Bachelorarbeit

Visualisierung von Warnsignalen

Eine empirische Studie zur Sicherheit mithilfe des Dual-Task-Paradigmas

Autor **Niclas Prock**

Informatik B. Sc.
Fachbereich 3
Mathematik und Informatik
Universität Bremen

Betreuer Dr.-Ing. Robert Porzel

Erster Begutachter Dr.-Ing. Robert Porzel
Zweiter Begutachter Prof. Dr. Rainer Koschke

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Bachelorarbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen oder anderen Quellen entnommen sind, sind als solche eindeutig kenntlich gemacht. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahren gewesen.

Bremen, den 8. September 2020

Niclas Prock

Danksagung

Ich möchte besonders meinem Betreuer Robert Porzel danken, der mir während des gesamten Prozesses hilfreiches und wertvolles Feedback gegeben hat. Mein Dank gilt außerdem Prof. Dr. Rainer Koschke, der sich als zweiter Begutachter bereiterklärt hat.
Auch möchte ich allen Testern und Freunden danken, die mir mit ihrer objektiven Kritik geholfen haben, diese Studie erfolgreich durchzuführen. Für die Korrekturarbeit möchte ich mich außerdem bei Ann-Sophie Flock bedanken.

Gender-Erklärung

Aus Gründen der besseren Lesbarkeit wird in dieser Bachelorarbeit die Sprachform des generischen Maskulinums angewandt. Es wird an dieser Stelle darauf hingewiesen, dass die ausschließliche Verwendung der männlichen Form geschlechtsunabhängig verstanden werden soll.

Abkürzungsverzeichnis

GUI Graphical User Interface

DMI Driver Machine Interface

RSE Redundant Signal Effect

QQ-Plot Quantile-Quantile Plot

NASA-TLX National Aeronautics and Space Administration-Task Load Index

1 Einleitung

1.1 Motivation

Das Steuern von Transportmitteln ist ein komplexer und wichtiger Prozess. Diese eher mechanischen Steuerungsmethoden werden immer weiter durch ein Graphical User Interface (GUI) ersetzt. Anstatt an Hebeln zu ziehen oder Knöpfe zu betätigen müssen Bildschirme bedient werden. Diese Entwicklung wirft neue Fragen für die Mensch-Technik-Interaktion auf. Bei der Entwicklung eines GUI für diese Transportmittel gibt es viele relevante Faktoren, die beachtet werden müssen. Besonders der Entwurf von Warnsystemen ist hierbei ein wichtiges Thema. Da es sich um sicherheitskritische Systeme handelt, wird ein besonderer Wert auf die Reaktionszeit und das einfache Verständnis der Signale gesetzt. Als Warnsignal oder Signal wird hier ein Element bezeichnet, das unvermittelt in Erscheinung tritt und Informationen transportiert. Dabei kann es sich um klassische Warnsignale wie Schilder im Straßenverkehr oder auch eine Warnleuchte im Armaturenbrett handeln.

Viele Systeme zur Visualisierung dieser Signale wurden bereits entwickelt und in verschiedenen Studien analysiert. Auf diese wird in Kapitel 2 genauer eingegangen. Häufig bezieht sich die Forschung hierbei auf die Automobilbranche. In der vorliegenden Arbeit wird der Fokus auf dem Lokomotivsektor liegen.

Bei der Entwicklung der entsprechenden Oberflächen muss entschieden werden, welche Darstellungsweise am besten für dieses Szenario geeignet ist. Dabei sollte der Verlust von Informationen sowie ein Ablenken von der eigentlichen Aufgabe - dem Führen des Fahrzeuges - verhindert werden. Auch der Einfluss von potenziell inkorrekten Warnmeldungen sollte möglichst gering gehalten werden. Die technische Komplexität der meisten Autos ist in den letzten Jahren stark angestiegen. Die dadurch möglicherweise bedingte Reizüberflutung der Fahrer kann dazu führen, dass diese vom Autofahren abgelenkt werden. Das Gleiche gilt auch für den Lokomotivverkehr. Der Lokführer hat ein komplexes Interface, welches er bedienen muss. Durch neue Technologien sollen zusätzliche Ausgabequellen für Warnsignale mit eingebunden werden. Die Darstellung dieser zusätzlichen Informationen stellt aufgrund der bereits enormen Datenmenge eine große Herausforderung dar. Besonders die Verwendung von Live-Videomaterial, wie zum Beispiel bei einer Rückfahrkamera, kann den Blickwinkel und somit den gesamten Fokus stark beeinflussen (Prekopcsák und Nagy 2011 und Brumby et al. 2007). Zusätzliche Information können ein Mehrwert für den Fahrer darstellen. Auf der anderen Seite sind zu viele Informationen jedoch möglicherweise auch verwirrend und reizüberflutend. In Bezug auf Lokomotiven könnte es sich dabei um wertvolle Informationen zur aktuellen Verkehrslage auf den Schienen oder Ähnliches handeln. Diese und weitere Faktoren gilt es bei der Visualisierung von Warnsignalen zu beachten.

Einige der technischen Neuerungen sind Warnsignale zur unmittelbaren Verkehrslage. Eine der einfachsten Formen dieses Signals ist das akustische ”Piepen” beim Einparken. Andere Arten von Warnsignalen können haptisch, wie ein vibrierendes Lenkrad, sein oder visuell in Form von Text oder Symbolen auf einem Display. Bei sicherheitskritischen Systemen werden häufig mehr als nur eine Form des Feedbacks genutzt, um den Fahrer auf die Gefahrensituation aufmerksam zu machen (Lee und Spence 2008 und Jeon et al. 2009).

1.2 Forschungsfrage

In dieser Arbeit liegt der Fokus auf der Evaluierung von unterschiedlichen Darstellungsformen von Warnsignalen und ihrer Auswirkung auf den Menschen in Stress-Situationen. Hierbei soll die Frage, welche Visualisierung in Gefahrensituationen am adäquatesten ist, beantwortet werden. Dafür wurden die zwei folgenden Hypothesen aufgestellt. In Kapitel 6 werden sie ausgewertet und interpretiert.

H_1 = *Die Fehlerrate beim Bearbeiten von Warnsignalen mit maximal einem Signal zur gleichen Zeit ist geringer als mit mehreren Signalen gleichzeitig.*

H_2 =*Die Effizienz beim Bearbeiten von Warnsignalen steigt, wenn eine größere Menge an Informationen gleichzeitig zur Verfügung steht.*

Es gibt viele Herangehensweisen an das Design eines Warnsignals. Dazu gehört unter anderem, ob das Signal ein auditives, visuelles oder haptisches Feedback zurückgibt. Bei der Verwendung von mehr als einer Art von Signal wird außerdem analysiert, inwiefern sich die Modalitäten gegenseitig beeinflussen. Hierzu gibt es bereits mehrere Studien. Man verwendet *sensor-bridging* oder *sensor-sharing*, um verschiedene Modalitäten aufeinander abzubilden (Prekopcsák und Nagy 2011). Problematisch wird es jedoch, wenn die Menge an sensorischem Feedback ausgereizt ist und zusätzliche Signale nicht erwünscht sind. Um dieses Problem zu berücksichtigen und das Ausmaß dieser Arbeit nicht zu überschreiten, wurde sich hier auf die Untersuchung eines rein visuellen Feedbacks beschränkt (Baranyi und Csapo 2010).

In der aktuellen Forschung zu diesem Thema wird viel mit intelligenten oder adaptiven Interfaces gearbeitet, um höhere Aufmerksamkeit auf ein Signal zu richten (Torok 2016 und Jämsä und Kaartinen 2015). Andere Herangehensweisen betrachten zum Beispiel Faktoren wie die Häufigkeit der Signale, die Darstellungsform oder die Position der angezeigten Informationen (Levy und Pashler 2008, Prekopcsák und Nagy 2011 und Izullah et al. 2016). In Kapitel 2 wird näher auf den Effekt der einzelnen Faktoren eingegangen. In dieser Arbeit soll untersucht werden, welche Auswirkung eine unterschiedliche Menge an Informationen auf die Leistung hat, da hierzu noch keine Forschungsliteratur zu existieren scheint. In den Arbeiten von Horowitz und Dingus und Velichkovsky et al. wurde bereits untersucht, welche Auswirkungen die Häufigkeit der Signale auf das Verhalten des Nutzers haben kann (1992, 2002). Diese Studien zeigen, was passiert, wenn Warnsignale sehr selten auftreten oder wenn ein häufiger Wechsel zwischen Evaluationen zweier Aufgaben stattfindet. Der Fokus meiner Arbeit liegt auf der reinen Menge der angezeigten Signale und der Auswirkung von aufeinanderfolgenden oder gleichzeitig dargestellten Signalen auf den Benutzer.

1.3 Herangehensweise

Aufgrund des Bezuges dieser Arbeit zur Steuerung von Lokomotiven soll es sich bei den Warnsignalen nicht um akute Warnungen zu Unfällen oder Ähnlichem handeln. Im KFZ-Bereich müssen häufig innerhalb von Bruchteilen einer Sekunde Entscheidungen getroffen werden, um einen Unfall zu vermeiden. Bei einer Lokomotive dagegen sind Entscheidungen über mögliche Probleme vermutlich nicht innerhalb einer sehr kurzen Zeit zu treffen. Der Zugfahrer ist nicht in der Lage, die Position oder Geschwindigkeit des Zuges rapide zu verändern. Die Entscheidung

darüber, ob ein Gegenstand oder Ähnliches die Schienen blockiert, muss zum Beispiel mithilfe von Kameras Sekunden oder Minuten vor Eintreffen des Zuges an dieser Position entschieden werden. Die auftretenden Probleme können also relativ komplex sein und erfordern kein rein reaktives Verhalten des Anwenders (Horrey et al. 2009).

Zur Untersuchung der Forschungsfrage soll empirisch vorgegangen werden. Dafür wird ein kontrolliertes Experiment durchgeführt, damit alle Variablen und Faktoren berücksichtigt werden können. Die Nutzer werden zwei verschiedene Versionen des Experiments durchlaufen. Bei jedem Durchlauf wird die Leistung des Probanden gemessen und im Anschluss verglichen, ob die Änderung einer Variable einen signifikanten Unterschied auf das Ergebnis hat oder nicht. Proband und Nutzer werden im Folgenden synonymisch verwendet.

Das kontrollierte Experiment wird in Form einer Online-Anwendung stattfinden. Dafür wird eine *gamifizierte* Anwendung entwickelt. Innerhalb dieser Arbeit werden dafür die Begriffe Anwendung oder System synonymisch verwendet. Es handelt sich nach Anderie nicht um ein klassisches Computerspiel, da das Ziel nicht „die Unterhaltung des Users“ ist (2016). Hierfür wird eine Website erstellt, auf der Probanden an dem Experiment teilnehmen können. Hier werden sie eine Haupt- und Nebenaufgabe haben. Letztere wird in mehreren Versionen vorhanden sein, die sich in der Menge der gleichzeitig verfügbaren Informationen unterscheiden. Die Versionen oder auch Durchläufe sind die Grundlage zur Messung einer Differenz. Durch die Messung verschiedener Faktoren wird evaluiert, welche Attribute oder Kombination dieser sich gut eignen. Die genaue Definition der Leistung wird in Kapitel 4 erläutert.

Für die Hauptaufgabe soll eine kognitiv anspruchsvolle Aufgabe gewählt werden, um den Fokus der Probanden hierauf zu lenken. In der Nebenaufgabe sollen komplexe Entscheidungen getroffen werden, die ebenfalls kognitive Eigenschaften beanspruchen. Dieses Dual-Task Setting sorgt dafür, dass immer eine der beiden Aufgaben bei Bearbeitung der anderen vernachlässigt wird, solange keine Automatisierung durch den Menschen möglich ist (Riby et al. 2004).

1.4 Aufbau der Arbeit

In diesem Abschnitt erläutere ich die Struktur der Arbeit, um einen besseren Überblick über die Inhalte zu ermöglichen. Zunächst werde ich im Kapitel Themenbezogene Arbeiten auf die Literatur eingehen, die Einfluss auf diese Arbeit hat. Im Anschluss wird in Systemdesign und Studiendesign erläutert, welche Entscheidungen für die Entwicklung der Studie und die dafür notwendige Anwendung getroffen wurden. Es folgt ein Auswerten und Darstellen der Daten aus der Studie im Kapitel Auswertung. In den letzten beiden Kapiteln Analyse und Fazit und Ausblick werden die vorher untersuchten Daten hinsichtlich der Hypothese interpretiert das Ergebnis reflektiert. Hier wird ebenfalls auf mögliche zukünftige Arbeiten und Fehler dieser Studie eingegangen. Um die Forschungsfrage zu beantworten wird zunächst die grundlegende Literatur in Kapitel 2 näher erläutert.

2 Themenbezogene Arbeiten

Nach der Einleitung und Motivation für die Arbeit wird im folgenden Kapitel nun der aktuelle Stand der Forschung in diesem Gebiet näher erläutert, um Entscheidungen für das Design des Systems, der Studie und die Analyse treffen zu können. Diese Arbeit stützt sich auf viele Ergebnisse von vorangegangenen Studien aus der Psychologie und Informatik, besonders aus dem Bereich der *Human-Computer-Interaction*. Im Gegensatz zum KFZ-Bereich gibt es im Bereich der Steuerung von Lokomotiven sehr wenig Studien und relevante Literatur. Es scheint hier aufgrund der geringeren Masse an Systemen und Herstellern weniger öffentlich zugängliche Forschungsliteratur zu geben. Ein großer Unterschied zwischen Auto- und Lokomotivführern ist, dass es sich im Kontext von Zügen immer um professionelle Nutzer handelt, die für ein bestimmtes Fahrzeug geschult wurden. Dadurch können komplexere Systeme und Interfaces verwendet werden.

2.1 Dual-Task Studie

Zur Analyse von Warnsignalen und ihrer Auswirkung auf den Nutzer können verschiedene Methoden angewendet werden. In dieser Arbeit wird eine Dual-Task-Studie durchgeführt, da die Warnsignale nicht rein reaktiv sondern kognitiv beanspruchende Aufgaben sein werden. Diese Form der Studie erlaubt es, eine Haupt- und Nebenaufgabe zu entwerfen, die sich dem realen Fahren einer Lokomotive annähert. Die Dual-Task Studie ermöglicht das Messen des Einflusses verschiedener Nebenaufgaben auf die Performance eines Nutzers. Eine Auswertung von lediglich der Nebenaufgabe könnte zu anderen Ergebnissen führen. Der Grund dafür ist das Dual-Task-Paradigma. Bei einer Dual-Task-Aufgabe gibt es eine zentrale, aufmerksamkeitsintensive Aufgabe, die sich im Zentrum des Blickfeldes befindet, und einen sekundären Stimulus, der in der Peripherie liegt. Das Dual-Tasks-Paradigma quantifiziert dabei, welche Art von Reizattributen in der Beinahe-Abwesenheit von räumlicher Aufmerksamkeit signalisiert und bewusst wahrgenommen werden kann (Ahmadi et al. 2011). Dual-Task-Studien zur Untersuchung der Reaktionszeit und allgemeinen Leistung beim Autofahren haben viele Faktoren aufgezeigt, die diese beeinträchtigen können (Unterkapitel 2.2, Unterkapitel 2.3.1). Besonders gilt jedoch, dass eine zusätzliche Aufgabe immer eine Einschränkung bedeutet, auch wenn darauf hingewiesen wird, diese situationsabhängig zu vernachlässigen (Levy und Pashler 2008).

Riby et al. haben in ihrer Studie untersucht, inwiefern der Umfang der Aufgabe einen Einfluss auf die Leistung des Probanden hat (2004). Aufgaben, die relativ einfach oder weitestgehend automatisierbar sind, haben demnach einen deutlich kleineren negativen Einfluss als andere. Zu den automatisierbaren Aufgaben zählen solche, die einen hohen Lerneffekt haben. Autofahren ist zum Beispiel eine komplexe Aufgabe mit hoher kognitiver Beanspruchung. Trotzdem ist es für den Menschen möglich, diese Aufgabe mit genügend Übung beinahe automatisch auszuführen. Wie fordernd die Tätigkeit ist wird dabei häufig falsch eingeschätzt und die tatsächlichen Daten weichen von den Selbsteinschätzungen der Probanden ab (Murata et al. 2013).

2.2 Visualisierung von Warnsignalen

Die Forschung im Bereich Visualisierung beschäftigt sich häufig mit der Frage, auf welche Art und Weise Informationen dargestellt werden sollten. Der Fokus dieser Arbeit liegt allerdings auf der Menge und Häufigkeit von Informationen. Das Kapitel Weiterführende Literatur zeigt einige Ergebnisse aus dem Bereich der Darstellungsform, während dieses Kapitel auf Studien eingeht, die sich mit der Menge und Häufigkeit der Informationen beschäftigen.

Warnsignale treten im Allgemeinen so selten auf, dass sie möglicherweise zu weiterem Stress führen können, wenn es zu einer tatsächlichen Gefahrensituation kommt. Auf der anderen Seite können zu häufige Warnsignale dazu führen, dass diese ignoriert werden, auch dann, wenn es sich um eine tatsächliche Gefahrensituation handelt. Ein weiterer Faktor, der die Wahrnehmung dieser Signale negativ beeinflussen kann, ist das häufige Wechseln des Fokus zwischen dem Evaluieren einer Gefahr und der Ausübung einer anderen Aufgabe (Horowitz und Dingus 1992 und Velichkovsky et al. 2002). Interessanterweise können zusätzliche fehlerhafte Signale die Leistung einer Person erhöhen. Beim Ausführen von einigen Aufgaben können auch inkorrekte Warnsignale die Aufmerksamkeit wieder auf die wichtigere Aufgabe richten und helfen somit, die Leistung der ausführenden Person zu verbessern (Yan et al. 2015).

2.3 Weiterführende Literatur

Viele Studien beschäftigen sich mit der Frage, wie man Informationen darstellen muss, damit diese besonders schnell und zuverlässig verarbeitet werden. Da sich diese Arbeit mit der Menge der Informationen und nicht ihrer Darstellung beschäftigt, wurden die Ergebnisse der weiterführenden Literatur nicht berücksichtigt. Für eine allgemein bessere User-Experience oder schnellere Verarbeitung der Nebenaufgabe könnten diese Ergebnisse verwendet werden. Das Ziel war es jedoch nicht, dem Nutzer eine möglichst einfache Aufgabe zu geben sondern den potentiell unterschiedlichen Einfluss einer geringen und einer größeren Menge an Signalen zu untersuchen.

2.3.1 Präattentive Attribute

Präattentive Wahrnehmung ist die unterbewusste Verarbeitung von elementaren visuellen Eigenschaften von Objekten (Few 2004). Mithilfe dieser Attribute ist es möglich, sich bereits unterbewusst auf ein Warnsignal vorzubereiten und früher eine adäquate Aktion auszuführen (Velichkovsky et al. 2002). Indem alle visuellen Eindrücke so einfach wie möglich gehalten werden und nur die relevanten Daten mit präattentiven Eigenschaften hervorgehoben werden, können diese deutlich effizienter verarbeitet werden.

2.3.2 Multimodale Interfaces

Die Verwendung von multimodalen Interfaces führt zu besseren Reaktionszeiten. Im Vergleich zu einzelnen Stimuli (visuell, auditorisch, haptisch) ist die Leistung eines Probanden bei der Kombination von mehreren Signalen besser. Aufgrund dessen eignen sich Warnsysteme besonders gut für ein multimodales Interface (Biondi et al. 2017a).

Die Ergebnisse der anderen Studien werden im nächsten Kapitel Auswirkung darauf haben, wie das zu entwickelnde System entworfen wird. Besonders geht es im Folgenden darum, die Erkenntnisse aus Unterkapitel 2.1 in die Designentscheidungen einfließen zu lassen.

3 Systemdesign

Im folgenden Kapitel wird die Entwicklung der Anwendung erläutert. Die Probanden sollen in Form einer Dual-Task-Studie auf Warnsignale reagieren. Es handelt sich hierbei um ein kontrolliertes Experiment. Dafür wird ein System entwickelt, das zwei kognitiv ansprechende Aufgaben anbietet. Der Nutzer soll diese parallel bearbeiten, um einen möglichst hohen Score (in dieser Arbeit synonymisch mit Punktzahl) zu erreichen. Der Aufbau der Studie richtet sich nach den Vorgaben aus anderen Studien, die bereits in Kapitel 2 näher erläutert wurden.

3.1 Anforderungsanalyse

Aus dem in Kapitel 2 gewonnenen Wissen über Dual-Task-Studien und Visualisierung wurden für die zu entwickelnde Anwendung Anforderungen zusammengefasst. Es folgt eine begründete Erläuterung zum Entstehungsprozess dieser. In den folgenden Sektionen wird detailliert auf die genaue Umsetzung jeder Anforderung eingegangen. Die finale Version des Prototypen ist unter Abbildung 2 einzusehen.

Da das Auswerten von Warnsignalen typischerweise keine alleinige Aufgabe sondern beim Steuern eines Zuges oder auch eines Automobils eine Nebenaufgabe ist, wird ein Dual-Task-Design für die Studie gewählt. Das System muss aufgrund dessen zwei kognitiv fordernde Aufgaben anbieten: die Haupt- und Nebenaufgabe. Die Hauptaufgabe hat dabei direkten Einfluss auf die Leistung der Nebenaufgabe und vice versa (Strayer und Johnston 2001). Erstere darf nicht zu stark automatisierbar sein, wie Riby et al. gezeigt haben, da sonst der Mehrwert einer Dual-Task-Studie verloren gehen kann (2004). Die Erfahrung mit Computerspielen kann dabei einen Einfluss auf die Leistung der Teilnehmer und wie schnell sie in der Lage sind, eine bestimmte Aufgabe zu automatisieren, haben. Eine 2006 durchgeführte Studie zeigt, dass Erfahrung mit Videospielen beispielsweise eine erhöhte Leistung beim Verfolgen von mehreren Objekten zur gleichen Zeit haben kann (Green und Bavelier). Damit die Aufgabe auch für Probanden mit Spielerfahrung nicht automatisierbar oder zu leicht wird, wurden diese Ergebnisse mit in Betracht gezogen. Die Nebenaufgabe darf ebenfalls nicht zu leicht gewählt werden. Sie beschäftigt sich mit der Visualisierung von Warnsignalen. Hier wird ein besonderer Mehrwert aus den Daten gewonnen, wenn eine hohe Fehlerquote bei den Probanden erreicht wird. Wenn alle Aufgaben vom Nutzer korrekt erfüllt werden, wird kein Unterschied zwischen den Versionen festgestellt. Deswegen war es beim Entwerfen der Anwendung wichtig, dass der Nutzer geistig gefordert wird, zeitlich unter Druck steht und die Erfahrung im Allgemeinen als anstrengend wahrnimmt.

Da das System für Lokomotivführer entwickelt wird, steht nicht die Unterhaltung des Users im Vordergrund und die Anwendung muss auch nicht für jeden Endverbraucher einfach zugänglich sein. Es geht hierbei darum, ein besonders sicheres, effizientes System zu erschaffen. Der Begriff der Sicherheit meint in diesem Kontext, dass wenig Fehler bei der Bearbeitung der Aufgaben gemacht werden. Die Bearbeitungsrate dieser Aufgaben wird als Effizienz bezeichnet. Da die Bearbeitung einer Aufgabe unmittelbar durch einen Klick oder einen Tastendruck erfolgt, kann die Zeit einer einzelnen Aufgabe nicht berechnet werden. Stattdessen wird die Menge an erfüllten Aufgaben pro Durchlauf gemessen. Außerdem ist es kein Ziel, die Anwendung besonders anschaulich für den Endbenutzer zu gestalten sondern sich hier nahe an der Domäne

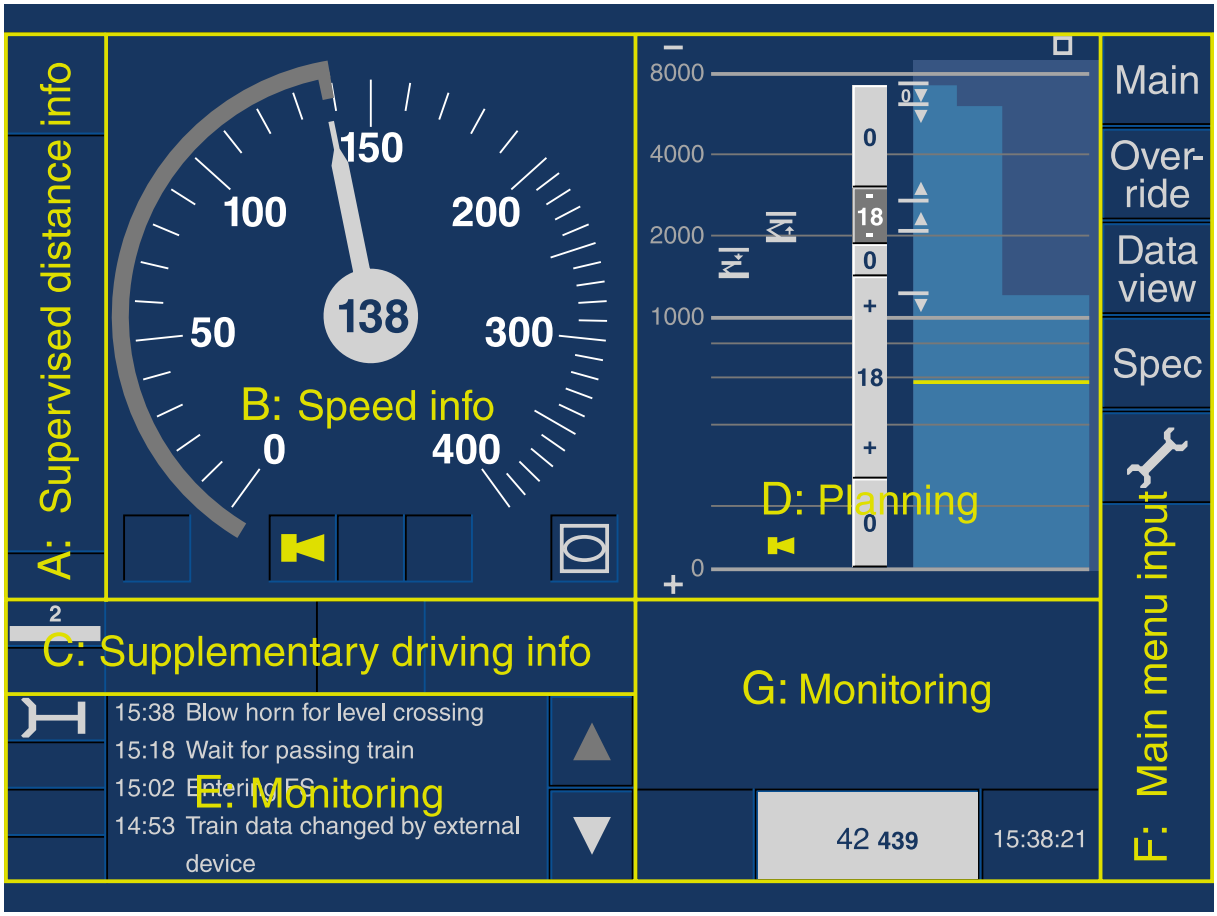

Abbildung 1: Originales Interface eines DMI (de Vries 2016 und Reynolds et al. 2017)

zu orientieren, auch wenn die Nutzerzufriedenheit und Produktivität durch ansprechenderes Design erhöht werden kann. Die genauen Spezifikationen zur Nebenaufgabe werden in Unterkapitel 3.4 erläutert.

3.2 Allgemeine Designentscheidungen

Das Driver Machine Interface (DMI) ist die Schnittstelle zwischen dem Lokführer und seinem Fahrzeug. Es bietet für diese Arbeit das grundlegende Layout für das zu entwickelnde System. Interessant sind dabei besonders die Bereiche ”B: Speed info”, ”D: Planning”, und ”G: Monitoring” von Abbildung 1. Das Steuern des Zuges findet hauptsächlich in den Bereichen B und D statt. Diese werden in meiner Anwendung in einer abstrakten Form dargestellt. Im Bereich G soll die Nebenaufgabe implementiert werden. Hilfreiche Informationen zur Nebenaufgabe sowie ein Score sollen in den Bereichen ”C: Supplementary driving info” und ”E: Monitoring” implementiert werden. Die Proportionen aus ungefähr zwei zu eins von oberem zu unterem Teil werden ebenfalls übernommen. Die Bereiche A und F werden komplett ignoriert, um nicht zusätzliche Informationen anzuzeigen, die den Nutzer verwirren könnten.

Die Anwendung wird für Deutschsprachige entwickelt, die im Allgemeinen mit der Benutzung von Computern und Websites vertraut sind. Viele computerversierte Menschen werden keine Schwierigkeiten mit der Verwendung und den Interaktionsmöglichkeiten haben. Darüber hinaus wird es keine besondere Unterstützung für Personen außerhalb dieser Gruppen geben. Das DMI wird typischerweise auf einem kleinen Bildschirm von ungefähr 10 Zoll angezeigt. Um diesen Zustand widerzuspiegeln wird das System eine feste Auflösung haben, die deutlich geringer als das gängige Full-HD Format mit 1920x1080 Pixeln ist. Dadurch sollte die Anwendung bei den meisten Probanden nur auf einem Teil ihres Bildschirms angezeigt werden, anstatt diesen zu füllen. Eine genaue Messung dieser Daten wird nicht stattfinden, da es technische Herausforderungen impliziert, die nicht im Verhältnis zu ihrem Nutzen stehen und den Rahmen dieser Arbeit überschreiten würden.

Es wird außer den visuellen Informationen, die sie bereitstellt, keine weiteren Interaktionsmöglichkeiten mit oder Feedback von der Anwendung geben. Besonders bei Warnsignalen gibt es bereits Studien, die untersuchen, ob auditorisches, visuelles oder haptisches Feedback besser geeignet ist (Lee und Spence 2008, Jeon et al. 2009 und Biondi et al. 2017b). Diese Studien beziehen sich auf die Automobilbranche und sind damit nicht deckungsgleich mit der Domäne dieser Studie. Die Auswirkungen verschiedener Modalitäten oder multimodaler Systeme für Lokomotiven sind kein Bestandteil dieser Arbeit. Darüber hinaus ist „Train driving [...] primarily a visual task “(Luke et al. 2006), was diese Entscheidung weiter unterstützt.

Die Farben der Anwendung werden sich ebenfalls nach dem DMI richten. Die Hintergrundfarbe wird auf Dunkelblau festgelegt. Text wird in Weiß angezeigt, um ihn auf dem dunklen Hintergrund besser lesbar zu machen. Ebenso werden alle Symbole ebenfalls in Weiß dargestellt. Auf der Abbildung 1 sind unter der Geschwindigkeitsanzeige in ”B: Speed Info” sowohl gelbe als auch weiße Symbole zu erkennen. Damit dies nicht zu zusätzlicher Verwirrung führt, werden alle einheitlich weiß gefüllt. Weitere Farben, die für besondere Anzeigen oder Ereignisse verwendet werden, sind Rot und Grün. Diese Farben werden allgemein mit richtig und falsch assoziiert. Innerhalb der Anwendung stehen sie ebenfalls für diese Anwendungsfälle.

Anstatt der für den Laien unverständlichen Symbole werden Symbole verwendet, die leicht zu unterscheiden sind. Um kulturelle Differenzen soweit wie möglich auszuschließen wurden Symbole verwendet, die international verständlich sind. Bei der Auswahl bezog ich mich auf die *Point shapes* vom Paket *ggplot2* zur Visualisierung der statistischen Programmiersprache *R*. Diese wurden unter den *ästhetischen Spezifikationen* zur Visualisierung von Daten beschlossen (Wickham et al. 1993).

Das Entwickeln der Anwendung wird in einem iterativen Prozess stattfinden, bei dem Testpersonen häufiger zu bestimmten Aspekten informell befragt werden. Aufgrund dieser Notizen wurde zum Beispiel die grundlegende Geschwindigkeit der Anwendung festgelegt. Diese Metrik legt fest, wie schnell sich Objekte innerhalb der Anwendung bewegen. Wichtig ist dabei, dass es nahezu unmöglich ist, eine perfekte Runde zu erzielen. Der erreichte Score ergibt sich aus den erfüllten Aufgaben. Er kann nicht ins Negative geraten, damit die Frustration bei den Nutzern nicht zu hoch ist. Diese Entscheidung wurde aufgrund eines Gesprächs mit einer Testperson beschlossen. Allgemein wurde sich aber dazu entschieden, einen Score anzuzeigen, um dem Nutzer direktes Feedback für sein Verhalten zu geben. Ohne eine Punkteanzeige können Verständnisprobleme während der Einführungsrunde nicht geklärt werden.

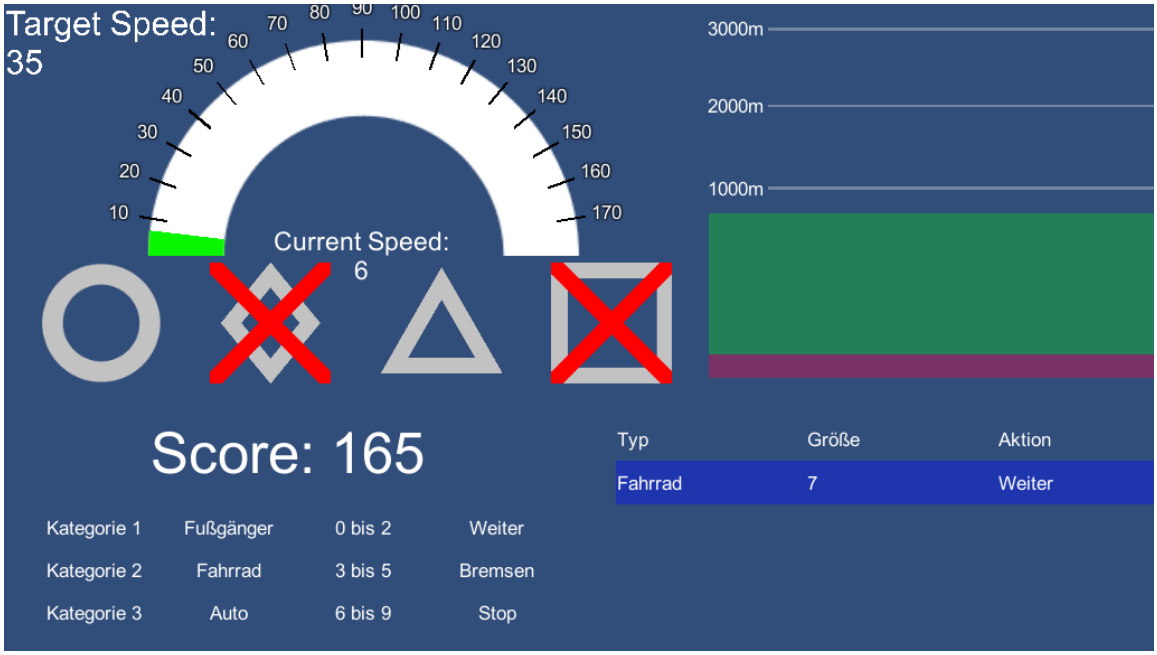

Abbildung 2: Aufbau der entwickelten Anwendung

Die Dauer der Runden wurde durch mehrfaches Testen auf drei Minuten für die Einführungsrunde und zwei Minuten für die weiteren Durchläufe festgelegt. Der Begriff Einführungsrunde wird in diesem Kontext synonymisch zu Tutorial verwendet. Innerhalb dieser Zeit schienen bereits genügend Fehler gemacht worden zu sein, aber der Nutzer war trotzdem gewillt, den Durchlauf nicht frühzeitig zu beenden oder aus Frust zufällige Eingaben zu tätigen. Um die Nutzererfahrung etwas zu steigern wäre eine Anzeige infrage gekommen, die angibt, wie lange die Runde noch dauert. Da das Interface jedoch bereits sehr gefüllt war und bei zwei Minuten typischerweise noch kein Bedürfnis zum Überprüfen der Restlaufzeit beim Nutzer auftrat, wurde sich dagegen entschieden.

3.3 Entwicklung der Hauptaufgabe

Die Hauptaufgabe soll sich kognitiv an die eines Lokführers anlehnen. Aus Platz- und Komplexitätsgründen entscheide ich mich, keinen objektiven Vergleich zwischen der kognitiven Komplexität des DMI und meinem System zu machen (McLeod et al. 2005). In den oberen zwei Dritteln von Abbildung 1 wird die Steuerung des Zuges dargestellt. Die Hauptaufgabe meiner Anwendung soll dies abstrakt widerspiegeln. In Abbildung 2 ist in den oberen zwei Dritteln die abstrakte Version der Hauptaufgabe meiner Anwendung zu sehen. Daraus ergeben sich zwei Teilaufgaben: das Anpassen der Geschwindigkeit und das Aktivieren verschiedener Schaltflächen aufgrund von Events. Im Folgenden werden Schaltflächen und Buttons synonymisch verwendet.

Die Abstraktion dieser Aufgaben erlaubt es mir, die Symbole und Schrift im Gegensatz zu Abbildung 1 zu vergrößern. Die Größe wurde so gewählt, dass trotz eines kleinen Bildschirmes

die Symbole noch gut erkennbar sind. Diese Größe wurde wieder mithilfe der Tester entwickelt. Nicht unterstützt wurden sehr hoch auflösende Displays mit 4K-Auflösung oder mobile Geräte mit sehr kleinen Bildschirmen. Der technische Mehraufwand hätte hier wieder den Rahmen der Arbeit überschritten.

Ein zusätzliches Problem, das mobile Geräte erfahren hätten, ist, dass die parallele Bearbeitung von mehreren Aufgaben mithilfe von Maus und Tastatur entworfen wurde. Die Hauptaufgabe wird mit der Maus und die Nebenaufgabe mithilfe der Tastatur bedient. Das bietet den verschiedenen Nutzern die Möglichkeit, die Aufgaben effizient zu bearbeiten und minimiert die Differenzen zwischen Benutzern mit schlechterer oder besserer Kontrolle über ihre Maus oder Tastatur individuell.

3.3.1 Geschwindigkeit anpassen

Mithilfe der Maus soll auf die gewünschte Geschwindigkeit geklickt werden. Die Geschwindigkeitsanzeige ändert sich nicht sofort sondern linear über einen geringen Zeitraum. Diese Funktion erschwert die Auswahl der korrekten Geschwindigkeit etwas. Die Zielgeschwindigkeit oder Target-Speed (Abbildung 2) wird direkt über der Geschwindigkeitsanzeige angezeigt. Damit der Nutzer nicht blind auf die Anzeige klickt, gibt es in Zehnerschritten Markierungen für die Geschwindigkeit. Die Zielgeschwindigkeit erhöht sich immer entweder um 5 oder um 10, sodass bis zu einer gewissen Genauigkeit die korrekte Geschwindigkeit relativ leicht erreicht werden kann. Diese Abstände und die linear steigende Geschwindigkeit haben sich erneut aus den ersten Tests ergeben. Damit die Frustration und Fehlerrate nicht zu hoch sind, ist eine Toleranz von einem Kmh implementiert.

Die Anzeige selber ist wieder schlicht weiß gehalten. Die Geschwindigkeit wird nicht durch eine Nadel angezeigt sondern, wie auch in Abbildung 1 zu sehen, durch das Füllen der Anzeige bis zur aktuellen Geschwindigkeit. Die Füllfarbe ist ein Farbgradient von Grün zu Rot. Je schneller der Zug fährt, desto schwerer wird es, alle Aufgaben zu erfüllen. Die Füllfarbe spiegelt das unterbewusst wider, hat aber ansonsten keinen weiteren Nutzen. In früheren Versionen war es angedacht, einen höheren Score für erfüllte Aufgaben zu geben, die bei einer höheren Geschwindigkeit ausgeführt wurden. Dies wurde allerdings zu komplex für die User, die zum ersten Mal die Anwendung ausführen. Im Allgemeinen ist die Anzeige sehr groß gewählt, da sich gezeigt hat, dass das korrekte Festlegen der Geschwindigkeit ansonsten mit der Maus zu schwer ist. Hier wurden über mehrere Iterationen versucht, ein gutes Verhältnis zwischen dieser Anzeige und den Schaltflächen für die andere Teilaufgabe zu finden.

3.3.2 Schaltflächen aktivieren

Wie in Abbildung 1 zu sehen, gibt es Events oder Ereignisse, auf die der Proband reagieren muss. Zu diesen Events gibt es eine Entfernung, die durch horizontale Linien angegeben wird. Diese Ereignisse bewegen sich auf der Anzeige immer weiter nach unten. In meiner Anwendung wird dieses Verhalten genauso abgebildet. Die Events „fallen “von oben nach unten und müssen unterhalb der 1000er Linie aktiviert werden (Abbildung 2).

Für jedes dieser Events gibt es ein Symbol, das ebenfalls als Schaltfläche unter der Geschwin-

digkeitsanzeige zu finden ist. Jedes Ereignis kann einen von zwei Zuständen haben - aktiviert oder deaktiviert. Ziel ist es, dass alle Ereignisse zwischen der 1000er Linie und dem roten Bereich mit den Schaltflächen unter der Geschwindigkeitsanzeige übereinstimmen. Durch das Klicken auf die Buttons kann man zwischen den beiden Zuständen wechseln. Wenn sich zum Beispiel ein deaktivierter Kreis (markiert durch ein rotes Kreuz, das diesen durchstreicht) innerhalb dieses Bereiches befindet, versucht der Nutzer auf die Schaltfläche mit dem Kreis zu klicken, um somit in den gleichen Zustand für dieses Symbol zu gelangen. Stimmen die Symbole überein, wird das Ereignis gelöscht und der Score wird erhöht. Sollte das Symbol die 0-Linie erreichen, wird es ebenfalls zerstört, aber es werden keine Punkte verteilt.

Die Position der Schaltflächen wurde so gewählt, da sich die Anwendung nahe am DMI orientiert und außerdem ein häufigerer Fokuswechsel durch das Springen zwischen der Anzeige auf der rechten und linken Seite stattfindet. Nach Horowitz und Dingus und Velichkovsky et al. kann dies dazu führen, dass häufiger Fehler gemacht werden (1992, 2002). Die Fehlerrate relativ hoch zu halten ist weiterhin ein Ziel zur Erhöhung der interessanten Datenmenge.

Die Anzahl der verschiedenen Ereignisse wurde auf vier gesetzt. Das hat einerseits den Vorteil, dass die Symbole relativ groß sein können und eine einfachere Bedienung erlauben. Andererseits wurde sich hier, wie bereits in Unterkapitel 3.2 erwähnt, nach den Point shapes von ggplot2 gerichtet. Hier gibt es vier Symbole, die auch noch gut erkennbar sind, wenn sie durchgestrichen sind. Da dies wichtige Anforderungen für die Events sind, wurde davon abgesehen, mehr Symbole zu verwenden.

Der Bereich zwischen der 1000er- und 0-Linie wurde grün markiert, da Tester beim Lesen der Anleitung ohne eine klare Abgrenzung der Bereiche Verständnisprobleme hatten, wann welche Schaltfläche aktiviert werden muss. Zur allgemeinen Kennzeichnung von korrektem und inkorrektem Verhalten wurde innerhalb der Anleitung immer Grün und Rot verwendet. Aufgrund dessen wurde dies auch für das eigentliche Spiel übernommen.

3.4 Entwicklung der Nebenaufgabe

Die Nebenaufgabe wird im Bereich unten rechts angezeigt. Sie beschäftigt sich mit der Auswertung der Warnsignale. In diesem Kontext werden Signal und Subtask synonymisch verwendet.

Auf Grundlage des Redundant Signal Effect (RSE) kann die Anzahl der Modalitäten eines Signals einen Einfluss auf die Leistung haben (Levy und Pashler 2008). Im Kontext dieser Arbeit wird eine ausschließliche Untersuchung der visuellen Aspekte durchgeführt. Das Ziel dieser Studie ist es, herauszufinden, inwiefern die Menge der gleichzeitigen Informationen einen Einfluss auf den Nutzer hat. Deswegen wird die Nebenaufgabe in der Anzahl der gleichzeitigen Subtasks für die verschiedenen Durchläufe beschränkt.

Die Signale sammeln sich nach und nach an, damit sichergestellt werden kann, dass alle Teilnehmer die Möglichkeit haben, in beiden Versionen gleich viele Signale zu verarbeiten. Die zu verarbeitenden Signale der Nebenaufgabe werden pro Durchlauf zufällig, allerdings in ihrer relativen Häufigkeit immer gleichmäßig auftauchen. Da die Aufgabe nicht vollständig korrekt lösbar sein sollte, wurde eine durchschnittliche Rate von einem Event pro sechs Sekunden

Typ	Größe	Hinweis/Vorschlag
Um was handelt es sich voraussichtlich bei dem erkannten Objekt. Zum Beispiel Auto, Fahrrad oder Fußgänger.	Die geschätze Größe des Gegenstandes in Meter.	Welche Aktion sollte vermutlich vorgenommen werden. (Stop, Bremsen, Weiter)

Tabelle 1: Erläuterung der Parameter der Nebenaufgabe

gewählt. Dieser Wert ergab sich aus dem mehrfachen Testen der Anwendung. Theoretisch besteht kein Zwang, die Nebenaufgabe zu bearbeiten und aufgrund dessen kann es hier zu großen Abweichungen zwischen den Probanden kommen. Diese Entscheidung wurde getroffen, da es in keinem Fall einen sichtbaren Nachteil für die Probanden beim inkorrektem Bearbeiten einer Aufgabe gibt. Wenn also keine Strafe vom System ausgesprochen werden kann, kann es auch keine Notwendigkeit zur Interaktion mit der Nebenaufgabe geben. Jede Entscheidung, die getroffen wird, ist final und kann nicht rückgängig gemacht werden. In diesem Fall ist die Strafe für fehlerhaftes Verhalten, dass keine Punkte vergeben werden.

Die Art und Weise, wie der Nutzer abgelenkt wird, ist relevanter als die Gesamtzeit zur Erledigung der Nebenaufgabe (Brumby et al. 2007). Aufgrund dessen wird der Nutzer gar nicht von der Hauptaufgabe durch die Subtasks abgelenkt sondern muss selber entscheiden, wie viele Ressourcen er für welches Problem aufwenden möchte beziehungsweise kann. Diese Entscheidung sollte es ermöglichen, zu messen, wie viele Aufgaben der Nutzer von sich aus bearbeitet oder ob er bereits mit der Hauptaufgabe überfordert ist. Aufgrund des *Rightward bias* befinden sich die Informationen der Nebenaufgabe auf der rechten Seite (Izullah et al. 2016). Dieser beschreibt das Phänomen, dass Informationen besser wahrgenommen werden, wenn sie sich auf der rechten Seite befinden. Da eine hohe Menge an verarbeiteten Subtasks für die spätere Untersuchung relevant ist, wurde dies mit in das Design einbezogen (Izullah et al. 2016). Adaptive oder intelligente Interfaces können einen ähnlichen Vorteil bieten, sind allerdings technisch schwer umzusetzen und würden den Rahmen der Arbeit überschreiten (Jämsä und Kaartinen 2015 und Torok 2016). Es wird darauf hingewiesen, dass diese Aufgabe sehr fordernd ist, auch wenn sich in anderen Studien gezeigt hat, dass es keinen signifikanten Unterschied macht, die Konzentration der Probanden auf etwas zu lenken (Levy und Pashler 2008).

Wie zu Anfang des Kapitels bereits erwähnt, soll die Nebenaufgabe eine kognitiv anspruchsvolle Aufgabe und nicht trivial sein. Durch eine komplexere Subtask sollen wieder einige Fehler beim Nutzer erzwungen werden, so wie es Horrey et al. (2009) gezeigt haben. Bei einer Subtask geht es darum, ein Ereignis in eine bestimmte Kategorie einzuordnen. Je nachdem zu welcher es gehört wird eine entsprechende Taste auf der Tastatur gedrückt. Wie die einzelnen Parameter dieser Ereignisse lauten und was sie bedeuten erläutert Tabelle 1.

Die Anzahl der Kategorien, Parameter und ihre Wertebereiche wurden iterativ mithilfe der Tester entwickelt und können auf Seite 14 genauer betrachtet werden. Damit man keinen Nachteil dadurch erhält, dass man wenig Erfahrung mit Computerspielen oder der Tastatur im Allge-

Kategorie	Typ	Größe	Hinweis/Vorschlag
1	Fußgänger	0 bis 2	Weiter
2	Fahrrad	3 bis 5	Bremsen
3	Auto	6 bis 9	Anhalten

Tabelle 2: Erläuterung der Kategorien der Nebenaufgabe

meinen hat, wurden drei nebeneinander liegende Tasten gewählt. Der Nutzer kann seine Hand dort ablegen und muss nicht in der Lage sein, schnell eine Taste zu finden. Die genauen Eigenschaften dieser Kategorien kann Tabelle 2 entnommen werden.

Zum erfolgreichen Bearbeiten einer Kategorie wird immer untersucht, ob zwei oder mehr Eigenschaften des Ereignisses mit den Angaben aus dieser Tabelle übereinstimmen. Im Falle, dass die drei Parameter des Ereignisses jeweils aus verschiedenen Kategorien stammen, wird die zweite Kategorie gewählt. Die Aufgabe bleibt damit einfach zu bearbeiten und ist dennoch komplex für eine Nebenaufgabe. Die genauen Parameter und Ähnliches wurden, wie bereits erwähnt, iterativ zusammen mit Testern ermittelt. Die Anleitung (Unterkapitel 3.4) wurde so gestaltet, dass sie dem Probanden genügend Möglichkeit gab, diese verstehen und anwenden zu können.

Zum besseren Verständnis folgt ein Beispielevent für die Nebenaufgabe:

Fußgänger	5	Bremsen

Tabelle 3: Beispiel eines Subtask-Events

Tabelle 3 zeigt ein Charakteristikum aus der Kategorie 1 und zwei aus der Kategorie 2. Die korrekte Aktion des Benutzers erfordert somit das Drücken der Taste 2.

Damit diese komplexen Regeln verstanden werden können, wurde eine Anleitung verfasst. Diese bietet zu allen Teilaufgaben der Anwendung interaktive Erklärungen der Regeln. Nutzer können sich so bereits mit den verschiedenen Mechaniken des Systems vertraut machen. Als Mechanik wird hier jede Interaktionsmöglichkeit des Nutzers mit dem System bezeichnet. Im Anschluss an die interaktive Sektion mit der Erläuterung aller Regeln wurde ein Testlauf beziehungsweise Tutorial gestartet. Bei der Einführungsrunde handelt es sich um eine voll funktionsfähige Version der Anwendung. Die Anzahl der zu erledigenden Events der Nebenaufgabe wird auf die gleiche Weise dargestellt wie in Durchlauf B. Dieser Durchlauf dauert drei Minuten und dient dazu, dass der Lerneffekt keinen oder weniger Einfluss auf das eigentlich Ergebnis der Studie hat.

3.5 Prototyp

Dieser Abschnitt beschreibt, welche Techniken bei der Entwicklung des Systems verwendet wurden. Es trägt somit nicht direkt zum Ergebnis der Studie bei und dient lediglich dazu, dass

möglichst viele Informationen für den Leser vorhanden sind.

3.5.1 Verwendete Technologien

Die verwendeten Technologien haben keinen direkten Einfluss auf die Ergebnisse und werden deswegen nicht ausführlicher aufgeführt.

Die Anwendung wird in Unity entwickelt und mithilfe von *WebGL* auf einer Website zur Verfügung gestellt. Die Seite wird in einem Azure WebApp Service freigegeben. Diese Technologien wurden gewählt, da ich in diesen Gebieten bereits Vorkenntnisse besitze. Der Code und die Daten der Auswertung sind unter https://github.com/prockUniBremen/Visualisierung-von-Warnsignalen zu finden.

3.5.2 Entwicklungsprozess

Wie bereits in den vorherigen Abschnitten (Unterkapitel 3.3, Unterkapitel 3.4) erwähnt wurde die Software mithilfe von Testern iterativ entwickelt. Das Grundprinzip, sich nach dem DMI zu richten, bot dabei das grundlegende Layout. Nachdem dieses nachgebildet wurde und die grundlegenden Mechaniken implementiert waren, begann die nächste Phase im Prozess. Als erster Tester habe ich selber mehrere Durchläufe absolviert und Notizen zu Problemen und Fehlern notiert. Nachdem ich mit dem Ergebnis zufrieden war, wurden externe Personen als Tester verwendet. Das Feedback dieser Nutzer war sehr hilfreich und notwendig für den Erfolg der Studie.

3.6 Zusätzliche Daten

Neben der Entwicklung des Systems mussten zusätzliche Daten von den Nutzern zur korrekten und besseren Auswertung erfasst werden. Außerdem wurde eine digitale Einverständniserklärung der Nutzer sowie ein subjektiver Test zur Aufzeichnung der Effektivität des Systems verlangt.

3.6.1 Einverständniserklärung

In vorangegangenen Kapiteln wurde bereits erwähnt, dass Entscheidungen getroffen wurden, die Effekte zwischen Personen mit Videospielerfahrung und ohne diese minimieren sollen. Damit diese und andere Faktoren ausgewertet können, wurden im Zusammenhang mit der Einverständniserklärung zusätzliche Daten aufgenommen. Dazu zählt das Alter, die Videospielerfahrung, der Besitz eines Führerscheins und ob Erfahrung mit dem Steuern von Zügen, auch virtuell in Simulatoren oder Ähnlichem, vorhanden ist.

Die Einverständniserklärung ist anonym und essentiell für eine ethisch korrekte Studie und wurde aufgrund dessen direkt am Anfang verlangt. Die Daten der verschiedenen Nutzer werden mit einzigartigen Identifikationsnummern gespeichert. Diese Nummer wird zufällig erzeugt und enthält keinerlei persönliche Informationen.

3.6.2 NASA TLX

”Der National Aeronautics and Space Administration-Task Load Index (NASA-TLX) [...] [ist] ein Instrument zur Erfassung der Arbeitsbelastung.” (Flägel et al. 2019). Die sechs verschiedenen Parameter, die aufgezeichnet werden, dienen in dieser Arbeit hauptsächlich der Überprüfung, ob die Aufgaben angemessen komplex waren. Dieser Test wurde auf Grundlage der Expertise meines Betreuers gewählt. Er sei ein sehr gängiger und adäquater Test, um die Belastung der Nutzer zu überprüfen. Die Ergebnisse werden in Kapitel 5 näher betrachtet. Im nächsten Kapitel wird jedoch zunächst erläutert, wie die Studie aufgebaut ist und welche Entscheidungen dazu geführt haben.

4 Studiendesign

Dieses Kapitel beschreibt, wie die Studie aufgebaut ist und wie sie durchgeführt wurde. Es werden die abhängigen und unabhängigen Variablen definiert und außerdem die fixen Variablen und Faktoren erläutert. Es gilt also die Fragen zu beantworten, was die Probanden durchführen müssen, wer an der Studie teilnehmen kann, wie Leistung definiert wird und welche Daten dafür erhoben werden.

4.1 Teilnehmer

Die Studie ist nicht dafür konzipiert, Experten bei Ihrer Arbeit zu evaluieren. Stattdessen ist jeder Laie in der Lage teilzunehmen. Diese Entscheidung wurde früh getroffen, da kein direkter Zugriff auf Domänenexperten zur Verfügung stand.

Es gibt Einschränkungen, die es nicht erlauben, dass jede Person an der Studie teilnehmen kann. Wie bereits erwähnt wurde die Anwendung auf Deutsch entwickelt und schließt alle Personen aus, die nicht über gute Deutschkenntnisse verfügen (Unterkapitel 3.2). Der technische Mehraufwand für die Unterstützung mehrerer Sprachen hätte den zeitlichen Rahmen überschritten. Darüber hinaus musste der teilnehmenden Person ein Computer mit Internet, Maus und Tastatur zur Verfügung stehen. Diese Entscheidung liegt dem Systemdesign zugrunde. Abgesehen von diesen Einschränkungen sollten keine großen Nutzergruppen von der Teilnahme ausgeschlossen sein.

Für die Teilnahme wurde eine Einverständniserklärung der Nutzer eingefordert. Jeder Teilnehmer, der bestimmte Informationen nicht teilen möchte, konnte nicht an der Studie teilnehmen. Außerdem war es den Teilnehmern immer möglich, die Anwendung zu schließen und ohne Konsequenzen aus der Umfrage auszusteigen. Ein Skript filtert bei der Auswertung später alle Ergebnisse heraus, die nicht das Ende erreicht haben. Dieses Vorgehen ist Standard bei jeder wissenschaftlichen Studie. Die Teilnehmer wurden per Mail oder über andere digitale Medien persönlich kontaktiert. Die Website ist für jeden Außenstehenden erreichbar und die Auswahl der Teilnehmer erfolgte dementsprechend nach der *Simple random sampling* Strategie (Wohlin et al. 2012). Da ich die Personen aber zum größten Teil persönlich kontaktiert habe, sind viele Probanden zum Beispiel aus einer ähnlichen Altersgruppe und demnach nicht repräsentativ für andere Gruppen. Mit mehr Zeit und einer größeren Bekanntmachung der Studie könnte eine

Gruppe von Probanden mit höherer Repräsentanz erreicht werden. An der Studie haben 19 Probanden teilgenommen. Damit trotz dieser geringen Teilnehmerzahlen kein Durchgang zu wenig Daten erhält, um daraus eine Schlussfolgerung zu ziehen, wird darauf geachtet, dass jeder Durchgang gleich viele Probanden erhält. Bei der Auswertung der demographischen Daten in Kapitel 5 wird dieser Aspekt erneut aufgegriffen.

4.2 Design

Beim grundsätzlichen Aufbau der Studie wurde sich an einem Standard-Design orientiert. Das *Paired comparison design* bietet die Möglichkeit, dass jeder Proband alle Durchgänge randomisiert anwendet (Wohlin et al. 2012). Das bedeutet, dass zufällig entschieden wird, welcher Durchlauf als erster durchlaufen wird. Einige Probanden werden zuerst Durchlauf A und andere Durchlauf B bearbeiten. Aufgrund der geringen Teilnehmerzahl kann so garantiert werden, dass genügend Daten zu den beiden verschiedenen Durchläufen gesammelt werden können. Eine spätere Auswertung der Differenzen zwischen den Durchgängen eines Probanden kann Aufschluss darüber geben, ob dieses Design einen Einfluss auf die Ergebnisse hat. Aufgrund des anstrengenden und frustrierenden Designs der Anwendung wurde außerdem davon abgesehen, mehr als zwei Durchläufe pro Proband durchzuführen.

Es wird zwei verschiedene Versionen oder Durchläufe geben. Diese werden sich, wie bereits in Kapitel 3 erwähnt, in der Menge der gleichzeitigen Subtasks unterscheiden. Es wurde festgelegt, dass es zwei verschiedene Durchläufe geben wird. Eine Version zeigt maximal eine Aufgabe an, während die andere bis zu vier ansammeln kann. Es gibt zwei verschiedene Versionen, damit jeder Proband für alle Durchläufe Daten sammeln kann, ohne zu stark zu ermüden oder frustriert zu werden. Die Anzahl der Subtasks pro Durchlauf wurde nach der größtmöglichen Differenz gewählt, um einen besonders starken Unterschied zu erzielen. Aufgrund des Layouts und der geringen Auflösung beziehungsweise Bildschirmgröße war die maximale Anzahl von Subtasks vier und dementsprechend wurde diese für den zweiten Durchlauf gewählt.

4.3 Aufbau

Aufgrund der aktuellen Pandemie wurde die Studie darauf ausgelegt, online durchgeführt zu werden. Die Menge der Daten, die damit erhoben werden kann, wird verringert, ist aber dennoch ausreichend zur Beantwortung einiger Fragen und Hypothesen. Auf weitere Probleme mit diesem Ansatz geht das Unterkapitel 4.7 weiter ein.

4.4 Messung der Daten

Während der Durchläufe werden automatisch eine Reihe an Daten erfasst. Allgemein wird die Leistung für diese Arbeit besonders nach der Sicherheit bewertet. Die Fehlerrate entscheidet also darüber, wie gut ein Proband abgeschnitten hat. Um diese zu messen sind in jedem Fall die Anzahl der bearbeiteten Aufgaben und ob sie richtig oder falsch bearbeitet wurden aufzuzeichnen.

In der Hauptaufgabe wird neben der bereits erwähnten Fehlerrate außerdem vermerkt, um welche Art von Ereignis es sich handelt. Dadurch kann später nachvollzogen werden, ob ein be-

stimmtes Event besser oder schlechter abgeschnitten hat. Dies sollte im Allgemeinen nicht der Fall sein. Über die notwendigen Daten zur Analyse der Sicherheit und Effizienz des Systems wurden weitere Daten erhoben. Diese dienen dem Zweck einer ausführlicheren Auswertung in Bezug auf Nebeneffekte, die durch diese Studie erfasst wurden. Dazu zählen unter anderem demographische Informationsangaben der Probanden. Diese haben, wie bereits erwähnt, Alter, Geschlecht, Besitz des Führerscheins und Erfahrung mit Zügen angegeben. Außerdem können die Daten aus dem NASA-TLX in der Auswertung verwendet werden. Die Anwendungsbereiche und Gründe für die Sammlung wurden bereits in vorangegangenen Kapiteln genannt (Unterkapitel 3.6.2).

4.5 Standpunkt

Diese Studie wird aus der Sicht von Forschern und nicht von Praktikern durchgeführt. Anstatt Experten in ihrer vertrauten Arbeitsumgebung zu testen oder eine Simulation mit realistischen Bedingungen zu erzeugen, wird hier allgemein überprüft, ob es einen signifikanten Unterschied zwischen den verschiedenen Durchläufen gibt. Der Grund für diese Entscheidung ist unter anderem der schwere Zugang zu Domänenexperten.

4.6 Abhängige und unabhängige Variablen

Die unabhängigen Variablen sind die Eingabeparameter des Experimentes. All diese sind im Kapitel 4 und Kapitel 3 bereits beschrieben worden. Die einzige Variable, die aktiv verändert wird, ist die Anzahl der Warnsignale. Alle anderen Parameter, wie beispielsweise die Farbgestaltung oder das Layout, werden nicht verändert. Damit soll sichergestellt werden, dass ein gemessener Unterschied zwischen den beiden Durchläufen an dem einen Parameter liegt, der verändert wurde.
Die abhängige Variable ist Ausgabeparameter der Studie und somit gleich dem Fokus oder Ziel dieser. Wie bereits erwähnt ist dies die Sicherheit des Systems (Unterkapitel 4.4). Aus der Sicht von Forschern gilt es hier nicht, in einer Simulation und unter echten Bedingungen zu überprüfen, welches System besser funktioniert, sondern allgemein herauszufinden, ob es einen signifikanten Unterschied bei verschiedenen Mengen an Subtasks gibt.

4.7 Unbeeinflussbare Faktoren

Zu den unbeeinflussbaren Faktoren gehören alle Parameter, die während des Experimentes nicht kontrolliert werden können. Einige dieser Faktoren wurden bereits in vorangegangenen Kapiteln erläutert und versucht, durch Designänderungen minimiert zu werden. Diese werden deswegen hier nicht erneut erwähnt.

Zusätzlich zu den bereits erwähnten gibt es allerdings weitere Faktoren, die im Folgenden kurz genannt und erläutert werden. Eine der offensichtlichsten ist dabei der Mensch selber. Jeder Proband hat unterschiedliche äußere Einflüsse erfahren, während er die Studie durchgeführt hat. Dadurch, dass diese vollständig online stattgefunden hat, konnten Störfaktoren von außerhalb, technische Einschränkungen wie eine schlechte Internetverbindung oder Ähnliches, nicht

kontrolliert werden. Darüber hinaus gibt es auch psychische Faktoren, die einen Einfluss auf die Ergebnisse jedes Probanden haben können.

Neben den eben genannten gibt es auch technische Probleme, die nicht vorausgesehen werden können. Die entwickelte Anwendung kann mögliche Bugs und kritische Fehler enthalten, die unabsehbar sind. Im Falle von fehlerhaften Daten wird der Datensatz komplett ignoriert. Im nächsten Kapitel wird die Auswertung der Ergebnisse stattfinden.

5 Auswertung

In diesem Kapitel werden die Ergebnisse der Datensammlung aufgezeigt und visualisiert. Auf Grundlage dieser wird im folgenden Kapitel 6 eine Hypothese zur Evaluation formuliert. An der Studie haben 19 Personen teilgenommen und ihre Daten zur Verfügung gestellt. Im Folgenden wird unter Betrachtung verschiedener Metriken erläutert, zu welchen Ergebnissen es jeweils kam.

5.1 Score

Der Score oder die Punktzahl eines Nutzers ermittelt sich aus der Anzahl der korrekt erledigten Aufgaben. Die Hauptaufgabe hat ungefähr doppelt so viele Events wie die Nebenaufgabe. Aufgrund dessen werden für sie doppelt so viele Punkte pro Event verteilt. Damit der Nutzer ein Erfolgserlebnis hat, wurden die Punkte jeweils mit fünf multipliziert. Für die Events der Hauptaufgabe werden also 10 und für die der Nebenaufgabe 5 Punkte verteilt. Der Score wirkt dadurch höher und soll dem Nutzer ein positiveres Feedback geben, ohne die Zahlen unnötig groß zu machen.

$$Score = Korrekt_{Hauptaufgabe} * 5 + Korrekt_{Nebenaufgabe} * 10 \quad (1)$$

Als erstes wird betrachtet, welchen Score die Teilnehmer in den verschiedenen Durchläufen erreicht haben. Hierbei werden die beiden Durchläufe und das Tutorial betrachtet, bei dem ebenfalls ein Score aufgezeichnet wurde. Der Box-Plot in Abbildung 3 weist folgende Charakteristika auf. Die horizontale Linie stellt den Median oder das 50% Perzentil dar. Die obere und untere Kante stellen jeweils das 75- und 25 Prozent Perzentil dar. Die von der Box ausgehenden vertikalen Linien sind definiert über einen Interquartilsabstand zwischen dem oberen und unteren Perzentil mal 1.5 plus oder minus das entsprechende Perzentil. Die *Whiskers* befinden sich immer am Ende dieser vertikalen Linie. Datenpunkte, die außerhalb des 1.5-fachen des Interquartilsabstandes liegen, werden als Punkt wie in Abbildung 5 dargestellt.

Wenn man die Durchläufe wie in Abbildung 3 chronologisch betrachtet, ergeben sich folgende Punktzahlen. Alle Angaben sind auf die Durchschnittswerte bezogen. Im Tutorial wurden 182.19 Punkte erreicht, im ersten Durchlauf 263.68 und im zweiten 265.92. In Abbildung 3 ist deutlich zu erkennen, dass zwischen dem Tutorial und den beiden Durchläufen ein größerer Unterschied besteht als zwischen diesen. Die Zunahme beträgt im Durchschnitt 82.61 Punkte.

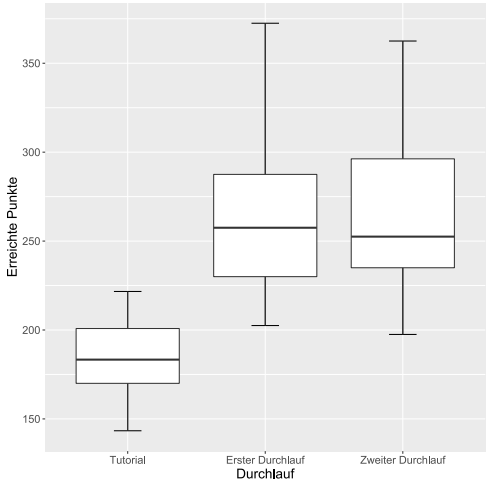

Abbildung 3: Erreichte Punktzahl pro Durchlauf (chronologisch)

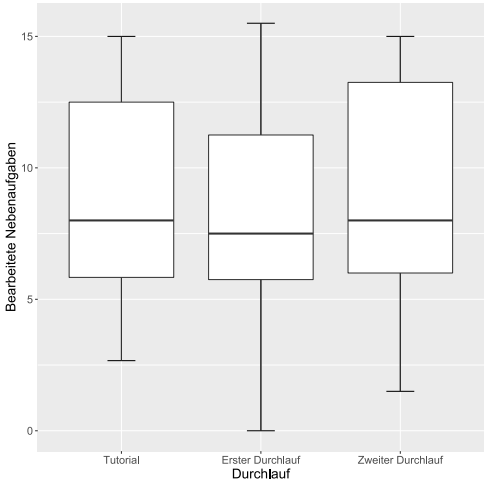

Abbildung 4: Bearbeitete Subtask-Events pro Durchlauf

Wenn man die Durchläufe nicht chronologisch sondern separat betrachtet und miteinander vergleicht, sind die durchschnittlichen Scores folgende. Im Tutorial wurden weiterhin 182.19 Punkte erreicht, während im Durchlauf A 258.42 und in Durchlauf B 271.18 Punkte erreicht wurden. Die Nebenaufgabe macht dabei durchschnittlich 0.78% der Gesamtpunkte jeder Runde aus. Die Wertebereiche reichen hier von 0.00% bis hin zu 0.75%.

5.2 Bearbeitungsrate

Die Bearbeitungsrate beschreibt die Menge der erledigten Nebenaufgaben eines Probanden. Da die Aufgaben theoretisch unmittelbar erledigt werden können, wird nicht berechnet, wie lange die Bearbeitung der einzelnen Aufgabe benötigt. Die Bearbeitungsrate ergibt sich stattdessen aus der Dauer eines Durchlaufes und der Gesamtzahl der bearbeiteten Events.

Abbildung 4 zeigt die Anzahl der bearbeiteten Subtask-Events für die beiden Durchläufe und die Einführungsrunde in chronologischer Reihenfolge. Die Unterschiede zwischen den Mittelwerten aller Durchläufe beträgt maximal 0.47. Wenn man die Unterschiede zwischen Durchlauf A und Durchlauf B betrachtet, anstatt in chronologischer Reihenfolge, ergibt sich eine Differenz von 5.00 im Mittelwert. Abbildung 9 zeigt die Unterschiede der beiden Durchläufe in Form eines Boxplots zur näheren Betrachtung. In Unterkapitel 6.2 wird das Ergebnis dieser Daten interpretiert.

5.3 Fehlerrate

Die Fehlerrate beschreibt die Anzahl der falsch interpretierten Nebenaufgaben der Benutzer im Verhältnis zur Gesamtzahl der bearbeiteten. Ein Fehler ist das Auswählen einer falschen Kategorie für die angegebenen Daten innerhalb des Subtask-Events. Daraus ergibt sich folgende

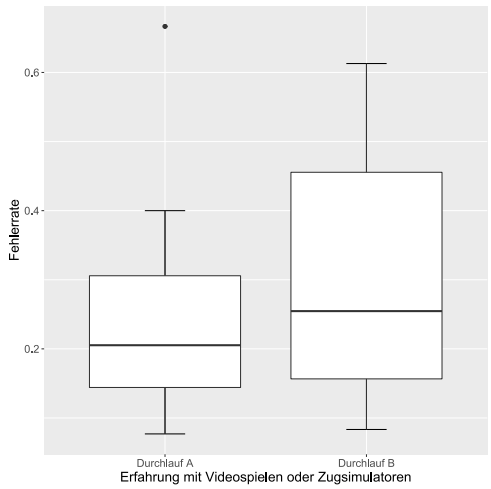

Abbildung 5: Fehlerrate pro Durchlauf

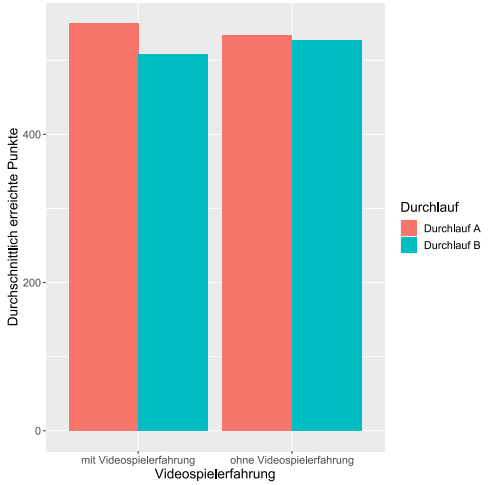

Abbildung 6: Erreichte Punkte im Verhältnis zur Videospielerfahrung gruppiert pro Durchlauf

Gleichung:

$$Fehlerrate = \frac{Fehler}{Bearbeitet_{Gesamt}} \tag{2}$$

Die Menge der Fehler ist in Abbildung 10 zu sehen. Die Fehlerraten der verschiedenen Durchläufe sind in Abbildung 5 visualisiert. In beiden Abbildungen ist zu erkennen, dass in Durchlauf B der Median höher ist als in Durchlauf A. Die Werte sind jeweils 0.31 und 0.26 für die beiden Durchläufe. Für die Anzahl der Fehler sind es jeweils 6.37 und 3.74. Eine genauere Analyse und Interpretation der Daten wird in Kapitel 6 vorgenommen.

5.4 Demographische Daten

Aufgrund des in Unterkapitel 4.1 beschriebenen Auswahlverfahrens haben die demographischen Daten keine hohe Varianz. Es wurden wenige Aspekte von den Nutzern abgefragt, sodass diese auch nicht weiter analysiert werden können. Das Alter der Teilnehmer liegt zwischen 20 und 31 mit einem Durchschnittsalter von 24.06 Jahren. Ein Proband hat einen inkorrekten Wert eingetragen und kann deswegen nicht betrachtet werden.

Es haben insgesamt 19 Probanden an der Studie teilgenommen, davon waren 12 männlich und 6 weiblich. Die Verteilung von männlichen zu weiblichen Nutzern liegt also jeweils bei 0.67 zu 0.33 (in Prozent). Von den weiblichen Teilnehmern hat keine angegeben, dass sie regelmäßig Videospiele spielt oder Erfahrung mit einem Zugsimulator hat. Bei den männlichen Teilnehmern konnten 10 angeben, dass sie Erfahrung mit Videospielen haben und davon wiederum 2, die darüber hinaus bereits Erfahrung mit Zugsimulatoren haben. Abbildung 6 zeigt die durchschnittlich erreichten Punkte aller Nutzer pro Durchlauf. Diese Ergebnisse sind unterteilt in die Gruppen mit und ohne Videospielerfahrung.

5.5 NASA-TLX

Der NASA-TLX wurde, wie bereits erwähnt, hauptsächlich dafür genutzt, um zu testen, ob die Probanden ausreichend von der Aufgabe gefordert werden. Dafür sind besonders die Werte für zeitliche Anforderung, geistige Anforderung und Frustration interessant. Diese haben Probanden im Durchschnitt mit jeweils 0.75%, 0.78 % und 0.56% angegeben. Die restlichen Durchschnittswerte waren körperliche Anforderung mit 0.17%, Anstrengung mit 0.67% und Leistung mit 0.46%. Inwiefern diese und alle anderen vorgestellten Ergebnisse mit dem übereinstimmen, was das System- und Studiendesign beabsichtigt hat, wird im nächsten Kapitel Analyse erläutert. Hier wird die Hypothese für die Analyse formuliert und die präsentierten Daten interpretiert.

6 Analyse

In diesem Kapitel wird die Hypothese für diese Arbeit formuliert und auf Grundlage der in Kapitel 5 präsentierten Daten beantwortet. Darüber hinaus werden Datensätze, die unabhängig von der Hypothese sind, aber einen interessanten Fund oder Ähnliches aufweisen, interpretiert.

6.1 Hypothese

Die Visualisierung von Warnsignalen ist eine wichtige Aufgabe beim Entwickeln einer GUI. Das System handhabt sicherheitsrelevante Daten und hat die Aufgabe, dem Nutzer besonders verständlich und effizient Signale zu übermitteln. Auf Grundlage dessen und der vorangegangen Ergebnisse aus dem Kapitel 5 wurde folgende Hypothese H_1 formuliert:

H_1 = *Die Fehlerrate beim Bearbeiten von Warnsignalen mit maximal einem Signal zur gleichen Zeit ist geringer als mit mehreren Signalen gleichzeitig.*

Zur Betrachtung dieses Problems wird die Nullhypothese H_0 formuliert. Diese geht davon aus, dass kein Unterschied in der Fehlerrate zwischen den beiden Durchläufen Durchlauf A und Durchlauf B existiert. Sollte H_0 abgelehnt werden, wird davon ausgegangen, dass die ursprüngliche Hypothese H_1 zutrifft. Die benötigten Daten für diese Analyse befinden sich auf der Verhältnisskala. Die Verteilung der Daten wird mithilfe des Shapiro-Wilk-Test durchgeführt (Royston 1982). Dieser setzt voraus, dass die Beobachtungen voneinander unabhängig sind und das zwischen 3 und 5000 Stichproben vorhanden sind. Da beide dieser Voraussetzungen erfüllt sind, kann er erfolgreich durchgeführt werden. Die Hypothese des Shapiro-Wilk-Tests ist, dass keine Normalverteilung bei den Daten vorliegt. Wird die Nullhypothese nun also akzeptiert, sind die Daten normalverteilt.

Mithilfe der Gleichung

$$W = \frac{(\sum_{i=1}^{n} \alpha_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{3}$$

kann der Wert der Teststatistik berechnet werden und mit dem kritischen Wert $W_{kritisch}$ verglichen werden. Das Signifikanzniveau wird allgemein auf 0.05 festgelegt, was einem $W_{kritisch}$

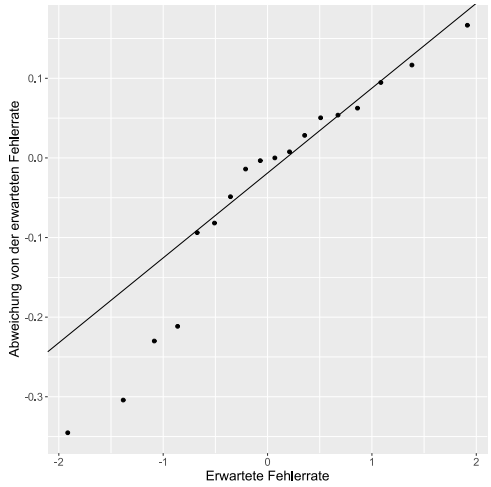

Abbildung 7: QQ-Plot der Fehlerraten in Durchlauf A und Durchlauf B

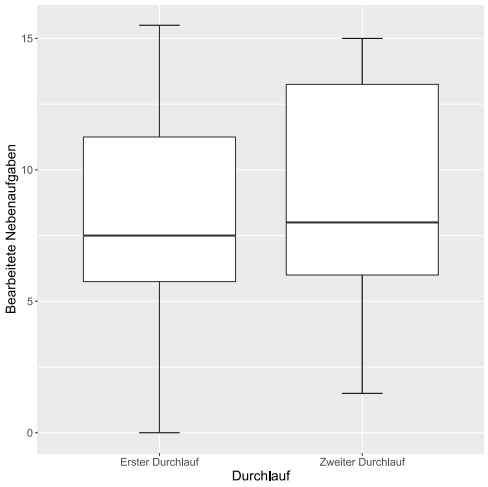

Abbildung 8: Anzahl bearbeiteter Nebenaufgaben pro Durchlauf (chronologisch)

von $0,842$ entspricht. Sollte dieser kleiner gleich W sein, wird die Nullhypothese abgelehnt. Die Anzahl der Beobachtungen sind der Wert n und deren Ergebnisse sind die Werte x_i. Die Gewichte α werden der Shapiro-Wilk-Tabelle entnommen. $\overline{x}$ beschreibt den Durchschnitt aller x. Der letzte fehlende Wert ist $x_{(i)}$, welcher die i-te geordnete Statistik für das jeweilige i angibt.

Das Ergebnis des Tests mit den Daten des Durchlauf A lautet $W = 0.83$ und für Durchlauf B $W = 0.91$. In einem Fall wird die Nullhypothese also abgelehnt und im anderen Fall akzeptiert. Die Daten befinden sich damit also insgesamt nicht in einer Normalverteilung. Werden die Daten logarithmisch transformiert, ergeben sich andere Werte für den Shapiro-Wilk-Test. Durchlauf A hat dann einen Wert von 0.96 und Durchlauf B 0.90. Mit so einer Transformation kann es laut Feng et al. zum Teil zu anderen Problemen kommen (2014).

Abbildung 7 zeigt einen QQ-Plot der Quantile der Fehlerraten von Durchlauf A und Durchlauf B. Wenn die beiden Datensätze eine ähnliche Verteilung haben, dann bilden die Punkte des Graphen eine Linie, die beinahe gerade ist. In Abbildung 7 ist bis auf kleinere Abweichungen eine Normalverteilung plausibel. Aufgrund des vertikalen Sprunges nach den ersten vier Datenpunkten ist dies weiterhin mit Bedacht zu betrachten.

Für das Paired comparison design gibt es drei typische Tests, die durchgeführt werden können. Der *Paired Students t-test* hat die größte statistische Macht oder auch Trennschärfe und vergleicht die Mittelwerte zweier unabhängiger Stichproben. Für die korrekte Anwendung wird eine Normalverteilung der Daten vorausgesetzt. Der *Wilcoxon-Vorzeichen-Rang-Test* ist ein nicht parametrisierter Test, der die Ränge zweier unabhängiger Stichproben vergleicht. Zuletzt gibt es noch den *Vorzeichentest*, welcher ähnlich wie Wilcoxon arbeitet, allerdings keine symmetrische Verteilung um den Mittelwert erwartet.

Trotz der nicht vorhandenen konkreten Normalverteilung wird ein Paired Students t-test durchgeführt (Student 1908). Das Ergebnis dieses Tests ist für 19 Teilnehmer ein Mittelwert von

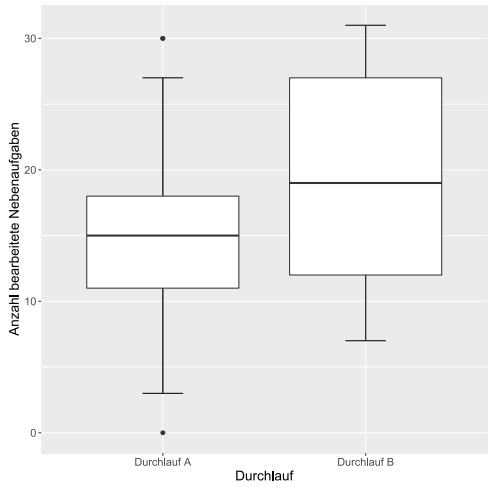

Abbildung 9: Bearbeitete Nebenaufgaben pro Durchlauf

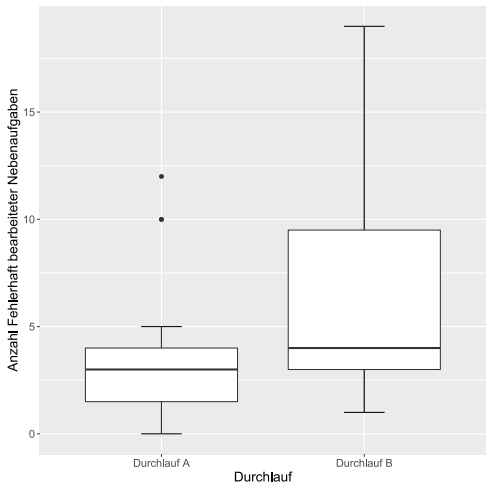

Abbildung 10: Anzahl Fehler pro Durchlauf

-0.04, eine Standardabweichung von 0.03 und einem p-Wert von 0.24. Da der p-Wert deutlich über dem bestimmten α von 0.05 liegt, wird die Nullhypothese nicht abgelehnt.

Der Wilcoxon Test verlangt den Daten keine Normalverteilung ab (Bauer 1972), benötigt aber trotzdem weitere Verarbeitung der Daten zur korrekten Verwendung. Es werden jeweils die Differenzenden der beiden Durchläufen auf einen neuen Vektor abgebildet, der dann für den Test verwendet wird. Es muss insbesondere darauf geachtet werden, dass zwei Wertepaare keine gleichen Werte haben, damit die Aussagekraft der Statistik weiterhin gegeben ist. Außerdem werden Wertepaare, deren Differenz null ist, bei der Berechnung verworfen, da sie keine weiteren Informationen geben. Gibt es viele solcher Fälle, sinkt die Aussagekraft ebenfalls ab. Beide Voraussetzungen wurden überprüft und stellen mit den Daten keine Probleme dar. Das Ergebnis des Wilcoxon Tests war ein p-Wert von 0.54. Die Nullhypothese wird damit wie beim T-Test bestätigt und die Hypothese H_1 wird abgelehnt.

Der Vorzeichentest wird aufgrund der vorangegangenen Ergebnisse nicht durchgeführt. Der auf Abbildung 5 zu erkennende Unterschied zwischen den beiden Durchläufen führt aufgrund von hoher Varianz oder der geringen Teilnehmerzahl also nicht zu einem signifikanten Ergebnis. Die Hypothese H_1, dass die Menge an parallelen Informationen einen Einfluss auf die Fehlerrate hat, wird abgelehnt.

6.2 Weiterführende Analyse

Wie bereits erwähnt spielt der Sicherheitsaspekt bei einer Anwendung zur Visualisierung von Warnsignalen eine signifikante Rolle. Darüber hinaus gibt es auch Interesse daran, die komplexen Signale besonders effizient zu verarbeiten. Im Folgenden soll genauer untersucht werden, ob es einen signifikanten Unterschied bei der Effizienz der beiden Durchläufe gibt. Dafür werden die vorangegangenen Daten zur Fehlerrate mit der allgemeinen Bearbeitungsrate (Anzahl der bearbeiteten Aufgaben pro Durchlauf) ins Verhältnis gesetzt.

Abbildung 9 zeigt, dass ein Unterschied zwischen den beiden Durchläufen in der Menge der bearbeiteten Aufgaben besteht. Mithilfe des T-Tests soll überprüft werden, ob dieser signifikant ist. Die Hypothese lautet in diesem Fall:

H_2 = *Die Effizienz beim Bearbeiten von Warnsignalen steigt, wenn eine größere Menge an Informationen gleichzeitig zur Verfügung steht.*

Um diese Hypothese zu beantworten wird genau wie in Unterkapitel 6.1 vorgegangen, indem die Nullhypothese in einem Signifikanztest untersucht wird. Der Shapiro-Wilk-Test hat für die Daten aus Durchlauf A für W = 0.95 ergeben und für Durchlauf B lag W bei 0.89. Diese Werte lassen vermuten, dass eine Normalverteilung vorliegt, da die Nullhypothese in beiden Fällen verworfen werden kann.

Mit den gleichen Daten wird nun ein T-Test durchgeführt. Die Ergebnisse sind ein Mittelwert von -5.00, Standardabweichung von 1.91 und ein p-Wert von 0.02. Der p-Wert liegt damit deutlich unter dem gewählten α von 0.05. Damit kann die Nullhypothese verworfen werden. Wenn die Hypothese H_2 wahr wäre, ist es wahrscheinlich, ähnliche Datensätze wie in dieser Studie wiederzufinden. Die Effizienz beim Bearbeiten von Warnsignalen steigt mit diesen Voraussetzungen signifikant, wenn eine größere Menge an Informationen gleichzeitig zur Verfügung steht.

Zum Vergleich werden noch einmal die Werte aus den chronologischen Stichproben miteinander verglichen. Für sie werden wieder die gleichen Bedingungen verwendet. Die Abbildung 8 zeigt einen Boxplot der Daten zur Veranschaulichung. Die Werten für den Shapiro-Test waren 0.96 für den ersten und 0.91 für den zweiten Durchlauf. Der p-Wert des T-Tests für diese Daten ergab 0.73. Die Nullhypothese könnte in diesem Fall nicht verworfen werden. Daraus lässt sich schließen, dass die vorangegangene Analyse valide ist.

Trotz der Tatsache, dass es keinen signifikanten Unterschied bei der Fehlerrate der Probanden gibt, ist ein signifikanter Unterschied zwischen den beiden Durchläufen erkennbar. Wenn die angegebene Hypothese wahr wäre, dann ist es sehr wahrscheinlich, die Daten aus dieser Studie wiederzufinden. Allgemein wird der Sicherheitsaspekt in keinem Fall verletzt, dennoch kann die Effizienz gesteigert werden, indem eine Mehrzahl an Informationen gleichzeitig zur Verfügung gestellt wird.

6.3 Lernbarkeit

Neben den Ergebnissen zur Analyse der Hypothese wurden im vorangegangenen Kapitel (Kapitel 5) zusätzliche Ergebnisse präsentiert. Die Interpretation findet im Folgenden ohne Untersuchung mit Signifikanztests statt.

Unter dem Aspekt der Lernbarkeit wurden Daten zur Bearbeitungsrate (Unterkapitel 5.2) und zur erreichten Punktzahl genannt (Unterkapitel 5.1). Die Annahme war, dass die Kombination aus ausführlicher, interaktiver Anleitung und einer Einführungsrunde für die Anwendung ausreichend zum Erlernen der Mechaniken ist. In Abbildung 3 ist gut zu erkennen, dass ein deutlicher Unterschied zwischen der Einführungsrunde und dem ersten Durchlauf vorhanden ist. Im darauffolgenden Durchlauf ist der Unterschied hingegen minimal. Neben dem deutlich unterschiedlichen Score ist in Abbildung 4 zu sehen, dass dies voraussichtlich nicht an der

Menge der bearbeiteten Subtask-Events lag sondern an der Fähigkeit des Nutzers. Die Diskrepanz der Punkte zwischen dem Tutorial und dem Durchlauf A ist 81.49 und 2.24 bei Durchlauf A und Durchlauf B. Die durchschnittliche Anzahl bearbeiteter Subtask-Events ist jeweils 0.47 und 0.39. Es wurden also deutlich mehr Punkte erreicht, obwohl von den meisten Nutzern kaum mehr Subtask-Events bearbeitet wurden. Diese Ergebnisse lassen darauf schließen, dass die Herangehensweise angebracht war.

6.4 Kritik / Probleme

Die Ergebnisse des NASA-TLX lassen darauf schließen, dass die Anwendung möglicherweise zu frustrierend gestaltet wurde. Während die geistige Anforderung allgemein mit 0.78 sehr hoch bewertet wurde, gibt es vor allem bei der Frustration einige Teilnehmer, die sehr hohe Werte angegeben haben. Eine Person hat nach dem Versuch erwähnt, dass sie die Nebenaufgabe nicht verstanden hat und einige Aufgaben zufällig erledigt hat. Eine andere Person hat in einem der Durchläufe keine Nebenaufgabe bearbeitet. Diese Ergebnisse lassen darauf schließen, dass die Frustration zu hoch oder das Tutorial nicht ausführlich genug war. Aufgrund der allgemeinen Ergebnisse lässt sich hier aber nicht direkt schließen, dass viele Ergebnisse invalide sind. So gibt es auch Probanden, die sogar Spaß an der Studie hatten, obwohl dies kein Ziel der Anwendung war.

Bei allen Teilnehmern ist eine mögliche Erschöpfung nach dem ersten Durchlauf festzustellen gewesen. Abbildung 3 zeigt, dass nach dem zweiten Durchlauf bereits eine leichte Abnahme der Punkte festzustellen ist. Im Allgemeinen wäre zu erwarten, dass bei einer Aufgabe die erreichten Punkte der Teilnehmer vorerst stetig steigen. Ich nehme an, dass aufgrund der Frustration und der hohen zeitlichen und geistigen Anforderung dies nicht der Fall ist und anstatt dessen in weiteren Durchläufen keine Steigerung zu erwarten ist. Anders wäre es hier, wenn die Teilnehmer erneut an der Studie teilnehmen würden.

Die demographischen Daten deuten darauf hin, dass eine geringe Transferierbarkeit der Daten vorliegt. Die Probanden sind etwa zwischen 20 und 30 Jahren alt und leben vermutlich in Deutschland. Da viel persönlicher Kontakt zu den Probanden aufgenommen wurde, wurde sich möglicherweise auch mehr Mühe gegeben als es unter anderen Umständen der Fall gewesen wäre. Die Teilnehmer mit Videospielerfahrung, die hier nur bei Männern vorhanden war, konnten im Durchschnitt nicht mehr Punkte erreichen (siehe Abbildung 6). Es waren also männliche Nutzer mit und weibliche ohne Videospielerfahrung an der Studie beteiligt. Da die Unterschiede aber nicht zu stark sind, wurden hier keine weiteren Schlüsse gezogen. Die Transferierbarkeit besonders auf andere Altersgruppen ist nicht gegeben und im Allgemeinen wird erwartet, dass weder ein Unterschied zwischen Männern und Frauen noch zwischen Personen mit und ohne Videospielerfahrung vorliegt.

6.5 Diskussion

Zur Analyse der Sicherheit und Effizienz von verschiedenen Visualisierungen von Warnsignalen wurde ein spezielles System entwickelt. Dieses war in der Lage, Kriterien wie Fehlerrate, Bearbeitungsrate und andere Faktoren zu messen. Probanden mussten in diesem kontrollierten Experiment zwei verschiedene Versionen oder Durchläufe der sonst gleichen Anwendung

durchlaufen. Es handelt sich hierbei um eine Dual-Task-Studie, die eine Haupt- und Nebenaufgabe im Kontext der Zugführung beinhaltet. Die Hauptaufgabe ist an das Führen eines Zuges angelehnt, während die Nebenaufgabe Warnsignale visualisiert, auf die es zu reagieren gilt. In einem Durchlauf wurde maximal eine Nebenaufgabe zur gleichen Zeit angezeigt, während bei dem anderen bis zu vier auf einmal angezeigt werden.

Eine Erhöhung der Sicherheit durch eine Veränderung dieser Eigenschaft konnte nicht festgestellt werden. Mit einem p-Wert von 0.24 im durchgeführten Student T-Test wurde die Hypothese abgelehnt. Daraus wird gefolgert, dass die Menge an gleichzeitigen Warnsignalen innerhalb eines angebrachten Rahmens, wie in Unterkapitel 3.4 beschrieben, keine Auswirkung auf die Sicherheit des Systems hat. Für Ansätze, die deutlich extremere Werte annehmen, kann hier keine Vermutung aufgestellt werden.

Die Ergebnisse der Bearbeitungsrate zeigen, dass verschiedene Mengen an gleichzeitigen Signalen zu unterschiedlicher Leistung der Probanden geführt hat. Der T-Test ergab hier einen Wert von 0.02 und damit weniger als der gewählte α-Wert von 0.05. Unter der Annahme, dass die Aussage: „Die Effizienz beim Bearbeiten von Warnsignalen steigt, wenn eine größere Menge an Informationen gleichzeitig zur Verfügung steht“ wahr ist, ist es wahrscheinlich, die Daten dieser Studie zu erhalten. Diese beiden Funde sind für die Entwicklung eines sicherheitskritischen Systems interessant. Es muss hierbei beachtet werden, dass im Tutorial die Anzahl der zu erledigenden Events der Nebenaufgabe auf die gleiche Weise dargestellt wurde, wie in Durchlauf B. Die Probanden haben also insgesamt mehr Zeit mit der Darstellungsweise von Durchlauf B verbracht und dadurch gegebenenfalls einen Lerneffekt erfahren. Inwiefern dies eine Auswirkung auf das Ergebnis hatte, ist nicht abzusehen.

Wichtig ist zu bedenken, dass in der Studie versucht wurde, sich nahe an dem DMI (Abbildung 1) zu orientieren. Für die Entwicklung von Systemen außerhalb dieser Domäne können andere Ergebnisse erwartet werden. Da dieses System aber hauptsächlich darauf ausgelegt war, Stress und Frustration bei den Nutzern hervorzurufen, gehe ich nicht davon aus, dass Systeme mit ähnlichen Voraussetzungen andere Ergebnisse erzielen, nur weil sie sich in einem anderen Kontext befinden. Dahingegen wichtiger ist die Betrachtung der Transferierbarkeit dieser Arbeit. Aufgrund der demographischen Daten (erläutert in Unterkapitel 5.4 und Unterkapitel 6.4) ist diese womöglich deutlich auf junge Menschen zwischen 20 und 30 Jahre, die voraussichtlich Erfahrung mit Technik oder sogar Videospielen haben, beschränkt. Im nächsten und letzten Kapitel werde ich ein persönliches Fazit aus der Studie ziehen und außerdem einen Ausblick über mögliche Anhaltspunkte für weitere Arbeiten in diesem Bereich geben.

7 Fazit und Ausblick

In diesem Kapitel betrachte ich rückwirkend die Ergebnisse der Auswertung und Evaluation der vorliegenden Arbeit. In einem persönlichen Fazit reflektiere ich Funde und Probleme der Studie und im darauffolgenden Ausblick sollen mögliche Lösungsansätze und weitere Forschungsfragen erläutert werden.

7.1 Fazit

Die Visualisierung von Warnsignalen ist ein Thema, mit der sich viele Studien auseinandersetzen. Eine Vielzahl der öffentlich verfügbaren Forschung in diesem Bereich kommt aus der Automobilbranche. In neueren Studien wird dabei häufig untersucht, wie die Auswirkung von Modalitäten, intelligenten Interfaces oder die Darstellungsform unter Einbezug von präattentiven Attributen Einfluss auf die Leistung eines Nutzers hat. Abgesehen von der Arbeit am Redundant Signal Effect (RSE) konnte ich wenig bis keine Informationen darüber finden, ob die Menge paralleler Signale einen Einfluss auf die Leistung eines Nutzers hat. Darüber hinaus gab es wenig öffentlich zugängliche Studien für den Lokomotivsektor. Deswegen ergab sich die Frage zur Untersuchung nach genau diesen beiden Kriterien: eine Studie zur Auswirkung der Menge paralleler Signale innerhalb der Zugdomäne.

Zur Untersuchung dieses Problems habe ich mich an Studien aus dem Bereich der Automobilbranche orientiert. Das Design des Systems wurde an das eines DMI angelehnt. Faktoren, die untersucht wurden, beziehen sich dabei besonders auf sicherheitsrelevante Bereiche, da die Führung eines Fahrzeuges immer in Bezug auf Sicherheit optimiert werden sollte.

Die Ergebnisse aus der Auswertung und Analyse haben in dieser Studie mit 19 Teilnehmern gezeigt, dass eine Erhöhung der Sicherheit unwahrscheinlich ist. Allerdings konnte festgestellt werden, dass ohne Erhöhung der Fehlerrate eines Nutzers die Bearbeitungsrate mit einer höheren Menge an parallelen Signalen zur Verarbeitung verbessert werden kann. Bei der Entwicklung eines sicherheitskritischen Systems kann diese Information sinnvoll sein, um die Leistung des Nutzers zu erhöhen.

Trotz der geringen Übertragbarkeit in andere Nutzergruppen präsentiert die vorliegende Arbeit ein interessantes Ergebnis und bietet ein solides Fundament für zukünftige Forschung in diesem Bereich. Die Funde sind aussagekräftig genug, um weitere Studien in diese Richtung zu begründen. Ich erwarte, dass eine Untersuchung anderer Altersgruppen andere Funde mit sich bringen würde, die Verwendung bei der Entwicklung von GUIs haben könnten.

Diese gehaltvollen Ergebnisse sind durch die Orientierung an anderen Studien und die Nähe zum Driver Machine Interface (DMI) erreicht worden. Eine genaue Planung des Systems und ausreichend Feedback von Testern ermöglichte diese erfolgreiche Studie. Die Entscheidung, eine interaktive Anleitung und eine Einführungsrunde zu implementieren, stellte sich als sehr hilfreich heraus. Auch die Verwendung von wenigen, aber eindeutigen Farben zum besseren Verständnis der Anwendung zusammen mit den sehr großen Symbolen und einfacher Bedienung mit wenigen Tasten, war entscheidend für den Erfolg. Doch neben den positiven Ansätzen für die Studie gibt es auch Bereiche, in denen Verbesserungen vorgenommen werden können. Trotz der Bemühung, eine verständliche Anleitung zu erstellen, gab es mindestens einen Nutzer, der die Nebenaufgabe nicht korrekt verstanden hat und aufgrund dessen zufällige Aktionen vorgenommen hat. Die Komplexität der Aufgabe könnte im Allgemeinen etwas vereinfacht werden, sodass die Einführungszeit bis hin zum ersten Durchlauf geringer gestaltet werden kann und die Motivation der Teilnehmer nicht bereits gesunken ist.

7.2 Ausblick

Wie bereits in Unterkapitel 6.4 erwähnt wird vermutet, dass eine Erschöpfung der Teilnehmer bereits nach dem zweiten Durchlauf zu erkennen beziehungsweise zu erwarten ist. In weiteren Untersuchungen könnte betrachtet werden, inwiefern die Menge paralleler Signale einen Einfluss auf die Erschöpfung hat. Ich vermute, dass aufgrund der höheren Bearbeitungsrate und Stimulanz des Nutzers bei größeren Mengen früher eine Erschöpfung eintritt.

Trotz der Ergebnisse der Studie von Horrey et al., dass die Fokussierung auf eine bestimmte Aufgabe keinen signifikanten Einfluss auf das Ergebnis hat, können hier weitere Untersuchungen vorgenommen werden (2009). Da es sich hier um Warnsignale handelt, kann das Hinweisen auf die Wichtigkeit einen anderen Einfluss haben. Im gleichen Zusammenhang kann negatives Feedback zum ausgeführten Verhalten ebenfalls Einfluss auf die Leistung der Nutzer haben. In einem Simulator zum Beispiel können Gefahren akkurater dargestellt werden. Die Konzentration der Benutzer auf diese Signale hat womöglich einen anderen Einfluss als den Hinweis zur Fokussierung auf eine bestimmte Aufgabe. Meine Annahme ist, dass aufgrund der stärkeren Auswirkung auf die Punkte oder die bestehende Gefahr in einem Simulator eine allgemein höhere Bearbeitungsrate zu erwarten ist. Inwiefern es Auswirkungen auf die Fehlerrate gibt, vermag ich nicht abzuschätzen. Wie bereits in Entwicklung der Nebenaufgabe und Diskussion erwähnt, wurde im Tutorial die gleiche Version wie in Durchlauf B ausgeführt. Aufgrund dessen wäre es interessant zu sehen, ob ein Wechsel zu Durchlauf A im Tutorial einen Unterschied auf die Ergebnisse hätte. Der hohe Lerneffekt des ersten Durchlaufes lässt aber vermuten, dass dieser keine direkte Auswirkung auf die Leistung der anderen Durchläufe hat.

Zur besseren Analyse einiger der betrachteten Probleme würde es sich anbieten, den NASA-TLX nach jedem erfolgreichem Durchlauf durchzuführen, um mehr Daten zur Anstrengung und Ähnlichem zu erhalten. Eine andere Maßnahme, die bei einer vergleichbaren Arbeit vorgenommen werden sollte, ist der konkrete Hinweis auf die Unterschiede zwischen den beiden Durchläufen. Diese Information muss keine direkte Auswirkung auf das Ergebnis haben, sollte aber dennoch für die Teilnehmer verfügbar sein.

Allgemein bietet die Arbeit eine gute Grundlage für die weitere Forschung in diesem Gebiet. Die aufgezeigten Probleme, Lösungsansätze und mögliche Änderungen können dazu dienen, zusätzliche Funde zu erzielen. Besonders sicherheitsrelevante Aspekte sollten in zukünftigen Arbeiten weiter untersucht werden. Die Hilfe von Domänenexperten zur Untersuchung der gefundenen Ergebnisse im realen Kontext ist darüber hinaus ausschlaggebend für die Verwendung der Funde in realen Anwendungen.

Literatur

Ahmadi, M., Pourhosein Gilakjani, A., und Ahmadi, S. (2011). The Relationship between Attention and Consciousness. *Journal of Language Teaching and Research*, 2.

Anderie, L. (2016). *Definition und Abgrenzung von Begrifflichkeiten und Märkten*, Seiten 21–22. Springer Berlin Heidelberg, Berlin, Heidelberg.

Baranyi, P. und Csapo, A. (2010). Cognitive infocommunications: CogInfoCom. In *2010 11th International Symposium on Computational Intelligence and Informatics (CINTI)*, Seiten 141–146.

Bauer, D. F. (1972). Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, 67(339):687–690.

Biondi, F., Strayer, D. L., Rossi, R., Gastaldi, M., und Mulatti, C. (2017a). Advanced driver assistance systems: Using multimodal redundant warnings to enhance road safety. *Applied Ergonomics*, 58:238 – 244.

Biondi, F., Strayer, D. L., Rossi, R., Gastaldi, M., und Mulatti, C. (2017b). Advanced driver assistance systems: Using multimodal redundant warnings to enhance road safety. *Applied Ergonomics*, 58:238 – 244.

Brumby, D., Salvucci, D., und Howes, A. (2007). An Empirical Investigation into Dual-Task Trade-offs while Driving and Dialing.

de Vries, M. (2016). Grundsätzlicher Aufbau des Driver Machine Interface (DMI) von ETCS. Online https://de.wikipedia.org/wiki/Driver_Machine_Interface#/media/Datei:DMI_ETCS-areas.svg; Accessed: 2020-06-06.

Feng, C., Hongyue, W., Lu, N., Chen, T., He, H., Lu, Y., und Tu, X. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, 26:105–9.

Few, S. (2004). Tapping the Power of Visual Perception. *Intelligent Enterprise*.

Flägel, K., Galler, B., Steinhäuser, J., und Götz, K. (2019). Der „National Aeronautics and Space Administration-Task Load Index“ (NASA-TLX) – ein Instrument zur Erfassung der Arbeitsbelastung in der hausärztlichen Sprechstunde: Bestimmung der psychometrischen Eigenschaften. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, 147-148:90 – 96.

Green, C. und Bavelier, D. (2006). Enumeration versus multiple object tracking: the case of action video game players. *Cognition*, 101(1):217—245.

Horowitz, A. D. und Dingus, T. A. (1992). Warning Signal Design: A Key Human Factors Issue in an In-Vehicle Front-To-Rear-End Collision Warning System. *Proceedings of the Human Factors Society Annual Meeting*, 36(13):1011–1013.

Horrey, W., Lesch, M., und Garabet, A. (2009). Dissociation between driving performance and drivers' subjective estimates of performance and workload in dual-task conditions. *Journal of Safety Research*, 40(1):7 – 12.

Izullah, F. R., Koivisto, M., Aho, A., Laine, T., Hämäläinen, H., Qvist, P., Peltola, A., Pitkäkangas, P., und Luimula, M. (2016). NeuroCar virtual driving environment: Simultaneous evaluation of driving skills and spatial perceptual-attentional capacity. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Seiten 000031–000036.

Jeon, M., Davison, B. K., Nees, M. A., Wilson, J., und Walker, B. N. (2009). Enhanced Auditory Menu Cues Improve Dual Task Performance and Are Preferred with In-Vehicle Technologies. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '09, Seite 91–98, New York, NY, USA. Association for Computing Machinery.

Jämsä, J. und Kaartinen, H. (2015). Adaptive user interface for assisting the drivers' decision making. In *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Seiten 17–18.

Lee, J.-H. und Spence, C. (2008). Assessing the Benefits of Multimodal Feedback on Dual-Task Performance under Demanding Conditions. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1*, BCS-HCI '08, Seite 185–192, Swindon, GBR. BCS Learning & Development Ltd.

Levy, J. und Pashler, H. (2008). Task prioritisation in multitasking during driving: opportunity to abort a concurrent task does not insulate braking responses from dual-task slowing. *Applied Cognitive Psychology*, 22(4):507–525.

Luke, T., Brook-Carter, N., Parkes, A., Grimes, E., und Mills, A. (2006). An investigation of train driver visual strategies. *Cognition, Technology & Work*, 8:15–29.

McLeod, R. W., Walker, G. H., und Moray, N. (2005). Analysing and modelling train driver performance. *Applied Ergonomics*, 36(6):671 – 680. Special Issue: Rail Human Factors.

Murata, A., Kanbayashi, M., und Hayami, T. (2013). Effectiveness of Automotive Warning System Presented with Multiple Sensory Modalities. In Duffy, V. G., Herausgeber, *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management. Healthcare and Safety of the Environment and Transport*, Seiten 88–97, Berlin, Heidelberg. Springer Berlin Heidelberg.

Prekopcsák, Z. und Nagy, G. I. (2011). Effective sensor-bridging with visual preattentive features. In *2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, Seiten 1–3.

Reynolds, S., Masson, C., und Barrett, P. (2017). RDG Guidance Note ETCS On-Board Equipment. *RDG GN/NTI/005*, 1.

Riby, L., Perfect, T., und Stollery, B. (2004). The effects of age and task domain on dual task performance: A meta-analysis. *European Journal of Cognitive Psychology*, 16(6):863–891.

Royston, J. P. (1982). Algorithm AS 181: The W Test for Normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2):176–180.

Strayer, D. L. und Johnston, W. A. (2001). Driven to Distraction: Dual-Task Studies of Simulated

Driving and Conversing on a Cellular Telephone. *Psychological Science*, 12(6):462–466. PMID: 11760132.

Student (1908). The Probable Error of a Mean. *Biometrika*, 6(1):1–25.

Torok, A. (2016). From human-computer interaction to cognitive infocommunications: A cognitive science perspective. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Seiten 000433–000438.

Velichkovsky, B. M., Rothert, A., Kopf, M., Dornhöfer, S. M., und Joos, M. (2002). Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, 5(2):145 – 156.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., und Dunnington, D. (1993). Aesthetic specifications. Online; Accessed: 2020-08-17.

Wohlin, C., Runeson, P., Hst, M., Ohlsson, M. C., Regnell, B., und Wessln, A. (2012). *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated.

Yan, X., Zhang, Y., und Ma, L. (2015). The influence of in-vehicle speech warning timing on drivers' collision avoidance performance at signalized intersections. *Transportation Research Part C: Emerging Technologies*, 51:231 – 242.