Objektivität erhöht. Es sind allerdings leichte Abweichungen bei der Auswertung möglich, da die Skala der Fragen 200 unterschiedliche Punkte aufweisen kann und die auswertenden Personen hier minimal voneinander abweichende Werte ablesen können. Die Interpretationsobjektivität kann ebenfalls als gegeben angesehen werden, da die Skalen des Fragebogens klar eingeteilt sind. Insgesamt kann das Gütekriterium der Objektivität als gegeben angesehen werden.

3.3.2 Reliabilität

Das Gütekriterium der Reliabilität ist gegeben, wenn ein Test das gewünschte Merkmal ohne Messfehler misst. Reliabilität ist also der Anteil der wahren Varianz an der Gesamtvarianz der Werte. Um die Reliabilität eines Tests zu überprüfen, wird der Reliabilitätskoeffizient berechnet, der Werte zwischen 0 und 1 annehmen kann. (Moosbrugger & Kelava, 2012). Zur Untersuchung der internen Konsistenz wurde Cronbachs Alpha genutzt, wobei die Korrelation zwischen den einzelnen Items betrachtet wird. Die Bewertung der Ergebnisse ist folgende: $\alpha > .9$ = exzellent; $\alpha > .8$ = gut; $\alpha > .7$ = akzeptabel (Blanz, 2015). Zu beachten ist allerdings, dass die interne Konsistenz lediglich Aussagen über die Korrelation der Werte trifft und keine inhaltlichen Aussagen macht. Außerdem steigt Cronbachs Alpha mit der Anzahl der Items. Eine Voraussetzung für Cronbachs Alpha ist außerdem, dass alle Items die gleiche Skala erfassen. Es ist zwar möglich, einen Gesamtwert beim DLR-WAT zu bilden, allerdings bilden die Items eigene Skalen. Dies ist zu berücksichtigen. Eine Folge vom Verstoß gegen diese Voraussetzung kann die Unterschätzung der Reliabilität sein. Es ist außerdem zu berücksichtigen, dass Cronbachs Alpha kein Maß für Eindimensionalität ist, sondern lediglich eine Inter-Item-Korrelation (Moosbrugger & Kelava, 2012). Für beide Studien erfolgte die Rechnung zwischen allen Skalen den DLR-WAT.

3.3.3 Validität

Die Validität bezieht sich auf die inhaltliche Übereinstimmung zwischen dem Ziel-Merkmal und dem Merkmal, was vom Test erfasst wird. Dies ist das wichtigste Gütekriterium, obwohl

die Objektivität und Reliabilität Voraussetzungen für die Validität sind. Ist die Validität gegeben, erlaubt dies eine Generalisierung der Testergebnisse auf Situationen außerhalb der Testsituation. Es gibt verschiedene Validitätsaspekte. Bei dieser Arbeit soll auf die Konstruktvalidität eingegangen werden, welche sich mit der theoretischen Fundierung des vom Test erfassten Merkmals beschäftigt (Moosbrugger & Kelava, 2012).

Es wurde die konvergente Validität untersucht, welche einen Unterpunkt der Konstruktvalidität darstellt. Bei konvergenter Validität geht man davon aus, dass unterschiedliche Operationalisierungen eines Konstrukts gleiche Messergebnisse hervorrufen (Moosbrugger & Kelava, 2012). Dementsprechend wurde in beiden Studien parallel zum DLR-WAT der NASA-TLX von den Versuchspersonen ausgefüllt. Dabei wurde die englische Variante benutzt, da diese besser validiert ist. Ein Überblick über die vielen Anwendungen des NASA-TLX findet sich bei Hart (2006) und eine Validitätsüberprüfung z.B. bei Rubio et al. (2004). Beim NASA-TLX wurde für die Berechnung der konvergenten Validität die Skala der Performance invertiert.

Die Leistungsparameter Übernahmezeit und die kumulierte Verspätung der ersten Studie sowie die Reaktionszeit und die Fehler der zweiten Studie wurden ebenfalls zur Überprüfung der Validität genutzt. Für die Einordnung der Korrelationswerte wurde Cohen (1988) zur Orientierung genutzt: $|r| = .10$ entspricht einem kleinen Effekt, $|r| = .30$ entspricht einem mittlerem Effekt und $|r| = .50$ entspricht einem großen Effekt.

3.3.4 Sensitivität

Sensitivität kann als Maß für die Entdeckungsleistung eines diagnostischen Instrumentes beschrieben werden. Es wird die Empfindlichkeit des Verfahrens auf Änderungen untersucht (Wirtz, 2021). Es wurde ein t-Test für abhängige Stichproben gerechnet, um zu überprüfen, ob sich der Testwert in Hinblick auf die Schwierigkeit des Experiments verändert hat. Für die erste Studie erfolgte ein t-Test für abhängige Stichproben für die Daten des DLR-WAT bei hoher und niedriger Belastung. Da es für die zweite Studie jeweils zwei Schwierigkeitsstufen pro Experiment gab, gibt es vier Paare, bei denen jeweils ein t-Test für abhängige Stichproben gerechnet wurde. Außerdem erfolgte eine deskriptive Analyse der Mittelwerte für die relevanten Variablen der t-Tests, um einen Überblick darüber zu erhalten, inwieweit Über- bzw. Unterbeanspruchung durch die Experimente induziert werden konnte. Als Voraussetzungsprüfung wurde der Shapiro-Wilk-Test gerechnet. Bei der ersten Studie sind die Daten normalverteilt. Für die zweite Studie zeigte sich, dass die Gesamtskalenwerte des DLR-WAT für das Experiment Entscheidungsfindung in der einfachen Bedingung ($p = .027$) nicht normalverteilt sind.

Die anderen Werte sind normalverteilt. Laut Stone (2010) ist der t-Test ab einer Stichprobengröße von > 30 allerdings robust gegenüber dieser Voraussetzungsverletzung, weshalb der t-Test dennoch gerechnet wurde.

3.3.5 Spezifität

Die Spezifität ist ebenfalls ein Maß für Entdeckungsleistung des diagnostischen Instrumentes und wird normalerweise in Verbindung mit der Sensitivität gesetzt (Wirtz, 2021). In dieser Arbeit wird das Maß allerdings benutzt, um die Güte der Skalenvielfalt zu überprüfen. Die Daten aus der zweiten Studie wurden für die Analyse genutzt. Pro Experiment wurden die Durchschnittswerte der Skalen des DLR-WAT über alle Versuchspersonen hinweg gebildet. Daraufhin wurde jeweils zwischen den Werten der schweren Bedingung und der leichten Bedingung die Differenz gebildet. Aufgrund des Betrages dieser Differenz wurde die Entscheidung über die Spezifität des DLR-WAT gefällt. Es sollte überprüft werden, ob der DLR-WAT spezifisch die Beanspruchung auf die Skalen deduzieren kann, auf welche die Experimente abzielten.

3.3.6 Konfirmatorische Faktorenanalyse

Anhand einer konfirmatorischen Faktorenanalyse (CFA) sollte geprüft werden, ob sich die von Grippenkoven et al. (2018) vorgeschlagene Struktur bestätigen lässt. Bei diesem Modell wird davon ausgegangen, dass die Skalen Informationsaufnahme, Wissensabruf und Entscheidungsfindung auf den Faktor mentale Beanspruchung laden. Mentale Beanspruchung, motorische und körperliche Beanspruchung, zeitliche Beanspruchung, Anstrengung, Frustration und Aufgabenbewältigung laden wiederum auf den latenten Faktor Beanspruchung. Es wurde eine schon vorhandene und theoretisch fundierte Struktur überprüft, weswegen das strukturprüfende Verfahren der konfirmatorischen Faktorenanalyse verwendet werden konnte (Moosbrugger et al., 2012).

Um eine konfirmatorische Faktorenanalyse berechnen zu können, wird eine Stichprobengröße von > 250 empfohlen (Bühner, 2011). Diese Stichprobengröße wurde nicht erreicht, weshalb die Aussagekraft stark eingeschränkt ist. Bei der konfirmatorischen Faktorenanalyse sollte zudem jeder Faktor mit mindestes drei Items erhoben werden. (Bühner, 2011). Dies ist nicht der Fall, da jeder Faktor, außer der mentalen Beanspruchung, eigenständig erhoben wird. Dies grenzt die Aussagekraft der Analyse erneut ein. Zudem sollte das Intervallskalenniveau vorliegen, was für den DLR-WAT als gegeben angesehen werden kann. Aufgrund des Skalenniveaus kann der Maximum-Likelihood-Schätzer verwendet werden (Moosbrugger et al., 2012). Hierbei werden die Parameter so geschätzt, dass die Likelihood für den Fall maximiert wird, dass

die empirische Kovarianzmatrix aus einer Population stammt, für welche die vom Modell vorgegebene Kovarianzmatrix gilt (Schermelleh-Engel et al., 2003). Der Maximum-Likelihood-Schätzer setzt voraus, dass den Daten eine Normalverteilung zugrunde liegt (Schermelleh-Engel et al., 2003). Mittels des Shapiro-Wilk-Tests bei dem ersten Datensatz wurde gezeigt, dass die Skala Frustration (p = .020) nicht normalverteilt ist. Bei den vorliegenden Daten aus der zweiten Studie wird ersichtlich, dass gemäß des Shapiro-Wilk-Tests die Daten der Skalen Entscheidungsfindung (p = .040) und zeitliche Beanspruchung (p = .041) nicht normalverteilt sind, die Daten der anderen Skalen jedoch normalverteilt sind. Eine Verletzung der Normalverteilungsannahme kann zu einer überhöhten Schätzung des Chi-Quadrat ($\chi 2$)-Wertes führen. Um dies zu vermeiden, gibt es eine Anzahl von angepassten Schätzmethoden, wovon der Maximum-Likelihood-Schätzer mit robusten Standardfehlern am häufigsten eingesetzt wird (Li, 2016 ; Satorra & Bentler, 1994).

Zur Überprüfung der Modellgüte wurden verschiedene Güte-Koeffizienten berechnet, wobei empfohlen wird, verschiede Kriterien zu betrachten, um auf Basis von mehr als nur einem Test den Modell-Fit zu beurteilen (Mueller, 1996). Für ein gutes Modell sollten alle Parameter in einem guten bis akzeptablen Wertebereich liegen (Schermelleh-Engel et al., 2003). Schermelleh-Engel et al. (2003) empfehlen die Gütekriterien: $\chi 2$ und der dazugehörige p-Wert, $\chi 2$/df, „root mean square error of approximation" (RMSEA), „standardized root mean square" (SRMR), Tucker Lewis Index (TLI), Comparative fit index (CFI) und Akaike Information Criterion (AIC). Bis auf das AIC wurden diese Kriterien in der vorliegenden Arbeit analysiert. Es folgen kurze Erklärungen zu den Kriterien.

Beim Chi-Quadrat (χ^2) -Test wird die Übereinstimmung zwischen der theoretische Kovarianzmatrix mit der empirischen Kovarianzmatrix überprüft (Moosbrugger et al., 2012). Bei einem nicht signifikanten Unterschied, wird die Annahme verfolgt, dass es keine Unterschiede zwischen den Matrizen gibt und es somit ein guter Modell-Fit ist (Schermelleh-Engel et al., 2003). Der χ^2-Wert wird allerdings von der Anzahl an Parametern (bei mehr Parametern wird der χ^2-Wert niedriger) und der Stichprobengröße (bei größerer Stichprobe wird der χ^2-Wert höher) beeinflusst, weshalb der χ^2-Test nicht als einziges Kriterium zur Bewertung der Modellgüte herangezogen werden sollte (Perry et al., 2015; Schermelleh-Engel et al., 2003).

Weitere Maße, die auf einem Vergleich der Kovarianzmatrizes basieren, sind der RMSEA und der SRMR. Der RMSEA bezeichnet die Diskrepanz des Modells zur empirischen Kovarianzmatrix pro Freiheitsgrad (Schermelleh-Engel et al., 2003). Es ist dabei abhängig von

Freiheitsgraden und der Stichprobengröße. Bei weniger Freiheitsgraden und kleinerer Stichprobe wird der RMSEA größer (Kenny et al., 2015). Für einen guten Fit sollte der RMSEA kleiner als .05 sein (Schermelleh-Engel et al., 2003). Der SRMR ist der standardisierte Mittelwert der quadrierten Diskrepanzen und sollte für einen guten Fit ebenfalls kleiner als .05 sein (Schermelleh-Engel et al., 2003).

Gütemaße, welche auf einen Modellvergleich zwischen Baseline-Modell und Zielmodell zurückgehen, sind der TLI und der CFI. Als Baseline-Modell wird dabei das Unabhängigkeitsmodell verwendet, bei dem nur die Varianzen der beobachteten Variablen geschätzt werden. Es wird betrachtet, ob sich das Zielmodell im Vergleich zum Baseline-Modell verbessert hat, wobei die Werte zwischen 0 (keine Verbesserung) und 1 (bestmögliche Verbesserung) variieren können (Schermelleh-Engel et al., 2003). In Tabelle 1 findet sich ein Überblick über die Einschätzung der Werte der Modellgütekriterien.

Tabelle 1: Bewertung der Modellgütekriterien nach Schermelleh-Engel et al. (2003)

	Gut	Akzeptabel
p-Wert von $\chi 2$	$.05 < p \leq 1.00$	$.01 \leq p \leq .05$
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 < \chi^2/df \leq 3$
RMSEA	$0 \leq \text{RMSEA} \leq .05$	$.05 < \text{RMSEA} \leq .08$
SRMR	$0 \leq \text{SRMR} \leq .05$	$.05 < \text{SRMR} \leq .10$
CFI, TLI	$.97 \leq \text{CFI/TLI} \leq 1.00$	$.95 \leq \text{CFI/TLI} < .97$

Anmerkungen. *df* = Freiheitsgrade. *p* = p-Wert. RMSEA = Root Mean Square Error of Approximation. SRMR = Standardized Root Mean Square Residual. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index.

Das AIC ist ein Modellgütekriterium der Sparsamkeit, wobei der $\chi 2$-Wert an die Anzahl der geschätzten Parameter angepasst wird. Das Ziel ist systematische und zufällige Fehler im Modell zu minimieren (Schermelleh-Engel et al., 2003). Der AIC sollte möglichst klein sein (Kaplan, 2009). Da in dieser Arbeit keine unterschiedlichen Modelle überprüft werden, wurde von der Berechnung des AIC abgesehen.

Zusätzlich zu den Gütekriterien wurde die Determinationskoeffizienten (R^2) berechnet, welche angeben, wie viel Varianz des latenten Faktors durch die jeweiligen manifesten Faktoren erklärt wird (Bortz & Schuster, 2010). R^2 kann Werte zwischen 0 und 1 annehmen, wobei Henseler et al. (2009) eine Einordnung der Werte in substanziell ($R^2 \geq .75$), moderat ($R^2 \geq .50$), und schwach ($R^2 \geq .25$) vorschlagen.

4 Ergebnisse

Im Folgenden werden die Ergebnisse aufgezeigt, die sich aufgrund der statistischen Analyse ergeben haben. Es wird auf die Reliabilität, Validität, Sensitivität, Spezifität und die Faktorenanalyse eingegangen.

4.1 Reliabilität

Für Cronbachs Alpha wurden für beide Studien jeweils die Skalenwerte des DLR-WAT über alle Sessions bzw. Experimente genutzt. Für die erste Studie beträgt Cronbachs Alpha $\alpha = .79$ und für die zweite Studie $\alpha = .88$, was gute bis exzellente Werte sind (Blanz, 2015).

4.2 Validität

Für die zweite Studie finden sich die Korrelationswerte von den einzelnen Skalen sowie des Gesamtwerts des DLR-WAT mit den einzelnen Skalen und dem Gesamtwert des NASA-TLX in der Tabelle 2.

Tabelle 2: Korrelationen der Skalen des DLR-WAT mit den Skalen des NASA-TLX für Studie eins

Skalen		r	95 % KI
DLR-WAT Informationsaufnahme	NASA-TLX Mental Demand	.58	-.03-.88
DLR-WAT Wissensabruf	NASA-TLX Mental Demand	.61	.02-.89
DLR-WAT Entscheidungsfindung	NASA-TLX Mental Demand	.59	.02-.88
DLR-WAT Motorische und körperliche Beanspruchung	NASA-TLX Physical Demand	.68	.13-.91
DLR-WAT Zeitliche Beanspruchung	NASA-TLX Temporal Demand	.94	.76-.98
DLR-WAT Anstrengung	NASA-TLX Effort	.93	.75-.98
DLR-WAT Frustration	NASA-TLX Frustration Level	.47	-.18-.83
DLR-WAT Aufgabenbewältigung	NASA-TLX Performance	.93	.73-.98
DLR-WAT Gesamt	NASA-TLX Gesamt	.81	.41-.95

Anmerkungen. *r* = Korrelation. KI = Konfidenzintervall.

Die Skala der zeitlichen Beanspruchung des DLR-WAT und die Temporal Demand Skala des NASA-TLX haben mit $r = .94$ die höchste Korrelation. Dies ist ein starker Effekt (Cohen, 1988). Danach kommen die Korrelationen der Skala Anstrengung mit der Effort-Skala ($r = .93$.) und die Skala der Aufgabenbewältigung mit der Performance-Skala ($r = .93$.) Diese Korrelationen sind laut Cohen (1988) als stark zu bewerten. Die niedrigste Korrelation hat die Frustrations-Skala mit der Frustration Level-Skala ($r = .47$) und dies ist laut Cohen (1988) ein mittelstarker Zusammenhang. Die Korrelationen von den Skalen Informationsaufnahme, Wissensabruf und Entscheidungsfindung mit Mental Demand sind als stark zu bewerten und

betragen r = .58, r = .61 bzw. r = .59. Motorische und körperliche Beanspruchung korreliert mit Physical Demand mit r = .68. Die Gesamtskalen beider Fragebogen korrelieren mit einem Wert von r = .81. Diese beiden Korrelationen sind ebenfalls als stark zu bewerten.

Für die zweite Studie finden sich die Korrelationswerte von den einzelnen Skalen sowie des Gesamtwerts des DLR-WAT mit den einzelnen Skalen und dem Gesamtwert des NASA-TLX in der Tabelle 3.

Tabelle 3: Korrelationen der Skalen des DLR-WAT mit den Skalen des NASA-TLX für Studie zwei

Skalen		*r*	95 % KI
DLR-WAT Informationsaufnahme	NASA-TLX Mental Demand	.69	.45-.84
DLR-WAT Wissensabruf	NASA-TLX Mental Demand	.43	.09-.68
DLR-WAT Entscheidungsfindung	NASA-TLX Mental Demand	.66	.41-.82
DLR-WAT Motorische und körperliche Beanspruchung	NASA-TLX Physical Demand	.71	.48-.85
DLR-WAT Zeitliche Beanspruchung	NASA-TLX Temporal Demand	.75	.54-.87
DLR-WAT Anstrengung	NASA-TLX Effort	.68	.44-.83
DLR-WAT Frustration	NASA-TLX Frustration Level	.54	.24-.75
DLR-WAT Aufgabenbewältigung	NASA-TLX Performance	.99	.87-1.00
DLR-WAT Gesamt	NASA-TLX Gesamt	.67	.43-.83

Anmerkungen. *r* = Korrelation. KI = Konfidenzintervall.

Die Aufgabenbewältigungsskala des DLR-WAT und die Performance-Skala haben mit r = .99 die höchste Korrelation, was als starker Effekt gilt. Die niedrigste Korrelation hat die Wissensabrufsskala mit der mental demand-Skala, r = .43, was als mittlerer Effekt zählt (Cohen, 1988). Die Korrelation von den Skalen Informationsaufnahme und Entscheidungsfindung mit mental demand beträgt r = .69 bzw. r = .66. Motorische und körperliche Beanspruchung korreliert mit physical demand mit r = .71 und zeitliche Beanspruchung mit der Skala des temporal demand mit r = .75. Die Anstrengungs- bzw. effort-Skala korreliert mit r = .68 zueinander. Die Frustrationsskalen korrelieren mit einem Wert von r = .54. Die Gesamtskalen beider Fragebogen korrelieren mit einem Wert von r = .67. Alle anderen Korrelationen sind somit als großer Effekt zu bewerten (Cohen, 1988).

Für beide Studien finden sich hauptsächlich starke Zusammenhänge zwischen den Skalen des DLR-WAT und des NASA-TLX. Lediglich die Korrelation zwischen zwei Skalenpaaren ist als mittelstarker Zusammenhang zu bewerten.

In der Tabelle 4 zeigen sich die Ergebnisse der Korrelationen des DLR-WAT Gesamtwertes mit den Leistungsmaßen aus der ersten Studie. Bei Aufzeichnung der kumulierten Verspätungen gab es in sechs Fällen technische Schwierigkeiten. Für diese Fälle wurde der Mittelwert der Verspätungen der anderen Versuchspersonen für diese Session genutzt. Für die Übernahmezeiten bei Tieren im Gleis (r = -.15) fällt die Korrelation schwach negativ aus. Mit der Zeit bei Langsamfahrstrecken (r = -.09) korreliert der DLR-WAT kaum. Dies bedeutet, dass mit höherer angegebener Beanspruchung weniger Zeit für die Übernahme des Zuges benötigt wurde. Die Korrelation zwischen dem DLR-WAT Gesamtwert und den Verspätungen ist schwach positiv (r = .15), dies bedeutet, dass mit höherer Beanspruchung mehr Verspätungen zusammenhingen.

Tabelle 4: Korrelation des DLR-WAT-Gesamtergebnisses mit Leistungsmaßen in Studie eins

Variable		*r*	*N*	Unteres 95 % KI	Oberes 95 % KI
DLR-WAT Gesamt	Übernahmezeit (s) bei Tieren	-.15	11	-.69	.50
	Übernahmezeit (s) bei Langsamfahrstrecke	-.09	11	-.66	.54
	Verspätungen (s)	.15	11	-.50	.69

Anmerkungen. *r* = Korrelation. *N* = Stichprobengröße. s = Sekunde.

In der Tabelle 5 zeigen sich die Ergebnisse der Korrelationen der Ergebnisse des DLR-WAT für die vier Zielskalen (Informationsaufnahme, Wissensabruf, Entscheidungsfindung, zeitliche Beanspruchung) der Experimente mit den dazugehörigen Leistungsmaßen, den Reaktionszeiten und Fehlern in Prozent. Es wurde dabei sowohl aus den DLR-WAT-Ergebnissen als auch bei den Leistungsmaßen der Mittelwert zwischen den beiden Experimenten gebildet, die besonders auf die jeweilige Skala abzielen sollten. Bei dem Experiment mit Schwerpunkt Informationsaufnahme (r = -.58) zeigt sich ein stark negativer Zusammenhang und bei Wissensabruf (r = -.20) ein schwach negativer Zusammenhang. Dies bedeutet, dass eine höhere Beanspruchung mit einer geringeren Reaktionszeit zusammenhängt. Bei der Skala Entscheidungsfindung zeigt sich kein Zusammenhang (r = -.01). Lediglich bei dem Experiment mit Schwerpunkt auf zeitlicher Beanspruchung zeigt sich eine positive Korrelation (r = .14). Dies bedeutet, dass eine höhere Beanspruchung mit einer höheren Reaktionszeit zusammenhängt. Bei dem Experiment mit Schwerpunkt Entscheidung wurden keine Fehler erhoben, aber bei den Experimenten mit den Schwerpunkten Informationsaufnahme (r = .48) und zeitliche Beanspruchung (r = .36) wird ein mittelstarker Zusammenhang zu der angegeben Beanspruchung ersichtlich. Die Korrelation

bei dem Experiment Wissensabruf und den Fehlern ist schwach positiv (r = .14). Dies bedeutet, dass eine höhere Beanspruchung mit einer höheren Fehlerquote zusammenhängt.

Tabelle 5: Korrelation der DLR-WAT-Ergebnisse pro Zielskala der Experimente mit Leistungsmaßen in Studie zwei

Experiment	Leistungsmaß	r	N	Unteres 95 % KI	Oberes 95 % KI
Info DLR-WAT	Reaktionszeit	-.58	32	-.77	-.28
	Fehler in Prozent (%)	.48	32	.16	.71
Wissen DLR-WAT	Reaktionszeit	-.20	32	-.51	.16
	Fehler in Prozent (%)	.14	32	-.22	.47
Entscheidung DLR-WAT	Reaktionszeit	-.01	32	-.36	.34
Zeit DLR-WAT	Reaktionszeit	.14	32	-.22	.47
	Fehler in Prozent (%)	.36	32	.01	.63

Anmerkungen. r = Korrelation. N = Stichprobengröße.

4.3 Sensitivität

Für die erste Studie finden sich die Ergebnisse des t-Tests für abhängige Stichproben zur Überprüfung der Sensitivität in der Tabelle 6. Es können die Signifikanzwerte für die einseitige Testung genutzt werden, da die Hypothese war, dass in den schwereren Experimenten die Beanspruchung höher ausfällt. Die Mittelwertdifferenz zwischen den DLR-WAT-Gesamtwerten zwischen den Sessions mit hoher Belastung und den Sessions mit geringer Belastung ist signifikant, $t(10) = 2.65$; $p = .012$; Hedges‘ $g = 0.74$. Dies ist laut Cohen (1988) ein mittlerer Effekt. Es zeigt sich, dass die Beanspruchung höher ist im Setting von höherer Belastung. Wie die deskriptiven Mittelwerte zeigen, liegen die Werte allerdings bei beiden Belastungsniveaus im Bereich der Unterbeanspruchung ($M_{\text{high\_demand}} = 46.71$; $M_{\text{low\_demand}} = 40.86$).

Tabelle 6: t-Test für abhängige Stichproben für die DLR-WAT-Gesamtwerte der ersten Studie

	Gepaarte Differenzen						
			95% KI				
	M	*SD*	Unterer Wert	Oberer Wert	*T*	*df*	einseitiges p
DLR-WAT gesamt für high demand - DLR-WAT gesamt für low demand	5.85	7.32	0.93	10.79	2.65	10	.012

Anmerkungen. *M* = Mittelwert. *SD* = Standardabweichung. KI = Konfidenzintervall. *t* = t-Wert. *df* = Freiheitsgrade. *p* = p-Wert.

Für die zweite Studie finden sich die deskriptiven Mittelwerte der verschiedenen Experimente in Tabelle 7. Die Ergebnisse des **Tabelle 8:** t-Test für abhängige Stichproben für die DLR-WAT-Ergebnisse der acht Experimente zur Überprüfung der Sensitivität finden sich in der Tabelle 8.

Tabelle 7: Mittelwerte des DLR-WAT für die acht Experimente aus Studie zwei

Experiment	*M*	SD
DLR-WAT Wissen schwer	114.18	17.86
DLR-WAT Wissen einfach	65.31	22.81
DLR-WAT Entscheidung schwer	96.75	27.21
DLR-WAT Entscheidung einfach	81.79	24.40
DLR-WAT Zeit schwer	88.80	26.94
DLR-WAT Zeit einfach	55.46	26.32
DLR-WAT Information schwer	89.43	24.07
DLR-WAT Information einfach	86.04	31.22

Anmerkungen. *M* = Mittelwert. *SD* = Standardabweichung.

Es lässt sich erkennen, dass die Mittelwerte für die schwierigere Bedingung höher sind als die für die einfache Bedingung. Allerdings ist lediglich bei Wissensabruf der Wert über 100, sonst sind alle Werte unter 100, sprich unter dem Wert der optimalen Beanspruchung.

Tabelle 8: t-Test für abhängige Stichproben für die DLR-WAT-Ergebnisse der acht Experimente aus Studie zwei

	Gepaarte Differenzen						
			95% KI				
	M	*SD*	Unterer Wert	Oberer Wert	*T*	*df*	einseitiges p
DLR-WAT Wissen schwer - DLR-WAT Wissen einfach	48.87	24.07	40.19	57.55	11.49	31	<.001
DLR-WAT Entscheidung schwer - DLR-WAT Entscheidung einfach	14.97	19.71	7.86	22.08	4.30	31	<.001
DLR-WAT Zeit schwer - DLR-WAT Zeit einfach	33.34	28.95	22.90	43.78	6.51	31	<.001
DLR-WAT Information schwer - DLR-WAT Information einfach	3.40	24.66	-5.49	12.29	0.78	31	.221

Anmerkungen. *M* = Mittelwert. *SD* = Standardabweichung. KI = Konfidenzintervall. *t* = t-Wert. *df* = Freiheitsgrade. *p* = p-Wert.

Es können die Signifikanzwerte für die einseitige Testung genutzt werden, da die Hypothese war, dass in den schwereren Experimenten die Beanspruchung höher ist. Für die Experimente zum Wissensabruf wurde der Mittelwertsunterschied signifikant, $t(31) = 11.49$; $p < .001$; $d = 2.03$ und ist somit laut Cohen (1988) als starker Effekt zu werten. Die Differenzen bei den Experimenten mit Schwerpunkt Entscheidungsfindung ($t(31) = 4.30$, $p < .001$, $d = 0.76$) und zeitlicher Beanspruchung ($t(31) = 6.51$, $p < .001$, $d = 1.15$) wurden ebenfalls signifikant und sind als mittlerer bzw. starker Effekt zu bewerten (Cohen, 1988). Für die Experimente mit Schwerpunkt Informationsaufnahme wurde das Ergebnis nicht signifikant, $t(31) = 3.40$, $p = .221$). Zusammenfassend bedeutet dies, dass bei den Experimenten mit Schwerpunkt Wissensabruf, Entscheidungsfindung und zeitliche Beanspruchung höhere Belastung als beanspruchender wahrgenommen wurde. Bei den Experimenten mit Schwerpunkt Informationsaufnahme ist diese Aussage nicht möglich.

4.4 Spezifität

Es wurden die Daten der zweiten Studie analysiert. Für die vier Experimentalgruppen mit den vier abgezielten Schwerpunkten folgen die Beträge der durchschnittlichen Mittelwertdifferenzen pro Skala des DLR-WAT. Die beiden Experimente, die besonders die Bewertung der Beanspruchung auf der Skala des Wissensabrufes verändern sollten, haben dies erfolgreich getan bzw. der DLR-WAT hat dies erkannt. Beim DLR-WAT wurde eine durchschnittliche Mittelwertdifferenzen auf der Skala Wissensabruf von 94.81 Punkten erreicht. Auf den Skalen

Informationsaufnahme (86.53 Punkte), Anstrengung (85.78 Punkte), Entscheidungsfindung (71.63 Punkte), zeitliche Beanspruchung (58.31 Punkte), Aufgabenbewältigung (42.69 Punkte), Frustration (37.31 Punkte) und motorische und körperliche Belastung (34.78 Punkte) fielen die durchschnittlichen Mittelwertdifferenzen jeweils geringer aus.

Bei den beiden Experimenten, die besonders die Bewertung der Beanspruchung auf der Skala der Entscheidungsfindung verändern sollten, zeigte sich auf dieser Skala auch die größte durchschnittliche Mittelwertdifferenz (40.56 Punkte). Auf den Skalen Wissensabruf (31.69 Punkte), zeitliche Beanspruchung (31.56 Punkte), Frustration (30.88 Punkte), Informationsaufnahme (30.72 Punkte), Anstrengung (28.53 Punkte), Aufgabenbewältigung (24.25 Punkte) sowie motorische und körperliche Belastung (17.19 Punkte) fielen die durchschnittlichen Mittelwertdifferenzen jeweils geringer aus.

Bei den beiden Experimenten, die besonders die Bewertung der Beanspruchung auf der Skala der zeitlichen Beanspruchung verändern sollten, zeigte sich auf dieser Skala auch die größte durchschnittliche Mittelwertdifferenz (69.75 Punkte). Auf den Skalen Anstrengung (61.66) Punkte, motorische und körperliche Belastung (61.09 Punkte), Informationsaufnahme (52.03 Punkte), Wissensabruf (39.38 Punkte), Entscheidungsfindung (38.25 Punkte), Aufgabenbewältigung (29.47 Punkte) sowie Frustration (28.34 Punkte) fielen die durchschnittlichen Mittelwertdifferenzen jeweils geringer aus.

Bei den beiden Experimenten, die besonders die Bewertung der Beanspruchung auf der Skala der Informationsaufnahme verändern sollten, zeigte sich die größte durchschnittliche Mittelwertdifferenz auf der Skala Wissensabruf (33.88 Punkte). Danach hatten die Skalen Entscheidungsfindung (31.34 Punkte) und Informationsaufnahme (30.88 Punkte) die größten Mittelwertdifferenzen. Auf den Skalen zeitliche Beanspruchung (26.19 Punkte), Anstrengung (25.44 Punkte), motorische und körperliche Belastung (21.03 Punkte), Aufgabenbewältigung (20.41 Punkte) und Frustration (15.44 Punkte) fielen die durchschnittlichen Mittelwertdifferenzen jeweils geringer aus.

Bei den Experimenten, die auf einen Beanspruchungsunterschied bei der zeitlichen Beanspruchung, bei der Entscheidungsfindung und bei dem Wissensabruf abzielen sollten, ist die Beurteilung durch den DLR-WAT dort auch am stärksten zwischen den unterschiedlichen Belastungsinduktionen verändert worden. Bei den Experimenten mit Schwerpunkt Informationsaufnahme war die Beurteilung beim DLR-WAT nicht besonders spezifisch auf diese Skala ausgerichtet.

4.5 Konfirmatorische Faktorenanalyse

Bei der konfirmatorischen Faktorenanalyse für die erste Studie wurden folgende Werte berechnet: χ^2 (19, N = 11) = 60.91, p < .001); χ^2/df = 3.21; RMSEA = .448; SRMR = .122; CFI = .555; TLI = .344. Der χ^2-Wert ist signifikant geworden, was eine signifikante Abweichung der empirischen von den theoretischen Modelldaten bedeutet. Keiner der Werte erreicht den akzeptablen Bereich für einen Modell-Fit (Schermelleh-Engel et al., 2003). Laut den Wertebereichen von Henseler et al. (2009) für die Determinationskoeffizienten ergaben sich substanzielle Werte für die Skalen Wissensabruf (R^2 = .94), zeitliche Beanspruchung (R^2 = .89) und Entscheidungsfindung (R^2 = .82). Ein moderater Determinationskoeffizienten ergab sich bei der Skala Informationsverarbeitung (R^2 = .58) und schwache Werte traten bei den Skalen motorische und körperliche Beanspruchung (R^2 = .38) und Anstrengung (R^2 = .34) auf. Kaum Varianz aufklären konnten die Skalen Aufgabenbewältigung (R^2 = .23) und Frustration (R^2 = .01). Für den Faktor mentale Beanspruchung konnten aufgrund von negativen Varianzen kein Wert berechnet werden.

Bei der konfirmatorischen Faktorenanalyse für die zweite Studie wurden folgende Werte berechnet: χ^2 (19, N = 32) = 35.64, p = .012); χ^2/df = 1.88; RMSEA = .165; SRMR = .078; CFI = .904; TLI = .859. Der χ^2-Wert ist signifikant geworden, was eine signifikante Abweichung der empirischen von den theoretischen Modelldaten bedeutet, der Wert liegt allerdings noch im akzeptablen Bereich. Der Quotient zwischen dem χ^2-Wert und den Freiheitsgraden liegt im guten Bereich für die Modellgüte. Der Wert vom SRMR ist als akzeptabel zu bewerten, allerdings können die Werte vom RMSEA, CFI und TLI nicht mehr als akzeptabel gewertet werden (Schermelleh-Engel et al., 2003). Laut CFI und TLI liefert das vorgeschlagene Faktorenmodell keinen signifikant besseren Fit als das Baseline-Modell.

Für die Determinationskoeffizienten ergaben sich schwache Werte für die Skalen Frustration (R^2 = .26) und körperliche und motorische Beanspruchung (R^2 = .47). Moderate Werte ergaben sich für die Skalen Informationsverarbeitung (R^2 = .69), Wissensabruf (R^2 = .62) und zeitliche Beanspruchung (R^2 = .55). Substanzielle Werte wurden für die Skala Entscheidungsfindung (R^2 = .77) und den latenten Faktor mentale Beanspruchung (R^2 = .91) berechnet. Die Aufgabenbewältigung konnte kaum Varianz aufklären (R^2 = .01) und für die Skala der Anstrengung war es aufgrund negativer Varianzen nicht möglich einen Wert zu errechnen.

Die von Schermelleh-Engel et al. (2003) geforderten Werte im akzeptablen bis guten Bereich über alle Gütekriterien konnten nicht erreicht werden. Dementsprechend kann aufgrund der

Faktorenanalysen bei beiden Studien die Aussage nicht getroffen werden, dass die postulierte Faktorenstruktur ein gutes Modell wäre.

5 Diskussion

Diese Arbeit hatte die Validierung des Fragebogens DLR-WAT zum Ziel. Im Folgenden werden die Ergebnisse zusammengefasst und interpretiert. Außerdem werden Grenzen dieser Studie aufgezeigt und es wird ein Ausblick auf die Zukunft gegeben.

5.1 Einordnung der Ergebnisse

Die Reliabilität ist mit einem Cronbachs Alpha von gerundet $\alpha = .8$ bzw. $\alpha = .9$ bei der zweiten Studie als gut bis exzellent zu bewerten (Blanz, 2015). Der DLR-WAT kann also als reliables Messinstrument gelten. Beim NASA-TLX wurden mit z.B. $\alpha = .83$ (Malekpour et al., 2014) oder $\alpha = .90$ (Mohammadi et al., 2013) ähnlich hohe Werte für die Reliabilität gefunden.

Bei der Überprüfung des DLR-WAT auf Sensitivität ist bei der ersten Studie das Ergebnis signifikant geworden, bei dieser Studie hat der DLR-WAT also empfindlich auf die Veränderung der Belastung reagiert. Bei der zweiten Studie wurden nur drei der vier t-Tests signifikant. Lediglich bei den Experimenten zum Wissensabruf, Entscheidungsfindung und zur zeitlichen Beanspruchung gab es einen signifikanten Unterschied zwischen den Schwierigkeiten. Hier wurde das Experiment, welches als beanspruchender wahrgenommen werden sollte, auch als beanspruchender bewertet. Bei den Experimenten zur Informationsaufnahme konnte dies nicht statistisch nachgewiesen werden. Ein möglicher Grund ist die fehlende sensitive Erfassung der Beanspruchung mit dem DLR-WAT auf der Skala der Informationsaufnahme. Eine alternative Begründung hierfür ist allerdings die fehlende theoretische Fundierung für die Programmierung der Experimente, da beide Experimente in der Bewertung der Versuchspersonen unter der optimalen Beanspruchung lagen. Es ist außerdem zu beachten, dass lediglich bei dem Experiment zum Wissensabruf (Studie zwei) die schwere Variante eine Beanspruchungsbewertung von den Versuchspersonen auf dem DLR-WAT über dem Optimum (100) erhielt. Bei den anderen waren es immer Werte unter 100 und befanden sich so im Bereich der Unterbeanspruchung. Dementsprechend waren die schwierigen Aufgaben in ihrer Bewertung näher an der optimalen Beanspruchung. Zusammenfassend kann gesagt werden, dass der DLR-WAT die Beanspruchung sensitiv erfasst.

Für die konvergente Validität wurde für beide Studien dieser Arbeit der Zusammenhang zwischen DLR-WAT mit dem NASA-TLX und den erhobenen Leistungsmaßen berechnet. Im

Hinblick auf die konvergente Validität des DLR-WAT in Bezug auf den NASA-TLX zeigte sich der größte Zusammenhang zwischen den Skalen Aufgabenbewältigung und Performance (r = .93, r = .99) sowie zwischen zeitlicher Beanspruchung und Temporal Demand (r = .94, r = .75) bei beiden Studien. Bei der ersten Studie ist außerdem die Skala der Anstrengung hoch korreliert mit der Skala des Efforts (r = .93). Die niedrigste Korrelation in der ersten Studie zeigt sich zwischen den Skalen Frustration und Frustration Level (r = .47) und bei der zweiten Studie zwischen Wissensabruf und Mental Demand (r = .43). Im Gegensatz zu allen anderen Korrelationen, die stark sind, sind dies die einzig mittelstarken (Cohen, 1988). Die Gesamtskalen korrelieren bei der ersten Studie mit r = .88 und bei der zweiten Studie mit r = .67. Hier werden Unterscheide zwischen den Studien in der Höhe der Korrelationen ersichtlich, allerdings sind beides starke Zusammenhänge (Cohen, 1988). Zusammenfassend kann gesagt werden, dass bei beiden Studien jede Korrelation hinsichtlich des Zusammenhangs zwischen DLR-WAT und NASA-TLX mittelstark bis stark positiv ausfällt. Somit kann die Annahme verfolgt werden, dass der DLR-WAT und der NASA-TLX das gleiche Konstrukt, die Beanspruchung, erfassen. Aus der Literatur lassen sich für den NASA-TLX starke Korrelationswerte von r = .97 und r = .98 bei der Korrelation mit der Subjective Workload Assessment Technique (SWAT) und dem Workload Profile ermitteln, welche ebenfalls Fragebogen zur Erfassung von Beanspruchung sind (Rubio et al., 2004). In der Studie von Xiao et al., (2005) zeigte sich eine mittelstarke Korrelation von r = .49 zwischen dem NASA-TLX und dem SWAT. Gefundene Korrelationen mit der Cooper-Harper-Scale, einem Fragebogen für Beanspruchung, welcher ursprünglich für Piloten entwickelt wurde, belaufen sich auf r = .90, r = .65 und r = .82 basierend auf den Kategorien von hoher, mittlerer und niedriger Performance der Versuchspersonen (Mansikka et al., 2019). So stimmen die hohen gefundenen Korrelationen in dieser Arbeit mit berichteten Korrelationen aus vorherigen Studien überein. Hinsichtlich der Leistungsmaße zeigte sich, dass sowohl bei Studie eins, als auch bei Studie zwei bei höherer angegebener Beanspruchung beim DLR-WAT meist die Übernahmezeit bzw. die Reaktionszeit sinkt. Zudem ist bei der ersten Studie der Zusammenhang zwischen der Beanspruchung und der Übernahmezeit bei Langsamfahrstrecken marginal. Lediglich bei der zweiten Studie bei den Experimenten mit dem Schwerpunkt zeitlicher Beanspruchung gibt es einen schwach negativen Zusammenhang, hier steigt die Reaktionszeit mit steigender Beanspruchung. Dies kann daran liegen, dass erst mit steigender Angabe von Beanspruchung ein besseres Niveau der Beanspruchung erreicht wurde und so die Reaktionszeit sank. Dagegen spricht jedoch, dass in Hinblick auf die zweite Studie bei allen Experimenten die Fehlerhäufigkeit mit steigender Beanspruchung steigt. Eine

Möglichkeit hierfür ist, dass die Beanspruchung in diesen Experimenten so gering war, dass die Versuchspersonen sich sehr beeilt haben und es so zu Flüchtigkeitsfehlern gekommen ist. Da der DLR-WAT erst im Nachhinein beantwortet wurde, kann es sein, dass die gemachten Fehler hier in die Bewertung mit eingeflossen sind. In der ersten Studie zeigte sich zudem eine Erhöhung der kumulierten Verspätungen der zu kontrollierenden Züge bei steigender Beanspruchung. Eine Erklärung dafür ist die Gestaltung des Experiments. In den Sessions mit der höheren Belastung musste häufiger auf Störungen reagiert werden. Dementsprechend ist es möglich, dass die Versuchspersonen eher mit aufkommenden Warnungen des Systems rechnen und dementsprechend schneller ihrer Kontrollfunktion nachkommen können. Ebenso bedeutet dieses Setting, dass unweigerlich durch häufigere Störungen auf der Strecke, für welche die Instruktion besagte, dass eine Geschwindigkeitsverringerung nötig sei, eine Erhöhung der Verspätung einhergeht. Für den NASA-TLX findet sich z.B. eine starke negative Korrelation von r = -.5 mit Leistungsmaßen bei einer chirurgischen Operation (Yurko et al., 2010). Hier zeigt sich, wie bei der zweiten vorliegenden Studie dieser Arbeit, dass bei höherer Beanspruchung die Leistung sinkt, bzw. die Fehler ansteigen. Weitere Korrelationen für die Validität sind z.B. r = -.15 mit Blinzelhäufigkeit und r = .45 mit Blinkdauer (Zheng et al., 2012). Somit wird die Annahme verfolgt, dass der DLR-WAT valide die Beanspruchung erfasst.

Aufgrund der deskriptiven Mittelwertdifferenzen wurde ersichtlich, dass bei den drei Experimentengruppen mit Schwerpunkt zeitlicher Beanspruchung, Entscheidungsfindung und Wissensabruf der DLR-WAT auf den jeweiligen Skalen die größten Ausschläge in der Bewertung der Beanspruchungsveränderung hatte. Hinsichtlich der Belastung, welche besonders auf der Skala Wissensabruf Beanspruchung induzieren sollte, hat dies nicht funktioniert. Dies kann daran liegen, dass der DLR-WAT auf dieser Skala nicht spezifisch die Beanspruchung erheben kann. Ein weiterer möglicher Grund ist, dass das Experiment nicht spezifisch genug diese Belastung ausgelöst hat. In der zweiten Studie dieser Arbeit hat der DLR-WAT Beanspruchung spezifisch auf drei von vier Skalen erkannt.

Die konfirmatorische Faktorenanalyse hat für die zweite Studie uneindeutige Ergebnisse geliefert. Laut dem Gütekriterium χ^2/df wurde ein guter Modellfit aufgezeigt. Der p-Wert des χ^2-Tests und der SRMR sprechen für eine akzeptable Passung. Jedoch wurde laut RMSEA, CFI sowie TLI selbst eine akzeptable Modellpassung verfehlt. In Hinblick auf die Determinationskoeffizienten klärt der Faktor mentale Beanspruchung die größte Varianz auf und die Skala Aufgabenbewältigung die kleinste. Auf Basis dessen könnte diese Skala dementsprechend aus dem Fragebogen entfernt werden, da sie nicht viel zur Varianzaufklärung beiträgt. Bei der CFA

für die erste Studie konnte kein akzeptabler Wert bei den Gütekriterien erreicht werden. Mit Blick auf die Determinationskoeffizienten klärt bei diesem Modell die Skala Wissensabruf auf meisten Varianz am latenten Faktor, der Beanspruchung, auf und die Skala Frustration die geringste. Die postulierte Faktorenstruktur konnte mit keiner der beiden Studien bestätigt werden. Ein Problem ist jedoch die zu geringe Stichprobengröße, welche die nötige Größe für eine CFA verfehlt. Aufgrund dessen ist es möglich, dass der Modellfit schlechter ausgefallen ist. Im Rahmen dieser Arbeit wurde darauf verzichtet, mögliche Verbesserungen des Modells zu untersuchen. In der Literatur finden sich ähnliche Werte für die konfirmatorische Faktorenanalyse des NASA-TLX (Hernandez et al., 2022). Dies zeigt die Notwendigkeit, die Faktorenstruktur in Zukunft noch besser zu untersuchen.

Abschließend kann die Forschungsfrage dieser Arbeit positiv beantwortet werden, der DLR-WAT eignet sich zur Erfassung von Beanspruchung, da er laut Analyse ein reliables und valides Instrument ist. Außerdem kann der DLR-WAT als objektiver Fragebogen beschrieben werden. Der DLR-WAT ist zudem sensitiv hinsichtlich unterschiedlicher Belastungen und kann Belastungen spezifisch einordnen.

5.2 Grenzen der Studie

Die individuelle Beanspruchbarkeit von Personen kann stark variieren. Dies liegt nicht nur an intraindividuelle Faktoren, sondern auch an externen, wie z.B. der Tageszeit und an Schlafmangel (Elmenhorst et al., 2018). Da die zweite Studie online durchgeführt wurde, konnten so einige mögliche Störvariablen nicht kontrolliert werden. Ein besseres Vorgehen ist hierbei eine strengere Überprüfung wie in der ersten Studie, was jedoch invasives Eingreifen, wie z.B. bei der Überprüfung des Schlafes, erfordert.

Es ist möglich, dass in der zweiten Studie das gewünschte Ziel der Experimente nicht erreicht wurde. So zeigte es sich, dass nur in einem von vier Experimenten mit abgezielter hoher Belastung dieses von den Versuchspersonen als höhere Beanspruchung als das Optimum bewertet wurde. Ein möglicher Grund hierfür ist die kurze Durchführungsdauer der einzelnen Experimente, die etwa nur zwei bis drei Minuten umfassten. Die Zeitspanne wurde gewählt, um die Motivation und Anstrengungsbereitschaft der Versuchspersonen zu erhalten. Es ist allerdings möglich, dass die Versuchspersonen die Aufgaben in diesen kurzen Zeiträumen nicht als besonders beanspruchend wahrgenommen haben, da sie es aus ihrem Alltag gewöhnt sind, an deutlich langfristigeren Aufgaben zu arbeiten und hier über einen längeren Zeitraum ein gutes Beanspruchungsniveau finden müssen.

Die Instruktion bei dem Experiment der Entscheidungsfindung der zweiten Studie muss deutlicher formuliert werden. Aufgrund von Rückmeldungen der Versuchspersonen fiel auf, dass einige eine hierarchische Reihenfolge der zu beachtenden Attribute für die Entscheidung angenommen haben. Dementsprechend wurde die Entscheidungsschwierigkeit und somit die Belastung und Beanspruchung umgangen. Ein weiteres Problem bei diesem Experiment war es, dass es als hohe Belastung induziert werden sollte, es laut Versuchsplan jedoch keine richtige Lösung gab. Hier wäre eine praktisch orientierte Aufgabe sinnvoll, bei der eine Entscheidung der Versuchspersonen praktische Konsequenzen nach sich zieht.

Der NASA-TLX war auf Englisch, sodass es möglich ist, dass alleine aufgrund der Sprachunterschiede unterschiedliche Ergebnisse im DLR-WAT im Vergleich zum NASA-TLX erzielt wurden. Dies sollte in zukünftigen Forschungen berücksichtigt werden, sodass z.B. die deutsche Variante des NASA-TLX zum Vergleich genutzt wird.

Im Hinblick auf das Experiment, welches besonders auf Wissensabruf abzielen sollte, ist fraglich, ob die Rahmenbedingen dafür gesorgt haben, dass die Versuchspersonen Wissen abrufen mussten, da lediglich Transformationen von Zahlen in Buchstaben angewandt werden mussten. Laut Rasmussen (1983) zeigt sich wissensbasiertes Verhalten in unbekannten Situationen, wenn keine Regeln zur Verfügung stehen. Auf diese theoretische Definition sollte in zukünftigen Studien zurückgegriffen werden, um zu überprüfen, wie gut der DLR-WAT die Beanspruchung in diesem Bereich erfassen kann. Generell war die theoretische Fundierung der Experimente in der zweiten Studie eingeschränkt. Um ein verbessertes Verständnis darüber zu erlangen, wie bestimmte Beanspruchungsanteile schwerpunktartig durch Experimente ausgelöst werden, könnten z.B. Fokusgruppen oder Experteninterviews durchgeführt werden. Hier ist weitere Forschung mit anderen Experimenten, welche möglichst nur auf eine der Skalen des DLR-WAT laden, nötig, um über das Kriterium der Spezifität Klarheit zu schaffen.

Zur weiteren Überprüfung der Sensitivität der Skalen sollten weitere Studien durchgeführt werden. In den zwei vorliegenden Studien konnte meist nur eine Unterbeanspruchung induziert werden. Mit weiteren Experimenten sollte dementsprechend versucht werden, Überbeanspruchung zu induzieren, um auch speziell in diesem Bereich eine gute Sensitivität des DLR-WAT überprüfen zu können.

Außerdem sollte in zukünftigen Studien die Anwendbarkeit des subjektiv optimalen Beanspruchungsniveaus überprüft werden. Im Umfang der vorliegenden Arbeit war es nicht möglich, die Belastung individuell auf die Versuchspersonen anzupassen, um so ein optimales

Beanspruchungsniveau bei den Aufgaben zu erreichen. Des Weiteren sollten die Nebengütekriterien (Kubinger & Jäger, 2003) in zukünftigen Studien untersucht werden, um die Validierung des DLR-WAT zu vervollständigen.

5.3 Zusammenfassung

Durch die Veränderung des Arbeitsfeldes verändern sich die Anforderungen und die Aufgaben der Menschen. Besonders im Verkehrssektor ist die Hochautomatisierung ein zentrales Thema. Dadurch kann es beim Menschen zu Unter- oder Überforderung kommen. Da sich die bisherigen Instrumente nur bedingt eignen, um ein optimales Beanspruchungsniveau für die Individuen zu finden, wurde der DLR-WAT entwickelt. Mit diesem Instrument soll zwischen unterschiedlichen Aufteilungen der Arbeit zwischen Mensch und Automation das optimale Setting hinsichtlich der Beanspruchung des Menschen gefunden werden (Grippenkoven et al., 2018).

Es wurde gezeigt, dass der DLR-WAT ein objektives, reliables, valides, sensitives und spezifisches Instrument für die Erfassung von Beanspruchung ist. Lediglich die Frage der Faktorenstruktur bleibt offen und sollte an größeren Stichproben getestet werden. Insgesamt ist es wünschenswert, dass sich die Nutzung des DLR-WAT ausbreitet, um mehr Erkenntnisse über die Beanspruchungsniveaus der Menschen zu erhalten und um bei Über- oder Unterbeanspruchung intervenieren zu können. Ein großes Anwendungsfeld stellt der Verkehrssektor dar, aber Menschen können auch in anderen Kontexten von dem DLR-WAT profitieren. Hier spielt, wie auch Evers (2009) schon betont hat, der Präventionsgedanke eine wichtige Rolle. Wenn die Menschen ein optimales Beanspruchungsniveau erreichen und aufrechterhalten können, können Unfälle vermieden und frühzeitig sicherheits- und gesundheitsfördernde Maßnahmen implementiert werden.

6 Literaturverzeichnis

Antoni, C., & Bungard, W. (1989). Beanspruchung und Belastung. In Erwin Roth (Hrsg.), *Enzyklopädie der Psychologie: Themenbereich D Praxisgebiete, Serie III Wirtschafts-, Organisations- und Arbeitspsychologie, Band 3 Organisationspsychologie* (S.431-458). Hogrefe.

Benderoth, S. (unveröffentlichter Bericht). Remote-Arbeitsplatz für Triebfahrzeugführer.

Bläsing, D. (2020). Mentale Beanspruchung in der Montage. In M. Bornewasser, S. Hinrichsen (Hrsg.), *Informatorische Assistenzsysteme in der variantenreichen Montage* (S. 65-87). Springer Vieweg.

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

Blanz, M. (2015). *Forschungsmethoden und Statistik für die Soziale Arbeit: Grundlagen und Anwendungen*. Kohlhammer.

Bortz, J. & Schuster, C. (2010a). Faktorenanalyse. In J. Bortz & C. Schuster (Hrsg.), *Springer-Lehrbuch. Statistik für Human- und Sozialwissenschaftler* (7. Aufl., S. 385–433). Springer. doi.org/10.1007/978-3-642-12770-0_23

Brandenburger, N., & Jipp, M. (2017). *Effects of expertise for automatic train operations. Cognition, Technology & Work, 1*(6), 59.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. Pearson.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., & Conway, D. (2016). *Robust multimodal cognitive load measurement* (S. 13-32). Springer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Auflage). Hillsdale.

de Waard, D. (1996). *The measurement of drivers' mental workload* [Dissertation, University of Groningen]. http://usd-apps.usd.edu/coglab/schieber/pdf/deWaard-Thesis.pdf

DIN EN ISO 10075-1:2018-01, Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung – Teil 1: Allgemeine Konzepte und Begriffe (ISO 10075-1:2017): Deutsche Fassung EN ISO 10075-1:2017. Beuth.

DIN EN ISO 9241-2:1992-06, Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten (VDTs); Teil 2: Leitsätze zur Aufgabegestaltung. Beuth.

Dunn, N., & Williamson, A. (2012). Driving monotonous routes in a train simulator: the effect of task demand on driving performance and subjective experience. *Ergonomics, 55*(9), 997-1008.

Eggemeier. F. T. (1988). Properties of workload assessment techniques. In P. A. Hancock and N. Meshkati (Hrsg.), *Human mental workload* (S. 41-62). Elsevier.

Elmenhorst, E. M., Elmenhorst, D., Benderoth, S., Kroll, T., Bauer, A., & Aeschbach, D. (2018). Cognitive impairments by alcohol and sleep deprivation indicate trait characteristics and a potential role for adenosine A1 receptors. *Proceedings of the National Academy of Sciences, 115*(31), 8009-8014.

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

Evers, C. (2009). A*uswirkungen von Belastungen und Stress auf das Verkehrsverhalten von Lkw-Fahrern* [Dissertation, Universität Bonn]. https://bonndoc.ulb.uni-bonn.de/xmlui/bitstream/handle/20.500.11811/3976/1843.pdf?sequence=1&isAllowed=y

Frey, D., Hoyos, C. G., & Stahlberg, D. (1992). *Angewandte Psychologie: Ein Lehrbuch*. Psychologie Verlags Union.

Georgiou, K., Larentzakis, A. V., Khamis, N. N., Alsuhaibani, G. I., Alaska, Y. A., & Giallafos, E. J. (2018). Can wearable devices accurately measure heart rate variability? *A systematic review. Folia medica, 60*(1), 7-20.

Ghanbary Sartang, A., Ashnagar, M., Habibi, E., & Sadeghi, S. (2016). Evaluation of Rating Scale Mental Effort (RSME) effectiveness for mental workload assessment in nurses. *Journal of Occupational Health and Epidemiology, 5*(4), 211-217.

Greil, H., Voigt, A. & Scheffler, C. (2008). *Optimierung der ergonomischen Eigenschaften von Produkten für ältere Arbeitnehmerinnen und Arbeitnehmer – Anthropometrie*. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.

Grippenkoven, J., Rodd, J., & Brandenburger, N. (2018). DLR-WAT: Ein Instrument zur Untersuchung des optimalen Beanspruchungsniveaus in hochautomatisierten Mensch-Maschine-Systemen. *AAET Automatisiertes & Vernetztes Fahren*, 199-213.

Groth, K., Riecker, A., & Steinbrink, C. (2011). Zeitverarbeitung deutscher Vokale bei Lese-Rechtschreib-Störung: Verhaltens-und fMRT--Experimente. In A. Heine, A. M. Jacobs (Hrsg.), *Lehr-Lern-Forschung unter neurowissenschaftlicher Perspektive* (S. 27-37). Waxmann.

Hancock, P. A., & Warm, J. S. (2003). *A dynamic model of stress and sustained attention. Journal of Human Performance in Extreme Environments, 7*(1), 4.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, S. 904-908). Sage publications.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (S. 139-183). North Holland Press.

Henseler, J., Ringle, C., and Sinkovics, R. (2009). The use of partial least squares path modeling in international marketing. In *Advances in International Marketing* (Vol. 20, S. 277-320). Emerald Publishing Limited. doi.org/10.1108/S1474-7979(2009)0000020014.

Hernandez, R., Roll, S. C., Jin, H., Schneider, S., & Pyatak, E. A. (2022). Validation of the National Aeronautics and Space Administration Task Load Index (NASA-TLX) Adapted for the whole day repeated measures context. *Ergonomics, 65*(7), 960-975.

Johannsen, G. (2013). *Mensch-Maschine-Systeme*. Springer.

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin Books.

Kaplan, D. (Hrsg.). (2009). *Structural equation modeling: Foundations and extensions* (2. Aufl.). SAGE. doi.org/10.4135/9781452226576

Kauffeld, S. (2019). *Arbeits-, Organisations-und Personalpsychologie für Bachelor*. Springer.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological methods & research, 44(*3), 486-507.

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing, 17*(2), 79-103.

Kubinger, K. D., & Jäger, R. S. (2003). *Schlüsselbegriffe der Psychologischen Diagnostik*. Beltz PVU.

Latin Square Generator (o. D.). https://hamsterandwheel.com/grids/index2d.php

IBM Corp. (2021). IBM SPSS Statistics for Windows (Version 28.0.) [Software]. IBM Corp.

Malekpour, F., Mehran, G., Mohammadian, Y., Mirzaee, V., & Malekpour, A. (2014). Assessment of mental workload in teachers of Hashtrud city using NASA-TLX mental workload index. *Pajoohandeh Journal, 19*(3), 157-161.

Mansikka, H., Virtanen, K., & Harris, D. (2019). Comparison of NASA-TLX scale, modified Cooper–Harper scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. *Ergonomics, 62*(2), 246-254.

Manzey D. (1997) Psychophysiologie mentaler Beanspruchung. In F. Rösler (Hrsg.): Ergebnisse und Anwendungen der Psychophysiologie. *Enzyklopädie der Psychologie* (S. 799-864). Hogrefe.

Marquart, G., & De Winter, J. (2015). Workload assessment for mental arithmetic tasks using the task-evoked pupillary response. *PeerJ Computer Science*, *1*, Artikel e16.

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition, 1*(1).

Mattsson, S., & Fast-Berglund, Å. (2016). How to support intuition in complex assembly?. *Procedia Cirp, 50*, 624-628.

Mehta, R. K., & Parasuraman, R. (2013). Neuroergonomics: a review of applications to physical and cognitive work. *Frontiers in human neuroscience, 7*, Artikel 889.

Mohammadi, M., Mazloumi, A., & Zeraati, H. (2013). Designing questionnaire of assessing mental workload and determine its validity and reliability among ICUs nurses in one of the TUMS's hospitals. *Journal of School of Public Health and Institute of Public Health Research, 11(*2), 87-96.

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (Vol. 2). Springer.

Mueller, R. O. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS. Springer texts in statistics*. Springer.

NASA TLX Paper and Pencil Version (2022, 15. Juni). *TLX @ NASA AMES – Home*. https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf

Nohl, J. (1989). Verfahren zur Sicherheitsanalyse - Eine prospektive Methode zur Analyse und Bewertung von Gefährdungen. Springer.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39*(2), 230–253. doi:10.1518/001872097778543886

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. doi:10.1109/3468.844354

Perry, J. L., Nicholls, A. R., Clough, P. J. & Crust, L. (2015). Assessing Model Fit: Caveats and Recommendations for Confirmatory Factor Analysis and Exploratory Structural Equation Modeling. *Measurement in Physical Education and Exercise Science, 19*(1), 12–21. doi.org/10.1080/1091367X.2014.952370

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological psychology, 40*(1-2), 187-195.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. Behavior *Research Methods, 48*(3), 936–949.

R Core Team. (2020). R: A language and environment for statistical computing. (Version 1.4.1103) [Software]. R Foundation for Statistical Computing.

Radlmayr, J., Körber, M., Feldhütter, A., & Bengler, K. (2016). 3 Methoden und Fahrermodelle für Hochautomatisiertes Fahren. In *Haus der Technik*. Expert Verlag GmbH.

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.

Richter, G. (2000). *Psychische Belastung und Beanspruchung. Streß, psychische Ermüdung, Monotonie, psychische Sättigung*. Schriftenreihe der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Heft Fa 36. Wirtschaftsverlag NW.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied psychology, 53*(1), 61-86.

Sammito, S., Thielmann, B., Seibt, R., Klussmann, A., Weippert, M., & Böckelmann, I. (2015). *Guideline for the application of heart rate and heart rate variability in occupational medicine and occupational science*. ASU Int 2015(06).

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Hrsg.), *Latent variables analysis: Applications for developmental research* (S. 399–419). Sage Publications, Inc.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *2003*(8), Artikel 2, 23–74.

Schlick, C., Bruder, R., & Luczak, H. (2018). *Arbeitswissenschaft*. Springer.

Sedlmeier, P., & Renkewitz, F. (2013). *Forschungsmethoden und Statistik – Ein Lehrbuch für Psychologen und Sozialwissenschaftler* (2. aktualisierte und erweiterte Aufl.) Hallbergmoos: Pearson.

Shahid, A., Wilkinson, K., Marcu, S. & Shapiro, C.M. (2011). Karolinska Sleepiness Scale (KSS). In Shahid, A., Wilkinson, K., Marcu, S., Shapiro, C. (Hrsg.), *STOP, THAT and One Hundred Other Sleep Scales* (S. 209-210). Springer.

Sperandio, J. C. (1971). Variation of operator's strategies and regulating effects on workload. *Ergonomics, 14*(5), 571-577.

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096-1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24-31.

Stone, E. R. (2010). t Test, Paired Samples. In N. J. Salkind (Hrsg.), *Encyclopedia of research design* (S. 1560–1565). Los Angeles: SAGE.

Warm, J. S., Matthews, G., & Finomore, V. S. (2008). Vigilance, Workload, and Stress. In P. A. Hancock & J. L. Szalma (Hrsg.), *Performance under Stress* (S. 115–141). Ashgate Publishing Limited.

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors, 50*(3), 449-455.

Wickens, C. D., & Carswell, C. M. (2012). Information Processing. In G. Salvendy (Hrsg.), *Handbook of human factors and ergonomics* (4. Auflage, S. 117-177). John Wiley & Sons.

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2021). *Engineering psychology and human performance*. Routledge.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54(*4), 389-393.

Wilson, J. R., & Rajan, J. A. (1995). Human-machine interfaces for systems control. In J. R. Wilson and E. N. Corlett (Hrsg.), *Evaluation of Human Work: a practical ergonomics methodology.* (S. 357-405). London: Taylor and Francis.

Wirtz, M. A. (Hrsg.) (2021). *Dorsch: Lexikon der Psychologie* (18. überarbeitete Auflage). Hogrefe.

Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index. *Chinese journal of industrial hygiene and occupational diseases, 23*(3), 178-181.

Yeh, Y. Y., & Wickens, C. D. (1988). *Dissociation of performance and subjective measures of workload. Human factors, 30*(1), 111-120.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). *State of science: mental workload in ergonomics. Ergonomics, 58*(1), 1-17.

Young, M. S., Stanton, N. A., & Walker, G. H. (2006). In loco intellegentia: human factors for the future European train driver. *International journal of industrial and systems engineering, 1*(4), 485-501.

Yurko, Y. Y., Scerbo, M. W., Prabhu, A. S., Acker, C. E., & Stefanidis, D. (2010). Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX tool. *Simulation in healthcare, 5*(5), 267-271.

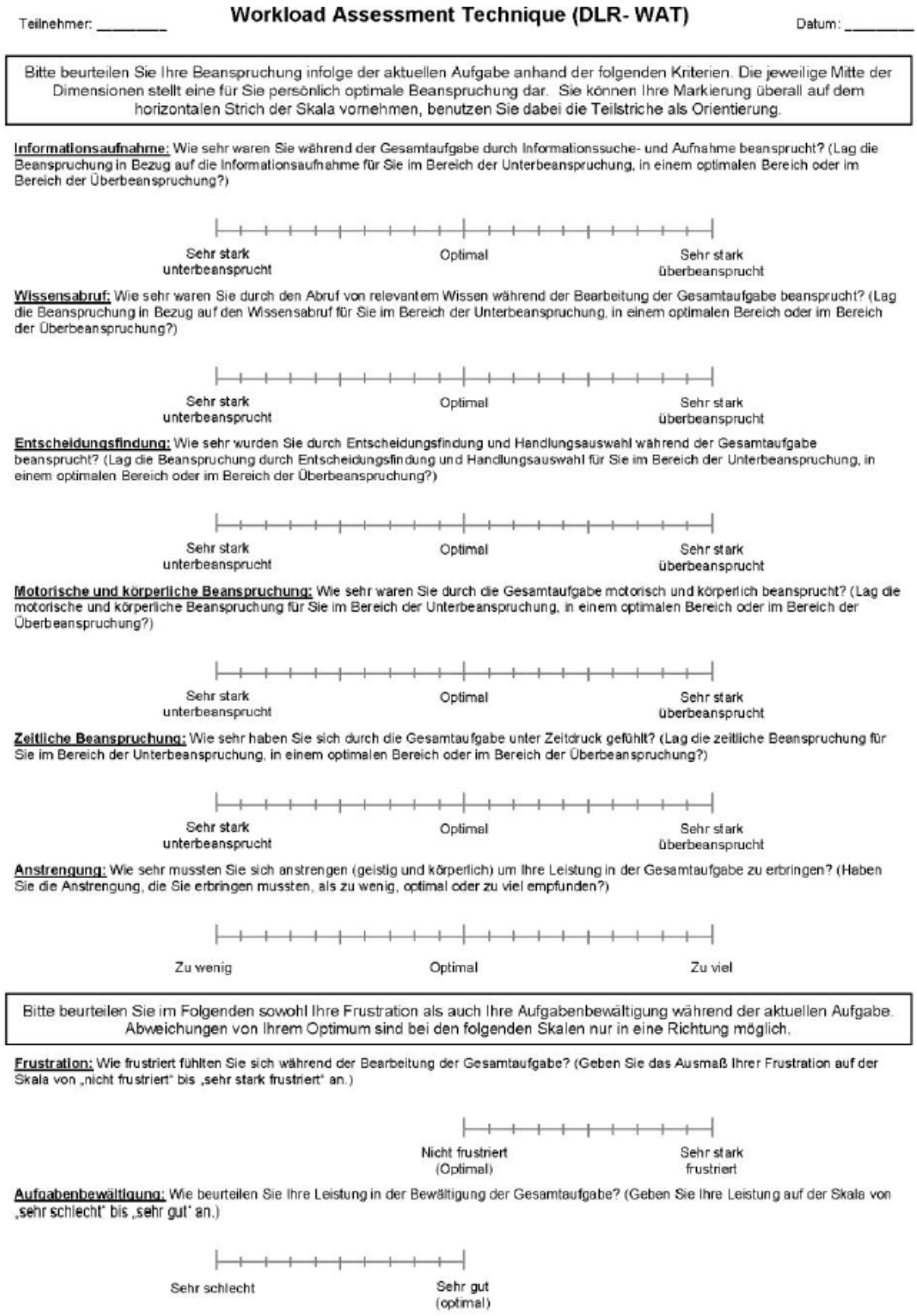
Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surgical endoscopy, 26*(10), 2746-2750.

Zijlstra, F., & van Doorn, L. (1985). *The construction of a scale to measure subjective effort*. (Technical Report). Delft University of Technology.

7 Anhang

Anhangsverzeichnis

A. DLR-WAT



Teilnehmer: _____ **Workload Assessment Technique (DLR- WAT)** Datum: _____

Bitte beurteilen Sie Ihre Beanspruchung infolge der aktuellen Aufgabe anhand der folgenden Kriterien. Die jeweilige Mitte der Dimensionen stellt eine für Sie persönlich optimale Beanspruchung dar. Sie können Ihre Markierung überall auf dem horizontalen Strich der Skala vornehmen, benutzen Sie dabei die Teilstriche als Orientierung.

Informationsaufnahme: Wie sehr waren Sie während der Gesamtaufgabe durch Informationssuche- und Aufnahme beansprucht? (Lag die Beanspruchung in Bezug auf die Informationsaufnahme für Sie im Bereich der Unterbeanspruchung, in einem optimalen Bereich oder im Bereich der Überbeanspruchung?)

Sehr stark unterbeansprucht — Optimal — Sehr stark überbeansprucht

Wissensabruf: Wie sehr waren Sie durch den Abruf von relevantem Wissen während der Bearbeitung der Gesamtaufgabe beansprucht? (Lag die Beanspruchung in Bezug auf den Wissensabruf für Sie im Bereich der Unterbeanspruchung, in einem optimalen Bereich oder im Bereich der Überbeanspruchung?)

Sehr stark unterbeansprucht — Optimal — Sehr stark überbeansprucht

Entscheidungsfindung: Wie sehr wurden Sie durch Entscheidungsfindung und Handlungsauswahl während der Gesamtaufgabe beansprucht? (Lag die Beanspruchung durch Entscheidungsfindung und Handlungsauswahl für Sie im Bereich der Unterbeanspruchung, in einem optimalen Bereich oder im Bereich der Überbeanspruchung?)

Sehr stark unterbeansprucht — Optimal — Sehr stark überbeansprucht

Motorische und körperliche Beanspruchung: Wie sehr waren Sie durch die Gesamtaufgabe motorisch und körperlich beansprucht? (Lag die motorische und körperliche Beanspruchung für Sie im Bereich der Unterbeanspruchung, in einem optimalen Bereich oder im Bereich der Überbeanspruchung?)

Sehr stark unterbeansprucht — Optimal — Sehr stark überbeansprucht

Zeitliche Beanspruchung: Wie sehr haben Sie sich durch die Gesamtaufgabe unter Zeitdruck gefühlt? (Lag die zeitliche Beanspruchung für Sie im Bereich der Unterbeanspruchung, in einem optimalen Bereich oder im Bereich der Überbeanspruchung?)

Sehr stark unterbeansprucht — Optimal — Sehr stark überbeansprucht

Anstrengung: Wie sehr mussten Sie sich anstrengen (geistig und körperlich) um Ihre Leistung in der Gesamtaufgabe zu erbringen? (Haben Sie die Anstrengung, die Sie erbringen mussten, als zu wenig, optimal oder zu viel empfunden?)

Zu wenig — Optimal — Zu viel

Bitte beurteilen Sie im Folgenden sowohl Ihre Frustration als auch Ihre Aufgabenbewältigung während der aktuellen Aufgabe. Abweichungen von Ihrem Optimum sind bei den folgenden Skalen nur in eine Richtung möglich.

Frustration: Wie frustriert fühlten Sie sich während der Bearbeitung der Gesamtaufgabe? (Geben Sie das Ausmaß Ihrer Frustration auf der Skala von „nicht frustriert" bis „sehr stark frustriert" an.)

Nicht frustriert (Optimal) — Sehr stark frustriert

Aufgabenbewältigung: Wie beurteilen Sie Ihre Leistung in der Bewältigung der Gesamtaufgabe? (Geben Sie Ihre Leistung auf der Skala von „sehr schlecht" bis „sehr gut" an.)

Sehr schlecht — Sehr gut (optimal)

Abbildung 6. Fragebogen DLR-WAT

B. NASA-TLX

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

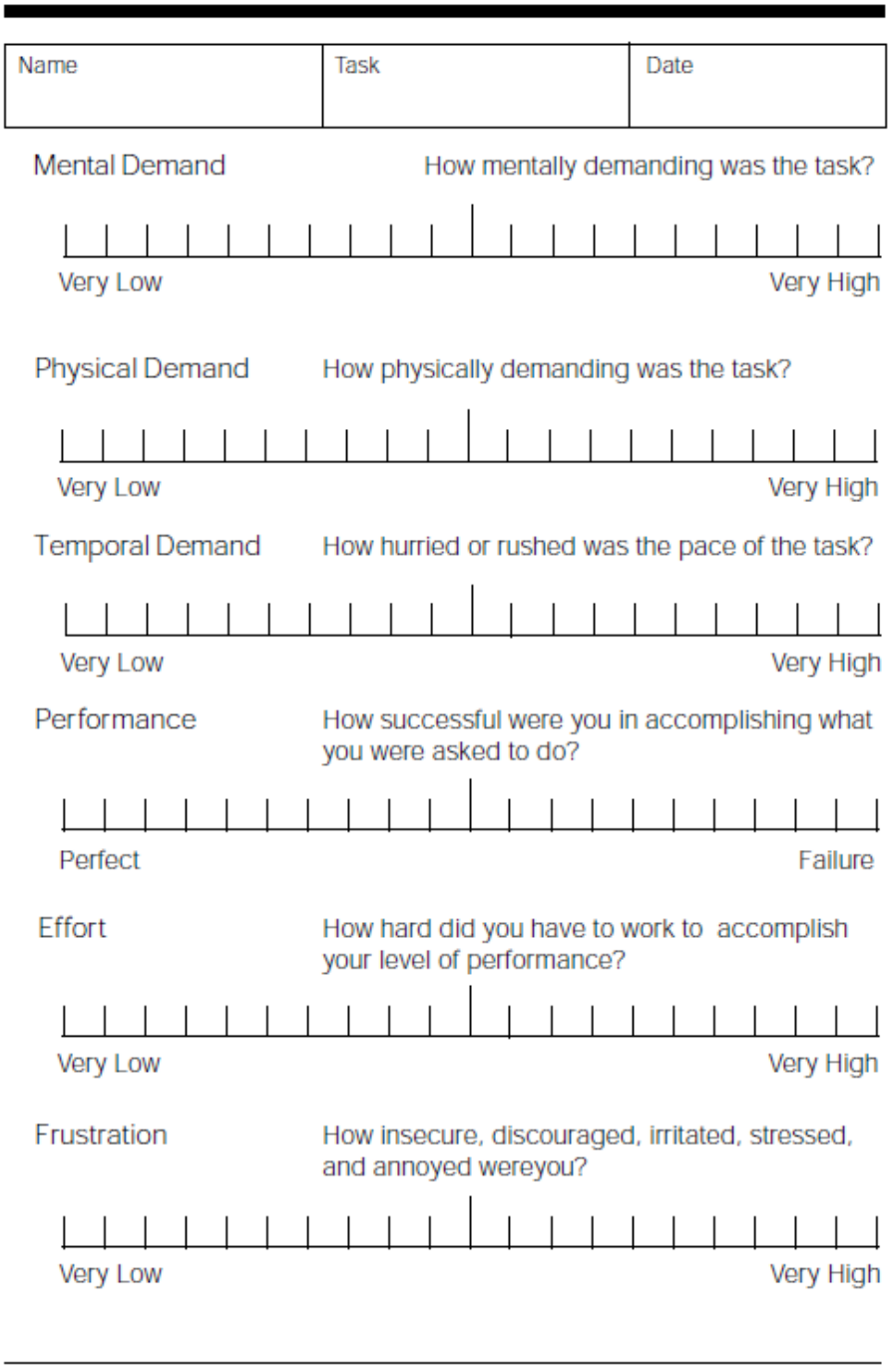


Abbildung 7. Fragbogen NASA-TLX RAW

Erklärung über die selbstständige Abfassung der Arbeit

Ich versichere, dass ich die vorliegende Masterarbeit selbständig verfasst und dabei keine anderen Hilfsmittel als die im Literaturverzeichnis genannten benutzt habe. Ich erkläre, dass die Arbeit in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden ist. Außerdem erkläre ich mich damit einverstanden, dass die von mir erstellte Abschlussarbeit in die Bibliothek der Technischen Universität Braunschweig aufgenommen und der Öffentlichkeit zugänglich gemacht wird.

Braunschweig, 30.09.2022 — Anne Seib

Ort, Datum — Unterschrift