Usability Studie – Fahrplanauskunft bei DB-Navigator und Webseite der Deutschen Bahn AG

Einzelleistung

im Studiengang
Medienwissenschaft, interdisziplinär

vorgelegt von

Kai Peters

am 30. September 2014
an der Universität Bielefeld

Abstract

Im Rahmen des Seminars „Der Nutzer im Fokus – Evaluation von User Interfaces“ an der Universität Bielefeld wurde eine Studie bezüglich der Usability von Online Produkten der Deutschen Bahn AG durchgeführt. Hierbei lag das Hauptaugenmerk auf der Fahrplanauskunft und Buchung mittels der Webseite und der App (DB-Navigator).

24 Probanden wurde dabei eine Aufgabe gestellt, die das Suchen nach einer bestimmten Verbindung und der gleichzeitigen Buchung beinhaltete. Hierbei kamen die Messmethoden Time-on-Task, Task Success sowie der SUS-Fragebogen zum Einsatz. Außerdem wurde der Zusammenhang zwischen dem Vorwissen der Testpersonen und den Ergebnissen untersucht. Schließlich wurden noch individuelle Beobachtungen in der Analyse berücksichtigt.

Zur statistischen Auswertung wurden Verfahren aus der Deskriptiven Statistik, der abhängige T-Test sowie die Korrelationsanalyse nach Spearman verwendet.

Dabei stellte sich heraus, dass es einen signifikanten, wenn auch schwachen, Unterschied der beiden Produkte hinsichtlich der Usability zu Gunsten der Webseite gibt. Mit verhältnismäßig einfachen Änderungen des App-Interfaces wäre es allerdings vorstellbar, dass diese Lücke geschlossen werden könnte.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

Die hier vorgelegte Studie befasst sich mit der Usability der beiden Online-Anwendungen der Deutschen Bahn AG (DB) zur Fahrplanauskunft. Dies ist zum einen die native App („DB-Navigator“) und zum anderen die Webseite www.bahn.de.

Usability Tests von Online Fahrplanauskünften sind beliebte Themen in Studien oder journalistischen Beiträgen. So setzte sich beispielsweise das Usability-Online Magazin des Frauenhofer Instituts mit den Online Auskünften regionaler Nahverkehrsbunde auseinander. Dabei wurden Usability-Probleme bezüglich des Buchungsprozesses sowie in unverständlicher Button-Gestaltung und Preisangaben festgestellt (Werhahn, 2008).

Auch die Online-Anwendungen der Deutschen Bahn AG waren bereits Gegenstand von Usability Studien. Die Markt- und Medienforschungsagentur Mediascore beschrieb bei einer früheren Version des DB-Navigators Probleme, die gleichzeitige Buchung von Hin- und Rückfahrt sowie der Preisauskunft betreffend (Reimamann & Kluge, 2012).

Auch der Webseite der DB wurden teilweise Mängel bezüglich der Button-Gestaltung oder Namensgebung von Links und Buttons innerhalb der Fahrplanauskunft und des Buchungsprozesses bescheinigt (Kutter, 2011).

Die hier vorliegende Studie widmet sich hauptsächlich dem Vergleich von App und Webseite bezogen auf die Fahrplanauskunft und Buchung von Verbindungen. Im Vordergrund der Untersuchung steht hierbei die Zeit, die die Probanden bezüglich einer bestimmten Aufgabe benötigen. Es wird erwartet, dass im Ergebnis der Webseite der höhere Usability zugeordnet werden kann, da hier die Möglichkeiten der Interfacegestaltung weitaus größer gegenüber nativen Apps sind.

Da neben dem Zeitaufwand auch andere Faktoren für die Usability Bewertung eine Rolle spielen, sollen zusätzlich zur Zeitmessung ebenfalls die Werte weitere Messmethoden zur Beurteilung der Usability einfließen. Dies sind die Werte des Task Success sowie das Ergebnis der genutzten SUS-Fragebögen.

1 Methode

1.1 Design

1.1.1 Forschungsfrage

Das Aufstellen einer Forschungsfrage legt den Grundstein einer Studie. Sie legt fest, welche Erkenntnisse die Studie liefern soll. Dabei sollte darauf geachtet werden, dass die Forschungsfrage auf eine bestimmte Antwort abzielt und daher nicht zu allgemein gehalten wird (Field & Hole, 2003).

Wie eingangs bereits beschrieben, soll die Usability von Webseite und App bezüglich der Fahrplanauskunft der Deutschen Bahn vornehmlich mittels des Zeitaufwands gemessen werden, der für eine Verbindungsrecherche aufgewendet werden muss. Daraus ergibt sich folgende Forschungsfrage:

„Mit welchem User Interface der Deutschen Bahn zur Fahrplanauskunft und zum Ticketkauf (Web, App) können Nutzer am schnellsten eine Route und das zugehörige Ticket heraussuchen?“

Anhand dieser Forschungsfrage lassen sich einfach abhängige und unabhängige Variablen ableiten.

1.1.2 Unabhängige und abhängige Variablen

Nach Tullis & Albert [2013] beschreiben unabhängigen Variablen einer Studie den Faktor des Experiments, der veränderbar ist und daher manipuliert werden kann. Beispielsweise können das Unterschiede hinsichtlich der Probanden (Geschlecht, Fähigkeiten) oder auch bezüglich des zu testenden Produkts (alte vs. neue Version) sein. Unabhängige Variablen können je nach Typ eine bestimmte Anzahl von Zuständen annehmen (Geschlecht: männlich vs weiblich; Versionen: 1. vs. 2. vs. 3.). Diese Eigenschaft wird als Level bezeichnet. Als abhängige Variablen werden die messbaren Elemente einer Studie beschrieben. Typisch sind hier Success Rates, Anzahl der Fehlermeldungen oder auch Zeiten. Die Daten abhängiger Variablen sollten in statistischen Tests analysiert werden können, also bspw. Zahlenwerte annehmen. Die Bestimmung dieser Variablen ist für eine Usability Studie notwendig, um eine erfolgreiche Durchführung zu garantieren. Die Variablen stehen im engen Verhältnis zur Forschungsfrage, da sie aus dieser meist schon abgelesen werden können (Tullis & Albert, 2013).

In der hier vorliegenden Studie wird eine unabhängige Variable definiert: das zu testende Interface. Dieses ist dahingehend manipulativ da dies die Webseite sowie die App

der Deutschen Bahn sein kann. Damit ist auch gleichzeitig bestimmt, dass das Level der unabhängigen Variable bei zwei liegt. Als abhängige Variablen werden die aufgewendete Zeit, die Bewertung der beiden Interfaces durch die Probanden sowie die erfolgreiche oder erfolglose Durchführung des Tests durch die Probanden festgelegt.

1.1.3 Experimentelles Design

Tullis & Albert [2013] beschreiben, dass bei der Bestimmung des experimentellen Designs einer Studie zwischen den Varianten ‚Within-Group' oder ‚Between-Group' unterschieden wird. Bei einem Within-Group Design testet jeder Proband jedes Produkt, während bei einem Between-Group Design die Probanden in Gruppen aufgeteilt werden und jede Gruppe nur eines der Produkte testet. Within-Group Designs eignen sich vor allem, wenn zwei Produkte miteinander verglichen werden sollen. Between-Group Designs finden vornehmlich bei dem Vergleich unterschiedlicher Probandengruppen Anwendung. Die Vorteile von Within-Groups liegen darin, dass der Studienumfang nicht zu hoch ausfallen muss und die unterschiedlichen Fähigkeiten der Teilnehmer meist nicht berücksichtigt werden müssen, da vornehmlich die Ergebnisse des einzelnen Probanden von Produkt A und Produkt B miteinander verglichen werden. Allerdings besteht die Gefahr, dass ein gewisser Lerneffekt innerhalb des Tests bei den Probanden auftritt, da die zu vergleichenden Produkte ja meist gewisse Ähnlichkeiten aufweisen. Dem kann allerdings entgegengewirkt werden, indem bei unterschiedlichen Probanden eine unterschiedliche Testreihenfolge festgelegt wird. Die Vor- und Nachteile bei Between-Groups liegen in einem größer benötigten Studienumfang und darin, dass die Fähigkeiten der Probanden bei der Auswahl mit in Betracht gezogen werden müssen. Ein Lerneffekt spielt hier jedoch keine Rolle (Tullis & Albert, 2013).

Da die Usability Studie von Webseite und App der Deutschen Bahn AG vor allem diese beiden Interfaces vergleichen sollte, wurde ein Within-Group-Design gewählt. Dies brachte ebenfalls den Vorteil, dass der Auswahlprozess der Probanden weniger eingeschränkt war, was sich für die Durchführung der Studie aufgrund ihres Rahmens als extrem hilfreich darstellte[1].

1.1.4 Metriken

Usability und User Experience können auf unterschiedliche Weisen gemessen werden. Man unterscheidet hier zwischen Performance Metrics und Self-Reported Metrics. Während bei ersteren die Messungen von einem Beobachter durchgeführt werden (z.B. Messung der Zeit, Fehler usw.) wird die Usability bei letzteren durch die Probanden

[1] Da die Studie innerhalb eines Seminars von Studenten durchgeführt wurde standen keinerlei gesonderten Mittel (etwas finanzielle) zur Verfügung, sodass bei der Auswahl von Probanden vor allem auf Freunde und Verwandte zurückgegriffen wurde.

selbst bewertet (Fragebögen). Die Art der Metriken ergibt sich meist schon aus den zuvor bestimmten abhängigen Variablen.

1.1.4.1 Performance Metrics

Nach Tullis & Albert [2013] basieren Performance Metrics auf dem Umgang der Nutzer mit einem zu testenden Produkt. Alle Aktionen, die ein Nutzer mit einem bestimmten Produkt durchführt, können in einer bestimmten Art und Weise gemessen werden. Allerdings wird dafür auch meist eine bestimmte Aufgabe benötigt, die der Proband bearbeitet. Performance Metrics eignen sich besonders, um die Effektivität und Effizienz eines Produktes zu bestimmen. Außerdem kann hiermit die Stärke eines Usability Problems identifiziert werden. Fünf grundlegenden Typen von Performance Metrics sind:

- Task Success: Misst die Erfolgsrate von Usability Tests
- Time-on-Task: Misst die benötigte Zeit von Usability Tests
- Errors: Misst die Anzahl der gemachten Fehler während eines Usability Tests
- Efficiency: Misst den Aufwand den ein User in einem Usability Test aufbringen muss
- Learnability: Misst die Veränderung der Leistung eines Nutzers bei einem Usability Test im Laufe der Zeit

(Tullis & Albert, 2013)

Für die hier vorliegende Studie wurden die Metriken ‚Task Success' und ‚Time-on-Task' ausgewählt, die sich als logische Konsequenz aus den abhängigen Variablen ableiten. Wichtig für das Messen der Erfolgsrate mittels Task Success ist es, dass der Erfolg eines Tests klar definiert werden muss. Dabei sollte nicht nur der erfolgreiche Abschluss einer Aufgabe, sondern auch der Weg dorthin in Betracht gezogen werden. Beim Task Success wird nochmals zwischen den Anwendungsmethoden ‚Binary Success' und ‚Level of Success' unterschieden. Während bei Binary Success entweder ein Score von 1 oder 0 bei erfolgreichem bzw. erfolglosen Abschließen einer Aufgabe vergeben wird, unterscheidet Level of Success noch zwischen verschiedenen Graden von Erfolg (1 ohne Probleme zum Erfolg; 2 mit kleinen Problemen zum Erfolg usw.) Mit diesen Werten lassen sich leicht Durchschnitte und Verhältnisse berechnen, die schließlich anschaulich präsentiert werden können (Tullis & Albert, 2013).

Hier wurde die Variante Binary Success Methode des Task Success' genutzt. Bei der gestellten Aufgabe (welche in Kap. 1.4 näher erläutert wird) wurde festgelegt, dass der Proband bei der Durchführung der Aufgabe bestimmte Zwischenschritte einhalten muss. Nur wenn diese auch eingehalten wurden, konnte schließlich ein Success Score von 1 vergeben werden.

Bezüglich der Time-on-Task Metrik wurde die Zeit vom Start bis zum Ende der Aufgabe jeweils bei Webseite und App gestoppt. Oftmals wird bei der Anwendung dieser Metrik lediglich darauf geachtet, ob ein Nutzer es schafft, innerhalb einer vorgegebenen Zeit die Aufgabe zu lösen (Tullis & Albert, 2013). Da dieser Punkt hier jedoch irrelevant war, da die Usability beider Produkte miteinander verglichen wurde, lag das Hauptaugenmerk auf den Durchschnittszeiten.

1.1.4.2 Self-Reported Metrics

Self-Reported Metrics zielen darauf ab, den Nutzer direkt zu fragen, inwieweit er die Usability eines bestimmten Produkts bewertet. Dies geschieht meist mittels Fragebögen, die allerdings sehr unterschiedlich aufgebaut sein können. Verschiedene Möglichkeiten sind:

- Likert Scales: Probanden stimmen einer Aussage zu oder lehnen sie ab
- Semantic Differentials: Probanden bewerten, wie sehr sie einem von zwei gegensätzlichen Adjektiven bezogen auf das getestete Produkt zustimmen
- Fill-In Questions: Probanden können freie Antworten notieren, welche allerdings eingeschränkt und/oder priorisiert werden können
- Checkbox Questions: Probanden wählen Optionen aus einer Liste aus
- Branching Questions: Die Antwort der Probanden auf eine Frage bestimmt, welche Frage sie als nächstes gestellt bekommen

Tullis & Albert [2008] beschreiben Self-Reported Metrics als das wichtigste Messinstrument, welches in Usability Tests eingesetzt werden kann. Dies liegt darin begründet, dass hier Emotionen und Gefühle von Probanden bezüglich eines bestimmten Produkts identifiziert werden können. So ist es beispielsweise möglich, dass die Tester zwar sehr viel Zeit und viele Klicks benötigen, um eine Aufgabe abzuschließen, sie aber trotzdem Freude im Umgang mit dem Produkt empfinden. Neben der Möglichkeit, einen eigenen Fragebogen zu entwickeln, existieren einige standardisierte Fragebögen, welche teilweise kostenlos genutzt werden können. Einer der beliebtesten ist der ‚System Usability Scale' Fragebogen (SUS), der den Semantic Differentials zuzuordnen ist. Dieser unterteilt sich in zehn Aussagen, welche abwechselnd positiv und negativ formuliert sind. Die Probanden haben die Möglichkeit, auf einer Skala von 0 bis 4 (‚strongly disagree' bis ‚strongly agree') diesen Aussagen zuzustimmen. Aus den Antworten kann schließlich ein Gesamtwert errechnet werden, der zwischen 0 und 100 liegt. Dafür werden die Codierungen der Antworten (0-4) auf bestimmte Weise umcodiert, aufsummiert und mit 2,5 multipliziert. Zwar wird die Art der Umcodierung in der Literatur unterschiedlich beschrieben, allerdings führen die unterschiedlichen Methoden zum selben Ergebnis. Sauro & Lewis [2012] beschreiben, dass von der Kodierung positiver Fragen der Wert 1 abgezogen und der Wert der Kodierung negativer Zahlen vom Wert 5 abgezogen wird.

Durch das Aufsummieren und der Multiplikation mit dem Faktor 2,5 erhält man dann einen Wert, der die allgemeine Usability der Anwendung bestimmt. Neben der Tatsache, dass der SUS kostenlos und lizenzfrei genutzt werden kann und sehr schnell abzuarbeiten ist, ist er auch noch äußerst populär[2]. Daher lässt sich der SUS Wert einer Studie gut mit denen anderer vergleichen. Der weltweite Durchschnitt des SUS liegt bei 68 (Tullis & Albert, 2013). Nach Bangor, Kortum & Miller [2008] können die Ergebnisse einer SUS-Auswertung folgendermaßen interpretiert werden:

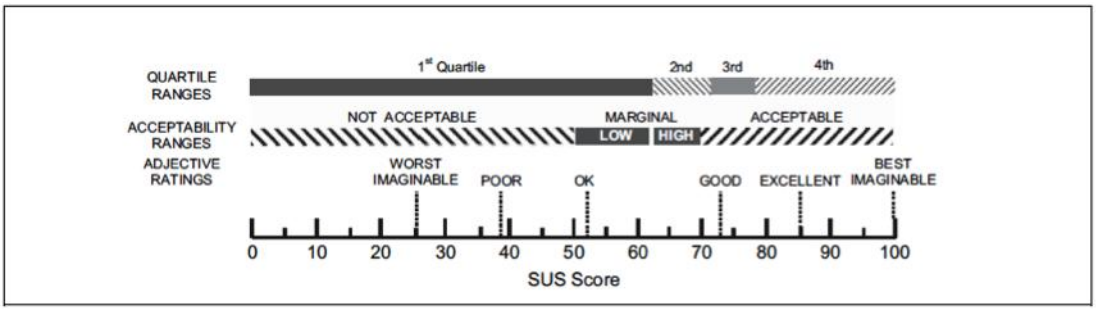


Abbildung 1: SUS-Score Interpretation (Bangor, Kortum, & Miller, 2008)

Laut Abbildung werden Produkte mit einem SUS Score von unter 50 als nicht akzeptierbar bewertet. Von 50 bis ca. 63 Punkten weist eine Anwendung eine gerade noch akzeptierbare Usability auf. Anwendungen mit einem darüber liegenden Score werden als akzeptabel bezeichnet. Dem SUS Fragebogen wird ein Cronbach Alpha Wert von .92 zugeschrieben, was einer exzellenten Zuverlässigkeit bezüglich der gelieferten Ergebnissen entspricht und damit sogar höher liegt als manche lizenz- oder gebührenpflichtige Fragebögen (Sauro, 2011). Da der SUS Fragebogen im Vergleich zu anderen eine recht simple und kostengünstige aber zugleich auch sehr effektive und effiziente Möglichkeit bietet, die Meinungen von Nutzern bezüglich eines bestimmten Produkts zu erheben, wurde dieser auch in der hier vorliegenden Studie verwendet[3].

1.2 Probanden

Der Auswahlprozess von Studienteilnehmern umfasst zwei Bereiche. Zum einen müssen Probanden mit geeigneten Eigenschaften gefunden werden. Diese beziehen sich auf die Zielgruppe des zu testenden Produkts. Würde man diese Einschränkung nicht festlegen, liefe man Gefahr unbrauchbare Ergebnisse zu erzielen, da die getesteten Personen sich mitunter anders verhalten als die Zielgruppe des Produkts (Rubin & Chisnell, 2008). Desweiteren besteht die Möglichkeit, den Studienumfang zu berechnen. Dies ist nicht unbedingt notwendig, da Usability Probleme teilweise schon bei ein oder zwei getesteten Probanden offensichtlich werden können. Allerdings wird so statistisch zu

[2]SUS wird bei 41% aller Post-Study Questionnaires in Usability Studien genutzt

[3] Für SUS-Beispielfragebogen siehe Anhang 1.

einem bestimmten Grad sichergestellt, dass die Ergebnisse bezogen auf die Gesamtpopulation aussagekräftig sind (Sauro & Lewis, 2012).

1.2.1 Probanden finden

Laut Rubin & Chisnell [2008] wird für den Auswahlprozess von Probanden oftmals ein User Profil erstellt. Dabei wird so gut wie möglich ein potentieller Nutzer eines Produkts beschrieben. Es setzt sich meist aus firmeninterne Zielgruppenbeschreibungen und auch Anforderungen des Produkts an potentielle Nutzer zusammen. Falls die Probanden in bestimmte Nutzergruppen unterteilt werden (bspw. ‚Anfänger' / ‚Experte'), ist es wichtig, diese Unterteilung genau zu definieren. Oft werden solche Bezeichnungen von unterschiedlichen firmeninternen Abteilungen auch unterschiedlich bewertet. Die Feststellung, in welche Kategorie ein potentieller Proband fällt, lässt sich gut mit Kompetenzmatrizen identifizieren. Eine einfachere Möglichkeit wäre, Probanden einfach nach der Nutzungsdauer bestimmter Produkte zu fragen. So könnte beispielsweise ein potentieller Teilnehmer eine tägliche Internetnutzung von 6 Stunden angeben, wodurch er als kompetenter User eingestuft werden würde. Dies birgt allerdings die Gefahr, dass der Nutzer eventuell lediglich voreingestellten Lesezeichen folgt und sich auf nicht mehr als drei Webseiten bewegt. Die Nutzungsdauer sagt hier also nicht viel über die Internetkompetenz aus. Eine Kompetenzmatrix dagegen fragt explizit nach sehr speziellen Nutzungsmustern und verbindet dies dann mit deren Frequenz. Aus der Beantwortung einer solchen Matrix ergibt sich dann eine bestimmte Punktzahl, aufgrund derer die Personen dann einer bestimmten Nutzergruppe zugeordnet werden kann. Bei der Verwendung von Matrizen bietet es sich an, zwischen ‚Domain Knowledge' und ‚Tool Knowledge' zu unterscheiden bzw. beides getrennt voneinander abzuprüfen. Domain Knowledge beschreibt das Wissen um einen Fachbereich während Tool Knowledge das Wissen über den Umgang mit bestimmten Werkzeugen wie Software o.ä. beschreibt. Außerdem fließt die Klassifikation der Probanden mit in die Einteilung von Nutzergruppen. Klassifikationen beschreiben Handlungsschemata, Vorlieben und Kenntnisse der Nutzer. Dies können im Rahmen einer Web-Anwendung beispielsweise die Nutzung von Premium-Angeboten, bevorzugte Browser und das Wissen um Alternativen sein. Typische Unterscheidungen bezüglich der Klassifikation wären beispielsweise „Chef/Sektretär" oder „Premiumnutzer/freier Nutzer". Die Erhebung der oben beschriebenen relevanten Daten erfolgt in einem Screening-Test, der sich meist als simpler Fragebogen darstellt. Die Antworten der potentiellen Probanden geben dann schlussendlich Aufschluss darüber, ob die jeweilige Person überhaupt an der Studie teilnehmen kann, und falls ja, welcher Nutzergruppe sie zuzuordnen ist (Rubin & Chisnell, 2008).

Die hier vorliegende Studie soll eine Aussage über die Usability von App und Webseite bezüglich aller möglichen Nutzer der Deutschen Bahn treffen. Daher war es nicht unbedingt nötig, ein spezielles User Profil zu erstellen, da theorietisch jede Person als mögli-

cher Nutzer gesehen werden kann. Daher spielte auch die Klassifikation der Nutzer eigentlich keine Rolle. Diese wurde allerding trotzdem durchgeführt, da sich dadurch eventuell weitere interessante Zusammenhänge ergeben könnten. Für die Einteilung der Probanden in bestimmte Nutzergruppen, der Hauptgrund für einen Screening Test, wurde dies allerdings nicht verwendet. Die Verwendung von Kompetenzmatrizen war hingegen notwendig. Zumindest die Matrix bezüglich der Tool-Knowledge sollte bei den potentiellen Probanden eine gewisse Punktzahl ergeben, da diese auf das Fachwissen bezüglich des Umgangs mit Smartphones, PCs, des Internets und Apps abzielte. Offensichtlich muss in diesen Bereichen ein gewisses Kompetenzlevel vorhanden sein, um die App und Webseite der Deutschen Bahn nutzen zu können.

1.2.2 Studienumfang berechnen

Zur Berechnung des Studienumfangs gibt es verschiedene Methoden. Welche davon genutzt wird, hängt vom Design der Studie ab. Es kommt vor, dass mehrere Methoden zu einem Studiendesign passen. Hier ist es wichtig eine Methode auszuwählen, die am besten auf das vorliegende Studiendesign passt. Jede Methode verwendet verschiedene Formeln aus der Statistik, etwa die zur Berechnung des T-Werts. Auf die Formeln soll im Weiteren nicht eingegangen werden. Es ist wichtig, dass bei der Planung der Studie sicher ist, welche Methode (und damit auch welche Formel) zum Studiendesign passt. Wenn dies der Fall ist, reicht es aus, die richtigen Werte zu identifizieren, die in die Formel(n) eingesetzt werden müssen. Sauro & Lewis [2012] beschreiben verschiedene Möglichkeiten, um den Stichprobenumfang auszurechnen, welche im Folgenden mit einfachen Beispielen erläutert werden:

- Einfache Schätzung bei nichtbinären Daten: Wird genutzt, wenn es theoretisch unendlich viele Ausprägungen der Daten gibt und diese vom Mittelwert nach oben und unten abweichen. Beispiel: Man möchte die durchschnittliche Zeit für den Arbeitsweg auf die komplette Zeit einer Anstellung ermitteln.
- Benchmark-Test bei nichtbinären Daten: Wird genutzt, wenn die Daten gegen einen bestimmten Maximalwert getestet werden sollen. Beispiel: Nach einigen Verbesserungen einer Webseite soll diese mindestens einen SUS-Score von 75 erreichen.
- Hypothesen-Test bei nichtbinären Daten: Wird genutzt, wenn eine Alternativhypothese gegen eine Nullhypothese getestet werden soll.
- Binominal-Test bei binären Daten und großem Stichprobenumfang: Wird genutzt, wenn in den Tests nur zwei Zustände erreicht werden können und die Mittel einen großen Stichprobenumfang zulassen. Beispiel: Die Erfolgsrate von Nutzer-Logins auf einer Webseite wird getestet (mögliche Zustände: ‚Login erfolgreich' und ‚Login erfolglos').

- Binominal-Test bei binären Daten und kleinem Stichprobenumfang: Ähnlich wie oben, allerdings bei kleinem Stichprobenumfang.
- Chi-Quadrat Test bei binären Daten. Wird genutzt, wenn zwei Alternativen gegeneinander für ein Between-Group Design getestet werden sollen. Beispiel: Zwei Versionen eines Programms werden auf eine gewünschte Erfolgsrate bei der Installation getestet.
- McNemar-Test bei binären Daten: Wird genutzt, wenn zwei Alternative gegeneinander für ein Within-Group Design getestet werden sollen. Beispiel siehe oben.

(Sauro & Lewis, 2012)

In dieser Studie zum Usability Vergleich von App und Webseite der deutschen Bahn bot sich zur Studienumfangsberechnung am ehesten der McNemar-Test an, da hier zwei alternative Interfaces bei einem Within-Group Design miteinander verglichen werden sollten. Die Größe des Studienumfangs hing schlussendlich vom Konfidenzintervall ab. Da diese Studie ohne weitere Mittel auskommen musste, war es nicht möglich, einen allzu großen Umfang zu bearbeiten. Dies bedeutete, dass Abstriche bezüglich des Konfidenzintervalls gemacht werden mussten. Schließlich wurde eine 80%ige Konfidenz in Kauf genommen, da sich der Stichprobenumfang hierbei auf annehmbare 24 Probanden belief.

1.3 Materialien

Je nachdem welche Methoden man innerhalb der Studie anwendet, werden verschiedenen Materialien benötigt. Im Allgemeinen lassen sich diese in die folgenden Kategorien einteilen:

- Orientation Script: Ein Orientation Script beschreibt den Ablauf einer Studie und dient dazu, den Teilnehmern vor Beginn einen umfassenden Ausblick auf die/den folgende/n Test/s zu geben. Durch die Anwendung dieses Skripts sollen alle möglichen Fragen der Teilnehmer im Vorfeld geklärt werden, sodass sich diese während des Tests vollkommen darauf konzentrieren können. Typische Elemente eines Orientation Scripts sind die Vorstellung aller Anwesenden, Kontext der Studie, Set-up der/des Tests und eine Beschreibung der Erwartungen an die Probanden.

  Diese Studie verwendete kein materielles Orientation Script, da dies in Anbetracht der Tatsache, dass alle Probanden aus Freundes- und Verwandtenkreisen zusammen kamen, als zu förmlich gewirkt hätte und der Studie damit wohl eher geschadet als geholfen hätte (bspw. durch Unverständnis bei den Probanden

über einen künstlich erzeugten, zu förmlichen Rahmen). Alle nötigen Informationen konnten zudem im Gespräch geklärt werden.

- Background Questionnaires: Hierbei handelt es sich meist um Fragebögen, die Hintergrundinformationen über den Probanden einholen sollen. Typischerweise sind dies Screening Tests, wie sie bereits in Kap. 1.2.1 beschrieben wurden. Wichtig hierbei ist vor allem eine einfache Gestaltung dieser Fragebögen, damit die Probanden diese schnell bearbeiten können und sich nicht überfordert fühlen. In der hier vorliegenden Studie kamen Background Questionaires in Form eines Screening Test Fragebogens zum Einsatz, wodurch die Möglichkeit gegeben war ein User Profil zu erstellen, sowie die Probanden anhand ihrer Fähigkeiten zu klassifizieren.
- Data Collection Instruments: Hierbei wird zwischen automatischen und manuellen Instrumenten unterschieden. Während manuelle Formen der Datenerfassung vor allem herkömmliche Materialien wie Notizblöcke oder Video- und Audioaufnahmegeräte sind, werden mit automatischen Instrumenten meist Programme bezeichnet, die es erlauben, die Anzahl von Tastendrücken und Mausklicks aufzuzeichnen.

  Bei dieser Studie wurden ausschließlich herkömmliche Materialien zur Datenerfassung verwendet, um die jeweiligen Zeiten sowie das Abschneiden bezüglich Task Success und SUS zu notieren. Die Verwendung automatisierter Instrumente wäre in Anbetracht der getesteten Produkte zwar auch möglich gewesen, hätte aber vornehmlich kein bedeutsames Ergebnis bezüglich der Forschungsfrage geliefert und wurde daher sowie aus Aufwandsgründen nicht berücksichtigt.
- Nondisclosure Agreements and Recordings Consent Form: Hierbei handelt es sich um Dokumente, die bei der Durchführung der Studie den rechtlichen Rahmen absichern. Etwa geben Teilnehmer hierbei beispielsweise ihr Einverständnis, durch Videokameras aufgezeichnet zu werden.

  In der vorliegenden Studie wurden etwaige Dokumente aufgrund des informellen Rahmens nicht benötigt.
- Pre-Test Questionnaires: Wie bei den Background Questionaires handelt es sich hierbei um Fragebögen, die vor dem eigentlichen Test auszufüllen sind. Allerdings beziehen sich diese auf die Erwartungen, die die Probanden auf das zu testende Produkt stellen. Hierbei besteht die Möglichkeit, erste Verbesserungsmöglichkeiten durch ein zu großes Abweichen von Erwartungen und Realität zu identifizieren.

  Die Studie zur Webseite und App der Deutschen Bahn verwendete keine Pre-Test Questionnaires.
- Task Scenarios: Hier wird die Aufgabe, die die Probanden zu bewältigen haben, beschrieben. Wichtige Punkte sind hierbei, dass das Ziel der Aufgabe klar be-

schrieben wird, die Motive für die Aufgabe verdeutlicht werden, echte Namen und Beschreibungen verwendet werden und, dass Anzeigen und Display- sowie Druckausgaben, die innerhalb der Aufgabe zu sehen sind, erläutert werden.
Für diese Studie wurde ein kurzer Absatz zu einem real möglichen Szenario beschrieben, dass den Teilnehmern vorgelesen wurde. Die Aufgabe der Teilnehmer, eine Fahrplanauskunft mit Hin- und Rückfahrt sowie einem Zwischenhalt zu einer bestimmten vorgegebenen Zeit mit Hilfe der App und der Webseite herauszusuchen und zu buchen, wurde in Form einer Liste ausgedruckt und an die Probanden ausgeteilt, nachdem es ihnen vorgelesen wurde.

- Post-Test Questionnaire: Diese Fragebögen werden den Probanden nach dem Test ausgehändigt. Sie sollen einen tieferen Einblick in das Meinungsfeld der Probanden liefern. Besonders wertvoll können Post-Tests im Verbund mit Pre-Tests sein, wenn diese an den Pre-Test anknüpfen.
  In dieser Studie wurde als Post-Test der SUS-Fragebogen in ausgedruckter Form verwendet.
- Debriefing Topics Guide: Nachbesprechungen mit den Probanden können den Versuchsleitern die Gründe für bestimmte Fehler aufzeigen, sollten diese innerhalb des Testdurchlaufs nicht geklärt worden sein. Zudem können weitere Probleme in Erfahrung gebracht werden, die ebenfalls während der Durchführung der Studie nicht aufgefallen sind. Dadurch kann neben der reinen Datensammlung vor allem eine rege Diskussion zu neuen Sichtweisen und anderen Betrachtungswinkeln führen.
  Eine Nachbesprechung fand innerhalb dieser Studie nicht statt.

(Rubin & Chisnell, 2008)

1.4 Prozedur

Der Ablauf der Testdurchläufe stellte sich wie folgt dar:

Zunächst wurde den Teilnehmern der Screening-Test ausgeteilt und von diesen ausgefüllt. Danach wurden mit Hilfe des Task Szenarios die inhaltlichen Aspekte der Studie, sowie die genaue Aufgabe der Probanden erläutert. Etwaige Fragen wurden im Folgenden geklärt. Daraufhin begannen die Teilnehmer mit dem Test. Hierbei wurde darauf geachtet, dass die benutzten Browser-Caches der benutzten Geräte (Smartphone, Laptop/PC) leer waren, da ansonsten durch zuvor gespeicherte Eingaben die Testdurchläufe nicht einheitlich gewesen wären. Da es sich um ein Within-Group Design handelte, wurde darauf geachtet, dass die Reihenfolge der beiden zu testenden Produkte von Tester zu Tester unterschiedlich waren, um einem möglichen Lernprozess entgegen zu wirken. Während der Testphase wurde die Zeit gestoppt. Den Start stellte das Aufrufen der Startseite (Webseite) bzw. des Home-Screens (App) dar. Das Ziel wurde mit dem Betä-

tigen des Buttons „Ticket / Reservierung“ bei der App, sowie „Zur Buchung“ bei der Webseite erreicht. Direkt im Anschluss an jeden Durchgang wurde von den Probanden ein SUS-Fragebogen bezüglich des gerade genutzten Interfaces ausgefüllt. Dieses Verfahren stellte sich in einer Pilotstudie als äußerst wichtig heraus, da bei einem Ausfüllen beider Bögen nach beiden Tests die Probanden meist mehr an das zuletzt getestete Interface denken und daher die Möglichkeit einer Verzerrung bezüglich der SUS-Fragebögen besteht. Nach beiden Durchläufen wurden die entsprechenden Ergebnisse notiert und gesammelt.

2 Ergebnisse

2.1 Deskriptive Statistik – Time-on-Task

Nach der Durchführung der Studie an 24 Probanden lagen folgende Zeiten (in Minuten) vor:

Nr.	App	Webseite	Nr.	App	Webseite	Nr.	App	Webseite
1	00:05:41	00:02:46	9	00:24:55	00:19:12	17	00:08:34	00:06:12
2	00:06:13	00:05:40	10	00:02:54	00:02:20	18	00:11:34	00:07:43
3	00:05:53	00:03:37	11	00:04:17	00:01:49	19	00:04:23	00:04:37
4	00:06:03	00:04:43	12	00:05:21	00:07:30	20	00:05:23	00:08:45
5	00:05:13	00:02:25	13	00:04:17	00:02:07	21	00:08:34	00:07:23
6	00:04:58	00:05:21	14	00:11:39	00:04:19	22	00:06:56	00:07:42
7	00:05:22	00:02:38	15	00:04:49	00:04:54	23	00:07:34	00:06:13
8	00:07:36	00:05:07	16	00:03:56	00:05:02	24	00:07:53	00:05:29

Tabelle 1: Gemessene Zeiten für Webseite und App

Von diesen Zeiten wurden zunächst die Mittelwerte gebildet:

App	Webseite
00:07:04	00:05:33

Tabelle 2: Mittelwerte der Zeiten von Webseite und App

Bereits hier wurde deutlich, dass die Probanden mehr Zeit bei der App für die Suche nach einer Fahrplanauskunft und die zugehörige Buchung aufwenden. Daher lag hier bereits die Vermutung nahe, dass die Webseite eine bessere Usability gegenüber der App aufweist. Allerdings mussten diese Werte noch näher analysiert werden, um sicher zu gehen, dass dieser Unterschied nicht einfach zufällig entstanden ist.

2.2 Mittelwertvergleiche für Within-Group Designs – Time-on-Task

Um festzustellen, ob der Unterschied zwischen den Mittelwerten zufällig entstanden ist oder tatsächlich signifikant ist, können die beiden Werte auf bestimmte Art miteinander verglichen werden. Nach Field [2003] kann für Mittelwertvergleiche bei Within-Group Designs der abhängige T-Test herangezogen werden. Um den abhängigen T-Test durchzuführen wird vorausgesetzt, dass die erhobenen Daten intervallskaliert, und bei Stichproben mit einem Umfang von weniger als 30, normalverteilt sind. Da es sich hierbei um intervallskalierte Daten handelt, mussten diese also vor der eigentlichen Anwendung des abhängigen T-Tests nur noch auf Normalverteilung getestet werden (Field & Hole, 2003).

Nach Ebermann [2010] können die Daten mit Hilfe des ‚Kolmogorov-Smirnov-Test' auf Normalverteilung getestet werden. Dieser eignet sich besonders bei Stichproben kleineren Umfangs und ohne Klasseneinteilung (Ebermann, 2010).

Daraus ergaben sich folgende Werte:

Kolmogorov-Smirnov-Anpassungstest

		Zeit_App	Zeit_Web
N		24	24
Parameter der Normalverteilung[a,b]	Mittelwert	0:07:04	0:05:33
	Standardabweichung	0:04:23	0:03:30
Extremste Differenzen	Absolut	,242	,187
	Positiv	,242	,187
	Negativ	-,195	-,143
Kolmogorov-Smirnov-Z		1,188	,915
Asymptotische Signifikanz (2-seitig)		,119	,372

Tabelle 3: Ergebnisse des Kolmogorov-Smirnov-Tests

Um zu bestimmen, ob es sich um eine Normalverteilung handelt oder nicht, werden folgend Werte betrachtet:

N = 24
Extremste Differenzen (Absolut) = ,242 (App) und ,187 (Web)
Asymptotische Signifikanz = ,119 (App) und ,372 (Web)

Die Werte von N und Extremste Differenzen (ED) werden mit den kritischen Werten des Kolmogorov-Smirnov-Tests verglichen. Diese gibt bei einer bestimmten Irrtumswahrscheinlichkeit die Grenzwerte für Stichproben zugehöriger Größe an. Wenn die Werte für ED die kritischen Werte aus der Tabelle überschreiten, liegt mit einer 95%igen Wahrscheinlichkeit keine Normalverteilung vor, wenn die Irrtumswahrscheinlichkeit – wie hier – bei 5% liegt. Bei N=24 liegt der zugehörige ED Wert bei 0,269. Weder der ED Wert der App noch der Webseite überschreiten diesen. Außerdem wird der Wert für Asymptotische Signifikanz mit dem Grenzwert von 0,05 verglichen. Hier liegen die Werte für App und Webseite deutlich darüber. Wäre dies nicht so, müsste man annehmen, dass lediglich 5% solcher hier vorliegenden Verteilungen tatsächlich normalverteilt sind (Ebermann, 2010). Man kann daher also annehmen, dass die vorliegenden Werte normalverteilt sind. Damit sind die Voraussetzungen für einen Mittelwertvergleich mit Hilfe des abhängigen T-Tests gegeben.

n	Wert	n	Wert
3	0,708	20	0294
4	0624	21	0,287
5	0,563	22	0,281
6	0,519	23	0,275
7	0,483	24	0,269
8	0,454	25	0,264
9	0,43	26	0,259
10	0,409	27	0,254
11	0,391	28	0,25
12	0,375	29	0,246
13	0,361	30	0,242
14	0,349	31	0,238
15	0,338	32	0,234
16	0,327	33	0,231
17	0,318	34	0,227
18	0,309	35	0,224
19	0,301		

Tabelle 4: Kritische Werte beim Kolmogorov-Smirnov-Test

Vor dem eigentlichen Test wurden eine Null-Hypothese (H0) sowie eine Alternativ-Hypothese (H1) aufgestellt. Durch den abhängigen T-Test kann am Ende festgestellt werden ob die Null-Hypothese beibehalten oder verworfen wird (Field & Hole, 2003).

H0: Es gibt keinen signifikanten Unterschied zwischen den Zeiten von App und Webseite.

H1: Es gibt einen signifikanten Unterschied zwischen den Zeiten von App und Webseite. Die Zeiten von App liegen signifikant über denen der Webseite.

Bei der Aufstellung der Alternativ-Hypothese muss entschieden werden, ob diese gerichtet oder ungerichtet formuliert wird. Bei einer ungerichteten Alternativ-Hypothese wird zuvor zwar angenommen, dass sich die Mittelwerte unterscheiden, allerdings ist nicht klar, welcher der beiden Werte über dem anderen liegt. Bei einer gerichteten Hypothese wird bereits die Erwartung ausgedrückt, dass ein bestimmter Wert über dem anderen liegt. Diese Festlegung bestimmt am Ende den Umgang mit dem Signifikanzwert bzw. des Konfidenzintervalls (Field & Hole, 2003). Da die Mittelwerte bereits bekannt sind, steht fest, dass die Werte in eine bestimmte Richtung gehen (Zeitwert App liegt höher). Daher ist die Alternativ-Hypothese gerichtet.

Bei der Durchführung des abhängigen T-Tests ergaben sich bei einem festgelegten Konfidenzintervall von 95% (Irrtumswahrscheinlichkeit α = 0,05) folgende Ergebnisse:

Test bei gepaarten Stichproben

		Gepaarte Differenzen					T	df	Sig. (2-seitig)
		Mittelwert	Standardabweichung	Standardfehler des Mittelwertes	95% Konfidenzintervall der Differenz Untere	Obere			
Paaren 1	Zeit_App - Zeit_Web	0:01:31	0:02:20	0:00:28	0:00:31	0:02:30	3,163	23	,004

Tabelle 5: Ergebnisse des abhängigen T-Tests

Um mit Hilfe dieser Werte eine Aussage über einen möglichen Unterschied zwischen der Zeit der App und der Webseite machen zu können, wird der Wert für die Signifikanz betrachtet und mit dem Wert für α verglichen. Bei ungerichteten Hypothesen würde man diesen mit α/2 vergleichen. Falls der Signifikanzwert kleiner als α ist, kann die Null-Hypothese verworfen und die Alternativ-Hypothese angenommen werden. (Field & Hole, 2003). In diesem Fall liegt der Signifikanzwert von ,004 deutlich unter α = ,05. Daher lässt sich mit 95%iger Wahrscheinlichkeit sagen, dass es einen signifikanten Unterschied zwischen den Zeiten von Webseite und App gibt. Dieser ist demnach nicht zufällig entstanden.

Alternativ ist es auch möglich, den errechneten T-Wert mit dem kritischen T-Wert aus der Tabelle für die T-Verteilung zu vergleichen. Den kritischen T-Wert findet man, indem man den Wert entsprechend der Freiheitsgrade (hier: df = 23) und des Konfidenzintervalls (genauer: 1-α; hier: 1-α = 0,95) aus der Tabelle abliest. In diesem Fall liegt dieser bei 1,741. Falls dieser Wert größer ist als der berechnete T-Wert, wird die Null-Hypothese beibehalten (Field & Hole, 2003). Da der T-Wert in diesem Fall bei 3,163 liegt, ist dieser deutlich größer als der kritische T-Wert. Damit wird, wie schon zuvor erkannt, die Null-Hypothese verworfen und die Alternativ-Hypothese angenommen.

Nun gilt es noch zu bestimmen, ob es sich um einen großen oder kleinen Unterschied handelt. Dazu kann mit Hilfe des T-Werts und der Freiheitsgrade die Effektgröße berechnet werden. Liegt diese über 0,1, spricht man von einem kleinen; über 0,5 von einem großen Effekt bzw. Unterschied (Field & Hole, 2003). Für die vorliegende Studie ergibt sich folgendes:

T-Wert	3,169
df	23
Effektstärke:	0,20851441

Tabelle 6: Effektstärke

Es lässt sich also zusammenfassen, dass es einen signifikanten Unterschied zwischen den Zeiten der Webseite und der App, zugunsten der Webseite gibt, dieser allerdings eher klein ausfällt.

2.3 SUS-Score

Neben der Zeitmessung, die vornehmlich für den Usability-Vergleich zwischen App und Webseite der Deutschen Bahn diente, wurden noch zwei weitere Messungen durchgeführt, um das Ergebnis des Mittelwertvergleichs der Zeiten als Aussage über die Usability möglichst unterstützen zu können. Zum einen wurde hier ein SUS-Fragebogen von den Teilnehmern beantwortet. Dieser lieferte folgende Ergebnisse für die einzelnen Probanden:

Nr.	App	Web	Nr.	App	Web	Nr.	App	Web
1	37,5	80	9	62,5	70	17	65	67,5
2	35	72,5	10	70	85	18	37,5	65
3	32,5	72,5	11	45	80	19	85	65
4	45	75	12	55	40	20	17,5	57,5
5	72,5	95	13	62,5	75	21	12,5	80
6	55	72,5	14	40	77,5	22	47,5	62,5
7	77,5	92,5	15	85	72,5	23	40	65
8	57,5	75	16	42,5	52,5	24	72,5	80

Tabelle 7: SUS-Scores für App und Webseite

Von diesen Werten wurde das arithmetische Mittel genommen, so dass sich für die App und Webseite folgende Gesamtergebnisse bezüglich des SUS-Scores ergaben:

App	Webseite
52,19	72,08

Tabelle 8: Durchschnittliche SUS-Scores App und Webseite

2.4 Task Success

Die dritte durchgeführte Messung betraf den Task Success der Aufgaben. Dabei erhielten die Probanden einen Score von 1 bei erfolgreichem Abschluss der Aufgabe oder einen Score von 0 bei erfolglosem Abschluss. Um zu bestimmen, wann eine Aufgabe als erfolgreich bzw. erfolglos deklariert werden konnte, wurden Definitionen für den Erfolg bzw. Misserfolg einer Aufgabe aufgestellt:

- Erfolg: Der Proband soll Hin- und Rückfahrt in einem Durchgang heraussuchen, sodass er nur ein einziges Mal auf den Button „Ticket/Reservierung“ bei der App, bzw. „Zur Buchung“ bei der Webseite klicken muss.

- Misserfolg: Wenn der Proband mehrere Fahrten buchen muss. Beispielsweise findet er nicht heraus, wie man einen Zwischenhalt eingibt oder eine Rückfahrt hinzufügt und muss daher öfter als ein Mal den entsprechenden Button klicken.
- Misserfolg: Wenn der Proband nicht weiter weiß. Beispielsweise ist er nicht in der Lage, einen Zwischeninhalt einzugeben oder eine Rückfahrt nicht hinzufügen und teilt dies dem Beobachter mit. Dann soll er die Aufgabe trotzdem weiter durchführen, bis er theoretisch alle Fahrten (wenn auch separat) gebucht hat. Diese Option sollte dem Probanden allerdings erst während des Tests mitgeteilt werden, da er sonst zu früh das eigentliche Ziel abbrechen könnte. Falls dieser Fall eintritt, wird auch dies als Misserfolg gewertet.

Die Messung des Task Succes lieferte folgende Ergebnisse:

Nr.	App	Web	Nr.	App	Web	Nr.	App	Web
1	1	1	9	0	1	17	1	0
2	1	0	10	1	1	18	0	1
3	1	1	11	1	1	19	1	0
4	1	1	12	1	0	20	0	0
5	1	1	13	1	0	21	0	0
6	1	1	14	0	1	22	1	0
7	1	1	15	1	1	23	1	0
8	0	1	16	1	1	24	1	1

Tabelle 9: Task Success Score der Probanden

Von diesen Daten wurden dann die prozentualen Häufigkeiten bestimmt:

Task Success - App

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Fail	6	25,0	25,0	25,0
	Success	18	75,0	75,0	100,0
	Gesamt	24	100,0	100,0	

Task Success - Webseite

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Fail	9	37,5	37,5	37,5
	Success	15	62,5	62,5	100,0
	Gesamt	24	100,0	100,0	

Tabelle 10: Häufigkeiten des Task Success (App & Webseite)

Während also bei der Webseite lediglich 62,5 % der Probanden die Aufgabe erfolgreich abschließen konnten, schafften dies bei der App sogar 75 %. Bereits hier wird deutlich, dass dieses Ergebnis im Gegensatz zu den Auswertungen von Time-on-Task sowie des SUS-Fragebogen steht. Eine genauere Betrachtung möglicher Ursachen dieses Gegensatzes ist im Kapitel ‚Diskussion' zu finden.

2.5 Weitere Ergebnisse

2.5.1 Korrelationsanalyse: Tool- und Domain-Knowledge

Wie zuvor schon beschrieben, wurde vor jedem Testdurchlauf ein Screening-Test durchgeführt, der unter anderem auch die jeweiligen Werte bezüglich Tool- und Domain-Knowledge abgefragte. Wie bereits erwähnt, dient ein solcher Test vornehmlich zur Bestimmung der Fähigkeiten von Probanden. Damit soll sichergestellt werden, dass nur Teilnehmer zum Test zugelassen werden, die in das entsprechende User Profil passen. Da dies hier jedoch nicht von Nöten war, diente dieser Screening Test lediglich dazu, eventuelle Zusatzergebnisse ermitteln zu können. Dabei stellte sich der Zusammenhang von Tool- und Domain-Knowledge und den erzielten Zeiten als interessanter Untersuchungsgegenstand heraus. Hier kam die Frage auf, ob sich hohe Werte bei der Tool-Knowledge (Umgang mit der Technik; Smartphone, PC, etc.) und/oder bei der Domain-Knowledge (Wissen über Prozesse von Reiseplanung, -buchung, etc.) eventuell auf ein gutes zeitliches Abschneiden bei den Testdurchläufen auswirken. Um dies zu untersuchen, wurde eine Korrelationsanalyse durchgeführt.

Bei einer Korrelationsanalyse werden der errechnete Korrelationskoeffizient sowie der p-Wert betrachtet. Liegt der Koeffizient unter 0,3, ist der Zusammenhang der untersuchten Merkmale vernachlässigbar. Ein Wert zwischen 0,3 und 0,7 spricht für einen schwachen Zusammenhang, während ein Wert über 0,7 einen starken Zusammenhang anzeigt. Ob der Zusammenhang als signifikant betrachtet werden kann, ergibt sich aus dem p-Wert, der dafür unter 0,05 liegen sollte. Der Korrelationskoeffizient kann auf unterschiedliche Weise bestimmt werden. Oftmals wird hier der Pearson-Koeffizient verwendet. Dieses Verfahren setzt allerdings normalverteilte Daten sowie Linearität voraus. Sollte eine dieser Bedingungen nicht gegeben sein, kann auch der Korrelationskoeffizient nach Spearman berechnet werden (Keller, 2013) (Bortz, 2005).

Wie die Tool- und Domain-Knowledge der Probanden bestimmt werden, wurde bereits in Kap. 1.2.1 beschrieben. Aus der Erhebung ergaben sich folgende Daten:

Nr.	Tool	Domain	Nr.	Tool	Domain	Nr.	Tool	Domain
1	10	6	9	4	0	17	11	3
2	3	2	10	10	2	18	4	1
3	8	1	11	8	0	19	10	1
4	8	1	12	11	3	20	2	4
5	11	2	13	13	3	21	4	5
6	9	1	14	6	4	22	6	4
7	13	7	15	7	4	23	9	3
8	12	5	16	11	6	24	10	6

Tabelle 11: Tool- und Domain-Knowledge der Probanden

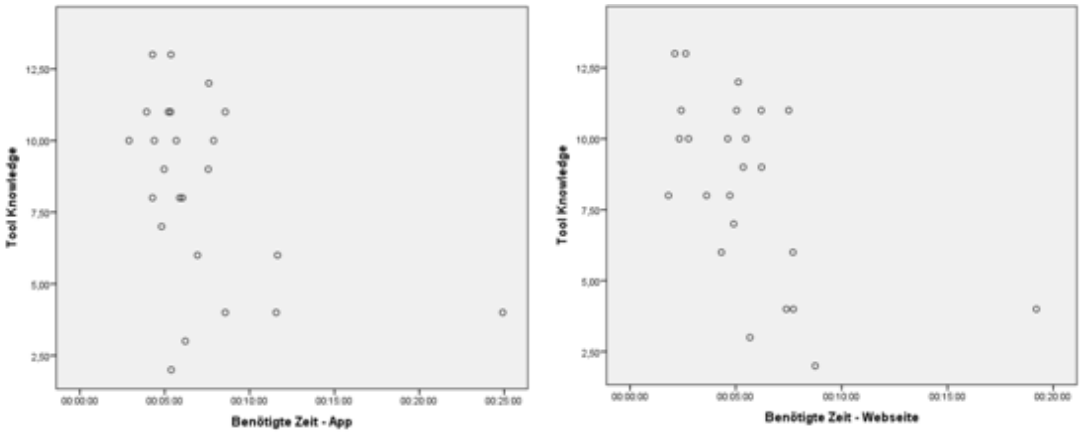
Diese Daten wurden zunächst auf Normalverteilung getestet. Dazu wurde wie zuvor beim Mittelwertvergleich der Kolmogorov-Smirnov-Test verwendet:

Kolmogorov-Smirnov-Anpassungstest

		Tool Knowledge	Domain Knowledge
N		24	24
Parameter der Normalverteilung[a,b]	Mittelwert	8,3333	3,0833
	Standardabweichung	3,19873	2,04124
Extremste Differenzen	Absolut	,157	,138
	Positiv	,121	,138
	Negativ	-,157	-,090
Kolmogorov-Smirnov-Z		,770	,676
Asymptotische Signifikanz (2-seitig)		,594	,751

Tabelle 12: Kolmogorov-Smirnov-Test für die Normalverteilung der Werte von Tool- und Domain-Knowledge

Wie bereits beschrieben, sind hier die Werte ‚Extremste Differenzen (Absolut)' sowie ‚Asymptotische Signifikanz' ausschlaggebend. In diesem Fall lassen beide auf eine Normalverteilung schließen. Im zweiten Schritt wurde die Linearität überprüft. Dazu wurde ein Streudiagramm zu jedem möglichen Zusammenhang (Tool-Knowledge & Zeit App, Tool-Knowledge & Zeit Webseite, Domain-Knowledge & Zeit App, Domain-Knowledge & Zeit Webseite) erstellt:



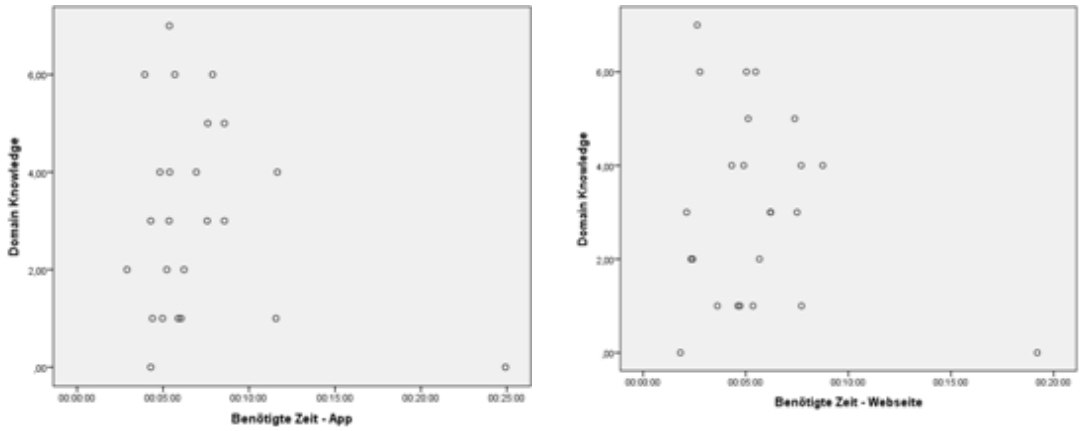


Abbildung 2: Streudiagramme bezüglich Tool-/Domain-Knowledge und der Zeiten bei App und Webseite

Auf der Y-Achse der Diagramme stehen die Werte für Tool- und Domain-Knowledge. Je höher der Wert, desto mehr Wissen ist in diesem Bereich vorhanden. Die X-Achse beschreibt die Zeit, die die Probanden bei den Testdurchläufen brauchten. Jeder Punkt innerhalb des Diagramms steht für einen Probanden. Würde man nun die Annahme zugrunde legen, dass bei einem stark ausgeprägten Wissen, eine kurze Zeit erzielt wird, müssten sich die einzelnen Punkte entlang einer Geraden von oben links nach unten rechts sammeln. Dies ist offensichtlich in keinem der Diagramme der Fall. Somit liegt keine Linearität vor, was in der Konsequenz dazu führt, dass für die Korrelationsanalyse die Spearman-Methode verwendet wird. Aus dieser Analyse ergaben sich folgende Werte:

Korrelationen

			Tool Knowledge	Domain Knowledge	Benötigte Zeit - Webseite	Benötigte Zeit - App
Spearman-Rho	Tool Knowledge	Korrelationskoeffizient	1,000	,307	-,510*	-,420*
		Sig. (2-seitig)	.	,145	,011	,041
		N	24	24	24	24
	Domain Knowledge	Korrelationskoeffizient	,307	1,000	,023	,069
		Sig. (2-seitig)	,145	.	,914	,750
		N	24	24	24	24
	Benötigte Zeit - Webseite	Korrelationskoeffizient	-,510*	,023	1,000	,607**
		Sig. (2-seitig)	,011	,914	.	,002
		N	24	24	24	24
	Benötigte Zeit - App	Korrelationskoeffizient	-,420*	,069	,607**	1,000
		Sig. (2-seitig)	,041	,750	,002	.
		N	24	24	24	24

Tabelle 13: Korrelationsanalyse nach Spearman (Tool-/Domain-Knowledge & Zeiten App/Webseite)

Wie zuvor beschrieben, sollte der p-Wert (Sig.) unter 0,05 liegen um von einem signifikanten Zusammenhang sprechen zu können. Der Wert des Korrelationskoeffizienten wird für unter 0,3 als vernachlässigbar, zwischen 0,3 und 0,7 als schwach und für über 0,7 als stark bewertet. Daher lässt sich zusammenfassend sagen, dass ein signifikanter Zusammenhang zwischen der Tool-Knowledge und der benötigten Zeit sowohl für die App als auch für die Webseite besteht, dieser allerdings als schwach bewertet werden

kann. Je besser sich die Probanden also auf den Umgang mit PC/Laptop, Smartphone, Internet und Apps im Allgemeinen verstanden, desto schneller konnten sie in beiden Fällen die Testdurchläufe absolvieren. Zwischen der Domain-Knowledge und den benötigten Zeiten besteht in beiden Fällen allerdings kein Zusammenhang. Das Wissen um Prozesse der Reiseplanung hat also keinen Einfluss auf den Umgang mit Online Anwendungen der Deutschen Bahn.

2.5.2 Subjektive Beobachtungen

Neben den geplanten Messungen sollen hier schließlich noch individuelle Beobachtung der Tester festgehalten werden. Während den Durchläufen ergab es sich, dass bestimmte Aktionen der Probanden den Testern immer wieder auffielen. Nach Absprache in der Gruppe war klar, dass viele Probanden ähnliche Schwierigkeiten bei bestimmten Punkten innerhalb des Testdurchlaufs hatten. Diese individuellen Beobachtungen waren zwar nicht in der Planung der Studie vorgesehen, sollen aber trotzdem aufgrund ihrer Auffälligkeit hier kurz stichwortartig vermerkt werden. Jedes hier beobachtete Problem hatte nach Einschätzung der Tester einen signifikanten Einfluss auf die Zeit:

- Probanden scheuen sich bei der App nach dem Finden der Hinfahrt auf den Button „Preise/Buchung“ zu klicken. Dies ist jedoch notwendig, um eine Rückfahrt hinzuzufügen. Nach einiger Zeit des Suchens betätigen die meisten Probanden den Button dann aber doch, da sie sonst keine weitere Möglichkeit sehen.
- Probanden suchen auf der ersten Seite des Eingabeformulars auf der Webseite eine Option, um den Zwischenhalt einzugeben und verwenden eine gewisse Zeit darauf. Diese Option wird allerdings erst in der erweiterten Suche sichtbar.
- Probanden suchen bei der finalen Überprüfung ihrer Reisedaten in der App die Angabe zu ihrem eingegebenen Zwischenhalt, welcher allerdings in der finalen Übersicht nicht angezeigt wird.

3 Diskussion

Bereits die Berechnung der Mittelwerte der Zeiten, die die Probanden jeweils für den Testdurchlauf mit der App und auf der Webseite benötigten, macht deutlich, dass die Studienteilnehmer deutlich besser mit der Webseite zurechtkommen. Im Durchschnitt waren die Teilnehmer hier knapp über 1,5 Minuten schneller als bei der App (App: 07:04; Webseite: 05:33). Der Vergleich der Mittelwerte ergab zusätzlich, dass dieses Ergebnis nicht zufällig entstanden ist. Man kann also davon ausgehen, dass dieser Unterschied bei allen Nutzern dieser beiden Anwendungen auftreten wird. Zwar ist der Unterschied nur gering, trotzdem existiert er. Aufgrund dessen kann der Webseite eine durchaus bessere Usability als der App bescheinigt werden.

Dieses Ergebnis spiegelt sich ebenfalls in der Auswertung der SUS-Fragebögen wider. Hier lässt sich ein noch sehr viel deutlicher Unterschied erkennen. Die Probanden bewerteten die App durchschnittlich mit 52,19, während die Webseite eine durchschnittliche Bewertung von 72,08 auf einer Skala von 0 bis 100 erhielt. Damit wird der Webseite eine allgemein gute Usability bescheinigt. Die App rangiert dagegen in einem gerade noch akzeptablen Bereich.

Bezüglich der Ergebnisse des Task Success ergibt sich allerdings ein Widerspruch zu den obigen Auswertungen. Während bei der Webseite 62,5% der Probanden einen erfolgreichen Durchlauf schafften, wurde bei der App eine 75% Erfolgsquote erreicht. Dies spricht gegen die zuvor gemachten Beobachtungen, dass die Webseite im Allgemeinen eine höhere Usability aufweist als die App, da mit diesem Ergebnis der App eine bessere Usability unterstellt wird.

Als Ursache für diesen Widerspruch kann hier die Verwendung eines ‚Binary Success' in Verbindung mit einer komplexen Aufgabe genannt werden. Innerhalb der Aufgabe mussten die Probanden viele kleinere Aktionen durchführen (Start- und Zielbahnhof, Datum, Uhrzeit, Zwischenhalt, etc. eingeben). Dadurch ergaben sich viele Möglichkeiten Fehler zu machen. Beispielsweise vergaßen die Probanden einen Zwischenhalt anzugeben oder fügten auch bei der Rückfahrt einen Zwischenhalt ein, was allerdings nicht vorgesehen war. Die Aufgabe wurde am Ende nur als erfolgreich angesehen – also mit einem Task Success Score von 1 bewertet – wenn all diese kleineren Aktionen korrekt ausgeführt wurden. Das führte dazu, dass Probanden, die nur geringfügige Fehler, wie beispielsweise das Eingeben einer falschen Uhrzeit, den gleichen Score erreichten wie Probanden, die sehr viel stärkere Fehler machten. Dies war beispielsweise der Umstand wenn keine Rückfahrt gebucht wurde. Daher wäre hier wohl die Anwendung der ‚Level of Success'-Methode angebrachter gewesen. Bei dieser Messmethode des Task Success' wird keine strikte Trennung zwischen Erfolg und Misserfolg, bezogen auf die

Gesamtaufgabe, unternommen. Stattdessen kann hier die Aufgabe in kleinere Teilaufgaben unterteilt werden, die dann hinsichtlich des Erfolgs gemessen werden. Daraus ergeben sich für jeden Probanden schlussendlich einzelne Erfolgsraten, die dann zusammengefasst werden. Erzielt ein Teilnehmer beispielsweise vier Erfolge bei fünf Teilaufgaben, würde man hier eine Erfolgsrate von 80% vergeben (Tullis & Albert, 2013).

Neben den hier bereits angesprochenen Ergebnissen wurde außerdem noch der Zusammenhang der gemessenen Zeiten und der bei den Probanden erhobenen Tool- sowie Domain-Knowledge ermittelt. Diese Auswertung war zwar in der Planung der Studie zunächst nicht vorgesehen, erschien jedoch im Verlauf als ein interessanter Punkt, den es wert war zu untersuchen. Es zeigte sich, dass Tool-Knowledge, also das Wissen um den Umgang mit der Technik, dazu führt, dass Fahrplanauskünfte und Buchungen mit der App und der Webseite der Deutschen Bahn schneller durchgeführt werden können. Dies erscheint nur logisch, da die beiden Anwendung natürlich auf die Verwendung von Smartphone und PC angewiesen sind. Allerdings lässt dieses Ergebnis auf der anderen Seite den Rückschluss zu, dass weniger technikversierte Nutzer Probleme beim Buchungsprozess haben könnten. Hier bestünde die Möglichkeit, diese User mit entsprechenden Hilf-Systemen zu unterstützen. Ob dies überhaupt nötig ist, und falls ja, inwiefern Verbesserungen durchgeführt werden müssten, kann durch dieses Ergebnis allerdings nicht bestimmt werden. Bezüglich des Zusammenhangs von Domain-Knowledge und der Zeiten lässt sich festhalten, dass dieser nicht existiert. Bahn-Nutzern bringt es also nichts, wenn sie besondere Vorkenntnisse in Sachen Reiseplanung haben. Der Buchungsprozess kann dadurch nicht schneller vollzogen werden. Hier muss allerdings zusätzlich angemerkt werden, dass die Werte der Tool-Knowledge durchweg höher ausfielen als die der Domain-Knowledge. Es wäre vorstellbar, dass ein Zusammenhang zwischen den Zeiten und der Domain-Knowledge bei höheren Domain-Knowledge Werten vorliegt. Dies kann durch die vorliegende Studie allerdings nicht beantwortet werden. Eine Aussage über die Usability aufgrund dieser Ergebnisse lässt sich nur schwer treffen. Am auschlaggebendsten dürfte hier die oben genannte Vermutung sein, dass Nutzer mit geringer Tool-Knowledge Schwierigkeiten bei der Benutzung haben könnten. Hier wären weiterführende Untersuchungen sinnvoll. Für User, die sich auf den grundlegenden Umgang mit den erforderlichen Technologien verstehen, kann die Usability allerdings als gut bewertet werden.

Eine zweite Reihe von Beobachtungen, die die Planungen der Studie ebenfalls nicht beinhalteten, waren die als ‚subjektive Beobachtungen' beschriebenen Punkte. Diese stellten sich als höchst nützlich heraus. Durch sie konnten Problemstellen innerhalb der Anwendungen identifziert werden, die mit den übrigen Messinstrumenten nicht zum Vorschein gekommen wären. Hier lässt sich zusammenfassen, dass Beschriftungen von Buttons innerhalb der Anwendungen höchst sensibel vorgenommen werden sollten, da

dies ein entscheidendes Kriterium dafür war, dass Nutzer mehr Zeit innerhalb der Testdurchläufe verwendeten. So war es nur möglich, eine Rückfahrt zu einer Buchung innerhalb des DB-Navigators hinzuzufügen, wenn zuvor der Button mit der Beschriftung „Preise/Buchung“ geklickt wurde. Die Probanden hatten allerdings mehrheitlich die Befürchtung, dass damit schon die finale Buchung unternommen werden würde. Eine andere Gestaltung würde hier zu einer besseren Usability Performance führen.

Desweiteren stellte sich das Nichtvorhandensein eines Eingabefeldes für einen Zwischenhalt auf der Startseite der Web-Anwendung als hinderlich dar. Hier verbrachten die Probanden ebenfalls eine erhebliche Zeitspanne damit, das entsprechende Eingabefeld zu suchen. Dieses wurde allerdings erst sichtbar, als die Probanden zum Interface der erweiterten Suche gelangten. Und selbst hier wurde die Option nur einzeilig dargestellt und lediglich mit einem kleinen Pfeil signalisiert, sodass auch hier vereinzelt Probleme mit der Eingabe des Zwischeninhalts auftauchten. Die Probanden sahen schlichtweg die Möglichkeit der Eingabe nicht. Eine Lösung für dieses Problem könnte beispielsweise eine Einbettung des Einzeilers der erweiterten Suche auf der Startseite sein, da dadurch nur verhältnismäßig wenig Platz weggenommen werden würde. In der erweiterten Suche könnte dafür ein bereits offenes Eingabefeld erscheinen, welches nicht erst durch das Klicken der Zeile „Zwischenhalt angeben“ aufgerufen werden müsste. Da allerdings keine Daten bezüglich der Nutzungshäufigkeit der Zwischenhalt Option vorlagen, besteht durchaus die Möglichkeit, dass aufgrund geringer Verwendung, eine eher zurückhaltende Gestaltung durchaus sinnvoll ist.

Ein drittes auffälliges Phänomen, welches bei den Testdurchläufen beobachtet werden konnte, war die Tatsache, dass der DB-Navigator in der finalen Übersicht der Reisedaten den eingegebenen Zwischenhalt nicht anzeigte. Auch dieses Problem stellte sich als zeitraubend heraus. Die Probanden waren sich unsicher, ob sie einen Zwischenhalt eingegeben hatten. Teilweise führte dies dazu, dass einige Teilnehmer mehrere Schritte zurück gingen, um ihre Daten zu überprüfen, obwohl sie nur den Button für die Buchung hätten klicken müssen. Hier sollte der Zwischenhalt in der finalen Übersicht der Reisedaten angezeigt werden, da dies zu einer erheblichen Verbesserung der Usability des DB-Navigators führen würde.

Gerade der erste und dritte Punkt der zuletzt angeführten Beobachtung wirkte sich stark auf die aufgewendete Zeit bei den Testdurchläufen der App aus. Durch einen vergleichsweise geringen Aufwand (Änderung von Bezeichnungen, Anzeigen der kompletten Reisedaten) würde sich die Usability des DB-Navigators enorm steigern.

Bezüglich der Durchführung der Studie kann abschließend festgehalten werden, dass zwei von drei der zu Beginn ausgewählten Metriken zur Messungen der Usability wertvolle Ergebnisse erbracht haben. Leider führte das dritte Messinstrument zu widersprüchlichen Ergebnissen. Hier wäre eine genauere Anpassung der Messmethode an den

Umfang der Aufgabe innerhalb der Studie sinnvoll gewesen. Allgemein kann festgehalten werden, dass bei komplexen Aufgabenstellungen geprüft werden sollte, ob binäre Messmethoden ausreichen oder eventuell unbrauchbare oder verzerrte Ergebnisse liefern. Außerdem wurde deutlich, dass individuelle Beobachtungen von Testern ein sehr ergiebiges Messinstrument darstellen. Besonders weil dadurch oftmals Probleme identifiziert werden können, die durch statistische Messungen nicht auffallen würden.

Hinsichtlich der Usability von Webseite und App (bezogen auf das Heraussuchen die Buchung von Verbindungen) lässt sich abschließend sagen, dass die Webseite nutzerfreundlicher ist. Allerdings ist es durchaus vorstellbar, dass durch geringere Verbesserungen der App der Unterschied zwischen beiden Anwendungen verringert werden kann. Da festgestellt wurde, dass der Unterschied eher schwach zu bewerten ist, könnte dies sogar dazu führen, dass die App – gerade bezogen auf die Zeiten – ein gleichwertiges Usability-Level erreichen kann. Ein Gleichziehen der App mit der Webseite bezüglich des SUS-Scores scheint weniger wahrscheinlich, da der Unterschied hier stärker ausgefallen ist. Hier würde es sich anbieten,in weiterführenden Studien andere Self-Reported Metrics zu verwenden, die einen detaillierteren Blick auf die Unterschiede zwischen App und Webseite ermöglichen.

Literaturverzeichnis

Bangor, A., Kortum, P., & Miller, J. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction* , 574-594.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler.* Heidelberg: Springer.

Ebermann, E. (18. Juni 2010). *Grundlagen statistischer Auswertungsverfahren*. Abgerufen am 12. August 2014 von Test auf Normalverteilung: http://www.univie.ac.at/ksa/elearning/cp/quantitative/quantitative-59.html

Field, A., & Hole, G. (2003). *How to design and report experiments.* London: SAGE Publications Ltd.

Keller, D. (16. Mai 2013). *Statistik und Beratung*. Abgerufen am 13. August 2014 von Analyse von Zusammenhängen: Korrelation: http://www.statistik-und-beratung.de/2013/05/analyse-von-zusammenhangen-korrelation/

Kutter, K. (28. Januar 2011). *Seibert Media Webblog*. Abgerufen am 8. August 2014 von Test Verbindungssuche: Mit der Bahn von Frankfurt nach London: http://blog.seibert-media.net/blog/2011/01/28/test-verbindungssuche-mit-der-bahn-von-frankfurt-nach-london/

Reimamann, A., & Kluge, U. (2012). Die Bahn macht mobil(e). *Internet World Business* , 48.

Rubin, J., & Chisnell, D. (2008). *Handbook of Usability Testing - How to plan, design and conduct effective tests.* Indianapolis: Wiley Publishing, Inc.

Sauro, J. (August 2011). *User Experience The Magazine of the User Experience Professionals Association*. Abgerufen am 8. August 2014 von SUStisfied? Little-Known System Usability Scale Facts: http://www.usabilityprofessionals.org/uxmagazine/sustified/

Sauro, J., & Lewis, J. (2012). *Quantifying the user experience - practical statistics for user research.* Waltham: Elsevier Inc.

Tullis, T., & Albert, B. (2013). *Measuring the User Experience - Collecting, Analyzing, and Presenting Usability Metrics.* Waltham: Elsevier Inc.

Werhahn, H. (24. September 2008). *Praxis & Wissen*. Abgerufen am 8. August 2014 von HVV und VRS: Vom Internet ins Ungewisse: http://www.fit-fuer-usability.de/archiv/hvv-und-vrs-vom-internet-ins-ungewisse/

Anhang

1. SUS Beispielfragebogen

SUS

1. **Ich kann mir sehr gut vorstellen, das System regelmäßig zu nutzen.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

2. **Ich empfinde das System als unnötig komplex.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

3. **Ich empfinde das System als einfach zu nutzen.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

4. **Ich denke, dass ich technischen Support brauchen würde, um das System zu nutzen.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

5. **Ich finde, dass die verschiedenen Funktionen des Systems gut integriert sind.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

6. **Ich finde, dass es im System zu viele Inkonsistenzen gibt**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

7. **Ich kann mir vorstellen, dass die meisten Leute das System schnell zu beherrschen lernen.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

8. **Ich empfinde die Bedienung als sehr umständlich.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

9. **Ich habe mich bei der Nutzung des Systems sehr sicher gefühlt.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu

10. **Ich musste eine Menge Dinge lernen, bevor ich mit dem System arbeiten konnte.**

	1	2	3	4	5	
Stimme überhaupt nicht zu	◯	◯	◯	◯	◯	Stimme vollkommen zu