

Evaluationsmethodik und Evaluationshandbuch

Deliverable AP 2.3

Projekt	USecureD – Usable Security by Design
Förderinitiative	Einfach intuitiv – Usability für den Mittelstand
Förderkennzeichen	01MU14002
Arbeitspaket	AP 2.3
Fälligkeit	31.10.2016
Autor	Hartmut Schmitt, Matthias Frauer (HK Business Solutions)
Status	final
Klassifikation	öffentlich



HK Business Solutions GmbH

Hartmut Schmitt
Mellinweg 20
66280 Sulzbach
schmitt@hk-bs.de

KMU
(Konsortialführer)



Technische Hochschule Köln

Prof. Dr.-Ing. Luigi Lo Iacono
Betzdorfer Straße 2
50679 Köln
luigi.lo_iacono@th-koeln.de

Hochschule
(Konsortialpartner)



Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Abstract

Im Projekt USecureD werden Musterlösungen und praxistaugliche Werkzeuge entwickelt, die kleine und mittlere Unternehmen (KMU) bei der Entwicklung bzw. bei der Auswahl betrieblicher Anwendungssoftware mit dem Qualitätsmerkmal „Usable Security“ unterstützen.

Ziel von Teilarbeitspaket 2.3 war es, eine leichtgewichtige Evaluationsmethodik zu entwickeln, die Softwarearchitekten und Entwickler bei kleinen und mittleren Softwareherstellern in die Lage versetzt, mit effizienten und praxistauglichen Werkzeugen im eigenen Kontext selbständig Usable-Security-Evaluationen von implementierten Konzepten und Softwareanwendungen durchzuführen. Es wurde ein variables Evaluierungskonzept entworfen, das in allen Entwicklungsstadien von Anwendungssoftware eine Bewertung des Qualitätsmerkmals Usable Security ermöglicht. In Kapitel 4 wird eine Auswahl geeigneter Werkzeuge zur Durchführung empirischer Evaluationen vorgestellt (inkl. weiterführender Informationen).

Schlagworte

Usable Security, IT-Security, Analytische Evaluation, Empirische Evaluation, Heuristik, Cognitive Walkthrough, Severity Rating, Nutzerstudie, Metrik, Fragebogen

Inhaltsverzeichnis

1	Aufgabenstellung und Vorgehensweise	4
2	Verwendung von Prototypen	5
3	Analytische Evaluation	6
3.1	Rahmenbedingungen	6
3.2	Methoden	6
3.2.1	Heuristische Evaluation	6
3.2.2	Cognitive Walkthrough	8
3.2.3	Konsistenz- und Patternprüfung	10
4	Empirische Evaluation	11
4.1	Rahmenbedingungen	11
4.1.1	Versuchsleiter und Probanden	11
4.1.2	Varianten	11
4.1.3	Testaufgaben	12
4.2	Metriken	12
4.3	Fragebögen	14
4.4.1	AttrakDiff	14
4.4.2	INTUI	15
4.4.3	PANAS	16
4.4.4	Bedürfnisskalen	17
4.4	Übersicht: Empirische Evaluation	18
5	Quellenverzeichnis	20

1 Aufgabenstellung und Vorgehensweise

Die Bewertung von Prototypen bzw. Softwaresystemen parallel zum Entwicklungsprozess trägt maßgeblich zur Qualität eines Produktes bei und reduziert durch frühes Erkennen von Fehlern und Schwachstellen die gesamte Entwicklungszeit. Dadurch werden insbesondere verspätete Auslieferungen und Nachbesserungen vermieden, welche zusätzlich Geld und Ressourcen kosten. Zielstellung innerhalb des USecureD-Projekts war es, eine leichtgewichtige Evaluationsmethodik zu entwickeln, die auf alle Entwicklungsstadien einer Anwendungssoftware angewandt werden kann und die aussagekräftige Resultate in Bezug auf das Qualitätsmerkmal Usable Security liefert. „Leichtgewichtig“ bedeutet in diesem Zusammenhang, dass der Methodeneinsatz je nach Evaluationsfall an die Erfordernisse des konkreten Projekts angepasst werden kann, um so einen optimalen Einsatz der personellen und zeitlichen Ressourcen zu ermöglichen.

Hierfür wurden zunächst relevante Qualitätskriterien identifiziert (vgl. Übersicht in Kapitel 4.4): Effizienz, Effektivität, intuitive Nutzung, pragmatische bzw. hedonische Qualität, Attraktivität des Produkts sowie die Erfüllung psychologischer Grundbedürfnisse nach Autonomie, Kompetenz und Sicherheit. Zur Überprüfung dieser Kriterien wurden geeignete Metriken und Messwerkzeuge ausgewählt. Darauf aufbauend wurde recherchiert, welche Methoden bzw. Werkzeuge zur Verfügung stehen, die für den aktuellen Entwicklungsstand eines Projekts erfahrungsgemäß gute Resultate erzielen. Der Fokus lag hierbei auf leicht anwendbaren und validierten Messwerkzeugen und Arbeitshilfen, aus denen insbesondere kleine und mittlere Softwarehersteller, denen bei der Evaluation oft nur eingeschränkte Personalressourcen, Zeitaufwände und Geldmittel zur Verfügung stehen, auswählen können.

Bei der Evaluierung eines Prototyps bzw. einer Software kann unterschieden werden zwischen einer analytischen Evaluation, bei der die Bewertung durch einen oder mehrere interne Mitarbeiter des Herstellers erfolgt, und einer empirischen Evaluation, bei der die Bewertung in der Regel durch mehrere Endnutzer oder externe Probanden erfolgt. Für beide Anwendungsbereiche werden die Rahmenbedingungen beschrieben und es werden die entsprechenden Methoden bzw. Werkzeuge vorgestellt. Die Anwendbarkeit der vorgestellten Werkzeuge unterscheidet sich vorwiegend dadurch, in welcher Entwicklungsstufe eines Produktes diese eingesetzt werden können und welche Zeit die Evaluation in Anspruch nimmt.

2 Verwendung von Prototypen

Die im Folgenden vorgestellten Werkzeuge können dazu verwendet werden, Evaluationen auf Test- bzw. Produktivsystemen oder anhand von Prototypen ([USABILITY BOK 2010], [FRAUNHOFER IESE 2015A]) vorzunehmen. Die Verwendung von Testsystemen oder Prototypen hat den Vorteil, dass hier – im Vergleich zu Produktivsystemen – einfacher Daten manipuliert werden können.

Prototypen dienen dazu, komplexe Anwendungslogik kostengünstig und anschaulich darzustellen. Man unterscheidet dabei zwischen zwei Arten: Low-Fidelity- Prototypen und High-Fidelity-Prototypen. **Low-Fidelity-Prototypen** [USABILITY FIRST 2015A] zeichnen sich dadurch aus, dass Sie die Anwendungslogik abstrahiert darstellen und sehr günstig zu erstellen sind (beispielsweise kann man auf einem Blatt Papier die grafische Oberfläche eines Produkts aufzeichnen, klickbare Bereiche entsprechend kennzeichnen und auf die darauffolgende Änderung der Oberfläche verweisen). Wird ein solcher Prototyp einem Probanden vorgelegt, sollte dieser die angedachte Logik schnell verstehen können. Zu beachten ist, dass bei dieser Art von Prototypen der Fokus des Probanden oft auf der Navigationslogik liegt und der Proband schnell bemerkt, falls wichtige Menüpunkte oder Funktionalitäten fehlen. Entwickelt man Software im Kundenauftrag, kann der Kunde somit frühzeitig erkennen, ob wesentliche Eigenschaften des Projekts fehlen oder falsch verstanden wurden. Diese abstrahierte Art der Darstellung hat allerdings den Nachteil, dass Details zu bestimmten Funktionen vom Probanden nicht weiter beachtet werden und dadurch nicht zwingend in die anschließende Diskussion miteinfließen. Der Einsatz von Low-Fidelity-Prototypen ist daher insbesondere in frühen Projektphasen empfehlenswert. **High-Fidelity-Prototypen** [USABILITY FIRST 2015B] demonstrieren bei etwas fortgeschrittenen Projekten eine möglichst aktuelle Projektphase oder sie simulieren bereits ein vollständiges Produkt. Der Großteil der Funktionen ist dabei „implementiert“ und suggeriert dem Probanden ein Produkt nahe der Fertigstellung. Hierdurch liegt der Fokus des Probanden oft auf den nicht implementierten Bereichen; grundlegende Menüstrukturen und Layouts werden dadurch oft nicht weiter beachtet und nicht hinterfragt. Die anschließende Diskussion wird oft unmittelbar auf Details der Anwendungsfunktionalitäten gelenkt.

Grundsätzlich sollte die Art des für Evaluationen verwendeten Prototyps an den Entwicklungsstand des Projekts angepasst werden. Betrachtet man beispielsweise ein Projekt in einer frühen Phase und entwirft zum Testen einen High-Fidelity-Prototypen, so sind Probanden sich oft über den tatsächlichen Stand des Projekts im Unklaren und vermuten, dass das Projekt fast abgeschlossen ist. Grundlegende Features wie Navigation oder Layout werden diese Probanden daher nicht mehr in Frage stellen, sie konzentrieren sich vielmehr auf spezielle Probleme, die sich daraus ergeben. Umgekehrt bewirkt ein Low-Fidelity-Prototyp in einem weit fortgeschrittenen Projekt den Eindruck, dass Änderungen in der Oberfläche und Logik noch schnell und einfach zu bewerkstelligen sind. Generell sollte also die Wahl des Prototyps dem Projektfortschritt entsprechen.

3 Analytische Evaluation

Die analytische Evaluation ist eine interne Analyse des Prototyps bzw. des Produkts durch einen Experten bzw. einen durch Weiterbildungen angeleiteten Mitarbeiter der Firma. Die analytische Evaluation besteht beispielsweise aus dem Anwenden von Heuristiken, einem Cognitive Walkthrough bzw. einer Pattern- und Konsistenzprüfung. Hierdurch erhält man ohne viel Aufwand eine schnelle Übersicht zur aktuellen Qualität des Prototyps bzw. des Produkts. Des Weiteren kann diese Art der Analyse bei jeder Entwicklungsiteration des Projekts erneut angewendet werden, sei es nach einem kleinen Update, um zu sehen wie bestimmte Mechaniken Auswirkungen auf das Gesamtprodukt haben, oder kurz vor der Auslieferung, um die Konsistenz und Fehlerfreiheit zu verifizieren.

3.1 Rahmenbedingungen

Die analytische Evaluation (vgl. z. B. [HTW CHUR 2013]) wird generell intern durchgeführt und es sind bei der Informationsgewinnung keine tatsächlichen Nutzer direkt involviert. Es wird mindestens ein Mitarbeiter mit Vorkenntnissen in der fachlichen Domäne benötigt. Alle im Folgenden vorgestellten Verfahren können mit relativ wenig Vorbereitung durchgeführt werden. Es ist nicht erforderlich, dass stets der gleiche Mitarbeiter die analytische Evaluation vornimmt, es sollte jedoch sichergestellt werden, dass die Personen nicht direkt in die Entwicklung des zu bewertenden Produkts involviert sind. Es ist von großem Vorteil, zwei bis drei Mitarbeiter für diese Aufgabe einzuteilen, da sich damit erfahrungsgemäß die Mehrzahl der Fehler finden lässt, während ein einzelner Proband manches übersehen kann. Auch eine abschließende Diskussion und Auswertung der Ergebnisse mit anderen Mitarbeitern bringt oft wichtige Erkenntnisse, die vor allem bei frühen Prototypen häufig zu Änderungen der zugrundeliegenden Projektlogik führen kann. Deutlich mehr als drei eingesetzte Mitarbeiter bringen bei der Fehlerfindung allerdings kaum einen weiteren Mehrwert.

Die Projektphase bzw. Reife des Prototyps ist bei dieser Art der Evaluation nicht relevant. Mit geeigneten Werkzeugen kann man bereits in frühen Phasen schnell behebbare Fehler finden, die im weiteren Entwicklungsverlauf schwerer zu beheben und dadurch wesentlich teurer in der Entwicklung sind. Im fortgeschrittenen Projektverlauf kann man jedoch bei einer Evaluation auf Schwierigkeiten stoßen, die erst in diesem späten Stadium sichtbar werden. Analytische Evaluationsmethoden zeichnen sich durch relativ geringen Aufwand und geringe zeitliche Dauer aus. Ist der Prüfer erfahren, kann er mit geeigneten Heuristiken oder einem Cognitive Walkthrough in wenigen Stunden den aktuellen Prototypen bewerten.

3.2 Methoden

3.2.1 Heuristische Evaluation

Ziel der heuristischen Evaluation [NIELSEN 1995b] ist es, generelle Probleme und Sachverhalte aufzudecken, die entweder bei der Entwicklung nicht bedacht wurden oder mit Schwachstellen umgesetzt wurden. Sie dient nicht der quantitativen Erfassung von Problemen innerhalb des zu testenden Produkts. Als Hilfestellung für die heuristische Evaluation kann man entsprechende Checklisten einsetzen, mit denen allgemeine Usability- und spezielle Security-relevante Fragestellungen abgearbeitet werden können. Hierbei müssen die Eigenschaften des Systems für alle Sichten und Zustände überprüft werden. Sind bei dieser Art der Evaluierung mehrere Personen beteiligt, müssen auch Sachverhalte festgehalten werden, die nur bei einzelnen Testern auftraten.

Die bekannten Heuristiken von Nielsen [NIELSEN 1995a], s. Tabelle 1, sind relativ leicht zu erlernen, einfach durchzuführen und bieten einen sehr guten Ansatz, um Benutzeroberflächen auf ihre Usability hin zu untersuchen. Die ITSM-Heuristiken (IT Security Management Tools) [JAFERIAN 2011], s. Tabelle 2, sind von Niensens Heuristiken abgeleitet. Sie gehen sowohl auf die Sicherheitsmechanismen eines Systems ein als auch auf die Besonderheiten, die durch kooperatives Arbeiten in Mehrbenutzersystemen entstehen. Dadurch sind sie sehr gut geeignet für die heuristische Evaluation im Usable-Security-Kontext. Nach Bedarf können die Heuristiken verschiedener Autoren, z. B. Nielsen und Jaferian, in Kombination verwendet werden, je nachdem, welche Anforderungen an das zu entwickelnde System bestehen.

Tabelle 1: Heuristiken von Nielsen [NIELSEN 1995a]

Sichtbarkeit des Systemstatus	Das System hält den Nutzer immer auf dem Laufenden, indem es angemessenes Feedback in einer angemessenen Zeit liefert.
Übereinstimmung zwischen System und realer Welt	Das System spricht die Sprache der Nutzer und vermeidet systemorientierte Terminologien. Informationen erscheinen in einer natürlichen und logischen Reihenfolge.
Benutzerkontrolle und -freiheit	Das System lässt die Nutzer nicht in Situationen geraten, aus denen sie nicht wieder zurückfinden.
Konsistenz und Standards	Gleiche Begriffe und Aktionen haben in unterschiedlichen Kontexten eine einheitliche Bedeutung. Plattform-Konventionen werden eingehalten.
Fehlerprävention	Besser als gute Fehlermeldungen ist ein gutes Design, welches das Eintreten von Fehlern erst gar nicht zulässt. Kritische Aktionen müssen vor der Durchführung vom Nutzer bestätigt werden.
Erkennen statt Erinnern	Objekte, Aktionen und Optionen sind sichtbar. Informationen, die an anderer Stelle wieder gebraucht werden, bleiben sichtbar oder sind einfach abrufbar.
Flexibilität und Effizienz der Nutzung	Um Fortgeschrittenen eine schnellere Bedienung zu erlauben, können häufig auftretende Aktionen vom Nutzer angepasst werden.
Ästhetik und minimalistisches Design	Dialoge enthalten keine irrelevanten oder selten gebrauchten Informationen. Wichtige Informationen werden nicht zwischen unwichtigen versteckt.
Hilfe beim Erkennen, Diagnostizieren und Beheben von Fehlern	Fehlermeldungen sind in einfacher Sprache (ohne Fehlercodes) gehalten, beschreiben das Problem präzise und bieten Lösungsvorschläge an.
Hilfe und Dokumentation	Wenn nötig, werden Hilfe und Dokumentation angeboten. Diese sind leicht durchsuchbar, fokussieren sich auf die Aufgabe der Nutzer und listen konkret die zu unternehmenden Schritte auf.

Tabelle 2: ITSM-Heuristiken [JAFERIAN 2011]

Sichtbarkeit des Aktivitätsstatus	Der Nutzer sieht, in welchen Prozessen er sich momentan wo befindet. Soweit dies notwendig ist, erhält er zudem Informationen, woran seine Kollegen gerade bzw. demnächst arbeiten.
Historie der Aktionen und Änderungen	Der Nutzer kann seine eigene Historie (Logs, Benutzerkommunikation, Aktivitäten, Änderungen u. ä.) und die Historie seiner Kollegen einsehen.
Anpassbare Darstellung von Informationen	Der Nutzer kann die Darstellungsform von Informationen ändern (in dem Maße, wie dies für die Zielgruppe und die Aufgabe angemessen ist), z. B. kann er wechseln zwischen Balken- und Kuchendiagrammen, Tabellen und Auflistungen, Kommandozeileninterface und GUI, Netzwerk-Log und Netzwerk-Graph.
Regeln und Einschränkungen	Im Rahmen der gesetzten Regeln und Einschränkungen hat der Nutzer Freiheiten bei der Ausführung seiner Handlungen, z. B. in Bezug auf die Reihenfolge, in der er bestimmte Aktionen ausführt.
Planung und Aufteilung der Arbeit zwischen Nutzern	Das System unterstützt die Arbeitsplanung, die Arbeitsteilung zwischen mehreren Nutzern sowie die Einbindung neuer Nutzer.
Erfassen, Teilen und Entdecken von Wissen	Nutzer können ihr Wissen speichern und mit anderen teilen, z. B. indem sie Dokumente, Skripte oder Notizen erstellen.
Überprüfung der Kenntnisse	Nutzer können ihr Wissen über sicherheitsrelevante Handlungen überprüfen, z. B. indem sie kritische Aktionen auf einem Testsystem ausprobieren, bevor sie diese auf dem Produktivsystem ausführen.

3.2.2 Cognitive Walkthrough

Beim Cognitive Walkthrough [UXPA 2010] versetzt sich ein Experte in einen Nutzer hinein und versucht, eine vorab definierte Liste von Aufgaben abzuarbeiten, welche mit der zu testenden Anwendung gelöst werden können sollen. Berücksichtigt der Experte die Eigenschaften der angestrebten Zielgruppe, so kann er gezielt spezielle Probleme aufdecken, wie etwa zu kleine Schriftgrößen bei Anwendungen für Senioren. Eine gute Hilfestellung, um sich besser in die Rolle eines potentiellen Nutzers zu versetzen, bieten Nutzerprofile oder sogenannte Personas [UIG 2016]. Personas sind fiktive Personen, die typische Anwender einer Zielgruppe repräsentieren.

Bei einem Cognitive Walkthrough im Usable-Security-Kontext ist zu beachten, dass die Liste der Aufgaben sicherheitskritische Schritte beinhalten sollte. Hierdurch kann der Experte beispielsweise erkennen, wie gut die Anwendung den zukünftigen Nutzern dabei hilft, Probleme im Umgang mit Sicherheitsfunktion zu vermeiden bzw. zu lösen. Beispielhaft kann die Use-Case-Übersicht 2.10.5 (Kosten- und Leistungsrechnung) aus der Use-Case-Sammlung des USecured-Projekts [SCHMITT & GORSKI 2015] eine Liste von Aufgaben liefern:

Finanzbuchhalter:

- Kostenstelle anlegen
- **Aufteilungsregel definieren***
- Kostenstellenbericht erstellen

Controller:

- Budget definieren
- **Budget zuteilen***
- **Verteilungsschlüssel definieren***
- Budget überwachen
- Budgetabweichungen ermitteln

***rote Schrift = sicherheitskritische Use Cases**

Auch beim Cognitive Walkthrough empfiehlt es sich, dass mehrere Mitarbeiter, die nicht in die Entwicklung involviert sind, sich in die Lage der Nutzer hineinversetzen und verschiedene Use Cases durchspielen. Durch das Nachvollziehen der kognitiven Prozesse der Nutzer fallen schnell potenzielle Usabilityprobleme auf, die dokumentiert und ausgewertet werden können. Falls interne Mitarbeiter vorhanden sind, die von ihrem Profil her selbst als Nutzer in Frage kommen, können diese bestenfalls an einem Walkthrough teilnehmen. Die gefundenen Probleme werden am Ende des Cognitive Walkthrough zusammengetragen und zeigen bestehende Schwachstellen auf. Hierbei bietet es sich an, ein Severity Rating ([NIELSEN 1995c], [FRAUNHOFER IESE 2015b]), also eine Priorisierung nach Schweregrad, s. Abbildung 1, vorzunehmen. Je detaillierter der Use Case beschrieben ist (Vor-/Nachbedingungen, Nutzungskontext, Qualitätsmerkmale), desto besser kann der Experte die im folgenden Flussdiagramm dargestellte Vorgehensweise beim Cognitive Walkthrough und Severity Rating nachvollziehen.

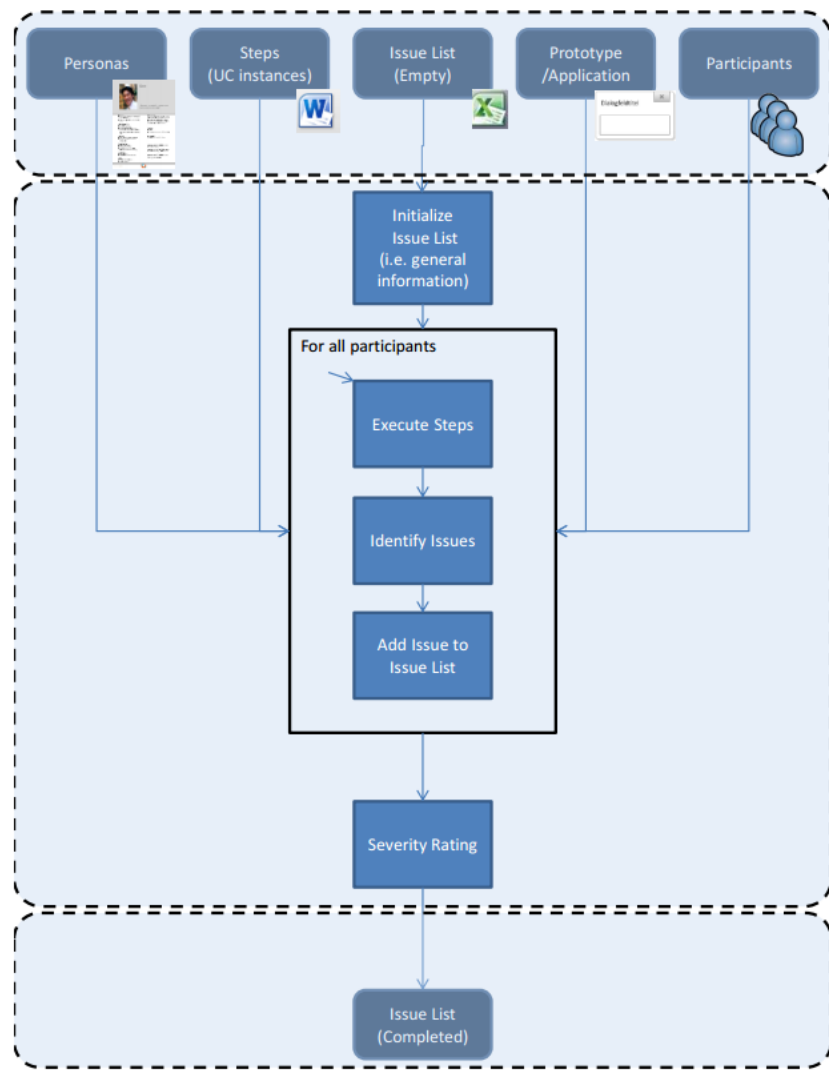


Abbildung 1: Severity Rating [FRAUNHOFER IESE 2015B]

Grundlage der im Diagramm gezeigten Vorgehensweise sind die oben angesprochenen Personas, typische Anwendungsfälle bzw. deren Ausprägungen (Steps bzw. Use Case instances), eine leere Usability-Issue-Liste, ein Untersuchungsgegenstand (Prototype/Application), der die zuvor definierten Anwendungsfälle zumindest prototypisch abdeckt, und die Evaluationsteilnehmer (Participants).

Nach der Initialisierung der Liste (Initialize Issue List), also dem Erfassen allgemeiner Informationen wie Datum/Uhrzeit und Namen der Teilnehmer, führen alle Teilnehmer die einzelnen Anwendungsfälle durch (Execute steps). Alle Usabilityprobleme, die dabei aufgedeckt werden (Identify Issues), werden in die Liste aufgenommen (Add Issue to Issue List). Die Usability-Issue-Liste enthält für jedes Problem eine eindeutige Nummer, jeweils die ID des Use Cases und des Screens, bei dem das Problem aufgetreten ist, eine Problembeschreibung, die Gesamtkritikalität des Problems und optional eine Kategorie sowie einen möglichen Lösungsvorschlag für das gefundene Problem.

Nachdem die Anwendungsfälle von allen Teilnehmern abgearbeitet wurden, bewerten diese die dokumentierten Probleme nach ihrer Kritikalität (Severity Rating). Die Gesamtkritikalität eines Problems setzt sich aus den Faktoren Häufigkeit (frequency), Problemeinfluß (impact), Persistenz (persistence) und Erscheinungsform (manifestation) zusammen. Sie berechnet sich aus dem gerundeten Mittelwert dieser vier Einzelfaktoren. Einzige Ausnahme ist hierbei der Faktor Manifestation: ist dieser Wert hoch, so ist die Gesamtkritikalität ebenfalls hoch. Die Liste ist nun fertiggestellt (Issue List Completed) und je nach verfügbarer Zeit können anschließend für die kritischen Probleme oder für alle gefundenen Probleme Lösungsvorschläge erarbeitet werden.

3.2.3 Konsistenz- und Patternprüfung

Insbesondere bei betrieblicher Anwendungssoftware ist eine konsistente Verwendung von Layoutelementen, Anwendungslogik und Strukturen unentbehrlich für ein gutes Usability-Erlebnis. Mittels einer Checkliste kann ein Experte schnell überprüfen, ob eine Anwendung alle genannten Eigenschaften durchgängig erfüllt. Werden bestimmte Einträge der verwendeten Checkliste in der Anwendung nicht konsistent verwendet, sollte dies analog zur Vorgehensweise beim Cognitive Walkthrough notiert und bewertet werden.

Eine beispielhafte Liste für das Prüfen der Konsistenz von Layoutelementen, Anwendungslogik und Strukturen lautet:

- Haben alle Eingabefelder die gleiche Größe / das gleiche Spacing?
- Befindet sich der Button für die nächste Seite immer an der gleichen Stelle?
- Bekomme ich immer an der gleichen Stelle das Menü eingeblendet?
- Verändert sich das Seitenlayout während der Abarbeitung von Aufgaben?
- Habe ich immer einen Indikator für meine aktuelle Position in der Menüstruktur?
- Bieten gleiche Interface-Elemente immer die gleichen Interaktionsmöglichkeiten?

Im Usable-Security-Kontext hilft insbesondere eine gute Kenntnis der im USecureD-Projekt entwickelten Patterns, Guidelines und Prinzipien dabei, solche Anforderungen zu überprüfen. Die Usable-Security-Patterns [DAS 2016] können checklistenartig auf ihre Umsetzung hin überprüft werden. Sind diese Patterns an (wichtigen) Stellen nicht oder nicht konsistent umgesetzt, bietet dies bereits einen guten Hinweis, wo noch Fehler zu beheben sind. Die Gesamtheit der USecureD-Werkzeuge ist auf der USecureD-Plattform frei zugänglich. Beispiele für Sachverhalte, die checklistenartig überprüft werden können, sind etwa:

- Werden dem Nutzer redundante Hinweise angezeigt?
- Kann der Nutzer nach einem Fehler sicher weiterarbeiten?
- Werden unterscheidbare Sicherheitsebenen verwendet?
- Bekommt der Nutzer nach dem Auftreten von Fehlern sichere Optionen vorgeschlagen?

4 Empirische Evaluation

Empirische Evaluationsmethoden beziehen Benutzer aktiv durch z. B. Befragungen oder Beobachtung mit in den Evaluationsprozess ein, wohingegen analytische Methoden auf Heuristiken, Experten, Modelle oder auf bereits vorhandenen Forschungsergebnissen aufbauen [SARODNICK ET AL. 2011, S. 119]. Der Begriff empirische Evaluation bezeichnet im Usability-Kontext häufig eine Nutzerstudie in kontrollierter Umgebung (z.B. einem Usability-Labor), bei der künftige bzw. reale Benutzer mit einem Produkt konfrontiert werden und während der Interaktion mit diesem Produkt beobachtet werden [SCHWEIBENZ & THISSEN 2003]; die Beobachtung der Nutzer ermöglicht Rückschlüsse über die Stärken und Schwächen des untersuchten Produkts. Generell ist diese Art der Analyse neben lauffähigen Produkten auch bei jeder Prototypstufe (Low-Fidelity oder High-Fidelity) durchführbar.

Um ohne allzu großen Aufwand eine empirische Evaluation im Bereich Usable Security durchzuführen, kann der Versuchsleiter mit mehreren Probanden bestimmte sicherheitsrelevante Testaufgaben (z. B. zuvor definierte Use Cases) durchspielen und deren Erfolg mit einem leicht modifizierten Usability-Fragebogen bewerten lassen. Wichtig ist dabei, dass der Versuchsleiter dem Probanden bei auftretenden Problemen nur kleine Hilfestellungen liefern darf (z.B. bei Missverständnissen in der Aufgabenstellung) und dass das Szenario so realistisch wie möglich ist.

Da die Aufwände für eine empirische Evaluation trotzdem nicht zu unterschätzen sind, lohnt sich eine solche Bewertung eher für spätere Projektphasen, da die Nutzer dann das gesamte Produkt bewerten können. Nichtsdestotrotz macht sich auch eine kleine Studie mit einem Low-Fidelity-Prototypen in einer frühen Projektphase bezahlt, da hierdurch frühzeitig Fehler im Konzept gefunden werden können und zudem weitere Ideen und Anregungen generiert werden können, die direkt in die weitere Entwicklung des Produkts einfließen können.

4.1 Rahmenbedingungen

4.1.1 Versuchsleiter und Probanden

Für die Durchführung der empirischen Evaluation wird mindestens ein Experte als Versuchsleiter benötigt. Für die Auswertung der Ergebnisse kann nach Bedarf ein weiterer Experte herangezogen werden. Ferner wird eine ausreichend große Gruppe zielgruppennaher Personen als Probanden benötigt. Zu dieser Gruppe können auch Mitarbeiter der eigenen Firma herangezogen werden, sofern diese keine Beteiligung am zu bewertenden Produkt haben und zur Zielgruppe des Produktes gehören.

4.1.2 Varianten

Abhängig vom Ort der Durchführung können zwei Varianten der empirischen Evaluation unterschieden werden, die Labor- und die Feldstudie.

Bei einer **Laborstudie** testen die Probanden das Produkt oder den Prototyp innerhalb eines Usability-Labors. Entsprechende Einverständniserklärungen vorausgesetzt, können hierbei mittels Videoaufnahmen, Pulsmessung, EKG usw. zahlreiche Daten erhoben werden. Hierdurch können viele Vermutungen über das Projekt verifiziert werden. Erfahrungsgemäß werden Probanden allerdings oftmals durch die Laborumgebung, die nicht der Umwelt der normalen Nutzung des Produkts entspricht, aus der Illusion einer alltäglichen Nutzung herausgerissen, wodurch es teilweise zu abweichenden Ergebnissen kommen kann.

Bei einer **Feldstudie** testen die Probanden das Produkt oder den Prototyp in der natürlichen Umgebung bzw. unter natürlichen Bedingungen. Dies führt oftmals zu realistischeren Ergebnissen. Bestimmte Messungen sind hierbei jedoch schwieriger zu erbringen als im Laborversuch, ebenso das Beobachten der Nutzer und das Erfassen von Notizen. Bei Feldversuchen ohne einen Versuchsleiter, der Beobachtungen durchführt und Notizen erfasst, ist es zwar oft möglich, mehr Ergebnisse in kürzerer Zeit zu generieren. Im Bereich Usable Security ist es jedoch in vielen Fällen wichtig, die unmittelbare Reaktion der Nutzer auf sicherheitsrelevante Ereignisse im System zu beobachten und zu analysieren, weshalb in solchen Fällen ein Versuchsleiter anwesend sein sollte.

4.1.3 Testaufgaben

In der Studie bearbeiten Probanden, die idealerweise das Profil eines typischen Nutzers haben, verschiedene Testaufgaben an dem aktuellen Prototyp bzw. System. Die Testaufgaben sollten ein breites Spektrum der bereits implementierten Funktionen und möglichst aller Sicherheitsmechanismen abdecken. Es sollten Szenarien konstruiert werden, bei denen der Nutzer auf Gefahren stößt, die abzuwenden sind.

Klassische Usability-Methoden beschreiben sehr oft, wie ein bestimmter Untersuchungsgegenstand bewertet werden kann, aber nicht, wann dieser bewertet werden sollte. Beim Thema Usable Security ist jedoch sehr oft die zeitliche Komponente relevant, beispielsweise wenn der Nutzer in einem bestimmten Moment eine sicherheitsrelevante Entscheidung zu treffen hat, oder wenn er ein Sicherheitsrisiko in einem bestimmten Moment nicht bemerkt, da er nicht darauf aufmerksam gemacht wurde. Um in solchen Fällen eine Bewertung der gebrauchstauglichen Sicherheit einer Anwendungssoftware zu erlangen, muss der Nutzer künstlich an ein solches Sicherheitsrisiko herangeführt werden.

Die Aufgaben bzw. die zugrundeliegenden Szenarios müssen unauffällig konstruiert werden, so dass der Nutzer von einer Warnmeldung oder einer Gefahr überrascht wird und eine möglichst realistische Reaktion hervorgerufen wird. Wie der Nutzer an ein solches Sicherheitsrisiko herangeführt werden kann, ist von Fall zu Fall verschieden. Werden in der Studie beispielsweise E-Mail-Systeme getestet, so kann der Nutzer im Test eine Phishing-Mail erhalten bzw. Malware-Attrappen als Anhang zugeschickt bekommen (vgl. z. B. [EGELMAN ET AL. 2008]). Wird im Unternehmen CRM-Software entwickelt, so kann der Testnutzer Warnmeldungen hinsichtlich Datenduplikaten bekommen und nicht in der Lage sein, eine Aktion durchzuführen, für die er keine entsprechenden Rechte hat (Berechtigungskonzept). Weitere solcher sicherheitsrelevanten Arbeitsschritte sind in der Use-Case-Sammlung des USecureD-Projekts [SCHMITT & GORSKI 2015] identifiziert und beschrieben worden. Eine Konstruktion wichtiger Testszenarios für betriebliche Software kann anhand dieser Sammlung erstellt werden.

4.2 Metriken

Metriken dienen im Bereich der Softwarequalität als Instrument, um verschiedene Eigenschaften einer Software bzw. ihrer Entwicklung oder Anwendung zu quantifizieren [ISO/IEC 2014]. Sie ermöglichen es, eine objektive Einschätzung der Qualität einer Software vorzunehmen oder mehrere Softwareprodukte hinsichtlich ihrer Qualität miteinander zu vergleichen.

Klassische Usability-Metriken zielen vor allem auf die Kriterien Effektivität (Erfolgsrate bzw. Fehlerrate bei den Testaufgaben), Effizienz (aufgewendete Zeit) und Zufriedenheit ab. Für den Bereich Usable Security wurde daher von Kainda, Flechais und Roscoe [KAINDA ET AL. 2010] ein Set von Qualitätsfaktoren (vgl. Abbildung 2) vorgeschlagen, das die Bereiche Usability und Security stärker integriert. Diese Qualitätsfaktoren umfassen zum einen die Usability-Definition des ISO-Standards 9241-11 [DIN 1999] sowie Qualitätsfaktoren zum Thema Sicherheit.

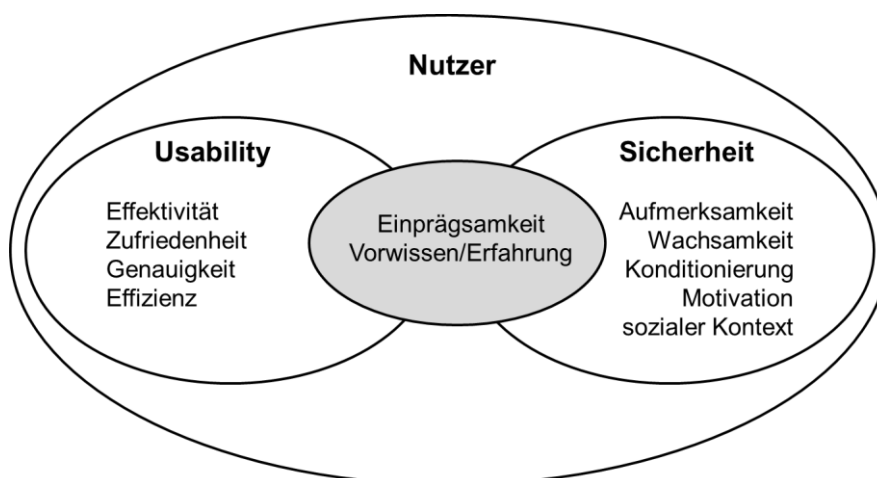


Abbildung 2: Security-usability threat model nach Kainda, Flechais und Roscoe [KAINDA ET AL. 2010]

Zur Überprüfung dieser Qualitätsfaktoren schlagen Kainda et al. ein Set von Metriken vor, mit denen die Reaktionen von Probanden erfasst werden können [vgl. Tabelle 3]. Im Rahmen einer empirischen Evaluation kann der Versuchsleiter anhand dieser Metriken z. B. die Bearbeitung von Testaufgaben durch einen Nutzer bewerten. Dadurch können Beobachtungen erfasst werden, die das Verhalten der Nutzer beschreiben. Notizen zu den entsprechenden Testaufgaben lassen in der Auswertung oftmals Erkenntnisse zu, die sonst verborgen geblieben wären.

Usability		Security	
Faktor	Metrik	Faktor	Metrik
Effektivität	erfolgreiche Bearbeitung	Aufmerksamkeit	Fehler durch fehlende Aufmerksamkeit (allgemein)
Zufriedenheit	Emotionen des Probanden	Wachsamkeit	Fehler durch fehlende Wachsamkeit (Sicherheitswarnungen, Risiken)
Genauigkeit	Fehlversuche	Konditionierung	Fehler durch Konditionierung
Effizienz	Aufgabenbearbeitung nach Zeit, Anzahl der Klicks	Motivation	Nutzer erkennt Vorteile durch Benutzung von Sicherheitsfunktionen, (Stör-)Anfälligkeit, Barrieren, Schwierigkeiten des Nutzers
Einprägsamkeit	Erinnerung des Probanden	Einprägsamkeit	Erinnerung des Probanden
Erlernbarkeit	sinkende Fehlerrate	Vorwissen / Erfahrung	erfolgreiche Bearbeitung, Übereinstimmung zwischen Produkt und mentalem Modell des Nutzers
		Sozialer Kontext	Sozialverhalten

Tabelle 3: Metriken nach Kainda, Flechais und Roscoe [KAINDA ET AL. 2010]

Anhand dieser Metriken kann der Versuchsleiter die „Leistungsdaten“ der Probanden während einer empirischen Evaluation aufzeichnen. Außerdem kann er die Reaktionen der Probanden in einem Ablaufprotokoll bewerten, z. B. kann er erfassen, ob der Nutzer überfordert, verwirrt oder fälschlicherweise selbstsicher war. Hierbei sollte er auch explizit Notizen machen, die auf positive Aspekte des Prototyps hindeuten und den Nutzer motivieren. Um solche Beobachtungen möglichst strukturiert zu erfassen bzw. zu klassifizieren, können sog. **Usability Tokens** [PICCIONI ET AL. 2013] verwendet werden. Es kann dann z. B. unterschieden werden zwischen „surprise“ (Nutzer ist überrascht von einem Element, das nicht erwartungskonform ist), „choice“ (Nutzer wird mit einer Auswahl zwischen mehreren Elementen konfrontiert), „missed“ (Nutzer vermisst ein Element), „incorrect“ (Nutzer wendet ein Element falsch an) und „unexpected“ (Nutzer wendet ein Element in unerwarteter Weise an). Handelt es sich um eine Feldstudie, können die Metriken auch mit einem Fragebogen abgefragt werden. Jedoch sind die resultierenden Ergebnisse mit Vorsicht zu verwerthen, da nicht jeder Nutzer seine Reaktionen richtig einschätzt, erkennt oder als erwähnenswert erachtet.

Neben der Verwendung der von Kainda et al. [KAINDA ET AL. 2010] vorgeschlagenen Metriken bietet es sich an, dass der Versuchsleiter auch **Zeitmessungen** durchführt (die benötigte Zeit gibt Aufschluss über die Effizienz des Probanden bei der Aufgabenlösung) und in strukturierter Weise die Vollständigkeit und Genauigkeit der Aufgabenlösung festhält. Die Effektivität wird laut ISO 9241-210 [DIN 2011] definiert durch die Vollständigkeit und Genauigkeit, mit der die Benutzer bestimmte Ziele erreichen. Angelehnt an diese Definition, kann die Effektivität bei der Durchführung von Aufgaben anhand einer Skala bewertet werden [SCHMITT ET AL. 2014]. Die Gesamteffektivität eines Systems wird nach der Durchführung aller Testaufgaben in Prozent angegeben und kann dadurch leicht mit anderen Produkten verglichen werden. Es kann vorkommen, dass der Versuchsleiter sich während des Tests über die Reaktionen eines Probanden nicht sicher ist. Beispielsweise kann ein Fehler

des Nutzers – er klickt versehentlich ein sicherheitsrelevantes Pop-up weg – sowohl durch den Faktor Wachsamkeit, als auch durch den Faktor Konditionierung entstanden sein. Der Versuchsleiter kann dann nachfragen, sollte dies allerdings erst nach dem Test tun, damit er den Probanden nicht aus der Versuchssituation herausreißt.

4.3 Fragebögen

Es gibt eine Reihe von Fragestellungen, die für das Thema Usable Security relevant sind, die jedoch durch die in Kapitel 4.3 vorgestellten Metriken nicht abgedeckt werden, z. B.: Wie schätzt der Nutzer die Produktqualität ein? Empfindet der Nutzer die Nutzung des Produkts als intuitiv? Wie fühlt sich der Nutzer während der Nutzung des Produkts? Sind bestimmte psychologische Grundbedürfnisse, die beim Nutzungserlebnis eine Rolle spielen, erfüllt?

Mit den Fragebögen, die im Folgenden vorgestellt werden, können im Rahmen einer empirischen Evaluation Informationen zu diesen Fragestellungen bzw. zu den entsprechenden Qualitätsdimensionen erhoben werden. In einem entsprechenden Testsetting sollten die Probanden nach dem Bearbeiten der Testaufgaben aufgefordert werden, die ausgewählten Fragebögen entsprechend ihrer Anweisungen auszufüllen.

Muster sämtlicher Fragebögen, Hinweise zur Auswertung und Interpretation der erhobenen Daten sowie Informationen zum theoretischen Hintergrund und zu weiterführender Literatur stehen im „Handbuch zur Fun-ni Toolbox: User Experience Evaluation auf drei Ebenen“ [DIEFENBACH & HASSENZAHL 2010] zur Verfügung.

4.4.1 AttrakDiff

Der AttrakDiff-Fragebogen [HASSENZAHL ET AL. 2003] ist ein Instrument zur Erfassung des wahrgenommenen Produktcharakters und zur globalen Bewertung interaktiver Produkte. Mit Hilfe von Wortpaaren wird erhoben, wie die Probanden ein Produkt wahrnehmen und bewerten. Diese Wortpaare stellen jeweils extreme Gegensätze dar (z.B. einfach–kompliziert, verwirrend–übersichtlich), zwischen denen eine Abstufung möglich ist (semantisches Differential).

Theoretisches Modell

Der Entwicklung des AttrakDiff-Fragebogens liegt ein theoretisches Modell zugrunde, das die Wahrnehmung, also den wahrgenommenen Produktcharakter, und die Bewertung, die globale Attraktivität des Produkts, voneinander unterscheidet. Der AttrakDiff ist ein etabliertes Evaluationsinstrument im Bereich Mensch-Computer-Interaktion. Er wurde konzipiert zur Evaluation interaktiver Produkte, kann aber auch im Bereich der Marktforschung und Konsumentenpsychologie eingesetzt werden.

Komponenten

Der AttrakDiff unterscheidet zwischen Attraktivität, pragmatischer Qualität und hedonischer Qualität. Bei der hedonischen Qualität, die sich auf die Bedürfnisse des Nutzers bezieht, die der Produktinteraktion zugrunde liegen, unterscheidet er wiederum die zwei Subkomponenten Identität und Stimulation. Insgesamt lassen sich mit dem AttrakDiff also vier Qualitätsdimensionen erfassen:

- **Attraktivität:** Die Attraktivität entspricht einem globalen Produkturteil des Nutzers. Typische Attribute sind Adjektive mit eindeutig positiver Konnotation wie „gut“, „angenehm“ oder „sympathisch“.
- **Pragmatische Qualität:** Die pragmatische Qualität eines Produkts entspricht im Wesentlichen der Gebrauchstauglichkeit im Sinne der DIN EN ISO 9241-11 [DIN 1999]. Sie beschreibt also „das Ausmaß in dem ein Produkt, System oder ein Dienst durch bestimmte Benutzer in einem bestimmten Anwendungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen.“ Typische pragmatische Attribute sind „einfach“, „praktisch“ und „übersichtlich“.
- **Hedonische Qualität – Identität:** Menschen bringen ihr Selbst nicht zuletzt durch die Wahl und Nutzung bestimmter Produkte zum Ausdruck. Die Ausprägung bei dieser Subkomponente gibt einen Hinweis, inwieweit sich die Nutzer mit dem Produkt identifizieren können und inwieweit das Produkt dazu dient, eine gewünschte Identität an andere zu kommunizieren. Typische Attribute sind „stilvoll“ oder „fachmännisch“.
- **Hedonische Qualität – Stimulation:** Menschen streben nach persönlicher Entwicklung und nach der Verbesserung von Kenntnissen und Fähigkeiten. Die Ausprägung bei dieser Subkomponente gibt einen Hinweis, inwieweit das Produkt stimulierend wirkt, z. B. durch neuartige, interessante Funktionalitäten

oder Interaktionsstile, und dieses Bedürfnis unterstützen kann. Typische Attribute sind „neuartig“ oder „innovativ“.

Fragebogen

Die vier oben beschriebenen Dimensionen sind durch jeweils sieben Adjektive repräsentiert. Der Fragebogen hat somit 28 Items. Außerdem existiert eine Kurzform des Fragebogens (AttrakDiff mini), die insgesamt lediglich zehn Items umfasst.

Weiterführende Informationen zum AttrakDiff sind auf der folgenden Website zu finden:

<http://www.attrakdiff.de>

Relevanz für Usable Security

Der AttrakDiff-Fragebogen lässt auf den vom Nutzer wahrgenommenen Charakter eines Produkts bzw. einer Produktkomponente schließen, z. B. einer Sicherheitsfunktion oder eines untersuchten Schutzmechanismus. Der wahrgenommene Produktcharakter ist beispielsweise bei der Bearbeitung von Warnhinweisen und dem Abwenden von Gefahren durch Interaktionen des Nutzers wichtig, da vor allem unerfahrene Nutzer an sicherheitsrelevante Probleme herangeführt werden müssen, um zu verstehen was gemeint ist und warum etwas geschieht.

4.4.2 INTUI

Der INTUI-Fragebogen [ULLRICH & DIEFENBACH 2010] ist ein Instrument zur Erfassung verschiedener Komponenten intuitiver Nutzung. Mit Hilfe gegensätzlicher Aussagenpaare geben die Probanden an, wie Sie die Interaktion mit einem Produkt wahrnehmen. Der Fragebogen basiert auf einem phänomenologischen Ansatz; es wurden verschiedene Aspekte berücksichtigt, die zusammengenommen abbilden, was Nutzer unter dem Begriff „intuitiv“ verstehen.

Theoretisches Modell

Bei der Konstruktion des Fragebogens wurden Nutzermeinungen und Intuitionstheorien aus der Entscheidungspsychologie herangezogen. Die Aussagenpaare lassen sich vier verschiedenen Komponenten zuordnen, deren jeweilige Ausprägung durch den entsprechenden Mittelwert veranschaulicht wird. Damit die Interaktion mit einem Produkt vom Nutzer als intuitiv wahrgenommen wird, muss nicht zwangsläufig auf jeder Komponente ein hoher Wert vorliegen. Bei welchen Komponenten ein hoher Wert angestrebt wird, sollte vielmehr von der Produktdomäne und dem Nutzungsszenario abhängig gemacht werden.

Komponenten

Der INTUI unterscheidet vier Komponenten intuitiver Nutzung; außerdem wird ein globales Urteil bezüglich der wahrgenommenen Intuitivität des Produkts erfasst:

- **Mühelosigkeit:** Ein hoher Wert zeigt an, dass der Nutzer die Interaktion mit dem Produkt als mühelos erlebt und nur ein geringes Ausmaß an Aufmerksamkeit aufbringen muss. Dies ist die Komponente, die die höchste Übereinstimmung mit dem klassischen Usabilitybegriff hat.
- **Bauchgefühl:** Ein hoher Wert zeigt an, dass der Nutzer die Interaktion mit dem Produkt eher als gefühlsgeleitet und weniger als verstandsgleitet erlebt. Dies ist eines der wichtigsten Merkmale intuitiver Entscheidungen in der Entscheidungspsychologie und im alltäglichen Sprachgebrauch.
- **Magisches Erleben:** Ein hoher Wert zeigt an, dass der Nutzer die Interaktion mit dem Produkt als außergewöhnlich und faszinierend wahrnimmt. Magisches Erleben beschreibt die Erlebensqualität und geht ähnlich wie der Begriff der User Experience (UX) über die Aufgabenerfüllung hinaus.
- **Verbalisierungsfähigkeit:** Ein hoher Wert zeigt an, dass der Nutzer den zeitlichen Verlauf der Interaktion gut erinnern und beschreiben kann. Dies spricht für eine logische oder einleuchtende Abfolge von Bedienschritten. Die Herkunft des relevanten Vorwissens, das die Basis für die intuitive Bedienung des Produkts bildet und möglicherweise aus einer ganz anderen Domäne stammt, ist allerdings oft nicht bewusst und damit nicht verbalisierbar.
- **Globales Intuitivitätsurteil:** Zusätzlich zu den vier Komponenten wird mit einem Item („gar nicht intuitiv“ – „sehr intuitiv“) ein globales Maß für die wahrgenommene Intuitivität des Produkts erfasst.

Fragebogen

Die vier Komponenten Mühelosigkeit, Bauchgefühl, magisches Erleben und Verbalisierungsfähigkeit werden mit insgesamt 16 Items abgefragt. Durch ein einzelnes, zusätzliches Item wird ein globales Intuitivitätsurteil erfragt.

Weiterführende Informationen zum INTUI-Fragebogen sind auf der folgenden Website zu finden:
<http://intuitiveinteraction.net>

Relevanz für Usable Security

Der INTUI-Fragebogen gibt Auskunft über die Qualität der intuitiven Benutzung des Prototyps. Neben Aspekten, die primär für die Usability relevant sind (z. B. „Mühelosigkeit“), können mit diesem Fragebogen auch sicherheitsrelevante Aspekte abgedeckt werden. Hierbei ist vor allem die Komponente „Verbalisierungsfähigkeit“ relevant, da z. B. gezeigt werden kann, ob der Nutzer bestimmte Abläufe oder Vorgehensweisen verstanden hat und ob er bei der Bearbeitung von Testaufgaben bewusst oder unbewusst vorhandenes Vorwissen angewendet hat.

4.4.3 PANAS

Der PANAS-Fragebogen (Positive And Negative Affect Schedule) [WATSON ET AL. 1988] ist ein Instrument zur Erfassung der Affektlage. Die Probanden geben mit Hilfe von Adjektiven an, wie sie sich fühlen oder zu einem bestimmten Zeitpunkt, z.B. während der Nutzung eines Produkts, gefühlt haben. Das Ausmaß an positivem und negativem Affekt wird hierbei getrennt erfasst.

Theoretisches Modell

Der PANAS basiert auf einem Modell, nach dem die Affektlage des Nutzers durch zwei voneinander unabhängige Dimensionen abgebildet wird, den positiven Affekt und den negativen Affekt.

Komponenten

Der PANAS unterscheidet zwei Komponenten (positiver und negativer Affekt), die mit den folgenden Adjektiven erfasst werden:

- Positiver Affekt: aktiv, angeregt, aufmerksam, begeistert, entschlossen, freudig erregt, interessiert, stark, stolz, wach.
- Negativer Affekt: ängstlich, bekümmert, beschämt, durcheinander, erschrocken, feindselig, gereizt, nervös, schuldig, verärgert.

Fragebogen

Anhand des PANAS-Fragebogens lassen sich Rückschlüsse auf die Affektlage während der Nutzung eines Produkts ziehen. Da die beiden Dimensionen voneinander unabhängig sind, können auch Affektlagen entstehen, in denen die Ausprägung beider Dimensionen hoch oder niedrig ist. Interessant sind vor allem die Extremwerte bei einzelnen Adjektiven, die Hinweise auf besonders charakteristische Empfindungen und Gefühle während der Interaktion mit dem beurteilten Produkt liefern.

Zur deutschen **Übersetzung** des PANAS vgl. [KROHNE ET AL. 1996].

Relevanz für Usable Security

Dadurch, dass der PANAS-Fragebogen die Affektlage des Nutzers in zwei Dimensionen (positiver und negativer Affekt) erfasst, können deutliche Erkenntnisse über den Prototypen bzw. über das generelle Befinden des Probanden während der Nutzung erlangt werden, die sonst eventuell unbemerkt blieben. Extremwerte bei den entsprechenden Adjektiven können Auskunft darüber geben, ob die Probanden bei der Bedienung von Sicherheitsfunktionen besonders verärgert, nervös, gereizt oder durcheinander sind.

4.4.4 Bedürfnisskalen

Die Bedürfnisskalen [SHELDON ET AL. 2001] sind ein Instrument zur Erfassung des Ausmaßes, in dem psychologische Grundbedürfnisse der Probanden erfüllt sind. Die Probanden geben mit Hilfe des Ausmaßes ihrer Zustimmung zu verschiedenen Aussagen an, zu welchem Grad ein bestimmtes psychologisches Grundbedürfnis erfüllt ist oder zu einem bestimmten Zeitpunkt (z. B. während der Nutzung eines Produkts) erfüllt war.

Theoretisches Modell

Basierend auf verschiedenen psychologischen Bedürfnistheorien zur Beschreibung von (positiven) Erlebnissen identifizierten die Autoren zehn psychologische Grundbedürfnisse: Autonomie, Kompetenz, Verbundenheit, Bedeutsamkeit, Körperlichkeit, Stimulation, Luxus, Sicherheit, Selbstwert und Popularität. Je nach Anwendungskontext bietet es sich an, einzelne Bedürfnisse abzufragen, die im vorliegenden Fall besonders interessieren. Je nach Fragestellung können auch einzelne Items eingesetzt werden.

Komponenten

Relevant für den Bereich Usable Security sind insbesondere die Skalen Autonomie, Kompetenz und Sicherheit:

- Autonomie bezieht sich auf das Bedürfnis, Dinge frei entscheiden zu können. Hierbei spielen Selbstbestimmtheit, Eigenständigkeit und Unabhängigkeit eine Rolle.
- Kompetenz bezieht sich auf das Bedürfnis, sich Herausforderungen zu stellen und sie zu bewältigen. Wichtig ist hierbei das Erleben von Erfolg und Selbstwirksamkeit.
- Sicherheit bezieht sich auf das Bedürfnis, Dinge planen zu können und sicher vor Bedrohung und Ungewissheit zu sein. Es geht also um ein Gefühl der Entspannung durch Planbarkeit und Struktur.

Fragebogen

Das Ausmaß der Erfüllung der zuvor beschriebenen Bedürfnisse wird in den Skalen durch jeweils drei Items abgefragt (auf einer Skala von eins bis fünf). Nach der Bewertung eines bestimmten Erlebnisses, z. B. der Interaktion mit einem Produkt, anhand der Bedürfnisskalen lassen sich durch Berechnung der jeweiligen Mittelwerte für die einzelnen Skalen Rückschlüsse auf die in dieser Situation erfüllten Bedürfnisse ziehen.

Zur **Anwendung der Bedürfnisskalen** im Bereich interaktiver Produkte vgl. [HASSENZAHL ET AL. 2010].

Relevanz für Usable Security

Relevant für den Bereich Usable Security sind insbesondere die drei beschriebenen Skalen „Autonomie“, „Kompetenz“ und „Sicherheit“. Anhand dieser Skalen kann abgefragt werden, zu welchem Grad die entsprechenden Grundbedürfnisse während der Nutzung eines Produkts erfüllt sind. Dies umfasst z. B. die Entscheidungsfreiheit und Selbstbestimmtheit des Nutzers im Umgang mit Sicherheitsfunktionen (sofern hier eine starke Ausprägung gewünscht ist), die Kompetenz des Nutzers zur korrekten Anwendung der Sicherheitsfunktionen und die Sicherheit des Nutzers vor Bedrohung und Ungewissheit, die ihm durch Struktur und angenehme Routinen vermittelt wird.

4.4 Übersicht: Empirische Evaluation

Die folgende Übersicht zeigt, welche Qualitätsmerkmale bzw. -kriterien in einer empirischen Evaluation durch welche der vorgeschlagenen Metriken bzw. durch welches Messverfahren (z. B. Fragebogen, Beobachtung oder Zeitmessung) abgedeckt werden können:

Merkmal	Kriterium	Metrik/Indikator	Messverfahren/ Instrument
Usability	Effektivität – Vollständigkeit	erfolgreiche Bearbeitung	Beobachtung/Erfassung durch Versuchsleiter, dreistufige Effektivitätsskala [SCHMITT ET AL. 2014]
Usability	Effektivität – Genauigkeit	erfolgreiche Bearbeitung	Beobachtung/Erfassung durch Versuchsleiter, vierstufige Effektivitätsskala [SCHMITT ET AL. 2014]
Usability	Effizienz – Erfüllungszeit	benötigte Zeit pro Aufgabe	Zeitmessung durch Versuchsleiter
Usability	Effizienz – Mentale Beanspruchung	Mühelosigkeit	INTUI [ULLRICH & DIEFENBACH 2010]
Usability	Zufriedenheit	Emotionen des Probanden, Metriken für Merkmale User Experience (s. unten)	Beobachtung durch Versuchsleiter, Messverfahren/ Instrumente für Merkmale User Experience (s. unten)
Security	Aufmerksamkeit	Fehler durch fehlende Aufmerksamkeit	Beobachtung, Befragung durch Versuchsleiter
Security	Wachsamkeit	Fehler durch fehlende Wachsamkeit	Beobachtung, Befragung durch Versuchsleiter
Security	Konditionierung	Fehler durch Konditionierung	Beobachtung, Befragung durch Versuchsleiter
Security	Motivation	Vorteile durch Benutzung von Sicherheitsfunktionen, (Stör-)Anfälligkeit, Barrieren, Schwierigkeiten des Nutzers	Beobachtung, Befragung durch Versuchsleiter
Security	Einprägsamkeit	Erinnerung des Probanden	Beobachtung, Befragung durch Versuchsleiter

Security	Vorwissen/Erfahrung	erfolgreiche Bearbeitung, Übereinstimmung zwischen Produkt und mentalem Modell des Nutzers	Beobachtung, Befragung durch Versuchsleiter
User Experience	Pragmatische Qualität	4 Fragebogen-Items	AttrakDiff mini [HASSENZAHL ET AL. 2003]
User Experience	Hedonische Qualität – Identität	2 Fragebogen-Items	AttrakDiff mini
User Experience	Hedonische Qualität – Stimulation	2 Fragebogen-Items	AttrakDiff mini
User Experience	Attraktivität	2 Fragebogen-Items	AttrakDiff mini
User Experience	Positiver/negativer Affekt	10 Fragebogen-Items (5 positiver Affekt, 5 negativer Affekt)	PANAS (Kurzversion) [WATSON ET AL. 1988]
User Experience	Erfüllung der psychologischen Grundbedürfnisse nach „Autonomie“, „Kompetenz“ und „Sicherheit“	9 Fragebogen-Items (je 3 pro Bedürfnis)	Bedürfnisskalen [SHELDON ET AL. 2001]
Intuitive Nutzung	Mühelosigkeit	5 Fragebogen-Items	INTUI
Intuitive Nutzung	Bauchgefühl	4 Fragebogen-Items	INTUI
Intuitive Nutzung	Magisches Erleben	4 Fragebogen-Items	INTUI
Intuitive Nutzung	Verbalisierungsfähigkeit	3 Fragebogen-Items	INTUI
Intuitive Nutzung	Globales Intuitivitätsurteil	1 Fragebogen-Item	INTUI

5 Quellenverzeichnis

- [DAS 2016] DAS-Group TH Köln (2016): Patterns. Verfügbar unter: <https://das.th-koeln.de/usecured/patterns> [15.11.2016]
- [DIEFENBACH & HASSENZAHL 2010] Sarah Diefenbach, Marc Hassenzahl (2010): Handbuch zur Fun-ni Toolbox: User Experience Evaluation auf drei Ebenen. Verfügbar unter: http://fun-ni.org/wp-content/uploads/Diefenbach+Hassenzahl_2010_HandbuchFun-niToolbox.pdf [09.01.2017]
- [DIN 1999] DIN Deutsches Institut für Normung e. V. (1999): Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten - Teil 11: Anforderungen an die Gebrauchstauglichkeit; Leitsätze (ISO 9241-11:1998); Deutsche Fassung EN ISO 9241-11:1998
- [DIN 2011] Ergonomie der Mensch-System-Interaktion - Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme (ISO 9241-210:2010); Deutsche Fassung EN ISO 9241-210:2010
- [EGELMAN ET AL. 2008] Serge Egelman, Lorrie Faith Cranor, Jason Hong (2008): You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (S. 1065-1074). New York: ACM
- [FRAUNHOFER IESE 2015A] Fraunhofer IESE (2015): Experimentelles und exploratives Prototyping durchführen. Verfügbar unter: <http://www.pq4agile.de/PQ4WP/wp-content/uploads/2015/02/PQ4Agile-AP-2.2-Experimentelles-und-exploratives-Prototyping-durchf%C3%BChren-V.11.pdf> [26.01.2017]
- [FRAUNHOFER IESE 2015B] Fraunhofer IESE (2015): Severity Rating durchführen. Verfügbar unter: <http://www.pq4agile.de/PQ4WP/wp-content/uploads/2015/02/PQ4Agile-AP-2.2-Severity-Rating-durchf%C3%BChren-V.11.pdf> [15.11.2016]
- [HASSENZAHL ET AL. 2003] Marc Hassenzahl, Michael Burmester, Franz Koller (2003): AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: J. Ziegler & G. Szwillus (Hrsg.), Mensch & Computer 2003: Interaktion in Bewegung (S. 187-196). Stuttgart, Leipzig: B. G. Teubner.
- [HASSENZAHL ET AL. 2010] Marc Hassenzahl, Sarah Diefenbach, Anja Göritz (2010): Needs, affect, and interactive products – Facets of user experience. In E. Lai-Chong Law & P. van Schaik (Hrsg.): Interacting with Computers 22(5), S. 353–362. Elsevier
- [HTW CHUR 2013] Hochschule für Technik und Wirtschaft Chur (2013): Usabilitymethoden. Verfügbar unter: <http://www.cheval-lab.ch/was-ist-usability/usabilitymethoden/#c299> [26.01.2017]
- [ISO/IEC 2014] International Organization for Standardization / International Electrotechnical Commission (2014): System und Software-Engineering - Qualitätskriterien und Bewertung von System- und Softwareprodukten (SQuaRE) - Leitfaden für SQuaRE
- [JAFERIAN 2011] Pooya Jaferian (2011): Heuristics for Evaluating IT Security Management Tools. In: Symposium on Usable Privacy and Security (SOUPS) 2011, July 20-22. Pittsburgh
- [KAINDA ET AL. 2010] Ronald Kainda, Ivan Flechais, A.W. Roscoe (2010): Security and Usability: Analysis and Evaluation. Oxford University Computing Laboratory, Oxford
- [KROHNE ET AL. 1996] Heinz Walter Krohne, Boris Egloff, Carl-Walter Kohlmann, Anja Tausch (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). Diagnostica 42 (2), S. 139-156. Göttingen: Hogrefe
- [NIELSEN 1995a] Jakob Nielsen (1995): 10 Usability Heuristics for User Interface Design. Verfügbar unter: <https://www.nngroup.com/articles/ten-usability-heuristics/> [15.11.2016]
- [NIELSEN 1995b] Jakob Nielsen (1995): How to Conduct a Heuristic Evaluation. Verfügbar unter: <https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/> [15.11.2016]
- [NIELSEN 1995c] Jakob Nielsen (1995): Severity Ratings for Usability Problems. Verfügbar unter: <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/> [15.11.2016]
- [PICCIONI ET AL. 2013] Marco Piccioni, Carlo A. Furia, Bertrand Meyer (2013): An Empirical Study of API Usability. In: 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. IEEE
- [SARODNICK ET AL. 2011] Sarodnick, F. und H. Brau (2011). Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung. 2., überarb. und aktualisierte Aufl. Wirtschaftspsychologie in Anwendung. Bern: Huber. 278 S. isbn: 978-3-456-84883-9 978-3-456-94883-6.
- [SCHMITT ET AL. 2014] Hartmut Schmitt, Anne Hess, Steffen Hess, Andreas Maier, Diana Löffler, Jörn Hurtienne (2014): Intuitive Benutzbarkeit messen: Eine Evaluationstoolbox für Software, Apps und technische Produkte. Verfügbar unter: <http://germanupa.de/events/mensch-und-computer-2014/shortpaper/intuitive-benutzbarkeit-messen.html> [15.11.2016]
- [SCHMITT & GORSKI 2015] Hartmut Schmitt & Peter Gorski (2015): Use-Case-Sammlung: Deliverable E1.2. Verfügbar unter: <https://www.usecured.de/UseWP/wp-content/uploads/2015/11/E-1.2-Usecured-Use-Case-Sammlung-V.1.pdf> [15.11.2016]
- [SCHWEIBENZ & THISSEN 2003] Werner Schweibenz, Frank Thissen (2002): Qualität im Web: Benutzerfreundliche Webseiten durch Usability Evaluation. Berlin: Springer

[SHELDON ET AL. 2001] Kennon M. Sheldon, Andrew J. Elliot, Youngmee Kim, Tim Kasser (2001): What is satisfying about satisfying events? Testing 10 candidate psychological needs. In: Journal of Personality and Social Psychology 80, S. 325–339.

[UIG 2016] Usability in Germany (UIG) e.V. (2016): Personas. Verfügbar unter: <http://www.usability-in-germany.de/definition/personas> [15.11.2016]

[ULLRICH & DIEFENBACH 2010] Daniel Ullrich, Sarah Diefenbach (2010): INTUI. Exploring the Facets of Intuitive Interaction. In: J. Ziegler & A. Schmidt (Eds.) Mensch & Computer 2010 (S. 251-260). München: Oldenbourg

[USABILITY BoK 2010] Usability Body of Knowledge (2010): Prototyping Methods. Verfügbar unter: <http://usabilitybok.org/prototyping-methods> [26.01.2017]

[USABILITY FIRST 2015A] Usability First (2015): Usability Glossary - low-fidelity prototype. Verfügbar unter: <http://www.usabilityfirst.com/glossary/low-fidelity-prototype/> [26.01.2017]

[USABILITY FIRST 2015B] Usability First (2015): Usability Glossary - high-fidelity prototype. Verfügbar unter: <http://www.usabilityfirst.com/glossary/high-fidelity-prototype/> [26.01.2017]

[UXPA 2010] The User Experience Professionals' Association (2010): Cognitive Walkthrough. Verfügbar unter: <http://www.usabilitybok.org/cognitive-walkthrough> [15.11.2016]

[WATSON ET AL. 1988] David Watson, Lee Anna Clark, Auke Tellegen (1988): Development and validation of brief measures of Positive and Negative Affect: The PANAS scales. In: Journal of Personality and Social Psychology (S.1063 - 1070)