Konstruktion eines Fragebogens zur Evaluation der User Experience bei Medizinprodukten

Construction of questionnaire on user experience for medical devices

Masterarbeit

im

Masterstudiengang Psychologie

der

Otto-Friedrich-Universität Bamberg

in Kooperation mit der Siemens Healthcare GmbH

Verfasserin: Hannah Luisa Bögler (Matrikelnummer: 2032563)

Betreuung Universität Bamberg: Prof. Dr. Astrid Schütz und Dr. Lukas Röseler

Betreuung Siemens Healthcare GmbH: M.Sc. Magdalena Feuerbach

Bamberg, 16. Dezember 2022

Zusammenfassung

Die Siemens Healthcare GmbH entwickelt Medizinprodukte, die von medizinischem Personal bedient werden. Für eine standardisierte Erfassung der User Experience (UX) soll ein UX-Fragebogen speziell für Medizinprodukte konstruiert werden. In Vorstudien wurde ermittelt, welche sechs UX-Aspekte bei Medizinprodukten besonders relevant sind. Vier der sechs wichtigsten Skalen aus der letzten Vorstudie (Steuerbarkeit, Effizienz, Vertrauen und Übersichtlichkeit) sind Teil eines bestehenden Fragebogens, des User Experience Questionnaire +, kurz UEQ+ (Schrepp & Thomaschewski, 2019a). Für die Skalen Sicherheit und Effektivität existierten dagegen noch keine Items. Zur Generierung und Konsolidierung neuer Items für diese beiden Skalen wurden drei interne Expert*innen-Workshops bei Siemens Healthineers durchgeführt. Anschließend wurden die Items im Rahmen einer externen Online-Studie an medizinischem Personal evaluiert. Im Rahmen der Studie wurde weiterhin überprüft, ob sich die Ergebnisse der letzten Vorstudie replizieren ließen. Fünf der sechs Skalen aus der Vorstudie fielen auch in der Replikationsstudie unter die sechs produktübergreifend wichtigsten Skalen. Eine Ausnahme bildete die Skala Sicherheit, welche durch die Skala Nützlichkeit ersetzt wurde. Eine produktübergreifende Version des UX-Fragebogens für Medizinprodukte, bestehend aus den sechs wichtigsten Skalen der Replikationsstudie, wird als angemessen betrachtet. Es werden optionale Zusatzskalen empfohlen. Die Evaluation der Sicherheitsitems ergab zwei optionale Sicherheitsskalen, eine allgemeine und eine Hardware-spezifische Skala im Format des UEQ+. Die Skala Effektivität erwies sich als eindimensional und kann nun als Bestandteil des produktübergreifenden UX-Fragebogens eingesetzt werden. Anhand eines Strukturgleichungsmodells wurde auf einen akzeptablen Modell-Fit geschlossen. Die interne Konsistenz der Skalen fiel hoch aus, während eine geringe Interrater-Reliabilität im Hinblick auf ein ausgewähltes Röntgensystem vorlag. Eine hohe Korrelation zur Gesamtzufriedenheit mit dem Produkt ließ auf konvergente Validität schließen. Es zeigten sich zudem erste Hinweise auf diskriminante Validität, jedoch sollte die Güte des UX-Fragebogens für Medizinprodukte in Zukunft noch genauer untersucht werden.

Stichwörter: User Experience, Usability, Fragebogen, quantitativ, UEQ+, Testkonstruktion, Medizinprodukte, medizinisches Personal

Abstract

Siemens Healthcare GmbH develops medical devices that are operated by medical personnel. For a standardized measurement of user experience (UX) a questionnaire specifically for medical devices is to be constructed. Preliminary studies have determined which six UX aspects are particularly relevant for medical devices. Four of the six most important scales from the last preliminary study (Dependability, Efficiency, Trust, and Clarity) are part of an existing questionnaire, the User Experience Questionnaire +, UEQ+ for short (Schrepp & Thomaschewski, 2019a). For the scales Safety and Effectiveness, however, no items existed yet. Three internal expert workshops were held at Siemens Healthineers to generate and consolidate new items for these two scales. Subsequently, the items were evaluated in an external online study involving medical personnel. The study also examined whether the results of the last preliminary study could be replicated. Five of the six scales from the preliminary study were also among the six most important scales across all products in the replication study. One exception was the Safety scale, which was replaced by the Usefulness scale. A cross-product version of the UX questionnaire for medical devices, consisting of the six most important scales of the replication study, is considered appropriate. Optional additional scales are recommended. The evaluation of the safety items resulted in two optional Safety scales, a general and a hardware-specific scale in the format of the UEQ+. The Effectiveness scale proved to be one-dimensional and can now be used as a component of the cross-product UX questionnaire. A structural equation model revealed an acceptable model fit. A high internal consistency of all scales could be demonstrated, whereas the interrater reliability regarding a selected X-ray system was low. High correlation with overall satisfaction with the product suggested convergent validity. Furthermore, first evidence of discriminant validity was found, but the quality of the UX questionnaire should be investigated in more detail in the future.

Keywords: user experience, usability, questionnaire, quantitative, UEQ+, test construction, medical devices, medical personnel

Abbildungsverzeichnis

Tabellenverzeichnis

Inhaltsverzeichnis

Konstruktion eines Fragebogens zur Evaluation der User Experience bei Medizinprodukten

Eine gute User Experience (UX) gewinnt zunehmend an Bedeutung, wenn es um die Entwicklung und Evaluation neuer Lösungen, Systeme und Produkte in verschiedenen Bereichen geht (Ariza & Maya, 2014; Robinson et al., 2018). Beispiele sind Geräte wie Smart TVs (Lim et al., 2019) oder Assistenzroboter (Kim et al., 2022), genauso wie Web-Applikationen (Mistry & Rajan, 2019) und Softwareanwendungen (Kashfi et al., 2017). Laut der Norm 9241 der International Organization for Standardization (ISO), Teil 210, die sich mit der menschzentrierten Gestaltung interaktiver Systeme beschäftigt, umfasst die UX „user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" (2019a, Abs. 3.15). Die Evaluation der UX in der Produktentwicklung verfolgt das Ziel, eine effiziente, intuitive sowie zielgerichtete Nutzung neuer Lösungen sicherzustellen. „Ein Produkt kann einfach oder schwierig zu benutzen sein. Es kann kompliziert oder intuitiv sein, verständlich oder unverständlich, effizient oder mühsam" (Richter & Flückiger, 2016, S. 10). Insbesondere in Bezug auf Medizinprodukte spielt die UX eine nicht zu unterschätzende Rolle, denn die Kosten von unverständlichen, umständlichen Produkten und daraus resultierenden Fehlern in der Anwendung können hoch sein (Clark & Israelski, 2012). UX-Testungen eröffnen die Möglichkeit, die Nutzung von Medizinprodukten effektiver und sicherer zu gestalten (Bitkina et al., 2020; International Organization for Standardization, 2015). Gerade wenn es um die Evaluation von komplexen Geräten geht, die durch medizinisches Personal bedient werden, müssen allerdings andere UX-Faktoren einbezogen werden als bei Softwareanwendungen oder Apps, die für die breite Bevölkerung ausgelegt sind. Bestehende UX-Fragebögen eignen sich daher nicht optimal für die Anwendung bei Medizinprodukten. Ziel der vorliegenden Arbeit ist es deshalb, einen standardisierten UX-Fragebogen speziell für Medizinprodukte zu entwickeln. Aufgrund der Ausrichtung der Siemens Healthcare GmbH (Siemens Healthineers), die diese Arbeit in Auftrag gibt, geht es insbesondere um die UX im Hinblick auf Medizinprodukte, die von Siemens Healthineers entwickelt werden. Die Produkte, auf die der UX-Fragebogen ausgelegt sein soll, werden von medizinischem Personal bedient, weshalb der Fragebogen auf die Erfassung der UX aus Sicht von Ärzt*innen, medizinisch-technischen Assistent*innen und weiteren Berufsgruppen dieser Sparte ausgerichtet ist.

1 Theoretischer Hintergrund

In diesem Kapitel werden die zentralen Begriffe der Usability und User Experience definiert und ihre Bedeutung für die Medizinbranche und Firmen wie Siemens Healthineers erläutert.

1.1 Usability und User Experience

Das Konzept der *Usability* geht bereits auf die 1940er Jahre zurück, als das US-amerikanische Militär begann, an der Verbesserung von Mensch-Maschine-Interaktionen bei komplexen Systemen zu forschen (Richter & Flückiger, 2016). Gemeint ist damit die Gebrauchstauglichkeit oder Benutzbarkeit von technischen Systemen, die häufig nur dann explizit wahrgenommen wird, wenn sie besonders schlecht erfüllt ist (Jacobson & Meyer, 2019). Die ISO definiert Usability als „extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use“ (2018, Abs. 3.1.1). Ob ein System gebrauchstauglich ist, lässt sich demnach nur unter Berücksichtigung der Ziele bewerten, für das ein*e Anwender*in es einsetzt (Richter & Flückiger, 2016). Qualitätskriterien für gute Usability sind nach der DIN EN ISO 9241-151 (International Organization for Standardization, 2008), hier allerdings bezogen auf die Gestaltung von Webseiten, beispielsweise die Aufgabenangemessenheit oder die Fehlertoleranz (Jacobson & Meyer, 2019).

Lange Zeit stand diese pragmatische und aufgabenbezogene Herangehensweise an Mensch-Technik-Schnittstellen im Vordergrund, jedoch mehrte sich Kritik an der engen Konzeptionalisierung von Usability. Neben der reinen Nützlichkeit eines Produkts wurden zunehmend auch Aspekte wie Ästhetik, Emotionen, Werte oder Einstellungen der Nutzer*innen einbezogen (Hassenzahl & Tractinsky, 2006; Minge & Riedel, 2013; Minge & Thüring, 2011; Thüring & Mahlke, 2007), um eine ganzheitliche Erfahrung mit einem Produkt zu beschreiben. Seit dieses erweiterte Konstrukt der *User Experience* (UX) in den 1990er Jahren aufkam (D. Norman et al., 1995), entstanden zahlreiche Modelle und Definitionen, mit dem Ziel, die verschiedenen Dimensionen der UX und Einflussgrößen auf sie greifbar zu machen (Ariza & Maya, 2014; Berni & Borgianni, 2021; Hinderks, Winter et al., 2019; Lallemand et al., 2015; Law et al., 2008). Die UX beschreibt damit die umfassende Nutzungserfahrung, wohingegen die Usability einen Teilaspekt der UX darstellt (Jacobson & Meyer, 2019).

Einige UX-Aspekte der verschiedenen Modelle überlappen sich inhaltlich, jedoch gibt es auch Unterschiede. Als modellübergreifend wesentliche Komponenten, die Einfluss auf die UX nehmen, lassen sich der oder die Nutzer*in, das System und der

Nutzungskontext betrachten (Ariza & Maya, 2014; Berni & Borgianni, 2021; Hassenzahl & Tractinsky, 2006; Roto et al., 2011; Thüring & Mahlke, 2007). Eine zentrale Rolle spielen hierbei nicht nur Ziele der Anwender*innen, sondern auch subjektive Variablen, wie ihre Erwartungen, Bedürfnisse, Motive und Emotionen (Hassenzahl et al., 2009; Hassenzahl, 2011; Hassenzahl & Tractinsky, 2006). Das System/Produkt beeinflusst die UX unter anderem durch seine Komplexität und den angedachten Verwendungszweck (Hassenzahl & Tractinsky, 2006). Ariza und Maya (2014) unterscheiden bezüglich des Produkts zwischen instrumentellen Eigenschaften, wie der Funktionalität und nicht-instrumentellen Merkmalen, wie der Ästhetik. Unter Kontextfaktoren fallen soziale und physikalische Aspekte der Situation oder der Umgebung. Beispielsweise könnte die UX dadurch beeinflusst werden, ob das Produkt im Bus oder während der Arbeit genutzt wird oder ob andere Aufgaben zeitgleich Aufmerksamkeit in Anspruch nehmen (Roto et al., 2011). Dass die UX umfassender ist als Usability wird auch daran deutlich, dass sie sich nicht nur auf die Wahrnehmung während der Interaktion mit dem Produkt bezieht, sondern auch auf Prozesse vor und nach der Interaktion (International Organization for Standardization, 2019a; Roto et al., 2011). Die UX kann während der einzelnen Interaktionsphasen evaluiert werden, oder kumulativ, wenn es um die Bewertung des Produkts geht, nachdem der/die Nutzende sich eine Zeit lang mit ihm auseinandersetzen konnte (Ariza & Maya, 2014; Roto et al., 2011). Es gibt noch weitere Dimensionen, nach denen sich UX klassifizieren oder einordnen lässt. Ariza und Maya (2014) differenzieren bezüglich der Art der Interaktion zwischen aktiver und passiver Nutzung, je nachdem, ob ein Produkt aktive physische Aktionen erfordert. Berni und Borgianni (2021) unterscheiden zusätzlich drei verschiedene Arten der Interaktionserfahrung, auf die in der Literatur Bezug genommen wird. Die ergonomische Erfahrung bezieht sich auf die reine Zielerreichung und den Aufwand, der mit der Interaktion verbunden ist. Die kognitive Erfahrung dagegen beinhaltet Gedanken und Wahrnehmungsprozesse, während die emotionale Erfahrung Freude oder andere Gefühle während der Interaktion umfasst.

Es ist festzustellen, dass UX facettenreich und dynamisch ist, weshalb es schwierig ist, eine einheitliche Definition zu formulieren (Lallemand et al., 2015; Law et al., 2008). Die gängigste Definition (Richter & Flückiger, 2016), die die bestehende Literatur außerdem am ehesten zusammenfasst (Berni & Borgianni, 2021), ist die der ISO (2019a). Nach der DIN EN ISO 9241-210 wird UX definiert als „user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" (Abs. 3.15). Unter Wahrnehmungen und Reaktionen fallen hierbei „emotions, beliefs, preferences,

perceptions, comfort, behaviours, and accomplishments that occur before, during and after use“ (International Organization for Standardization, 2019a, Abs. 3.15). Die Nutzungserfahrung ist dabei laut dieser Definition das Ergebnis verschiedener Einflüsse, wie dem Markenimage oder der Systemleistung. Darüber hinaus nehmen laut der ISO-Definition auch der psychische und physische Zustand der Benutzerin oder des Benutzers Einfluss auf die UX. Dieser wird wesentlich durch „prior experiences, attitudes, skills, abilities and personality“ (Abs. 3.15) sowie den Nutzungskontext bestimmt.

Die Erfassung und Verbesserung von UX ist immer dann von Interesse, wenn eine Interaktion mit technischen Systemen gegeben oder geplant ist (Richter & Flückiger, 2016). Eine wesentliche Annahme, welche die menschzentrierte Gestaltung technischer Systeme begründet, ist, dass eine positive User Experience nicht zufällig entsteht, sondern durch gezieltes Design verbessert werden kann (Winter, 2020). Dazu müssen die Perspektiven und Bedürfnisse von Anwender*innen als Gestaltungsgrundlage für interaktive Produkte dienen (Hassenzahl, 2011). Im Zuge der fortschreitenden Digitalisierung gewinnt die UX somit an Bedeutung und breitet sich in verschiedenen Bereichen aus (Robinson et al., 2018). Insbesondere im Hinblick auf Medizinprodukte, die zunehmend komplexer werden, werden Maßnahmen zur Evaluation und Verbesserung der UX verstärkt (Bitkina et al., 2020).

1.2 User Experience und Usability bei Medizinprodukten

Entsprechend der Definition der World Health Organization (WHO) werden Medizinprodukte wie folgt definiert:

> An article, instrument, apparatus or machine that is used in the prevention, diagnosis or treatment of illness or disease, or for detecting, measuring, restoring, correcting or modifying the structure or function of the body for some health purpose. Typically, the purpose of a medical device is not achieved by pharmacological, immunological or metabolic means. […] This category includes medical equipment, implantables, single use devices, in vitro diagnostics and some assistive technologies. (2017, S. 10)

Medizinprodukte werden häufig in direktem Kontakt zu Patient*innen eingesetzt (Bundesinstitut für Arzneimittelsicherheit und Medizinprodukte, 2022) und kommen in der medizinischen Versorgung immer mehr zum Einsatz. Der technische Fortschritt ermöglicht eine rasche Weiterentwicklung, bringt allerdings auch zunehmend anspruchsvollere Bedienweisen der Produkte mit sich (Kennedy, 2018). Schwierigkeiten bei der Anwendung werden somit wahrscheinlicher, gerade wenn Medizinprodukte nicht unter

Gesichtspunkten der Gebrauchstauglichkeit und Nutzerfreundlichkeit entwickelt wurden (International Organization for Standardization, 2015). Anders als bei Freizeitprodukten gehen mit anwenderunfreundlichen Medizinprodukten Risiken einher, die weitreichende Folgen haben können (Hyman, 2018; Kennedy, 2018; Rose et al., 2018; Zhang et al., 2005). Fehler in der Anwendung sind eine der Hauptursachen für Komplikationen im medizinischen Bereich. Eine Analyse von Montag (2012) zeigte, dass von den untersuchten 1222 Vorkommnissen während Operationen 34.3 % auf reine Usability-Ursachen zurückzuführen waren. Anwendungsfehler beruhten beispielsweise darauf, dass der Zustand des Geräts für den/die Anwender*in nicht ersichtlich war, dass Kennzeichnungen fehlten oder das Gerät auf Fehler nicht robust reagierte. Ein besonders häufiger Fehler, der von Anwender*innen ausging, war eine falsche, nicht bestimmungsgemäße Anwendung des Geräts. Weitere Anwenderfehler umfassten Defizite bei der Wartung, der Reinigung oder unterlassene Funktionstests. Folgen der Vorkommnisse beliefen sich teilweise auf eine Verlängerung der Operationszeit, in anderen Fällen wurden Personen gefährdet oder sogar verletzt.

Die Zunahme an Studien und internationalen Standards zu den Themen Usability und User Experience bei Medizinprodukten untermauern ihre Bedeutung (Bitkina et al., 2020; International Organization for Standardization, 2016). In der IEC 62366 der ISO zur Anwendung von Usability Engineering bei Medizinprodukten werden Usability-Anforderungen an Medizinprodukte festgehalten (2015). Die Norm regelt beispielsweise, wie Hersteller den Prozess zur Gewährleistung der Usability bei Medizinprodukten gestalten sollten, um Sicherheit zu garantieren. Einen Leitfaden für Hersteller von Medizinprodukten gibt es auch von der US-amerikanischen Food and Drug Administration (FDA) (2016a). Es wird dabei sowohl auf die Anwender*innen (z.B. mindestens erforderliche Körpergröße) als auch die Anwendungsumgebung (z.B. Lichtverhältnisse) und die Benutzeroberfläche (z.B. Kennzeichnungen) Bezug genommen, um alle Komponenten im Design-Prozess zu berücksichtigen. Die ISO-Norm 14971 (2019b), die Empfehlungen zum Risikomanagement bei Medizinprodukten beinhaltet, spezifiziert, dass der Hersteller definieren sollte, worin eine intendierte Nutzung des Systems besteht. Neben der medizinischen Indikation sollte in der Dokumentation zum Risikomanagement, übereinstimmend mit dem Leitfaden der FDA, festgehalten werden, welche Voraussetzungen Nutzer*innen erfüllen müssen und wie eine geeignete Anwendungsumgebung gestaltet sein sollte.

Es kann festgehalten werden, dass die Verbesserung der UX einen Mehrwert für Anwender*innen bieten kann. Insbesondere im Hinblick auf Medizinprodukte ist eine intuitive, sichere und effiziente Bedienung von großem Wert. Fehler und Schwierigkeiten bei der Anwendung können durch UX-Evaluationen frühzeitig erkannt und ausgeräumt werden (Bitkina et al., 2020; Kennedy, 2018). Welche Methoden zur Evaluation der UX zur Verfügung stehen wird im Folgenden beschrieben.

1.3 Methoden zur Erfassung von UX

Ansätze zur Beurteilung der UX lassen sich in quantitative und qualitative Methoden unterteilen (Bargas-Avila & Hornbæk, 2011; Lanius et al., 2021). Vorteile quantitativer Methoden bestehen darin, dass ihr Einsatz kosten- und zeiteffizient ist und das hohe Maß an Standardisierung eine Vergleichbarkeit von Produkten sowie Verlaufskontrollen ermöglicht (Bitkina et al., 2020). Andererseits geben quantitative Methoden keinen Aufschluss darüber, welche Störfaktoren konkret bestehen und wie die UX verbessert werden kann. Hier liefern qualitative Methoden detaillierte Informationen (Budiu, 2017). Um der Komplexität von UX gerecht zu werden und Bedürfnisse der Nutzer*innen umfassend zu verstehen, hat sich eine Kombination beider Herangehensweisen bewährt (Darin et al., 2019; Robinson et al., 2018; Sauro, 2016).

1.3.1 Qualitative Methoden.

Zu qualitativen Methoden zählen beispielsweise Usability-Tests, Interviews, Beobachtungen, Expert*innen-Reviews oder Card Sorting-Verfahren (Jacobson & Meyer, 2019). Bei *Usability-Tests* werden Personen der Zielgruppe bei der Erledigung typischer Aufgaben am Produkt beziehungsweise an einem Prototyp beobachtet und befragt. Verhaltensweisen während der Interaktion sowie mögliche Schwierigkeiten können so nachvollzogen werden. Für *Expert*innen-Reviews* wird häufig ein aufgabenbasierter Ansatz (*Cognitive Walkthrough*) herangezogen, bei dem Expert*innen ebenfalls typische Nutzungsszenarien durchlaufen und reflektieren. *Heuristische Evaluationen* im Rahmen von Expert*innen-Reviews erfüllen den Zweck, zu ermitteln, ob die Gestaltung des Systems UX-Heuristiken und Richtlinien entspricht (Held et al., 2019; Jacobson & Meyer, 2019; Nielsen, 1994, 1994). UX-Heuristiken sind gängige und anerkannte UX-Prinzipien. Ein Beispiel wäre das Vorhandensein eines Hilfe-Buttons und einer Dokumentation über sämtliche Interaktionsmöglichkeiten (Nielsen, 2020). Eine weitere qualitative Methode, das sogenannte *Card Sorting*, ist hilfreich für eine sinnvolle Strukturierung von Informationen. Mithilfe von Karten sollen Begriffe Kategorien zugeordnet werden, wodurch ein besseres Verständnis des mentalen Modells der Nutzer*innen gewonnen werden kann

(Jacobson & Meyer, 2019). Einen Überblick über qualitative UX-Methoden, auf die an dieser Stelle nicht detaillierter eingegangen wird, bieten zum Beispiel Richter und Flückiger (2016) oder Jacobson und Meyer (2019).

1.3.2 Quantitative Methoden.

Es existiert darüber hinaus eine Reihe quantitativer UX-Methoden, es werden jedoch nur einige Beispiele zur Veranschaulichung herausgegriffen. Eine quantitative Herangehensweise zur Evaluation der UX ist beispielsweise das *virtuelle Eye-Tracking*, bei dem die Augenbewegungen automatisiert durch Software ausgewertet werden. So kann nachempfunden werden, auf welche Bereiche eines Produkts sich die Aufmerksamkeit der Testpersonen konzentriert (Marucci, 2019). Eine weitere Möglichkeit zur Beschaffung quantitativer Daten sind sogenannte *Clickstream-Analysen* (Rohrer, 2022). Hierbei werden zum Beispiel Informationen darüber gesammelt, welche Elemente einer Webseite typischerweise nacheinander angeklickt werden. Eine beliebte quantitative Methode ist außerdem die Nutzung von UX-Fragebögen, auf die vor dem Hintergrund der vorliegenden Arbeit genauer eingegangen wird. Es existiert bereits eine Reihe wissenschaftlich entwickelter und validierter Fragebögen zur quantitativen Beurteilung der UX und Usability. Diese lassen sich im Wesentlichen in zwei Gruppen einordnen. Im Zeitraum von 1986 bis 1999 wurden überwiegend Usability-Fragebögen entwickelt, wohingegen ab dem Jahr 2003 neben Usability-Kriterien zunehmend die User Experience im Ganzen betrachtet wurde (Hinderks, Winter et al., 2019). Fragebögen, welche insbesondere Usability-Aspekte erfassen, sind beispielsweise die *System Usability Scale* (SUS) (Brooke, 1996), der *Questionnaire for User Interface Satisfaction* (QUIS) (Chin et al., 2014) oder das *Software Usability Measurement Inventory* (SUMI) (Kirakowski, 1996). Zu neueren Fragebögen, die häufig zusätzlich Emotionen und hedonische Qualität berücksichtigen, zählen unter anderen der *AttrakDiff 2* (Hassenzahl et al., 2003) oder der *User Experience Questionnaire* (UEQ) (Laugwitz et al., 2008). Eine Übersicht über bestehende Usability- und UX-Fragebögen bietet Tabelle 1.

Tabelle 1

Übersicht über bestehende UX-Fragebögen

Autor*in	Bezeichnung des Fragebogens
Brooke (1996)	System Usability Scale (SUS)
Chin et al. (2014)	Questionnaire for User Interface Satisfaction (QUIS)
Davis (1989)	Perceived Usefulness and Ease of Use (PUEU)
Kirakowski & Corbett (1993)	Software Usability Measurement Inventory (SUMI)
Prümper & Anft (1993)	IsoNorm (IsoNorm)
Lewis (1995)	Post Study System Usability Questionnaire (PSSUQ)
Lin et al. (1997)	Purdue Usability Testing Questionnaire (PUTQ)
Gediga et al. (1999)	IsoMetrics usability inventory
Hassenzahl et al. (2003)	AttrakDiff2
Laugwitz et al. (2008)	User Experience Questionnaire (UEQ)
Moshagen & Thielsch (2010)	Visual Aesthetics of Websites Inventory (VisAWI)
Minge & Riedel (2013)	Modular Evaluation of Key Components of User Experience (meCUE)
Kothgassner et al. (2013)	Technology Usage Inventory (TUI)
Sauro (2015)	Standardized User Experience Percentile Rank Questionnaire (SUPR-Q)
Schrepp & Thomaschweski (2019b)	User Experience Questionnaire + (UEQ+)
Zhou et al. (2019)	mHealth App Usability Questionnaire (MAUQ)

Anmerkung. Die Tabelle ist angelehnt an eine Übersicht von Hinderks, Winter et al. (2019, S. 1723), wurde jedoch verändert und ergänzt.

Das Problem der bisher aufgeführten Fragebögen besteht darin, dass relevante UX-Aspekte sich je nach Produkt und Kontext deutlich unterscheiden (Winter et al., 2015; Winter et al., 2017). Bisherige Fragebögen beinhalten Skalen, die für manche Produkte überflüssig sind, während andere Aspekte fehlen (Hinderks, Winter et al., 2019). Beispielsweise sollte die Haptik bei der Evaluation von Haushaltsgeräten einbezogen werden, wohingegen sie bei der Beurteilung einer Softwareanwendung ungeeignet ist (Schrepp & Thomaschewski, 2020). Der UEQ (Laugwitz et al., 2008) bietet beispielsweise zur Erfassung der pragmatischen Qualität die drei Skalen Effizienz, Durchschaubarkeit und Steuerbarkeit. Darüber hinaus beinhaltet er zur Erfassung der hedonischen Qualität die beiden Skalen Stimulation und Neuheit sowie zusätzlich die Skala Attraktivität. Diese Skalen lassen sich bereits auf zahlreiche Produkte anwenden, jedoch gibt es Anwendungskontexte, wie Medizinprodukte, auf die diese Skalen nicht optimal angepasst sind.

Aus diesem Grund entwickelten Schrepp und Thomaschweski (2019a, 2019b, 2019c) Skalen zur modularen Erweiterung des UEQs, den sogenannten *User Experience Questionnaire +* (UEQ+). In Abhängigkeit des Produkts lässt sich ein individueller Fragebogen aus verschiedenen Skalen zusammenstellen. Es wurden zu diesem Zweck bereits 20 Skalen mit je vier Items entworfen (Boos & Brau, 2017; Hinderks, 2016; Klein et al., 2020; Laugwitz et al., 2008; Otten et al., 2020; Schrepp & Thomaschewski, 2019b). Die Autoren empfehlen aus Gründen der Ökonomie eine Auswahl von fünf oder sechs Skalen. Jede Skala des UEQ+ setzt sich aus einem Einleitungssatz und vier Items zusammen. Die Items wiederum bestehen jeweils aus gegensätzlichen Begriffen als Polen. Sie bilden demnach sogenannte semantische Differentiale (Osgood et al., 1975). Die Bewertung erfolgt, wie beim UEQ (Laugwitz et al., 2008), anhand einer siebenstufigen Ratingskala. Ein weiteres Item ermittelt, wie wichtig der jeweilige Aspekt in Bezug auf das vorliegende Produkt ist, wodurch eine Gewichtung der abgefragten Produkteigenschaften möglich ist.

Der UEQ+ deckt mit den 20 bestehenden Skalen bereits einige Produktkategorien ab. Das Handbuch des UEQ+ (Schrepp & Thomaschewski, 2020) enthält Empfehlungen, welche Skalen für welche Produkte verwendet werden sollten. Für Medizinprodukte existieren allerdings noch keine Empfehlungen. Es ist außerdem anzunehmen, dass für Medizinprodukte noch weitere Faktoren einbezogen werden sollten als jene, die der UEQ+ derzeit abdeckt. Anders als bei reinen Software-Anwendungen stehen bei Medizinprodukten andere UX-Aspekte im Fokus. Die bestimmungsgemäße, sichere und effiziente

Nutzung medizinischer Geräte ist besonders wichtig, weshalb häufig primär Usability-Aspekte betrachtet werden (Bitkina et al., 2020; Mentler, 2021). Des Weiteren unterscheiden sich die Nutzergruppen von Medizinprodukten von üblichen Zielgruppen. Komplexe medizinische Geräte, wie CT- oder MRT-Geräte, werden durch medizinisches Personal mit einer entsprechenden Ausbildung bedient. Zu Expert*innen auf diesem Gebiet gehören zum Beispiel Radiolog*innen und medizinisch-technische-Radiologieassistent*innen (MTRA). Um relevante Aspekte hinsichtlich der UX bei Medizinprodukten zu ermitteln, wurden von Siemens Healthineers bereits Studien und Expert*innenbefragungen durchgeführt, die im folgenden Abschnitt erläutert werden.

1.4 User Experience bei Siemens Healthineers

Unter Medizinprodukte fallen auch Apps und Produkte, die von Patient*innen selbst angewendet werden. Der UX-Fragebogen, der im Rahmen dieser Arbeit entwickelt wird, soll jedoch insbesondere auf Produkte des Auftraggebers Siemens Healthineers abgestimmt sein. Explizit soll es um Produkte gehen, die von medizinischem Personal angewendet werden, wohingegen die Wahrnehmung von Patient*innen durch den Fragebogen vorerst nicht erfasst werden soll. Der UX-Fragebogen soll Produkte der folgenden sechs Business Lines von Siemens Healthineers abdecken, die jeweils mehrere Produktkategorien umfassen:

1. Magnetic Resonance (MR): Magnetresonanztomographie
2. Computed Tomography (CT): Computertomographie, SPECT und PETCT
3. X-ray Products (XP): Radiographie/Röntgen-Geräte, Fluoroskopie, Urologie und Mammographie
4. Digital & Automation (D&A): Modalitätenübergreifende Software für Reading & Reporting
5. Advanced Therapy (AT)/Advanced Therapy Surgery (AT-SU): Angiographie-Systeme und C-Bögen
6. (Optional) Strahlentherapie: Bestrahlungssysteme

Die Business Line Strahlentherapie ist optional, da sie im Rahmen der Fusionierung mit der Firma Varian erst 2021 Teil von Siemens Healthineers wurde. Bisher wurden bei Siemens Healthineers zur Erfassung der UX bei Medizinprodukten neben qualitativen Methoden wie Usability-Tests auch quantitative Daten mithilfe von Fragebögen erhoben, vor allem mit der System Usability Scale (SUS) (Brooke, 1996) und dem UEQ (Laugwitz et al., 2008). Wie zuvor ausgeführt, sind diese Fragebögen nicht optimal auf Medizinprodukte zugeschnitten. Es wurden deshalb bereits Schritte von Siemens Healthineers

unternommen, einen auf Medizinprodukte abgestimmten UX-Fragebogen zu entwickeln. Zunächst werden für eine Einordnung allgemeine Informationen zur Siemens Healthcare GmbH vermittelt.

1.4.1 Allgemeine Informationen zur Siemens Healthcare GmbH.

Die Siemens Healthcare GmbH (Siemens Healthineers) ist ein internationales, börsennotiertes Medizintechnikunternehmen, das 2017 aus der Siemens AG hervorging (Siemens Healthcare GmbH, 2022b). Mit 66.000 Mitarbeiter*innen in 75 Ländern gehört das Unternehmen zu einem der größten Hersteller für Medizintechnikprodukte weltweit. Siemens Healthineers bietet Lösungen für die bildgebende Diagnostik, Labordiagnostik, bildgestützte Therapie sowie Ansätze zur Krebsbehandlung. Fortschrittliche Softwarelösungen, die auf künstlicher Intelligenz basieren, ergänzen das Portfolio. Jährlich investiert das Unternehmen 1,5 Milliarden Euro in Entwicklung und Forschung (Siemens Healthcare GmbH, 2022a).

Die Siemens Healthcare GmbH hat eine eigene Abteilung für Design und User Experience, welche sich einer menschzentrierten Entwicklung der Produkte annimmt. Ein interdisziplinäres Team arbeitet daran, die Produkte an die sensiblen Situationen im klinischen Alltag anzupassen. Ein einheitliches Designkonzept soll die Einarbeitung in neue Systeme erleichtern und Vertrautheit schaffen. Das Design soll ein Markenimage vermitteln, das für hohe Qualität und Wert steht, sodass das Vertrauen in die Produkte gestärkt wird (Siemens Healthcare GmbH, 2022c). Die Abteilung für Design und User Experience setzt sich aus mehreren Teams zusammen. Das Team User Research, Testing, Prototypes & Implementation beschäftigt sich unter anderem mit der Entwicklung des UX-Fragebogens.

Der Produktentwicklungsprozess bei Siemens Healthinners entspricht den Empfehlungen der U.S. Food & Drug Administration (2016a) und soll im Folgenden knapp skizziert werden. Am Anfang des Prozesses steht eine Erkundung der Bedürfnisse und Wünsche der Zielgruppe. Wesentliche Aufgaben und Schwierigkeiten der Anwender*innen müssen identifiziert werden. Auf Basis dieser Einblicke werden Lösungsmöglichkeiten abgewogen und diskutiert. Die geeignetste Lösung wird dann in Form eines Prototyps umgesetzt, der an einer Nutzergruppe getestet wird. Bei solchen UX-Testungen sollten alle relevanten Aufgaben aus der vorangegangenen Analyse am Prototypen durchgeführt werden. Der Prototyp wird daraufhin iterativ weiterentwickelt, detaillierter ausgearbeitet und erneut getestet. Wenn ein zufriedenstellendes Ergebnis erreicht ist, das heißt, wenn keine Schwierigkeiten oder Unklarheiten bei der Anwendung mehr auftreten, folgt die

Produktion und Markteinführung. Auch in diesem letzten Stadium ist die Evaluation nicht abgeschlossen, denn es erfolgt eine weitere Überprüfung, wie das Produkt von der Zielgruppe angenommen wird. In allen Testphasen, wie auch bei der abschließenden Evaluation werden verschiedene Methoden zur Evaluation der UX angewandt und in jedem Schritt könnte ein UX-Fragebogen hilfreich sein, um den aktuellen Stand der Entwicklung zu ermitteln. Ferner ermöglicht ein Fragebogen Vergleiche zu bisherigen Produkten.

1.4.2 Projekt 2015/16 - Protokoll zur UX-Evaluation bei Medizinprodukten.

In den Jahren 2015/16 wurde bei Siemens Healthineers ein Projekt zur Entwicklung eines geeigneten Ansatzes zur Bewertung der UX bei Medizinprodukten ins Leben gerufen (Braun, 2016). Ein wesentliches Ziel des Projekts bestand darin, ein Protokoll für den Ablauf von UX-Studien zu entwerfen, welches zum Beispiel eine Videoaufzeichnung des Interaktionsablaufs vorsieht. Neben der Anwendung qualitativer Methoden ist im letzten Abschnitt des Protokolls auch die Anwendung eines UX-Fragenbogens geplant, der hierfür zusätzlich entwickelt wurde. Unter Bezugnahme auf bisherige UX-Fragebögen wurden sechs zentrale UX-Dimensionen extrahiert. Es handelte sich um die Dimensionen *Ergebnisqualität*, *Sicherheit*, *Vertrauen & Kontrolle*, *Emotionen*, *Erlernbarkeit* und *Flüssigkeit*. Die Auswahl relevanter Skalen erfolgte allerdings rein theoriegeleitet und wurde nicht empirisch überprüft. Items für die entsprechenden Dimensionen wurden aus bestehenden Fragebögen übernommen. Es wurden darüber hinaus Items ergänzt, nachdem ihre Eignung von Expert*innen bestätigt wurde. Anschließend wurde in den USA, Österreich und China eine Pilotstudie durchgeführt, um die Adäquatheit des Ansatzes zu überprüfen. Es wurden insgesamt jedoch nur 15 Teilnehmende verschiedener Berufsgruppen befragt.

Es existiert somit bereits ein Fragebogen, der allerdings keine vollständige empirische Überprüfung durchlief. Ein weiterer Kritikpunkt an diesem ersten Fragebogen ist, dass er nicht unter Einbezug aller Business Lines von Siemens Healthineers entworfen wurde. Dagegen konzentrierte er sich auf die Bereiche CT, MR und Diagnostik-Software. In Kapitel 1.4 (User Experience bei Siemens Healthineers) wurde beschrieben, welche weiteren Produkte für die Entwicklung des neuen Fragebogens einbezogen werden sollen. Dazu kommt, dass die für Medizinprodukte relevanten Skalen nicht empirisch ermittelt wurden, wie von Winter (2015; 2017) empfohlen. Ebenso konnten die bestehenden Skalen des UEQ+ (Schrepp & Thomaschewski, 2019b) nicht beachtet werden, denn dieser existierte zum Zeitpunkt des früheren Projekts 2015/16 noch nicht. Die zusätzlichen Items wurden zwar von Expert*innen akzeptiert, jedoch wurden sie nicht von Expert*innen

entwickelt, wodurch relevante Aspekte fehlen könnten. Die Ziele der Projekte überschneiden sich nichtsdestotrotz deutlich mit jenen der vorliegenden Arbeit, weshalb die Ergebnisse des Projektes für die Entwicklung des neuen Fragebogens verwertet wurden. Eine Kombination des empfohlenen Protokolls für UX-Studien und einer Erhebung von Daten anhand des neuen UX-Fragebogens könnte besonders wertvolle Einblicke liefern. Es könnten dann sowohl detaillierte qualitative Informationen als auch ein umfassendes quantitatives Bild im Hinblick auf die UX gewonnen werden.

1.4.3 Interne Pilotstudie 2020 – Überprüfung der Relevanz der Skalen des UEQ+ an Siemens Healthineers Mitarbeitenden.

Als der UEQ+ (Schrepp & Thomaschewski, 2019a) veröffentlicht wurde, wurde die Entwicklung eines einheitlichen UX-Fragebogens für Medizinprodukte wieder aufgegriffen. Zur Ermittlung der Eignung der bestehenden Skalen des UEQ+ wurde im Herbst 2020 eine Siemens Healthineers-interne Pilotstudie durch einen Praktikanten umgesetzt. Die befragten 117 Mitarbeiter*innen sollten hierzu mittels eines Online-Fragebogens die Module des UEQ+ hinsichtlich ihrer Wichtigkeit für die Erfassung der UX bei Medizinprodukten einschätzen. Sie sollten sich dabei auf jene Medizinproduktkategorie beziehen, mit der sie durch ihre Arbeit am vertrautesten waren. Produktkategorien sind zum Beispiel Fluoroskopie- oder Mammographie-Geräte, die beide unter die Business Line XP fallen. In dieser Pilotstudie wurden alle in Kapitel 1.4 (User Experience bei Siemens Healthineers) aufgelisteten Business Lines bis auf Strahlentherapie mit den jeweiligen Produktkategorien einbezogen. Aus der Produktkategorie durfte dann ein konkretes Referenzprodukt/-modell frei angegeben werden, für das die Studie ausgefüllt wurde. Es wurden sowohl Skalenbeschreibungen als auch die Items der Skalen präsentiert. Zuerst sollten die zehn wichtigsten Skalen des UEQ+ für das gewählte Medizinprodukt anhand ihrer Beschreibungen ausgewählt werden. Aus den übrigen zehn Skalen sollten die Proband*innen beliebig viele Skalen auswählen, die ihnen besonders irrelevant erschienen und ihre Auswahl jeweils in einem kurzen Satz begründen. In einem nächsten Schritt sollten die zehn wichtigsten Skalen in einer Rangfolge nach ihrer Wichtigkeit angeordnet werden. Für die fünf wichtigsten Skalen sollte auch hier eine kurze Begründung für die Einschätzung ihrer hohen Relevanz abgegeben werden. Zuletzt wurden die einzelnen Items der Skalen vorgelegt, welche ebenfalls bezüglich ihrer Wichtigkeit bewertet werden sollten. Diese Bewertung erfolgte auf siebenstufigen Skalen (1 = *völlig unwichtig* bis 7 = *sehr wichtig*). In Tabelle 2 sind in der dritten Spalte von links die sechs Skalen abgetragen, die sich im Rahmen der Pilotstudie 2020 als besonders wichtig erwiesen.

Abgebildet sind jene Skalen, die bei der Rangfolge der 10 relevantesten Skalen im Mittel als am wichtigsten eingestuft wurden.

Tabelle 2

Vergleich der wichtigsten UX-Aspekte für Medizinprodukte aus früheren Projekten

UX-Aspekte	Projekt 2015/16	Pilotstudie 2020	Interviews [c] 2021	Externe Studie 2021
Sicherheit [a]	X [b]			X
Steuerbarkeit	X [b]	X		X
Effizienz		X	X	X
Effektivität [a]	X [b]			X
Übersichtlichkeit			X	X
Vertrauen			X	X
Intuitive Bedienung		X	X	
Anpassbarkeit			X	
Durchschaubarkeit				
Vertrauenswürdigkeit		X		
Inhaltsqualität		X		
Nützlichkeit		X		
Emotionen	X			
Flüssigkeit [a]	X			
Erlernbarkeit	X			

Anmerkung. Teile der Tabelle wurden von Müßig (2022) übernommen.

[a] Diese UX-Aspekte sind keine Skalen des UEQ+ (Schrepp & Thomaschewski, 2019b), sondern wurden auf Basis von Literaturrecherche hinzugefügt.

[b] Die Skalen wurden im ursprünglichen Projekt anders benannt, überschneiden sich jedoch mit den Skalen des UEQ+ und wurden daher zur Übersichtlichkeit in die Tabelle eingeordnet.

[c] Die Auswertung basiert darauf, wie oft der UX-Aspekt bei der abschließenden Frage, welche sechs Skalen am wichtigsten sind, genannt wurde. Details werden im Text berichtet.

1.4.4 Masterarbeit 2021/22 – Überprüfung der Relevanz der Skalen des UEQ+ und von Zusatzskalen.

Aufbauend auf der beschriebenen internen Pilotstudie aus dem Jahr 2020 führte Müßig (2022) in ihrer Masterarbeit die Konzeption eines modularen UX-Fragebogens für Medizinprodukte fort. Eine umfassende Literaturrecherche im Rahmen von Müßigs Arbeit diente der Überprüfung, welche Skalen des UEQ+ eine theoretische Bedeutsamkeit haben könnten. Anhand der recherchierten Literatur wurden zudem weitere Aspekte anderer UX-Fragebögen hinzugefügt, die sich auf Medizinprodukte anwenden lassen könnten. Unternehmensleitlinien von Siemens Healthineers und Heuristiken der Abteilung für Design und User Experience wurden ebenfalls herangezogen, um weitere relevante Aspekte zu identifizieren. Es entstand so eine Sammlung von UX-Aspekten, bestehend aus den Skalen des UEQ+ und den neu recherchierten Aspekten. In einem nächsten Schritt galt es nun zu ermitteln, welche Aspekte für Medizinprodukte am wichtigsten sind.

Im Sommer 2021 wurden dazu acht Expert*innen von Siemens Healthineers aus den Bereichen CT, MR und XP online via Microsoft Teams (Microsoft, 2022) interviewt. Nach der Präsentation der Definition von UX wurden die Proband*innen gebeten, in Bezug auf Medizinprodukte relevante UX-Aspekte frei zu nennen. Ziel dieses Vorgehens war es, mögliche weitere relevante Aspekte zu sammeln. Anschließend wurden dem/der Interviewten nacheinander die Skalen des UEQ+ inklusive zugehöriger Items sowie die ermittelten Zusatzskalen vorgelegt, die hinsichtlich ihrer Bedeutsamkeit bewertet werden sollten. Am Ende der Interviews wurde ein Überblick über alle Skalen präsentiert und der/die Interviewte sollte sich für die sechs wichtigsten UX-Aspekte für Medizinprodukte entscheiden. Intuitive Bedienung wurde bei dieser abschließenden Frage fünfmal genannt, genauso häufig wie Effizienz. Übersichtlichkeit, Anpassbarkeit und Vertrauen erhielten je vier Stimmen. Inhaltsqualität und Konsistenz wurden je dreimal aufgezählt.

Im Rahmen der von Müßig (2022) zuletzt durchgeführten externen Online-Studie, die von August bis Oktober 2021 dauerte, wurden Radiolog*innen, medizinisch-technische Radiologieassistent*innen (MTRAs) und weitere medizinische Berufsgruppen aus verschiedenen Ländern weltweit befragt. Der Großteil der Teilnehmenden, 53 von insgesamt 74 Teilnehmenden kam aus Deutschland. Bis zu diesem Zeitpunkt wurden ausschließlich interne Expert*innen zur Wichtigkeit der Skalen für Medizinprodukte befragt. Die externe Studie zielte darauf ab, medizinisches Personal mit einem höheren Praxisbezug einzubeziehen. Die Aufgabe der Befragten war auch hier, die Relevanz der Skalen des UEQ+ und der Zusatzskalen für die UX von Medizinprodukten zu beurteilen. Die

Skalen wurden dazu jeweils in einem Satz beschrieben. Auch hier sollte bei der Beurteilung auf ein Referenzprodukt einer beliebigen Produktkategorie Bezug genommen werden, welches der/die Teilnehmende am häufigsten nutzte. Es wurden Produkte aus den Business Lines CT, MR, XP und AT/AT-SU angegeben. Die Beurteilung der Wichtigkeit der Skalen erfolgte auf einer siebenstufigen Skala (von -3 = *total unwichtig* bis +3 = *total wichtig*). Außerdem sollte im Anschluss an die Bewertung der Wichtigkeit der Skala, ebenfalls auf einer siebenstufigen Skala (von -3 = *sehr schlecht* bis +3 = *sehr gut*), bewertet werden, wie gut die Eigenschaft beim Referenzprodukt bereits umgesetzt war. Detaillierte Angaben, welche Skalen mit welchen Beschreibungen abgefragt wurden, folgen in Kapitel 3.2.2 (Variablen der Replikation). Drei optionale Skalen zu Sprachsteuerung wurden nur dann dargeboten, wenn Teilnehmende zuvor angegeben hatten, dass ihr Gerät über eine Sprachsteuerung verfügt. Abschließend sollten die Skalen, die zuvor mit +2 oder +3 bewertet wurden, in eine absteigende Rangfolge nach ihrer Wichtigkeit gebracht werden. Für die Auswertung des Online-Fragebogens wurde betrachtet, welche UX-Aspekte auf den siebenstufigen Skalen als durchschnittlich am wichtigsten bewertet wurden. Es zeigte sich, dass die sechs Skalen *Steuerbarkeit*, *Effizienz*, *Übersichtlichkeit*, *Vertrauen*, *Sicherheit* und *Effektivität* von allgemeiner, produktübergreifender Bedeutung waren. Bei separater Analyse ergaben sich Unterschiede zwischen den Business Lines, die sich jedoch bei der Berechnung einer MANOVA nicht als signifikant herausstellten.

In Tabelle 2 sind die Ergebnisse des Projekts 2015/16, der Pilotstudie, der Interviews und der externen Studie von Müßig (2022) im Vergleich dargestellt. Es zeigen sich Überschneidungen darin, welche UX-Aspekte von internen sowie externen Studienteilnehmenden als besonders wichtig betrachtet wurden. Effizienz und Übersichtlichkeit wurden in allen drei Analysen als sehr wichtig für die UX von Medizinprodukten eingestuft. Auch die Skala Vertrauen erwies sich sowohl in den Interviews als auch in der externen Studie als relevant. Intuitive Bedienung wurde in der Vorstudie und in den Interviews sehr hoch eingestuft, gehörte in der externen Studie dagegen nicht zu den Top 6 Skalen, sondern belegte den 15. Rangplatz. Auch im Hinblick auf andere UX-Aspekte ergaben sich Abweichungen zwischen den Analysen, welche zum Teil aber darauf zurückzuführen sind, dass in der ersten Vorstudie die aus der Literaturrecherche abgeleiteten Zusatzskalen noch nicht abgefragt wurden. Da nur die externe Studie von medizinischem Personal aus der Praxis bearbeitet wurde, welches die eigentliche Zielgruppe darstellt, wird sich an ihren Ergebnissen orientiert.

Abgeleitet aus Müßigs Masterarbeit (2022) wird bei Siemens Healthineers aktuell ein produktübergreifender UX-Fragebogen verwendet, der sich aus den wichtigsten Skalen der externen Studie zusammensetzt. Es existieren darüber hinaus Empfehlungen für optionale Zusatzskalen für die einzelnen Business Lines, die neben der produktübergreifenden Version ergänzend abgefragt werden können. Rein statistisch betrachtet zeigten sich jedoch keine signifikanten Unterschiede in den Wichtigkeitsbewertungen der Skalen zwischen den Business Lines. Optional werden jene Skalen empfohlen, die sich unter den Top 6 Skalen der einzelnen Business Lines befanden, aber von den sechs produktübergreifend wichtigsten Skalen abwichen. Die ersten vier der sechs produktübergreifend wichtigsten Skalen aus der externen Studie sind Teil des UEQ+ und können somit direkt eingesetzt werden. Die beiden Skalen Sicherheit und Effektivität, die sich gleichfalls als produktübergreifend relevant herausstellten, sind nicht Teil des UEQ+, für sie existieren dementsprechend noch keine Items. Der nächste Schritt in der Entwicklung des Ziel-Fragebogens zur Evaluation der UX bei Medizinprodukten besteht folglich darin, Items für diese Skalen zu entwickeln, die anschließend im Rahmen einer weiteren Studie überprüft werden. Im Rahmen dieser geplanten Studie sollen Medizinprodukte anhand der bestehenden Items des UEQ+ und der neu entwickelten Items bewertet werden, um den produktübergreifenden UX-Fragebogen auf seine Eignung zu überprüfen.

1.5 Hintergrund zur Fragebogenkonstruktion

Es wird davon ausgegangen, dass allgemeines Wissen zur psychologischen Testkonstruktion gegeben ist. Hintergrundinformationen können zum Beispiel bei Bühner (2011) nachgelesen werden.

1.5.1 Anwendbarkeit der psychologischen Testtheorie.

Die psychologische Testtheorie mit ihren Konstruktionsprinzipien und Gütekriterien für Messinstrumente bildet derzeit die Basis für die Konstruktion von UX-Fragebögen. Es ist allerdings fragwürdig, ob sich das Vorgehen ohne Weiteres übertragen lässt, da durchaus Unterschiede zwischen UX- und Persönlichkeitsfragebögen oder anderen psychologischen Instrumenten bestehen. Schrepp und Rummel (2018) führen an, dass eine wesentliche Abweichung darin besteht, dass psychologische Fragebögen in der Regel Rückschlüsse auf Eigenschaften einer Person zulassen sollen. Bei UX-Fragebögen sind einzelne Personen dagegen gar nicht von Interesse, das Ziel besteht vielmehr darin zu erfassen, welche Meinung „eine repräsentative Gruppe von Nutzern von einem Produkt hat“ (Rummel & Schrepp, 2018, S. 726). Die möglichst exakte Bestimmung eines

Personenwerts ist demnach nicht die interessierende Variable, sondern die mittlere Bewertung von Eigenschaften eines Produkts (Skalen) über mehrere Personen hinweg.

Dass die Beantwortung eines UX-Fragebogens stets nur unter Bezugnahme auf ein bestimmtes Produkt erfolgen kann, wirkt sich laut Schrepp und Rummel (2018) auf das Reliabilitätskonzept aus. Zur Beurteilung der Reliabilität von Skalen im Sinne der internen Konsistenz wird häufig Cronbachs α eingesetzt (Cronbach, 1951), das die Interkorrelation der Items abbildet. Die Autoren argumentieren, dass je nach Produkt die Items entsprechend anders interpretiert und beantwortet werden, weshalb das Cronbachs α der Skalen bei verschiedenen Produkten variieren kann. „Damit ist Reliabilität keine Eigenschaft des Fragebogens, sondern immer abhängig vom gemessenen Produkt“ (Rummel & Schrepp, 2018, S. 727). Dagegen lässt sich jedoch einwenden, dass ein Produkt, das beispielsweise als gut steuerbar eingeschätzt wird, zu einer hohen Zustimmung zu allen Items dieser Skala führen sollte. Speziell bei den Skalen des UEQ+ (Schrepp & Thomaschewski, 2019a), bei denen jeweils ein Einleitungssatz vorangestellt ist, sind die Items inhaltlich ähnlich. Wenn man zum Beispiel die Skala Steuerbarkeit heranzieht (s. Abbildung 3, Kapitel 3.2.3 – Items zur Fragebogenentwicklung), ist es unwahrscheinlich, dass die Reaktionen des Produkts auf Eingaben als „vorhersagbar“ (Item 1) und gleichzeitig als „nicht erwartungskonform“ (Item 4) beurteilt werden. Die Interkorrelation der Items sollte sich bei verschiedenen Produkten demnach zumindest annähern, denn sie zielen alle auf *eine* Eigenschaft des Produkts ab, die für sich genommen besser oder schlechter erfüllt sein sollte. Tatsächlich könnte es in manchen Fällen aber vorkommen, dass Items je nach Produkt anders interpretiert werden. Schrepp und Rummel (2018) führen hier als extremes Beispiel das Item „Spart mir Zeit“ beziehungsweise „Kostet mich Zeit“ (S. 729) an. Das Item könnte im Kontext einer betriebswirtschaftlichen Software als Arbeitszeitersparnis, im Kontext eines sozialen Netzwerks als „Ich verbringe zu viel Zeit in diesem Netzwerk“ (S. 729) interpretiert werden. In solchen Fällen könnte sich die Interkorrelation der Items einer Skala ändern. Da der Fokus des zu entwickelnden UX-Fragebogens dieser Arbeit ausschließlich auf Medizinprodukten liegt, sollte eine einheitliche Interpretation der Items in diesem Fall gleichwohl gegeben sein. Die Berechnung von Cronbachs α sollte deshalb theoretisch betrachtet möglich sein. Es stellt sich indes die Frage, ob Reliabilität im Sinne der internen Konsistenz der Skalen überhaupt ein praktisch relevanter Kennwert ist. Schrepp und Rummel (2018) konnten anhand einer Simulation zeigen, dass die Genauigkeit der Schätzung der Skalenmittelwerte kaum mit Cronbachs α zusammenhängt und daher auch bei geringem Cronbachs α Schlussfolgerungen bezüglich der

UX möglich sein sollten. Solange der Skalenmittelwert über eine ausreichend große, repräsentative Stichprobe betrachtet wird, dürfte es vertretbar sein, Abstriche bezüglich Cronbachs α hinzunehmen. Die Berechnung der internen Konsistenz der Skalen ist aus diesem Grund als hinreichendes, nicht jedoch als notwendiges Kriterium für die Eignung des Fragebogens anzusehen.

Die Interkorrelation der Items ist abgesehen von der Reliabilität für eine faktorenanalytische Konstruktion von Bedeutung. Basierend auf der Argumentation von eben sollten aber auch hier über verschiedene Produkte hinweg manifeste Variablen (Items) immer auf die gleichen Faktoren (Skalen) laden. Die Items, die zu einer Skala gehören, sollten in jedem Fall korrelieren. Was sich dennoch ändern könnte, ist die Korrelation zwischen den Skalen. Zum Beispiel könnten die Skalen Vertrauen und Effizienz je nach Produkt mehr oder weniger korrelieren. Es ist vorstellbar, dass ein Produkt aufgrund vieler Berechtigungsprüfungen wenig effizient ist, was aber das Vertrauen erhöhen könnte, dass die Daten nicht missbraucht werden. Bei anderen Produkten wiederum könnten beide Eigenschaften erfüllt sein. Dass Skalen je nach Produkt unterschiedlich hoch korreliert sein können, sollte der Aussagekraft des Fragebogens beziehungsweise der Skalenmittelwerte jedoch keinen Abbruch tun.

Diese Annahme lässt sich damit erklären, dass die UX als Gesamtkonstrukt als eine formative latente Variable aufgefasst werden kann (Edwards & Bagozzi, 2000). Formative Variablen umfassen mehrere Indikatoren. In der vorliegenden Arbeit sind die Indikatoren die einzelnen UX-Skalen für Medizinprodukte, die sich wiederum aus Items zusammensetzen. Indikatoren einer formativen latenten Variablen müssen nicht, können aber miteinander zusammenhängen (Bühner, 2011). Es ist zum Beispiel möglich, dass die Übersichtlichkeit sich durch Weiterentwicklung eines Produktes verbessert, das Vertrauen in Bezug auf die Datensicherheit jedoch gleichbleibt. „Jede Veränderung in der Ausprägung eines einzelnen Indikators geht mit einer Veränderung der Ausprägung in der latenten Variablen einher“ (Bühner, 2011, S. 18). Die Ausprägung der formativen latenten Variablen kann sich dementgegen verändern, in diesem Beispiel durch die Zunahme der Übersichtlichkeit gesteigert werden, ohne dass sich die Ausprägung aller Indikatoren ändern muss. Es ist daher für die Interpretation der Ergebnisse des Fragebogens nicht problematisch, wenn die Korrelation zwischen den Skalen je nach Produkt variiert.

Anders als mit formativen latenten Variablen verhält es sich mit reflektiven latenten Variablen (Edwards & Bagozzi, 2000). Das Modell reflektiver latenter Variablen basiert auf der Annahme, dass die reflektive latente Variable die manifesten Variablen

kausal beeinflusst. „Die Items spiegeln die latente Variable wider“ (Bühner, 2011, S. 16). Ändert sich die Ausprägung einer reflektiven latenten Variablen, so ändern sich alle Itemantworten, da sie als zusammenhängend angenommen werden. Skalen, die nach psychologischer Testtheorie beziehungsweise faktorenanalytisch konstruiert werden, basieren auf reflektiven latenten Variablen (Brandt, 2020; Bühner, 2011; Gäde et al., 2020). Es wird angenommen, dass den einzelnen UX-Skalen, die jeweils vier Items umfassen, ebenfalls reflektive latente Variablen zugrunde liegen.

Ein erheblicher Unterschied zu psychologischen Fragebögen besteht in der Interpretation der latenten Variablen (Schrepp & Rummel, 2018). Bei psychologischen Fragebögen wird davon ausgegangen, dass eine zugrundeliegende, nicht direkt beobachtbare Eigenschaft oder Fähigkeit der untersuchten Person die Itemantworten steuert (Bühner, 2011). Wie bereits beschrieben, sind Personeneigenschaften aber nicht die interessierende Variable im Fall von UX-Fragebögen. Die Erhebung eines UX-Fragebogens ergibt nur dann Sinn, wenn man einen Ansatzpunkt zur Verbesserung der UX ableiten kann. Die einzige Variable, die sich im Designprozess beeinflussen lässt, ist die Produktbeschaffenheit. Es wäre demnach hilfreich, wenn Itemantworten von UX-Fragebögen Rückschlüsse auf die Eigenschaften des Produkts zulassen würden (Schrepp & Rummel, 2018). Die UX ist jedoch ein Konglomerat aus Eigenschaften des Anwenders oder der Anwenderin, des Produkts und Kontextes, wie zuvor verdeutlicht wurde (Ariza & Maya, 2014; Berni & Borgianni, 2021; Hassenzahl & Tractinsky, 2006; Roto et al., 2011; Thüring & Mahlke, 2007). Persönlichkeitsmerkmale, Vorwissen, Einstellungen der Nutzer*innen und viele weitere Variablen beeinflussen die Beantwortung der Items (Hassenzahl et al., 2009; Hassenzahl, 2011; Hassenzahl & Tractinsky, 2006; Rummel & Schrepp, 2018). Dass die Itemantworten ausschließlich auf die Eigenschaften des Produkts zurückzuführen sind, so wie beispielsweise in Intelligenztests die Antworten im Idealfall ausschließlich auf die zugrundeliegenden kognitiven Fähigkeiten der untersuchten Person zurückzuführen sind, ist bei UX-Fragebögen nicht unbedingt der Fall. Einzelne Skalenwerte (Mittelwerte der Items einer Skala) einer Person liefern deshalb keine eindeutigen Aussagen über Produkteigenschaften. Eine aussagekräftige Interpretation von UX-Fragebögen ermöglichen nur die Skalenmittelwerte über eine repräsentative Gruppe von Nutzer*innen hinweg. Es ist somit besonders wichtig festzulegen, für welche Nutzergruppe und welchen Anwendungsbereich ein UX-Fragebogen entwickelt wurde und entsprechend einsetzbar ist. Wurde der Fragebogen nur an Student*innen und einem bestimmten Produkt getestet, sind anhand des Fragebogens nur zu diesem Anwendungsbereich valide Aussagen

möglich (Thielsch & Hirschfeld, 2018). Wenn der Fragebogen für verschiedene Produkte eingesetzt werden soll, sollte eine entsprechend breite Validierung erfolgen. Im Vergleich zu psychologischen Fragebögen ist die Validierung eines UX-Fragebogens so betrachtet aufwändiger, wenn ein Einsatz für viele Nutzer- und Produktgruppen erwünscht ist (Thielsch & Hirschfeld, 2018).

Schrepp und Rummel (2018) legen außerdem nahe, vorab zu überlegen, welche Skalen aus Design- beziehungsweise Anwendersicht wichtig sind und darauf aufbauend passende Items zu generieren, anstatt empirisch geleitet vorzugehen und Skalen faktorenanalytisch zu extrahieren. Entsprechend dieser Empfehlung wurde bei der Entwicklung des UX-Fragebogens für Medizinprodukte bislang vorgegangen. Psychologische Fragebögen, zum Beispiel das NEO-PI-R (Costa & McCrae, 2008; Ostendorf & Angleitner, 2004) wurde induktiv, datengeleitet konstruiert, um Menschen in ihren Eigenschaften möglichst umfassend zu beschreiben (Andresen & Beauducel, 2008). Aus UX-Fragebögen sollten sich dagegen vor allem praxisrelevante und im Design umsetzbare Informationen ableiten lassen, eine umfassende Beschreibung der UX ist dafür nicht nötig.

Der psychologischen Testtheorie liegen zusammengenommen zwar teilweise Annahmen zugrunde, die sich nicht optimal auf die Anforderungen von UX-Fragebögen übertragen lassen, dennoch gibt es auch Überschneidungen. Die Gütekriterien der psychologischen Testtheorie sind weit verbreitet (Schrepp & Rummel, 2018) und ein Anhaltspunkt für die Qualität von Fragebögen, gerade im Hinblick auf die Validierung. Sie werden deshalb auch für die Evaluation des zu entwickelnden UX-Fragebogens für Medizinprodukte eingesetzt.

1.5.2 Gütekriterien.

Auf das psychologische Gütekriterium der Reliabilität wurde im vorangegangenen Abschnitt bereits eingegangen. Im folgenden Kapitel wird noch einmal auf verschiedene Reliabilitätsmaße und die Umsetzung der Reliabilitätsmessung in der vorliegenden Arbeit eingegangen. Im Anschluss wird die Validierung und Objektivität des UX-Fragebogens beleuchtet.

1.5.2.1 Reliabilität. In bisherigen Studien zur Entwicklung von UX-Fragebögen wurde überwiegend Cronbachs α (Cronbach, 1951) berichtet (Schrepp & Rummel, 2018; Thielsch & Hirschfeld, 2018). Abgesehen davon, dass ohnehin fraglich ist, wie aussagekräftig dieses Maß bei der Anwendung auf UX-Fragebögen ist, wird Cronbachs α für die Bestimmung der Reliabilität zunehmend kritisch betrachtet (Cho & Kim, 2015; McNeish, 2018; Moosbrugger & Kevala, 2020). Voraussetzung für die Verwendung von Cronbachs

α ist streng genommen das Vorliegen von essentieller τ-Äquivalenz, die beinhaltet, dass alle Items gleich hoch auf einen Faktor laden (Berger, 2019; Eid & Schmidt, 2014). Realistischerweise ist es allerdings häufig so, dass Items unterschiedlich hohe Ladungen aufweisen. Einige Autor*innen empfehlen deshalb die Verwendung von McDonalds Omega (ω) (McDonald, 1970, 1999), welches nur τ-Kongenerität und somit keine homogenen Ladungen voraussetzt (Hayes & Coutts, 2020; Moosbrugger & Kevala, 2020). Cronbachs α stellt somit einen Spezialfall von McDonalds ω dar. Liegt essentielle τ-Äquivalenz vor, so kommen beide Schätzer zum gleichen Ergebnis (McNeish, 2018). Die Verwendung von McDonalds ω liegt deshalb für die Reliabilitätsbestimmung des UX-Fragebogens näher.

Ein weiteres Maß, das sich zur Evaluation des UX-Fragebogens als hilfreich erweisen könnte, ist die Interrater-Reliabilität (Gisev et al., 2013). Bewerten Personen das gleiche Produkt, sollte sich ein Zusammenhang der Ratings zeigen, wenn Produkteigenschaften Einfluss auf die Beantwortung des UX-Fragebogens haben. Zumindest sollten Bewertungen von Personen, die das gleiche Referenzprodukt beurteilen, ähnlicher sein als Bewertungen von Befragten, die verschiedene Produkte einschätzen. Allerdings muss auch dieses Kriterium für die Eignung des Ziel-Fragebogens nicht zwingend erfüllt sein, da individuelle, personenspezifische Einflüsse auf die Beantwortung überwiegen könnten. Sollte sich aber eine hohe Interrater-Reliabilität der Bewertungen zeigen, spricht das für die Eignung des UX-Fragebogens. Ob dieses Maß überhaupt berechnet werden kann, hängt davon ab, ob das gleiche Referenzprodukt von mehreren Personen bewertet wird. Die Ermittlung weiterer Reliabilitätsmaße, wie zum Beispiel die Erfassung von Retest-Reliabilität in Bezug auf Produktbewertungen ist darüber hinaus empfehlenswert (Thielsch & Hirschfeld, 2018), lässt sich jedoch im Rahmen dieser Masterarbeit nicht umsetzen.

1.5.2.2 Validität. Die inhaltliche Validität des UX-Fragebogens sollte dadurch gegeben sein, dass die Items von Expert*innen generiert und konsolidiert werden. Außerdem werden die Skalen nicht faktorenanalytisch deduktiv gebildet, was zu Interpretationsschwierigkeiten der Faktoren führen könnte, sondern sie basieren auf dem UEQ+ (Schrepp & Thomaschewski, 2019a) und Literaturrecherche (Müßig, 2022). Eine Stärke des UX-Fragebogens besteht darin, dass die Skalen im Rahmen einer separaten Vorstudie zunächst auf ihre empirische Wichtigkeit hin überprüft wurden. Da die Vorstudie sowie auch die in dieser Arbeit beschriebene Studie an klinischem Personal durchgeführt wurden, lässt sich von einer hohen externen Validität des Fragebogens ausgehen (Döring &

Bortz, 2016). Die Konstruktvalidität des Ziel-Fragebogens soll über zwei Wege geprüft werden. Zum einen wird die Struktur mithilfe einer Faktorenanalyse überprüft, zum anderen wird die Korrelation mit verwandten Konstrukten berechnet (Rammstedt, 2010). Zur Berechnung letzterer, der konvergenten Validität, werden zwei Zusatzitems erhoben. Angelehnt an die Validierung des UEQ+ (Schrepp & Thomaschewski, 2019b) wird mit einem Item die Gesamtzufriedenheit mit der Nutzerfreundlichkeit des Produkts abgefragt. Als weiteres Item wird in der vorliegenden Studie, angelehnt an den *Net Promoter Score* (Reichheld, 2003) abgefragt, wie wahrscheinlich es ist, dass der/die Nutzer*in das Produkt einem/einer Kolleg*in weiterempfiehlt. Die Validierungsitems sollen, wie die anderen Items auch, auf einer siebenstufigen Ratingskala bewertet werden. Falls die beiden Items hoch korrelieren ($\geq .70$), werden sie zusammengefasst. Es wird dann geprüft, ob der Mittelwert der beiden Items mit den Skalen zur Produktbewertung korreliert (s. Kapitel 3.2.3 – Items zur Fragebogenentwicklung).

Ein valider UX-Fragebogen müsste des Weiteren im Sinne der diskriminanten Validität zwischen Produkten, zum Beispiel zwischen verschiedenen MRT-Modellen, differenzieren (Thielsch & Hirschfeld, 2018). Es ist jedoch unklar, ob Proband*innen das konkrete Referenzprodukt, für das sie den Studienfragebogen ausfüllen, erstens benennen können und ob zweitens die gleichen Modelle häufig genug bewertet werden, damit eine sinnvolle statistische Auswertung erfolgen kann. Als Alternative wird daher untersucht, ob sich Produkte verschiedener Hersteller in ihrer Bewertung auf den Items des UEQ+ und den neu entwickelten Items unterscheiden. Die Produktkategorien (CT-Geräte, Angiographie-Geräte, etc.) werden hierfür für jeden Hersteller zusammengefasst, um eine ausreichend große Anzahl an bewerteten Produkten pro Hersteller zu erhalten. So kommt es allerdings zu einer Vermischung verschiedener Produktkategorien und Referenzprodukte, die sich in ihrer UX deutlich unterscheiden können, aber gemeinsam betrachtet werden. Wenn der UX-Fragebogen also nicht zwischen Produkten verschiedener Hersteller differenziert, wäre das kein Beleg für mangelnde Güte, weil sich ähnliche mittlere Produktbewertungen auch auf die Zusammenfassung verschiedener Produkte zurückführen lassen könnten. Dagegen würde eine Differenzierung zwischen verschiedenen Herstellern einen Hinweis auf die Eignung des Fragebogens geben. Es handelt sich deshalb ebenfalls um ein hinreichendes, aber nicht notwendiges Kriterium.

1.5.2.3 Objektivität. Die Erfüllung von Objektivität ist Voraussetzung für Reliabilität und Validität (Thielsch & Hirschfeld, 2018). Im Rahmen der Entwicklung des UX-Fragebogens nimmt die Objektivität noch keine große Rolle ein, vor allem bei der

praktischen Anwendung des UX-Fragebogens wird sie zum Tragen kommen. Die Durchführungsobjektivität des zu entwickelnden UX-Fragebogens sollte gegeben sein, da jede Skala zu Beginn einen Einleitungssatz vorgibt, der Fehlinterpretationen unwahrscheinlich macht. Hilfestellungen oder Instruktionen von Seiten der Testleiter*in sollten deshalb nicht nötig sein, wenn der UX-Fragebogen später eingesetzt wird. Durch das geschlossene Antwortformat der Items ist Auswertungsobjektivität sichergestellt (Rammstedt, 2010). Diese wird zusätzlich unterstützt durch ein Excel-Auswertungstool, das für die bestehenden Skalen des UEQ+ bereits existiert (Schrepp & Thomaschewski, 2019d) und auch für die Fragebogenversion für Medizinprodukte verfügbar sein soll. Mit dem Ziel Interpretationsobjektivität zu gewährleisten, sollte ein Vergleichsmaßstab für erhobene Werte vorliegen. Die internationale Erhebung einer Benchmark an einer großen Stichprobe kann erfolgen, sobald der UX-Fragebogen entwickelt, validiert und in andere Sprachen übersetzt wurde. Ergebnisse des UX-Fragebogens im Rahmen von UX-Testungen an neuen Prototypen oder Produkten lassen sich dann mit bisherigen Produkten der entsprechenden Produktkategorie vergleichen und einordnen (s. Kapitel 4.6.4 – Weiterführende Schritte bei Siemens Healthineers).

1.6 Zielsetzung und Hypothesen der vorliegenden Arbeit

Zusammenfassend möchte Siemens Healthineers sowohl entwicklungsbegleitend als auch entwicklungsabschließend standardisiert und valide erfassen können, wie gut die User Experience der entwickelten Medizinprodukte ist. Es werden dabei ausschließlich Produkte einbezogen, die von medizinischem Personal bedient werden. Zu diesem Zweck soll im Rahmen der vorliegenden Arbeit ein auf den Medizinkontext abgestimmter UX-Fragebogen entwickelt werden, welcher auf dem UEQ+ von Schrepp und Thomaschewski (2019a) basiert.

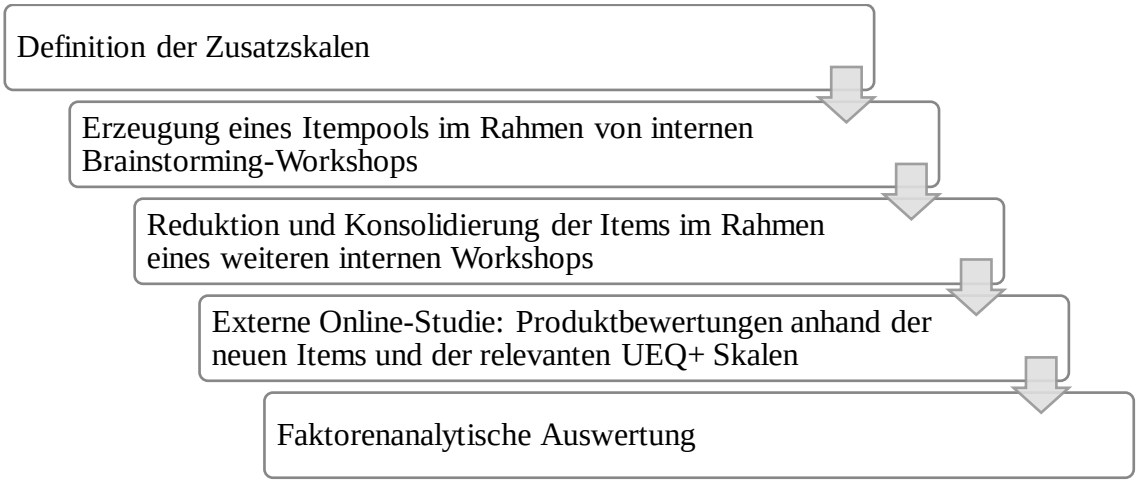
Ein erster Bestandteil der vorliegenden Arbeit ist eine Teilreplikation der vorangegangenen externen Studie von Tanja Müßig (2022). Für die Studie konnten zwar 161 Teilnehmende rekrutiert werden, jedoch musste ein Teil ausgeschlossen werden, sodass am Ende nur 74 Personen in die Analysen eingingen. Ein Großteil der Proband*innen brach den Studienfragebogen bereits vor der eigentlich interessanten Einschätzung der Wichtigkeit der Skalen ab. Weitere 25 Teilnehmende mussten ausgeschlossen werden, weil sie bei Siemens Healthineers arbeiteten, ein auffälliges Antwortverhalten vorlag oder eine Berufsgruppe angegeben wurde, die nicht Teil der Zielgruppe war. Es wird deshalb zunächst überprüft, ob sich in einer weiteren Studie dieselben sechs Skalen als produktübergreifend besonders relevant für die Evaluation der UX bei Medizinprodukten

herausstellen. Für die Business Line Strahlentherapie nahm in Müßigs Studie keine Person teil, für die Business Line D&A nur eine Person. Auch für die Business Lines XP und AT/AT-SU ließen sich nur sieben beziehungsweise acht Personen rekrutieren. Aufgrund der geringen Größen der Substichproben lässt sich nicht ausschließen, dass es Abweichungen in den wichtigsten Skalen zwischen den Business Lines geben könnte, selbst wenn sich in der Vorstudie keine signifikanten Unterschiede zeigten. Falls signifikante Unterschiede vorliegen, sollte es Business Line-spezifische Versionen des UX-Fragebogens geben. Es handelt sich um eine Teilreplikation der Studie, da Teile des ursprünglichen Studienfragebogens gekürzt wurden. Details zu den Abweichungen von der Vorstudie werden in Kapitel 3.2.2 (Variablen der Replikation) erläutert. Sowohl die Entwicklung der Items sowie deren Überprüfung im Rahmen der externen Studie werden auf deutscher Sprache erfolgen, da die exakte Formulierung und Wortwahl der Items einen wesentlichen Einfluss auf ihre Beantwortung haben könnte. Um sicherzustellen, dass die Items auf anderen Sprachen gleich interpretiert werden, genügt eine einfache Übersetzung nicht. Es wird deshalb zuerst eine deutsche Version des UX-Fragebogens entworfen. Die Items werden anschließend durch Übersetzung und Rückübersetzung in andere Sprachen überführt, sodass der UX-Fragebogen auch in anderen Ländern zum Einsatz kommen kann (Brislin, 1970).

Das Hauptziel der Arbeit ist die Entwicklung von Items für die Zusatzskalen Sicherheit und Effektivität, welche sich in der vorangegangenen Studie als wichtig herausgestellt haben. Die Items sollen im oben beschriebenen Format des UEQ+ entwickelt werden. Das Vorgehen ist der Konstruktion des UEQ und des UEQ+ nachempfunden (Boos & Brau, 2017; Laugwitz et al., 2008; Schrepp & Thomaschewski, 2019b). Die geplanten Schritte in Bezug auf die Itementwicklung sind in Abbildung 1 dargestellt.

Abbildung 1

Geplantes Vorgehen zur Skalenkonstruktion basierend auf der Entwicklung des UEQ (Laugwitz et al., 2008)



Aus dem geplanten Vorgehen werden folgende Hypothesen abgeleitet:

Hypothesen zur Replikation:

1. Es zeigen sich die gleichen sechs Skalen wie in der Vorstudie als produktübergreifend wichtig.
2. Die Business Lines (MR, CT, XP, AT/AT-SU, D&A und Strahlentherapie) unterscheiden sich nicht signifikant darin, welche UX-Skalen am wichtigsten sind. Es gibt demnach eine produktübergreifende Empfehlung für die Skalen, die verwendet werden sollten.

Hypothesen zur Fragebogenentwicklung:

3. Skalenkonstruktion

   3.1 Skala Sicherheit

   3.1.1 Es lässt sich mittels einer Parallelanalyse Eindimensionalität der Skala Sicherheit nachweisen. Als Vergleichsmaßstab werden unkorrelierte, multivariat normalverteilte Daten simuliert. Faktoren mit Eigenwerten, die über den zufälligen liegen, werden extrahiert.

   3.1.2 Es lassen sich vier geeignete Items für die Skala Sicherheit finden. Die Itemladungen auf den Faktor sind größer als .30 (Brown, 2015) und signifikant von null verschieden.

   3.1.3 Detaillierte Analysen der durch die Faktorenanalyse vorausgewählten Items ergeben eine angemessene Schwierigkeit. Die Itemschwierigkeit

liegt zwischen .20 und .80, wobei die Items eine möglichst breite Streuung der Schwierigkeit aufweisen sollten (Döring & Bortz, 2016). Die Trennschärfe hängt eng zusammen mit der Ladung auf den Faktor (Eid et al., 2013). Da für die Ladung bereits ein Grenzwert festgelegt wurde, wird dieser auch zur Gewährleistung der Trennschärfe herangezogen. Bei faktorenanalytischer Konstruktion resultieren zwangsläufig homogene Skalen, weshalb hier ebenfalls auf eine separate Prüfung verzichtet wird.

3.2 Skala Effektivität

Es gelten die gleichen Kriterien wie für die Skala Sicherheit.

3.2.1 Es lässt sich Eindimensionalität der Skala Effektivität nachweisen.

3.2.2 Es lassen sich vier geeignete Items für die Skala Effektivität finden.

3.2.3 Detaillierte Analysen der durch die Faktorenanalyse vorausgewählten Items ergeben eine angemessene Schwierigkeit.

4. Die angenommene Struktur des Fragebogens (s. Abbildung 2) lässt sich nachweisen.

4.1 Globale Modellgüte

4.1.1 Der $\chi 2$-Test zur Beurteilung der Abweichung zwischen der beobachteten und der modellimplizierten Kovarianz-Varianz-Matrix ist nicht signifikant. Wegen der Stichprobenabhängigkeit ist dieses Kriterium hinreichend, jedoch nicht notwendig.

4.1.2 Fit Indizes deuten auf eine hohe Modellpassung hin. Folgende Kriterien werden herangezogen: Root-Mean-Square-Error of Approximation (RMSEA) ≤ 0.08; Tucker-Lewis-Index (TLI) und Comparative-Fit-Index (CFI) > 0.95; Standardized Root Mean Square Residual (SRMR) ≤ 0.09 (Bühner, 2011)

4.2 Lokale Modellgüte

4.2.1 Die Steigungsparameter (Ladungen) sind signifikant.

4.2.2 Es liegen laut Modifikationsindizes keine Fehlspezifikationen vor. Zur Identifikation von Fehlspezifikationen wird das Vorgehen von Saris et al. (2009) herangezogen (s. Kapitel 3.5.4 – Ergebnisse der Fragebogenentwicklung). Auch dieses Kriterium ist hinreichend, jedoch nicht notwendig, da laut Aichholzer (2017) bei Modellen mit vielen Parametern einzelne Abweichungen zu erwarten sind.

5. Psychometrische Gütekriterien

5.1 Der Fragebogen ist reliabel.

McDonalds Omega (ω) (McDonald, 1970, 1999) der Skalen ist größer oder gleich .80 (Lance et al., 2006). Auch hier widerlegt eine geringere Reliabilität wie zuvor beschrieben nicht die Eignung des Fragebogens. Falls mehrere Personen das gleiche Produkt bewerten, wird außerdem die Interrater-Reliabilität berechnet.

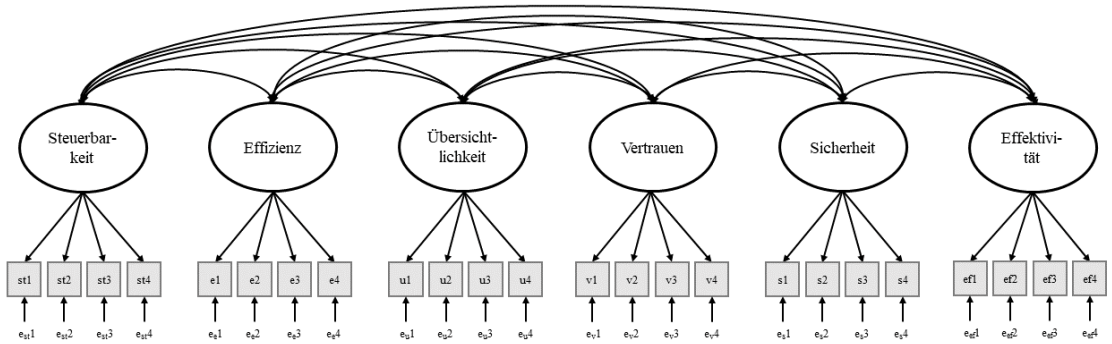
5.2 Der Fragebogen ist valide.

5.2.1 Es zeigt sich eine signifikante Korrelation zwischen den Skalen des produktübergreifenden UX-Fragebogens und der Gesamtzufriedenheit mit dem Produkt, die durch zwei zusätzlich abgefragte Items erfasst werden soll. Die Korrelation sollte mindestens .30 betragen, was nach Cohen (1992) einem mittleren Effekt entspricht.

5.2.2 Produkte verschiedener Hersteller unterscheiden sich signifikant in ihrer durchschnittlichen Bewertung. Dieses Kriterium ist ebenso hinreichend, jedoch nicht notwendig.

Abbildung 2

Angenommene Struktur des zu entwickelnden UX-Fragebogens für Medizinprodukte



2 Methode Schritt 1 – Itemgenerierung und -konsolidierung

In den vorangegangenen Kapiteln wurden der theoretische Hintergrund und zuvor durchgeführte Studien zur Konzeption eines UX-Fragebogens für Medizinprodukte bei Siemens Healthineers erläutert. Auf dieser Grundlage wird nun das methodische Vorgehen für die Generierung von Items für die Zusatzskalen Sicherheit und Effektivität berichtet.

2.1 Workshops zur Itemgenerierung

Zunächst wurde recherchiert, was Sicherheit und Effektivität im Kontext von Medizinprodukten ausmacht. Angelehnt an die Entwicklung des UEQ und des UEQ+ wurden daraufhin Expert*innenworkshops zur Itemgenerierung durchgeführt (Boos & Brau, 2017; Laugwitz et al., 2008; Schrepp & Thomaschewski, 2019b). Die Ergebnisse der Literaturanalyse wurden in die Workshops und deren Auswertung einbezogen und flossen so in die Itemgenerierung ein. Mehr Informationen zur Literaturanalyse befinden sich in Anhang A. Der Ablauf und die Auswertung der Workshops werden im Folgenden beschrieben.

2.1.1 Ablauf der Workshops zur Itemgenerierung.

Für eine erste Sammlung von Ideen, wurden Mitarbeiter*innen von Siemens Healthineers per E-Mail zu Workshops eingeladen. Bei den eingeladenen Teilnehmer*innen handelte es sich um Expert*innen der verschiedenen Business Lines. Ihre Arbeitsfelder lagen in den Bereichen Innovationsmanagement, Customer Service/Support Center, Forschung und Entwicklung, Vertrieb und Trainingscenter und hatten somit einen hohen Praxisbezug. Es konnte kein*e Vertreter*in der Business Line AT/AT-SU teilnehmen, zwei Mitarbeiter*innen aus diesem Bereich übermittelten aber Informationen per Nachricht beziehungsweise in einem Einzelgespräch. Vier der Teilnehmer*innen waren ehemalige medizinisch-technische Angestellte (MTA). Somit konnte sichergestellt werden, dass die Teilnehmer*innen mit den Geräten vertraut waren und Sicherheits- und Effektivitätsaspekte auch aus Praxisperspektive bewerten konnten. Zusätzlich wurden Mitarbeitende der User Experience-Abteilung eingeladen, um auch ihre Erfahrungen und Wissen über Anwenderfreundlichkeit einzubeziehen.

Es fanden zwei Workshops mit je sechs Teilnehmer*innen statt. Der erste Workshop bezog sich auf die Skala Sicherheit, der zweite auf die Skala Effektivität. Die Workshops wurden online über Microsoft Teams (Microsoft, 2022) durchgeführt und dauerten jeweils 90 Minuten. Nach einer kurzen Vorstellungsrunde aller Anwesenden wurden Hintergrundinformationen zu User Experience, dem UEQ+ und zum Aufbau von Fragebögen

präsentiert. Einige PowerPoint-Folien befinden sich zur Veranschaulichung in Anhang A (Abbildung A1). Zunächst wurde die Definition der jeweiligen Zusatzskala erörtert, um zu prüfen, ob ein ähnliches Verständnis des Konstrukts gegeben war. Die Teilnehmenden wurden gebeten, sich zuerst ihre eigene Definition von Sicherheit beziehungsweise Effektivität bei Medizinprodukten zu überlegen und diese auf einem vorbereiteten Conceptboard (Conceptboard, 2022) einzutragen. Conceptboard ist eine Web-Applikation, die wie ein unbegrenzt großes Whiteboard aufgebaut ist und es mehreren Personen gleichzeitig ermöglicht Einträge vorzunehmen. Ausschnitte aus den ausgefüllten Conceptboards zu den beiden Workshops sind ebenfalls in Anhang A (Abbildung A2) ersichtlich.

Erst nach der Sammlung von Definitionen und Schlagworten wurde die jeweilige Definition von Sicherheit beziehungsweise Effektivität nach der ISO vorgestellt und besprochen (s. folgende Abschnitte). Für die eigentliche Itemgenerierung wurden die Teilnehmenden in Gruppen von zwei Personen aufgeteilt. Dazu wurde die Funktion für „breakout rooms“ in Microsoft Teams genutzt. In den Gruppen sollten sich die Teilnehmer*innen überlegen, was Medizinprodukte unsicher/sicher oder uneffektiv/effektiv macht und woran Anwender*innen das festmachen würden. Die Ergebnisse wurden auch hier auf dem Conceptboard in Form von Stichpunkten festgehalten, das Format der UEQ+ Items wurde noch nicht berücksichtigt. Anschließend, nach etwa 15 Minuten, wurden die Teilnehmer*innen zurück in den Hauptraum geholt und es wurde über die gesammelten Ideen diskutiert. Zusätzlich wurden aus der Literatur erschlossene Sicherheits-/Effektivitätsaspekte präsentiert, um die Diskussion anzuregen (s. Anhang A, Abbildungen A3 und A4). Zum Abschluss wurde die weitere Verwendung der gesammelten Ideen erläutert.

2.1.2 Auswertung und Ergebnisse der Workshops zur Itemgenerierung.

Die von der Moderatorin handschriftlich festgehaltenen Diskussionspunkte wurden gleich nach den Workshops digitalisiert. Die im Workshop genannten Definitionen und Schlagworte zum Thema Sicherheit und Effektivität wurden auf Gemeinsamkeiten untereinander sowie auf Übereinstimmung mit der Definition der ISO überprüft. Eine einheitliche, konsensfähige Definition war erforderlich, damit später ein treffender Einleitungssatz für die Items im Format des UEQ+ erarbeitet werden konnte. Aus den Notizen zur Diskussion und den Einträgen des Conceptboards für die Items der jeweiligen Skala wurden dann Kategorien abgeleitet, denen die gesammelten Beiträge inhaltlich zugeordnet wurden. Im nächsten Schritt war es möglich, Items für jede Kategorie zu bilden, die jedoch vorerst als ganze Sätze formuliert wurden. Eine Erstellung der Items im Format des UEQ+ hätte die frühzeitige Festlegung eines Einleitungssatzes erfordert, was die

Generierung von Items eingeschränkt und erschwert hätte. Die aus der Literatur abgeleiteten Sicherheits- und Effektivitätsaspekte wurden ebenfalls einbezogen, indem sie mit den Kategorien des Workshops abgeglichen wurden. Aspekte, die im Workshop nicht genannt wurden, aber in der Literatur vorkamen, wurden zum Itempool hinzugefügt (s. Anhang A, Tabellen A1 und A2). Genauso wurde mit den Items des früheren Projekts von Siemens Healthineers aus den Jahren 2015/16 verfahren. Daraus resultierten zwei vorläufige Itemlisten für die beiden Zusatzskalen.

2.1.2.1 Definition und generierte Items der Skala Sicherheit. Zunächst wird auf die Definition sowie die generierten Items der Skala Sicherheit eingegangen. Der Fokus der Definitionen und Schlagworte aus dem Workshop lag darauf, dass Gefährdungen für alle Beteiligten ausgeschlossen sein müssen. Das stimmt im Wesentlichen mit der ISO-Definition überein, die Sicherheit allgemein als „freedom from risk […] which is not tolerable“ (2014, Abs. 3.14) definiert. Im Hinblick auf Usability bei Medizinprodukten spezifiziert die ISO: „Unacceptable risk can arise from use error, which can lead to exposure to direct physical hazards or loss or degradation of clinical functionality“ (2015, Abs. 1). Ein weiterer wichtiger Punkt, der in den Workshops mehrmals genannt wurde, war das Thema Datensicherheit. Ein Teilnehmer konkretisierte zum Beispiel, dass Geräte nicht unerkannt verändert werden dürften. Die wahrgenommene Datensicherheit wird allerdings bereits durch die Skala Vertrauen des UEQ+ abgedeckt, die ebenfalls zu den produktübergreifend wichtigsten Skalen für Medizinprodukte zählt, weshalb Datensicherheit für die Skala Sicherheit ausgespart wurde.

Die aus dem Workshop abgeleiteten Items zur Skala Sicherheit sind in untenstehender Tabelle 3 abgetragen. Einige Items wurden ausgeschlossen, weil sie sich im Rahmen von UX-Testungen nicht beurteilen lassen, eine zu starke Überschneidung mit anderen Skalen bestand oder sie nicht spezifisch genug formuliert waren. Ausgeschlossene Items sind am Ende von Tabelle 3 aufgelistet. Andere Items überschnitten sich zwar auch mit Skalen des UEQ+, sie wurden aber beibehalten, wenn die Skala nicht zu den sechs UX-Skalen gehörte, die sich in der externen Vorstudie als besonders wichtig für Medizinprodukte erwiesen hatten. Dieses Vorgehen wurde gewählt, um zu verhindern, dass manche Punkte, die den Teilnehmenden des Workshops wichtig erschienen im endgültigen UX-Fragebogen gar nicht abgefragt werden. Es lässt sich außerdem kaum vermeiden, dass UX-Skalen sich inhaltlich überschneiden (Schrepp & Thomaschewski, 2019c). Ein Beispiel für ein solches Item ist „Ich fühle mich nicht gut eingewiesen in die Funktionsweise des Systems“, das zur Skala Erlernbarkeit gezählt werden könnte, jedoch auch die

Sicherheit beeinflusst. Es verblieben 29 Items in der vorläufigen Itemliste für die Skala Sicherheit. Der fertige UX-Fragebogen soll, wie erwähnt, für alle Business Lines von Siemens Healthineers, einschließlich Reading & Reporting-Software, eingesetzt werden. Für die Skala Sicherheit wurden allerdings einige Items gesammelt, die sich ausschließlich auf Hardware-Bestandteile beziehen, wie zum Beispiel das Item „Es besteht Kollisionsgefahr“. Diese Tatsache erforderte Abweichungen im geplanten Vorgehen, worauf später genauer eingegangen wird.

Tabelle 3

Vorläufige Items der Skala Sicherheit, die aus dem Workshop zur Itemgenerierung abgeleitet wurden

Item	Negativer Pol	Positiver Pol
Allgemeine Items (auf Soft- und Hardwarekomponenten anwendbar)		
1	Es sind unzulässige Veränderungen möglich.	Es können keine unzulässigen Veränderungen vorgenommen werden.
2	Das System erscheint/wirkt unsicher.	Das System erscheint/wirkt sicher.
3	Der Ablauf von Diagnostik/Intervention ist fehleranfällig.	Der Ablauf von Diagnostik/Intervention beugt Fehlern vor/ist reibungslos.
4	Ich fühle mich unsicher in der Bedienung.	Ich fühle mich sicher in der Bedienung.
5	Ich fühle mich nicht gut eingewiesen in die Funktionsweise des Systems.	Ich fühle mich ausreichend gut vorbereitet auf die Nutzung des Systems.
6	Bei einem Fehler muss alles abgebrochen werden.	Einzelne Schritte können rückgängig gemacht werden.
7	Ich kann mir wichtige Sicherheitsfunktionen nur schwer merken.	Ich kann mir wichtige Sicherheitsfunktionen leicht merken.
8	Nutzungsfehler sind wahrscheinlich.	Das Risiko für Nutzungsfehler ist minimal.

Item	Negativer Pol	Positiver Pol
9	Das System stürzt bei Fehlern ab.	Das System reagiert robust auf Fehler.
10	Akustische und optische Warnungen können nicht eindeutig interpretiert werden.	Akustische und optische Warnungen sind in ihrer Bedeutung verständlich.
11	Das System gibt keine oder zu spät Hinweise auf mögliche Gefahren.	Das System weist mich rechtzeitig auf mögliche Gefahren hin.
12	Es fehlen sicherheitskritische Informationen.	Es werden alle sicherheitsrelevanten Informationen und Arbeitsschritte angezeigt.
13	Fehlermeldungen sind nichtssagend.	Fehlermeldungen sind aussagekräftig.
14	Wurden Aktionen einmal in Gang gesetzt, lassen sie sich nicht stoppen.	Aktionen können, wenn nötig, unterbrochen werden.
15	Ich habe keine Möglichkeit mich in Ausnahmesituationen über Sicherheitsmechanismen hinwegzusetzen.	Sicherheitsmechanismen können in Ausnahmesituationen außer Kraft gesetzt werden.
16	Ich habe den Eindruck, dass Hard- oder Softwareausfälle sich nur schwer beheben lassen.	Ich habe den Eindruck, dass Hard- oder Softwareausfälle schnell behoben werden können.
17	In Notfällen habe ich keinen Handlungsspielraum.	Es gibt ausreichend Absicherung im Notfall, wie zum Beispiel Notfalltasten.
18	Es bestehen Verwechslungsgefahren.	Alle Kennzeichnungen und Funktionen sind eindeutig.
19	Es gibt keine Kontrollmechanismen (um z.B. das versehentlich endgültige Löschen von Daten zu verhindern oder Eingaben auf Plausibilität zu prüfen).	Kontrollmechanismen sorgen für mehr Sicherheit (indem sie z.B. prüfen, ob Eingaben plausibel sind oder einen zusätzlichen Schritt zum Löschen von Daten erfordern).

Item	Negativer Pol	Positiver Pol
20	Das System hindert mich nicht daran, falsche Entscheidungen zu treffen.	Falsche Entscheidungen werden unterbunden.
Hardware-Items		
21	Es besteht Kollisionsgefahr.	Kollisionen mit dem Gerät sind ausgeschlossen.
22	Aktionen und Bewegungen des Geräts können versehentlich ausgelöst werden.	Das versehentliche Auslösen von Aktionen oder Bewegungen ist nicht möglich.
23	Das Produkt ist schwer zu reinigen/desinfizieren.	Das Produkt ist leicht zu reinigen/desinfizieren.
24	Das Produkt wirkt anfällig und fragil.	Das Produkt wirkt robust und solide verarbeitet.
25	Das Gerät könnte Schäden an anderen Objekten verursachen.	Schäden an anderen Objekten sind ausgeschlossen.
26	Ich habe das Gefühl, dass physische Gefahren, wie zu hohe Strahlung, gegeben sind.	Ich habe das Gefühl, dass keine physischen Gefahren, wie zu hohe Strahlung, gegeben sind.
27	Es könnte Kontakt zu gefährlichen Bereichen entstehen.	Gefährliche Bereiche des Systems sind geschützt.
28	Es könnten Patient*innen und/oder Anwender*innen verletzt werden.	Eine Verletzung von Patient*innen und/oder Anwender*innen durch das Gerät ist sehr unwahrscheinlich.
29	Es ist keine Barrierefreiheit gegeben.	Das Produkt ist barrierefrei.
Ausgeschlossene Items [a]		
30	Das Gerät lässt sich schlecht manövrieren. (Eher Effizienz)	Das Gerät ist mobil (kann z.B. an alle Körperbereiche gelangen). (Eher Effizienz)

Item	Negativer Pol	Positiver Pol
31	Das Manual ist unverständlich. (Das Manual ist i.d.R. nicht bekannt, wenn ein Produkt bei einem UX-Test vorgestellt wird)	Das Manual ist sehr verständlich. (Das Manual ist i.d.R. nicht bekannt, wenn ein Produkt bei einem UX -Test vorgestellt wird)
32	Das System erlegt mir teils hinderliche Sicherheitsmechanismen auf. (Eher Effizienz)	Sicherheitsmechanismen behindern meine Arbeit nicht. (Eher Effizienz)
33	Das Produkt basiert auf unüblichen Interaktionsprinzipien. (Eher Effizienz)	Die Verwendung des Produkts basiert auf gängigen Interaktionsprinzipien. (Eher Effizienz)
34	Die Funktionsweise des Produkts erscheint ungewohnt und fremd. (Eher Effizienz)	Die Funktionsweise des Produkts erscheint vertraut. (Eher Effizienz)
35	Es erfolgen keine regelmäßigen Sicherheitsupdates. (Lässt sich nicht beantworten, wenn ein Produkt bei einem UX -Test vorgestellt wird)	Sicherheitsupdates erfolgen regelmäßig und automatisch. (Lässt sich nicht beantworten, wenn ein Produkt bei einem UX -Test vorgestellt wird)
36	Die kognitiven Anforderungen sind so hoch, dass Fehler begünstigt werden. (Nicht direkt Sicherheit)	Die kognitiven Anforderungen sind nicht zu hoch. (Nicht direkt Sicherheit)
37	Ich habe das Gefühl, dass wichtige Daten verloren gehen können (z.B. Bildverluste). (Eher Vertrauen)	Ich habe den Eindruck, dass wichtige Daten selbst bei Fehlbedienung gesichert sind. (Eher Vertrauen)

Anmerkung. Items wurden nach dem Workshop teilweise umformuliert, um ihre Verständlichkeit zu verbessern oder um Aspekte aus der Literatur zu integrieren. In Anhang A ist zu sehen, welche Aspekte aus der Literatur integriert wurden (Tabelle A1).

[a] Der Grund für den Ausschluss wird in Klammern erläutert.

2.1.2.2 Definition und generierte Items der Skala Effektivität. Die Auswertung und die Ergebnisse des Workshops zur Generierung von Items für die Skala Effektivität werden in diesem Abschnitt berichtet. Wie für die Skala Sicherheit wird zuerst auf die Definition von Effektivität bei Medizinprodukten und anschließend auf die entwickelten Items eingegangen. Nach der ISO ist Effektivität die „accuracy and completeness with which users achieve specified goals“ (2018, Abs. 3.1.12). Die Zielerreichung im Kontext von Medizinprodukten wurde von den Teilnehmenden des Workshops als Diagnosesicherheit oder als Erfolg einer medizinischen Intervention spezifiziert. Um diese zu gewährleisten, ist eine optimale und konstant hohe Bild- und Befundqualität entscheidend. Bei der Frage danach, wie die Teilnehmenden Effektivität definieren würden, wurde zudem mehrmals genannt, dass eine stabile Performance und das Nicht-Auftreten von Störungen ein medizinisches Gerät effektiv machen würden. Das Gerät müsse unmittelbar einsatzbereit sein, wenn es benötigt werde. Es wurde auch darauf eingegangen, dass eine Untersuchung oder Intervention aus mehreren Teilschritten bestehe, die einzeln vollständig bewältigt werden müssten. Eine sinnvolle, logische Anordnung der Arbeitsschritte sei dazu hilfreich. Gerade die letzten Punkte zeigen, dass es schwierig ist, Effektivität von Effizienz zu trennen. Ebenso wie Effektivität betrachtet auch Effizienz die erzielten Ergebnisse, allerdings anders als Effektivität im Verhältnis zu den eingesetzten Ressourcen (International Organization for Standardization, 2018, Abs. 3.1.13). Ressourcen sind beispielsweise die aufgewendete Zeit, Kosten und das Material. Die Teilnehmenden konnten diese Unterscheidung nach der ISO zwar nachvollziehen, während der Ideensammlung kam dennoch immer wieder die Diskussion auf, was nun unter Effizienz und was unter Effektivität fällt. Eine Teilnehmerin begründete das mit gemeinsamen Einflussfaktoren. Eine logische Anordnung der Arbeitsschritte zum Beispiel könnte den Prozess beschleunigen und gleichzeitig das Ergebnis verbessern.

Die vorläufige Itemliste nach Auswertung des zweiten Workshops ist in Tabelle 4 dargestellt. Auch für die Skala Effektivität wurden einige Items aufgrund einer zu starken Überschneidung mit der Skala Effizienz ausgeschlossen. Für die Skala Effektivität verblieben 26 Items in der vorläufigen Itemliste.

Tabelle 4

Vorläufige Items der Skala Effektivität, die aus dem Workshop zur Itemgenerierung abgeleitet wurden

Item	Negativer Pol	Positiver Pol
1	Das Gerät ist leistungsschwach.	Das Gerät ist leistungsfähig.
2	Ergebnisse stehen nicht zeitnah zur Verfügung.	Ergebnisse stehen unmittelbar zur Verfügung.
3	Wichtige Informationen fehlen oder sind schwer zugänglich.	Wichtige Informationen sind leicht zugänglich.
4	Es werden irrelevante Informationen angezeigt.	Es werden nur relevante und häufig gebrauchte Informationen angezeigt.
5	Es werden redundante Informationen angezeigt.	Es gibt keine unnötige Redundanz.
6	Es fehlen notwendige Funktionen.	Das Gerät bietet alle für die Aufgabenbearbeitung notwendigen Funktionen.
7	Manche Funktionen sind überflüssig.	Es gibt keine überflüssigen Funktionen.
8	Manche Funktionen sorgen für Verwirrung.	Es ist klar, welche Funktionen wofür gedacht sind.
9	Das Gerät ermöglicht keine eindeutige Diagnosestellung (bzw. keine erfolgreiche Intervention).	Das Gerät ermöglicht eine eindeutige Diagnosestellung (bzw. eine erfolgreiche Intervention).
10	Das gewünschte Ergebnis kann nicht erzielt werden.	Das gewünschte Ergebnis kann mit dem Gerät erzielt werden.
11	Das Gerät unterstützt mich nicht beim Erreichen meiner Ziele.	Das Gerät unterstützt das Erreichen meiner Ziele.
12	Bilder sind schwer zu interpretieren.	Bilddaten lassen sich eindeutig interpretieren.

Item	Negativer Pol	Positiver Pol
13	Die Ergebnisse müssen aufwendig nachbearbeitet werden.	Die Ergebnisse können direkt genutzt werden.
14	Das System bietet keine Entscheidungsunterstützung.	Das System unterstützt die Entscheidungsfindung.
15	Die Qualität der Ergebnisse ist mangelhaft.	Die Ergebnisse sind von hoher Qualität.
16	Die Qualität der Ergebnisse ist wechselhaft.	Die Qualität der Ergebnisse ist konstant.
17	Die Bildqualität lässt teilweise zu wünschen übrig.	Die Bildqualität ist auch bei Bewegungen hoch.
18	Prozeduren müssen wiederholt werden, um zum gewünschten Ergebnis zu führen.	Prozeduren führen direkt zum gewünschten Ergebnis.
19	Meine Aufgaben lassen sich mit dem System nicht vollständig erfüllen.	Ich kann meine Aufgaben mit dem System vollständig erfüllen.
20	Das System ist nicht praxistauglich.	Das System lässt sich problemlos in die Praxis übertragen.
21	Das System schneidet nicht besser ab als vergleichbare Systeme.	Das System bietet einen Mehrwert gegenüber vergleichbaren Systemen.
22	Eine Anpassung des Systems an Abläufe in der Klinik ist nicht möglich.	Das System kann individualisiert und angepasst werden.
23	Der Ablauf der Arbeitsschritte ist zu unflexibel.	Der Ablauf der Arbeitsschritte ist adaptiv.
24	Begriffe und Bezeichnungen sind nicht auf meine Arbeitstätigkeit abgestimmt.	Die verwendeten Begriffe und Bezeichnungen entsprechen denen meiner Arbeitstätigkeit.

Item	Negativer Pol	Positiver Pol
25	Das System ist nicht auf die von mir zu bearbeitenden Aufgaben zugeschnitten.	Das System ist auf die von mir zu bearbeitenden Aufgaben zugeschnitten.
26	Das Gerät ist mit anderen Geräten inkompatibel.	Das Gerät ist mit anderen Geräten kompatibel.
Ausgeschlossene Items [a]		
27	Die Arbeitsschritte folgen einem unlogischen Ablauf. (Eher Effizienz)	Die Arbeitsschritte folgen einem logischen Ablauf. (Eher Effizienz)
28	Eine vollständige Bearbeitung zusammenhängender Arbeitsabläufe ist nicht möglich. (Eher Effizienz)	Ich kann zusammenhängende Arbeitsabläufe vollständig bearbeiten. (Eher Effizienz)
29	Die Handhabung des Geräts ist schwerfällig und nicht intuitiv. (Eher Effizienz)	Das Gerät ist intuitiv bedienbar und leicht zu handhaben. (Eher Effizienz)

Anmerkung. Items wurden nach dem Workshop teilweise umformuliert, um ihre Verständlichkeit zu verbessern oder um Aspekte aus der Literatur zu integrieren. In Anhang A ist zu sehen, welche Aspekte aus der Literatur integriert wurden (Tabelle A2).

[a] Der Grund für den Ausschluss wird in Klammern erläutert.

2.2 Workshop zur Konsolidierung und Reduktion des Itempools

Es wurde ein weiterer Workshop durchgeführt, um die Ergebnisse der vergangenen beiden Workshops zu konsolidieren. Das Vorgehen war angelehnt an die Entwicklung des UEQ von Laugwitz et al. (2008), die ebenfalls Expert*innen zur Konsolidierung und Vorauswahl der Items befragten, bevor die Items in einer Studie evaluiert wurden.

2.2.1 Ablauf des Workshops zur Konsolidierung und Reduktion des Itempools.

Für den Workshop wurden erneut vier Siemens-Healthineers-interne Expert*innen eingeladen. Zwei der Teilnehmenden führten regelmäßig UX-Testungen zur Evaluation neuer Prototypen und Produkte durch. Die anderen beiden Teilnehmenden waren Expert*innen für Produkte aus den Bereichen MR und XP und hatten beide früher als MTRAs gearbeitet. Der etwa 90-minütige Workshop wurde wieder über Microsoft Teams (Microsoft, 2022) durchgeführt. Nachdem sich die Anwesenden kurz vorgestellt hatten, wurden auch hier Hintergrundinformationen bezüglich der User Experience, des UEQ+ und zum Aufbau von Fragebögen präsentiert. Auf einem Conceptboard (Conceptboard, 2022), wovon Ausschnitte in Anhang B (Abbildung B1) einsehbar sind, wurden diesmal die gesammelten Items aus den vergangenen Workshops präsentiert. Daneben wurden die Definitionen von Sicherheit und Effektivität sowohl nach der ISO als auch die im Workshop gesammelten Definitionen auf dem Conceptboard abgebildet, jedoch nicht erneut diskutiert. Die Teilnehmenden wurden in Gruppen von je zwei Personen aufgeteilt, so dass pro Team ein*e UX-Expert*in und ein*e Produktexpert*in zusammenarbeiteten. Die Aufgabe der Teilnehmenden war es, die Items per Drag & Drop in eine Rangfolge zu bringen, mit den wichtigsten beziehungsweise für einen Fragebogen geeignetsten Items an oberster Stelle. Es wurde darauf hingewiesen, dass Aspekte, die relevant sind, aber sehr selten auftreten, für einen Fragebogen unter Umständen nicht geeignet sind und daher nicht hoch priorisiert werden müssten. Die Aufgabe der UX-Expert*innen war es außerdem, darauf zu achten, dass sich die Items im Rahmen einer UX-Testung überhaupt beurteilen lassen. Die Teilnehmenden hatten die Möglichkeit Kommentare zu unpassenden Begriffen zu notieren und unpassende Items in einen Veto-Kasten zu verschieben. Die erste Gruppe beschäftigte sich mit den Items der Skala Sicherheit, die zweite mit den Items der Skala Effektivität.

2.2.2 Auswertung des Workshops zur Konsolidierung und Reduktion des Itempools.

Den Teilnehmenden fiel es schwer, die Items in eine eindeutige Rangfolge zu bringen. Beide Gruppen belegten Rangplätze mit mehreren Items, wodurch sich keine eindeutigen Rangfolgen bilden ließen. Die Teilnehmenden fügten Kommentare hinzu, die in die Auswertung einbezogen wurden. So wurde zum Beispiel an manchen Items der Skala Sicherheit kritisiert, dass sie zu subjektiv seien. Beispielsweise das Item „Ich fühle mich unsicher/sicher in der Bedienung“ hänge zu stark von der bedienenden Person ab. Für die Items der Skala Effektivität wurde bei mehreren Items eine Überschneidung mit Effizienz angemerkt. Manche Items wurden deshalb auf Basis der Kommentare oder Anmerkungen in der anschließenden Diskussion ausgeschlossen. Der zuvor generierte Itempool konnte so auf Eignung und Verständlichkeit der Items hin überprüft und bereinigt werden. Bei der Entwicklung des UEQ (Laugwitz et al., 2008) wurde eine deutlich größere Anzahl an Items generiert. Die von Expert*innen angegebene Rangfolge der UEQ-Items diente deshalb vor allem auch dazu, eine Reduktion der initialen 229 Items vorzunehmen. So mussten in der anschließenden Studie nur die 80 wichtigsten anstelle der ursprünglichen 229 Items evaluiert werden. In der vorliegenden Arbeit wurde in den ersten beiden Workshops eine deutlich geringere Anzahl an Items generiert, weshalb der Workshop vorrangig der Konsolidierung und weniger der Reduktion der Items diente.

2.2.2.1 Konsolidierte Items der Skala Sicherheit. Nach dem Workshop zur Konsolidierung verblieben für die Skala Sicherheit 25 Items, wovon sich 10 ausschließlich auf Hardware-Komponenten anwenden ließen. Es wurden deshalb keine Items aufgrund der Rangfolge für diese Skala ausgeschlossen. Die Rangfolge war dennoch hilfreich, um herauszufinden, ob Hardware-Items für die Teilnehmenden einen hohen Stellenwert hatten. Wäre das nicht der Fall gewesen, hätte man die Skala auf allgemeine Sicherheitsitems (Items, die sich auf Soft- und Hardwarekomponenten anwenden lassen) beschränken können, die dann auch für reine Software-Anwendungen der Business Line D&A nutzbar gewesen wären. Die Items, die sich ausschließlich auf Hardware-Komponenten bezogen, wurden im Workshop jedoch als sehr wichtig für die Skala Sicherheit bewertet, wie das Ranking der wichtigsten Items für die Skala Sicherheit in Tabelle 5 zeigt. Es ist somit nötig, Hardware-Items in den UX-Fragebogen aufzunehmen. Ein mögliches Vorgehen wäre, der Skala Sicherheit zusätzliche Hardware-Items hinzuzufügen, welche für die Business Line D&A weggelassen werden. So könnte eine Skala beibehalten werden, die für alle Business Lines vier allgemeine Sicherheitsitems umfasst und für Business Lines mit

Hardware-Komponenten zwei oder drei zusätzliche Hardware-spezifische Items beinhaltet. Alternativ könnte man eine separate Sicherheitsskala bezogen auf Hardwarekomponenten einführen. Welches Vorgehen sinnvoller ist, wird die geplante explorative Faktorenanalyse zeigen. Sollten die Items auf verschiedene Faktoren laden, ist es nicht sinnvoll alle Items in eine Skala aufzunehmen und für die Auswertung zu mitteln.

Tabelle 5

Rangfolge nach Wichtigkeit und Eignung der Sicherheitsitems aus dem Workshop zur Konsolidierung und Itemreduktion

Rang-folge	Negativer Pol	Positiver Pol
1	Es könnten Patient*innen und/oder Anwender*innen verletzt werden. [a]	Eine Verletzung von Patient*innen und/oder Anwender*innen durch das Gerät ist sehr unwahrscheinlich. [a]
	Ich habe das Gefühl, dass physische Gefahren, wie zu hohe Strahlung, gegeben sind. [a]	Ich habe das Gefühl, dass keine physischen Gefahren, wie zu hohe Strahlung, gegeben sind. [a]
2	Es besteht Kollisionsgefahr. [a]	Kollisionen mit dem Gerät sind ausgeschlossen. [a]
	Das Gerät könnte Schäden an anderen Objekten verursachen. [a]	Schäden an anderen Objekten sind ausgeschlossen. [a]
3	Das System erscheint/wirkt unsicher.	Das System erscheint/wirkt sicher.
	Das Produkt wirkt anfällig und fragil. [a]	Das Produkt wirkt robust und solide verarbeitet. [a]
4	Nutzungsfehler sind wahrscheinlich.	Das Risiko für Nutzungsfehler ist minimal.
5	Das System gibt keine oder zu spät Hinweise auf mögliche Gefahren.	Das System weist mich rechtzeitig auf mögliche Gefahren hin.
6	Fehlermeldungen sind nichtssagend.	Fehlermeldungen sind aussagekräftig.

Rang-folge	Negativer Pol	Positiver Pol
	Akustische und optische Warnungen können nicht eindeutig interpretiert werden.	Akustische und optische Warnungen sind in ihrer Bedeutung verständlich.
	Es bestehen Verwechslungsgefahren.	Alle Kennzeichnungen und Funktionen sind eindeutig.
7	Aktionen und Bewegungen des Geräts können versehentlich ausgelöst werden. [a]	Das versehentliche Auslösen von Aktionen oder Bewegungen ist nicht möglich. [a]
8	In Notfällen habe ich keinen Handlungsspielraum. [a]	Es gibt ausreichend Absicherung im Notfall, wie zum Beispiel Notfalltasten. [a]
	Ich habe keine Möglichkeit mich in Ausnahmesituationen über Sicherheitsmechanismen hinwegzusetzen. [a]	Sicherheitsmechanismen können in Ausnahmesituationen außer Kraft gesetzt werden. [a]
9	Der Ablauf von Diagnostik/Intervention ist fehleranfällig.	Der Ablauf von Diagnostik/Intervention beugt Fehlern vor/ist reibungslos.
10	Wurden Aktionen [b] einmal in Gang gesetzt, lassen sie sich nicht stoppen.	Aktionen [b] können, wenn nötig, unterbrochen werden.
11	Es sind unzulässige Veränderungen möglich.	Es können keine unzulässigen Veränderungen vorgenommen werden.
	Das System hindert mich nicht daran, falsche Entscheidungen zu treffen.	Falsche Entscheidungen werden unterbunden.

Rang-folge	Negativer Pol	Positiver Pol
12	Es gibt keine Kontrollmechanismen (um z.B. das versehentlich endgültige Löschen von Daten zu verhindern oder Eingaben auf Plausibilität zu prüfen).	Kontrollmechanismen sorgen für mehr Sicherheit (indem sie z.B. prüfen, ob Eingaben plausibel sind oder einen zusätzlichen Schritt zum Löschen von Daten erfordern).
	Es fehlen sicherheitskritische Informationen.	Es werden alle sicherheitsrelevanten Informationen und Arbeitsschritte angezeigt.
13	Das Produkt ist schwer zu reinigen/desinfizieren. [a]	Das Produkt ist leicht zu reinigen/desinfizieren. [a]
14	Es ist keine Barrierefreiheit gegeben. [a]	Das Produkt ist barrierefrei. [a]
15	Bei einem Fehler muss alles abgebrochen werden.	Einzelne Schritte können rückgängig gemacht werden.
	Das System stürzt bei Fehlern ab.	Das System reagiert robust auf Fehler.
	Es könnte Kontakt zu gefährlichen Bereichen [b] entstehen. [a]	Gefährliche Bereiche [b] des Systems sind geschützt. [a]

Anmerkung. Die Rangfolge wurde aus dem Workshop übernommen, Hardware- und Software-bezogene Items sind daher nicht separiert. Oben stehen die wichtigsten/geeignetsten Items. In mehreren Fällen belegten Items denselben Rangplatz, sie sind untereinander abgetragen.

[a] Items, die sich ausschließlich auf Hardware beziehen.

[b] Diese Wörter wurden von Teilnehmenden des Workshops als zu unkonkret kommentiert.

2.2.2.1 Konsolidierte Items der Skala Effektivität. Die zweite Gruppe, die sich mit der Skala Effektivität beschäftigte, brachte die Items nicht direkt in eine Rangfolge. Sie wählte ein anderes Vorgehen, indem zuerst bunte Kreise an besonders wichtige Items geheftet wurden und dann inhaltliche Kategorien gebildet wurden. Die Teilnehmenden sortierten die Kategorien anschließend nach ihrer Wichtigkeit. Innerhalb der Kategorien wurde keine Rangordnung festgelegt. Für die Items der Skala Effektivität wurden die nach Wichtigkeit geordneten Kategorien genutzt, um, wie bei der Entwicklung des UEQ, die Anzahl der Items zu reduzieren. Es wurden so viele Effektivitätsitems wie allgemeine Sicherheitsitems aufgenommen. Das entsprach den drei wichtigsten Kategorien der Skala Effektivität, die zusammen 15 Items umfassten. Die drei wichtigsten Kategorien mit ihren zugehörigen Items sind in Tabelle 6 dargestellt.

Tabelle 6

Rangfolge der Wichtigkeit und Eignung der Effektivitätsitems aus dem Workshop zur Konsolidierung und Itemreduktion

Rangfolge	Negativer Pol	Positiver Pol
1. Kategorie „generelle Zielerreichung“	Das gewünschte Ergebnis kann nicht erzielt werden.	Das gewünschte Ergebnis kann mit dem Gerät erzielt werden.
	Meine Aufgaben lassen sich mit dem System nicht vollständig erfüllen.	Ich kann meine Aufgaben mit dem System vollständig erfüllen.
	Es fehlen notwendige Funktionen.	Das Gerät bietet alle für die Aufgabenbearbeitung notwendigen Funktionen.
	Das Gerät ermöglicht keine eindeutige Diagnosestellung (bzw. keine erfolgreiche Intervention).	Das Gerät ermöglicht eine eindeutige Diagnosestellung (bzw. eine erfolgreiche Intervention).
2. Kategorie „keine Behinderung“	Das System ist nicht praxistauglich.	Das System lässt sich problemlos in die Praxis [a] übertragen.
	Wichtige Informationen fehlen oder sind schwer zugänglich.	Wichtige Informationen sind leicht zugänglich.

Rangfolge	Negativer Pol	Positiver Pol
	Begriffe und Bezeichnungen sind nicht auf meine Arbeitstätigkeit abgestimmt.	Die verwendeten Begriffe und Bezeichnungen entsprechen denen meiner Arbeitstätigkeit.
	Das System ist nicht auf die von mir zu bearbeitenden Aufgaben zugeschnitten.	Das System ist auf die von mir zu bearbeitenden Aufgaben zugeschnitten.
	Das Gerät unterstützt mich nicht beim Erreichen meiner Ziele.	Das Gerät unterstützt das Erreichen meiner Ziele.
	Bilder sind schwer zu interpretieren.	Bilddaten lassen sich eindeutig interpretieren.
	Das System bietet keine Entscheidungsunterstützung.	Das System unterstützt die Entscheidungsfindung.
3. Kategorie „Qualität“	Die Qualität der Ergebnisse ist mangelhaft.	Die Ergebnisse sind von hoher Qualität.
	Die Bildqualität lässt teilweise zu wünschen übrig.	Die Bildqualität ist auch bei Bewegungen [a] hoch.
	Die Qualität der Ergebnisse ist wechselhaft.	Die Qualität der Ergebnisse ist konstant.
	Das System schneidet nicht besser ab als vergleichbare Systeme.	Das System bietet einen Mehrwert gegenüber vergleichbaren Systemen.

Anmerkung. Es wurden Kategorien von den Teilnehmenden gebildet, welche in eine Rangfolge gebracht wurden. Oben steht die wichtigste Kategorie. Die Items innerhalb der Kategorien wurden nicht nach ihrer Wichtigkeit oder Eignung sortiert.

[a] Diese Wörter wurden von Teilnehmenden des Workshops als zu unkonkret kommentiert.

2.3 Überführung der Items in das Format des UEQ+

Auf Basis der Ergebnisse der Workshops wurden die Ergebnisse zuletzt in das Format des UEQ+ übertragen, welches aus einem Einleitungssatz und semantischen Differentialen besteht (Schrepp & Thomaschewski, 2019a). Für die Bildung eines Einleitungssatzes wurden die Skalendefinitionen herangezogen und es wurde betrachtet, welche Wörter besonders häufig in den Items vorkamen, sodass sie sich möglichst einfach an den Einleitungssatz anpassen ließen. Es fand eine interne Besprechung mit Kolleg*innen der UX-Abteilung und der ehemaligen Masterandin Tanja Müßig statt, während der Einleitungssätze diskutiert und die Items entsprechend umgeformt wurden. Für einige Items ließ sich diese Umformung allerdings nicht sinnvoll umsetzen, weshalb sie umformuliert oder sogar ausgeschlossen werden mussten. Die Anzahl der Items veränderte sich durch dieses Vorgehen, sodass 13 allgemeine und sieben Hardware-spezifische Items für die Skala Sicherheit verblieben. Für die Skala Effektivität verblieben 14 Items. Die Items im Format des UEQ+ sind in den Abbildungen 7 und 8 abgebildet, die sich im Methodenteil der externen Studie (Kapitel 3.2.3) befinden. Eine Abbildung, welche die Überführung der Items in das Format des UEQ+ verdeutlicht ist im Anhang einsehbar (Anhang C, Tabellen C1 und C2).

3 Methode Schritt 2 – Externe Studie

Im Rahmen der externen Online-Studie wurden die durch die Workshops erhaltenen Items an medizinischem Personal evaluiert. Darüber hinaus sollten die Ergebnisse der Vorstudie von Tanja Müßig (2022) repliziert werden. Die Methodik der externen Online-Studie wird in diesem Kapitel erläutert.

3.1 Stichprobe

Im ersten Abschnitt der Methodenbeschreibung werden zunächst die Stichprobenplanung, der Ablauf der Rekrutierung und die Stichprobenbeschreibung aufgegriffen.

3.1.1 Ermittlung der Stichprobengröße.

Für Faktorenanalysen ist unter den geplanten Auswertungsverfahren die größte Stichprobe erforderlich. Die Empfehlungen für eine optimale Stichprobengröße unterscheiden sich allerdings erheblich. Manche Autor*innen geben absolute Werte vor, Comrey und Lee (1992) geben beispielsweise eine Stichprobengröße von 200 Personen als angemessen an, eine Größe von 300 als gut. Ebenso empfiehlt Bühner (2011) eine Stichprobengröße von etwa 250 Personen. Andere Autor*innen betrachten das Verhältnis der Anzahl der Proband*innen zu Variablen, wobei manche zu einem von Verhätnis von 3:1 (Cattell, 1978) und andere zu einem Verhältnis von 10:1 (Everitt, 1975) raten. Vorschläge

gehen somit deutlich auseinander (MacCallum et al., 1999). Weitere Empfehlungen gehen dahin, die Kommunalitäten der Variablen (Kopp & Lois, 2014, S. 95; Mundfrom et al., 2005) oder das Verhältnis von manifesten Variablen zu Faktoren heranzuziehen (Preacher & MacCallum, 2002). Pro Faktor sollten mindestens drei, besser mehr Items gegeben sein (Kopp & Lois, 2014). Allgemein ist festzuhalten, dass die Anzahl der Proband*innen umso größer sein sollte, „je größer die Anzahl der Faktoren, je niedriger die Kommunalitäten und je kleiner die Anzahl der Items pro Faktor“ (Kopp & Lois, 2014, S. 95). Angesichts der kaum zu überblickenden bestehenden Empfehlungen, denen teils wenig empirische Evidenz zugrunde liegt (Mundfrom et al., 2005), ist es nicht einfach, eine geeignete Stichprobengröße abzuleiten. Da einige Empfehlungen im Bereich von 200 bis 300 Personen liegen (Bühner, 2011; Comrey & Lee, 1992; Gagne & Hancock, 2006; Goretzko et al., 2021), wird sich an dieser Zahl orientiert. Eine größere Stichprobe ist für Faktorenanalysen immer von Vorteil (Mundfrom et al., 2005), allerdings muss auch die praktische Umsetzbarkeit berücksichtigt werden, denn in der vorliegenden Arbeit werden nicht Personen aus der Allgemeinbevölkerung rekrutiert, sondern es wird ausschließlich medizinisches Personal einbezogen. Es ist deshalb geplant, pro Business Line 40 Proband*innen zu rekrutieren. In Abhängigkeit davon, ob Teilnehmende für Strahlentherapie rekrutiert werden können, ergeben sich so 200 bis 240 Personen.

Es wurde mithilfe der Software G*Power (Faul et al., 2007), Version 3.1.9.7, überprüft, wie hoch die Power bei einer Stichprobengröße von 200 Personen für die Berechnung der MANOVA für Hypothese 1 wäre. Die Analyse wurde für fünf anstelle von sechs Gruppen berechnet, denn die Business Line Strahlentherapie wurde vorerst nicht einbezogen. Angenommen wurde, dass ein mittlerer Effekt groß genug wäre, um den Aufwand zu rechtfertigen, für jede Business Line eine eigene Fragebogenversion zu entwerfen. Als Effektstärkemaß ließ in G*Power nur Cohen’s f^2 angeben, ein mittlerer Effekt entspricht $f^2 = 0.15$ (Ellis, 2010). Da für die Berechnung einer MANOVA das partielle Eta-Quadrat ($\eta^2_{part.}$) das gängigste Effektstärkemaß ist, wird dieses jedoch für die Auswertung herangezogen. Die Power für 200 Personen würde für die angenommenen Werte bei einem angesetzten α-Fehlerniveau von 5 % über 99 % betragen. Die für die Faktorenanalyse eingeplante Stichprobengröße ist also mehr als ausreichend für die Berechnung der MANOVA. Die Wahrscheinlichkeit einen (mittleren) Effekt zu finden, wenn er tatsächlich vorliegt, ist sehr hoch.

3.1.2 Rekrutierung der Stichprobe.

Über das Collaboration Management von Siemens Healthineers konnten 51 Proband*innen rekrutiert werden, darunter 18 Personen der Business Line CT, 10 Personen der Business Line MR, 22 Personen der der Business Line XP und eine Person der Business Line D&A. Der Auftrag zur Rekrutierung weiterer 40 Personen aus dem Bereich AT/AT-SU und 20 Personen aus dem Bereich Strahlentherapie wurde an eine Rekrutierungsagentur übergegeben. Es wurde davon ausgegangen, dass die Rücklaufquote über das Collaboration Management höher sein würde, es waren 180 Teilnehmende eingeplant (je 40 Personen für die Business Lines CT, MR, XP, D&A und 20 Personen für die Business Line Strahlentherapie). Im Verlauf der Rekrutierung stellte sich jedoch heraus, dass sich über diesen Weg weniger Teilnehmende anwerben ließen als erwartet. Es wurde zusätzlich versucht, Teilnehmende über die sozialen Netzwerke XING (XING, 2022) und LinkedIn (LinkedIn, 2022) zu rekrutieren, über diesen Weg nahmen aber keine Personen teil. Aus diesem Grund wurde die Anzahl der Teilnehmenden, die über die Agentur rekrutiert werden sollte, erhöht. Die Agentur sollte zusätzlich je 10 Personen für die Business Lines CT und XP sowie je 20 weitere Personen für die Business Lines MR und D&A rekrutieren, um unter Berücksichtigung des eingeplanten Budgets auf annähernd gleich viele Proband*innen für jede Business Line zu kommen. Zwar wurden die Teilnehmenden gezielt für die Business Lines rekrutiert, einige füllten die Online-Studie allerdings für eine andere Business Line aus als jene, für die sie rekrutiert worden waren, sodass die Anzahl der Teilnehmenden für die Business Lines von der geplanten Anzahl abwich. In Tabelle 7 ist zusammengefasst, wie viele Personen für jede Business Line teilnahmen. Für die Business Lines Strahlentherapie und D&A war es besonders schwierig, Proband*innen zu rekrutieren, sodass für diese Bereiche weniger Personen einbezogen werden konnten als für andere Business Lines. Je nach Business Line wurden repräsentative Berufsgruppen rekrutiert, die direkt mit den Geräten arbeiteten. Für die meisten Business Lines waren das vorwiegend medizinisch-technische (Radiologie-)Assistent*innen. Für SPECT und PETCT (der Business Line CT) konnten außerdem nuklearmedizinisch-technische Assistent*innen teilnehmen. (Interventionelle) Radiolog*innen, Kardiolog*innen und Chirurg*innen sowie operationstechnische Angestellte wurden für die Business Line AT/AT-SU rekrutiert. Mit Reading- und Reporting-Software (Business Line D&A) arbeiten ausschließlich Ärzt*innen, was die Rekrutierung für diese Business Line erschwerte. Es wurde nicht festgelegt, zu welchen Anteilen welche Berufsgruppen rekrutiert werden sollten.

Tabelle 7

*Anzahl der Proband*innen unterteilt nach Business Line und Produktkategorie*

Business Line	*n*	Produktkategorie	*n*
CT	33	CT	31
		SPECT/PETCT	2
MR	31	MRT	31
XP	34	Urologie	0
		Mammographie	6
		Radiographie/Röntgen	26
		Fluoroskopie	2
D&A	19	Reading- und Reporting-Software	19
AT/AT-SU	40	Angiographie	16
		C-Bögen	24
Strahlentherapie	16	Bestrahlungssysteme	16
Gesamt	173	alle Produktkategorien	173

Innerhalb der Business Lines, für die Proband*innen gezielt rekrutiert wurden, stand es ihnen frei zu wählen, für welche Produktkategorie dieser Business Lines sie die Umfrage bearbeiten wollten. Für die Anteile der Produktkategorien innerhalb der Business Lines wurden ebenfalls keine Quoten festgelegt.

3.1.3 Beschreibung der Stichprobe.

Insgesamt wurde der Link zur Studie 245-mal aufgerufen, jedoch verblieben nach Ausschluss nicht abgeschlossener Studienfragebögen und Siemens Healthineers Mitarbeitenden nur 173 Personen in der Stichprobe. Unter den Befragten befanden sich 97 weibliche Teilnehmende, 72 männliche Teilnehmende und vier Personen, die divers als Geschlecht angaben. Das durchschnittliche Alter in der Stichprobe betrug 38.91 Jahre (*SD* = 9.88). Die jüngste Person war zum Zeitpunkt der Durchführung 22 Jahre alt, die älteste 63 Jahre. Alle Befragten hatten laut eigenen Angaben gute (*n* = 4), sehr gute (*n* = 11) oder verhandlungssichere (*n* = 4) Deutschkenntnisse. Weitere 154 Teilnehmende teilten mit, Muttersprachler*innen zu sein. Keine*r der Teilnehmenden hatte demnach keine oder nur Grundkenntnisse der deutschen Sprache.

Unter den rekrutierten Personen waren 102 medizinisch-technische Radiologieassistent*innen, eine radiologische Krankenpflegekraft, eine medizinisch-technische Angestellte, 12 operationstechnische Angestellte und kein*e nuklearmedizinisch-technische Assistent*in. Es konnten außerdem 52 Ärzt*innen rekrutiert werden, wovon 30 als (interventionelle) Radiolog*innen, 10 als (interventionelle) Kardiolog*innen, sieben als Chirurg*innen und fünf in sonstigen Fachrichtungen arbeiteten. Weitere fünf Personen gaben an, einer anderen als der aufgeführten Berufsgruppen anzugehören. Im Hinblick auf die Berufserfahrung gab keine*r der Teilnehmenden an, weniger als ein Jahr Erfahrung zu haben. Achtzehn Personen hatten 1 bis 3 Jahre Berufserfahrung, 40 Personen 4 bis 7 Jahre, 19 Personen 8 bis 10 Jahre, 52 Personen 11 bis 20 Jahre, 25 Personen 21 bis 30 Jahre und 19 Personen mehr als 30 Jahre Erfahrung. Einige der Personen hatten Weiterbildungen absolviert, 42 den (kleinen) Röntgenschein, 50 hatten eine Ausbildung zur Schnittbild-Radiologie, 18 eine Zusatzausbildung für Mammographie sowie 44 Personen sonstige Ausbildungen. Die Mehrheit der Teilnehmenden arbeitete 36 bis 40 (n = 72) oder mehr als 40 (n = 72) Stunden pro Woche. Elf Personen berichteten 31 bis 35 Stunden und 14 Personen 21 bis 30 Stunden in der Woche zu arbeiten. Eine Person gab eine wöchentliche Arbeitszeit von 11 bis 20 Stunden an sowie drei Teilnehmende 1 bis 10 Stunden.

Ein großer Anteil der Befragten arbeitete in einer Universitätsklinik (n = 66) oder einem Krankenhaus ohne Anbindung an eine Universität (n = 64). In Praxen arbeiteten 34 der Teilnehmenden und sieben in sonstigen Einrichtungen. Zwei der Befragten machten zu dieser Frage keine Angabe. Die Größe der Krankenhäuser und Unikliniken, in denen Befragte arbeiteten, umfasste in der Regel mehr als 500 Betten (n = 81). Häufig wurde auch eine Anzahl von 300 bis 399 Betten genannt (n = 25). Nur wenige Befragte gaben an, in einer Klinik mit weniger als 300 Betten ($n_{<100\_Betten}$ = 1, $n_{100\text{-}199\_Betten}$ = 3, $n_{200\text{-}299\_Betten}$ = 9) oder einer Klinik mit 400 bis 500 Betten (n = 6) zu arbeiten. Zur Finanzierung der Institution machten 52 der Teilnehmenden keine Angabe, 44 gaben an, dass die Institution, in der sie arbeiteten, privat und 77, dass die Institution staatlich finanziert werde.

Die Mehrheit der Personen gab die Radiologie als Fachrichtung beziehungsweise Abteilung an (n = 125). Am zweit- und dritthäufigsten wurden Chirurgie (n = 20) und Kardiologie (n = 17) angegeben. Die Fachrichtungen beziehungsweise Abteilungen Intensivmedizin (n = 10), Forschung (n = 8), Gynäkologie (n = 8), Neurologie (n = 5), Pädiatrie (n = 5), Onkologie (n = 4) und Dermatologie (n = 1) wurden seltener genannt. Weitere 24 Personen teilten mit, in einer anderen als der aufgeführten Fachrichtung zu

arbeiten. Es gab bei dieser Frage die Möglichkeit, mehr als eine Antwortoption auszuwählen, weshalb die Summe 173 Personen übersteigt. Der Fokus der Abteilung beziehungsweise Praxis lag bei 44 der Befragten auf reiner Diagnostik, bei 12 auf reinen Interventionen. In den meisten Fällen, bei 95 Personen, spielten sowohl Diagnostik als auch Interventionen eine Rolle. Weitere 15 Befragte wählten bei dieser Frage die Kategorie *sonstiges* aus und sieben machten keine Angabe.

3.2 Erhobene Variablen

Die einzelnen Bestandteile der externen Studie werden nun beschrieben. Für eine verständliche Abgrenzung werden die Fragen, die den Teilnehmenden in der externen Studie vorgelegt wurden, als *Studienfragebogen* bezeichnet. Einer Verwechslung mit dem *UX-Fragebogen* für Medizinprodukte, dessen Entwicklung Ziel der vorliegenden Masterarbeit ist, soll so vorgebeugt werden.

3.2.1 Soziodemografische, berufs- und produktbezogene Variablen.

Den ersten Teil des Studienfragebogens bildeten Fragen dazu, mit Produkten welcher Hersteller bereits gearbeitet wurde und mit welchen Produktkategorien (Fluoroskopie-, CT-, Angiographie-Geräte, etc.) die Teilnehmenden Erfahrung hatten. Es folgte dann die Frage, für welche Produktkategorie der/die Teilnehmende rekrutiert worden war. Die Business Lines bei Siemens Healthineers sind, wie oben beschrieben, in Produktkategorien untergliedert, die als Antwortoptionen auf diese Frage präsentiert wurden. Die Teilnehmenden durften frei wählen, für welche Produktkategorie der Business Line, für die sie rekrutiert worden waren, sie den Studienfragebogen ausfüllen wollten. Personen, die für die Business Line XP rekrutiert worden waren, durften sich beispielweise aussuchen, ob sie den Studienfragebogen bezogen auf Fluoroskopie-, Mammographie-, Urologie- oder Radiographie-/Röntgensysteme beantworten wollten. Die Mail, in welcher der Link zur Studie verschickt wurde, beinhaltete, für welche Business Line der/die Teilnehmende rekrutiert wurde und zwischen welchen Produktkategorien er/sie demnach wählen durfte. Unter manche Business Lines fiel nur eine Produktkategorie, hier konnten die Teilnehmenden den Studienfragebogen entsprechend nur mit Bezug auf diese Produktkategorie beantworten (s. Kapitel 1.4 – User Experience bei Siemens Healthineers). Die anschließenden Fragen wurden an die ausgewählte Produktkategorie angepasst. Es wurde abgefragt, wie viele verschiedene Geräte der Produktkategorie bisher bedient worden waren und welches konkrete Modell zum Zeitpunkt der Studie am häufigsten genutzt wurde. Wenn das exakte Modell unbekannt war, konnte hier alternativ eine Beschreibung des Modells eingegeben werden. Der Hersteller und das Anschaffungsjahr des

Referenzprodukts, die Häufigkeit seiner Nutzung und wie zufrieden der/die Befragte mit der Einarbeitung in das Gerät war (-3 = *überhaupt nicht zufrieden* bis +3 = *äußerst zufrieden)*, wurden ebenfalls erhoben.

Am Ende der Studie wurden das Geschlecht, Alter und die Deutschkenntnisse der Teilnehmenden abgefragt. Als berufsbezogene Variablen wurden außerdem die Berufsgruppe, die Berufserfahrung in Jahren, zusätzliche medizinische Weiterbildungen, die Arbeitsstunden pro Woche und die Art der Institution (Universitätsklinikum, Praxis, etc.) erfasst. Es wurde in diesem Kontext auch die Anzahl der Betten im Krankenhaus (sofern in einem gearbeitet wurde) sowie die Finanzierung der Institution (staatlich oder privat) abgefragt. Zusätzlich wurden die Abteilungen beziehungsweise Fachrichtungen der Teilnehmenden sowie der Fokus ihrer Abteilung (Diagnostik, Intervention oder beides) erfasst. Alle Fragen, bis auf das Alter, das Herstellungsjahr und das konkrete Modell wurden durch vorgegebene Antwortkategorien abgefragt, welche zum Teil bereits in der Beschreibung der Stichprobe aufgeführt wurden (3.1.3) oder im Ergebnisteil (3.5.2 – Deskriptive Statistik) ersichtlich sind. Die Antwortkategorien wurden bis auf jene zur Frage nach der Anzahl bisher bedienter Produkte von Müßig (2022) übernommen. Einige der aufgeführten Variablen waren für die Hypothesen nicht von Bedeutung, wurden aber als mögliche Einflussgrößen auf die interessierenden Variablen beziehungsweise für eine präzise Beschreibung der Stichprobe mit erhoben.

3.2.2 Variablen der Replikation.

Neben den Skalen des UEQ+ (Schrepp & Thomaschewski, 2019a) wurden, wie in der Vorstudie (Müßig, 2022), die durch Literaturrecherche ermittelten Zusatzskalen zur Bewertung ihrer Wichtigkeit vorgelegt. Die Skalenbeschreibungen wurden aus der Vorstudie übernommen. Für die Skalen des UEQ+ lieferte die Webseite des UEQ+ bereits Beschreibungen, die auch Müßig verwendete (Schrepp & Thomaschewski, 2019d). Die Beschreibungen für die hinzugefügten Skalen waren von Müßig aus der Literatur abgeleitet worden. Eine Übersicht über alle Skalen mit ihren Beschreibungen kann Tabelle 8 entnommen werden. Drei Skalen des UEQ+ zur Sprachsteuerung wurden in der Vorstudie nur jenen Teilnehmenden präsentiert, die zuvor angegeben hatten, dass ihr Gerät über eine Sprachassistenz verfügt. Eine Sprachsteuerung ist bisher allerdings nur an sehr wenigen Produkten implementiert, weshalb die zugehörigen Skalen für die vorliegende Studie nicht abgefragt wurden. Für die Skalen Haptik und Akustik hatten Proband*innen die Möglichkeit „passt nicht zum Produkt“ anzukreuzen. Die Beschreibung der Skala Vertrauen wurde nach interner Diskussion geändert. Die ursprüngliche Beschreibung lautete

„Meine eingegebenen Daten sind in sicheren Händen. Die Daten werden nicht missbraucht, um mich zu schädigen“. Die Beschreibung bezog sich somit auf Daten der Anwender*innen. Im Fall von Medizinprodukten ist jedoch wahrscheinlicher, dass (auch) an Patientendaten gedacht wird. Aus diesem Grund wurde die Beschreibung neutraler formuliert, sodass verschiedene Daten gemeint sein konnten (s. Tabelle 8). Im Anschluss wurde erfragt, ob im Kontext der Frage an Patient*innendaten, eigene Daten (von medizinischem Personal) oder allgemein alle Daten gedacht wurde.

Auf einer siebenstufigen Skala (von -3 = *total unwichtig* bis +3 = *total wichtig*) sollten die Teilnehmenden einschätzen, wie wichtig die einzelnen Eigenschaften für das Nutzungserleben von Medizinprodukten sind. Die Proband*innen sollten bei der Beantwortung an das konkrete Modell (Referenzprodukt) denken, welches sie zum Zeitpunkt der Studie am häufigsten nutzten. In der Vorstudie wurde darüber hinaus abgefragt, wie gut die jeweilige Eigenschaft im Hinblick auf das Produkt bereits umgesetzt ist. Diese Fragen wurden in der vorliegenden Studie weggelassen, da anhand der Items zur Fragebogenentwicklung ohnehin ermittelt wurde, wie gut die Produkte abschneiden und sich sonst einige Fragen gedoppelt hätten. Zuletzt wurde den Proband*innen in der Vorstudie eine Liste aller Skalen vorgelegt, die sie im Hinblick auf ihre Wichtigkeit auf der Ratingskala mit +2 oder +3 bewertet hatten. Ihre Aufgabe war es, die wichtigsten Eigenschaften über Drag & Drop in eine absteigende Rangfolge nach ihrer Wichtigkeit zu bringen. Zudem wurde ein offenes Feld präsentiert, welches zusätzliche Angaben zu weiteren relevanten UX-Aspekten ermöglichte. Dieser Teil wurde für die neue Studie ebenfalls gekürzt, da für die Analysen die abschließende Rangfolge nicht von Interesse war, sondern von Müßig (2022) nur die Ratings auf den siebenstufigen Skalen für die Berechnung der MANOVA (Hypothese 2) verwendet wurden. Eine weitere Abweichung zu Müßigs Vorstudie bestand außerdem darin, dass Proband*innen der Vorstudie nicht wie in der vorliegenden Studie gezielt für eine bestimmte Business Line rekrutiert wurden. Stattdessen sollten sie die Fragen für eine Produktkategorie einer beliebigen Business Line beantworten, mit der sie sich während ihrer Arbeit am meisten beschäftigten.

Tabelle 8

Für die Replikation übernommene Skalen und ihre Beschreibung (Müßig, 2022)

Skala	Beschreibung
Effizienz	Ich kann meine Ziele mit minimalem zeitlichem und physischem Aufwand erreichen. Das Produkt reagiert schnell auf meine Eingaben.
Originalität	Das Design ist kreativ und weckt das Interesse.
Steuerbarkeit	Das Produkt reagiert immer vorhersehbar und konsistent auf meine Eingaben. Ich habe stets die volle Kontrolle über die Interaktion.
Durchschaubarkeit	Es ist einfach, sich mit dem Produkt vertraut zu machen und zu lernen, wie es zu benutzen ist.
Stimulation	Ich finde das Produkt anregend und spannend. Es macht Spaß, sich damit zu beschäftigen.
Anpassungsfähigkeit	Ich kann das Produkt an meine persönlichen Vorlieben bzw. meinen persönlichen Arbeitsstil anpassen. Vorhandene Anpassungsmöglichkeiten sind einfach zu finden und zu verstehen.
Intuitive Bedienung	Ich kann das Produkt unmittelbar und ohne jegliche Einarbeitung oder Anleitung durch andere bedienen.
Wertigkeit	Das Produkt macht einen hochwertigen und professionellen Eindruck. Es strahlt eine gewisse Exklusivität aus.
Inhaltsseriosität	Die Informationen, die das Produkt bietet, sind von guter Qualität und zuverlässig.
Attraktivität	Der Gesamteindruck des Produkts.
Inhaltsqualität	Die Informationen, die das Produkt bietet, sind aktuell und gut aufbereitet.
Nützlichkeit	Die Benutzung des Produkts bringt mir Vorteile. Es spart mir Zeit und Mühe und macht mich produktiver.
Haptik	Gefühle, die beim Berühren des Produkts entstehen.
Erwartungskonformität	Das System verhält sich, wie ich es erwarte.
Visuelle Ästhetik	Das Produkt ist schön und ansprechend gestaltet.

Skala	Beschreibung
Akustik	Auswirkungen von Geräuschen oder Betriebsgeräuschen des Produkts auf das Nutzererlebnis.
Übersichtlichkeit	Eindruck von Ordnung, Struktur und visueller Komplexität einer grafischen Benutzeroberfläche.
Vertrauen [a]	Eingegebene Daten sind in sicheren Händen. Die Daten werden nicht missbraucht, um jemanden zu schädigen.
Zusatzskalen	
Erlernbarkeit	Das System unterstützt mich beim Lernen.
Komplexität	Das System zeigt mir nur relevante und häufig gebrauchte Informationen und keine Redundanzen.
Konsistenz	Das System ist einheitlich gestaltet und konform mit bereits bestehenden Konzepten.
Effektivität	Das System unterstützt mich beim Erledigen meiner Aufgaben.
Sympathiewert	Die Marke des Produkts strahlt Wertigkeit und ein Qualitätsversprechen aus.
Innovationen	Das System ist innovativ und bringt z.B. hilfreiche Funktionen bei der Bildauswertung oder bei der Positionierung der Patienten.
Sicherheit	Das Produkt erzeugt ein körperliches Sicherheitsgefühl bei der Nutzung.

Anmerkung. Für die Skalen des UEQ+ (alle Skalen außer die Zusatzskalen) lieferte die Webseite des UEQ+ bereits Beschreibungen, die auch Müßig (2022) heranzog (Schrepp & Thomaschewski, 2019d).

[a] Die Beschreibung der Skala wurde aufgrund von möglichen Missverständnissen geändert. Die ursprüngliche Beschreibung lautete: *Meine* eingegebenen Daten sind in sicheren Händen. Die Daten werden nicht missbraucht, um *mich* zu schädigen.

3.2.3 Items zur Fragebogenentwicklung.

Für die, laut der externen Vorstudie, sechs produktübergreifend relevantesten Skalen zur Evaluation der UX von Medizinprodukten wurden die zugehörigen Items abgefragt. Auch bei ihnen sollte auf ein konkretes Referenzmodell Bezug genommen werden. Die Items der vier fertigen Skalen des UEQ+ (Steuerbarkeit, Übersichtlichkeit, Effizienz und Vertrauen) wurden aus einer Vorlage übernommen, die auf der Webseite des UEQ+ von Schrepp und Thomaschewski (2019d) zur Verfügung gestellt wird. Die Skalen sind in den Abbildungen 3 bis 6 ersichtlich. Der Einleitungssatz der Skala Vertrauen wurde genauso wie die Skalenbeschreibung geändert und neutraler formuliert. Für die Skalen Sicherheit und Effektivität wurden die aus den Workshops abgeleiteten Items im Format des UEQ+ präsentiert, welche die Abbildungen 7 und 8 zeigen. Die siebenstufigen Skalen reichten auch hier von -3 (negativer Pol) bis +3 (positiver Pol), die Enden wurden mit den gegensätzlichen Begriffen beschriftet.

Nach der Präsentation der Skalen wurden zwei weitere Items zur Validierung abgefragt. Es wurde zum einen die Gesamtzufriedenheit mit dem Produkt erfragt („Insgesamt bin ich mit der Nutzerfreundlichkeit des Produkts…“ -3 = *sehr unzufrieden* bis +3 = *sehr zufrieden)*, zum anderen, angelehnt an den Net Promoter Score (Reichheld, 2003), wie wahrscheinlich eine Weiterempfehlung an Kolleg*innen ist („Wie wahrscheinlich ist es, dass Sie das Produkt einem Kollegen/einer Kollegin weiterempfehlen?“ -3 = *sehr unwahrscheinlich* bis +3 = *sehr wahrscheinlich*). Die Bewertung erfolgte, wie die der anderen Items, auf einer siebenstufigen Skala. Um eine Äquidistanz der Abstufungen nahezulegen, wurden Zahlen von -3 bis +3 über den Ratingstufen aller Items hinzugefügt. Es wurde außerdem bei der Präsentation der neuen Items darauf hingewiesen, dass Zeilen mit unverständlichen oder unpassenden Begriffen leer gelassen werden konnten. In Kapitel 4.3 (Abweichungen von der Präregistrierung) wird dieses Vorgehen begründet.

Abbildung 3

Skala Steuerbarkeit des UEQ+ (Schrepp & Thomaschewski, 2020)

Steuerbarkeit

Die Reaktion des Produkts auf meine Eingaben und Befehle empfinde ich als								
	-3	-2	-1	0	1	2	3	
unberechenbar	○	○	○	○	○	○	○	vorhersagbar
behindernd	○	○	○	○	○	○	○	unterstützend
unsicher	○	○	○	○	○	○	○	sicher
nicht erwartungskonform	○	○	○	○	○	○	○	erwartungskonform

Abbildung 4

Skala Vertrauen des UEQ+ (Schrepp & Thomaschewski, 2020)

Vertrauen

In Bezug auf die Verwendung persönlicher Informationen und Daten ist das Produkt [a]								
	-3	-2	-1	0	1	2	3	
unsicher	○	○	○	○	○	○	○	sicher
unseriös	○	○	○	○	○	○	○	seriös
unzuverlässig	○	○	○	○	○	○	○	zuverlässig
intransparent	○	○	○	○	○	○	○	transparent

[a] Der Einleitungssatz wurde verändert, indem das Wort „meine“ vor „persönliche Daten“ entfernt wurde, um eine Passung auf Medizinprodukte zu gewährleisten, bei denen insbesondere auch Patient*innendaten geschützt werden sollen.

Abbildung 5

Skala Effizienz des UEQ+ (Schrepp & Thomaschewski, 2020)

Effizienz

Für das Erreichen meiner Ziele empfinde ich das Produkt als								
	-3	-2	-1	0	1	2	3	
langsam	○	○	○	○	○	○	○	schnell
ineffizient	○	○	○	○	○	○	○	effizient
unpragmatisch	○	○	○	○	○	○	○	pragmatisch
überladen	○	○	○	○	○	○	○	aufgeräumt

Abbildung 6

Skala Übersichtlichkeit des UEQ+ (Schrepp & Thomaschewski, 2020)

Übersichtlichkeit

Die Benutzeroberfläche des Produkts empfinde ich als								
	-3	-2	-1	0	1	2	3	
schlecht gegliedert	○	○	○	○	○	○	○	gut gegliedert
unstrukturiert	○	○	○	○	○	○	○	strukturiert
ungeordnet	○	○	○	○	○	○	○	geordnet
unorganisiert	○	○	○	○	○	○	○	organisiert

Abbildung 7

Eigens entwickelte Skala Sicherheit im Format des UEQ+

Sicherheit

Anwendungsfehler und Risiken, die bei der Nutzung des Produkts entstehen können, empfinde ich als								
	-3	-2	-1	0	1	2	3	
wahrscheinlich	○	○	○	○	○	○	○	unwahrscheinlich
nicht rechtzeitig rückgemeldet	○	○	○	○	○	○	○	rechtzeitig rückgemeldet
schwer verständlich angezeigt	○	○	○	○	○	○	○	leicht verständlich angezeigt
schwer erkennbar	○	○	○	○	○	○	○	leicht erkennbar
begünstigt	○	○	○	○	○	○	○	unterbunden
schwer behebbar	○	○	○	○	○	○	○	leicht behebbar
nicht vom System kontrolliert	○	○	○	○	○	○	○	vom System kontrolliert
durch fehlende Informationen bedingt	○	○	○	○	○	○	○	nicht durch fehlende Informationen bedingt
unaufhaltbar	○	○	○	○	○	○	○	aufhaltbar
nicht gekennzeichnet	○	○	○	○	○	○	○	gekennzeichnet
ungesichert	○	○	○	○	○	○	○	abgesichert
unveränderbar	○	○	○	○	○	○	○	beeinflussbar
schwerwiegend	○	○	○	○	○	○	○	unerheblich
Hardware-Items (gleicher Einleitungssatz)								
bedrohlich	○	○	○	○	○	○	○	harmlos
gesundheitsgefährdend	○	○	○	○	○	○	○	nicht gesundheitsgefährdend
kollisionsbegünstigend	○	○	○	○	○	○	○	nicht kollisionsbegünstigend
schädigend	○	○	○	○	○	○	○	nicht schädigend

durch die Verarbeitung des Produkts begünstigt	○	○	○	○	○	○	○	durch die Verarbeitung des Produkts unwahrscheinlich
durch Systembewegungen möglich	○	○	○	○	○	○	○	nicht durch Systembewegungen möglich
durch umständliche Reinigung begünstigt	○	○	○	○	○	○	○	durch einfache Reinigung vermieden

Abbildung 8

Eigens entwickelte Skala Effektivität im Format des UEQ+

Effektivität

Die Ergebnisse, die sich mit dem Produkt/System erzielen lassen, empfinde ich als								
	-3	-2	-1	0	1	2	3	
unzweckmäßig	○	○	○	○	○	○	○	zweckmäßig
nutzlos	○	○	○	○	○	○	○	nützlich
für meine Anforderungen unpassend	○	○	○	○	○	○	○	auf meine Anforderungen abgestimmt
nicht zielerfüllend	○	○	○	○	○	○	○	zielerfüllend
unvollständig	○	○	○	○	○	○	○	vollständig
ungenau	○	○	○	○	○	○	○	genau
uneindeutig	○	○	○	○	○	○	○	eindeutig
inadäquat	○	○	○	○	○	○	○	adäquat
entscheidungshinderlich	○	○	○	○	○	○	○	entscheidungsunterstützend
minderwertig	○	○	○	○	○	○	○	hochwertig
nicht bedarfsgerecht	○	○	○	○	○	○	○	bedarfsgerecht
wechselhaft	○	○	○	○	○	○	○	konstant
unterdurchschnittlich	○	○	○	○	○	○	○	überdurchschnittlich
untauglich	○	○	○	○	○	○	○	praxistauglich

3.3 Prozedur

Die Online-Studie wurde über den Anbieter QuestionPro (QuestionPro, 2022) erstellt und im Zeitraum zwischen dem 27. Juli und dem 14. Oktober 2022 zur Teilnahme freigeschaltet. Die Durchführung nahm im Durchschnitt 19 Minuten in Anspruch. Der Zugang zur Studie wurde über einen Link ermöglicht, der per Mail versendet wurde. Vorab erhielten die Teilnehmenden Angaben zur Thematik der Studie. Sie wurden informiert, dass der Zweck die Entwicklung eines UX-Fragebogens war und dass die Studie von Siemens Healthineers in Auftrag gegeben wurde. Vor Beginn der eigentlichen Studie musste den Teilnahmebedingungen zugestimmt werden. Für jene Proband*innen, die über die Agentur rekrutiert wurden, gab es ein Feld zur Eingabe einer Identifikationsnummer, um die Bezahlung abzuwickeln. Um die Bezahlung der Proband*innen abzuwickeln, die über das Collaboration Management von Siemens Healthineers rekrutiert wurden, sollte der Name der Institution, in der die Proband*innen arbeiteten, angegeben werden. Es wurde erfragt, ob die teilnehmende Person bei Siemens Healthineers arbeitete, falls ja wurde die Studie abgebrochen. Zu Beginn der eigentlichen Umfrage wurden oben beschriebene Informationen zur Produktkategorie und dem Referenzprodukt eingeholt. In einem ersten Abschnitt wurden anschließend die Fragen der Replikationsstudie vorgegeben, nachdem das Konzept der User Experience vorgestellt und die Instruktion eingeblendet wurden. Es folgten in einem zweiten Abschnitt die Items der sechs relevantesten Skalen für Medizinprodukte laut Müßigs Vorstudie, darunter die neu entwickelten Sicherheits- und Effektivitätsitems. Die Items sollten mit Bezug auf das angegebene Referenzprodukt beantwortet werden. Danach wurden die beiden Items zur Validierung abgefragt. Soziodemografische und berufsbezogene Variablen wurden am Ende der Umfrage platziert. In einem letzten Abschnitt wurde ein Text zur Danksagung eingeblendet und eine E-Mail-Adresse für Interesse an weiteren Studien angegeben. Nach der Teilnahme erhielten die Teilnehmenden eine finanzielle Entschädigung per Überweisung.

3.4 Studiendesign

Es handelte sich um eine nicht-experimentelle Online-Studie mit einem Erhebungszeitpunkt (Querschnitt). Manche Proband*innen waren mit mehreren Produktkategorien verschiedener Business Lines vertraut. Sie wurden den Business Lines jedoch nicht randomisiert zugewiesen, sondern gezielt für diese rekrutiert, um eine ausreichend hohe Teilnehmerzahl für jede Business Line sicherzustellen. Die Fragen zur Bewertung der Wichtigkeit der Skalen sowie die Items der Skalen des UEQ+ und der neuen Skalen Sicherheit und Effektivität wurden randomisiert präsentiert. Aufgrund des Between-

Subject-Designs lagen unabhängige Messwerte vor. Die Teilnehmenden konnten den Fragebogen nur einmalig, für eine Produktkategorie einer Business Line ausfüllen.

3.5 Ergebnisse

Die Datenanalyse wurde mittels der open source Statistik-Software R (The R Foundation, 2022), Version 4.2.1 aus dem Jahr 2022 vorgenommen.

3.5.1 Vorbereitung der Daten, Umgang mit fehlenden Werten und Ausreißern.

Die siebenstufigen Skalen von -3 bis +3 wurden für die Berechnungen zu Werten von 1 bis 7 kodiert. Die Mitte der Skala entspricht somit dem kodierten Wert 4. Die einzelnen Items der Skalen des UEQ+ (Schrepp & Thomaschewski, 2019a) beziehungsweise später auch die vier geeignetsten Items der Zusatzskalen wurden für jede Skala gemittelt, sodass sich ein Skalenwert pro Person ergab. Die Produktkategorien wurden entsprechend der Business Lines bei Siemens Healthineers zusammengefasst. So wurden zum Beispiel CT-, PETCT- und SPECT-Geräte zur übergeordneten Kategorie CT zusammengefasst (s. Kapitel 1.4 – User Experience bei Siemens Healthineers).

Wann immer möglich wurde paarweiser Ausschluss fehlender Werte angewendet. Für die Berechnungen der multivariaten Varianzanalysen (MANOVAs) war jedoch nur die Anwendung fallweisen Ausschlusses möglich. Die meisten der für die Hypothesen relevanten Fragen waren ohnehin als Pflichtfelder eingestellt, weshalb sie nicht ausgelassen werden konnten. Die Mehrheit der soziodemografischen, berufs- und produktbezogenen Fragen war dagegen optional. Bei den Replikationsitems zur Bewertung der Wichtigkeit der Skalen waren alle Items als verpflichtend eingestellt. Ausnahmen bildeten nur die Skalen Haptik und Akustik, da davon ausgegangen wurde, dass diese nicht auf alle Referenzprodukte anwendbar waren. Für die Skala Haptik gaben 15 Teilnehmende diesen Aspekt als nicht zu ihrem Referenzprodukt passend an, für die Skala Akustik 14 Teilnehmende. Eine weitere Ausnahme bildeten die neu entwickelten Items für die Zusatzskalen. Hier wurden Proband*innen angewiesen, unverständliche Items auszulassen. Dieses Vorgehen wird in Kapitel 4.4 (Abweichungen von der Präregistrierung) begründet. Personen, welche die Umfrage nicht bis zum Ende beantworteten, wurden ausgeschlossen, sodass keine Werte bei den neuen Items fehlten, weil Personen die Umfrage abgebrochen hatten. Falls Antworten auf Items der Skala Sicherheit oder Effektivität fehlten, wurde zur Bildung von Skalenwerten der Durchschnitt aus den verbliebenen Items berechnet. Die Items der Skalen des UEQ+ (Steuerbarkeit, Effizienz, Vertrauen, Übersichtlichkeit) wurden als Pflichtfelder eingestellt.

Es wurde geprüft, ob Ausreißer für die Variablen Alter und Herstellungsjahr des Referenzprodukts gegeben waren. Es handelte sich hierbei um die einzigen beiden Variablen, bei denen freie Angaben möglich waren. Ausreißer wurden auf Basis von Box-Plots identifiziert, wobei als Grenzen für die Whisker das erste Quartil abzüglich des 1.5-fachen Interquartilsabstands (Q1 - 1.5 x IQR) beziehungsweise das dritte Quartil zuzüglich des 1.5-fachen Interquartilsabstands (Q3 + 1.5 x IQR) betrachtet wurden. Auf diese Weise identifizierte Ausreißer wurden anschließend auf inhaltliche Plausibilität geprüft. Zwei der Teilnehmenden wurden für die Berechnung des mittleren Herstellungsjahres ausgeschlossen, da sie hier unplausible Werte eingegeben hatten. Es wurde für die Produktkategorien Radiographie/Röntgen (Herstellungsjahr von 0) und Bestrahlungssysteme (Herstellungsjahr von 1900) je eine Person ausgeschlossen. Für alle folgenden Analysen wurde das α-Fehler-Niveau auf 5 % festgelegt.

3.5.2 Deskriptive Statistik.

Angaben bezüglich einiger berufsbezogener sowie soziodemografischer Variablen wurden bereits in Kapitel 3.1.3 (Beschreibung der Stichprobe) berichtet. Es werden an dieser Stelle nun Angaben zu produktbezogenen Variablen dargelegt. Die Teilnehmenden gaben durchschnittlich an, bereits mit Produkten von 2.6 verschiedenen Herstellern gearbeitet zu haben (*SD* = 1.1, *Mdn* = 3). Es werden aus Gründen der Vertraulichkeit keine konkreten Namen der Hersteller berichtet. Im Durchschnitt hatten die Teilnehmenden Erfahrung mit 3.6 verschiedenen Produktkategorien (*SD* = 1.1, *Mdn* = 4). Besonders häufig wurde Erfahrung mit CT-Geräten (*n* = 129), Röntgen-Geräten (*n* = 123) und MRT-Geräten (*n* = 118) genannt. Mit Angiographie-Systemen hatten 96 der Befragten Erfahrung, mit Fluoroskopie-Systemen 78 und mit C-Bögen 74 der Befragten. Mit den anderen Produktkategorien hatte ein geringerer Anteil der Teilnehmenden Erfahrung ($n_{\text{Mammographie\_Erfahrung}} = 54$, $n_{\text{SPECT/PETCT\_Erfahrung}} = 38$, $n_{\text{Bestrahlung\_Erfahrung}} = 27$, $n_{\text{Urologie\_Erfahrung}} = 12$).

Alle weiteren Fragen der Studie bezogen sich auf das Referenzprodukt. Es handelte sich hierbei um ein konkretes Modell, welches der/die Teilnehmende am häufigsten nutzte. Im Mittel wurden die Referenzprodukte über alle Produktkategorien hinweg im Jahr 2016 (*M* = 2016.10, *SD* = 4.84) hergestellt. Das Herstellungsjahr war nicht als verpflichtendes Feld eingestellt, weshalb hier insgesamt nur 140 Teilnehmende in die Berechnung einbezogen werden konnten. Für die Frage, wie viele verschiedene Produkte der jeweiligen Produktkategorie (z.B. CT-Geräte) der/die Befragte bereits bedient habe, waren die Antwortkategorien 1) *1*, 2) *2 bis 5*, 3) *6 bis 10*, 4) *mehr als 10* und 5) *keine*

Angabe vorgegeben. Befragte aller Produktkategorien hatten bisher am häufigsten mit zwei bis fünf Produkten gearbeitet. Es wurde außerdem erfragt, wie häufig das Referenzprodukt genutzt werde. Vorgegeben waren die von Müßig (2022) übernommenen, allerdings nicht gleichabständigen Antwortkategorien 1) *mehrmals am Tag*, 2) *einmal täglich*, 3) *mehrmals in der Woche*, 4) *einmal die Woche*, 5) *einmal im Monat*, 6) *seltener*. Für alle Produktkategorien wurde am häufigsten die Antwortoption *mehrmals am Tag* ausgewählt. Ausnahmen bildeten die Produktkategorien Fluoroskopie und SPECT/PETCT, für die jeweils nur zwei Personen teilnahmen. In beiden Produktkategorien gab je eine Person an, das Referenzprodukt mehrmals am Tag zu nutzen, die andere gab einmal in der Woche (Fluoroskopie) beziehungswiese mehrmals in der Woche (SPECT/PETCT) an. Produktübergreifend gaben 122 der 173 Teilnehmenden an, das Referenzprodukt mehrmals täglich zu nutzen. Bezüglich der Zufriedenheit mit der Einarbeitung in das Referenzprodukt, die ebenfalls auf einer siebenstufigen Skala bewertet wurde, betrug der Durchschnitt 5.3 (*SD* = 1.5). Der Median belief sich auf den Wert 6.

Die deskriptive Statistik bezüglich der Bewertung der Wichtigkeit der Skalen aus der vorliegenden Studie zeigt untenstehende Tabelle 9. Auf die Frage, an welche Art von Daten bei der Skala Vertrauen gedacht wurde, gaben 69.36 % der Personen an, an Patient*innendaten gedacht zu haben und 27.17 % verbanden mit Daten in diesem Kontext alle Arten von Daten. Nur 2.31 % der Befragten dachten an eigene Daten (Daten von medizinischem Personal). Eine Person dachte außerdem an andere Daten und eine weitere Person bezog zu dieser Frage keine Stellung.

Die Mittelwerte, Standardabweichungen, und Mediane für die einzelnen Items der abgefragten Skalen des UEQ+ (Schrepp & Thomaschewski, 2019a) Steuerbarkeit, Vertrauen, Effizienz und Übersichtlichkeit sind in Tabelle 10 abgetragen. Für eine bessere Übersichtlichkeit wurden die Items der neuen Skalen Sicherheit und Effektivität in separaten Tabellen (Tabelle 11 und 12) dargestellt. Im Hinblick auf die beiden Validierungsitems zur Erfassung der Gesamtzufriedenheit zeigte sich für das Item „Insgesamt bin ich mit der Nutzerfreundlichkeit des Produkts…“ (*sehr unzufrieden* bis *sehr zufrieden*) ein Mittelwert von 5.4 (*SD* = 1.5, *Mdn* = 6). Für das Item „Wie wahrscheinlich ist es, dass Sie das Produkt einem Kollegen/einer Kollegin weiterempfehlen?“ (*sehr unwahrscheinlich* bis *sehr wahrscheinlich*) ergab sich ebenfalls ein Mittelwert von 5.4 (*SD* = 1.7, *Mdn* = 6). Für alle Items, sowohl die bereits existierenden Items des UEQ+ als auch die neu entwickelten und die Validierungsitems, umfasste der Range alle Werte der Skala von 1 bis 7 (bzw. -3 bis +3).

Tabelle 9

Deskriptive Statistik bezüglich der produktübergreifenden Bewertung der Wichtigkeit der Skalen

Skala	*M*	*SD*	*Mdn*
Steuerbarkeit	6.60	0.69	7
Effizienz	6.57	0.65	7
Nützlichkeit	6.50	0.90	7
Übersichtlichkeit	6.49	0.75	7
Effektivität	6.41	0.82	7
Vertrauen	6.40	0.97	7
Erwartungskonformität	6.36	0.87	7
Durchschaubarkeit	6.30	0.92	7
Inhaltsseriosität	6.23	0.95	6
Sicherheit	6.06	1.19	6
Inhaltsqualität	6.06	0.94	6
Innovationen	6.05	1.00	6
Erlernbarkeit	5.76	1.24	6
Konsistenz	5.73	1.23	6
Intuitive Bedienung	5.51	1.50	6
Wertigkeit	5.50	1.27	6
Anpassungsfähigkeit	5.38	1.40	6
Komplexität	5.32	1.29	6
Attraktivität	5.25	1.24	5
Sympathiewert	5.15	1.49	5
Akustik	5.12	1.63	5
Stimulation	5.05	1.39	5
Visuelle Ästhetik	4.85	1.51	5
Haptik	4.52	1.63	5
Originalität	4.27	1.57	4

Anmerkung. $N = 173$. $n_{Haptik} = 158$. $n_{Akustik} = 159$. Die Skalen sind absteigend nach ihrer Wichtigkeit sortiert. Auf einer siebenstufigen Skala (von -3 = *total unwichtig* bis +3 = *total wichtig*, kodiert von 1 bis 7) sollten die Teilnehmenden einschätzen, wie wichtig die einzelnen Eigenschaften für das Nutzungserleben von Medizinprodukten sind. Die Werte wurden statt auf eine

Nachkommastelle, wie bei Integer-Skalen laut APA vorgesehen (American Psychological Association, 2022), auf zwei Nachkommastellen gerundet, um eine eindeutige Rangfolge darzustellen.

Tabelle 10

Deskriptive Statistiken für die Items der relevanten Skalen des UEQ+ (Schrepp & Thomaschewski, 2019a)

Skala	Item [a]	*M*	*SD*	*Mdn*
Steuerbarkeit	Vorhersagbar	5.4	1.4	6
	Unterstützend	5.1	1.6	5
	Sicher	5.5	1.6	6
	Erwartungskonform	5.4	1.6	6
Vertrauen	Sicher	5.8	1.4	6
	Seriös	6.0	1.4	6
	Zuverlässig	5.9	1.4	6
	Transparent	5.5	1.5	6
Effizienz	Schnell	4.8	1.8	5
	Effizient	5.3	1.7	6
	Pragmatisch	5.0	1.6	6
	Aufgeräumt	5.1	1.5	5
Übersichtlichkeit	Gut gegliedert	5.1	1.7	6
	Strukturiert	5.2	1.8	6
	Geordnet	5.2	1.7	6
	Organisiert	5.2	1.7	6

Anmerkung. *N* = 173. Die semantischen Differentiale wurden auf einer siebenstufigen Skala von -3 bis +3 (kodiert von 1 bis 7) bewertet.

[a] Es wird jeweils nur der positive Pol genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) eingesehen werden.

Tabelle 11

Deskriptive Statistik für die erhobenen Items der Zusatzskala Sicherheit

Item [a]	*M*	*SD*	*Mdn*	*NA* [b]
Allgemeine Items (auf Soft- und Hardwarekomponenten anwendbar)				
Unwahrscheinlich	4.5	1.4	5	11
Rechtzeitig rückgemeldet	4.6	1.6	5	13
Leicht verständlich angezeigt	4.6	1.6	5	8
Leicht erkennbar	4.7	1.5	5	5
Unterbunden	4.4	1.2	4	19
Leicht behebbar	4.5	1.5	5	5
Kontrolliert	4.7	1.6	5	13
Nicht durch fehlende Informationen bedingt	4.5	1.6	5	12
Aufhaltbar	4.8	1.5	5	12
Gekennzeichnet	4.6	1.5	5	10
Abgesichert	4.7	1.5	5	14
Beeinflussbar	5.0	1.5	5	11
Unerheblich	4.0	1.5	4	10
Hardware-Items				
Harmlos	4.2	1.6	4	4
Nicht gesundheitsgefährdend	4.2	1.8	4	5
Nicht kollisionsbegünstigend	4.4	1.7	4	12
Nicht schädigend	4.2	1.7	4	10
Nicht durch unbeabsichtigte Bewegungen möglich	4.2	1.6	4	13
Durch einfache Reinigung vermieden	4.3	1.5	4	14
Durch die Verarbeitung des Produkts unwahrscheinlich	4.6	1.7	5	9

Anmerkung. $N_{maximal}$ = 173 (ohne Berücksichtigung der fehlenden Werte, welche in der rechten Spalte stehen). $n_{Hardware\_maximal}$ = 154 (ohne Personen der Business Line D&A). Die semantischen Differentiale wurden auf einer siebenstufigen Skala von -3 bis +3 (kodiert von 1 bis 7) bewertet.
[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

[b] NA = not available; Anzahl der Werte, die aufgrund von bewusster Auslassung bei den Items fehlten.

Tabelle 12

Deskriptive Statistik für die erhobenen Items der Zusatzskala Effektivität

Item [a]	*M*	*SD*	*Mdn*	*NA* [b]
Zweckmäßig	5.8	1.4	6	4
Nützlich	5.9	1.3	6	2
Auf meine Anforderungen abgestimmt	5.3	1.4	6	5
Zielerfüllend	5.9	1.3	6	3
Vollständig	5.7	1.3	6	4
Genau	5.6	1.3	6	2
Eindeutig	5.7	1.3	6	4
Adäquat	5.6	1.4	6	5
Entscheidungsunterstützend	5.4	1.4	6	7
Hochwertig	5.5	1.4	6	1
Bedarfsgerecht	5.6	1.5	6	6
Konstant	5.4	1.5	6	2
Überdurchschnittlich	4.9	1.4	5	4
Praxistauglich	5.8	1.4	6	2

Anmerkung. $N_{maximal}$ = 173 (ohne Berücksichtigung der fehlenden Werte, welche in der rechten Spalte stehen). Die semantischen Differentiale wurden auf einer siebenstufigen Skala von -3 bis +3 (kodiert von 1 bis 7) bewertet.

[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

[b] NA = not available; Anzahl der Werte, die aufgrund von bewusster Auslassung bei den Items fehlten.

3.5.3 Ergebnisse der Replikation.

Die ersten beiden Hypothesen bezogen sich auf jene Items, die aus der Vorstudie übernommen wurden. Sie erfragten die Wichtigkeit der UX-Aspekte für Medizinprodukte. Die Ergebnisse zu den Hypothesen 1 und 2 werden in diesem Kapitel erläutert.

3.5.3.1 Hypothese 1. Die erste Hypothese postulierte, dass sich die gleichen sechs Skalen wie in der Vorstudie als produktübergreifend am wichtigsten erweisen. Für diese Hypothese wurde keine Berechnung vorgenommen. Stattdessen wurde die durchschnittliche Bewertung der Wichtigkeit der einzelnen Skalen über alle Proband*innen und Produkte hinweg mit den Ergebnissen der Vorstudie verglichen. Untenstehende Tabelle 13 zeigt die sechs produktübergreifend am wichtigsten bewerteten UX-Skalen der Vorstudie (Müßig, 2022) und der in dieser Arbeit beschriebenen Replikationsstudie im Vergleich.

Tabelle 13

Gegenüberstellung der sechs produktübergreifend wichtigsten Skalen aus der Vorstudie und der Replikationsstudie

Wichtigkeitsrang	Erste externe Studie 2021 (Müßig, 2022)	Zweite externe Studie 2022 (der vorliegenden Arbeit)
1	Sicherheit (M = 6.61, SD = 0.74)	Steuerbarkeit (M = 6.60, SD = 0.69)
2	Steuerbarkeit (M = 6.56, SD = 0.77)	Effizienz (M = 6.57, SD = 0.65)
3	Effizienz (M = 6.50, SD = 0.87)	Nützlichkeit (M = 6.50, SD = 0.90)
4	Effektivität (M = 6.45, SD = 0.78)	Übersichtlichkeit (M = 6.49, SD = 0.75)
5	Übersichtlichkeit (M = 6.39, SD = 0.96)	Effektivität (M = 6.41, SD = 0.82)
6	Vertrauen (M = 6.31, SD = 1.22)	Vertrauen (M = 6.40, SD = 0.97)

Anmerkung. $N_{\text{Studie 2021}}$ = 65-68. $N_{\text{Studie 2022}}$ = 173. Die Werte wurden statt auf eine Nachkommastelle, wie bei Integer-Skalen laut APA vorgesehen (American Psychological Association, 2022), auf zwei Nachkommastellen gerundet, um eine eindeutige Rangfolge darzustellen.

Wie Tabelle 13 zu entnehmen ist, überschnitt sich der Großteil der produktübergreifend wichtigsten Skalen, wenn sich auch die exakte Rangfolge unterschied. Die einzige Ausnahme bildete die Skala Nützlichkeit, welche in der Replikationsstudie als sehr wichtig bewertet wurde und die Skala Sicherheit aus der Rangfolge der sechs wichtigsten Skalen verdrängte. Hypothese 1 konnte damit nicht vollständig angenommen werden, die meisten Skalen stimmten jedoch wie postuliert überein.

3.5.3.2 Hypothese 2. Die Annahme der zweiten Hypothese bestand darin, dass die Business Lines (MR, CT, XP, AT/AT-SU, D&A und Strahlentherapie) sich nicht signifikant darin unterscheiden, welche UX-Skalen am wichtigsten sind und es demnach eine produktübergreifende Empfehlung des UX-Fragebogens gibt. Zur inferenzstatistischen Prüfung der Hypothese wurde eine einfaktorielle MANOVA mithilfe des „stats"-Pakets in R durchgeführt. Wie in Müßigs Arbeit (2022) wurden als abhängige Variablen die Bewertungen der Wichtigkeit der 25 UX-Skalen herangezogen. Als Faktorstufen der unabhängigen Variable dienten die sechs Business Lines.

Vor der eigentlichen Analyse wurden die Daten auf die Erfüllung der notwendigen Voraussetzungen überprüft. Die Unabhängigkeit der Residuen war gegeben, da Befragte den Fragebogen nur einmalig für eine Business Line ausfüllten. Die Business Lines waren unabhängig voneinander und nominalskaliert. Anhand von Histogrammen wurde die Normalverteilung der abhängigen Variablen unterteilt nach Business Lines überprüft. Ein Großteil der abhängigen Variablen zeigte sich als linksschief verteilt. Auch Mardia-Test und Q-Q-Plot sprachen gegen multivariate Normalverteilung. Die Ergebnisse zu diesem und einigen folgenden Voraussetzungstests sind in Tabelle D1 (Anhang D) dargestellt. Transformationen brachten keine wesentlich verbesserte Annäherung an die Normalverteilung und wurden daher nicht angewendet. Mithilfe von Streudiagrammen wurde überprüft, ob die Beziehungen zwischen den abhängigen Variablen für jede Gruppe der unabhängigen Variablen linear waren. Das eindeutige Vorliegen linearer Beziehungen war anzuzweifeln. Multikollinearität der Variablen ließ sich dagegen ausschließen. Der Levene-Test wurde bei den Wichtigkeitsbewertungen der Skalen Haptik, intuitive Bedienung, Komplexität und Sicherheit signifikant. Für alle anderen Variablen ließ sich Homoskedastizität annehmen. Der geplante Boxsche M-Test auf Homogenität der Varianz-Kovarianz Matrizen ließ sich nicht sinnvoll durchführen, da für manche Business Lines weniger Beobachtungen als abhängige Variablen vorlagen, was insgesamt eine Einschränkung für die Berechnung einer MANOVA darstellt. Aufgrund der mangelnden

Erfüllung der Voraussetzungen wurde entschieden neben einer MANOVA einen nicht-parametrischen multivariaten Kruskal-Wallis-Test zu berechnen.

Die MANOVA ergab einen signifikanten Unterschied zwischen den Business Lines für die kombinierten abhängigen Variablen, $F(5, 143) = 1.92$, $p < .001$, partielles $\eta^2 = .28$, Wilk's $\Lambda = 0.19$. Alle vier standardmäßig ausgegebenen Statistiken (Pillai-Spur, Wilks Lamda, Hotelling-Spur und die größte charakteristische Wurzel nach Roy) kamen zu dem gleichen Ergebnis, dass ein signifikanter Effekt vorlag. Mithilfe des R-Pakets „MNM" wurde, wie von Oja (2010) beschrieben, eine Erweiterung des Kruskal-Wallis-Tests mit mehreren abhängigen Variablen durchgeführt. Als Teststatistik wird Q^2 ausgegeben, was der Pillai-Spur entspricht. Auch dieser Test erwies sich als signifikant, $Q^2(125) = 208.33$, $p < .001$. Es konnte demnach gefolgert werden, dass sich die Wichtigkeitsbewertungen der Skalen zwischen den Business Lines signifikant unterschieden.

Es stellte sich indes die Frage, ob der gefundene Effekt jene Skalen betraf, welche in den UX-Fragebogen für Medizinprodukte aufgenommen werden sollten oder Skalen am unteren Ende der Rangfolge, welche ohnehin nicht in den UX-Fragebogen übernommen würden. Folglich wurden Post hoc Tests berechnet, um zu ermitteln, ob sich Unterschiede zwischen den Business Lines auf relevante Skalen bezogen und somit eine praktische Bedeutsamkeit hatten. Es wurden mithilfe des „stats"-Pakets zuerst einfaktorielle Varianzanalysen (ANOVAs) für jene Skalen als abhängige Variablen berechnet, die sich produktübergreifend oder für die einzelnen Business Lines unter den sechs wichtigsten Skalen befanden. Um der Kumulierung des α-Fehler-Niveaus entgegenzuwirken wurde eine Bonferroni-Holm-Korrektur (Holm, 1979) angewendet. Anhand der ANOVAs zeigten sich keine signifikanten Unterschiede zwischen den Business Lines in Bezug auf die mittlere Wichtigkeitsbewertung der relevanten Skalen, wie Tabelle 14 zeigt. In Abbildung 9 sind darüber hinaus die Mittelwerte und Standardabweichungen der sechs wichtigsten Skalen für jede Business Line und deren Abweichungen von den produktübergreifend wichtigsten Skalen abgetragen.

Tabelle 14

Post-hoc einfaktorielle Varianzanalysen zur Überprüfung von Unterschieden in der mittleren Wichtigkeitsbewertung der Skalen für die verschiedenen Business Lines

Wichtigkeitsbewertung der Skala	$F(5, 167)$	p_{korr} [c]	η^2
Produktübergreifend wichtig [a]			
Steuerbarkeit	2.10	.54	.06
Effizienz	1.16	> .999	.03
Nützlichkeit	1.57	> .999	.04
Übersichtlichkeit	1.16	.996	.03
Effektivität	0.99	.848	.03
Vertrauen	1.21	> .999	.04
Business Lines [b]			
Erwartungskonformität	3.96	.103	.08
Durchschaubarkeit	0.51	.771	.01
Inhaltsseriosität	1.49	> .999	.04

Anmerkung. Als Faktorstufen der unabhängigen Variable fungierten die sechs Business Lines (CT, MR, XP, D&A, AT/AT-SU und Strahlentherapie).

[a] Sechs produktübergreifend wichtigste Skalen.

[b] Skalen, die sich unter den sechs wichtigsten Skalen einzelner Business Lines befanden.

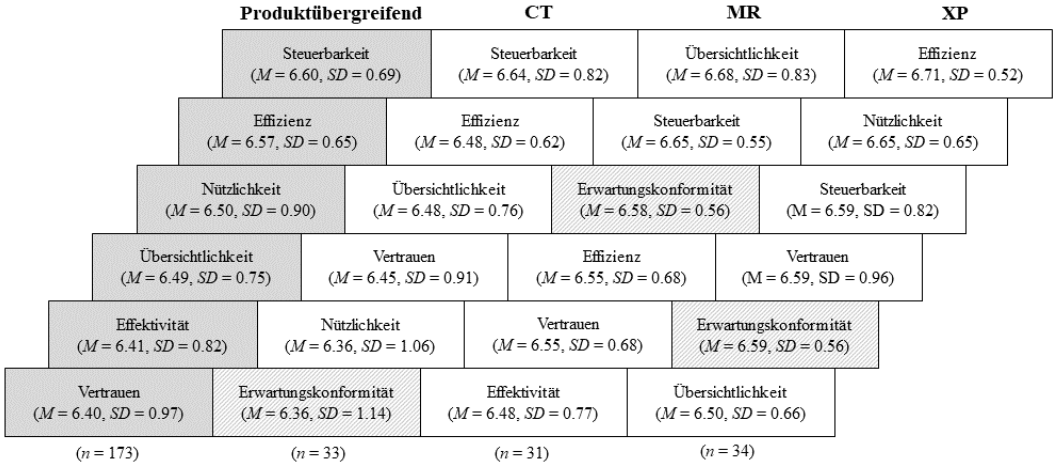
[c] Bonferroni-Holm-korrigierter p-Wert.

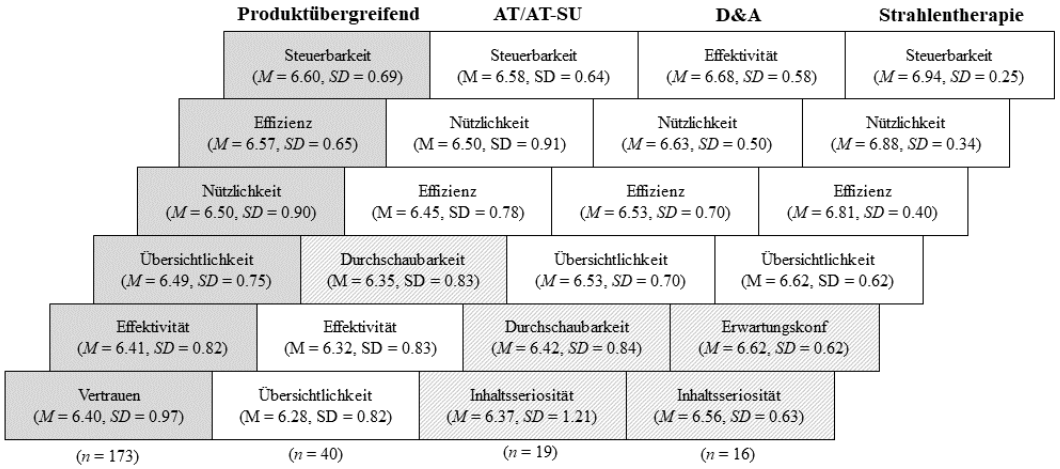
Abbildung 9

Vergleich der sechs wichtigsten Skalen der Business Lines mit den produktübergreifend wichtigsten Skalen (Ergebnisse der Replikationsstudie)

☐ Produktübergreifendes Wichtigkeitsranking der Skalen

☒ Von den sechs produktübergreifend wichtigsten Skalen abweichend





Anmerkung. Diese Abbildung ist einer Abbildung von Müßig (2022) nachempfunden und wurde entsprechend angepasst.

Wegen der nicht optimal erfüllten Voraussetzungen wurden, ebenfalls mit dem „stats“-Paket, zusätzlich nichtparametrische post-hoc Tests (Kruskal-Wallis-Tests) berechnet, deren Ergebnisse in Tabelle 15 zu sehen sind. Für die meisten Skalen zeigten sich bei korrigierter Überschreitungswahrscheinlichkeit auch hier keine signifikanten Unterschiede zwischen den Business Lines in Bezug auf die Wichtigkeitsbewertung der Skalen. Eine Ausnahme bildete die Skala Steuerbarkeit, für welche daraufhin paarweise Gruppenvergleiche mittels Wilcoxon-Tests durchgeführt wurden („rstatix“-Paket). Die Überschreitungswahrscheinlichkeiten wurden wie zuvor mithilfe der Bonferroni-Holm-Korrektur angepasst. Nur die Business Lines D&A ($Mdn = 6$) und Strahlentherapie ($Mdn = 7$) unterschieden sich signifikant in der Bewertung der Wichtigkeit der Skala Steuerbarkeit, $z = 56.5$, $p_{korr} = .004$, $r = .62$. Während Steuerbarkeit für die Business Line Strahlentherapie den Wichtigkeitsrang 1 belegte, belegte sie für die Business Line D&A Rang 8.

Es wurde neben den durchgeführten inferenzstatistischen Tests als grafisches Kriterium eine Gegenüberstellung aller Wichtigkeitsbewertungen der Business Lines erstellt (s. Abbildung 10). In Anhang D ist zusätzlich ein Vergleich der sechs wichtigsten Skalen für die einzelnen Business Lines von Müßigs Vorstudie aus dem Jahr 2021 (Müßig, 2022) und der Replikationsstudie abgebildet (Abbildung D1).

Tabelle 15

Nichtparametrische post-hoc Kruskal-Wallis-Tests zur Überprüfung von Unterschieden in der Wichtigkeitsbewertung der Skalen für die verschiedenen Business Lines

Wichtigkeitsbewertung der Skala	$\chi^2(5)$	p_{korr} [c]	Ordinales η^2
Produktübergreifend wichtig [a]			
Steuerbarkeit	16.63	.047*	.07
Effizienz	5.12	.803	< .01
Nützlichkeit	6.23	> .999	< .01
Übersichtlichkeit	7.95	.954	.02
Effektivität	5.61	> .999	.04
Vertrauen	5.57	> .999	< .01
Business Lines [b]			
Erwartungskonformität	15.50	.067	.06
Durchschaubarkeit	2.03	.845	< .01
Inhaltsseriosität	8.09	> .999	.02

Anmerkung. Als Faktorstufen der unabhängigen Variable fungierten die sechs Business Lines (CT, MR, XP, D&A, AT/AT-SU und Strahlentherapie).

[a] Sechs produktübergreifend wichtigste Skalen.

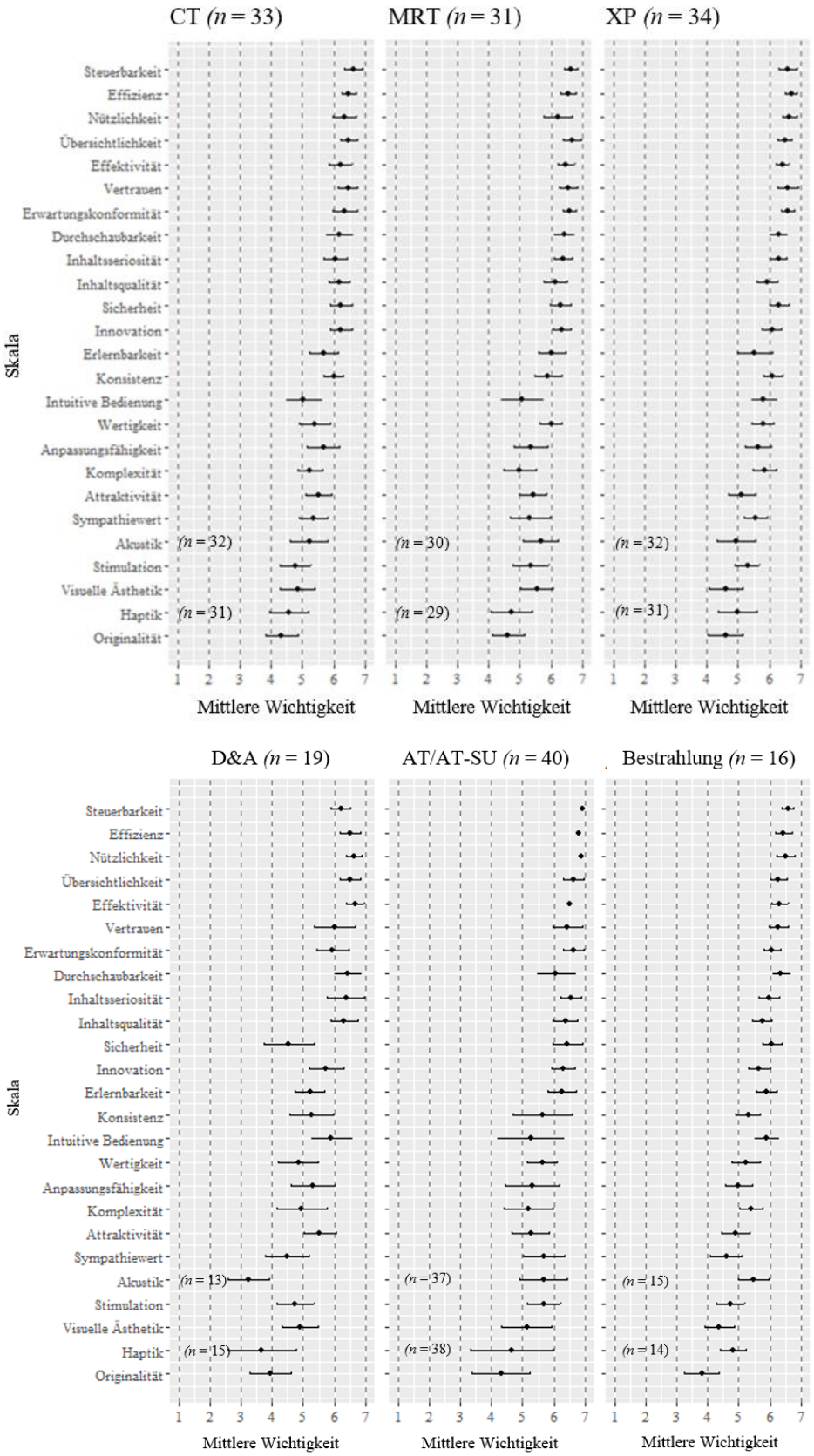
[b] Skalen, die sich unter den sechs wichtigsten Skalen einzelner Business Lines befanden.

[c] Bonferroni-Holm-korrigierter p-Wert.

* $p < .05$.

Abbildung 10

Wichtigkeitsbewertungen der Skalen unterteilt nach Business Lines



Die durchgeführten Post-hoc-Tests sprachen insgesamt dafür, dass ähnliche Skalen für alle Business Lines wichtig waren und sich die signifikanten Effekte der MANOVA sowie des erweiterten Kruskal-Wallis-Tests vor allem auf für den UX-Fragebogen irrelevante Skalen bezogen. Das Ranking wies für alle Business Lines den gleichen Trend auf (s. Abbildung 10), es gab nur geringfügige Abweichungen in den exakten Rangplätzen. Unterschiede in der mittleren Wichtigkeitsbewertung fielen so gering aus, dass eine Differenzierung zwischen den Rangplätzen in vielen Fällen erst bei Einbezug der zweiten Nachkommastelle möglich war. Da außer des Kruskal-Wallis-Tests in Bezug auf die Skala Steuerbarkeit keiner der Post-hoc-Tests auf signifikante Unterschiede in der mittleren Wichtigkeitsbewertung im Hinblick auf relevante Skalen hindeutete, wurde eine produktübergreifende Version des UX-Fragebogens als akzeptabel angenommen. Für diese produktübergreifende Version des Fragebogens wurden die Skalen Übersichtlichkeit, Effektivität, Effizienz, Steuerbarkeit, Vertrauen und Nützlichkeit herangezogen, welche sich in der vorliegenden Studie als sechs produktübergreifend wichtigste Skalen herausstellten. Alle Skalen, bis auf die Skala Nützlichkeit, zeigten sich auch in Müßigs Studie als produktübergreifend besonders relevant. Nützlichkeit belegte in Müßigs Studie produktübergreifend dagegen nur den 13. Rang (M = 6.12, SD = 1.20, Mdn = 7, N = 67). Weil in der vorliegenden Arbeit jedoch mehr als doppelt so viele Teilnehmende einbezogen wurden, wurden die Ergebnisse der Replikationsstudie als repräsentativer gewertet und entsprechend die Skala Nützlichkeit anstelle von Sicherheit in den produktübergreifenden UX-Fragebogen für Medizinprodukte aufgenommen. Bei der Skala Nützlichkeit handelt es sich um eine bereits bestehende Skala des UEQ+ (Schrepp & Thomaschewski, 2019a), für die entsprechend Items existieren. In Anhang E sind noch einmal alle sechs Skalen des produktübergreifenden UX-Fragebogens mit ihren Items zusammengefasst.

Da sich auch andere Skalen für mehrere Business Lines als relevant erwiesen, weil sie sich unter den sechs wichtigsten Skalen einzelner Business Lines befanden, wurden, wie bereits von Müßig (2022) vorgeschlagen, Empfehlungen für optionale Zusatzskalen abgeleitet. Diese können zusätzlich zu den sechs Skalen des produktübergreifenden UX-Fragebogens je nach Business Line und Fragestellung eingesetzt werden. Maßgeblich für die Auswahl der optionalen Zusatzskalen waren die Ergebnisse der Replikationsstudie, allerdings traten viele der empfohlenen optionalen Zusatzskalen auch bei Müßig (2022) unter den Top 6 Skalen einiger Business Lines auf (Anhang D, Abbildung D1). Die Skala Erwartungskonformität zeigte sich in der Replikationsstudie für die Business Lines CT, MR, XP und Strahlentherapie als sinnvolle optionale Zusatzskala. Sie ist

allerdings nicht Bestandteil des UEQ+ und erfordert deshalb die Entwicklung neuer Items. Weiterhin könnte die Skala Inhaltsseriosität wertvolle Zusatzinformationen für die Business Lines D&A und Strahlentherapie liefern sowie die Skala Durchschaubarkeit für die Business Lines AT/AT-SU sowie D&A. Diese beiden Skalen können direkt eingesetzt werden, denn sie sind Teil des UEQ+. Nachdem Sicherheit in Müßigs Studie produktübergreifend als sehr wichtig bewertet wurde und in der Replikationsstudie produktübergreifend immerhin (gemeinsam mit der Skala Inhaltsqualität) den 10. Rang belegte, könnte auch dieser Aspekt je nach Fragestellung eine hilfreiche Ergänzung darstellen. Die Evaluation der entwickelten Sicherheits- und Effektivitätsitems wird im nächsten Kapitel beschrieben. In Anhang F sind optionale Zusatzskalen zusammengefasst.

3.5.4 Ergebnisse der Fragebogenentwicklung.

Es wird nun im Folgenden die Prüfung der dritten und vierten Hypothesen beschrieben, die sich auf die neu entwickelten Sicherheits- und Effektivitätsitems sowie die Struktur des UX-Fragebogens insgesamt bezogen.

3.5.4.1 Hypothese 3. Zuerst werden die Ergebnisse bezüglich der entwickelten Sicherheitsitems, anschließend jene in Bezug auf die Effektivitätsitems berichtet.

3.5.4.1.1 Skala Sicherheit. Hypothese 3.1 beschäftigte sich mit der Entwicklung der Skala Sicherheit. Durch den Umstand, dass Sicherheit nicht mehr unter die sechs produktübergreifend wichtigsten Skalen fiel, wurde die Skala nicht mehr als grundlegender Bestandteil des UX-Fragebogens für Medizinprodukte entwickelt. Um die Informationen aus den Workshops zur Itementwicklung zu nutzen, wurde die Skala trotzdem gebildet, sodass sie sich in Zukunft optional je nach Fragestellung einsetzen lässt. Für die Sicherheitsitems wurden zwei separate Analysen berechnet, da einige der Items sich nur auf Hardware-Produkte anwenden ließen und somit nicht zu Softwareanwendungen (insbesondere der Business Line D&A) passten. In eine erste Faktorenanalyse wurden nur allgemeine Sicherheitsitems einbezogen, die alle Proband*innen beantworten konnten. In eine zweite Analyse gingen alle Sicherheitsitems (allgemeine und Hardware-spezifische) ein und sie wurde nur mit denjenigen 154 Teilnehmenden berechnet, die nicht zur Business Line D&A gehörten, da letztere Hardware-Items nicht beantworten konnten. Für die Berechnungen wurde das „psych“-Paket in R genutzt.

Zunächst wurden die Voraussetzungen zur Durchführung einer Faktorenanalyse mit den allgemeinen Sicherheitsitems untersucht. Der Mardia-Test deutete auf mangelnde multivariate Normalverteilung der allgemeinen Sicherheitsitems hin, wohingegen sich viele der Histogramme einer Normalverteilung annäherten und der Q-Q-Plot keine allzu

großen Abweichungen andeutete. Wegen der Uneindeutigkeit bezüglich der Annahme von Normalverteilung wurden deshalb darüber hinaus Schiefe und Kurtosis der Items ausgegeben. Nach einer Daumenregel von West, Finch und Curran (1995, zitiert nach Bühner, 2011) sollte die Schiefe im Betrag kleiner als zwei sowie die Kurtosis im Betrag kleiner als sieben ausfallen, damit keine zu große Abweichung von einer Normalverteilung vorliegt. Diese Bedingungen waren für alle allgemeinen Sicherheitsitems erfüllt, sodass ausreichende Annäherung an eine Normalverteilung als gegeben betrachtet wurde. Die Items korrelierten laut Kaiser-Meyer-Olkin-Kriterium (KMO) und Bartlett-Test ausreichend hoch (s. Anhang D, Tabelle D1). Streudiagramme deuteten auf lineare Zusammenhänge zwischen den Variablen hin. Für die zweite geplante Faktorenanalyse unter Einbezug der Hardware-Items wurden diese Voraussetzungen separat geprüft. Es zeigten sich vergleichbare Voraussetzungen wie für die allgemeinen Sicherheitsitems, sodass die gleichen Verfahren angewendet werden konnten. Schrepp und Thomaschewski (2019a) hatten für die Entwicklung des UEQ+ Hauptkomponentenanalysen berechnet. Maximum-Likelihood-Faktorenanalysen bieten allerdings den Vorteil, dass Fit-Indizes ausgegeben werden, mit denen sich die Passung des angenommenen Modells zu den beobachteten Daten einschätzen lässt. Die Ergebnisse der Verfahren wurden daher verglichen.

In einem ersten Schritt wurde eine Parallelanalyse ausgegeben, welche in Anhang D eingesehen werden kann (Abbilung D2). Die Software gab sowohl eine Extraktionsempfehlung anhand einer Faktorenanalyse (Hauptachsenanalyse) als auch anhand einer Hauptkomponentenanalyse aus. Es wurde die Extraktion von einer Hauptkomponente, jedoch von zwei Faktoren empfohlen. Der grafische Eigenwerteverlauf legte nach dem Elbow-Kriterium die Extraktion eines Faktors nahe (Moosbrugger & Kevala, 2020). Aufgrund dieses Widerspruchs wurde an dieser Stelle noch keine Entscheidung gefällt, wie viele Faktoren extrahiert werden sollten. Stattdessen wurde direkt die zweite Parallelanalyse für alle Sicherheitsitems (inklusive der Hardware-Items), entsprechend ohne Proband*innen der Business Line D&A durchgeführt. Die Parallelanalyse führte diesmal sowohl auf Grundlage einer Hauptkomponentenanalyse als auch auf Grundlage einer Faktorenanalyse zu einer Empfehlung von zwei Faktoren (beziehungsweise Hauptkomponenten), was auch grafisch nachvollziehbar war (s. Anhang D, Abbildung D3). In einem nächsten Schritt wurde untersucht, ob die Ladungen der Faktorenanalyse (Maximum-Likelihood) beziehungsweise der Hauptkomponenten bei Betrachtung aller Sicherheitsitems auf eine Trennung von allgemeinen und Hardware-Items hindeuteten. Die Ladungen sprachen eindeutig für die Annahme eines allgemeinen und eines Hardware-

spezifischen Faktors (bzw. einer Hauptkomponente) und somit gegen eine Unterteilung der allgemeinen Sicherheitsitems, wie Tabelle 16 zu entnehmen ist. Auf Basis der bisherigen Ergebnisse wurde demnach auf eine 2-Faktor-Lösung geschlossen. Die Annahme von Eindimensionalität der Skala Sicherheit nach Hypothese 3.1.1 war somit nicht erfüllt.

Tabelle 16

Vergleich der Ergebnisse der explorativen Maximum-Likelihood-Faktorenanalyse sowie der Hauptkomponentenanalyse inklusive der Hardware-Sicherheitsitems

Sicherheitsitem [a]	Faktorladung [b]			
	ML-Faktoren-analyse [c]		Hauptkomponen-tenanalyse [d]	
	1	2	1	2
Allgemeine Sicherheitsitems				
Unwahrscheinlich	**.68***	.14	**.67**	.18
Rechtzeitig rückgemeldet	**.79***	-.01	**.81**	-.01
Leicht verständlich angezeigt	**.75***	.08	**.72**	.14
Leicht erkennbar	**.89***	-.08	**.86**	-.03
Unterbunden	**.71***	.01	**.76**	.00
Leicht behebbar	**.57***	.12	**.66**	.06
Kontrolliert	**.64***	.15	**.62**	.20
Nicht durch fehlende Informationen bedingt	**.50***	.03	**.58**	-.04
Aufhaltbar	**.65***	-.03	**.76**	-.12
Gekennzeichnet	**.84***	-.17	**.83**	-.12
Abgesichert	**.68***	.14	**.64**	.22
Beeinflussbar	**.64***	-.03	**.74**	-.10
Unerheblich	**.44***	**.37***	**.43**	**.40**
Hardware-Items				
Harmlos	-.06	**.87***	.09	**.89**
Nicht gesundheitsgefährdend	-.04	**.89***	-.02	**.86**
Nicht kollisionsbegünstigend	.08	**.56***	-.02	**.79**

Sicherheitsitem [a]	Faktorladung [b]			
	ML-Faktoren-analyse [c]		Hauptkomponen-tenanalyse [d]	
	1	2	1	2
Nicht schädigend	.04	**.89***	.04	**.85**
Nicht durch unbeabsichtigte Bewegungen möglich	.23	**.48***	.09	**.68**
Durch einfache Reinigung vermieden	.19	**.40***	.09	**.55**
Durch die Verarbeitung des Produkts unwahrscheinlich	**.32***	**.41***	.22	**.58**

Anmerkung. N = 154. Es wurde für beide Analysen eine Oblimin-Rotation verwendet. Faktorladungen > .30 sind hervorgehoben.

[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

[b] Auf Basis der Parallelanalysen wurden 2 Faktoren/Hauptkomponenten extrahiert.

[c] ML = Maximum-Likelihood-Faktorenanalyse.

[d] Für die Hauptkomponentenanalyse gab es keinen Signifikanztest.

* p < .001 (Bonferroni-adjustierter Wert, um die Wahrscheinlichkeit für einen Fehler 1. Art auf einem Niveau von .05 zu halten).

Bei Betrachtung der Fit-Indizes, die im Rahmen der Maximum-Likelihood-Faktorenanalyse ausgegeben wurden, fiel auf, dass der Fit des Modells mit zwei Faktoren (bezüglich des Modells inklusive Hardware-Items) nicht zufriedenstellend war. Als ein Fit-Index wurde der Root-Mean-Square-Error of Approximation (RMSEA) herangezogen, bei dem Werte von kleiner oder gleich 0.08 als akzeptabel angesehen werden (Hu & Bentler, 1998, 1999). Ein weiterer Index, der Aufschluss darüber gibt, welche Verbesserung des Fits ein angenommenes Modell gegenüber einem Nullmodell bringt, ist der Tucker-Lewis-Index (TLI). Hier sollten Werte größer oder gleich 0.95 sein (Hu & Bentler, 1998, 1999). Tabelle 17 vergleicht die Fit-Indizes und Varianzaufklärungen bei Variation der Anzahl an Faktoren beziehungsweise Hauptkomponenten. Es ist erkennbar, dass sowohl für das Modell ohne Hardware-Items als auch für das Modell mit Hardware-Items der Fit bei Hinzunahme eines Faktors überlegen war. Die Konfidenzintervalle des RMSEA überschnitten sich allerdings sowohl beim Vergleich der beiden Modelle ohne Hardware-Items (mit

einem bzw. zwei Faktoren) als auch beim Vergleich der Modelle mit Hardware-Items (mit zwei bzw. drei Faktoren). Es lag somit laut dem RMSEA keine signifikante Überlegenheit der Modelle mit einem zusätzlichen Faktor vor.

Tabelle 17

Vergleich verschiedener Faktoren- und. Hauptkomponentenanalysen

Anzahl Faktoren/HK [a]	Methode	TLI	RMSEA	95 % KI		Varianzaufklärung	Korrelation der Faktoren/HK
				UG [d]	OG [e]		
Allgemeine Sicherheitsitems							
1 Faktor/HK	ML [b]	0.88	0.10	0.08	0.12	48 %	–
	PCA [c]	–	–	–	–	52 %	–
2 Faktoren/	ML	0.96	0.06	0.03	0.09	55 %	.67
HK	PCA	–	–	–	–	61 %	.57
Allgemeine und Hardware-Sicherheitsitems							
2 Faktoren/	ML	0.86	0.09	0.08	0.11	53 %	.50
HK	PCA	–	–	–	–	57 %	.49
3 Faktoren/	ML	0.89	0.08	0.06	0.10	57 %	.27/.48/.60
HK	PCA	–	–	–	–	63 %	.29/.45/.55

Anmerkung. $N_{allgemein}$ = 173. $N_{inkl.Hardware}$ = 154 (ohne D&A-Teilnehmende). Es wurde Oblimin-Rotation verwendet.

[a] HK = Hauptkomponente(n).

TLI = Tucker-Lewis-Index.

RMSEA = Root-Mean-Square-Error of Approximation.

[b] ML = Maximum-Likelihood-Faktorenanalyse.

[c] PCA = Hauptkomponentenanalyse.

[d] UG = Untere Grenze des Konfidenzintervalls.

[e] OG = Obere Grenze des Konfidenzintervalls.

Gegen die Hinzunahme eines weiteren Faktors für die allgemeinen Sicherheitsitems sprach außerdem, dass sich für eine mögliche weitere Skala keine vier geeigneten Items finden ließen. Die Items „schwer behebbar vs. leicht behebbar“, „unaufhaltbar vs. aufhaltbar“ und „unveränderbar vs. beeinflussbar“ zeigten im Rahmen der Maximum-Likelihood-Faktorenanalyse des Modells ohne Hardware-Items mit zwei Faktoren Ladungen von .55, .79 und .86 auf den zweiten Faktor, auf den ersten Faktor luden sie mit .22, .03 und -.05. Das Item „begünstigt vs. unterbunden“ jedoch wies eine Doppelladung auf. Es lud auf den zweiten Faktor mit .43 und auf den ersten mit .37, wodurch keine eindeutige Zuordnung zu einem Faktor möglich war (s. Anhang D, Tabelle D2). Bei Unterteilung der allgemeinen Sicherheitsitems hätte deshalb eine zweite unvollständige Skala resultiert. Dazu kam, dass keine sinnvolle inhaltliche Abgrenzung der Faktoren möglich war. Weil es demnach sowohl theoretisch nahe lag die allgemeinen Sicherheitsitems nicht in zwei Faktoren zu unterteilen als auch die grafischen Anhaltspunkte und die Mehrzahl der Ausgaben der Parallelanalysen dagegensprachen, wurde die Annahme eines Faktors für die allgemeinen Sicherheitsitems trotz des schlechteren Fits beibehalten. Die beiden Sicherheitsskalen wurden *Allgemeine Sicherheit* und *Hardware-Sicherheit* genannt.

Im Folgenden war das Ziel, geeignete Items für die beiden Skalen zu finden. In bereits erwähnter Tabelle 16 ist zu sehen, welche Items im zweifaktoriellen Modell mit allen Sicherheitsitems signifikante Ladungen aufwiesen. Die Tabelle zeigt auch, dass sich die Ladungsmuster der Maximum-Likelihood-Methode sowie der Hauptkomponentenanalyse stark ähneln. Auf den Faktor der Hardware-Items luden die Items „bedrohlich vs. harmlos“, „gesundheitsgefährdend vs. nicht gesundheitsgefährdend“, „kollisionsbegünstigend vs. nicht kollisionsbegünstigend“ und „schädigend vs. nicht schädigend“ am höchsten. Die Ladungen waren im Einklang mit Hypothese 3.1.2 signifikant von null verschieden. Für die Beurteilung der Itemschwierigkeit wurde zum einen der Mittelwert der Items und zum anderen der prozentuale Schwierigkeitsindex betrachtet (Döring & Bortz, 2016, S. 477). Die Mittelwerte waren bereits in Tabelle 11 (3.5.2 Deskriptive Statistik) abgetragen. Alle Hardware-Items wiesen mit einem Mittelwert von 4.2 bis 4.6 eine ähnliche, mittlere Schwierigkeit auf. Der prozentuale Schwierigkeitsindex wurde nur für die am höchsten ladenden Items berechnet. Ein Bereich von 20 – 80 % gilt als sinnvoll, damit die Items differenzieren (Döring & Bortz, 2016). Das Item „bedrohlich vs. harmlos“ wies einen Schwierigkeitsindex von 55.94 %, das Item „gesundheitsgefährdend vs. nicht gesundheitsgefährdend“ von 55.71 %, das Item „kollisionsbegünstigend vs. nicht

kollisionsbegünstigend“ von 56.18 % und das Item „schädigend vs. nicht schädigend“ einen Schwierigkeitsindex von 54.43 % auf. Damit erwiesen sich alle Hardware-Items bezüglich ihrer Schwierigkeit als geeignet (im Einklang mit Hypothese 3.1.3). Ideal wäre es zwar gewesen, die Items so auszuwählen, dass sie möglichst unterschiedliche Schwierigkeiten aufweisen, allerdings war bereits anhand der Mittelwerte zu erkennen, dass die Schwierigkeit aller Items ähnlich war. Ein weiteres Kriterium, das in die Auswahl der Items einbezogen werden sollte, war die Anzahl fehlender Werte für jedes Item, da diese Aufschluss über seine Verständlichkeit gab. Bei allen Items fehlten einige Werte, wie Tabelle 11 (3.5.2 Deskriptive Statistik) zeigt. Bei den Items „bedrohlich vs. harmlos“ und „gesundheitsgefährdend vs. nicht gesundheitsgefährdend“ fehlten mit vier beziehungsweise fünf Werten die wenigsten Werte im Vergleich zu den anderen Items. Auch das Item „schädigend vs. nicht schädigend“ wies weniger fehlende Werte als manch andere Hardware-Items auf. In Bezug auf das Item „kollisionsbegünstigend vs. nicht kollisionsbegünstigend“ ergaben sich 12 fehlende Werte. Nachdem sich für die anderen Items aber keine deutliche Überlegenheit in dieser Hinsicht zeigte, wurde das Item beibehalten.

Bezüglich des Faktors der allgemeinen Sicherheitsitems gab es für das Modell inklusive Hardware-Items geringfügig unterschiedliche Ergebnisse zwischen der Maximum-Likelihood-Faktorenanalyse und der Hauptkomponentenanalyse im Hinblick auf die vier höchstladenden Items. Es gab eine Vielzahl signifikanter Ladungen größer als .30 (im Einklang mit Hypothese 3.1.2). Die drei Items „schwer erkennbar vs. leicht erkennbar“, „nicht gekennzeichnet vs. gekennzeichnet“ und „nicht rechtzeitig rückgemeldet vs. rechtzeitig rückgemeldet“ luden in beiden Verfahren am höchsten auf den Faktor beziehungsweise die Hauptkomponente (s. Tabelle 16). Nur das Item mit der viert höchsten Ladung wich ab. In der Maximum-Likelihood-Faktorenanalyse hatte die viert höchste Ladung das Item „schwer verständlich angezeigt vs. leicht verständlich angezeigt“, in der Hauptkomponentenanalyse teilten sich die Items „begünstigt vs. unterbunden“ und „unaufhaltbar vs. aufhaltbar“ die viert höchste Ladung. Die Ergebnisse wurden mit dem Modell ohne Hardware-Items verglichen. Hier stimmten für die Maximum-Likelihood-Faktorenanalyse und die Hauptkomponentenanalyse die Items mit den vier höchsten Ladungen überein, wie Tabelle 18 aufzeigt. Es handelte sich um die Items „nicht rechtzeitig rückgemeldet vs. rechtzeitig rückgemeldet“, „schwer verständlich angezeigt vs. leicht verständlich angezeigt“, „schwer erkennbar vs. leicht erkennbar“ und „ungesichert vs. abgesichert“.

Tabelle 18

Vergleich der Ergebnisse der explorativen Maximum-Likelihood-Faktorenanalyse sowie der Hauptkomponentenanalyse anhand der allgemeinen Sicherheitsitems

Allgemeines Sicherheitsitem [a]	Faktorladung [b]	
	ML-Faktoren-analyse [c]	Hauptkomponen-tenanalyse [d]
Unwahrscheinlich	.68*	.71
Rechtzeitig rückgemeldet	.79*	.81
Leicht verständlich angezeigt	.78*	.79
Leicht erkennbar	.84*	.84
Unterbunden	.71*	.75
Leicht behebbar	.65*	.69
Kontrolliert	.71*	.73
Nicht durch fehlende Informationen bedingt	.48*	.54
Aufhaltbar	.65*	.69
Gekennzeichnet	.70*	.72
Abgesichert	.75*	.77
Beeinflussbar	.62*	.67
Unerheblich	.60*	.64

Anmerkung. $N = 173$.

[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

[b] Auf Basis der Parallelanalysen wurde ein Faktor/eine Hauptkomponente extrahiert.

[c] ML = Maximum-Likelihood-Faktorenanalyse.

[d] Für die Hauptkomponentenanalyse gab es keinen Signifikanztest.

* $p < .002$ (Bonferroni-adjustierter Wert, um die Wahrscheinlichkeit für einen Fehler 1. Art auf einem Niveau von .05 zu halten).

Die Items „schwer erkennbar vs. leicht erkennbar“ und „nicht rechtzeitig rückgemeldet vs. rechtzeitig rückgemeldet“ waren zusammengenommen in allen Verfahren unter den höchst ladenden Items und wurden deshalb in die Skala Allgemeine Sicherheit aufgenommen. Für die Auswahl der fehlenden beiden Items wurden die Schwierigkeit und die Anzahl der fehlenden Werte der Items in der engeren Auswahl herangezogen. Diese sind in Tabelle 19 ersichtlich. Die Schwierigkeit aller Items lag in einem passablen, mittleren Bereich (im Einklang mit Hypothese 3.1.3). Allerdings war eine gewünschte breite Streuung der Schwierigkeit für die allgemeinen Sicherheitsitems ebenso wie bei den Items für Hardware-Sicherheit nicht realisierbar. Das Item „schwer verständlich angezeigt vs. leicht verständlich angezeigt“ wies eine vergleichsweise geringe Anzahl fehlender Werte auf und wurde deshalb in die Skala Allgemeine Sicherheit aufgenommen. Da das Item „unaufhaltbar vs. aufhaltbar“ am ehesten einen abweichenden Schwierigkeitsindex gegenüber den anderen Items aufwies, wurde es als geeignet befunden. Ein weiterer Vorteil dieses Items bestand darin, dass es einen etwas anderen inhaltlichen Aspekt erfragte, womit sich seine Aufnahme trotz der relativ hohen Anzahl fehlender Werte ebenfalls begründen ließ. Eine Alternative zu diesem Item wäre aufgrund der geringeren Anzahl fehlender Werte das Item „nicht gekennzeichnet vs. gekennzeichnet“ gewesen. Dieses Item überschneidet sich jedoch inhaltlich mit dem Item „schwer erkennbar vs. leicht erkennbar“ und wurde deshalb aussortiert.

Zusammenfassend ließ sich im Hinblick auf die entwickelten Sicherheitsitems keine Eindimensionalität nachweisen (entgegen Hypothese 3.1.1). Es wurden zwei Faktoren (Skalen) extrahiert, sodass eine Skala für Allgemeine Sicherheit und eine für Hardware-Sicherheit entstand. Für beide Skalen fanden sich je vier geeignete Items mit signifikanten Ladungen größer als .30 und einer geeigneten Schwierigkeit (im Einklang mit den Hypothesen 3.1.2 und 3.1.3). Eine Übersicht der ausgewählten Items für die beiden Skalen befindet sich in Anhang F.

Tabelle 19

Schwierigkeit und weitere Kennwerte der allgemeinen Sicherheitsitems in der engeren Auswahl

Allgemeines Sicherheitsitem [a]	*M*	*SD*	*P* [b]	*NA* [c]	Auswahl [d]	Grund für Auswahl/Ausschluss
Leicht erkennbar	4.7	1.5	61.31 %	5	Ja	Übereinstimmend hohe Ladung [e]
Rechtzeitig rückgemeldet	4.6	1.6	60.10 %	13	Ja	Übereinstimmend hohe Ladung [e]
Gekennzeichnet	4.6	1.5	59.82 %	10	Nein	Überschneidung mit „leicht erkennbar“
Leicht verständlich angezeigt	4.6	1.6	59.29 %	8	Ja	Geringe Anzahl fehlender Werte
Unterbunden	4.4	1.2	57.14 %	19	Nein	Hohe Anzahl fehlender Werte
Aufhaltbar	4.8	1.5	63.87 %	12	Ja	Schwierigkeit und inhaltlicher Zusatzaspekt abgedeckt
Abgesichert	4.7	1.5	61.95 %	14	Nein	Hohe Anzahl fehlender Werte

Anmerkung. *N* = 173. Ausführlichere Erläuterungen befinden sich im Text.

[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

[b] Prozentualer Schwierigkeitsindex (Döring & Bortz, 2016).

[c] NA = not available (Anzahl an Auslassungen).

[d] Hiermit ist die Aufnahme des Items in die Skala der allgemeinen Sicherheitsitems gemeint.

[e] Die Ladungen wurden zuvor im Text berichtet.

3.5.4.1.2 Skala Effektivität. Hypothese 3.2 beschäftigte sich mit der Entwicklung der Skala Effektivität. Es wurde genauso vorgegangen wie bei der Evaluation der Sicherheitsitems. Bei der Voraussetzungsprüfung sprachen Histogramme, Q-Q-Plot und der Mardia-Test gegen die Annahme multivariater Normalverteilung (s. Anhang D, Tabelle D1). Viele der Items zeigten sich linksschief verteilt. KMO und Bartlett-Test ergaben eine ausreichend hohe Korrelation der Items, um eine Faktorenanalyse durchführen zu können. Lineare Zusammenhänge der Variablen ließen sich auf Basis von Streudiagrammen annehmen. Es wurden, wie für die Items der Skala Sicherheit, eine Maximum-Likelihood-Faktorenanalyse und eine Hauptkomponentenanalyse berechnet. Aufgrund der nicht erfüllten Normalverteilung wurden die Ergebnisse zusätzlich mit denen einer Hauptachsenanalyse verglichen, welche robuster gegenüber Verletzungen der Normalverteilung ist (Schnaudt, 2020).

Eine Parallelanalyse, welche in Anhang D (Abbildung D4) zu sehen ist, führte eindeutig zur Annahme eines Faktors beziehungsweise einer Hauptkomponente. Eindimensionalität der Skala Effektivität im Sinne von Hypothese 3.2.1 konnte somit angenommen werden. Als nächstes wurden die Ladungen der Items betrachtet, welche Tabelle 20 zusammenfasst. Die Varianzaufklärung betrug bezüglich der Maximum-Likelihood-Faktorenanalyse und der Hauptachsenanalyse 73 %, für die Hauptkomponentenanalyse lag sie bei 74 %. Der RMSEA und TLI sprachen mit einem Wert von 0.07, 95 % KI [0.05, 0.09], beziehungsweise 0.97 für einen sehr hohen Fit.

Tabelle 20

Vergleich der Ergebnisse der explorativen Maximum-Likelihood-Faktorenanalyse, der Hauptachsenanalyse und der Hauptkomponentenanalyse anhand der Effektivitätsitems

Effektivitätsitem [a]	Faktorladung [b]		
	ML-Faktoren-analyse [c]	Hauptachsen-analyse	Hauptkompo-nentenanalyse [d]
Zweckmäßig	.87*	.87*	.88
Nützlich	.89*	.89*	.90
Auf meine Anforderungen abgestimmt	.81*	.81*	.83
Zielerfüllend	.89*	.88*	.89
Vollständig	.88*	.87*	.88
Genau	.87*	.87*	.88
Eindeutig	.91*	.90*	.91
Adäquat	.89*	.89*	.89
Entscheidungsunterstützend	.74*	.75*	.77
Hochwertig	.85*	.85*	.86
Bedarfsgerecht	.90*	.91*	.91
Konstant	.76*	.76*	.79
Überdurchschnittlich	.76*	.76*	.79
Praxistauglich	.88*	.87*	.88

Anmerkung. $N = 173$.

[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

[b] Auf Basis der Parallelanalysen wurde ein Faktor/eine Hauptkomponente extrahiert.

[c] ML = Maximum-Likelihood.

[d] Für die Hauptkomponentenanalyse gab es keinen Signifikanztest.

* $p < .004$ (Bonferroni-adjustierter Wert, um die Wahrscheinlichkeit für einen Fehler 1. Art auf einem Niveau von .05 zu halten).

Wie zu erkennen ist, kamen alle drei Verfahren zu sehr hohen Ladungen. Die Items “uneindeutig vs. eindeutig“, „nicht bedarfsgerecht vs. bedarfsgerecht“, „nutzlos vs. nützlich“ und „inadäquat vs. adäquat“ zeigten mit geringem Unterschied zu den anderen Items die höchsten Ladungen, wenn man die Verfahren gemeinsam betrachtete. In Tabelle 12 im Abschnitt 3.5.2 (Deskriptive Statistik) sind die Itemmittelwerte zur Beurteilung der Schwierigkeit abgetragen. Insgesamt wiesen alle Items eine eher geringe Schwierigkeit auf, die Mittelwerte reichten von 4.9 bis 5.9 und lagen damit über dem mittleren Skalenwert von 4. Der prozentuale Schwierigkeitsindex des Items “uneindeutig vs. eindeutig“ lag bei 78.01 %, der des Items „nicht bedarfsgerecht vs. bedarfsgerecht“ bei 76.25 % und der des Items „inadäquat vs. adäquat“ bei 76.59 %. Das Item „nutzlos vs. nützlich“ erfüllte den Anforderungsbereich von 20 – 80 % mit einem Schwierigkeitsindex von 81.29 % nicht und wurde daher aussortiert. Das Item „unterdurchschnittlich vs. überdurchschnittlich“ wies den niedrigsten Mittelwert von 4.9 auf. Auch wenn die Ladung des Items niedriger ausfiel als die Ladung anderer Items, wurde es aufgrund des Schwierigkeitsindex von 64.60 % aufgenommen, sodass die Schwierigkeit der Items unter den gegebenen Umständen möglichst breit streute. Jedes der Effektivitätsitems wurde von mindestens einer Person wegen Unverständlichkeit ausgelassen. Die Anzahl fehlender Werte wurde für die Skala Effektivität jedoch nicht in die Auswahl der Items einbezogen, da keines der vorausgewählten Items durch herausstechend viele Auslassungen auffiel. Zusammengefasst erwies sich die Skala Effektivität als eindimensional (im Einklang mit Hypothese 3.2.1) und es konnten vier geeignet schwere Items, mit signifikanten Ladungen größer als .30 gefunden werden (im Einklang mit Hypothese 3.2.2 und 3.2.3). Die ausgewählten Items für die Skala Effektivität sind in Anhang E zusammengefasst dargestellt.

3.5.4.2 Hypothese 4. Im Rahmen von Hypothese 4 sollte nun die Struktur des produktübergreifenden UX-Fragebogens für Medizinprodukte insgesamt untersucht werden. Die Skala Nützlichkeit zählte zwar zu den sechs produktübergreifend wichtigsten Skalen und somit zum UX-Fragebogen, für sie wurden allerdings keine Items erhoben, weshalb sie nicht einbezogen werden konnte. Die Sicherheitsskalen wurden wegen der Ergebnisse der Replikationsstudie nicht mehr in den produktübergreifenden UX-Fragebogen aufgenommen, sie waren demnach ebenfalls kein Bestandteil des Strukturgleichungsmodells. Erwartet wurde anstelle der 6-Faktoren-Struktur somit eine 5-Faktoren-Struktur unter Einbezug der neuen Skala Effektivität sowie der Skalen Steuerbarkeit,

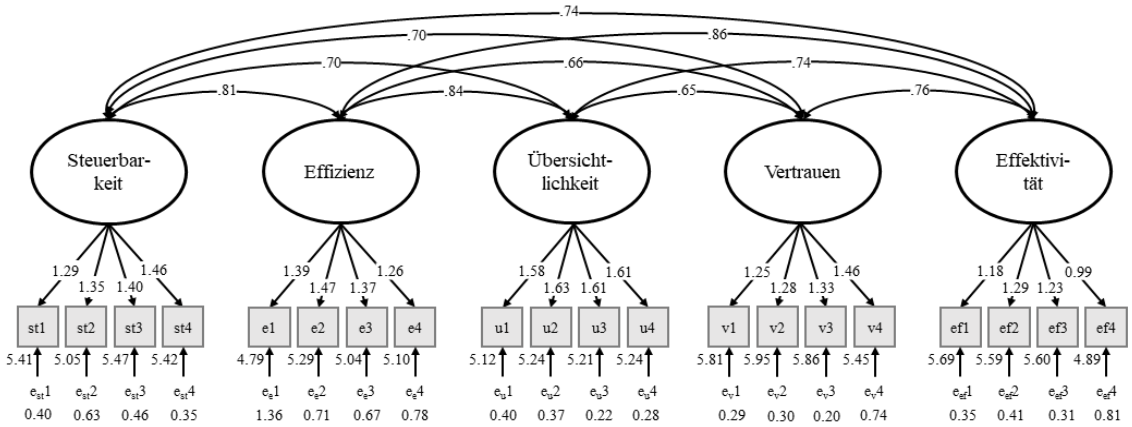
Übersichtlichkeit, Vertrauen und Effizienz des UEQ+ (Schrepp & Thomaschewski, 2019a).

Ein Mardia-Test und Q-Q-Plot zeigten mangelnde multivariate Normalverteilung an, weshalb auf eine robustere Variante der Maximum-Likelihood-Schätzmethode, die „mlr"-Schätzmethode des R-Pakets „lavaan" zurückgegriffen wurde. Streudiagramme deuteten auf lineare Beziehungen zwischen den Variablen hin. Bei der Aufstellung des Strukturgleichungsmodells wurde ein tau-kongenerisches Messmodell für alle Skalen angenommen, bei dem Ladungen (Steigungsparameter), Itemschwierigkeiten (Intercepts) sowie Fehlervarianzen nicht gleichgesetzt, sondern frei geschätzt werden. Der χ^2-Test zur Beurteilung der Abweichung zwischen der beobachteten und der modellimplizierten Kovarianz-Varianz-Matrix wurde entgegen der Annahme von Hypothese 4.1.1 signifikant, $\chi^2(160) = 242.90$, $p < .001$, was für eine Abweichung der Matrizen sprach. Wegen der Stichprobenabhängigkeit sollte der χ^2-Test jedoch mit Vorsicht interpretiert werden. Die Fit-Indizes deuteten im Gegensatz zum χ^2-Test auf einen guten globalen Modell-Fit hin. Der robuste TLI betrug 0.97 und der robuste Comparative Fit Index (CFI) 0.97, was akzeptablen Werten von größer oder gleich 0.95 entsprach (Hu & Bentler, 1998, 1999). Der robuste RMSEA fiel mit einem Wert von 0.06, 90% KI [0.05, 0.08], ebenfalls in einen akzeptablen Wertebereich, genauso wie das Standardized Root Mean Square Residual (SRMR), das mit 0.04 unter der Grenze von kleiner oder gleich 0.09 lag. Die Fit-Indizes deuteten demnach, wie von Hypothese 4.1.2 postuliert, auf eine angemessene Passung des angenommenen Modells zu den beobachteten Daten hin. Abbildung 11 stellt die beobachteten Kovarianzen, Steigungsparameter, Intercepts und Fehlervarianzen des Strukturgleichungsmodells dar.

Der Abbildung des Strukturgleichungsmodells ist des Weiteren zu entnehmen, dass die latenten Variablen sehr hoch korrelierten. Am höchsten erwies sich mit .86 die Korrelation zwischen Effizienz und Effektivität. Die Korrelationen zwischen latenten Variablen im Strukturgleichungsmodell sind um Messfehler bereinigt. Einfache Rangkorrelationen nach Spearman oder Produkt-Moment-Korrelationen nach Pearson fallen geringer, jedoch immer noch sehr hoch aus, wie die Tabellen D3 und D4 in Anhang D zeigen.

Abbildung 11

Unstandardisierte Lösung des Strukturgleichungsmodells



Anmerkung. Abgetragen sind die Kovarianzen, unstandardisierten Steigungsparameter, Intercepts und Fehlervarianzen (von oben nach unten). Die latenten Variablen wurden durch Fixierung des Mittelwerts auf null und der Standardabweichung auf eins standardisiert. Die Kovarianzen zwischen den latenten Variablen entsprechen damit den Korrelationen. Alle Kovarianzen, Steigungsparameter, Intercepts und Fehlervarianzen zeigten sich signifikant von null verschieden ($p \leq .001$). Die Abkürzungen der Items (in den Quadraten) sind unten erläutert. Es wird nur der positive Pol der Items genannt.

st1	vorhersagbar	u3	geordnet
st2	unterstützend	u4	organisiert
st3	sicher	v1	sicher
st4	erwartungskonform	v2	seriös
e1	schnell	v3	zuverlässig
e2	effizient	v4	transparent
e3	pragmatisch	ef1	eindeutig
e4	aufgeräumt	ef2	bedarfsgerecht
u1	gut gegliedert	ef3	adäquat
u2	strukturiert	ef4	überdurchschnittlich

Auf Ebene der lokalen Modellgüte zeigten sich im Einklang mit Hypothese 4.2.1 für alle Items signifikante Steigungsparameter (Ladungen) mit einer Überschreitungswahrscheinlichkeit von kleiner als .001 für alle Ladungen. Die Modifikationsindizes deuteten dagegen auf einige Fehlspezifikationen hin, was gegen die Annahme von Hypothese 4.2.2 sprach. Modifikationsindizes „zeigen an, wie stark sich der Chi-Quadrat-Wert des globalen Hypothesentests verringert, wenn der entsprechende Parameter frei geschätzt werden würde“ (Bühner, 2011, S. 555). Für die Beurteilung von Fehlspezifikationen, wurde das Vorgehen von Saris et al. (2009) gewählt, das Bühner (2011) zusammenfasst. Vereinfacht wird für dieses Vorgehen zunächst betrachtet, ob die Teststärke des Parameters über 75 % liegt. Falls dies der Fall ist, der Modifikationsindex größer oder gleich 3.84 und der expected parameter change (EPC) größer als 0.10 ist, liegt eine Fehlspezifikation vor. Liegt der Modifikationsindex bei einer Teststärke von über 75% unter einem Wert von 3.84, liegt keine Fehlspezifikation vor. Bei einer Teststärke unter 75 % und einem Modifikationsindex größer oder gleich 3.84 ist ebenfalls eine Fehlspezifikation anzunehmen, während bei einem Modifikationsindex unter 3.84 bei einer Teststärke von unter 75 % nicht zwangsläufig eine Fehlspezifikation gegeben ist. Tabelle 21 fasst zusammen, für welche Parameter sich nach diesem Vorgehen Fehlspezifikationen ergaben. Für eine Vielzahl von Items hätte man Zusammenhänge modellieren können, um so einen höheren Fit zu erreichen. In Bezug auf Items der Skalen Effizienz und Steuerbarkeit deuteten hohe Modifikationsindizes auf nicht modellierte Ladungen auf andere Faktoren hin. Vor allem für das Item „überladen vs. aufgeräumt“ der Skala Effizienz wurden Ladungen auf die Faktoren Übersichtlichkeit, Steuerbarkeit und Vertrauen nahegelegt. Es wurde außerdem eine Ladung des Items „ineffizient vs. effizient“ auf den Faktor Übersichtlichkeit sowie eine Ladung des Items „unpragmatisch vs. pragmatisch“ auf den Faktor Steuerbarkeit durch entsprechende Modifikationsindizes angezeigt. Auch für die Items „behindernd vs. unterstützend“ und „nicht erwartungskonform vs. erwartungskonform“ der Skala Steuerbarkeit bildeten sich Doppelladungen auf die Skalen Effizienz, Effektivität und Übersichtlichkeit ab. Items anderer Skalen zeigten keine oder nur vereinzelt nicht modellierte Ladungen auf andere Faktoren und die Modifikationsindizes fielen geringer aus (s. Tabelle 21).

Tabelle 21

Modifikationsindizes zum Strukturgleichungsmodell

Parameter [a]	MI [b]	EPC [c]	Power	Fehlspezifikation
St. unterstützend ~~ Ef. bedarfsgerecht	21.11	0.22	.54	Ja
Ü. Faktor =~ E. aufgeräumt	20.56	0.75	.09	Ja
Ü. gut gegliedert ~~ Ü. organisiert	18.39	0.16	.75	Ja
E. schnell ~~ E. effizient	16.72	0.39	.19	Ja
Ü. strukturiert ~~ Ü. geordnet	16.62	0.16	.75	Ja
St. vorhersagbar ~~ V. sicher	14.43	-0.13	.86	Ja
E. Faktor =~ St. unterstützend	14.04	0.49	.12	Ja
V. zuverlässig ~~ Ef. eindeutig	13.37	0.11	.93	Ja
V. transparent ~~ E. effizient	12.13	-0.22	.34	Ja
Ü. Faktor =~ St. erwartungskonform	9.96	-0.27	.22	Ja
E. effizient ~~ E. aufgeräumt	9.90	-0.24	.26	Ja
V. sicher ~~ Ef. adäquat	8.48	-0.09	.91	Nein
Ü. strukturiert ~~ Ü. organisiert	8.18	-0.11	.74	Ja
V. sicher ~~ Ef. bedarfsgerecht	8.04	0.10	.83	Nein
St. vorhersagbar ~~ V. zuverlässig	7.90	0.09	.91	Nein
St. unterstützend ~~ V. sicher	7.32	0.11	.71	Ja
St. Faktor =~ E. aufgeräumt	7.31	0.41	.10	Ja
Ü. Faktor =~ E. effizient	6.45	-0.44	.09	Ja
St. vorhersagbar ~~ Ü. geordnet	6.30	-0.08	.91	Nein
St. unterstützend ~~ Ef. eindeutig	6.17	-0.11	.62	Ja
Ef. Faktor =~ St. unterstützend	5.97	0.27	.15	Ja
Ü. Faktor =~ St. unterstützend	5.96	0.24	.17	Ja
St. Faktor =~ E. pragmatisch	5.76	-0.36	.10	Ja
Ü. gut gegliedert ~~ Ü. geordnet	5.75	-0.09	.77	Nein
Ü. gut gegliedert ~~ E. effizient	5.63	-0.12	.52	Ja
V. transparent ~~ Ef. überdurchschnittlich	4.46	-0.15	.34	Ja
Ü. geordnet ~~ Ü. organisiert	5.29	-0.09	.77	Nein
E. Faktor =~ St. erwartungskonform	5.15	-0.26	.14	Ja
E. aufgeräumt ~~ Ef. überdurchschnittlich	5.12	-0.15	.31	Ja

Parameter [a]	MI [b]	EPC [c]	Power	Fehlspezifikation
St. vorhersagbar ~~ Ef. bedarfsgerecht	5.11	-0.09	.71	Ja
V. zuverlässig ~~ Ef. bedarfsgerecht	5.02	-0.07	.88	Nein
V. Faktor =~ E. aufgeräumt	4.97	0.24	.15	Ja
Ü. Faktor =~ V. seriös	4.96	-0.15	.32	Ja
Ü. organisiert ~~ Ef. eindeutig	4.96	0.07	.89	Nein
Ü. Faktor =~ V. transparent	4.71	0.21	.19	Ja
Ü. Faktor =~ Ef. überdurchschnittlich	4.69	0.25	.14	Ja
Ü. geordnet ~~ Ef. bedarfsgerecht	4.42	0.07	.88	Nein
V. transparent ~~ E. aufgeräumt	4.34	0.14	.34	Ja
St. vorhersagbar ~~ Ü. organisiert	4.28	0.07	.86	Nein
St. vorhersagbar ~~ Ef. eindeutig	4.21	0.08	.78	Nein
Ef. Faktor =~ St. erwartungskonform	4.15	-0.20	.18	Ja
St. unterstützend ~~ Ef. adäquat	4.14	-0.09	.65	Ja
V. zuverlässig ~~ Ü. organisiert	4.01	0.05	.97	Nein
V. transparent ~~ Ü. strukturiert	3.95	-0.09	.58	Ja

Anmerkung. Die Modifikationsindizes sind der Größe nach absteigend sortiert. Die Skalennamen sind den Items vorangestellt, um sie den Skalen zuordnen zu können. Sie sind abgekürzt mit „St.“ für Steuerbarkeit, „Ef.“ für Effektivität, „E.“ für Effizienz, „V.“ für Vertrauen und „Ü.“ für Übersichtlichkeit.

[a] „~~“ steht für Korrelationen und „=~“ für Ladungen, die im angenommenen Modell nicht modelliert wurden.

[b] MI = Modifikationsindex.

[c] EPC = expected parameter change (Geschätzter Wert, wenn man den Parameter frei schätzen würde).

Insgesamt ließ sich die globale Güte des angenommenen Modells nachweisen (im Einklang mit Hypothese 4.1). Wegen der Stichprobenabhängigkeit des χ^2-Tests wurden die berichteten Fit-Indizes, die auf angemessene globale Güte hindeuteten, als aussagekräftiger bewertet als die Signifikanz des χ^2-Tests, was dem präregistrierten Vorgehen entsprach. Die lokale Modellgüte wurde durch das Vorliegen signifikanter Ladungen nachgewiesen (Hypothese 4.2.1). Das Kriterium lokaler Modellgüte, dass keine Fehlspezifikationen laut Modifikationsindizes gegeben sind, konnte dagegen nicht angenommen werden (entgegen Hypothese 4.2.2). Laut Aichholzer (2017) sind bei Modellen mit vielen Parametern Abweichungen jedoch erwartbar. Dass keine Fehlspezifikationen vorliegen, war deshalb bereits in der Präregistrierung als nicht notwendiges Kriterium festgehalten.

3.5.5 Psychometrische Gütekriterien.

Die fünfte Hypothese beschäftigte sich mit den psychometrischen Gütekriterien des UX-Fragebogens. Die Ergebnisse bezüglich der Reliabilität und Validität werden im folgenden Abschnitt berichtet. Die Skala Nützlichkeit konnte auch hier aufgrund der nicht erhobenen Items nicht einbezogen werden. Ergebnisse zu den psychometrischen Gütekriterien für die beiden Sicherheitsskalen befinden sich im Anhang D (Tabelle D5), da sie nicht Teil der produktübergreifenden Version des UX-Fragebogens sind.

3.5.5.1 Ergebnisse bezüglich der Reliabilität. Zunächst wurde mithilfe des „MBESS“-Pakets McDonalds Omega (McDonald, 1970, 1999) zur Beurteilung der Reliabilität für die einzelnen Skalen berechnet. Für die Skala Steuerbarkeit ergab sich ein Wert von .94, 95 % KI [.92, .96], ebenso wie für die Skala Vertrauen, 95 % KI [.92, .97]. Für die Skala Übersichtlichkeit zeigte sich ein noch höheres McDonalds Omega von .97, 95 % KI [.96, .98]. Auch für die Skalen Effizienz und Effektivität ergaben sich sehr hohe Werte von .90, 95 % KI [.87, .93] und .92, 95 % KI [.90, .94]. Im Einklang mit Hypothese 5.1 ließ sich somit für alle Skalen ein McDonalds Omega von größer oder gleich .80 feststellen, was für eine hohe Reliabilität der Skalen sprach (Lance et al., 2006; Nunnally, 1978).

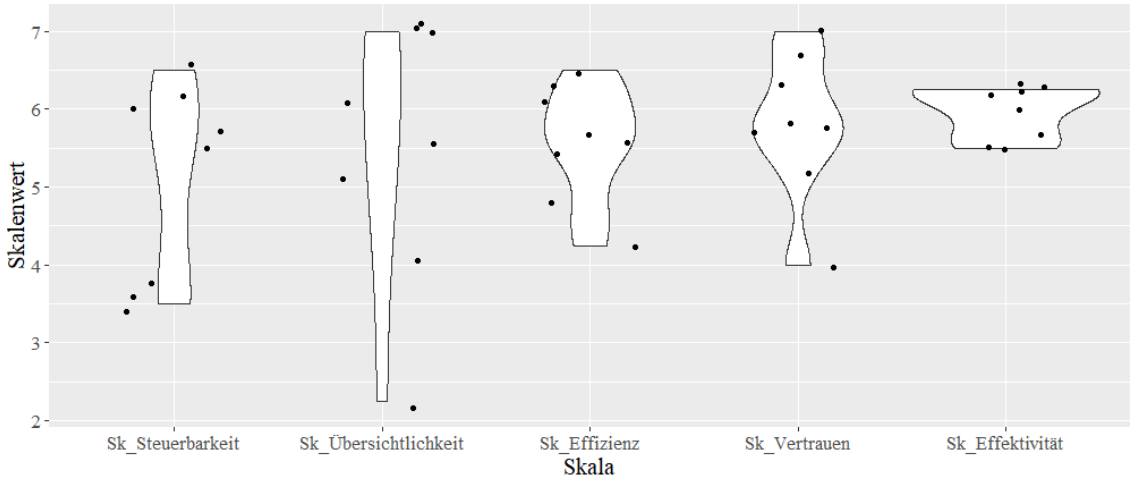
Neben McDonalds Omega wurden die Interrater-Reliabilität sowie die absolute Interrater-Übereinstimmung für ausgewählte Produkte betrachtet. Diese beiden Aspekte der Reliabilität sind zu unterscheiden (Gisev et al., 2013). Die Interrater-Reliabilität bezieht sich auf die relative Bewertung der Skalen, im vorliegenden Fall zum Beispiel darauf, ob Personen die Skala Steuerbarkeit als besser erfüllt bewerten als die Skala Übersichtlichkeit in Bezug auf ein Referenzprodukt. Betrachtet wird also die Konsistenz der Bewertungen, nicht die absolute Bewertung. Falls eine Person insgesamt kritischer

bewertet hat, kann die Konsistenz trotzdem hoch sein. Die absolute Übereinstimmung (interrater agreement) erfasst dagegen, ob exakte Skalenwerte für verschiedene Bewerter*innen übereinstimmen (Gisev et al., 2013). Für beide Maße wurde aus dem Durchschnitt der Items jeder Skala ein Skalenwert für jede Person berechnet. Es wurde ein Radiographie-System ausgewählt, das von acht Teilnehmenden als Referenzprodukt bewertet wurde. Keine der herangezogenen Personen wies einen fehlenden Skalenwert auf. Die Berechnung der Interrater-Reliabilität wurde mithilfe des „irr"-Pakets vorgenommen. Die Intraklassen-Korrelation zeigte eine geringe Übereinstimmung der Teilnehmenden in Bezug auf die Bewertung der fünf produktübergreifend relevanten Skalen unter Verwendung eines „two-way mixed effects"-Modells und der "average-Rater"-Einheit. Die Konsistenz betrug 0.12, 95 % KI [-1.88, 0.90], $F(4,28) = 1.14$, $p = .358$. Nach Koo und Li (2016) liegen Werte ab .50 in einem akzeptablen Bereich, womit das Ergebnis deutlich außerhalb des akzeptablen Bereichs lag. Berechnete man die Intraklassen-Korrelation unter Hinzunahme der beiden Sicherheitsskalen, fiel auf, dass diese mit .78, 95 % KI [.40, .96], $F(6,42) = 4.54$, $p = .001$, deutlich höher ausfiel (s. Anhang D, Tabelle D5).

Ergänzend wurde zur Beurteilung der absoluten Übereinstimmung untersucht, ob die Standardabweichungen im Hinblick auf die einzelnen Skalen kleiner unter den Befragten mit dem gleichen Referenzprodukt waren als in der Gesamtstichprobe. Für die Skalen Steuerbarkeit ($SD_{\text{Produkt}} = 1.29$, $SD_{\text{Gesamt}} = 1.42$), Effizienz ($SD_{\text{Produkt}} = 0.75$, $SD_{\text{Gesamt}} = 1.45$), Vertrauen ($SD_{\text{Produkt}} = 0.93$, $SD_{\text{Gesamt}} = 1.31$) und Effektivität ($SD_{\text{Produkt}} = 0.34$, $SD_{\text{Gesamt}} = 1.23$) war das erfüllt. Für die Skala Übersichtlichkeit ($SD_{\text{Produkt}} = 1.69$, $SD_{\text{Gesamt}} = 1.64$) war dies jedoch nicht der Fall. Zur Veranschaulichung wurden Violinplots ausgegeben. Für eine hohe absolute Übereinstimmung der Teilnehmenden sollten die Werte für jede Skala nahe beieinander liegen, zwischen den Skalen dürfen sie voneinander abweichen. Wie in Abbildung 12 zu sehen ist, lagen einige Werte der Skalen Steuerbarkeit und Übersichtlichkeit weit voneinander entfernt. Die Plots der Skalen Effizienz, Vertrauen und Effektivität deuteten dagegen auf eine höhere absolute Übereinstimmung der Befragten hin. Für die anderen Produktkategorien hatten maximal vier Personen das gleiche Referenzprodukt bewertet. Einige Personen hatten außerdem mehr als ein Referenzprodukt eingetragen, sodass unklar war, für welches genaue Produkt sie die Studie ausgefüllt hatten. Für die Bewertung der Interrater-Reliabilität wurden ausschließlich Personen herangezogen, die nur ein Referenzprodukt eingegeben hatten. Weil die Berechnung der Interrater-Reliabilität für vier oder weniger Personen nicht aussagekräftig war, wurde darauf verzichtet, die Berechnung für weitere Referenzprodukte vorzunehmen.

Abbildung 12

Violinplots der Skalen zur Bewertung eines Radiographie-Systems im Vergleich



Anmerkung. Dargestellt sind die Ergebnisse in Bezug auf jene Skalen, die Teil des produktübergreifenden UX-Fragebogens sind. Die Skala Nützlichkeit ist zwar Bestandteil des UX-Fragebogens, für sie wurden allerdings keine Items erhoben, weshalb sie hier nicht aufgeführt ist.

Zusammenfassend lässt sich festhalten, dass die Reliabilität der Skalen gemessen an McDonalds Omega in einem sehr guten Bereich liegt (im Einklang mit Hypothese 5.1). Die Interrater-Reliabilität im Sinne der Konsistenz schnitt für das ausgewählte Referenzprodukt dagegen nicht gut ab (entgegen Hypothese 5.1). Die grafisch und anhand der Standardabweichungen beurteilte absolute Übereinstimmung der Teilnehmenden war für einige Skalen akzeptabel, für andere zeigte sich allerdings eine geringe absolute Übereinstimmung. In Bezug auf die Reliabilität ließ sich deshalb keine eindeutige Schlussfolgerung ableiten.

3.5.5.2 Ergebnisse bezüglich der Validität. Zur Bewertung der Validität wurden die beiden Zusatzitems „Insgesamt bin ich mit der Nutzerfreundlichkeit des Produkts…“ (*sehr unzufrieden* bis *sehr zufrieden*) und das Item „Wie wahrscheinlich ist es, dass Sie das Produkt einem Kollegen/einer Kollegin weiterempfehlen?“ (*sehr unwahrscheinlich* bis *sehr wahrscheinlich*) erfragt. Erwartet wurde eine signifikante Korrelation von mindestens .30 zwischen den Validierungsitems und den Skalenwerten (Hypothese 5.2.1). Es gab eine Person, die keines der ausgewählten Effektivitätsitems beantwortet hatte, sodass für diese Personen kein Skalenwert berechnet werden konnte. Zunächst wurde ermittelt, ob eine Normalverteilung der beiden Items zur Validierung bestand, was auf Basis des Shapiro-Wilk-Tests und einer grafischen Überprüfung anhand von

Histogrammen abgelehnt wurde. Lineare Beziehungen ließen sich annehmen. Es wurde aufgrund der mangelnden Normalverteilung, welche eine Voraussetzung der Signifikanztests darstellt, auf Spearmans Roh (ρ) zurückgegriffen. Die beiden Validierungsitems zeigten eine Rangkorrelation von .84 (Spearmans ρ). Wie in der Präregistrierung festgehalten, wurden sie daher durch Mittelung zu einer neuen Variable, der *Gesamtzufriedenheit* zusammengefasst. Tabelle 22 fasst die Rangkorrelationen zwischen den Skalenwerten und der Gesamtzufriedenheit nach Spearman zusammen. Pearsons Produkt-Moment-Korrelationen werden ergänzend berichtet. Es zeigten sich für alle Skalen signifikante Korrelationen, die größer als .30 waren, was nach Cohen (1992) einem mittleren Effekt entsprach. Diese Ergebnisse sprachen demnach für eine hohe konvergente Validität der Skalen (im Einklang mit Hypothese 5.2.1). Die Ergebnisse bezüglich der neu entwickelten Sicherheitsskalen sind auch hier in Anhang D (Tabelle D5) abgebildet.

Tabelle 22

Korrelationen zwischen den Skalenwerten und der Gesamtzufriedenheit

Skalenwerte [a]	Gesamtzufriedenheit	
	Pearsons *r*	Spearmans ρ
Steuerbarkeit	.51***	.60***
Vertrauen	.41***	.46***
Effizienz	.68***	.71***
Übersichtlichkeit	.67***	.68***
Effektivität	.63***	.67***

Anmerkung. $N = 173$. $N_{\text{Effektivität}} = 172$. Aufgrund von Rangbindungen konnten p-Werte für Spearmans ρ nicht exakt berechnet werden.

[a] Durchschnitt der beantworteten Items jeder Skala.

* $p < .05$. *** $p < .001$.

Hypothese 5.2 postulierte, dass Produkte verschiedener Hersteller sich in ihrer durchschnittlichen Bewertung auf den Skalen des UEQ+ (Schrepp & Thomaschewski, 2019a) sowie den neu entwickelten Skalen unterscheiden. Diese Fragestellung sollte wie Hypothese 2 mit einer MANOVA getestet werden. Als unabhängige Variable diente in diesem Fall der Hersteller der Medizinprodukte, als abhängige Variablen die zuvor ermittelten Skalenwerte der fünf relevantesten Skalen. Sechs Antwortkategorien wurden für den Hersteller des Referenzprodukts vorgegeben, wovon unter allen Teilnehmenden aber nur vier vorkamen. Diese vier Hersteller dienten als Faktorstufen. Die Namen der Hersteller werden aus Gründen der Vertraulichkeit nicht genannt. Ebenso werden aus diesem Grund keine konkreten Skalenmittelwerte unterteilt nach Herstellern berichtet. Hersteller 1 wurde von 112 Personen, Hersteller 2 von 15 Personen, Hersteller 3 von 18 und Hersteller 4 von vier Personen angegeben. Die Gruppen der unabhängigen Variablen waren somit sehr unterschiedlich groß. Für Hersteller 4 lagen weniger Beobachtungen als abhängige Variablen vor. Die Analyse wurde trotzdem vorgenommen, die Ergebnisse sind jedoch mit Vorsicht zu interpretieren. Die Überprüfung von Normalverteilung anhand von Histogrammen, Q-Q-Plot und Mardia-Test deutete auf mangelnde Erfüllung dieser Voraussetzung hin. Auch hier erwiesen sich einige Variablen als linksschief verteilt. Lineare Beziehungen ließen sich aufgrund der geringen Substichproben teils schwer beurteilen. Multikollinearität lag nicht vor. Weiterhin konnte Homoskedastizität mithilfe des Levene-Tests festgestellt werden. Der Boxsche M-Test ließ sich aufgrund der geringen Proband*innenanzahl in manchen Gruppen nicht durchführen. Wie bereits für Hypothese 2 wurde auf Basis der Voraussetzungsprüfung auch hier entschieden, einen nichtparametrischen multivariaten Kruskal-Wallis-Test zum Vergleich zu berechnen.

Die MANOVA ergab signifikante Unterschied zwischen den Herstellern für die kombinierten abhängigen Variablen, $F(3, 144) = 1.80$, $p = .033$, partielles $\eta^2 = .06$, Wilk's $\Lambda = 0.83$. Alle ausgegebenen Teststatistiken stimmten darin überein, dass ein signifikanter Unterschied vorlag. Der multivariate Kruskal-Wallis-Test lieferte gleichfalls ein signifikantes Ergebnis, $Q^2(15) = 26.67$, $p = .032$. Die Ergebnisse konnten wegen der ungleichen und geringen Gruppengrößen nur unter Vorbehalt interpretiert werden, jedoch legten sie zumindest nahe, dass es signifikante Unterschiede in der mittleren Skalenbewertung zwischen den Herstellern gab. Dieses Ergebnis entsprach der Annahme von Hypothese 5.2.2. Es ist indessen anzumerken, dass bei Berechnung der MANOVA inklusive der optionalen Sicherheitsskalen keine signifikanten Unterschiede zwischen verschiedenen Hersteller festzustellen waren (s. Anhang D Tabelle D5). Da es sich bei den

Sicherheitsskalen nur um optionale Zusatzskalen handelte, wurde dieses Ergebnis jedoch nicht als maßgeblich betrachtet. Bezüglich der Validität der Skalen ist festzuhalten, dass konvergente Validität aufgrund einer signifikanten Korrelation zur Gesamtzufriedenheit für alle produktübergreifend wichtigen Skalen gegeben war (im Einklang mit Hypothese 5.2.1). Produkte verschiedener Hersteller unterschieden sich zudem signifikant in ihrer durchschnittlichen Bewertung, was unter Vorbehalt auf diskriminante Validität hindeutete (im Einklang mit Hypothese 5.2.2).

3.5.6 Explorative Analysen.

Ein Punkt, dem bisher keine Beachtung geschenkt wurde, ist die auffällig geringe Wichtigkeitsbewertung der Skala Sicherheit von Befragten der Business Line D&A. Für D&A belegte sie den 21. Rang, für alle anderen Business Lines lag sie mit Rängen von 7 bis 10 deutlich höher. Diese Bewertung war wahrscheinlich darauf zurückzuführen, dass die Beschreibung der Skala nicht ideal auf Reading- & Reporting-Software anwendbar war. Sie wurde von Müßig (2022) übernommen und lautete „Das Produkt erzeugt ein körperliches Sicherheitsgefühl bei der Nutzung“. Eine Änderung der Beschreibung, sodass sie auf Software-Produkte anwendbar gewesen wäre, hätte eine deutliche Veränderung der Bedeutung mit sich gebracht. Um sich möglichst an das Material von Müßigs Vorstudie zu halten, wurde deshalb keine Änderung der Beschreibung vorgenommen. Allerdings nahm an der Vorstudie nur eine Person der Business Line D&A teil, wodurch es hier zu einer Verzerrung in der Wichtigkeitsbewertung der Skala Sicherheit gekommen sein könnte. Die mittlere Wichtigkeitsbewertung in Müßigs Studie betrug 6.61 (SD = 0.74), wohingegen sie in der vorliegenden Replikationsstudie nur bei 6.06 (SD = 1.19) lag. Es wurde aus diesem Grund explorativ berechnet, welche durchschnittlichen Wichtigkeitsbewertung sich ohne die 19 Proband*innen der Business Line D&A ergeben hätte. Das Ergebnis zeigte, dass die Skala Sicherheit so eine durchschnittliche Bewertung von 6.25 (SD = 0.96, n =154) erreicht hätte. Zu den sechs produktübergreifend wichtigsten Skalen hätte sie mit dem 9. Rang auch ohne Teilnehmende der Business Line D&A nicht gehört.

4 Diskussion

In diesem Abschnitt werden die Ergebnisse zusammengefasst und diskutiert. Es wird auf Abweichungen von der Präregistrierung, Limitationen und Implikationen eingegangen. Abschließend wird ein Ausblick gegeben.

4.1 Zusammenfassung der Ergebnisse

Ziel der vorliegenden Arbeit war die Entwicklung eines UX-Fragebogens speziell für Medizinprodukte, welche von medizinischem Personal bedient werden. Die Ergebnisse der letzten Vorstudie (Müßig, 2022) gaben bereits Aufschluss darüber, welche sechs Skalen besonders relevant sein könnten. Einige Skalen (Steuerbarkeit, Vertrauen, Effizienz und Übersichtlichkeit) konnten aus dem User Experience Questionnaire + (UEQ+) (Schrepp & Thomaschewski, 2019a) übernommen werden. Die Skalen Sicherheit und Effektivität wurden in der Vorstudie ebenfalls als sehr wichtig für Medizinprodukte eingeschätzt. Für diese Skalen existierten allerdings noch keine Items, da sie nicht Bestandteil des UEQ+ sind, sondern auf Basis von Literaturrecherche ergänzt wurden. Im Rahmen von internen Expert*innen-Workshops wurden daher Items für Sicherheit und Effektivität entwickelt und konsolidiert. In einer neuen externen Online-Studie im Rahmen dieser Arbeit wurden die entwickelten Items sowie die Struktur des Fragebogens an medizinischem Personal überprüft. Außerdem wurde untersucht, ob sich die Ergebnisse der letzten Vorstudie (Müßig, 2022) replizieren ließen. Die Ergebnisse konnten zum Teil repliziert werden, fünf der sechs wichtigsten Skalen befanden sich auch in der Replikationsstudie unter den sechs produktübergreifend wichtigsten Skalen. Die Skala Sicherheit wurde jedoch durch die Skala Nützlichkeit ersetzt. Im Einklang mit den Ergebnissen der Vorstudie erwies sich eine produktübergreifende Version des UX-Fragebogens für Medizinprodukte als akzeptabel. Für den finalen UX-Fragebogen für Medizinprodukte wurden aufgrund der größeren Stichprobe die sechs produktübergreifend wichtigsten Skalen der Replikationsstudie herangezogen. Für verschiedene Medizinprodukte beziehungsweise Fragestellungen werden optionale Zusatzskalen empfohlen. Die Skala Sicherheit sollte als eine der optionalen Zusatzskalen entwickelt werden, jedoch ließ sich basierend auf Faktorenanalysen keine Eindimensionalität annehmen. Die entwickelten Sicherheitsitems wurden in zwei Skalen, Allgemeine Sicherheit sowie Hardware-Sicherheit unterteilt. Für beide Varianten fanden sich vier geeignete Items im Format des UEQ+. Die neu entwickelten Effektivitätsitems erwiesen sich als eindimensional und es konnten auch für diese Skala vier geeignete Items gefunden werden. Der globale Modell-Fit wurde anhand eines Strukturgleichungsmodells nachgewiesen. Die lokale Modellgüte erwies sich als

akzeptabel, auch wenn einige Fehlspezifikationen vorlagen. Untersuchungen der Reliabilität ergaben eine hohe interne Konsistenz aller Skalen. Die Interrater-Reliabilität bezüglich eines ausgewählten Referenzprodukts fiel dagegen gering aus. Absolute Übereinstimmung der Teilnehmenden im Hinblick auf das genannte Referenzprodukt ließ sich ebenfalls nicht für alle Skalen nachweisen. Konvergente Validität war für die Skalen der produktübergreifenden Version des UX-Fragebogens gegeben. Produkte verschiedener Hersteller unterschieden sich signifikant in ihrer durchschnittlichen Bewertung, was ein Hinweis auf diskriminante Validität des UX-Fragebogens war, aufgrund der deutlich abweichenden Gruppengrößen jedoch mit Vorsicht interpretiert werden sollte.

4.2 Diskussion der Ergebnisse

Die Diskussion der Ergebnisse ist wie der Ergebnisteil untergliedert in Replikationsergebnisse (Hypothesen 1 und 2), Ergebnisse der Fragebogenentwicklung (Hypothesen 3 und 4) und die Betrachtung der psychometrischen Gütekriterien (Hypothese 5). Zusätzlich wird zu Beginn des Kapitels auf die Itementwicklung im Rahmen der Workshops eingegangen.

4.2.1. Diskussion der Itementwicklung.

Für die Entwicklung von Items für die neuen Skalen Sicherheit und Effektivität, welche im Verlauf der UX-Fragebogenentwicklung zu den Skalen Allgemeine Sicherheit, Hardware-Sicherheit und Effektivität wurden, wurden die Definitionen nach der ISO (International Organization for Standardization, 2014, 2015, 2018) herangezogen (s. Kapitel 2.1 – Workshops zur Itemgenerierung). Alternativ hätte man direkt auf Müßigs (2022) Skalenbeschreibungen aus der Vorstudie zurückgreifen können, welche schließlich zu einer hohen Wichtigkeitsbewertung dieser Skalen geführt hatten (s. Kapitel 3.2.2 – Variablen der Replikation). Müßig hatte die Skalenbeschreibungen aus Siemens Healthineers internen UX-Heuristiken und anderen Fragebögen, wie dem PSSUQ (Lewis, 1995) abgeleitet. Auf Basis von Müßigs Skalenbeschreibung der Skala Sicherheit hätte man die Business Line D&A für die Entwicklung der Sicherheitsitems auch ausschließen können, da ein körperliches Sicherheitsgefühl für Reading- und Reporting-Software nicht relevant ist. Die Definition der ISO beinhaltet dagegen, dass Risiken aufgrund von Anwendungsfehlern ausgeschlossen sind (International Organization for Standardization, 2015). Wenn man diese Definition heranzieht, fällt unter Sicherheit bei Medizinprodukten auch, dass anhand von Software keine falschen Diagnosen gestellt werden, die ebenfalls ein Risiko darstellen, wenn auch nicht unmittelbar. Um ein einheitliches Verständnis von Sicherheit und Effektivität sicherzustellen, wurden in den Workshops zur

Itemgenerierung die Definitionen aufgegriffen und diskutiert. Im Rahmen der Workshops wurde deutlich, dass Teilnehmende bei Sicherheit im Einklang mit der Definition nach der ISO nicht nur an körperliche Sicherheit dachten. Auch für die Skala Effektivität weicht die Definition nach der ISO von Müßigs Skalenbeschreibung ab. Nach der ISO besteht Effektivität darin, dass Ziele vollständig und genau erreicht werden (International Organization for Standardization, 2018), während bei Müßig die Unterstützung beim Erreichen der Ziele im Vordergrund steht. Die Abweichung ist hier aber im Vergleich zu Sicherheit geringfügig und beide Effektivitätsdefinitionen sowie genannte Aspekte der Teilnehmenden passen zu allen Business Lines.

Um nicht zu stark von Müßigs Vorstudie abzuweichen, wurden die Skalenbeschreibungen zur Abfrage der Wichtigkeit der Skalen für die Replikationsstudie beibehalten. Für die Itementwicklung wurden allerdings die Definitionen nach der ISO als wesentlich betrachtet, da sie erstens international anerkannt und zweitens (vor allem im Fall von Sicherheit) allgemein gehalten sind, sodass es leichter möglich war, Items auf dieser Grundlage zu generieren. Dieses Vorgehen widerspricht sich in gewisser Weise, es konnte allerdings keine optimale Lösung gefunden werden, da in jedem Fall ein Kompromiss nötig gewesen wäre. Sicherheit wurde in der Replikationsstudie anhand von Müßigs Beschreibung von Teilnehmenden der Business Line D&A als wenig wichtig eingestuft, nach der Definition der ISO spielt sie aber durchaus auch für diese Business Line eine Rolle. Die Skala Allgemeine Sicherheit wird deshalb trotzdem als optionale Zusatzskala in Abhängigkeit von der Fragestellung für alle Business Lines empfohlen.

Abgesehen von den Skalendefinitionen kann auch der Ablauf der Workshops hinterfragt werden. Die Items wurden, anders als bei der Entwicklung der Items des UEQ+ (Schrepp & Thomaschewski, 2019a), zunächst nicht im Format des UEQ+ generiert. Stattdessen wurde ohne die Vorgabe eines Einleitungssatzes frei diskutiert und es wurden Schlagworte, Stichpunkte und Sätze gesammelt. Bei der Überführung der Items in semantische Differentiale brachte dieses Vorgehen den Nachteil mit sich, dass Items umformuliert, gekürzt oder sogar ausgeschlossen werden mussten. Allerdings wurden im Rahmen der Workshops trotz der wenigen Vorgaben nicht sehr viele Ideen für Items gesammelt. Dazu trug vermutlich die Tatsache bei, dass anders als bei der Entwicklung des UEQ+ die Itemgenerierung in dieser Arbeit mit Ausrichtung auf eine spezielle Art von Produkten (Medizinprodukten) erfolgte. Bei Vorgabe eines Einleitungssatzes wäre das Brainstorming zusätzlich erschwert worden.

4.2.1 Diskussion der Replikationsergebnisse.

Unter den sechs produktübergreifend wichtigsten Skalen befand sich in der Replikationsstudie Nützlichkeit anstelle von Sicherheit. Eine Erklärung dafür könnte sein, dass die Beschreibung der Skala (körperliches Sicherheitsgefühl) nicht zu Softwareanwendungen passte und daher Teilnehmende der Business Line D&A sie entsprechend als unwichtiger bewerteten. An Müßigs Vorstudie 2021 (Müßig, 2022) nahm nur eine Person dieser Business Line teil, weshalb die produktübergreifende Wichtigkeitsbewertung der Skala höher ausgefallen sein könnte als in der Replikationsstudie, in die 19 Teilnehmende der Business Line einbezogen wurden. Man hätte die Beschreibung der Skala entsprechend breiter fassen können, was allerdings eine wesentliche Abweichung von Müßigs Vorstudie dargestellt hätte. Es wurde deshalb explorativ geprüft, ob die Skala Sicherheit ohne Teilnehmende der Business Line D&A unter die sechs wichtigsten Skalen gefallen wäre, was jedoch abgelehnt wurde. Die größere Anzahl an Teilnehmenden der Business Line D&A ist somit nicht allein ausschlaggebend dafür, dass sich Sicherheit nicht mehr unter den sechs wichtigsten Skalen befindet. Was an dieser Stelle berücksichtigt werden sollte, ist dass der Großteil der Skalen insgesamt als sehr wichtig bewertet wurde, was eine Differenzierung der Rangplätze erschwerte. In einigen Fällen konnte erst auf die zweite Nachkommastelle ein Rangplatz zugewiesen werden und sogar dann teilten sich manche Skalen denselben Rang. Die Skala Sicherheit wurde auch in der Replikationsstudie als sehr wichtig bewertet, lag aber wie berichtet nicht mehr unter den Top 6 Skalen. Man könnte deshalb auch argumentieren, dass generell mehr als sechs Skalen abgefragt werden sollten, da sie alle als wichtig empfunden werden. Das wäre jedoch in der Praxis nicht ökonomisch (Schrepp & Thomaschewski, 2020).

Aufgrund der geringen Unterschiede in den Wichtigkeitsbewertungen wäre auch die umständlichere Handhabung von Business Line-spezifischen Versionen des UX-Fragebogens nicht gerechtfertigt, zumal auch die Post-hoc Tests im Wesentlichen gegen Business Line-spezifische Versionen sprachen. Die beste Lösung scheint deshalb die Empfehlung einer produktübergreifenden Version des UX-Fragebogens für Medizinprodukte bestehend aus sechs Skalen und je nach zeitlichen Ressourcen, Business Line und Fragestellung die Abfrage von optionalen Zusatzskalen zu sein. Diese Empfehlung leitete auch Müßig (2022) aus ihren Ergebnissen ab. Für die Auswahl der finalen produktübergreifend wichtigsten Skalen sowie der optionalen Zusatzskalen wurde den Ergebnissen der Replikationsstudie Vorrang gewährt, da für sie eine mehr als doppelt so große Anzahl an Personen befragt wurde. Ideal wäre es zwar gewesen, die Daten von Müßigs Vorstudie und

der Replikationsstudie zusammenzuführen und gemeinsam auszuwerten, es ergibt sich hier jedoch das Problem, dass manche Personen eventuell an beiden Studien teilgenommen haben und dann überrepräsentiert wären. Aufgrund der anonymisierten Durchführung der Studien ist das jedoch nicht mehr nachvollziehbar.

Was die optionalen Zusatzskalen betrifft, existieren für die Skalen Durchschaubarkeit und Inhaltsseriosität als Teil des UEQ+ bereits Items. Für die Sicherheitsskalen wurden im Rahmen dieser Arbeit Items entwickelt. Sie könnten insbesondere für den Prozess des Usability-Engineerings in Bezug auf Medizinprodukte interessant sein, bei dem die Gewährleistung von Sicherheit von besonderem Interesse ist (International Organization for Standardization, 2015). Die optionale Zusatzskala Erwartungskonformität hat im Gegensatz zu den bisher genannten optionalen Zusatzskalen noch keine Items. Ein Item der Skala Steuerbarkeit lautet allerdings „nicht erwartungskonform vs. erwartungskonform“. Es ist deshalb fraglich, ob Erwartungskonformität nicht bereits durch die Skala Steuerbarkeit abgedeckt wird. Würde man Items für die Skala Erwartungskonformität entwickeln, wäre anzunehmen, dass die Skalen sich inhaltlich stark ähneln würden. Aus diesem Grund ist die Entwicklung der optionalen Zusatzskala Erwartungskonformität nicht ratsam. Laut Schrepp und Thomaschewski (2019c) ist auch die Skala Inhaltsseriosität „eine Spezialisierung von Vertrauen auf eine bestimmte Art von Anwendung“ (S.152). Sie argumentieren, dass Inhaltsseriosität sich für die Anwendung auf beispielsweise Webseiten eignet, während für Online-Banking-Applikationen eher die Skala Vertrauen geeignet ist. Im Hinblick auf professionelle, seriöse Anwendungen wird Inhaltsseriosität als solche vermutlich weniger in Frage gestellt. Nachdem auch für Medizinprodukte eine gewisse Seriosität vorausgesetzt werden kann, könnte auf diese optionale Zusatzskala anhand dieser Argumentation ebenfalls verzichtet werden, weil sie durch die Skala Vertrauen bereits abgedeckt wird. Die optionale Verwendung der Skala wird daher nicht empfohlen.

Zusammengefasst gibt es nun die produktübergreifende Empfehlung des UX-Fragebogens für Medizinprodukte bestehend aus den sechs Skalen Steuerbarkeit, Vertrauen, Effizienz, Übersichtlichkeit, Nützlichkeit und Effektivität. Als optionale Zusatzskalen werden je nach Fragestellung für alle Business Lines die Skala Allgemeine Sicherheit sowie für alle Business Lines bis auf D&A die Skala Hardware-Sicherheit empfohlen. Die optionale Zusatzskala Durchschaubarkeit wird des Weiteren für die Business Lines D&A und AT/AT-SU nahegelegt, sofern ausreichende zeitliche Ressourcen zur Verfügung stehen.

4.2.2 Diskussion der Ergebnisse zur Fragebogenentwicklung.

Im Rahmen der eigentlichen Fragebogenentwicklung sollten die generierten Items für die Skalen Sicherheit und Effektivität sowie die Struktur des UX-Fragebogens evaluiert werden. Eindimensionalität der Sicherheitsitems ist nicht gegeben. Nach der Entwicklung der Items in den Workshops kristallisierte sich bereits heraus, dass manche der Items nur auf Hardware anwendbar sind. Auf Basis der Faktorenanalysen, die für eine Unterteilung der allgemeinen und Hardware-Sicherheitsitems sprachen, wurde die Möglichkeit ausgeschlossen, eine Sicherheitsskala mit optionalen Hardware-Items zu entwerfen, was zunächst zur Debatte stand. Dass die Sicherheitsitems nicht eindimensional sind, ist also einerseits auf die Hardware-spezifischen Items zurückzuführen. Andererseits stand weiterhin die Unterteilung der allgemeinen Sicherheitsitems in zwei Skalen im Raum. Aufgrund der widersprüchlichen Ergebnisse musste priorisiert werden, welche Kriterien die Extraktionsentscheidung bestimmen sollten. Es wurde trotz des schlechteren Fits zugunsten der inhaltlichen und theoretischen Plausibilität ein Faktor für die allgemeinen Sicherheitsitems extrahiert. Inhaltlich deuteten die Faktorladungen bei einer Unterteilung der allgemeinen Sicherheitsitems darauf hin, dass ein Faktor tendenziell die Erkennung und Kontrolle von Anwendungsfehlern und Risiken zusammenfasst, ein möglicher zweiter Faktor deren Behebung beziehungsweise Beeinflussung. Allerdings ist diese Abgrenzung nicht eindeutig möglich, denn eines der vier Items mit Ladung auf den zweiten Faktor der allgemeinen Sicherheitsitems zeigte eine Doppelladung. Die Definition der ISO im Hinblick auf Sicherheit bei Medizinprodukten sagt aus, dass ein inakzeptables Risiko durch Nutzungsfehler nicht gegeben sein darf (2014). Wenn man diese Definition aufspaltet, beinhaltet das sowohl die Erkennung von Fehlern, deren Vermeidung, Kontrolle durch das System sowie ihre Behebung. Man könnte Sicherheit bei Medizinprodukten also durchaus in verschiedene Facetten untergliedern, jedoch bedingen sich diese Facetten zum Teil. Fehler können nur behoben oder aufgehalten werden, wenn sie zuvor erkannt wurden. Theoretisch und aus Praxisperspektive betrachtet, ist demnach eine Gesamtskala für die allgemeinen Sicherheitsitems sinnvoll und ökonomisch, solange die empirischen Ergebnisse dem nicht widersprechen. Sowohl die Ergebnisse der Hauptkomponentenanalyse als auch die Verläufe der Eigenwerte in den Screeplots sprechen aber dafür, dass die Extraktion einer Skala aus den allgemeinen Sicherheitsitems auch empirisch begründet ist. Die Hinzunahme eines Faktors würde keine signifikante Verbesserung des Fits bewirken. Nichtsdestotrotz ist zu beachten, dass der Fit nicht ideal ist und eine Untersuchung der Sicherheitsskalen, insbesondere der Skala Allgemeine Sicherheit, anhand

neuer Daten sinnvoll wäre. Da es sich bei den Skalen Allgemeine Sicherheit und Hardware-Sicherheit allerdings um optionale Zusatzskalen handelt, beeinträchtigt diese Tatsache den Einsatz des produktübergreifenden UX-Fragebogens nicht.

Die Evaluation der Effektivitätsitems entwickelte sich im Einklang mit der Hypothese (3.2), sodass eine eindimensionale Skala mit vier geeigneten Items aus ihr hervorgingen. Sowohl die allgemeinen und Hardware-Sicherheitsitems als auch die Effektivitätsitems weisen jedoch vor allem untereinander, aber auch insgesamt ähnliche Itemschwierigkeiten auf. Die Sicherheitsitems liegen in einem mittleren bis leichten Schwierigkeitsbereich, während die Effektivitätsitems alle eher leicht sind. Eine breite Streuung der Schwierigkeit ist für eine gute Differenzierung der User Experience zwischen verschiedenen Produkten eigentlich wünschenswert (Döring & Bortz, 2016). Dass die Items alle eher leicht ausfallen, könnte damit zusammenhängen, dass manche positiven Pole als zu neutral wahrgenommen werden. Ob die „Ergebnisse, die sich mit dem Produkt/System erzielen lassen“ beispielsweise als „bedarfsgerecht“ oder „adäquat“ empfunden werden (Items der Skala Effektivität), hängt außerdem von dem oder der Anwender*in ab. Ein*e erfahrene*r Radiolog*in wird beispielsweise wahrscheinlich auch bei einem wenig effektiven Produkt in der Lage sein, eine Diagnose zu stellen und deshalb angeben, dass die Ergebnisse „bedarfsgerecht“ sind, obwohl die Diagnosestellung mit anderen Produkten eventuell eindeutiger möglich wäre. Die Mitte der Skala könnte dann eventuell so interpretiert werden, dass eine Diagnosestellung gar nicht möglich ist, womit der negative Pol des Items unter Umständen sogar als Vorliegen eines Schadens verstanden werden könnte. Diese denkbare Interpretation mancher Items könnte durch die numerischen Bezeichnungen der Ratingstufen von -3 bis +3 zusätzlich verstärkt worden sein. Bisherige Forschung zeigt, dass die Auswahl negativer numerischer Bezeichnungen bei bipolaren Skalen vermieden wird (Kunz, 2015; Schwarz et al., 1991). Schwarz et al. (1991) liefern als mögliche Erklärung für dieses Phänomen, dass negative Werte andeuten, dass positive Merkmale nicht nur ausbleiben, sondern negative Merkmale explizit gegeben sind, was im Einklang mit oben genanntem Beispiel steht. Dass die Items eher als erfüllt bewertet werden, ist nicht nur für die Skala Effektivität zu beobachten. Auch für andere Skalen, zum Beispiel für die Skala Vertrauen, lagen die Itemmittelwerte über dem Skalenmittelwert von 4. Wie die Differenzierungsfähigkeit der Items verbessert werden könnte, wird später aufgegriffen (s. Kapitel 4.6.2 – Verbesserung der Differenzierungsfähigkeit der Items). Es wäre allerdings auch denkbar, dass die Referenzprodukte der Studie tatsächlich eine gute User Experience erzeugten und deshalb positiv bewertet

wurden, immerhin wurden vor allem neue Produkte bewertet, die im Durchschnitt im Jahr 2016 hergestellt wurden. Es ist außerdem anzumerken, dass die Schwierigkeit der Items zwar nicht optimal ist, nichtsdestotrotz aber in einem akzeptablen Bereich liegt. Ebenso sind die Items angesichts der hohen Ladungen trennscharf und die Skalen infolge der faktorenanalytischen Konstruktion homogen, was für die Eignung der Items spricht.

Die erwartete Struktur des produktübergreifenden UX-Fragebogens lässt sich anhand des Strukturgleichungsmodells als gültig annehmen, wobei beachtet werden muss, dass die Skala Nützlichkeit nicht enthalten war. Der χ^2-Test wurde zwar signifikant und deutete damit auf eine schlechte Passung des angenommenen Modells zu den beobachteten Daten hin, jedoch ist diese Signifikanz unter Umständen auf die große Stichprobe zurückzuführen, denn Fit-Indizes belegen eine akzeptable globale Modellpassung. Bei Betrachtung der lokalen Modellgüte stellten sich indessen einige Modifikationsindizes als auffällig heraus. Bühner (2011) betont jedoch, dass die praktische Bedeutsamkeit von Modifikationen „im jeweiligen Forschungskontext betrachtet werden“ (S. 558) muss. Es ist nur dann sinnvoll, Veränderungen am Modell vorzunehmen, wenn diese auch theoretisch begründet sind. Schrepp und Thomaschewski (2019c) räumen selbst ein, dass UX-Skalen häufig korrelieren. Es ist somit schwierig, Items zu entwerfen, die keine Ladungen auf mehrere Faktoren aufweisen und nicht über Skalen hinweg korreliert sind. Wie von Schrepp und Rummel (2018) empfohlen, wurde für die Entwicklung der Skalen des UEQ+ sowie der neuen Skalen zuerst ermittelt, welche UX-Aspekte aus Design- beziehungsweise Anwendersicht eine praktisch relevante Rolle spielen. Für diese Skalen wurden anschließend Items generiert. Die Zugehörigkeit der Items zu den Skalen ist damit theoretisch begründet und nicht rein explorativ faktorenanalytisch abgeleitet. Eine Umstellung der Struktur des UX-Fragebogens aufgrund der Modifikationsindizes ist deshalb nicht sinnvoll.

Anhand des Strukturgleichungsmodells war ferner erkennbar, dass die Skalen, beziehungsweise die latenten Variablen, hoch korrelierten. Man könnte sich deshalb fragen, ob es wirklich nötig ist, alle Skalen zu erheben. Je weniger Skalen erhoben werden, desto ökonomischer ist der Fragebogen. Vor allem die Skala Effizienz zeigt hohe Zusammenhänge zu anderen Skalen. Im Abschnitt zur Itemgenerierung (Kapitel 2.1) wurde bereits deutlich, dass es Teilnehmenden insbesondere schwerfiel, die Skalen Effektivität und Effizienz abzugrenzen. Anhand der Items lassen sich die Skalen inhaltlich allerdings dadurch abgrenzen, dass Effizienz als einzige Skala den zeitlichen Aufwand erfragt (s. Anhang E). Es könnte zudem sein, dass die Skalen je nach Produkt unterschiedlich

stark zusammenhängen, wie in Kapitel 1.5.1 (Anwendbarkeit der psychologischen Testtheorie) ausgeführt wurde. Aus diesen Gründen ist es nicht sinnvoll, Skalen voreilig auszuschließen. Sollte sich anhand weiterer Daten jedoch zeigen, dass Korrelationen zwischen Skalen, beispielsweise jene der Skala Effizienz zu anderen Skalen, stabil sehr hoch ausfallen, könnte man darüber nachdenken den produktübergreifenden UX-Fragebogen um die Skala Effizienz zu kürzen. Da insbesondere die Begriffe Effizienz und Effektivität schwer abzugrenzen waren, wurde am Ende des Projekts auf Grundlage interner Absprachen entschieden, die Skala Effektivität für eine erleichterte Trennung in der Praxis in *Ergebnisqualität* umzubenennen. Die Skalen Allgemeine Sicherheit und Hardware-Sicherheit wurden ebenfalls umbenannt in *Risikohandhabung* und *Physische Sicherheit*, um eine leichte Unterscheidung zu ermöglichen. Für die vorliegende Arbeit wurden aus Gründen der Einheitlichkeit und Verständlichkeit konsistent die bisherigen Bezeichnungen Effektivität und Allgemeine beziehungsweise Hardware-Sicherheit herangezogen. Der finale UX-Fragebogen für Medizinprodukte insgesamt wird intern als *User Experience Questionnaire for Medical Devices* (UEQ MD) bezeichnet.

4.2.3 Diskussion der Ergebnisse zu psychometrischen Gütekriterien.

Im Hinblick auf die Reliabilität besteht für alle Skalen, jene des produktübergreifenden UX-Fragebogens und auch die zusätzlich entwickelten Sicherheitsskalen, eine hohe interne Konsistenz. Die Interrater-Reliabilität im Sinne der Konsistenz wurde für ein ausgewähltes Röntgengerät berechnet. Für die produktübergreifenden Skalen fiel sie gering aus. Dieses Ergebnis spricht für wenig Einigkeit unter den Befragten im Hinblick auf das Referenzprodukt. Der erhaltene Wert muss allerdings nicht zwangsläufig auf Uneinigkeit zurückzuführen sein, er könnte auch durch die geringen Unterschiede in der mittleren Bewertung der Skalen bedingt sein. Darauf deutet auch die höhere Interrater-Reliabilität bei Hinzunahme der neu entwickelten Sicherheitsskalen hin, denn die mittlere Bewertung der Sicherheitsskalen fiel, wie die deskriptive Statistik bezüglich der Items erkennen lässt, schlechter aus als die der anderen Skalen. Zudem ist zu beachten, dass für die Berechnung nur acht Personen einbezogen werden konnten und das Konfidenzintervall einen großen Wertebereich einschloss. Der berechnete Wert der Interrater-Reliabilität ist somit sehr unsicher.

Die absolute Übereinstimmung zwischen den Befragten kann ebenfalls nicht eindeutig beurteilt werden. Geht man nach grafischen Anhaltspunkten, zeigt sich für die Skalen Effizienz, Effektivität und Vertrauen eine deutlich höhere Übereinstimmung als für andere Skalen. Möglicherweise ist das darauf zurückzuführen, dass manche Skalen

stärker individuellen Einflüssen unterliegen als andere. Die Skala Vertrauen beispielsweise zielt auf die Erhebung von Datensicherheit ab, die sich unter Umständen objektiver einschätzen lässt als die Übersichtlichkeit der Benutzeroberfläche. Wenn eine Person bereits Erfahrung mit ähnlichen Benutzeroberflächen gesammelt hat, könnte sie die Übersichtlichkeit als hoch einstufen, eine Person mit wenig Berufserfahrung oder Erfahrung mit anderen Benutzeroberflächen schätzt diese dagegen eventuell schlechter ein. Auch hier besteht, wie bereits an mehreren Stellen erwähnt, eine wesentliche Herausforderung darin, dass die UX sich aus verschiedenen Einflüssen zusammensetzt. Wie im Anfangsteil der Arbeit ausgeführt (1.5.1 – Anwendbarkeit der psychologischen Testtheorie), sollten deshalb immer Skalenmittelwerte über eine repräsentative Stichprobe betrachtet werden, um individuelle Einflussfaktoren möglichst auszugleichen (Rummel & Schrepp, 2018).

Die Korrelationen der Skalen mit der Gesamtzufriedenheit bezüglich des Produkts lassen auf hohe konvergente Validitäten schließen. Nur für die Skala Hardware-Sicherheit, die aber ohnehin kein Bestandteil des produktübergreifenden UX-Fragebogens ist, zeigte sich eine deutlich geringere Korrelation zur Gesamtzufriedenheit im Vergleich zu den anderen Skalen. Eine denkbare Erklärung dafür könnte sein, dass die Hardware-Items sich auf den Ausschluss von Gefahren beziehen. Positive Pole der Skala lauten zum Beispiel „nicht gesundheitsgefährdend“ oder „nicht schädigend“. Der reine Ausschluss von Gefahren bringt aber unter Umständen keine Zufriedenheit mit dem Produkt mit sich, sondern vielmehr die Abwesenheit von Unzufriedenheit. Die Gesamtzufriedenheit ist daher eventuell kein geeignetes Kriterium zur Überprüfung der konvergenten Validität dieser Skala. Die Wahrnehmung einer bedenkenlosen Anwendbarkeit des Produkts wäre hier womöglich ein geeigneteres Kriterium gewesen.

Aufgrund der geringen Anzahl an Personen, die das gleiche Referenzprodukt bewertet hatten, wurde die diskriminante Validität nicht anhand konkreter Modelle, sondern anhand von Herstellern untersucht. Die Bewertungen von Produkten verschiedener Hersteller unterschieden sich signifikant, was jedoch wegen der teils sehr geringen Anzahl an Teilnehmenden pro Faktorstufe mit Vorsicht zu interpretieren ist. Insgesamt sollte dieses Ergebnis mit Vorbehalt betrachtet werden, denn für jeden Hersteller wurde eine Vielzahl verschiedener Produkte zusammengenommen, um überhaupt eine sinnvolle statistische Auswertung möglich zu machen. Nichtsdestotrotz kann das Ergebnis zumindest als Anhaltspunkt für diskriminante Validität des produktübergreifenden UX-Fragebogens gewertet werden.

Es ist insgesamt zu beachten, dass für die Untersuchung der psychometrischen Gütekriterien, wie auch schon zuvor für die Überprüfung der Struktur des UX-Fragebogens, die Skala Nützlichkeit nicht einbezogen werden konnte, da für sie keine Items erhoben wurden. Die Güte des UX-Fragebogens konnte somit nicht vollständig überprüft werden. Nachdem die Ergebnisse nun ausführlich dargestellt und diskutiert wurden, werden als Nächstes Abweichungen von der Präregistrierung aufgezeigt.

4.3 Abweichungen von der Präregistrierung

Eine erste Präregistrierung wurde drei Wochen vor der Datenerhebung eingereicht. In einem Update der Präregistrierung wurden Änderungen festgehalten. Die Datenerhebung lief zu diesem Zeitpunkt seit drei Wochen und es wurden bereits 32 Proband*innen rekrutiert, jedoch fand noch kein Einblick in die Daten statt. Was im Update der Präregistrierung beziehungsweise danach geändert wurde, wird im Folgenden geschildert.

4.3.1 Formale Abweichungen sowie Änderungen des Studienfragebogens.

Zunächst wurde die Reihenfolge der ersten beiden Hypothesen im Update der Präregistrierung geändert. Ursprünglich postulierte die erste Hypothese, dass die Business Lines sich nicht signifikant in den wichtigsten UX-Skalen unterscheiden und die zweite Hypothese, dass sich die gleichen sechs UX-Skalen wie in der Vorstudie als am wichtigsten erweisen. Andersherum ist der Aufbau jedoch logischer und die Ergebnisse wurden durch die Änderung nicht beeinflusst. Eine weitere Abweichung bestand in der Abwandlung von Begriffen, womit ein vereinfachtes Verständnis erzielt werden sollte. Die neuen Begriffe wurden im Update erläutert. In der vorliegenden Arbeit wurden die Begriffe konsistent angepasst, um Verwirrung zu vermeiden. Der Begriff *Produktkategorie* wurde ersetzt durch *Business Lines*, womit die sechs Bereiche MR, CT, XP, AT, D&A und Strahlentherapie gemeint sind. Diese Business Lines wiederum werden unterteilt in insgesamt 11 Produktkategorien. Zum Beispiel fallen unter die Business Line CT neben CT-Geräten auch SPECT- und PETCT-Geräte. Die Unterteilung kann in Kapitel 1.4 (User Experience bei Siemens Healthineers) nachgelesen werden. Die Business Lines wurden den Proband*innen im Rekrutierungsprozess vorgegeben, sie durften aber frei wählen, für welche Produktkategorie sie die Umfrage beantworten wollten. Aus der gewählten Produktkategorie sollte dann ein *Referenzprodukt* angegeben werden, mit dem am häufigsten gearbeitet wurde. Für dieses Referenzprodukt wurde der Großteil der Umfrage ausgefüllt. Die Business Line DH/Syngo wurde außerdem während der Bearbeitung der Masterarbeit bei Siemens Healthineers umbenannt in D&A (Digital & Automation). In

der Präregistrierung taucht noch die alte Bezeichnung der Business Line auf, für die Masterarbeit wurde die neue Bezeichnung D&A verwendet.

Es wurden zudem einzelne Änderungen am Aufbau des Studienfragebogens vorgenommen, die im Update festgehalten wurden. Die berufsbezogenen Fragen wurden aus der Vorstudie (Müßig, 2022) übernommen, allerdings wurden sie geringfügig verändert beziehungsweise wurden Fragen gekürzt. Die für die vorliegende Arbeit nicht relevante Frage „Haben Sie bisher ein Training für Siemens Healthineers Produkte erhalten?“ wurde entfernt, um die Abbruchquote durch eine geringere Gesamtanzahl an Fragen zu reduzieren. Auch das Einfügen von Antwortkategorien bei der Frage „Wie viele verschiedene [xxx]-Geräte haben Sie schon bedient?“ sollte die Beantwortung beschleunigen und die Abbruchquote so verringern. Diese Änderungen hatten keinen Einfluss auf die Ergebnisse, da diese Fragen nicht in die Prüfung der Hypothesen einbezogen wurden.

4.3.2 Abweichungen bezüglich der Stichprobe.

Es ergaben sich darüber hinaus weitere Abweichungen von der Präregistrierung in Bezug auf die Stichprobe, die im Update noch nicht festgehalten wurden. Die Rekrutierung verlief anders als vorgesehen. Für jede Business Line waren 40 Proband*innen eingeplant, was sich jedoch nicht umsetzen ließ. Über das Collaboration Management von Siemens Healthineers konnten weniger Personen als geplant rekrutiert werden, weshalb anstelle der vorgesehenen 200-240 Teilnehmenden nur 173 Personen an der externen Studie teilnahmen. Ferner füllten einige Teilnehmende die Studie nicht für jene Business Line aus, für die sie rekrutiert worden waren. Es wurden zusätzlich potenzielle Teilnehmende über die sozialen Netzwerke Xing (XING, 2022) und LinkedIn (LinkedIn, 2022) angeschrieben, was ursprünglich nicht geplant war. Diese Abweichung in der Rekrutierung hatte jedoch ohnehin keine Auswirkungen, da über sich über soziale Netzwerke keine Personen rekrutieren ließen. Der genaue Ablauf der Rekrutierung wurde im Abschnitt 3.1.2 (Rekrutierung der Stichprobe) berichtet. Die damit geringer ausgefallene Stichprobengröße war insbesondere für die Berechnung der Faktorenanalysen und des Strukturgleichungsmodells ungünstig, da für diese Verfahren sehr große Stichproben erforderlich sind. Allerdings ergab sich auch für die MANOVAs das Problem, dass in manchen Gruppen weniger Personen als abhängige Variablen vorlagen. Dazu kam, dass Gruppengrößen unterschiedlich groß waren, was für die Berechnung von MANOVAs ebenfalls nachteilig ist. Die kleinere Stichprobe schränkt somit die Interpretierbarkeit der Ergebnisse ein. Weiterhin sollte die Studie eigentlich für ein gezieltes Referenzprodukt der ausgewählten Produktkategorie ausgefüllt werden, was einige Teilnehmende

allerdings missverstanden. Sie trugen in das Feld, in welches das Referenzprodukt eingegeben werden sollte, mehrere Modelle ein, sodass unklar war, für welches Produkt genau sie die Studie beantwortet hatten. Für die meisten Analysen, mit Ausnahme der Interrater-Reliabilität, spielte das exakte Modell keine Rolle. Es kann jedoch nicht beurteilt werden, inwieweit dieser Umstand die Beantwortung der Items beeinflusst haben könnte.

4.3.3 Abweichungen im Umgang mit fehlenden Werten und der Prüfung von Voraussetzungen.

Eine wesentliche Abweichung von der ersten Präregistrierung bestand im Umgang mit fehlenden Werten. Entgegen dem geplanten fallweisen Ausschluss wurde im Update paarweiser Ausschluss festgehalten. Der Großteil der Items, die für die Hypothesen relevant waren, war ohnehin als verpflichtend vorgegeben. Details dazu, welche Items keine Pflichtfelder waren, wurden im Kapitel 3.5.1 (Vorbereitung der Daten) berichtet. Eine dort beschriebene Ausnahme stellten die neu entwickelten Items der Skalen Effektivität und Sicherheit dar. Für sie wurden Proband*innen angewiesen, unverständliche Items auszulassen. Dieses Vorgehen wurde gewählt, da in QuestionPro, dem Online-Tool zur Studienerstellung (*QuestionPro*, 2022), keine andere Einstellung möglich war. Bei Items mit zwei Polen konnte man hier eine Antwortkategorie für „keine Angabe" beziehungsweise „Ich finde das Item unverständlich“ nur zwischen den Polen einfügen, was bei der Beantwortung der Items gestört hätte. Bei einer Auslassung erschien dafür die Nachfrage, ob man das Item wirklich unbeantwortet lassen wolle. Personen, die die Umfrage abbrachen, wurden ausgeschlossen. So konnte sichergestellt werden, dass keine Werte fehlten, weil Personen das Item übersehen oder die Umfrage nicht bis zum Ende bearbeitet hatten. Durch paarweisen Ausschluss sollte vermieden werden, Personen mit einzelnen fehlenden Antworten vollständig auszuschließen. So konnten alle verfügbaren Werte genutzt werden. Die Anwendung paarweisen Ausschlusses war im Update der Präregistrierung für alle Analyseverfahren angedacht. Bei der Auswertung der Daten stellte sich jedoch heraus, dass die Anwendung paarweisen Ausschlusses für die Durchführung der MANOVAs und der multivariaten Kruskal-Wallis-Tests nicht möglich war. Stattdessen wurde automatisch fallweiser Ausschluss angewandt. Welche denkbaren Auswirkungen dieser Umgang mit fehlenden Werten auf die Ergebnisse haben könnte, wird in Kapitel 4.4.3 (Umgang mit fehlenden Werten) diskutiert. Die Bewertung der Wichtigkeit der Skalen Haptik und Akustik war als optional eingestellt, weil sie zu manchen Referenzprodukten nicht passten. Die Wichtigkeitsbewertung der Skala Haptik ließen 15 Personen aus, welche dann entsprechend nicht in die Berechnung der MANOVA für die

zweite Hypothese einbezogen werden konnten. Es wäre nachvollziehbar gewesen, wenn Personen der Business Line D&A die Skala Haptik ausgelassen hätten, weil sie sich nicht auf Software anwenden ließ. Dies war jedoch nicht der Fall. Zwar wurde die Bewertung von vier Personen der Business Line D&A ausgelassen und damit geringfügig von mehr Personen als von anderen Business Lines, die entstandenen fehlenden Werte verteilten sich jedoch relativ gleichmäßig auf die verschiedenen Business Lines. Zum Beispiel ließen auch drei Personen der Business Line XP die Skala Haptik aus. Für die Skala Akustik verhielt es sich ähnlich. Unter den 14 Personen, welche die Wichtigkeitsbewertung der Skala Akustik übersprangen, befanden sich mit sechs Teilnehmenden etwas mehr Personen der Business Line D&A als von anderen Business Lines, aber auch hier verteilten sich die fehlenden Werte auf alle Business Lines. Nachdem somit von keiner der Business Lines deutlich mehr Personen ausgeschlossen wurden als von anderen, dürfte es in dieser Hinsicht zu keiner systematischen Verzerrung der Ergebnisse der MANOVA für Hypothese 2 gekommen sein. Für die MANOVA im Rahmen der Prüfung von Hypothese 5.2.2 (Unterschiede in der Bewertung von Produkten verschiedener Hersteller) wurde durch den fallweisen Ausschluss eine Person ausgeschlossen, für die sich kein Skalenwert für Effektivität bilden ließ. Der fallweise Ausschluss hatte auf diese Berechnung somit keine erheblichen Auswirkungen. Auch für die Berechnung der Intraklassen-Korrelation (Hypothese 5.1) wurden fehlende Werte fallweise ausgeschlossen. Es wies jedoch für diese Berechnung keine der acht herangezogenen Personen einen fehlenden Skalenwert auf.

Was ebenso erst nach dem Update der Präregistrierung abgewandelt wurde, war die Prüfung der Voraussetzungen. Diese wurden ausführlicher geprüft, weil präregistrierte Methoden teils keine eindeutigen Schlüsse zuließen. Neben inferenzstatistischen Methoden, wie zum Beispiel dem Mardia-Test wurden auch grafische Kriterien, beispielsweise Histogramme, einbezogen. Für die neu entwickelten Sicherheitsitems war erst nach Betrachtung der Schiefe und Kurtosis eine Schlussfolgerung im Hinblick auf die Verteilung möglich. Die genauere Prüfung der Voraussetzungen dürfte keine Auswirkungen auf die Ergebnisse gehabt haben.

4.3.4 Abweichungen in der Prüfung der Hypothesen.

Neben den bisher genannten Änderungen waren auch für die Prüfung der Hypothesen Abweichungen vom geplanten Vorgehen beziehungsweise nicht präregistrierte Entscheidungen nötig. Für die ersten beiden Hypothesen, die sich auf die Replikation der Vorstudie (Müßig, 2022) bezogen, wurde vorab nicht festgelegt, wie mit abweichenden Ergebnissen zwischen den Studien umgegangen wird. Aufgrund der größeren Stichprobe

wurde entschieden, sich an den Wichtigkeitsrankings der Skalen aus der Replikationsstudie zu orientieren, wobei sich die Wichtigkeitsrankings der Studien ohnehin ähnelten.

In Bezug auf die dritte Hypothese ergaben sich im Verlauf der Itemgenerierung Items, die sich nur auf Hardware-Komponenten anwenden lassen. Die dritte Hypothese, welche die Extraktion der vier geeignetsten Items für die beiden Zusatzskalen vorsah, unterlag somit notwendigen Änderungen. Es wurden anstelle von einer Faktorenanalyse zwei Faktorenanalysen für die Skala Sicherheit durchgeführt und es war die Prüfung von zwei Strukturgleichungsmodellen geplant. Zum einen wurden allgemeine Sicherheitsitems unter Einbezug aller Proband*innen herangezogen. Zum anderen wurden alle Sicherheitsitems und nur Personen einbezogen, die Hardware-Items beantworten konnten. Im Update der Präregistrierung wurde diese neue Herangehensweise festgehalten. Geplant war, im Fall von Eindimensionalität aller Sicherheitsitems (inklusive der Hardware-Items) wie vorgesehen eine Sicherheitsskala zu konstruieren, die dann vier allgemeine Items und zusätzliche, optionale Hardware-Items umfasst hätte. Andernfalls sollte die Skala geteilt werden. Es wurden wie geplant separate Faktorenanalysen mit und ohne Hardware-Items berechnet. Bei der Auswertung der Wichtigkeitsbewertungen der Skalen fiel Sicherheit jedoch nicht mehr unter die sechs produktübergreifend wichtigsten Skalen, weshalb, anders als im Update vorgesehen, nur ein Strukturgleichungsmodell ohne Sicherheitsitems aufgestellt wurde.

Bezüglich der Faktorenanalysen zur Prüfung der dritten Hypothese wurde in der ersten Version der Präregistrierung festgehalten, dass Maximum-Likelihood-Faktorenanalysen berechnet werden sollten. Im Fall von nicht optimal erfüllter Normalverteilung sollten weiterhin robustere Hauptachsenanalysen durchgeführt werden (Schnaudt, 2020). Da Schrepp und Thomaschewski (2019a) für die Entwicklung des UEQ+ jedoch Hauptkomponentenanalysen berechnet hatten, wurde entschieden, auch ebensolche zu Vergleichszwecken anzuwenden. Maximum-Likelihood-Faktorenanalysen waren zunächst die Methode der Wahl, weil sie die Berechnung verschiedener Gütekriterien für die Faktorenlösung ermöglichen (Fabrigar et al., 1999). Allerdings empfiehlt es sich, mehrere Methoden miteinander zu vergleichen und zu sehen, ob sich für alle die gleiche Faktorenlösung ergibt (Williams et al., 2010). Wie im Update festgehalten, wurden verschiedene Methoden verglichen. Allerdings führte die Parallelanalyse bezüglich der allgemeinen Sicherheitsitems laut Hauptkomponentenanalyse zur Extraktion einer Hauptkomponente, laut Faktorenanalyse (Hauptachsenanalyse) zu zwei Faktoren. Es wurde nicht präregistriert, wie in solch einem Fall weiter vorgegangen wird. Die Parallelanalyse wurde

mit jener inklusive Hardware-Items verglichen. Da hier Hauptkomponentenanalyse und Hauptachsenanalyse das gleiche Ergebnis brachten, wurde sich daran orientiert und ein Faktor, beziehungsweise im Modell mit Hardware-Items zwei Faktoren (bzw. Hauptkomponenten) angenommen. Es zeigte sich indessen ein schlechterer Fit für diese Modelle im Vergleich zu Modellen mit einem zusätzlichen Faktor (bzw. einer Hauptkomponente). Aus theoretischen Gründen und auf Basis grafischer Anhaltspunkte wurden trotzdem jene Modelle mit einem Faktor (bzw. einer Hauptkomponente) weniger angenommen, die Entscheidung war allerdings nicht eindeutig. Für die Auswahl geeigneter Items wurden wie präregistriert die Ladungen sowie die Itemschwierigkeiten betrachtet. Darüber hinaus wurde auch die Anzahl fehlender Werte einbezogen, denn diese ermöglichte eine Einschätzung der Verständlichkeit der Items. Dieses Vorgehen wurde nicht präregistriert und es wurde entsprechend auch nicht festgelegt, welche maximale Anzahl an Auslassungen hinnehmbar ist. Sowohl für die Bestimmung der Faktorenanzahl als auch die Auswahl der Items wäre es besser gewesen, vorab Entscheidungskriterien festzulegen, zumal sich die Entscheidungen wesentlich auf den erhaltenen finalen UX-Fragebogen für Medizinprodukte auswirkten. Um Entscheidungen nachvollziehbar zu machen, wurde auf eine ausführliche Darstellung wert gelegt.

Im Hinblick auf Hypothese 4 ergab sich im Rahmen vertiefter Recherchen bezüglich der Modifikationsindizes aufgrund eines Missverständnisses des Vorgehens von Saris et al. (2009) eine Änderung der Entscheidungsregel, wann Fehlspezifikationen vorliegen. Diese Änderung wurde auch im Update der Präregistrierung noch nicht dokumentiert. Die Beurteilung wurde allerdings sogar strenger als in der Präregistrierung beziehungsweise dem Update festgehalten, sodass mehr Fehlspezifikationen identifiziert wurden. Der Prozess wurde im Ergebnisteil ausführlich dargelegt (s. Kapitel 3.5.4 Ergebnisse der Fragebogenentwicklung).

Zur Beurteilung der Interrater-Reliabilität (Hypothese 5.1) wurde die Intraklassen-Korrelation für ein ausgewähltes Referenzprodukt berechnet. Im Update war geplant, die Intraklassen-Korrelation für mehr als ein Referenzprodukt zu berechnen, da sich aber nur ein Produkt mit einer größeren Anzahl an Bewerter*innen finden ließ (acht anstatt vier oder weniger Personen), wurde darauf verzichtet, die Interrater-Reliabilität für mehrere Produkte zu berechnen. Des Weiteren wurde nicht präregistriert, welche Art der Intraklassen-Korrelation berechnet werden sollte, da im Vorhinein nicht berücksichtigt wurde, dass verschiedene Arten zur Verfügung stehen. Im Abschnitt 3.5.5 (Psychometrische Gütekriterien) wurden Überlegungen und das Vorgehen diesbezüglich dargelegt.

Neben der Intraklassen-Korrelation wurde deskriptiv verglichen, ob die Standardabweichungen jener Skalenbewertungen von Personen, die das gleiche Referenzprodukt beurteilt hatten, geringer ausfielen als die Gesamtstandardabweichungen. Diese Beurteilung der absoluten Übereinstimmung wurde im Update der Präregistrierung noch nicht festgehalten. Neben den rein numerischen Kennwerten wurden als grafischer Anhaltspunkt für die absolute Übereinstimmung der Befragten zusätzlich Violinplots ausgegeben. Die Abweichung zur Präregistrierung bestand demnach darin, dass die Untersuchung der Inter-rater-Reliabilität umfangreicher vorgenommen wurde, was jedoch zu keinem eindeutigeren Ergebnis führte.

In der Präregistrierung waren außerdem einige explorative Analysen vorgesehen. Diese wurden indes nicht umgesetzt, weil das den Umfang der Masterarbeit überstiegen hätte. Insgesamt hätte die Präregistrierung an einigen Stellen ausführlicher ausfallen sollen, denn durch fehlende vorab festgelegte Kriterien mussten teils auf Basis der vorliegenden Daten Entscheidungen getroffen werden. Diese unterliegen Interpretationsspielraum, auch wenn versucht wurde, Gedankengänge und Entscheidungen nachvollziehbar darzustellen.

4.4 Limitationen

Trotz aller Sorgfalt bei der Entwicklung des UX-Fragebogens gibt es Kritikpunkte und Einschränkungen, die in diesem Kapitel beleuchtet werden sollen. Einige Schwachpunkte, wie beispielsweise die nicht erreichte Anzahl geplanter Proband*innen, wurden bereits im vorangegangenen Kapitel aufgegriffen und werden an dieser Stelle daher nicht erneut thematisiert.

4.4.1 Einfluss numerischer Bezeichnungen der Ratingstufen auf die Itemantworten.

Was hinterfragt werden sollte, ist das Einfügen von Bezeichnungen über den sieben einzelnen Ratingstufen der Items. In der Vorstudie von Müßig (2022), die hier teilrepliziert wurde, waren über den Items zur Bewertung der Wichtigkeit der Skalen Zahlen von -3 bis +3 präsentiert worden. Diese numerischen Label wurden für mehr Einheitlichkeit in der zuletzt durchgeführten Studie auch für die Items der Skalen zur Produktbewertung (die wichtigsten Skalen des UEQ+ und die Zusatzskalen) aufgenommen, obwohl sie kein Bestandteil des UEQ+ sind (Schrepp & Thomaschewski, 2019a). Allerdings könnten die Zahlen durchaus Einfluss auf die Beantwortung der Items gehabt haben. Es konnte mehrfach gezeigt werden, dass der Wertebereich von numerischen Bezeichnungen die Beantwortung von Items beeinflusst. Wenn Zahlen, wie in der vorliegenden Arbeit, von

negativ bis positiv reichen, werden extremere Antworten begünstigt (Moors et al., 2014). Negative Zahlen scheinen vermieden zu werden, was zu einer Antworttendenz in die positive Richtung führen kann (Kunz, 2015; Schwarz et al., 1991; Schwarz & Hippler, 1995). Schwarz et al. (1991) zeigten zum Beispiel, dass bei Vorgabe einer bipolaren Skala von -5 bis +5 die Werte zwischen -5 und 0 von 13 % der Befragten ausgewählt wurden, während bei einer Skala von 0 bis 10 äquivalente Werte von 0 bis 5 von 34 % der Befragten ausgewählt wurden. Die Autor*innen schlossen daraus, dass Werte von 0 bis 10 als (Nicht-)Vorliegen eines Merkmals interpretiert werden, während negative Werte als Vorliegen eines expliziten Fehlers interpretiert werden. Diese Interpretation könnte auch auf die vorliegende Arbeit zutreffen. Die angegebenen Zahlen von -3 bis +3 könnten dazu geführt haben, dass Teilnehmende negative Werte als Vorliegen eines eindeutigen Mangels verstanden. In der Folge könnten sich die ausgewählten Antworten in den positiven Bereich verschoben haben, was sich ungünstig auf die Differenzierungsfähigkeit der Items ausgewirkt haben könnte. Abgesehen von numerischen Bezeichnungen der Ratingstufen sind auch andere Arten der Beschriftung denkbar. Im Ausblick (Kapitel 4.6.1) wird eine Empfehlung bezüglich der Verwendung von Bezeichnungen abgeleitet.

4.4.2 Skalenniveau der verwendeten Skalen.

Likert-Skalen finden großen Anklang in den Sozialwissenschaften und werden unter den psychometrischen Skalen am häufigsten eingesetzt (Döring & Bortz, 2016). Nichtsdestotrotz gibt es Uneinigkeit darüber, auf welchem Skalenniveau die mit ihnen erhobenen Daten vorliegen (Awang et al., 2016; Carifio & Perla, 2008; Jamieson, 2004; Joshi et al., 2015; Knapp, 1990; Liddell & Kruschke, 2018). Die für die Hypothesen dieser Arbeit relevanten Variablen wurden mittels siebenstufiger Skalen beziehungsweise semantischer Differentiale erhoben. Semantische Differentiale und Likert-Skalen sind zwar nicht das gleiche (Föhl & Friedrich, 2022), da sie jedoch beide mehrere Abstufungen (in der Regel fünf oder sieben) vorgeben, werden sie hier gemeinsam betrachtet.

Häufig werden die mit Likert-Skalen erhobenen Daten als intervallskaliert behandelt, was allerdings Gleichabständigkeit der einzelnen Ratingstufen voraussetzen würde, die in der Regel, wie auch in dieser Arbeit, gar nicht untersucht wird (Liddell & Kruschke, 2018; Wu & Leung, 2017). Durch das Einfügen der Zahlen von -3 bis +3 über den einzelnen Ratingstufen sollte die Wahrnehmung von Äquidistanz in der vorliegenden Studie verbessert werden (Rohrmann, 2007), wie die Abstände wahrgenommen wurden, kann jedoch nicht beurteilt werden. Einige Autor*innen empfehlen die Annahme einer Intervallskalierung nur dann, wenn mehrere Items zu Skalen zusammengefasst werden und

die Annahme von Ordinalskalierung, wenn einzelne Items betrachtet werden (Joshi et al., 2015; Subedi, 2016). Liddell und Kruschke (2018) sowie Kuzon et al. (1996) und Jamieson (2004) sind sich allerdings einig, dass die Berechnung des Durchschnitts von Variablen bereits voraussetzt, dass diese intervallskaliert sind, denn nur bei Intervallskalierung ist der Mittelwert überhaupt aussagekräftig.

Andere Autor*innen schlagen die Verwendung von mehr Abstufungen, zum Beispiel 11-stufigen Skalen vor, weil durch sie eine bessere Annäherung an eine Normalverteilung erzielt werden könne (Leung, 2011; Wu & Leung, 2017). Normalverteilung ist eine Bedingung für die Anwendung parametrischer Verfahren, die sich bei Likert-Skalen häufig als problematisch erweist (Liddell & Kruschke, 2018). G. Norman (2010) kritisiert wiederum, dass Studien nicht ignoriert werden dürften, die die Robustheit parametrischer Verfahren gegenüber Voraussetzungsverletzungen wie mangelnder Normalverteilung aufzeigen (z.B. Glass et al., 1972). Es existieren zudem auch Arbeiten, die darauf hindeuten, dass parametrische und nichtparametrische Verfahren bezüglich Likert-Skalen in vielen Fällen zu vergleichbaren Ergebnissen führen (Kühnel, 1993; Mircioiu & Atkinson, 2017; Murray, 2013; Rhemtulla et al., 2012). Es gibt somit Argumente für und gegen die Verwendung (nicht-)parametrischer Verfahren in Bezug auf Likert-Skalen oder semantische Differentiale. Als Fazit ist festzuhalten, dass die für die Auswertung dieser Studie überwiegend verwendeten parametrischen Verfahren zwar nicht unangebracht waren, jedoch ein Vergleich der Ergebnisse von parametrischen und nichtparametrischen Verfahren empfehlenswert ist. Für die berechneten MANOVAs und Korrelationen wurde das umgesetzt, für die Faktorenanalysen und das Strukturgleichungsmodell dagegen nicht.

4.4.3 Umgang mit fehlenden Werten.

Der Umgang mit fehlenden Werten in der vorliegenden Masterarbeit wurde im Kapitel 3.5.1 (Vorbereitung der Daten, Umgang mit fehlenden Werten und Ausreißern) beschrieben. Es wurde jedoch noch nicht darauf eingegangen, welche Auswirkungen das beschriebene Vorgehen auf die Ergebnisse gehabt haben könnte. Wenn Werte nicht völlig zufällig fehlen, kann es bei einem Ausschluss der fehlenden Werte zu Verzerrungen in den Parameterschätzungen kommen (Brown, 2015; Lüdtke & Robitzsch, 2010; Rubin, 1976). Imputationsbasierte Verfahren, die fehlende Werte durch einen oder mehrere plausible Werte ersetzen, sind daher in der Regel vorzuziehen (Lüdtke & Robitzsch, 2010). In der vorliegenden Arbeit lagen fehlende Werte allerdings teilweise aufgrund dessen vor, dass Teilnehmende sich zu manchen Items nicht sinnvoll äußern konnten. Die Hardware-Items der Skala Sicherheit konnten Personen, die sich auf reine Software-Anwendungen

bezogen, nicht beantworten, weshalb eine Imputation hier keinen Sinn ergeben hätte. Es wurde deshalb vorgezogen, zwei Faktorenanalysen zu berechnen. Ein ähnlicher Fall lag bei den Replikationsitems zur Wichtigkeitsbewertung der Skalen Haptik und Akustik vor, die wie in Müßigs Studie als optional eingestellt wurden. Hier wäre eine Imputation der fehlenden Werte ebenfalls nicht plausibel gewesen. Das führte durch den fallweisen Ausschluss fehlender Werte im Rahmen der MANOVAs allerdings zu einer Verringerung der einbezogenen Fälle um 15 Personen und möglicherweise zu einer Verzerrung der Ergebnisse.

Ein weiterer Kritikpunkt, der an dieser Stelle anzuführen ist, ist der Ausschluss abgebrochener Studienfragebögen. Ziel dieses Vorgehens war, Personen nicht mehrfach in die Analysen einzubeziehen, die den Studienfragebogen begannen, abbrachen und später erneut von Anfang an ausfüllten. Das war durch den anonymisierten Link möglich. Zwar ist es wahrscheinlich, dass dies mehrfach vorkam, dennoch wäre es ebenfalls möglich, dass so Personen ausgeschlossen wurden, welche die Studie nur einmal begannen und dann unterbrochen wurden. Falls dem so war, wäre anzunehmen, dass vor allem Personen mit einer hohen Arbeitsbelastung nicht erneut an der Umfrage teilnahmen. Folglich könnte es zu Selektionseffekten in der Stichprobe gekommen sein.

4.4.4 Mängel bezüglich der Sicherung der Güte.

Es wurden verschiedene Kennwerte zur Beurteilung der psychometrischen Güte des UX-Fragebogens für Medizinprodukte herangezogen, welche die Qualität des UX-Fragebogens belegen. Eine zentrale Einschränkung ist hierbei, dass die Skala Nützlichkeit weder in das Strukturgleichungsmodell zur Überprüfung der Konstruktvalidität noch in die Untersuchung anderer Gütekriterien einbezogen werden konnte, da für sie keine Items erhoben wurden. Im Hinblick auf die diskriminante Validität, für welche die Skala Nützlichkeit ebenfalls nicht inkludiert werden konnte, wurden bislang außerdem nur verschiedene Hersteller verglichen. Es sollten jedoch einzelne Modelle von einer entsprechend großen Stichprobe bewertet und verglichen werden, andernfalls kann nicht geschlussfolgert werden, dass der UX-Fragebogen zwischen konkreten Modellen differenziert. Gerade das wäre allerdings wichtig, denn in der Praxis dürfte vor allem interessant sein, ob beispielsweise ein ausgewähltes MRT-Modell anderen MRT-Modellen über- oder unterlegen ist.

Was weiterhin kritisiert werden kann, ist dass die Interrater-Reliabilität nur für ein ausgewähltes Referenzprodukt berechnet wurde. Dieses Produkt wurde zudem nur von acht Personen bewertet. Das Konfidenzintervall des erhaltenen Werts ist so breit, dass

eigentlich keine Aussage bezüglich der Interrater-Reliabilität in Bezug auf das Referenzprodukt möglich ist. Was bisher ebenso nicht berücksichtigt wurde, ist die zeitliche Stabilität der Messung (Retest-Reliabilität), wofür eine zweite Erhebung des Fragebogens anhand derselben Stichprobe mit denselben Referenzprodukten nötig wäre (Moosbrugger & Kevala, 2020). Bezüglich der Retest-Reliabilität muss allerdings beachtet werden, dass sich die UX durchaus über die Zeit verändern könnte. Auch wenn das Produkt gleichbleibt, könnte es sein, dass Personen die Interaktion anders erleben, beispielweise weil in der Zwischenzeit technische Fortschritte den Vergleichsmaßstab angehoben haben könnten. Da die UX per se vom Kontext abhängt (Ariza & Maya, 2014) und bei einem sich verändernden Kontext womöglich nicht konstant ist, ist auch hier die Frage, inwieweit die Retest-Reliabilität ein Eignungskriterium von UX-Fragebögen darstellt. Wenn zwei Testergebnisse mit größerem zeitlichem Abstand nicht übereinstimmen, liegt das unter Umständen nicht daran, dass der UX-Fragebogen nicht zuverlässig und genau misst, sondern an einer tatsächlichen Veränderung der UX über die Zeit. In Anbetracht des raschen technischen Fortschritts könnte es so schon bei kurzen zeitlichen Abständen zu deutlichen Veränderungen der UX kommen.

Grundsätzlich stellt sich Frage, ob gängige Konstruktionsprinzipien, die aus der psychologischen Testtheorie übernommen werden, im Hinblick auf UX-Fragebögen überhaupt angebracht sind (s. Kapitel 1.5.1 – Anwendbarkeit der psychologischen Testtheorie). Generell besteht eine Schwierigkeit darin, dass nicht wie bei psychologischen Fragebögen Personeneigenschaften von Interesse sind, sondern primär, inwieweit Produkteigenschaften die UX beeinflussen. Was die Überprüfung der Güte des UX-Fragebogens erschwert ist, dass, wie bereits mehrfach genannt, Personen-, Produkt- und Kontexteigenschaften die Bewertung der Produkte beeinflussen (Hassenzahl et al., 2009; Hassenzahl, 2011; Hassenzahl & Tractinsky, 2006; Rummel & Schrepp, 2018). Bei UX-Fragebögen sind nicht wie bei psychologischen Fragebögen einzelne Personenwerte von Interesse, sondern die mittlere Bewertung eines Produkts über mehrere Personen hinweg (Rummel & Schrepp, 2018). Wie Skalenwerte, sprich latente Variablen bei UX-Fragebögen genau zu interpretieren sind, ist wegen der Vermischung verschiedener Einflüsse unklar. Die Übertragung von Prinzipien der psychologischen Testtheorie kann demnach angezweifelt werden, auch wenn sie derzeit bei der Entwicklung von UX-Fragebögen die methodische Grundlage bildet. Welche Methoden allerdings geeigneter wären, ist bislang ungewiss (Schrepp & Rummel, 2018).

Aufgrund des Fehlens qualitativer Informationen konnten in der durchgeführten Studie keine Hinweise zu fehlenden Werte bezüglich der neu entwickelten Items eingeholt werden. Ausnahmslos alle neuen Items wiesen Auslassungen auf, was auf Verständnisschwierigkeiten hindeutet. Auch wenn nur wenige Werte fehlten, wäre es aufschlussreich gewesen, qualitative Anmerkungen zu den ausgelassenen Items zu erhalten. So hätte man die Augenscheinvalidität der Items evaluieren können. Augenscheinvalidität ist für den entwickelten UX-Fragebogen für Medizinprodukte wünschenswert, um die Teilnahmebereitschaft möglichst zu erhöhen (Döring & Bortz, 2016). Wenn Testpersonen die Items unpassend oder unverständlich finden, ist es unwahrscheinlich, dass sie bei der Beantwortung des UX-Fragebogens motiviert sind. Insgesamt lässt der entwickelte UX-Fragebogen für Medizinprodukte angesichts der rein quantitativen Datenerhebung nur eingeschränkt Schlussfolgerungen zu. Er erleichtert zwar den Vergleich von Medizinprodukten und Verlaufskontrollen, konkrete Vorschläge der Testpersonen jedoch nicht erfasst werden. Durch den UX-Fragebogen lässt sich ermitteln, ob oder wie gut ein Medizinprodukt der Zielgruppe gefällt. Was jedoch stört oder besonders positiv auffällt, lässt sich nicht beurteilen. Gerade solche Informationen wären für Designer*innen allerdings wichtig, um gezielt Verbesserungen vornehmen zu können.

4.5 Implikationen und Beitrag der Arbeit

Trotz einiger Einschränkungen liefert der entwickelte UX-Fragebogen für Medizinprodukte einen wesentlichen Beitrag zur Erfassung von UX. Die entwickelten Skalen Effektivität, Allgemeine Sicherheit und Hardware-Sicherheit ermöglichen die standardisierte Erfassung von UX-Aspekten, die sich so bislang nicht messen ließen, in Bezug auf Medizinprodukte aber nachweislich relevant sind. Neben ihrer Anwendung auf Medizinprodukte könnten sie auch als Erweiterung des UEQ+ (Schrepp & Thomaschewski, 2019a) eine Bereicherung darstellen. Die Skala Hardware-Sicherheit ist sehr spezifisch auf Medizinprodukte angepasst und daher unter Umständen nicht auf andere Produkte anwendbar, aber Effektivität und Allgemeine Sicherheit könnten auch für Produkte anderer Branchen interessant sein. Die Skala Allgemeine Sicherheit könnte insbesondere in risikoreichen Branchen Anwendung finden. Ein Vorschlag wäre zum Beispiel der Einsatz der Skala für UX-Testungen von Piloten- oder Fluglotsensoftware.

Die vorliegende Arbeit liefert außerdem Erkenntnisse darüber, welche Daten aus Sicht von medizinischem Personal besonders schützenswert sind. Es konnte ermittelt werden, dass sich Datensicherheit aus Perspektive des Personals primär auf Patient*innendaten bezieht. Ein Drittel der Befragten versteht unter Datensicherheit alle Arten von

Daten, eigene Daten (von medizinischem Personal) scheinen dagegen eine untergeordnete Rolle zu spielen. Der veränderte, verallgemeinerte Einleitungssatz der Skala Vertrauen sollte deshalb beibehalten werden (s. Kapitel 3.2.3 – Items zur Fragebogenentwicklung).

Es sollte darüber hinaus auch in Zukunft die Information genutzt werden, welche Skalen für das jeweils untersuchte Medizinprodukt besonders wichtig sind. Der neu entwickelte UX-Fragebogen sieht, wie der UEQ+, am Ende jeder Skala eine erneute Abfrage der Wichtigkeit durch ein zusätzliches Item vor. Der Vergleich von Wichtigkeitsbewertung und dem erreichten Skalenmittelwert gibt Aufschluss darüber, ob Schwerpunkte im Design richtig gesetzt wurden. Falls eine Skala im Vergleich zu den anderen Skalen als weniger wichtig für das spezifische Produkt bewertet wird, ist es weniger dringlich, diesen Aspekt zu verbessern. Aspekte, die dagegen als sehr wichtig eingestuft werden, aber zum Zeitpunkt der Erhebung nicht gut erfüllt sind, sollten im Designprozess fokussiert werden. Anschließend könnten dann qualitative Erhebungen Aufschluss darüber geben, was konkret verbessert werden sollte. Falls sich im Verlauf einer Produktentwicklung bereits zu Beginn ein wichtiger Aspekt als sehr gut erfüllt herausstellen sollte, ist es jedoch ratsam ihn auch in späteren Stadien der Entwicklung zu erheben. Die Skalen waren zwar in der vorliegenden Studie positiv korreliert, aber es wäre durchaus denkbar, dass sich in manchen Fällen durch Veränderung einer Komponente die Bewertung einer Skala verbessert, während sich eine andere verschlechtert. Das wäre zum Beispiel möglich, wenn zusätzliche Mechanismen zur Gewährleistung des Datenschutzes eingebaut werden. Die Skala Vertrauen könnte so besser abschneiden, wohingegen die Übersichtlichkeit der Benutzeroberfläche durch zusätzliche Hinweise eventuell abnimmt. An die genannten Implikationen der Arbeit anknüpfend wird im Folgenden ein Ausblick über weitere Schritte und Herausforderungen im Prozess der UX-Fragebogenentwicklung gegeben.

4.6 Ausblick

Im Abschnitt Limitationen (Kapitel 4.4) wurde dargelegt, dass einige Aspekte bei der Entwicklung des UX-Fragebogens für Medizinprodukte nicht ausreichend berücksichtigt werden konnten. Im nun folgenden Abschnitt werden Vorschläge unterbreitet, wie man diese in Zukunft noch einfließen lassen könnte. Außerdem wird auf weitere Herausforderungen und zukünftige Aufgaben eingegangen.

4.6.1 Bezeichnungen der Ratingstufen.

Im Abschnitt 4.4.1 (Einfluss numerischer Bezeichnungen der Ratingstufen auf die Itemantworten) wurden mögliche Verzerrungen aufgrund numerischer Bezeichnungen

der Ratingstufen aufgezeigt. Es stellt sich nun die Frage, ob Bezeichnungen für die Abstufungen eingefügt werden sollen oder nicht. Es wurde bereits ausgeführt, dass insbesondere der Wertebereich numerischer Bezeichnungen Itemantworten beeinflusst (Kunz, 2015; Schwarz et al., 1991; Schwarz & Hippler, 1995). Für die Klärung der Frage, in welchem Wertebereich numerische Bezeichnungen liegen sollten, ist entscheidend, ob die erhobenen Konstrukte uni- oder bipolar sind (Schwarz et al., 1991; Tourangeau et al., 2007). Rossiter (2011) betrachtet evaluative Ansichten als bipolar, denn sie können negativ, neutral oder positiv ausfallen. Es ist naheliegend, semantische Differentiale als bipolar anzunehmen, da gegensätzliche Begriffe vorgegeben werden, allerdings lässt sich diese Annahme für einige Items durchaus diskutieren (Heise, 1969; Rossiter, 2011). Manche Autor*innen schlagen für semantische Differentiale rein positive Nummerierungen, zum Beispiel von 1 bis 7 vor, was Unipolarität andeutet (Jacob et al., 2014; Sreejesh et al., 2014). Saris und Gallhofer (2007) zeigten, dass die Verwendung unipolarer Skalen für bipolare Konstrukte die Reliabilität nicht mindert, was die Verwendung rein positiver Zahlen für semantische Differentiale begründen würde.

Abgesehen von numerischen Bezeichnungen der Abstufungen können auch andere Arten der Beschriftung in Betracht gezogen werden. Empirisch deuten einige Arbeiten darauf hin, dass bei Skalen mit verbaler Beschriftung der Endstufen extreme Antworten wahrscheinlicher sind als bei Beschriftung aller Abstufungen (DeCastellarnau, 2018; Eutsler & Lang, 2015; Moors et al., 2014). Darüber hinaus belegt eine Mehrheit der Studien, dass vollständige verbale Beschriftungen sich günstig auf die Reliabilität auswirken (Alwin, 2007; Alwin & Krosnick, 1991; Krosnick & Berent, 1993; Menold et al., 2014; Saris & Gallhofer, 2007). Manche Autor*innen empfehlen eine Kombination verbaler und numerischer Beschriftungen, um zum Beispiel die wahrgenommene Äquidistanz der Abstufungen zu verbessern (Krosnick & Fabrigar, 1997; Rohrmann, 1978, 2007). Eine Arbeit von Moors et al. (2014) zeigte, dass auch das Einfügen von Zahlen zusätzlich zu verbalen Beschriftungen der Endstufen im Vergleich zu Beschriftungen der Endstufen ohne numerische Bezeichnungen die Tendenz zu extremen Antworten verringert. Andere Studien bestätigen jedoch nicht, dass numerische Bezeichnungen einen Vorteil bringen (Christian et al., 2009; Tourangeau et al., 2000).

Insgesamt spricht der Forschungsstand für den Einsatz verbaler Bezeichnungen für alle Abstufungen (Alwin, 2007; Alwin & Krosnick, 1991; DeCastellarnau, 2018; Krosnick & Berent, 1993; Menold et al., 2014; Moors et al., 2014; Saris & Gallhofer, 2007). Auch eine Studie von Garland (1990) ergab, dass Befragte verbale Bezeichnungen

semantischer Differentiale gegenüber keinen oder numerischen Beschriftungen bevorzugen. Allerdings ist bei verbalen Beschriftungen darauf zu achten, dass Gleichabständigkeit der Stufen nahegelegt wird, weshalb empirisch ermittelt werden sollte, welche Bezeichnungen geeignet sind. Rohrmann (1978, 2007) unterbreitet Vorschläge, welche Beschriftungen empirisch nachweislich als gleichabständig wahrgenommen werden und somit eingesetzt werden können. Diese Empfehlungen beziehen sie sich jedoch auf fünf Abstufungen. Der entwickelte UX-Fragebogen für Medizinprodukte weist zum einen sieben Abstufungen auf, zum anderen soll er international eingesetzt werden, weshalb eine Übersetzung der Beschriftungen in diverse Sprachen erforderlich wäre. Zahlen lassen sich universell einsetzen (Sauro & Lewis, 2020), jedoch lässt sich auf Grundlage bisheriger Forschung kein eindeutiger Mehrwert durch den Einsatz von numerischen Beschriftungen feststellen (DeCastellarnau, 2018). Da auch der UEQ+ (Schrepp & Thomaschewski, 2019a) keine Beschriftung der Ratingstufen enthält, wird zunächst beschlossen, für den UX-Fragebogen für Medizinprodukte ebenfalls keine Beschriftungen einzusetzen.

4.6.2 Verbesserung der Differenzierungsfähigkeit der Items.

Durch das Weglassen der numerischen Bezeichnungen der Ratingstufen könnte sich die Schwierigkeit der Items günstig verändern. Wie oben beschrieben, werden negative Werte tendenziell vermieden (Kunz, 2015; Schwarz et al., 1991; Schwarz & Hippler, 1995). Wenn es keine negativen Bezeichnungen über den Abstufungen mehr gibt, könnte die möglicherweise vorliegende Verzerrung der Antworten in die positive Skalenrichtung aufgehoben werden. Alle Items, sowohl jene der bestehenden UEQ+ Skalen als auch die neu entwickelten Items weisen einen ähnlichen, leichten Schwierigkeitsbereich auf, worunter ihre Differenzierungsfähigkeit leidet. Diese ist indessen wichtig, um Produkte abzugrenzen und Verbesserungen im Verlauf der Produktentwicklung feststellen zu können.

Ein möglicher weiterer Ansatzpunkt zur Verbesserung der Differenzierungsfähigkeit könnte neben dem Weglassen der Bezeichnungen eine Anpassung der Instruktion vor der Bearbeitung des UX-Fragebogens sein, sodass verdeutlicht wird, dass die Mitte der Skala als neutral zu betrachten ist. Wenn ein Produkt genauso gut abschneidet wie andere, beispielsweise indem es vergleichbar adäquate Ergebnisse liefert, sollte die Mitte der Skala, nicht der rechte Pol angekreuzt werden. Hilfreich könnten auch Beispielanker sein, welche für ausgewählte Items, zum Beispiel für ein Item jeder Skala präsentiert werden könnten. Eine bessere Differenzierungsfähigkeit könnte unter Umständen auch durch mehr Abstufungen der Ratingskala erreicht werden. Zum Beispiel Leung (2011) empfiehlt die Verwendung 11-stufiger Skalen, weil neben einer höheren Genauigkeit auch

eine Annäherung an Intervallskalenniveau gegeben ist. Allerdings erfordern mehr Abstufungen eine höhere Differenzierungsfähigkeit der Befragten, weshalb der allgemeine Konsens fünf bis sieben Abstufungen nahelegt (Steiner & Benesch, 2021).

Mithilfe der Vorschläge könnte erreicht werden, dass die Items weniger leicht ausfallen, die Schwäche der geringen Schwierigkeitsunterschiede wäre jedoch weiterhin anzunehmen. Es sollte deshalb überlegt werden, ob nicht für einige Items extremere Begriffe als positive Pole eingesetzt werden sollten. Eine Zustimmung sollte dann weniger wahrscheinlich sein und einige Items würden vermutlich schwerer ausfallen. Die Entwicklung neuer Items würde allerdings eine erneute Evaluation an einer neuen Stichprobe erfordern. Es ist deshalb ratsam, zunächst zu überprüfen, ob sich die Items in der Praxis auch so bewähren.

4.6.3 Weitere Maßnahmen zur Sicherung der Güte.

Im Abschnitt 4.4 (Limitationen) wurde erläutert, dass die Güte des entwickelten UX-Fragebogens für Medizinprodukte bislang nicht umfänglich gesichert werden konnte. Zum einen sollte die Verständlichkeit der neu entwickelten Items mithilfe qualitativer Informationen überprüft werden, um Augenscheinvalidität sicherzustellen. Es sollte ermittelt werden, ob die Items von verschiedenen Testpersonen gleich interpretiert werden. Falls nicht, müssen sie unter Umständen angepasst und erneut evaluiert werden. Zum anderen sollte die Struktur und Güte des produktübergreifenden UX-Fragebogens erneut getestet werden, sobald Daten für alle Skalen, inklusive der Skala Nützlichkeit, vorliegen. Eine geeignete Methode, um der Vermischung von Personen-, Produkt- und Kontexteigenschaften (Hassenzahl et al., 2009; Hassenzahl, 2011; Hassenzahl & Tractinsky, 2006; Rummel & Schrepp, 2018) entgegenzuwirken, könnte darin bestehen, jede Testperson mehrere Produkte unter Konstanthaltung der Umgebung beurteilen zu lassen. So könnte sichergestellt werden, dass sich Unterschiede in den Skalenwerten tatsächlich auch auf Produkteigenschaften zurückführen lassen, die für den Designprozess wesentlich sind. Bisher wurden nur verschiedene Hersteller in Bezug auf die mittlere Produktbewertung verglichen, zur Gewährleistung von diskriminanter Validität ist das oben beschriebene Vorgehen jedoch aussagekräftiger. Die konvergente Validität wurde bereits insofern betrachtet, dass Korrelationen zwischen den UX-Skalen und der Gesamtzufriedenheit mit dem Produkt berechnet wurden. Was darüber hinaus interessant wäre, sind Zusammenhänge mit externen Kriterien, die praktisch relevant sind und Anhaltspunkte für Designer*innen liefern könnten. Denkbar wäre zum Beispiel ein Zusammenhang zwischen der

UX und der Bildqualität. Es wäre zu erwarten, dass die UX, insbesondere die wahrgenommene Effektivität, sich mit höherer Auflösung verbessert.

Die Berechnung psychometrischer Gütekriterien gibt wertvolle Anhaltspunkte in Bezug auf die Qualität des entwickelten UX-Fragebogens für Medizinprodukte. Es gilt jedoch, wie bereits mehrfach verdeutlicht, zu beachten, dass sich gängige Maßnahmen zur Sicherung der Güte aus der psychologischen Testkonstruktion eventuell nicht ohne Weiteres auf UX-Fragebögen übertragen lassen. Da geeignete, standardisierte Methoden die Grundlage für qualitativ hochwertige UX-Fragebögen bilden, sollten sie in Bezug auf UX-Fragebögen genauer erforscht und hinterfragt werden.

4.6.4 Weiterführende Schritte bei Siemens Healthineers.

In den Entwicklungsprozess des UX-Fragebogens für Medizinprodukte floss eine Vielzahl an verschiedenen Medizinprodukten ein, die von Siemens Healthineers und anderen Firmen hergestellt werden. Die Produktkategorie Ultraschall, die ebenfalls bei Siemens Healthineers vertreten ist, wurde indessen nicht einbezogen, da sie sich keiner Business Line zuordnen lässt. Es muss noch überprüft werden, ob sich der UX-Fragebogen auch auf Ultraschall-Geräte anwenden lässt, was aufgrund der allgemein gehaltenen Formulierung der Items aber anzunehmen ist. Neben Ultraschall-Geräten blieb eine Besonderheit mancher Produkte bislang ebenfalls außen vor: Sprachsteuerungsfunktionen. Der UEQ+ beinhaltet seit 2020 Skalen zu Sprachsteuerungen (Klein et al., 2020). Allerdings gaben in der Vorstudie von Müßig (2022) nur knapp 10 % aller Befragten an, dass ihr Gerät über eine solche Funktion verfügt, etwa ein Viertel war sich unsicher. Unter diesen Personen berichteten 76 % mit Sprachsteuerungsfunktionen überhaupt nicht oder nur geringfügig vertraut zu sein. Momentan sind Sprachassistenzen in Bezug auf Medizinprodukte somit noch nicht weit verbreitet oder werden wenig genutzt. Das könnte sich aber in Zukunft ändern, wenn Sprachsteuerung im Alltag Einzug hält und die Bedienweise vertrauter wird. Dann sollte überprüft werden, ob die Skalen zur Sprachsteuerung ergänzend eingesetzt werden sollten.

Eine Ausweitung des UX-Fragebogens für Medizinprodukte ist indessen nicht nur auf weitere Produktkategorien, sondern ebenso auf andere Länder vorgesehen. Die Siemens Healthcare GmbH ist international tätig, daher steht als nächstes die Übersetzung der neu entwickelten Items in andere Sprachen an. Die Skalen des UEQ+ stehen bereits in zahlreichen Sprachen zur Verfügung (Schrepp & Thomaschewski, 2019d), eine Übersetzung ist somit nur für die neuen Items der Skala Effektivität und der Sicherheitsskalen nötig. Um sprachliche Feinheiten einzufangen, sollten die Items sowie auch die

Einleitungssätze in die jeweils andere Sprache übersetzt und anschließend rückübersetzt werden, im Idealfall von bilingualen Muttersprachler*innen (Brislin, 1970). So kann dieselbe Bedeutung der Skalen auf allen Sprachen sichergestellt werden.

Neben beziehungsweise nach einer Übersetzung der Items erfolgt außerdem die Erhebung einer Benchmark. Eine Benchmark ist zentral, um mittels des UX-Fragebogens erhobene Werte einordnen und interpretieren zu können (Schrepp et al., 2017). Nur so lässt sich beurteilen, ob beispielsweise ein neu entwickeltes MRT-Gerät im Vergleich zu bisherigen Produkten auf dem Markt gut abschneidet. Dafür muss der entwickelte UX-Fragebogen im Rahmen möglichst vieler Usability- und UX-Testungen eingesetzt werden, sodass eine umfangreiche, repräsentative Datenbank von Produktbewertungen entsteht, die als Vergleichsmaßstab dienen kann. Nach Müßigs Vorstudie (2022) wurde bei Siemens Healthineers bereits damit begonnen, die produktübergreifend wichtigsten Skalen des UEQ+ (Steuerbarkeit, Effizienz, Übersichtlichkeit, Vertrauen) sowie optionale Zusatzskalen, die sich in der Vorstudie für die einzelnen Business Lines als wichtig herausstellten, bei Usability- und UX-Testungen einzusetzen. Bei ersten Tests konnte im Verlauf der Weiterentwicklung eines Prototyps anhand der wichtigsten UEQ+ Skalen eine Verbesserung der UX festgestellt werden, wenn auch nur an einer sehr kleinen Stichprobe von sieben Personen. Trotzdem gibt dieses Ergebnis einen ersten Hinweis darauf, dass der UX-Fragebogen eine Verbesserung der UX aufgrund von vorgenommenen Veränderungen am Produkt durchaus erfassen kann. Für Verlaufskontrollen kann der UX-Fragebogen somit bereits herangezogen werden. Mit dem Ziel, verschiedene Modelle auch untereinander vergleichen zu können, ist der nächste Schritt die Erhebung einer umfangreichen Benchmark.

Es wird darüber hinaus bereits daran gearbeitet, die Auswertung des UX-Fragebogens zu vereinfachen, sodass er häufiger eingesetzt wird und die Erhebung der Benchmark schnell vonstattengeht. Schrepp und Thomaschewski (2019d) stellen auf der Webseite zum UEQ+ ein Excel-Tool zur Auswertung zur Verfügung. Hier müssen nur die Itemwerte jeder Testperson eingegeben werden, woraufhin die deskriptiven Itemstatistiken, Cronbachs α sowie ein *Key Performance Indicator*, kurz KPI (Hinderks, Schrepp et al., 2019), automatisch ausgegeben werden. Dieser KPI berechnet sich aus der relativen Wichtigkeit der Skalen für jede*n Proband*in multipliziert mit dem jeweiligen Skalenwert. Diese Werte werden wiederum gemittelt, sodass man einen Gesamtwert für die UX über alle Skalen und alle Proband*innen erhält. Der KPI gibt zwar keinen Aufschluss darüber, wie gut die einzelnen UX-Aspekte erfüllt sind, in der Praxis kann es aber

hilfreich sein, einen Kennwert für die Gesamt-UX zu haben. Aus diesem Grund wird für den UX-Fragebogen für Medizinprodukte ebenfalls ein solches Excel-Tool entworfen. Für Designer*innen ist der KPI dagegen weniger aufschlussreich, da er nur einen Vergleich der Produkte oder eine Fortschrittsmessung erlaubt, aber keine Rückschlüsse ermöglicht, welche UX-Aspekte verbessert werden müssen. Qualitative Informationen sollten deshalb nicht vernachlässigt werden, denn nur mit ihnen kann ein umfassendes Bild der UX einer Zielgruppe abgeleitet werden. Quantitative und qualitative Methoden sollten grundsätzlich kombiniert eingesetzt werden (Darin et al., 2019; Robinson et al., 2018; Sauro, 2016). Besonders hilfreich könnte eine Kombination des konstruierten UX-Fragebogens und des Protokolls sein, das im Rahmen des Projekts 2015/16 (Braun, 2016) entwickelt wurde (s. Kapitel 1.4.2). Das Protokoll wurde speziell für UX-Testungen von Medizinprodukten entwickelt und sieht zum Beispiel eine Videoaufzeichnung des Interaktionsablaufs vor, sodass auch qualitative Beobachtungsdaten verfügbar sind.

4.7 Fazit

Mithilfe des entwickelten Fragebogens wird es in Zukunft möglich sein, die User Experience von Medizinprodukten, welche von medizinischem Personal bedient werden, produktübergreifend, standardisiert und ökonomisch zu erfassen. Medizinproduktübergreifend relevante, durch den UX-Fragebogen abgedeckte Skalen sind Steuerbarkeit, Übersichtlichkeit, Effizienz, Vertrauen, Effektivität und Nützlichkeit. Der UX-Fragebogen stellt eine Ergänzung zu qualitativen Methoden dar und wird dazu beitragen, die UX der Produkte gezielt zu verbessern, wovon medizinisches Personal sowie auch Patient*innen profitieren können. Die psychometrischen Gütekriterien des Fragebogens belegen größtenteils seine Eignung, wobei in einigen Bereichen weitere Untersuchungen angebracht sind. Insgesamt stellt sich die Frage, inwieweit sich Prinzipien der psychologischen Testtheorie auf die Konstruktion von UX-Fragebögen anwenden lassen, was zukünftig genauer betrachtet werden sollte. Der nächste Schritt, um Ergebnisse des UX-Fragebogens einordnen zu können, besteht in der Erhebung einer umfangreichen Benchmark als Vergleichsmaßstab. Weiterhin wird eine Übersetzung der Items in andere Sprachen und die Erstellung eines Excel-Tools zur Unterstützung der Auswertung erfolgen. Der UX-Fragebogen wurde auf Grundlage einer Vielfalt an Medizinprodukten entwickelt. Seine Anwendbarkeit auf weitere Medizinprodukte sollte in Zukunft geprüft werden. Darüber hinaus könnten die entwickelten Zusatzskalen eine Ergänzung des Repertoires an Skalen des User Experience Questionnaire + (Schrepp & Thomaschewski, 2019a) darstellen und somit auch für andere Produktklassen und Branchen einsetzbar werden.

5 Literaturverzeichnis

Aichholzer, J. (2017). *Einführung in lineare Strukturgleichungsmodelle mit Stata*. *Lehrbuch*. Springer. https://doi.org/10.1007/978-3-658-16670-0

Almquist, E., Senior, J., & Bloch, N. (2016). The elements of value. *Harvard Business Review*, 94(9), 47–53. https://www.bain.com/insights/the-elements-of-value-hbr

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. *Wiley series in survey methodology*. Wiley. https://doi.org/10.1002/9780470146316

Alwin, D. F., & Krosnick, J. (1991). The Reliability of Survey Attitude Measurement. *Sociological Methods & Research*, *20*(1), 139–181. https://doi.org/10.1177/0049124191020001005

American Psychological Association. (2022). *APA style numbers and statistics guide*. https://apastyle.apa.org/instructional-aids/numbers-statistics-guide.pdf

Andresen, B., & Beauducel, A. (2008). TBS-TK Rezension: NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R). *Report Psychologie*, *33*(10), 543–544. https://www.bdp-verband.de/binaries/content/assets/beruf/testrezensionen/neopir.pdf

Ariza, N., & Maya, J. (2014). Proposal to identify the essential elements to construct a user experience model with the product using the thematic analysis technique. In D. Marjanović, M. Štorga, N. Pavković & N. Bojčetić (Hrsg.), *Proceedings of the DESIGN 2014 13th International Design Conference* (S. 11–22). The Design Society. https://www.academia.edu/72704134

Awang, Z., Afthanorhan, A., & Mamat, M. (2016). The Likert scale analysis using parametric based Structural Equation Modeling (SEM). *Computational Methods in Social Sciences*, *4*(1), 13–21. https://eprints.unisza.edu.my/5106/

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole & W. A. Kellogg (Hrsg.), *CHI '11: CHI Conference on Human Factors in Computing Systems* (S. 2689–2698). Association for Computing Machinery. https://doi.org/10.1145/1978942.1979336

Berger, U. (2019). Interne Konsistenz von Fragebogen-Skalen: Alpha oder Omega? *Psychotherapie, Psychosomatik, medizinische Psychologie*, *69*(8), 348–349. https://doi.org/10.1055/a-0878-1270

Berni, A., & Borgianni, Y. (2021). From the definition of user experience to a framework to classify its applications in design. In Proceedings of the Design Society (Hrsg.), *Proceedings of the International Conference on Engineering Design (ICED21)* (S. 1627–1636). Cambridge University Press. https://doi.org/10.1017/pds.2021.424

Bitkina, O. V., Kim, H. K., & Park, J [Jaehyun]. (2020). Usability and user experience of medical devices: An overview of the current state, analysis methodologies, and future challenges. *International Journal of Industrial Ergonomics*, *76*, 1–11. https://doi.org/10.1016/j.ergon.2020.102932

Boos, B., & Brau, H. (2017). Erweiterung des UEQ um die Dimensionen Akustik und Haptik. In S. Hess & H. Fischer (Hrsg.), *Mensch & Computer 2017 - Usability Professionals.* Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2017-up-0236

Brandt, H. (2020). Exploratorische Faktorenanalyse (EFA). In H. Moosbrugger & A. Kevala (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 575–612). Springer. https://doi.org/10.1007/978-3-662-61532-4_23

Braun, C. (2016). *UX Measurement for Healthcare products: Instrument Development.* Siemens Healthcare GmbH.

Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Brooke, J. (1996). SUS: a 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Hrsg.), *Usability Evaluation in Industry* (S. 189–194). Taylor & Francis. https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2. Aufl.). Guilford Press. https://doi.org/10.5860/choice.44-2769

Budiu, R. (2017, 01. Oktober). *Quantitative vs. Qualitative Usability Testing.* Nielsen Norman Group. https://www.nngroup.com/articles/quant-vs-qual/

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). *Psychologie*. Pearson Studium.

Bundesinstitut für Arzneimittelsicherheit und Medizinprodukte. (2022, 09. September). *Medizinprodukte*. https://www.bfarm.de/DE/Medizinprodukte/_node.html

Burns, A. J., Johnson, M. E., & Honeyman, P. (2016). A brief chronology of medical device security. *Communications of the ACM*, *59*(10), 66–72. https://doi.org/ 10.1145/2890488

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Cattell, R. B. (1978). *Scientific Use of Factor Analysis in Behavioral and Life Sciences* (1. Aufl.). Springer. https://doi.org/10.1007/978-1-4684-2262-7

Chin, J. P., Diehl, V. A., & Norman, L. K. (2014). Development of an instrument measuring user satisfaction of the human-computer interface. In M. Jones, P. Palanque, A. Schmidt & T. Grossman (Hrsg.), *CHI '14: CHI Conference on Human Factors in Computing Systems* (S. 213–218). Association for Computing Machinery. https://doi.org/10.1145/57167.57203

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha: Well known but poorly understood. *Organizational Research Methods*, *18*(2), 207–230. https://doi.org/10.1177/1094428114555994

Christian, L. M., Parsons, N. L., & Dillman, D. A. (2009). Designing Scalar Questions for Web Surveys. *Sociological Methods & Research*, *37*(3), 393–425. https://doi.org/10.1177/0049124108330004

Clark, S., & Israelski, E. (2012). Total Recall: The Consequence of Ignoring Medical Device Usability User Experience Magazine. *User Experience Magazine*, *11*(1). https://uxpamagazine.org/total-recall/

Cohen, J. (1992). Quantitative methods in psychology: A Power Primer. *Psychological bulletin*, *112*(1), 155–159. http://users.cla.umn.edu/~nwaller/prelim/cohenpowerprimer.pdf

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2. Aufl.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315827506

Conceptboard. (2022). *Conceptboard - Collaborative Online Whiteboard*. https://conceptboard.com/

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews & D. Saklofske (Hrsg.), *The SAGE Handbook of Personality Theory and Assessment: Volume 2 – Personality Measurement and Testing* (S. 179–198). SAGE Publications Ltd. https://doi.org/10.4135/9781849200479.n9

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Darin, T., Coelho, B., & Borges, B. (2019). Which Instrument Should I Use? Supporting Decision-Making About the Evaluation of User Experience. In A. Marcus & W. Wang (Hrsg.), *Design, User Experience, and Usability: Practice and Case Studies* (S. 49–67). Springer. https://doi.org/10.1007/978-3-030-23535-2_4

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Dhillon, B. S. (2012). *Safety and human error in engineering systems*. CRC Press. https://doi.org/10.1201/b12534

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer. https://doi.org/10.1007/978-3-642-41089-5

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989x.5.2.155

Eid, M., Gollwitzer, M., & Schmitt, M. (2013). *Statistik und Forschungsmethoden* (3. Aufl.). Beltz.

Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. *Bachelorstudium Psychologie*. Hogrefe.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press. https://doi.org/10.1017/CBO9780511761676

Eutsler, J., & Lang, B. (2015). Rating Scales in Accounting Research: The Impact of Scale Points and Labels. *Behavioral Research in Accounting*, *27*(2), 35–51. https://doi.org/10.2308/bria-51219

Everitt, B. S., (1975). Multivariate analysis: the need for data, and other problems. *British Journal of Psychiatry*, *126*(3), 237–240. https://doi.org/10.1192/bjp.126.3.237

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989x.4.3.272

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Föhl, U., & Friedrich, C. (2022). Formate von Fragen. In *Quick Guide Onlinefragebogen* (S. 31–55). Springer. https://doi.org/10.1007/978-3-658-36291-1_3

Ford, E. C., & Evans, S. B. (2018). Incident learning in radiation oncology: A review. *Medical physics*, *45*(5), 100–119. https://doi.org/ 10.1002/mp.12800

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In T. Turner, G. Szwillus, M. Czerwinski, F. Paternò & S. Pemberton (Hrsg.), *CHI 2000: Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (S. 345–352). Association for Computing Machinery. https://doi.org/10.1145/332040.332455

Furniss D., Masci P., Curzon P., Mayer A., & Blandford A. (2014). 7 Themes for guiding situated ergonomic assessments of medical devices: a case study of an inpatient glucometer. *Applied Ergonomics*, *45*(6), 1668–1677. https://doi.org/10.1016/j.aperg o.2014.05.012

Gäde, J., Schermelleh-Engel, K., & Brandt, H. (2020). Konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kevala (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl., S. 615–657). Springer. https://doi.org/10.1007/978-3-662-61532-4_24

Gagne, P., & Hancock, G. R. (2006). Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. *Multivariate behavioral research*, *41*(1), 65–83. https://doi.org/10.1207/s15327906mbr4101_5

Garland, R. (1990). A comparison of three forms of the semantic differential. *Marketing Bulletin*, *1*(4), 19–24. http://marketing-bulletin.massey.ac.nz/v1/mb_v1_a4_garland.pdf

Gediga, G., Hamborg, K.-C., & Düntsch, I. (1999). The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems. *Behaviour & Information Technology*, *18*(3), 151–164. https://doi.org/10.1080/014492999119057

GE Healthcare (2022). *Computertomographie.* https://www.gehealthcare.de/Products/Computed-Tomography

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in social & administrative pharmacy*, *9*(3), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.3102/00346543042003237

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Hassenzahl, M. (2011). User Experience and Experience Design. In M. Soegaard & R. F. Dam (Hrsg.), *Encyclopedia of Human-Computer Interaction* (2. Aufl.). The Interaction Design Foundation. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/user-experience-and-experience-design

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003* (S. 187–196). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-80058-9_19

Hassenzahl, M., Eckoldt, K., & Thielsch, M. T. (2009). User Experience und Experience Design – Konzepte und Herausforderungen. In Brau, H., Diefenbach, S., Hassenzahl, M., Kohler, K., Koller, F., Peissner, M., K. Petrovic, M. Thielsch, D. Ullrich & D. Zimmermann (Hrsg.), *Tagungsband UP09* (S. 233–237). Fraunhofer Verlag. https://dl.gi.de/handle/20.500.12116/5497

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Hayes, A. F., & Coutts, J. J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But… *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Heise, D. R. (1969). Some methodological issues in semantic differential research. *Psychological Bulletin*, *72*(6), 406–422. https://doi.org/10.1037/h0028448

Held, T., Schrepp, M., & Mayalidag, R. (2019). User Experience Review: Ein einfaches und flexibles Verfahren zur Beurteilung der User Experience durch Experten. In H. Fischer & S. Hess (Hrsg.), *Mensch & Computer 2019 - Usability Professionals* (S. 4–9). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2019-up-0117

Hinderks, A. (2016). *Modifikation des User Experience Questionnaire (UEQ) zur Verbesserung der Reliabilität und Validität.* https://doi.org/10.13140/RG.2.2.31619.50722

Hinderks, A., Schrepp, M., Mayo, F. J. D., Escalona, M. J., & Thomaschewski, J. (2019). Developing a UX KPI based on the user experience questionnaire.

Computer Standards & Interfaces, *65*, 38–44. https://doi.org/10.1016/j.csi.2019.01.007

Hinderks, A., Winter, D., Thomaschewski, J., & Schrepp, M. (2019). Applicability of User Experience and Usability Questionnaires. *Journal of Universal Computer Science*, *25*(13), 1717–1735. https://doi.org/10.3217/jucs-025-13-1717

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://www.jstor.org/stable/4615733?origin=JSTOR-pdf

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989x.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hyman, W. A. (2018). Errors in the Use of Medical Equipment. In M. S. Bogner (Hrsg.), *Human Error and Safety. Human Error in Medicine* (S. 327–347). CRC Press. https://doi.org/10.1201/9780203751725-15

International Organization for Standardization (2008). *Ergonomics of human-system interaction — Part 151: Guidance on World Wide Web user interfaces* (ISO Standard Nr. 9241-151:2008). https://www.iso.org/standard/37031.html

International Organization for Standardization (2014). *Safety aspects – Guidelines for their inclusion in standards* (ISO/IEC Standard Nr. 51:2014). https://www.iso.org/standard/53940.html

International Organization for Standardization (2015). *Medical devices - Part 1: Application of usability engineering to medical devices* (IEC Standard Nr. 62366-1:2015). https://www.iso.org/standard/63179.html

International Organization for Standardization (2016). *Medical devices - Part 2: Guidance on the application of usability engineering to medical devices* (IEC/TR Standard Nr. 62366-2:2016). https://www.iso.org/standard/69126.html

International Organization for Standardization (2018). *Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts* (ISO Standard Nr. 9241-11:2018). https://www.iso.org/standard/63500.html

International Organization for Standardization (2019a). *Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems* (ISO Standard Nr. 9241-210:2019). https://www.iso.org/standard/77520.html

International Organization for Standardization (2019b). *Medical devices - Application of risk management to medical devices* (ISO Standard Nr. 14971:2019). https://www.iso.org/standard/72704.html

Jacob, R., Heinz, A., & Décieux, J. P. (2014). *Umfrage: Einführung in die Methoden der Umfrageforschung* (3. Aufl.). Oldenbourg Verlag. https://doi.org/10.1524/9783486736175

Jacobson, J., & Meyer, L. (2019). *Praxisbuch Usability und UX* (2. Aufl.). Rheinwerk Computing.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert Scale: Explored and Explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Kashfi, P., Nilsson, A., & Feldt, R. (2017). Integrating User eXperience practices into software development processes: implications of the UX characteristics. *PeerJ Computer Science*, *3*, 1–40. https://doi.org/10.7717/peerj-cs.130

Kennedy, B. (2018, 17. Mai). *How good UX can save lives.* User Zoom. https://www.userzoom.com/ux-library/how-good-ux-can-save-lives/

Kim, H. K., Jeong, H., Park, J [Jangwoon], Park, J [Jaehyun], Kim, W.-S., Kim, N., Park, S., & Paik, N.-J. (2022). Development of a Comprehensive Design Guideline to Evaluate the User Experiences of Meal-Assistance Robots considering Human-Machine Social Interactions. *International Journal of Human–Computer Interaction*, *38*(1), 1–14. https://doi.org/10.1080/10447318.2021.2009672

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and Usage. In P. W. Jordan, B. Thomas, I. L. McClelland & B. Weerdmeester (Hrsg.), *Usability Evaluation in Industry* (S. 169–178). CRC Press. https://doi.org/10.1201/9781498710411

Kirakowski, J., & Corbett, M. (1993). SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology*, *24*(3), 210–212. https://doi.org/10.1111/j.1467-8535.1993.tb00076.x

Klein, A. M., Hinderks, A., Schrepp, M., & Thomaschewski, J. (2020). Construction of UEQ+ scales for voice quality: measuring user experience quality of voice interaction. In B. Preim, A. Nürnberger & C. Hansen (Hrsg.), *MuC '20: Proceedings of the Conference on Mensch und Computer* (S. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3404983.3410003

Knapp, T. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing Research*, *39*(2), 121–123. https://doi.org/10.1097/00006199-199003000-00019

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kopp, J., & Lois, D. (2014). *Sozialwissenschaftliche Datenanalyse: Faktorenanalyse und Skalierung* (2. Aufl.). Springer. https://doi.org/10.1007/978-3-658-02300-3_5

Kothgassner, O. D., Felnhofer, A., Hauk, N., Kastenhofer, E., Gomm, J., & Kryspin-Exner, I. (2013). *Technology Usage Inventory (TUI): Manual.* Österreichische Forschungsförderungsgesellschaft. https://www.researchgate.net/publication/259292979

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format. *American Journal of Political Science*, *37*(3), 941–964. https://doi.org/10.2307/2111580

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In L. E. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Hrsg.), *Wiley Series in Probability and Statistics. Survey Measurement and Process Quality* (S. 141–164). Wiley. https://doi.org/10.1002/9781118490013.ch6

Kühnel, S. M. (1993). Lassen sich ordinale Daten mit linearen Strukturgleichungsmodellen analysieren? *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, *33*, 29–51. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-201496

Kunz, T. (2015). *Rating Scales in Web Surveys: A Test of New Drag-and-Drop Rating Procedures* [Dissertation]. Technische Universität Darmstadt. https://tuprints.ulb.tu-darmstadt.de/5151/

Kuzon, W. M., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery*, *37*(3), 265–272. https://doi.org/10.1097/00000637-199609000-00006

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior*, *43*, 35–48. https://doi.org/10.1016/j.chb.2014.10.048

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria. *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Lang, A. R., Martin, J. L., Sharples, S., & Crowe, J. A. (2013). The effect of design on the usability and real world effectiveness of medical devices: a case study with adolescent users. *Applied ergonomics*, *44*(5), 799–810. https://doi.org/10.1016/j.apergo.2013.02.001

Lanius, C., Weber, R., & Robinson, J. (2021). User Experience Methods in Research and Practice. *Journal of Technical Writing and Communication*, *51*(4), 350–379. https://doi.org/10.1177/00472816211044499

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In R. Nugent (Hrsg.), *HCI and Usability for Education and Work* (S. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Law, E., Roto, V., Vermeeren, A. P. O. S., Kort, J., & Hassenzahl, M. (2008). Towards a shared definition of user experience. In M. Czerwinski, A. Lund & D. Tan (Hrsg.), *CHI EA '08: CHI '08 Extended Abstracts on Human Factors in Computing Systems* (S. 2395–2398). Association for Computing Machinery. https://doi.org/10.1145/1358628.1358693

Leung, S.-O. (2011). A Comparison of Psychometric Properties and Normality in 4-, 5-, 6-, and 11-Point Likert Scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*(6), 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lim, J. H., Rhie, Y. L., & Park, J [Jaehyun] (2019). Applying Semantic Network Analysis to Develop User Experience Assessment Model for Smart TV. *Applied Sciences*, *9*(16), 1–16. https://doi.org/10.3390/app9163307

Lin, H. X., Choong, Y.-Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology*, *16*(4-5), 267–277. https://doi.org/10.1080/014492997119833

LinkedIn. (2022). *Willkommen in deiner beruflichen Community*. https://www.linkedin.com

Lüdtke, O., & Robitzsch, A. (2010). Umgang mit fehlenden Daten in der empirischen Bildungsforschung. In S. Maschke & L. Stecher (Hrsg.), *Enzyklopädie Erziehungswissenschaften Online*. Beltz Juventa. https://doi.org/10.3262/EEO07100150

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989x.4.1.84

Marucci, V. (2019, 07. Januar). *Leitfaden zum Identifizieren geeigneter UX-Methoden.* Testing Time. https://www.testingtime.com/blog/leitfaden-ux-methoden/

McDonald, R. P. (1970). The Theoretical Foundations of Principal Factor Analysis, Canonical Factor Analysis, and Alpha Factor Analysis. *British Journal of Mathematical and Statistical Psychology*, *23*(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410601087

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Medicines & Healthcare products Regulatory Agency (2021). *Safety Guidelines for Magnetic Resonance Imaging Equipment in Clinical Use*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/958486/MRI_guidance_2021-4-03c.pdf

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How Do Respondents Attend to Verbal Labels in Rating Scales? *Field Methods*, *26*(1), 21–39. https://doi.org/10.1177/1525822X13508270

Mentler, T. (2021). Usability Engineering und User Experience Design sicherheitskritischer Systeme. In C. Reuter (Hrsg.), *Sicherheitskritische Mensch-Computer-Interaktion: Interaktive Technologien und Soziale Medien im Krisen- und Sicherheitsmanagement* (2. Aufl., S. 47–66). Springer Vieweg. https://doi.org/10.1007/978-3-658-32795-8_3

Miclăuş, T., Valla, V., Koukoura, A., Nielsen, A. A., Dahlerup, B., Tsianos, G. I., & Vassiliadis, E. (2020). Impact of design on medical device safety. *Therapeutic Innovation & Regulatory Science*, *54*(4), 839–849. https://doi.org/10.1007/s43441-019-00022-4

Microsoft. (2022). *Microsoft Teams*. https://www.microsoft.com/de-de/microsoft-teams

Minge, M., & Riedel, L. (2013). meCUE – Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In S. Boll, S. Maaß & R. Malaka (Hrsg.), *Mensch & Computer 2013 - Tagungsband* (S. 89–98). Oldenbourg Wissenschaftsverlag. https://doi.org/10.1524/9783486781229.89

Minge, M., & Thüring, M. (2011). Der Einfluss von Gebrauchstauglichkeit und Gestaltung auf Kognition und Emotion. In C. Lutsch & F. Adler (Hrsg.), *Der kurze Weg zum Glück* (S. 109–119). Forum für Gestalten. https://www.researchgate.net/publication/263733282

Mircioiu, C., & Atkinson, J. (2017). A Comparison of Parametric and Non-Parametric Methods Applied to a Likert Scale. *Pharmacy*, *5*(2), 1–12. https://doi.org/10.3390/pharmacy5020026

Mistry, A., & Rajan, A. P. (2019). Evaluation of Web Applications based on UX Parameters. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(4), 2564–2570. https://doi.org/10.11591/ijece.v9i4

Money, A. G., Barnett, J., Kuljis, J., Craven, M. P., Martin, J. L., & Young, T. (2011). The role of the user within the medical device design and development process: medical device manufacturers' perspectives. *BMC medical informatics and decision making*, 11(1), 1–12. https://doi.org/10.1186/1472-6947-11-15

Montag, K. (2012). *Untersuchung zur Gebrauchstauglichkeit von Medizinprodukten im Anwendungsbereich OP* [Dissertation]. Eberhard-Karls-Universität Tübingen. https://d-nb.info/116051786x/34

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The Effect of Labeling and Numbering of Response Scales on the Likelihood of Response Bias. *Sociological Methodology*, *44*(1), 369–399. https://doi.org/10.1177/0081175013516114

Moosbrugger, H., & Kevala, A. (Hrsg.). (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Springer. https://doi.org/10.1007/978-3-662-61532-4

Moshagen, M., & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies*, *68*(10), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Murray, J. (2013). Likert data: what to use, parametric or non-parametric? *International Journal of Business and Social Science*, *4*(11), 258–264. https://www.academia.edu/77150308

Müßig, T. (2022). *The challenge of creating a unified User Experience Questionnaire for medical devices* [unveröffentlichte Masterarbeit]. Otto-Friedrich-Universität, Bamberg.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In B. Adelson, S. Dumais & J. Olson (Hrsg.), *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (S. 152–158). Association for Computing Machinery. https://doi.org/10.1145/191666.191729

Nielsen, J. (1994, 01. November). *How to conduct a heuristic evaluation.* Nielsen Norman Group. https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/

Nielsen, J. (2020, 15. November). *10 Usability Heuristics for User Interface Design.* Nielsen Norman Group. https://www.nngroup.com/articles/ten-usability-heuristics/

Norman, D., Miller, J., & Henderson, A. (1995). What you see, some of what's in the future, and how we go about doing it. In J. Miller (Hrsg.), *ACM Conferences, Conference Companion on Human Factors in Computing Systems* (S. 155). Association for Computing Machinery. https://doi.org/10.1145/223355.223477

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Nunnally, J. C. (1978). *Psychometric theory* (2. Aufl.). McGraw-Hill.

Oja, H. (2010). *Multivariate Nonparametric Methods with R: An approach based on spatial signs and ranks*. *Lecture Notes in Statistics*. Springer. https://doi.org/10.1007/978-1-4419-0468-3

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1975). *The measurement of meaning* (9. Aufl.). University of Illinois Press.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung*. Hogrefe.

Otten, R., Schrepp, M., & Thomaschewski, J. (2020). Visual clarity as mediator between usability and aesthetics. In B. Preim, A. Nürnberger & C. Hansen (Hrsg.), *MuC '20: Proceedings of the Conference on Mensch und Computer* (S. 11–15). Association for Computing Machinery. https://doi.org/10.1145/3404983.3409990

Patel, V. L., & Zhang, J., (2007). Patient safety in health care. In F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandowsky & T. J. Perfect (Hrsg.), *Handbook of applied cognition* (S. 307–331). Wiley. https://doi.org/10.1002/9780470713181.ch12

Philips GmbH. (2022). *Ingenia MR-RT*. https://www.philips.de/healthcare/product/HC781439/Ingenia-MR-RT-Simulation-Plattform#documents

Pickup, L., Nugent, B., & Bowie, P. (2019). A preliminary ergonomic analysis of the MRI work system environment: Implications and recommendations for safety and design. *Radiography*, *25*(4), 339–345. https://doi.org/10.1016/j.radi.2019.04.001

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: factor recovery with small sample sizes. *Behavior Genetics, 32*(2), 153–161. https://doi.org/10.1023/A:1015210025234

Prümper, J., & Anft, M. (1993). Die Evaluation von Software auf Grundlage des Entwurfs zur internationalen Ergonomie-Norm ISO 9241 Teil 10 als Beitrag zur partizipativen Systemgestaltung — ein Fallbeispiel. In K. Rödiger (Hrsg.), *Software-Ergonomie '93: Berichte des German Chapter of the ACM.* (S. 145–156). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-82972-6_12

QuestionPro (2022). *Plattform für Marktforschung und Experience Management.* https://www.questionpro.de/

Rahn, J. (2010). *Usability und User Experience* [Seminararbeit]. Universität Ulm. https://www.mathematik.uni-ulm.de/sai/ss10/guidesign/downloads/ux_arbeit.pdf

Rammstedt, B. (2010). Reliabilität, Validität, Objektivität. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 239–258). Springer. https://doi.org/10.1007/978-3-531-92038-2_11

Reichheld, F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–54. https://hbr.org/2003/12/the-one-number-you-need-to-grow

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Richter, M., & Flückiger, M. (2016). *Usability und UX kompakt: Produkte für Menschen*. Springer. https://doi.org/10.1007/978-3-662-49828-6

Robinson, J., Lanius, C., & Weber, R. (2018). The past, present, and future of UX empirical research. *Communication Design Quarterly, 5*(3), 10–23. https://doi.org/10.1145/3188173.3188175

Rohrer, C. (2022, 17. Juli). *When to Use Which User-Experience Research Methods.* Nielsen Norman Group. https://www.nngroup.com/articles/which-ux-research-methods/

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie, 9*(3), 222–245. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=12743193

Rohrmann, B. (2007). *Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data* [Projektbericht]. University of Melbourne. http://www.rohrmannresearch.net/pdfs/rohrmann-vqs-report.pdf

Rose, T., Blake, C., Norris, B., & Lowe, C. (2018). Improving Patient Safety through Device Design and Usability. In P. D. Bust (Hrsg.), *Contemporary Ergonomics 2007: Proceedings of the International Conference on Contemporary Ergonomics (CE2007)* (S. 469–474). CRC Press. https://doi.org/10.1201/9781315106595-74

Rossiter, J. R. (2011). *Measurement for the Social Sciences: The C-OAR-SE Method and Why It Must Replace Psychometrics*. Springer. https://doi.org/10.1007/978-1-4419-7158-6

Roto, V., Law, E., Vermeeren, A., & Hoonhout, J. (2011). *User Experience White Paper – Bringing clarity to the concept of user experience*. https://www.academia.edu/853674

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rummel, B., & Schrepp, M. (2018). UX-Fragebögen: Was steckt in der Varianz? In R. Dachselt & G. Weber (Hrsg.), *Mensch & Computer 2018 - Workshopband* (S. 769–777). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2018-ws16-0337

Russell-Rose, T., Blake, C., Norris, B., & Lowe, C. (2007). Improving patient safety through device design and usability. In P. D. Bust (Hrsg.), *Proceedings of the International Conference on Contemporary Ergonomics* (S. 469–474). Taylor & Francis. https://www.researchgate.net/publication/272023491

Saris, W. E. & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. *Wiley series in survey methodology*. Wiley. https://doi.org/10.1002/9780470165195

Saris, W. E., Satorra, A. & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Sauro, J. (2015). SUPR-Q: A comprehensive measure of the quality of the website user experience. *Journal of Usability Studies*, *10*(2), 68–86. https://doi.org/10.5555/2817315.2817317

Sauro, J. (2016). The challenges and opportunities of measuring the user experience. *Journal of Usability Studies*, *12*(1), 1–7. https://doi.org/10.5555/3040226.3040227

Sauro, J., & Lewis, J. (2020, 01. Januar). *Should All Scale Points Be Labeled?* MeasuringU. https://measuringu.com/scale-points-labeled/

Schnaudt, C. (2020). Explorative Faktorenanalyse und Skalenkonstruktion. In M. Tausendpfund (Hrsg.), *Fortgeschrittene Analyseverfahren in den Sozialwissenschaften: Ein Überblick* (S. 205–242). Springer. https://doi.org/10.1007/978-3-658-30237-5_7

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(4), 40–44. https://doi.org/10.9781/ijimai.2017.445

Schrepp, M., & Rummel, B. (2018). UX Fragebögen: Verwenden wir die richtigen Methoden? In R. Dachselt & G. Weber (Hrsg.), *Mensch und Computer 2018 - Workshopband* (S. 725–731). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2018-ws16-0325

Schrepp, M., & Thomaschewski, J. (2019a). *Construction and first Validation of Extension Scales for the User Experience Questionnaire (UEQ)* [Projektbericht]. Hochschule Emden/Leer. https://doi.org/10.13140/RG.2.2.19260.08325

Schrepp, M., & Thomaschewski, J. (2019b). Design and Validation of a Framework for the Creation of User Experience Questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(7), 88–95. https://doi.org/10.9781/ijimai.2019.06.006

Schrepp, M., & Thomaschewski, J. (2019c). Eine modulare Erweiterung des User Experience Questionnaire. In H. Fischer & S. Hess (Hrsg.), *Mensch & Computer 2019 – Usability Professionals* (S. 148–156). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2019-up-0108

Schrepp, M., & Thomaschewski, J. (2019d). *UEQ+: A Modular Extension of the User Experience Questionnaire*. http://ueqplus.ueq-research.org/

Schrepp, M., & Thomaschewski, J. (2020, 21. Juli). *Handbook for the modular extension of the User Experience Questionnaire*. http://ueqplus.ueq-research.org/Material/UEQ+_Handbook_V2.pdf

Schwarz, N., & Hippler, H.-J. (1995). The numeric values of rating scales: a comparison of their impact in mail surveys and telephone interviews. *International Journal of Public Opinion Research*, *7*(1), 72–74. https://doi.org/10.1093/ijpor/7.1.72

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating Scales: Numeric Values May Change the Meaning of Scale Labels. *Public Opinion Quarterly*, *55*(4), 570. https://doi.org/10.1086/269282

Siemens Healthcare GmbH. (2022a). *How we innovate*. https://www.siemens-healthineers.com/innovations/how-we-innovate

Siemens Healthcare GmbH. (2022b). *Über Siemens Healthineers*. https://www.siemens-healthineers.com/de/about

Siemens Healthcare GmbH. (2022c). *UX-Design: Nutzer*innen im Fokus*. https://www.siemens-healthineers.com/deu/innovations/ux-design-focus-on-the-user

Siemens Healthcare GmbH. (2022d). *Computertomographie*. https://www.siemens-healthineers.com/de/computed-tomography

Sreejesh, S., Mohapatra, S., & Anusree, M. R. (2014). Scales and Measurement. In S. Sreejesh, S. Mohapatra & M. R. Anusree (Hrsg.), *Business Research Methods: An Applied Orientation* (S. 107–142). Springer. https://doi.org/10.1007/978-3-319-00539-3_4

Steiner, E., & Benesch, M. (2021). *Der Fragebogen: Von der Forschungsidee zur SPSS-Auswertung* (6. Aufl.). utb. https://doi.org/10.36198/9783838587882

Subedi, B. P. (2016). Using Likert type data in social science research: Confusion, issues and challenges. *International journal of contemporary applied sciences*, *3*(2), 36–49. https://www.academia.edu/67787449

The R Foundation. (2022, 23. Juni). *R: The R Project for Statistical Computing*. https://www.r-project.org/

Thielsch, M. T., & Hirschfeld, G. (2018). Woran erkenne ich einen guten User Experience Fragebogen? In R. Dachselt & G. Weber (Hrsg.), *Mensch & Computer 2018 - Workshopband* (S. 749–755). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2018-ws16-0489

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, *42*(4), 253–264. https://doi.org/10.1080/00207590701396674

Tice, J.A., Helfand, M., & Feldman, M.D. (2010). *Clinical Evidence for Medical Devices: Regulatory Processes Focusing on Europe and the United States of America*. World Health Organization. https://apps.who.int/iris/handle/10665/70454?show=full

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly*, *71*(1), 91–112. https://doi.org/10.1093/poq/nfl046

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

U.S. Food & Drug Administration (2016a). *Applying Human Factors and Usability Engineering to Medical Devices: Guidance for Industry and Food and Drug Administration Staff* (FDA-2011-D-0469). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices

U.S. Food and Drug Administration. (2016b). *Factors to Consider Regarding Benefit-Risk in Medical Device Product Availability, Compliance, and Enforcement Decisions* (FDA-2016-D-1495). https://www.fda.gov/media/98657/download

U.S. Food and Drug Administration. (2022). *Medical Device Recalls.* https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRes/res.cfm

Wetzlinger, W., Auinger, A., & Dörflinger, M. (2014). Comparing Effectiveness, Efficiency, Ease of Use, Usability and User Experience When Using Tablets and Laptops. In: A. Marcus (Hrsg.), *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience* (S. 402–412). Springer. https://doi.org/10.1007/978-3-319-07668-3_39

Wieland, A., & Wallenburg, C. M. (2012). Dealing with supply chain risks: Linking risk management practices and strategies to performance. *International journal of physical distribution & logistics management*, *42*(10), 887–905. https://doi.org/10.1108/09600031211281411

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care*, *8*(3), 1–13. https://doi.org/10.33151/ajp.8.3.93

Winter, D. (2020). Maßnahmen zur Steigerung der organisationalen UX-Kompetenz. In C. Hansen, A. Nürnberger & B. Preim (Hrsg.), *Mensch und Computer 2020 - Workshopband*. Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2020-ws03-002

Winter, D., Hinderks, A., Schrepp, M., & Thomaschewski, J. (2017). Welche UX Faktoren sind für mein Produkt wichtig? In S. Hess & H. Fischer (Hrsg.), *Mensch & Computer 2017 - Usability Professionals* (S. 191–200). Gesellschaft für Informatik e.V. https://doi.org/10.18420/muc2017-up-0002

Winter, D., Schrepp M., & Thomaschewski, J. (2015). Faktoren der User Experience: Systematische Übersicht über produktrelevante UX-Qualitätsaspekte. In A. Endmann, H. Fischer & M. Krökel (Hrsg.), *Mensch & Computer 2015 – Usability*

Professionals (S. 33–42). De Gruyter. https://doi.org/10.1515/9783110443882-005

Witt, C. M., Treszl, A., & Wegscheider, K. (2011). Comparative Effectiveness Research: Externer Validität auf der Spur. *Deutsches Ärzteblatt*, *108*(46). 2468–2474. https://www.aerzteblatt.de/archiv/113438

World Health Organization. (2017). *Human resources for medical devices: The role of biomedical engineers*. https://www.who.int/publications-detail-redirect/9789241565479

Wu, H., & Leung, S.-O. (2017). Can Likert Scales be Treated as Interval Scales? – A Simulation Study. *Journal of Social Service Research*, *43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

XING. (2022). *Zeit, Dich beruflich zu entfalten.* https://www.xing.com

Zhang, J., Patel, V. L., Johnson, T. R., Chung, P., & Turley, J. P. (2005). Evaluating and predicting patient safety for medical devices with integral information technology. In K. Henriksen, J. B. Battles, E. S. Marks & D. I. Lewin (Hrsg.), *Advances in Patient Safety: From Research to Implementation* (2. Aufl., S. 323–336). Agency for Healthcare Research and Quality. https://doi.org/10.1037/e448232006-001

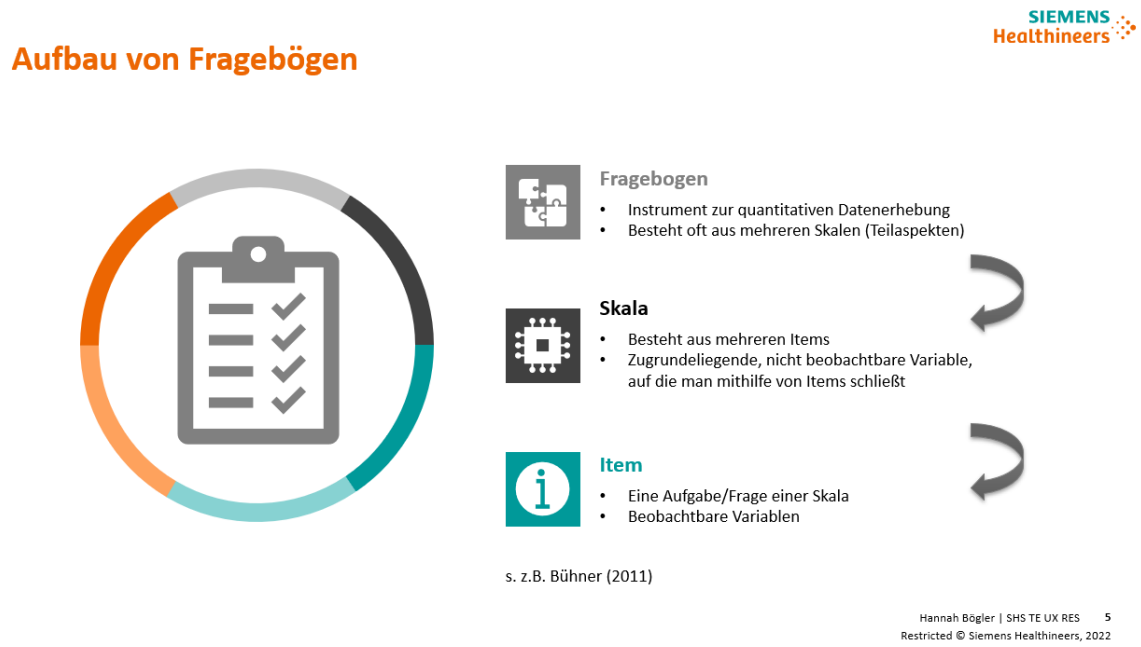
Zhou, L., Bao, J., Setiawan, I. M. A., Saptono, A., & Parmanto, B. (2019). The mHealth App Usability Questionnaire (MAUQ): Development and Validation Study. *JMIR MHealth and UHealth*, *7*(4), 1–15. https://doi.org/10.2196/11500

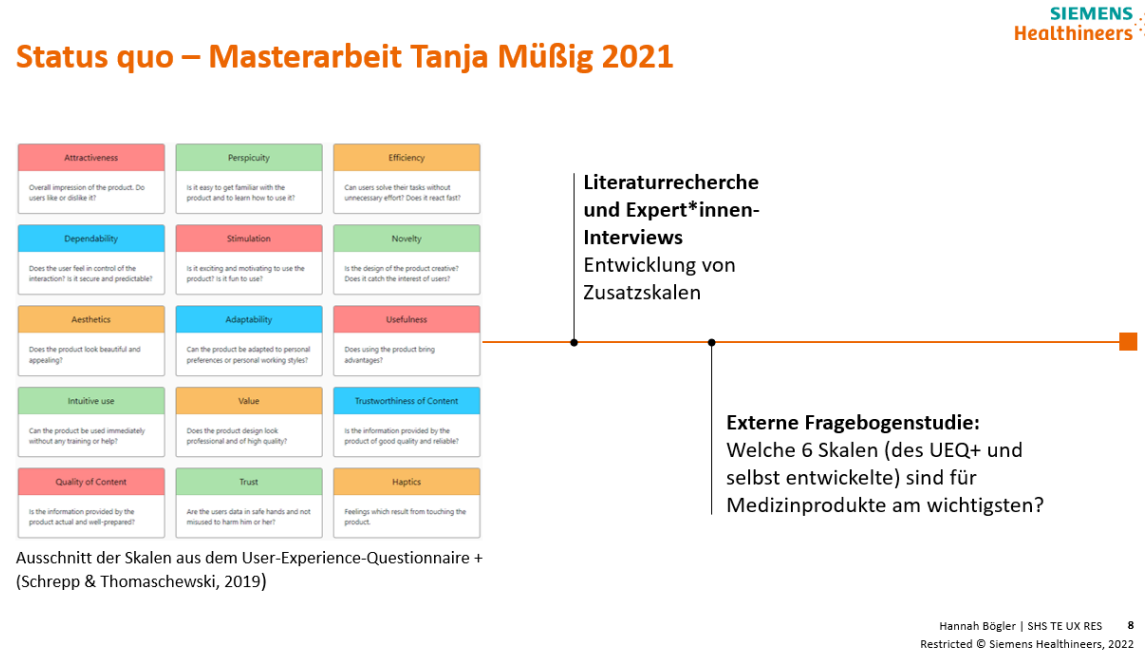
Anhang

Anhang A – Material der Workshops zur Itemgenerierung

Abbildung A1

Ausschnitte der Präsentationen zur Vermittlung von Hintergrundinformationen zu User Experience, dem UEQ+ und dem Aufbau von Fragebögen im Rahmen der Workshops





Ausschnitt der Skalen aus dem User-Experience-Questionnaire + (Schrepp & Thomaschewski, 2019)

Anmerkung. In allen drei durchgeführten Workshops wurden die gleichen Einführungsfolien präsentiert.

Abbildung A2

Ausschnitte des Conceptboards zur Durchführung der Workshops zur Itemgenerierung am Beispiel der Skala Sicherheit

Abbildung A3

Weitere Aspekte aus der Literatur zur Skala Sicherheit, die Teilnehmenden präsentiert wurden

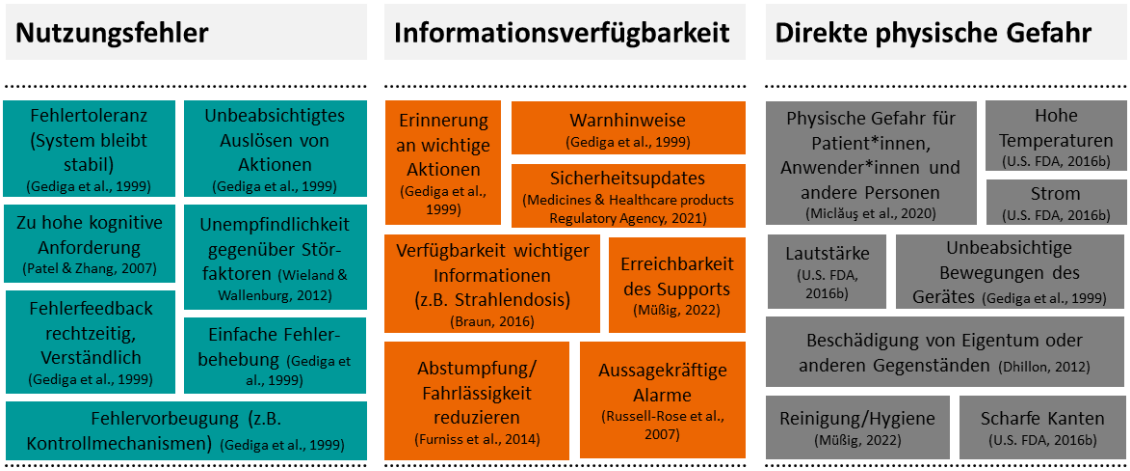


Abbildung A4

Weitere Aspekte aus der Literatur zur Skala Effektivität, die Teilnehmenden präsentiert wurden

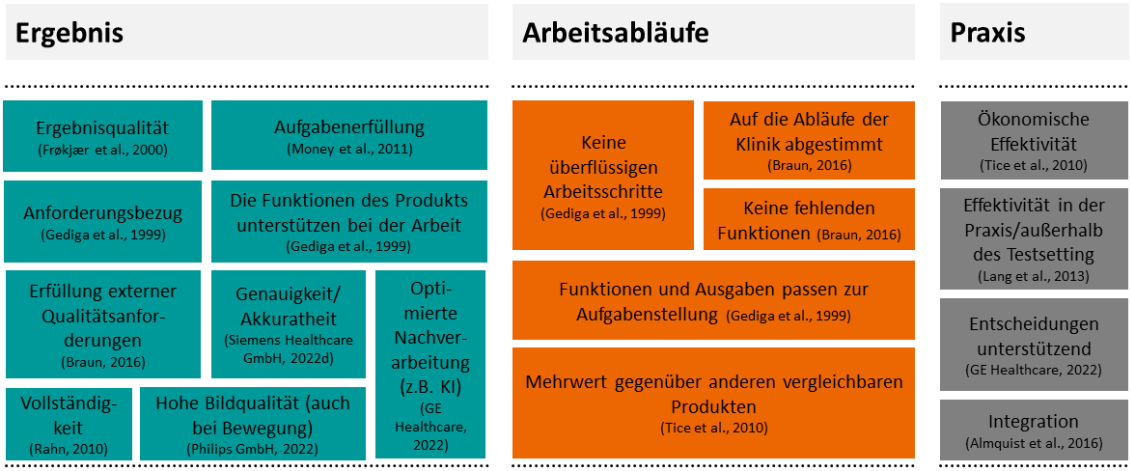


Tabelle A1

Auswertung des ersten Workshops zur Itemgenerierung der Skala Sicherheit

Item	Gesammelte Items im Workshop am 31.05.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	Negativer Pol (unsicher)	Positiver Pol (sicher)	Inhalt	Quelle	Inhalt	Negativer Pol (unsicher)	Positiver Pol (sicher)
1		Es besteht Kollisionsgefahr.	Kollisionen mit dem Gerät sind ausgeschlossen.				*Es besteht Kollisionsgefahr.*	*Kollisionen mit dem Gerät sind ausgeschlossen.*
2	Unbeabsichtigte Bewegungen/ Kontakte *(nur auf Hardware anwendbar)*	Unbeabsichtigter Bewegungen des Geräts sind möglich.	Bewegungen des Systems sind ausreichend langsam und antizipierbar.	Es sollten keine unbeabsichtigten Bewegungen des Geräts ausgelöst werden können, Kollisionsgefahr; Die Software ist so gestaltet, dass das versehentliche Auslösen von Aktionen verhindert wird (z.B. durch Sicherheitsabstände zwischen kritischen Tasten).	U.S. Food and Drug Administration (2022) Gediga et al. (1999)	Die Software ist so gestaltet, dass das versehentliche Auslösen von Aktionen verhindert wird (aus Isometrics)	*Aktionen und Bewegungen des Geräts können versehentlich ausgelöst werden.*	*Das versehentliche Auslösen von Aktionen oder Bewegungen ist nicht möglich.*
3	Hygiene/Reinigung *(nur auf Hardware anwendbar)*	Das Produkt ist schwer zu reinigen/desinfizieren.	Das Produkt ist leicht zu reinigen/desinfizieren.	Reinigung muss problemlos möglich sein	U.S. Food and Drug Administration (2022)		*Das Produkt ist schwer zu reinigen/desinfizieren.*	*Das Produkt ist leicht zu reinigen/desinfizieren.*
4		Das Produkt wirkt anfällig und fragil (z.B. bei häufiger Reinigung)	Das Produkt wirkt mechanisch robust.				*Das Produkt wirkt anfällig und fragil (z.B. bei häufiger Reinigung).*	*Das Produkt wirkt robust und solide verarbeitet.*
5	Cybersecurity/ Datensicherheit (Eher Vertrauen?)	Es sind unzulässige Veränderungen möglich.	Der Zugriff auf die Software ist ausreichend eingeschränkt.	Cybersecurity	Burns et al. (2016)		Es sind unzulässige Veränderungen möglich. (nicht nur Cybersecurity/Vertrauen, wird deshalb beibehalten)	Es können keine unzulässigen Veränderungen vorgenommen werden. (nicht nur Cybersecurity/Vertrauen, wird deshalb beibehalten).
6	Verhinderung von Schäden an anderen Geräten *(nur auf Hardware anwendbar)*			Verhinderung von Schäden an Gegenständen gemäß den Anforderungen der Aufgabe; Risks can also be related to damage to property (for example objects, data, other equipment) or the environment.	Dhillon (2012) ISO 14971:2019 International Organization for Standardization (2019b)		*Das Gerät könnte Schäden an anderen Objekten verursachen.*	*Schäden an anderen Objekten sind ausgeschlossen.*
7	*Gefühl* von Sicherheit	Das Gerät *erscheint/wirkt* unsicher.	Das Gerät *erscheint/wirkt* sicher und vertrauenswürdig.			subjektive Sicherheit vor Bedrohungen	Das System *erscheint/wirkt* unsicher.	Das System *erscheint/wirkt* sicher und vertrauenswürdig.
8	Workflow	Das Gerät ist starr und lässt sich schlecht manövrieren.	Das Gerät ist mobil (kann z.B. an alle Körperbereiche gelangen).		ISO/IEC GUIDE 51:2014 International Organization for Standardization (2014)		Das Gerät lässt sich schlecht manövrieren. (Eher Effizienz)	Das Gerät ist mobil (kann z.B. an alle Körperbereiche gelangen). (Eher Effizienz)
9		Der Ablauf von Diagnostik/Intervention ist fehleranfällig.	Der Ablauf von Diagnostik/Intervention beugt Fehlern vor/ist reibungslos/übersichtlich.				Der Ablauf von Diagnostik/Intervention ist fehleranfällig.	Der Ablauf von Diagnostik/Intervention beugt Fehlern vor/ist reibungslos.

Item	Gesammelte Items im Workshop am 31.05.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	Negativer Pol (unsicher)	Positiver Pol (sicher)	Inhalt	Quelle	Inhalt	Negativer Pol (unsicher)	Positiver Pol (sicher)
10	Einweisung/Manual	Ich fühle mich unsicher in der Bedienung.	Ich fühle mich sicher in der Bedienung.				Ich fühle mich unsicher in der Bedienung.	Ich fühle mich sicher in der Bedienung.
11		Ich fühle mich nicht gut eingewiesen in die Funktionsweise des Geräts.	Ich fühle mich ausreichend gut vorbereitet auf die Nutzung des Geräts.				Ich fühle mich nicht gut eingewiesen in die Funktionsweise des Systems.	Ich fühle mich ausreichend gut vorbereitet auf die Nutzung des Systems.
12		Das Manual ist unverständlich.	Das Manual ist sehr verständlich.				Das Manual ist unverständlich (Manual bei UX-Testungen nicht bekannt).	Das Manual ist sehr verständlich. (Manual bei UX-Testungen nicht bekannt).
13	Fehler-/Problembehebung	Bei einem Fehler muss alles abgebrochen werden.	Einzelne Schritte können rückgängig gemacht werden.	Fehler bei der Eingabe von Daten können leicht rückgängig gemacht werden; Ich empfinde den Korrekturaufwand bei Fehlern als gering.	Gediga et al. (1999)		Bei einem Fehler muss alles abgebrochen werden.	Einzelne Schritte können rückgängig gemacht werden.
14		Ich kann mir wichtige Sicherheitsfunktionen nur schwer merken.	Ich kann mir wichtige Sicherheitsfunktionen leicht merken.				Ich kann mir wichtige Sicherheitsfunktionen nur schwer merken.	Ich kann mir wichtige Sicherheitsfunktionen leicht merken.
15				Risiken können durch Nutzungsfehler entstehen; Sicherheit als Nicht-Auftreten von Anwendungsfehlern; Wahrscheinlichkeit von Fehlern minimieren und die Wahrscheinlichkeit maximieren, dass sie abgefangen werden, wenn sie auftreten.	ISO/IEC GUIDE 51:2014 International Organization for Standardization (2014) Dhillon et al. (2012) Patel & Zhang (2007)	Fehlervermeidung: Die Software ist so gestaltet, dass Anwendungsfehler nur sehr selten vorkommen	Nutzungsfehler sind wahrscheinlich.	Das Risiko für Nutzungsfehler ist minimal.
16				Eine Eingabe von mir hat noch nie zu einem Systemfehler (z.B. Absturz) geführt.	Gediga et al. (1999)	Fehlerrobustheit	Das System stürzt bei Fehlern ab.	Das System reagiert robst auf Fehler.
17	Fehlerverständlichkeit/-feedback und Informationsverfügbarkeit	Ich finde es schwierig, akustische und optische Alarme/Warnungen eindeutig zu interpretieren.	Akustische und optische Alarme und Warnungen sind in ihrer Bedeutung verständlich.	Warnhinweise müssen sofort als solche erkannt werden können.	Pickup et al. (2019)		Akustische und optische Warnungen können nicht eindeutig interpretiert werden.	Akustische und optische Warnungen sind in ihrer Bedeutung verständlich.
18		Das System gibt keine oder zu spät Hinweise auf mögliche Gefahren.	Das System weist mich rechtzeitig auf mögliche Gefahren hin. (z.B. durch Signaltöne, Lichter)	Vor der Ausführung möglicherweise problematischer Aktionen gibt die Software eine Warnung aus; Bei fehlerhaften Eingaben gibt die Software in einigen Fällen zu spät Rückmeldungen.	Gediga et al. (1999)		Das System gibt keine oder zu spät Hinweise auf mögliche Gefahren.	Das System weist mich rechtzeitig auf mögliche Gefahren hin.
19						Wahrnehmen kritischer Informationen: Die Software ist so gestaltet, dass ich keine wichtige Aktion vergessen kann; Die Software zeigt mir für alle Arbeitsschritte die sicherheitsrelevanten Informationen.	Es fehlen sicherheitskritische Informationen.	Es werden alle sicherheitsrelevanten Informationen und Arbeitsschritte angezeigt.

Item	Gesammelte Items im Workshop am 31.05.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	Negativer Pol (unsicher)	Positiver Pol (sicher)	Inhalt	Quelle	Inhalt	Negativer Pol (unsicher)	Positiver Pol (sicher)
20	Fehlerverständlichkeit/-feedback und Informationsverfügbarkeit	Fehlermeldungen sind nichtssagend.	Fehlermeldungen sind aussagekräftig.	In einer Fehlersituation gibt die Software konkrete Hinweise, wie der Fehler behoben werden kann; Die Fehlermeldungen sind gut verständlich und hilfreich.	Gediga et al. (1999) Russell-Rose et al. (2007)	In einer Fehlersituation gibt die Software konkrete Hinweise, wie der Fehler behoben werden kann. (aus Isometrics)	Fehlermeldungen sind nichtssagend.	Fehlermeldungen sind aussagekräftig.
21	Freiheiten in der Bedienung	Wurden Aktionen einmal in Gang gesetzt, lassen sie sich nicht stoppen.	Aktionen können, wenn nötig, unterbrochen werden.				Wurden Aktionen einmal in Gang gesetzt, lassen sie sich nicht stoppen.	Aktionen können, wenn nötig, unterbrochen werden.
22		Ich habe keine Möglichkeit mich in Ausnahmesituationen über Sicherheitsmechanismen hinwegzusetzen.	Sicherheitsmechanismen können in Ausnahmesituationen außer Kraft gesetzt werden.				Ich habe keine Möglichkeit mich in Ausnahmesituationen über Sicherheitsmechanismen hinwegzusetzen.	Sicherheitsmechanismen können in Ausnahmesituationen außer Kraft gesetzt werden.
23		Das System erlegt mir teils hinderliche Sicherheitsmechanismen auf.	Sicherheitsmechanismen behindern meine Arbeit nicht.				Das System erlegt mir teils hinderliche Sicherheits-mechanismen auf. (Eher Effizienz)	Sicherheitsmechanismen behindern meine Arbeit nicht. (Eher Effizienz)
24	Hard-/Softwareausfälle und Verlust von Daten	Ich habe das Gefühl, dass wichtige Daten verloren gehen können (z.B. Bildverluste).	Ich habe den Eindruck, dass wichtige Daten in jedem Fall gesichert sind.	Eingegebene Informationen gehen selbst bei einer Fehlbedienung nicht verloren; Befehle, die Daten unwiderruflich löschen, sind mit einer Sicherheitsabfrage gekoppelt.	Gediga et al. (1999)		Ich habe das Gefühl, dass wichtige Daten verloren gehen können (z.B. Bildverluste).	Ich habe den Eindruck, dass wichtige Daten selbst bei Fehlbedienung gesichert sind.
25		Ich habe den Eindruck, dass Hard- oder Softwareausfälle sich nur schwer beheben lassen.	Ich habe den Eindruck, dass Hard- oder Softwareausfälle schnell behoben werden können.				Ich habe den Eindruck, dass Hard- oder Softwareausfälle sich nur schwer beheben lassen.	Ich habe den Eindruck, dass Hard- oder Softwareausfälle schnell behoben werden können.
26	Gewohnheit/Vertrautheit	Das Produkt basiert auf unüblichen Interaktionsprinzipien.	Die Verwendung des Produkts basiert auf gängigen Interaktionsprinzipien.				Das Produkt basiert auf unüblichen Interaktionsprinzipien. (Eher Effizienz)	Die Verwendung des Produkts basiert auf gängigen Interaktionsprinzipien. (Eher Effizienz)
27		Die Funktionsweise des Produkts erscheint ungewohnt und fremd.	Die Funktionsweise des Produkts erscheint vertraut.				Die Funktionsweise des Produkts erscheint ungewohnt und fremd. (Eher Effizienz)	Die Funktionsweise des Produkts erscheint vertraut. (Eher Effizienz)
28	Physische Gefahr *(nur auf Hardware anwendbar)*	Ich habe das Gefühl, dass physische Gefahren, wie zu hohe Strahlung, gegeben sind.	Ich habe das Gefühl, dass keine physische Gefahren, wie zu hohe Strahlung, gegeben sind.		ISO/IEC GUIDE 51:2014 International Organization for Standardization (2014)		*Ich habe das Gefühl, dass physische Gefahren, wie zu hohe Strahlung, gegeben sind.*	*Ich habe das Gefühl, dass keine physische Gefahren, wie zu hohe Strahlung, gegeben sind.*
29		Es könnte Kontakt zu gefährlichen Bereichen entstehen.	Gefährliche Bereiche des Systems sind geschützt.				*Es könnte Kontakt zu gefährlichen Bereichen entstehen.*	*Gefährliche Bereiche des Systems sind geschützt.*
30				Risks can be related to injury, not only to the patient, but also to the user and other persons. (auch im Workshop allgemeine Meinung)	ISO 14971:2019 International Organization for Standardization (2019b)		*Es könnten Patient*innen und/oder Anwender*innen verletzt werden.*	*Eine Verletzung von Patient*innen und/oder Anwender*innen durch das Gerät ist sehr unwahrscheinlich.*
31	Notschalter/Handlungsmög-lichkeiten im Notfall	In Notfällen habe ich keinen Handlungsspielraum.	Es gibt ausreichend Absicherung im Notfall, wie zum Beispiel Notfalltasten.				In Notfällen habe ich keinen Handlungsspielraum.	Es gibt ausreichend Absicherung im Notfall, wie zum Beispiel Notfalltasten.

Item	Gesammelte Items im Workshop am 31.05.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	**Negativer Pol (unsicher)**	**Positiver Pol (sicher)**	**Inhalt**	**Quelle**	**Inhalt**	**Negativer Pol (unsicher)**	**Positiver Pol (sicher)**
32	Weitere Aspekte	Es bestehen Verwechslungsgefahren.	Alle Kennzeichnungen und Funktionen sind eindeutig.				Es bestehen Verwechslungsgefahren.	Alle Kennzeichnungen und Funktionen sind eindeutig.
33		Es ist keine Barrierefreiheit gegeben.	Das Produkt ist barrierefrei.				*Es ist keine Barrierefreiheit gegeben.*	*Das Produkt ist barrierefrei.*
34	Sicherheitsmechanismen/ -updates			Eingaben, die ich mache, werden auf ihre Richtigkeit hin überprüft, bevor die Daten weiter verarbeitet werden; Befehle, die Daten unwiderruflich löschen, sind mit einer Sicherheitsabfrage gekoppelt; Einbauen von Kontroll- / Sicherheits-mechanismen	Gediga et al. (1999)		Es gibt keine Kontrollmechanismen (um z.B. das versehentlich endgültige Löschen von Daten zu verhindern oder Eingaben auf Plausibilität zu prüfen).	Kontrollmechanismen sorgen für mehr Sicherheit (indem sie z.B. prüfen, ob Eingaben plausibel sind oder einen zusätzlichen Schritt zum Löschen von Daten erfordern).
35						Die Software ist so gestaltet, dass ich keine falschen Entscheidungen treffen kann.	Das System hindert mich daran, falsche Entscheidungen zu treffen.	Falsche Entscheidungen werden unterbunden.
36				Sicherheitsupdates/ Aktualisierung der Hinweise	Medicines & Healthcare products Regulatory Agency (2021)		Es erfolgen keine regelmäßigen Sicherheitsupdates. (In Usability-Tests nicht prüfbar)	Sicherheitsupdates erfolgen regelmäßig und automatisch. (In Usability-Tests nicht prüfbar).
37	Kognitive Anforderungen			Cognitive biases können Nutzerfehler begünstigen.	Ford & Evans (2018) Patel & Zhang (2007)		Die kognitiven Anforderungen sind so hoch, dass Fehler begünstigt werden. (Nicht direkt Sicherheit)	Die kognitiven Anforderungen sind nicht zu hoch. (Nicht direkt Sicherheit)

Anmerkung. Kursiv Geschriebenes lässt sich nur auf Hardware-Komponenten anwenden. Grau Geschriebenes markiert Aspekte, die nicht geeignet sind, weil sie entweder bereits durch eine andere Skala abgedeckt werden oder sich im Rahmen von Usability-/UX-Testungen nicht beurteilen lassen.

Tabelle A2

Auswertung des zweiten Workshops zur Itemgenerierung der Skala Effektivität

Item	Gesammelte Items im Workshop am 02.06.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	Negativer Pol (uneffektiv)	Positiver Pol (effektiv)	Inhalt	Quelle	Inhalt	Negativer Pol (uneffektiv)	Positiver Pol (effektiv)
1	Logische Abfolge der Arbeitsschritte	Die Arbeitsschritte folgen einem unlogischen Ablauf.	Die Arbeitsschritte folgen einem logischen Ablauf.			Die Software zwingt mir eine unpraktische Arbeitsweise auf.	Die Arbeitsschritte folgen einem unlogischen Ablauf. (Eher Effizienz)	Die Arbeitsschritte folgen einem logischen Ablauf. (Eher Effizienz)
2				Mit der Software kann ich zusammenhängende Arbeitsabläufe vollständig bearbeiten.	Gediga et al. (1999)		Eine vollständige Bearbeitung zusammenhängender Arbeitsabläufe ist nicht möglich. (Eher Effizienz)	Ich kann zusammenhängende Arbeitsabläufe vollständig bearbeiten. (Eher Effizienz)
3	Störungsfreiheit/ Zuverlässigkeit (Eher Effizienz?/Sicherheit?)	Das Gerät ist leistungsschwach.	Das Gerät ist leistungsfähig und funktioniert reibungslos.				Das Gerät ist leistungsschwach. (Wird vorerst beibehalten, da Bezug zur Ergebnisqualität/Effektivität)	Das Gerät ist leistungsfähig. (Wird vorerst beibehalten, da Bezug zur Ergebnisqualität/Effektivität)
4	Schnelle Verfügbarkeit der Ergebnisse (Eher Effizienz?)	Ergebnisse stehen nicht zeitnah zur Verfügung.	Ergebnisse stehen unmittelbar zur Verfügung.				Ergebnisse stehen nicht zeitnah zur Verfügung. (Wird vorerst beibehalten, da Bezug zur Ergebnissen)	Ergebnisse stehen unmittelbar zur Verfügung. (Wird vorerst beibehalten, da Bezug zur Ergebnissen)
5	Handhabung	Die Handhabung des Geräts ist schwerfällig und unintuitiv.	Das Gerät ist intuitiv bedienbar und leicht zu handhaben.				Die Handhabung des Geräts ist schwerfällig und unintuitiv. (Eher Effizienz)	Das Gerät ist intuitiv bedienbar und leicht zu handhaben. (Eher Effizienz)
6	Verfügbarkeit relevanter Informationen (Eher Sicherheit?)	Wichtige Informationen fehlen oder sind schwer zugänglich.	Wichtige Informationen sind leicht zugänglich.				Wichtige Informationen fehlen oder sind schwer zugänglich.	Wichtige Informationen sind leicht zugänglich.
7							Es werden irrelevante Informationen angezeigt.	Es werden nur relevante und häufig gebrauchte Informationen angezeigt.
8				Features beinhalten nur relevante und häufig gebrauchte Infos und keine unnötige Redundanz; Die vom Programm erzeugten Ausgaben passen zu meinen Aufgabenstellungen, d.h. sie erhalten keine überflüssigen, zu knappen oder unverständlich formulierten Informationen.	SHS UX Heuristik "Minimalismus" Gediga et al. (1999)		Es werden redundante Informationen angezeigt. (Wird vorerst beibehalten, da Bezug zur Ergebnisqualität/ Effektivität)	Es gibt keine unnötige Redundanz. (Wird vorerst beibehalten, da Bezug zur Ergebnisqualität/Effektivität)
9	Verfügbarkeit und Verständlichkeit von Funktionen; Die Funktionen des Produkts unterstützen bei der Arbeit	Es fehlen notwendige Funktionen.	Das Gerät bietet alle notwendigen Funktionen.	Die Software bietet mir alle Möglichkeiten, die ich für die Bearbeitung meiner Aufgaben benötige.	Gediga et al. (1999)	Die Software bietet mir alle Möglichkeiten, die ich für die Bearbeitung meiner Aufgaben benötige. [Isometrics Item]/ Die in der Software implementierten Funktionen unterstützen mich bei der Durchführung meiner Arbeit; Das System stellt mir diejenigen Funktionen zur Verfügung, um meine Arbeitsziele (in der geforderten Qualität) zu erreichen	Es fehlen notwendige Funktionen.	Das Gerät bietet alle für die Aufgabenbearbeitung notwendigen Funktionen.
10		Manche Funktionen sind überflüssig.	Es gibt keine überflüssigen Funktionen.				Manche Funktionen sind überflüssig.	Es gibt keine überflüssigen Funktionen.

Item	Gesammelte Items im Workshop am 02.06.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	Negativer Pol (uneffektiv)	Positiver Pol (effektiv)	Inhalt	Quelle	Inhalt	Negativer Pol (uneffektiv)	Positiver Pol (effektiv)
11	Verfügbarkeit und Verständlichkeit von Funktionen; Die Funktionen des Produkts unterstützen bei der Arbeit	Manche Funktionen sorgen für Verwirrung.	Dem/der Anwender*in ist klar, welche Funktionen wofür gedacht sind.			Der Aufbau des Systems eindeutig und dessen Funktionsweise ist gut zu verstehen/vorherzusagen.	Manche Funktionen sorgen für Verwirrung.	Es ist klar, welche Funktionen wofür gedacht sind.
12							Das Gerät ermöglicht keine eindeutige Diagnosestellung (bzw. keine erfolgreiche Intervention).	Das Gerät ermöglicht eine eindeutige Diagnosestellung (bzw. eine erfolgreiche Intervention).
13		Das Gerät ermöglicht keine eindeutige Diagnosestellung (bzw. keine erfolgreiche Intervention).	Das Gerät ermöglicht eine eindeutige Diagnosestellung (bzw. eine erfolgreiche Intervention).				Das gewünschte Ergebnis kann nicht erzielt werden.	Das gewünschte Ergebnis kann mit dem Gerät erzielt werden.
14				Das Gerät unterstützt mich beim Erreichen meiner Ziele.	Definition von Müßig (2022)		Das Gerät unterstützt mich nicht beim Erreichen meiner Ziele.	Das Gerät unterstützt das Erreichen meiner Ziele.
15	Eindeutige Diagnose und Therapieempfehlung (bzw. erfolgreiche Intervention) (= Zielerreichung); Entscheidungsunterstützung	Nachtrag von AT/AT-SU: Uneindeutige, artefaktbelastete Bilder.	Eindeutige Interpretation von Bilddaten.				Bilder sind schwer zu interpretieren.	Bilddaten lassen sich eindeutig interpretieren.
16		Die Ergebnisse müssen noch nachbearbeitet werden, um eine Diagnose zu stellen.	Die Ergebnisse können direkt zur Diagnosestellung genutzt werden.				Die Ergebnisse müssen aufwendig nachbearbeitet werden. (noch Intervention einbringen?)	Die Ergebnisse können direkt genutzt werden. (noch Intervention einbringen?)
17				Entscheidungsunterstützung; Methoden und Datenquellen, die zur anstehenden Entscheidung passen; Optimierte Nachverarbeitung (z.B. KI), sodass unabhängig von der Erfahrung des Anwenders/der Anwenderin reproduzierbare Ergebnisse resultieren.	GE Healthcare (2022) Witt et al. (2011) Philips (2022)		Das System bietet keine Entscheidungsunterstützung.	Das System unterstützt die Entscheidungsfindung.
18	Konstante Qualität der Ergebnisse	Die Qualität der Ergebnisse ist teilweise mangelhaft.	Die Ergebnisse entsprechend konstant hoher Qualität.	Indicators of effectiveness include quality of solution; the level and quality/accuracy of completed tasks	Frøkjær et al. (2000) Wetzlinger et al. (2014)	Insgesamt bin ich mit der Qualität der Ergebnisse, die ich mit dieser Software erzielen kann, zufrieden. Mit der Software kann ich die Qualität der Ergebnisse liefern, die das Krankenhaus von mir verlangt.	Die Qualität der Ergebnisse ist mangelhaft.	Die Ergebnisse sind von hoher Qualität.
19							Die Qualität der Ergebnisse ist wechselhaft.	Die Qualität der Ergebnisse ist konstant.
20				Hohe Bildqualität, auch bei Bewegung	Philips (2022)		Die Bildqualität lässt teilweise zu wünschen übrig.	Die Bildqualität ist auch bei Bewegungen hoch.
21	Aufgabenerfüllung	Prozeduren müssen wiederholt werden, um zum gewünschten Ergebnis zu führen.	Prozeduren führen direkt zum gewünschten Ergebnis.				Prozeduren müssen wiederholt werden, um zum gewünschten Ergebnis zu führen.	Prozeduren führen direkt zum gewünschten Ergebnis.
22				Effectiveness is concerned with whether the user is able to successfully accomplish a given task; effective for the intended purpose; Effektivität bedeutet, dass der Benutzer sein Ziel vollständig erreichen kann.	Money et al. (2011) World Health Organization (2017) Rahn (2010)	das System ist nützlich für alle meine Arbeitsziele	Meine Aufgaben lassen sich mit dem System nicht vollständig erfüllen.	Ich kann meine Aufgaben mit dem System vollständig erfüllen.
23	Praxistauglichkeit			effective in the real world	Lang et al. (2013)		Das System ist nicht praxistauglich.	Das System lässt sich problemlos in die Praxis übertragen.

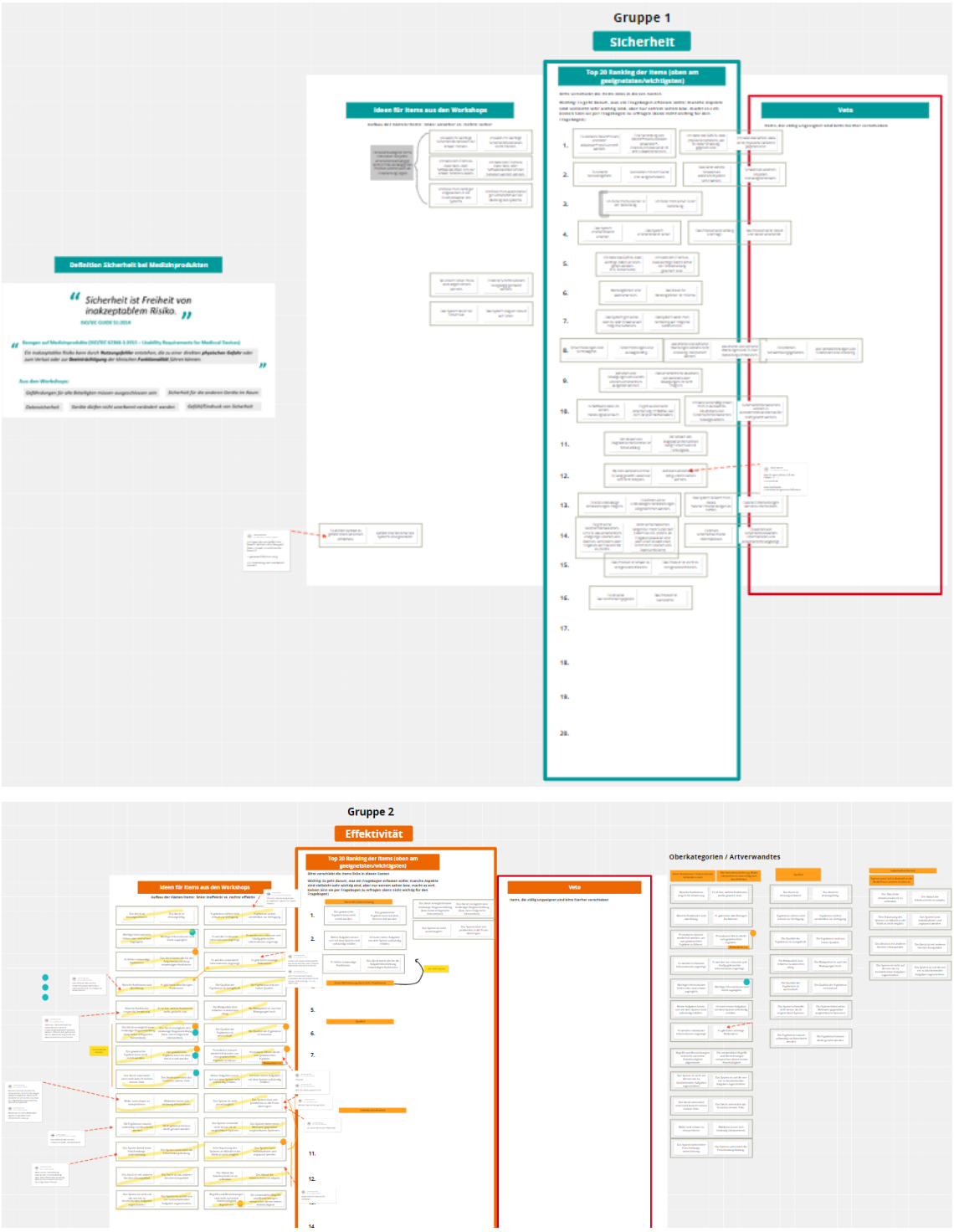
Item	Gesammelte Items im Workshop am 02.06.22			Aus der Literatur		Projekt 2015/16 (Braun)	Abgeleitete Items	
	Aspekt	Negativer Pol (uneffektiv)	Positiver Pol (effektiv)	Inhalt	Quelle	Inhalt	Negativer Pol (uneffektiv)	Positiver Pol (effektiv)
24	Mehrwert gegenüber bisherigen Produkten			incremental cost-effectiveness of a new device compared with the accepted standard ; Comparative Effectiveness Research	Tice et al. (2010) Witt et al., (2011)		Das System schneidet nicht besser ab als vergleichbare Systeme.	Das System bietet einen Mehrwert gegenüber vergleichbaren Systemen.
25	Anpassbarkeit/ Angepasstheit an die Aufgabenstellung/ Anforderungsbezug	Eine Anpassung des Geräts (an Abläufe in der Klinik) ist nicht möglich.	Das Gerät kann individualisiert und angepasst werden.			Die Funktionsweise der Software passt gut zu den Abläufen und Denkweisen in der Klinik.	Eine Anpassung des Systems an Abläufe in der Klinik ist nicht möglich.	Das System kann individualisiert und angepasst werden.
26		Der Ablauf der Arbeitsschritte ist zu unflexibel.	Der Ablauf der Arbeitsschritte ist adaptiv.				Der Ablauf der Arbeitsschritte ist zu unflexibel.	Der Ablauf der Arbeitsschritte ist adaptiv.
27				Die in der Software verwendeten Begriffe und Bezeichnungen entsprechen denen meiner Arbeitstätigkeit.	Gediga et al. (1999)		Begriffe und Bezeichnungen sind nicht auf meine Arbeitstaätigkeit abgestimmt.	Die verwendeten Begriffe und Bezeichnungen entsprechen denen meiner Arbeitstätigkeit.
28				Die Software ermöglicht es mir, Daten so einzugeben, wie es von der Aufgabenstellung gefordert wird; Die Software ist auf die von mir zu bearbeitenden Aufgaben zugeschnitten; Die mit der Software erzeugten Ergebnisse lassen sich meinen Anforderungen entsprechend darstellen bzw. ausgeben.	Gediga et al. (1999)		Das System ist nicht auf die von mir zu bearbeitenden Aufgaben zugeschnitten.	Das System ist auf die von mir zu bearbeitenden Aufgaben zugeschnitten.
29	Gerätekompatibilität	Das Gerät ist mit anderen Geräten inkompatibel.	Das Gerät ist mit anderen Geräten kompatibel.				Das Gerät ist mit anderen Geräten inkompatibel.	Das Gerät ist mit anderen Geräten kompatibel.

Anmerkung. Grau Geschriebenes markiert Aspekte, die nicht geeignet sind, weil sie entweder bereits durch eine andere Skala abgedeckt werden oder sich im Rahmen von Usability-/UX-Testungen nicht beurteilen lassen.

Anhang B – Material des Workshops zur Itemreduktion

Abbildung B1

Ausschnitte des Conceptboards nach der Durchführung des Workshops zur Itemkonsolidierung und -reduktion



Anmerkung. Das Bildschirmfoto oben zeigt einen Ausschnitt des Conceptboards für die Skala Sicherheit, das untere für die Skala Effektivität.

Anhang C – Überführung der Items in das Format des UEQ+

Tabelle C1

Überführung der Skala Sicherheit in das Format des UEQ+

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Anwendungsfehler und Risiken empfinde ich als…	
1	Es könnten Patient*innen und/ oder Anwender*-innen verletzt werden. [a]	Eine Verletzung von Patient*innen und/oder Anwender*innen durch das Gerät ist sehr unwahrscheinlich. [a]	gesundheitsgefährdend	nicht gesundheitsgefährdend
			schwerwiegend	unerheblich
	Ich habe das Gefühl, dass physische Gefahren, wie zu hohe Strahlung, gegeben sind.	Ich habe das Gefühl, dass keine physischen Gefahren, wie zu hohe Strahlung, gegeben sind.	begünstigt	unterbunden
2	Es besteht Kollisionsgefahr.	Kollisionen mit dem Gerät sind ausgeschlossen.	kollisionsbegünstigend	nicht kollisionsbegünstigend
	Das Gerät könnte Schäden an anderen Objekten verursachen.	Schäden an anderen Objekten sind ausgeschlossen.	schädigend	nicht schädigend
3	Das System erscheint/wirkt sicher.	Das System erscheint/wirkt sicher.	bedrohlich	harmlos

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Anwendungsfehler und Risiken empfinde ich als…	
	Das Produkt wirkt anfällig und fragil.	Das Produkt wirkt robust und solide verarbeitet.	durch die Verar-beitung des Pro-dukts begünstigt	durch die Verar-beitung des Pro-dukts unwahr-scheinlich
4	Nutzungsfehler sind wahrschein-lich.	Das Risiko für Nutzungsfehler ist minimal.	wahrscheinlich	unwahrscheinlich
5	Das System gibt keine oder zu spät Hinweise auf mögliche Gefah-ren.	Das System weist mich rechtzeitig auf mögliche Ge-fahren hin.	nicht rechtzeitig rückgemeldet	rechtzeitig rückgemeldet
	Fehlermeldungen sind nichtssagend.	Fehlermeldungen sind aussagekräf-tig.	schwer erkennbar	leicht erkennbar
6	Akustische und optische Warnun-gen können nicht eindeutig interpre-tiert werden.	Akustische und optische Warnun-gen sind in ihrer Bedeutung ver-ständlich.	schwer verständ-lich angezeigt	leicht verständlich angezeigt
	Es bestehen Ver-wechslungsgefah-ren.	Alle Kennzeich-nungen und Funk-tionen sind ein-deutig.	nicht gekenn-zeichnet	gekennzeichnet

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Anwendungsfehler und Risiken empfinde ich als…	
7	Aktionen und Bewegungen des Geräts können versehentlich ausgelöst werden.	Das versehentliche Auslösen von Aktionen oder Bewegungen ist nicht möglich.	durch Systembewegungen möglich	nicht durch Systembewegungen möglich
	In Notfällen habe ich keinen Handlungsspielraum.	Es gibt ausreichend Absicherung im Notfall, wie zum Beispiel Notfalltasten.	unveränderbar	beeinflussbar
8	Ich habe keine Möglichkeit mich in Ausnahmesituationen über Sicherheitsmechanismen hinwegzusetzen.	Sicherheitsmechanismen können in Ausnahmesituationen außer Kraft gesetzt werden.	unveränderbar	beeinflussbar
9	Der Ablauf von Diagnostik/ Intervention ist fehleranfällig.	Der Ablauf von Diagnostik/Intervention beugt Fehlern vor/ist reibungslos.	wahrscheinlich	unwahrscheinlich
10	Wurden Aktionen einmal in Gang gesetzt, lassen sie sich nicht stoppen.	Aktionen können, wenn nötig, unterbrochen werden.	unaufhaltbar	aufhaltbar

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Anwendungsfehler und Risiken empfinde ich als…	
11	Es sind unzulässige Veränderungen möglich.	Es können keine unzulässigen Veränderungen vorgenommen werden.	ungesichert	abgesichert
	Das System hindert mich nicht daran, falsche Entscheidungen zu treffen.	Falsche Entscheidungen werden unterbunden.	nicht vom System kontrolliert	vom System kontrolliert
12	Es gibt keine Kontrollmechanismen, um z.B. das versehentlich endgültige Löschen von Daten zu verhindern oder Eingaben auf Plausibilität zu prüfen.	Kontrollmechanismen sorgen für mehr Sicherheit (indem sie z.B. prüfen, ob Eingaben plausibel sind oder einen zusätzlichen Schritt zum Löschen von Daten erfordern).	nicht vom System kontrolliert	vom System kontrolliert
	Es fehlen sicherheitskritische Informationen.	Es werden alle sicherheitsrelevanten Informationen und Arbeitsschritte angezeigt.	durch fehlende Informationen bedingt	nicht durch fehlende Informationen bedingt

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Anwendungsfehler und Risiken empfinde ich als…	
13	Das Produkt ist schwer zu reinigen/desinfizieren.	Das Produkt ist leicht zu reinigen/desinfizieren.	durch umständliche Reinigung begünstigt	durch einfache Reinigung vermieden
14	Es ist keine Barrierefreiheit gegeben. [b]	Das Produkt ist barrierefrei. [b]		
15	Bei einem Fehler muss alles abgebrochen werden.	Einzelne Schritte können rückgängig gemacht werden.	schwer behebbar	leicht behebbar
	Das System stürzt bei Fehlern ab.	Das System reagiert robust auf Fehler.	wahrscheinlich	unwahrscheinlich
	Es könnte Kontakt zu gefährlichen Bereichen entstehen.	Gefährliche Bereiche des Systems sind geschützt.	ungesichert	abgesichert

Anmerkung. Einige neu entwickelte Items decken mehrere Aspekte der Workshops ab und sind daher in mehr als einer Zeile aufgeführt.

[a] Für dieses Workshop-Item wurden zwei Items im Format des UEQ+ gebildet.

[b] Dieser Aspekt ließ sich durch den Einleitungssatz nicht aufnehmen.

Tabelle C2 Überführung der Skala Effektivität in das Format des UEQ+

Überführung der Skala Effektivität in das Format des UEQ+

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Die Ergebnisse, die sich mit dem Produkt/System erzielen lassen, empfinde ich als…	
1. Kategorie „generelle Zielerreichung“	Das gewünschte Ergebnis kann nicht erzielt werden.	Das gewünschte Ergebnis kann mit dem Gerät erzielt werden.	nicht zielerfüllend	zielerfüllend
	Meine Aufgaben lassen sich mit dem System nicht vollständig erfüllen.	Ich kann meine Aufgaben mit dem System vollständig erfüllen.	unvollständig	vollständig
	Es fehlen notwendige Funktionen. [a]	Das Gerät bietet alle für die Aufgabenbearbeitung notwendigen Funktionen. [a]	unzweckmäßig	zweckmäßig
			nicht bedarfsgerecht	bedarfsgerecht
	Das Gerät ermöglicht keine eindeutige Diagnosestellung (bzw. keine erfolgreiche Intervention).	Das Gerät ermöglicht eine eindeutige Diagnosestellung (bzw. eine erfolgreiche Intervention).	uneindeutig	eindeutig

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Die Ergebnisse, die sich mit dem Produkt/System erzielen lassen, empfinde ich als…	
2. Kategorie „keine Behinderung“	Das System ist nicht praxistauglich.	Das System lässt sich problemlos in die Praxis übertragen.	untauglich	praxistauglich
	Wichtige Informationen fehlen oder sind schwer zugänglich. [b]	Wichtige Informationen sind leicht zugänglich. [b]		
	Begriffe und Bezeichnungen sind nicht auf meine Arbeitstätigkeit abgestimmt.	Die verwendeten Begriffe und Bezeichnungen entsprechen denen meiner Arbeitstätigkeit.	inadäquat	adäquat
	Das System ist nicht auf die von mir zu bearbeitenden Aufgaben zugeschnitten.	Das System ist auf die von mir zu bearbeitenden Aufgaben zugeschnitten.	für meine Anforderungen unpassend	auf meine Anforderungen abgestimmt

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Die Ergebnisse, die sich mit dem Produkt/System erzielen lassen, empfinde ich als…	
3. Kategorie „Qualität“	Das Gerät unterstützt mich nicht beim Erreichen meiner Ziele.	Das Gerät unterstützt das Erreichen meiner Ziele.	nutzlos	nützlich
	Bilder sind schwer zu interpretieren.	Bilder lassen sich eindeutig interpretieren.	ungenau	genau
	Das System bietet keine Entscheidungsunterstützung.	Das System unterstützt die Entscheidungsfindung.	entscheidungshinderlich	entscheidungsunterstützend
	Die Qualität der Ergebnisse ist mangelhaft.	Die Ergebnisse sind von hoher Qualität.	minderwertig	hochwertig
	Die Bildqualität lässt teilweise zu wünschen übrig. [b]	Die Bildqualität ist auch bei Bewegungen hoch. [b]		
	Die Qualität der Ergebnisse ist wechselhaft.	Die Qualität der Ergebnisse ist konstant.	wechselhaft	konstant

Ergebnisse des Workshops zur Itemreduktion			Items im Format des UEQ+	
Rang-folge	Negativer Pol	Positiver Pol	Negativer Pol	Positiver Pol
			Die Ergebnisse, die sich mit dem Produkt/System erzielen lassen, empfinde ich als…	
	Das System schneidet nicht besser ab als vergleichbare Systeme.	Das System bietet einen Mehrwert gegenüber vergleichbaren Systemen.	unterdurch-schnittlich	überdurch-schnittlich

[a] Für dieses Workshop-Item wurden zwei Items im Format des UEQ+ gebildet.

[b] Dieser Aspekt ließ sich durch den Einleitungssatz nicht aufnehmen.

Anhang D – Zusätzliche Ergebnisse der externen Studie

Tabelle D1

Ergebnisse zu Voraussetzungstests

Test	Anzahl V [a]	*n*	Teststatistik
Hypothese 2			
Mardia-Test	25	149	$z = 17.96$***
Hypothese 3			
Allgemeine Sicherheitsitems	13		
Mardia-Test	13	145	$z = 15.18$***
KMO	13	173	0.93
Bartlett-Test	13	173	$\chi^2(78) = 1224.88$***
Alle Sicherheitsitems (inkl. Hardware-Items)	20		
Mardia-Test	20	119	$z = 13.72$***
KMO	20	154	0.91
Bartlett-Test	20	154	$\chi^2(190) = 1381.09$***
Effektivitätsitems	14		
Mardia-Test	14	160	$z = 20.31$***
KMO	14	173	0.97
Bartlett-Test	14	173	$\chi^2(91) = 2639.8$***
Hypothese 4			
Mardia-Test	20	164	$z = 34.57$***
Hypothese 5			
Shapiro-Wilk-Test (Zufriedenheit) [b]	1	173	$W = 0.78$***
Shapiro-Wilk-Test (Weiterempfehlung) [b]	1	173	$W = 0.81$***
Mardia-Test	5	148	$z = 11.76$***

Anmerkung. Informationen dazu, welche Voraussetzungstests durchgeführt und welche Variablen einbezogen wurden, befinden sich im Ergebnisteil (Kapitel 3.5).

[a] Anzahl (abhängiger) Variablen.

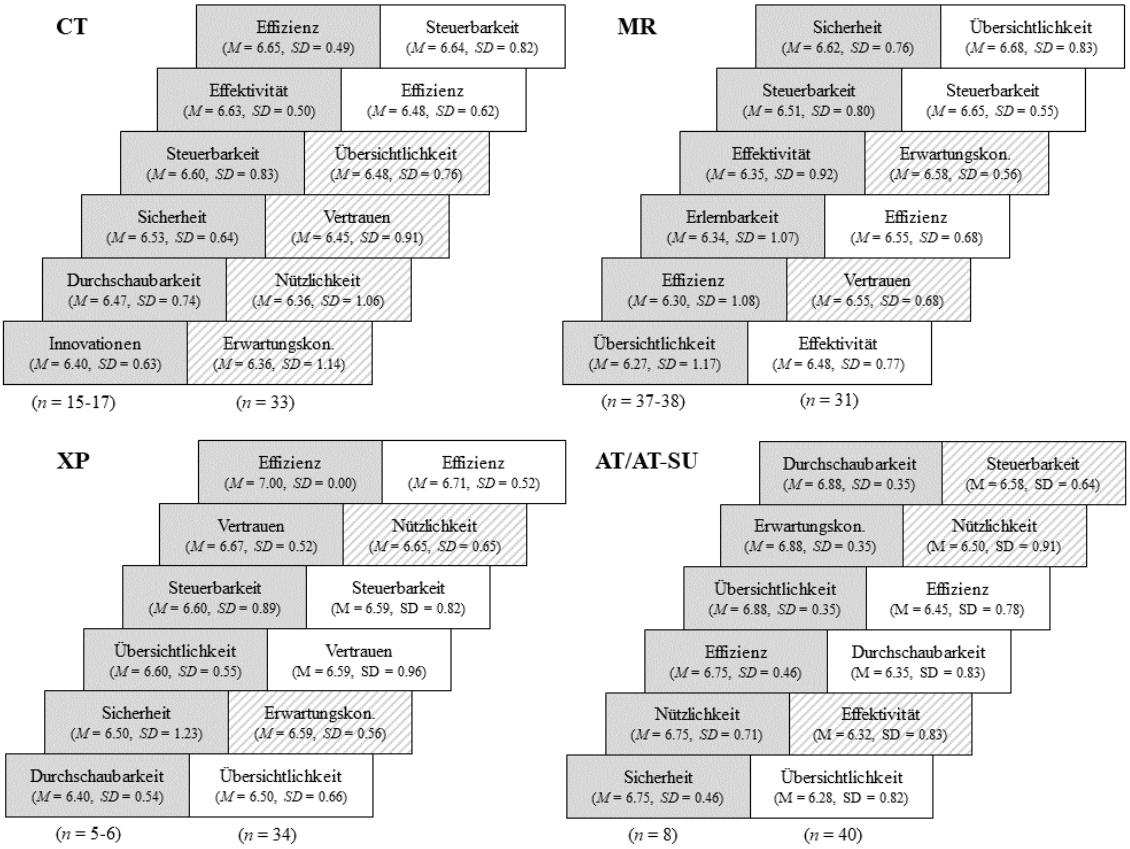
[b] Validierungsitems.

*** $p < .001$.

Abbildung D1

Vergleich der sechs am wichtigsten bewerteten Skalen der Business Lines aus der Vorstudie und der Replikationsstudie

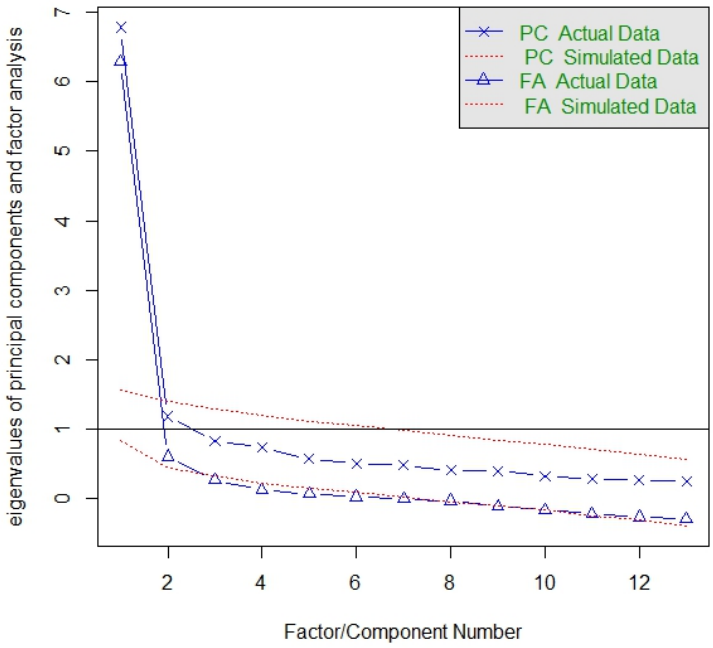
- Wichtigkeitsranking der Skalen aus Müßigs Vorstudie (2021)
- Wichtigkeitsranking der Skalen aus der Replikationsstudie (2022)
- Abweichende Skalen, die bei Müßigs Vorstudie (2021) nicht unter den sechs wichtigsten Skalen für die jeweilige Business Line waren



Anmerkung. Diese Abbildung ist einer Abbildung von Müßig (2022) nachempfunden und wurde entsprechend angepasst. In ihrer Vorstudie wurden Personen, die den Fragebogen abbrachen, nicht ausgeschlossen, wodurch unterschiedliche Stichprobengrößen zustande kamen. Proband*innen der Business Line Strahlentherapie wurden nur für die Replikationsstudie rekrutiert und für die Business Line D&A nahm in Müßigs Studie nur eine Person teil, weshalb für diese Business Lines keine Vergleiche dargestellt werden.

Abbildung D2

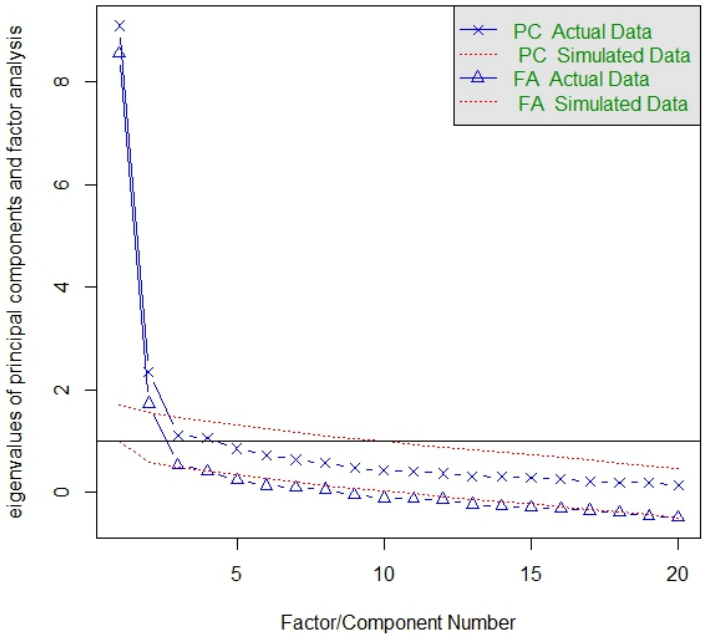
Parallelanalyse der allgemeinen Sicherheitsitems



Anmerkung. *N* = 173. PC = Hauptkomponentenanalyse. FA = Faktorenanalyse (Hauptachsenanalyse).

Abbildung D3

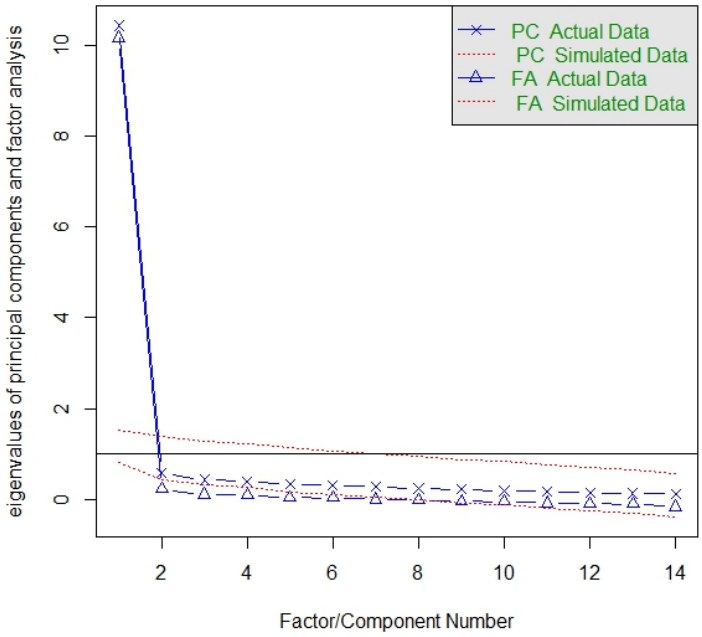
Parallelanalyse aller Sicherheitsitems (inklusive Hardware-Items)



Anmerkung. *N* = 154. PC = Hauptkomponentenanalyse. FA = Faktorenanalyse (Hauptachsenanalyse).

Abbildung D4

Parallelanalyse der Effektivitätsitems



Anmerkung. *N* = 173. PC = Hauptkomponentenanalyse. FA = Faktorenanalyse (Hauptachsenanalyse).

Tabelle D2

Ergebnisse der explorativen Maximum-Likelihood-Faktorenanalyse anhand der allgemeinen Sicherheitsitems bei Extraktion von zwei Faktoren

Allgemeines Sicherheitsitem [a]	Faktorladung	
	1	2
Unwahrscheinlich	**.57***	.15
Rechtzeitig rückgemeldet	**.73***	.10
Leicht verständlich angezeigt	**.79***	.00
Leicht erkennbar	**.74***	.14
Unterbunden	**.37***	**.43***
Leicht behebbar	.22	**.55***
Kontrolliert	**.82***	-.12
Nicht durch fehlende Informationen bedingt	**.46***	.04
Aufhaltbar	.03	**.79***
Gekennzeichnet	**.69***	.03
Abgesichert	**.78***	-.02
Beeinflussbar	-.05	**.86***
Unerheblich	**.69***	-.10

Anmerkung. N = 173. Es wurde eine Maximum-Likelihood-Faktorenanalyse mit Oblimin-Rotation berechnet. Faktorladungen > .30 sind in dicker Schrift hervorgehoben.

[a] Es wird jeweils nur der positive Pol des Items genannt. Die vollständigen Items können in Kapitel 3.2.3 (Items zur Fragebogenentwicklung) nachgesehen werden.

* $p < .002$ (Bonferroni-adjustierter Wert, um die Wahrscheinlichkeit für einen Fehler 1. Art auf einem Niveau von .05 zu halten).

Tabelle D3

Pearsons Produkt-Moment-Korrelationen zwischen den erhobenen produktübergreifend wichtigsten Skalen

Skala	1	2	3	4	5
1. Effektivität	–				
2. Effizienz	.77***	–			
3. Steuerbarkeit	.69***	.75***	–		
4. Übersichtlichkeit	.70***	.78***	.68***	–	
5. Vertrauen	.69***	.61***	.66***	.63***	–

Anmerkung. $N = 173$. $N_{\text{Effektivität}} = 172$.

*** $p < .001$.

Tabelle D4

Rangkorrelationen nach Spearman zwischen den erhobenen produktübergreifend wichtigsten Skalen

Skala	1	2	3	4	5
1. Effektivität	–				
2. Effizienz	.73***	–			
3. Steuerbarkeit	.58***	.69***	–		
4. Übersichtlichkeit	.60***	.76***	.63***	–	
5. Vertrauen	.52***	.57***	.57***	.58***	–

Anmerkung. $N = 173$. $N_{\text{Effektivität}} = 172$. Aufgrund von Rangbindungen konnten p-Werte nicht exakt berechnet werden.

*** $p < .001$.

Tabelle D5

Ergebnisse zu psychometrischen Gütekriterien bezüglich der Sicherheitsskalen

Gütekriterium	Skala	
	Allgemeine Sicherheit	Hardware-Sicherheit
Reliabilität		
McDonalds Omega (McDonald, 1970, 1999)	.85, 95 % KI [.83, .90]	.90, 95 % KI [.87, .93]
Interrater-Reliabilität [a]	.78, 95 % KI [.40, .96], $F(6,42) = 4.54$, $p = .001$	
Standardabweichungen im Vergleich	$SD_{Produkt} = 1.55$ $SD_{Gesamt} = 1.28$	$SD_{Produkt} = 1.24$ $SD_{Gesamt} = 1.51$
Validität		
Korrelation mit der Gesamtzufriedenheit	Spearmans ρ = .50*** Pearsons $r = .46$***	Spearmans ρ = .21* Pearsons $r = .19$*
MANOVA [a b]	$F(3, 128) = 1.51$, $p = .071$, partielles $\eta^2 = .08$, Wilk's $\Lambda = 0.78$	
Multivariater Kruskal-Wallis-Test [a b]	$Q^2(21) = 32.10$, $p = .057$	

Anmerkung. $N_{Sicherheit\_allgemein} = 170$. $N_{Sicherheit\_Hardware} = 152$. Hintergründe zu dieser Tabelle können im Kapitel 3.5.5 (Psychometrische Gütekriterien) nachgelesen werden.

[a] Für diese Berechnungen wurden die produktübergreifend wichtigsten Skalen sowie die beiden Sicherheitsskalen gemeinsam herangezogen.

[b] Prüfung der Fragestellung, ob Produkte verschiedener Hersteller sich in ihrer durchschnittlichen Bewertung auf den Skalen unterscheiden.

* $p < .05$. *** $p < .001$.

Anhang E – Skalen des UX-Fragebogens für Medizinprodukte im Überblick

Steuerbarkeit

Die Reaktion des Produkts auf meine Eingaben und Befehle empfinde ich als								
unberechenbar	○	○	○	○	○	○	○	vorhersagbar
behindernd	○	○	○	○	○	○	○	unterstützend
unsicher	○	○	○	○	○	○	○	sicher
nicht erwartungskonform	○	○	○	○	○	○	○	erwartungskonform

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020).

Vertrauen

In Bezug auf die Verwendung persönlicher Informationen und Daten ist das Produkt [a]								
unsicher	○	○	○	○	○	○	○	sicher
unseriös	○	○	○	○	○	○	○	seriös
unzuverlässig	○	○	○	○	○	○	○	zuverlässig
intransparent	○	○	○	○	○	○	○	transparent

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020).

Effizienz

Für das Erreichen meiner Ziele empfinde ich das Produkt als								
langsam	○	○	○	○	○	○	○	schnell
ineffizient	○	○	○	○	○	○	○	effizient
unpragmatisch	○	○	○	○	○	○	○	pragmatisch
überladen	○	○	○	○	○	○	○	aufgeräumt

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020).

Übersichtlichkeit

Die Benutzeroberfläche des Produkts empfinde ich als								
schlecht gegliedert	○	○	○	○	○	○	○	gut gegliedert
unstrukturiert	○	○	○	○	○	○	○	strukturiert
ungeordnet	○	○	○	○	○	○	○	geordnet
unorganisiert	○	○	○	○	○	○	○	organisiert

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020).

Nützlichkeit

Die Möglichkeit das Produkt zu nutzen empfinde ich als								
nutzlos	○	○	○	○	○	○	○	nützlich
nicht hilfreich	○	○	○	○	○	○	○	hilfreich
nicht vorteilhaft	○	○	○	○	○	○	○	vorteilhaft
nicht lohnend	○	○	○	○	○	○	○	lohnend
Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020).

Effektivität

Die Ergebnisse, die sich mit dem Produkt/System erzielen lassen, empfinde ich als								
uneindeutig	○	○	○	○	○	○	○	eindeutig
nicht bedarfsgerecht	○	○	○	○	○	○	○	bedarfsgerecht
inadäquat	○	○	○	○	○	○	○	adäquat
unterdurchschnittlich	○	○	○	○	○	○	○	überdurchschnittlich
Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Die Skala wurde später auf Grundlage interner Absprachen umbenannt in *Ergebnisqualität*, mit dem Ziel in der Praxis Verwechslungen mit der Skala Effizienz vorzubeugen. Um Verwirrung zu vermeiden, wurde in dieser Arbeit konsistent von Effektivität gesprochen.

Anhang F – Optionale Zusatzskalen

Allgemeine Sicherheit

Anwendungsfehler und Risiken, die bei der Nutzung des Produkts entstehen können, empfinde ich als								
schwer erkennbar	○	○	○	○	○	○	○	leicht erkennbar
nicht rechtzeitig rückgemeldet	○	○	○	○	○	○	○	rechtzeitig rückgemeldet
schwer verständlich angezeigt	○	○	○	○	○	○	○	leicht verständlich angezeigt
unaufhaltbar	○	○	○	○	○	○	○	aufhaltbar
Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Die Anwendung dieser Skala wird optional, zusätzlich zur produktübergreifenden Version des UX-Fragebogens für Medizinprodukte für alle Business Lines empfohlen, insofern sie einen Mehrwert zur Beantwortung der konkreten Fragestellung liefert. Die Skala wurde später auf Grundlage interner Absprachen umbenannt in *Risikohandhabung*. Um Verwirrung zu vermeiden, wurde in dieser Arbeit konsistent von Allgemeiner Sicherheit gesprochen.

Hardware-Sicherheit

Anwendungsfehler und Risiken, die bei der Nutzung des Produkts entstehen können, empfinde ich als								
bedrohlich	○	○	○	○	○	○	○	harmlos
gesundheitsgefährdend	○	○	○	○	○	○	○	nicht gesundheitsgefährdend
schädigend	○	○	○	○	○	○	○	nicht schädigend
kollisionsbegünstigend	○	○	○	○	○	○	○	nicht kollisionsbegünstigend

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Die Anwendung dieser Skala wird optional, zusätzlich zur produktübergreifenden Version des UX-Fragebogens für Medizinprodukte für alle Business Lines bis auf D&A empfohlen, insofern sie einen Mehrwert zur Beantwortung der konkreten Fragestellung liefert. Sie ist nur für Produkte und Prototypen geeignet, die eine Hardware-Komponente beinhalten. Die Skala wurde später auf Grundlage interner Absprachen umbenannt in *Physische Sicherheit*. Um Verwirrung zu vermeiden, wurde in dieser Arbeit konsistent von Allgemeiner Sicherheit gesprochen.

Durchschaubarkeit

Die Bedienung des Produkts empfinde ich als								
unverständlich	○	○	○	○	○	○	○	verständlich
schwer zu lernen	○	○	○	○	○	○	○	leicht zu lernen
kompliziert	○	○	○	○	○	○	○	einfach
verwirrend	○	○	○	○	○	○	○	übersichtlich

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020). Die Anwendung dieser Skala wird optional, zusätzlich zur produktübergreifenden Version des UX-Fragebogens für Medizinprodukte für die Business Lines AT/AT-SU und D&A empfohlen.

Inhaltsseriosität

Die Informationen und Daten, die mir das Produkt bereitstellt sind								
nutzlos	○	○	○	○	○	○	○	nützlich
unglaubwürdig	○	○	○	○	○	○	○	glaubwürdig
unseriös	○	○	○	○	○	○	○	seriös
ungenau	○	○	○	○	○	○	○	genau

Die durch diese Begriffe beschriebene Produkteigenschaft ist für mich								
völlig unwichtig	○	○	○	○	○	○	○	sehr wichtig

Anmerkung. Diese Skala wurde aus dem UEQ+ übernommen (Schrepp & Thomaschewski, 2020). Sie könnte aufgrund der Ergebnisse der Replikation optional, zusätzlich zur produktübergreifenden Version des UX-Fragebogens für Medizinprodukte für die Business Lines D&A und Strahlentherapie eingesetzt werden, allerdings stellt sie nach Schrepp und Thomaschewski (2019c) eine Verallgemeinerung der Skala Vertrauen dar. Ihre Verwendung wird daher nicht empfohlen.

Erwartungskonformität

Für diese Skala existieren keine Items.

Anmerkung. Die Skala könnte aufgrund der Ergebnisse der Replikation optional, zusätzlich zur produktübergreifenden Version des UX-Fragebogens für Medizinprodukte für die Business Lines CT, MR, XP und Strahlentherapie abgefragt werden. Es existieren allerdings noch keine Items für sie. Nachdem eine Überschneidung mit der Skala Steuerbarkeit wegen des Items „nicht erwartungskonform vs. erwartungskonform“ wahrscheinlich ist, wird ihre Entwicklung und Verwendung nicht empfohlen.

Eigenständigkeitserklärung

Ich erkläre hiermit gemäß § 10 Abs. 4 APO, dass ich die vorstehende Masterarbeit selbständig verfasst bzw. erbracht habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt worden sind und die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht wurden. Ferner, dass die digitale Fassung der gedruckten Ausfertigung ausnahmslos in Inhalt und Wortlaut entspricht und dass zur Kenntnis genommen wurde, dass diese digitale Fassung, einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Bamberg, 16.12.2022