Evaluation von Software über TLX und SUS hinaus Einsatz von Fragebögen

Vincenz Arlt
Proseminararbeit
Proseminar Service and Cloud Computing
TU Dresden

***Zusammenfassung*—In dieser Arbeit werden verschiedene Fragebögen zur Evaluation verglichen. Nach einer Einführung in die einzelnen Fragebögen (System Usability Scale, Task Load Index, User Experience Questionnaire und Unified Theory of Acceptance and Use of Technology) wird auf ihre Besonderheiten und Eigenschaften eingegangen. Dabei werden auf ihre Ziele, erfasste Werte, Umfang, Einsatzzweck und Auswertung hingewiesen. Jeder der Fragebögen hat spezifische Vor- und Nachteile, auf die eingegangen wird, um ihre Verwendungsmöglichkeiten zu beschreiben. Abschließend werden Empfehlungen gegeben, wie sich jeder der Fragebögen in der Entwicklung einer Software nutzbringend einsetzen lässt.**

I. Einführung

Fragebögen stellen eine häufig eingesetzte Methode zur Evaluation dar. Probanden bearbeiten mithilfe der zu testenden Software eine vorher definierte Aufgabe und füllen im Anschluss einen Fragenkatalog aus. Diese Vorgehensweise bietet einige klare Vorteile:

- Die Fragen können schnell beantwortet werden, nachdem der Proband seine Aufgabe beendet hat.
- Der Proband muss nicht weiter kontrolliert oder beobachtet werden und kann den Fragebogen von jedem Ort aus ausfüllen.
- Der Fragebogen kann auf verschiedene Arten verteilt werden. Er kann im Papierformat ausgefüllt ein eingereicht werden, er kann über ein Computerprogramm ausgefüllt und sofort verschickt und ausgewertet werden oder über eine Befragung des Probanden vom Entwickler stellvertretend ausgefüllt werden.
- Die Auswertung kann im Nachhinein vorgenommen werden.
- Es ist einfacher, eine große Anzahl an Testern zu finden.
- Die größere Datenmenge verbessert die Sicherheit der Ergebnisse.

Auf der anderen Seite gehen damit einige Probleme einher. Es ist leicht, einem Probanden einen Fragebogen zur Bearbeitung zu geben, aber wie soll dieser Bogen im Detail aussehen? Die Daten, die der Fragebogen liefern kann, hängen von seinen Fragen, ihrem Aufbau und der Kontrolle der Antworten ab. Wie diese Fragen formuliert sein sollten, hängt von dem Ziel und den Fragen ab, die der Steller des Fragebogens hat. Außerdem ist nicht klar, wie der ausgefüllte Fragebogen ausgewertet werden soll. Antworten auf offene Fragestellungen etwa müssen vom Auswertendem langatmig durchgelesen und gedeutet werden. Werden konkrete Werte oder Meinungen abgefragt, so ist andererseits zu klären, wie und ob diese verrechnet werden. Auch die Kontrolle der Antworten ist wichtig. Ein unaufmerksamer Teilnehmer könnte mit falschen oder nicht genauen Antworten das Ergebnis verfälschen. Ist der Fragebogen selbst erstellt, sind zwar individuelle Fragestellungen möglich, aber es fehlen Tests und Validierungen des eigenen Fragebogens. Damit ist unklar, wie verwertbar und fehlerbehaftet die ermittelten Daten sind. Genauso wie es leicht ist, einen Fragebogen zu benutzen, so lassen sich auch Fehler machen, die den Nutzen wieder zu Nichte machen.

Die Lösung dieser Probleme stellen standardisierte Fragebögen dar. Mit klarer Aufgabenstellung und vorgegebenen Aufbau sind sie einfacher einzusetzen. Die Methoden zur Auswertung und die Verwendungsmöglichkeiten der ermittelten Werte sind ebenfalls gegeben. Idealerweise wurden sie bereits von Anderen verwendet und haben Daten, die ihre Validität und Genauigkeit bestätigen. Damit geben sie die Vorteile, die Evaluation auf Basis von Fragebögen hat, machen es aber auch gleichzeitig dem Anwender einfacher, sie korrekt einzusetzen und klare Antworten zu erhalten. Das neue Problem, das dabei aufkommt: welcher Fragebogen genau soll nun verwendet werden?

Im Folgenden werden vier verschiedene Fragebögen vorgestellt. Sie messen unterschiedliche Werte und unterscheiden sich in ihrem Umfang, ihrer Methodik und ihrer Auswertung. Ihre Verwendung hängt von der genauen Fragestellung ab, die ihr Verwender beantwortet haben will. Je nachdem, wie diese aussieht, bieten sich bestimmte Fragebögen zu unterschiedlichen Zeitpunkten in der Entwicklung an. Die Unterschiede zwischen den Fragebögen machen es notwendig, ihre Eigenschaften genauer zu beleuchten. Daraus kann geschlossen werden, welche Informationen sie dem Verwender geben, welche Besonderheiten für ihren Einsatz gelten und auf welche Schwächen geachtet werden muss. Schließlich kann daraus abgeleitet werden, wie sich die Fragebögen in der Praxis einsetzen lassen.

Die hier vorgestellten Fragebögen und ihre Hauptgrößen sind die System Usability Scale für die Usability, der Task Load Index für die Arbeitsbelastung des Anwenders, das User Experience Questionnaire für seine Nutzererfahrung und die Unified Theory of Acceptance and Use of Technology für die allgemeine Akzeptanz einer Technologie.

II. Vorstellung der Fragebögen

A. System Usability Scale (SUS)

Die System Usability Scale [1]stellt eine äußerst schnelle Möglichkeit dar, die Usability, zusammengesetzt aus Effizienz und Effektivität des Systems und Zufriedenheit des Anwenders, zu bestimmen. Sie wurde aus einem Fragenkatalog aus insgesamt fünfzig Fragen entwickelt. Aus diesen wurden diejenigen ausgewählt, die unter den Probanden die größten Antwortabweichungen aufweisen. Aus diesen Fragen wurden Likert Skalen entwickelt. Bei diesen wird vom Nutzer angegeben, wie sehr er einer Aussage zustimmt oder nicht zustimmt. Die Aussagen beziehen sich auf die Nutzungshäufigkeit, unnötige Komplexität, Einfachheit in der Handhabung, nötige Hilfe durch Experten, Funktionsintegration, Inkonsistenzen, Sicherheit im Umgang, Lerngeschwindigkeit, Vertrauen und nötiges Vorwissen für das System. Dabei wurde darauf geachtet, dass die Fragen unterschiedliche Aspekte der Usability abdecken, wie zum Beispiel Lerneffekte oder Lernaufwand, nötige und vorhandene Unterstützung bei Verwendung der Software und geplante Weiterbenutzung.

Die SUS besteht aus zehn Fragen, die über eine Likert Skala von 1 bis 5 bewertet werden. Dabei wechseln sich positive („das System war einfach zu verwenden“) mit negativen („Das System war umständlich zu verwenden”) Aussagen ab, um die Anzahl von unüberlegten Antworten zu reduzieren und den Probanden dazu zu bewegen, sich jede Frage durchzulesen. Zudem wiederholen sich teilweise die abgefragten Eigenschaften der Aufgabe, was eine weitere Kontrollmöglichkeit gibt. Sie wird angewendet, direkt nachdem der Tester das zu evaluierende System verwendet hat.

Bei der Auswertung wird nach einem Schema jeder Antwort eine Position von Null bis Vier zugewiesen. Positive und negative Aussagen werden dabei anders ausgewertet. Bei positiven Aussagen wird der Positionswert Minus 1 gerechnet, bei negativen gilt 5 Minus dem Positionswert. Damit ergeben Positive Fragen Werte von -1 bis 3, negative Werte von 5 bis 1. Die Punktwerte werden aufsummiert und mit 2,5 multipliziert, um einen SUS Wert von 0 bis 100 zu erhalten. Mit diesem Wert lässt sich anschließend die Usability verschiedener Systeme vergleichen.

B. Task Load Index (TLX)

Der Task Load Index dient dazu, die Arbeitsbelastung eines Probanden bei einer vorher definierten Aufgabe zu bewerten. Der TLX wurde von der NASA entworfen und diente ursprünglich für die Luft- und Raumfahrt. Er wird manchmal auch als NASA-TLX bezeichnet.

Die Arbeitsbelastung ist auf das menschliche Verhalten während der Bearbeitung der Aufgabe zentriert. Die subjektiv vom Probanden empfundene Arbeitslast wird in Faktoren unterteilt, die von der eigentlichen Aufgabe abhängen. Die im TLX verwendeten Faktoren wurden zu Skalen zusammengefasst, die in Experimenten an verschiedenen Aufgaben getestet und bestätigt wurden. Dabei sind Skalen entstanden, die sich auf die Anforderungen der Aufgabe, auf das Verhalten des Probanden und auf seine subjektive Meinung beziehen.

Für den TLX werden zuerst sechs Parameter vom Probanden abgefragt, durch die die Aufgabe beschrieben werden soll. Diese sind Geistige, Körperliche und Zeitliche Anforderung, die eingeschätzte Gesamtleistung, Frustration und die Anstrengung, um die Aufgabe zu erfüllen. Der Proband kann im ersten Schritt der Erhebung diese sechs Parameter auf einer 100-Punkte Skala bewerten, wobei nur Bewertungen in Fünferschritten erlaubt sind. im Zweiten wird eine Gewichtung der Parameter vorgenommen, in dem sie paarweise verglichen werden. Dabei soll sich der Proband entscheiden, welcher der beiden Werte wichtiger für die Arbeitsbelastung bei dieser Aufgabe war. Um alle möglichen Parameterpaare abzudecken, wird dies fünfzehnmal durchgeführt. Die Häufigkeit, mit der ein Aspekt gewählt wurde, bildet die Wichtung. Aus der Wichtung und Bewertung wird der Task Load Index gebildet, bestehend aus der gewichteten Beanspruchung durch die einzelnen Parameter und die aufsummierte Gesamtbeanspruchung [2].

C. User Experience Questionnaire (UEQ)

Das User Experience Questionnaire soll die User Experience messen, also den „subjektiv empfundenen Eindruck, den der Benutzer in Bezug auf das Produkt entwickelt hat“ [3]. Hier wurde anfangs mit 229 Aussagen gearbeitet, aus denen die achtzig aussagekräftigsten ausgewählt wurden, die mit Tests aus interaktiven Produkten weiter reduziert wurden. Dabei wurde auf verschiedene Softwareanwendungen und Plattformen wie Mobilgeräte geachtet.

Dieser Fragebogen besteht aus insgesamt sechsundzwanzig Fragen, die über eine siebenstufige Likert Skala beantwortet werden. Die Fragen verteilen sich auf sechs Dimensionen: Attraktivität, Effizienz, Durchschaubarkeit, Steuerbarkeit, Stimulation und Originalität. Je Dimension sind mindestens vier Aussagen zu treffen, um die Robustheit der Dimension zu verbessern und eine Kontrolle der Antworten zu ermöglichen. Antworten aus einer Dimension sollten dabei ähnliche Werte aufweisen. Aus den Antworten auf die Fragen werden je Dimension anschließend die statistischen Kennwerte berechnet. Aus ihnen werden die Validität der Erhebung sowie die Wertung der einzelnen Dimensionen ersichtlich.

D. Unified theory of acceptance and use of technology (UTAUT)

Auch die Akzeptanz, also ob der Nutzer das Produkt überhaupt annimmt und verwendet, ist entscheidend. Eine Methode, dies zu messen, stellt die UTAUT dar. Sie entstand als Vereinigung von acht einzelnen psychologischen Modellen von menschlichem Verhalten und Handlungsweisen. Diese sind:

- die Theory of Reasoned Action, die menschliches Verhalten beschreibt,
- das Technology Acceptance Model, das Vorhersagen zu Akzeptanz und Nutzung trifft,
- das Motivational Model, das Verhalten aus den Motivationen heraus modelliert,

- die Theory of Planned Behavior, die mit neuen Verhaltenskontrolltechniken die Theory of Reasoned Action erweitert,
- die Kombination aus Theory of Reasoned Action und Theory of Planned Behavior als Hybridmodell des menschlichen Verhaltens,
- das Model of PC Utilization, die die vorherigen Verhaltensmodelle beim Umgang mit einer Technologie betrachtet,
- Die Innovation Diffusion Theory, die den Begriff der Innovation einbringt
- und Die Social Cognitive Theory, die Lernvorgänge anhand von Vorbildern und Nachahmung beschreibt.

Aus ihnen wurden die wichtigsten Daten ausgewählt, um einen zusammenfassenden Fragebogen zu erstellen. In Tests wurde ermittelt, dass die UTAUT dabei bessere Ergebnisse liefert als die ursprünglichen Modelle [4]. Gerichtet ist die UTAUT dabei vor allem an Manager, die testen wollen, ob neue Produkte überhaupt erfolgsversprechend sein können und vom potentiellen Nutzer angenommen werden.

Es werden jeweils vier Fragen in insgesamt acht Kategorien gestellt. Die ersten vier (Leistungserwartungen, Aufwandserwartungen, sozialer Einflüsse und Einsatzbedingungen) stellen dabei die Schlüsselkategorien dar, um Nutzungsintention und Nutzungsverhalten zu ermitteln. Weiterhin werden Einstellung, Selbstbeschreibungsfähigkeit, Sorgen bei der Verwendung und geplante Verwendung des getesteten Produkts abgefragt. Die genaue Auswertung ist nicht vorgegeben. Wurde eine Auswertungsmethode gewählt, wie zum Beispiel eine Likert Skala, werden aus ihr je Kategorie die Kennwerte berechnet.

III. Vergleich

Mit diesen Fragebögen lassen sich verschiedene Größen einer Software überprüfen. Zur Abgrenzung der Fragebögen untereinander werden ihre wichtigsten Eigenschaften im Folgenden verglichen.

Dabei soll darauf eingegangen werden, welche Ziele mit den Fragebögen erreicht werden können, also welche Werte und Erkenntnisse aus ihnen gewonnen werden können und wie nützlich diese sind. Auch ihr Gebrauch in der Praxis ist von Bedeutung. Je nachdem, wie leicht oder umständlich ihre Durchführung und Auswertung sind, ändern sich ihre Verwendungsmöglichkeiten. Auch auf die Kontrolle ihrer Ergebnisse muss geachtet werden. Jeder Fragebogen verwendet andere Kontrollmechanismen zur Sicherung der Qualität der erhaltenen Daten. Aus diesen Eigenschaften ergibt sich auch, welche Effekte der Einsatz dieser Fragebögen auf die weitere Softwareentwicklung haben kann.

A. Eigenschaften von SUS

Die SUS misst ausschließlich die Usability eines Programms und hat nur gerade so viele Fragen, wie zum Erstellen eines Usability Wertes nötig ist. Unter den hier vorgestellten Fragebögen ist er mit zehn Fragen deutlich der kürzeste. Das macht ihn, zusammen mit den verständlich formulierten Fragen und einfachem Aufbau, ohne viel Aufwand verwendbar. Auch die Auswertung ist dementsprechend einfach.

Bei der Erstellung der Methode wurde aus einem großen Fragenkatalog genau die Menge an Fragen ausgewählt, die die Usability am effektivsten ermitteln kann. Die Verwendbarkeit und Validität wurde in weiteren Arbeiten kontrolliert und bestätigt [5]. Zur Kontrolle der Antworten sind die Aussagen abwechselnd positiv und negativ. Dies fängt unaufmerksame Probanden leicht ab, falls diese nur auf die Richtung ihrer Antworten achten, aber nicht die Fragen genau durchlesen. Der Fragentyp wird auch in der Auswertung beachtet.

Der Usability Wert ist aber für sich stehend wenig aussagekräftig und muss mit den Usability Werten anderer Anwendungen und Aufgaben verglichen werden. Außerdem ist der Wert selbst wenig nützlich, um die Usability zu verbessern, da nur angegeben werden kann, ob sie vergleichsweise gut oder schlecht ist. Zwar werden verschiedene Aspekte der Usability abgefragt, diese sind aber aus dem SUS Wert nicht ersichtlich. Um sie zu erhalten, müssten die einzelnen Aussagen noch einmal betrachtet werden und die Auswertung genauer erfolgen. Dabei stört aber, dass die Aussagen nur wenige Aspekte abdecken. Mehrere Fragen decken den gleichen Fakt ab, nur einmal in positiver und anschließend in negativer Formulierung. Zum Beispiel ist eine Aussage, das System sei leicht zu nutzen, eine weitere darauf, das System sei unhandlich in der Benutzung. Damit schrumpft die effektive Aussagekraft, die die 10 Aussagen haben, weiter. Daher bringt auch eine genauere Betrachtung dieser kaum einen Mehrwert.

Bei Einsatz von SUS ist ein iterativer Entwicklungsansatz passend. Veränderungen in der Software können mit SUS überprüft werden und die daraus folgenden Erkenntnisse sofort wieder in den Entwicklungsprozess einfließen.

Die System Usability Scale ist damit einfach und leicht zu verwenden, hat aber auch nur dementsprechend viel Wirkung. Ihr Einsatz bietet sich eher zu Beginn der Entwicklung an, um aus verschiedenen Prototypen denjenigen mit guten Usability Werten zu finden. Durch die Einfachheit der Methode ist es leicht, die SUS Bestimmung zu wiederholen und den Usability Wert während der Entwicklung zu verfolgen. Auch die Kombination mit anderen Usability Messungen ist möglich. Dabei gilt, dass die SUS mit anderen Usability Messungen korreliert [1].

B. Eigenschaften von TLX

Der Task Load Index ist umfangreicher, weil er neben einem Endergebnis (der Gesamtbeanspruchung) auch angibt, wie er sich aus den sechs im Verfahren abgefragten Parametern und ihren Wichtungen zusammensetzt. Damit gibt der TLX mehr Informationen, aber benötigt einen höheren Aufwand in der Verwendung.

Die Fragenanzahl ist an sich zwar gering, aber sie wird durch die fünfzehn zu treffenden Wichtungen erhöht, wodurch insgesamt einundzwanzig Entscheidungen zu treffen sind. Zudem müssen die zu ermittelnden Größen dem Probanden eventuell zuerst erklärt werden. Die Wichtung der Größen

ist zudem langatmig, da alle möglichen paarweisen Kombinationen abgefragt werden. Außerdem müssen die Werte im Anschluss noch gewichtet und summiert werden.

Der lange Prozess bei der Erfassung macht auch eine Kontrolle der Ergebnisse nötig. Bei der Durchführung kann der Proband leicht die Konzentration verlieren und unaufmerksam werden, vor allem bei dem Wichten der Parameter. Fünfzehnmal ähnliche Entscheidungen treffen zu müssen ist monoton und kann die Geduld des Probanden auf die Probe stellen. Dafür bietet der TLX in seiner Methode keine Hilfestellungen.

Der TLX liefert auch abhängig von der Art, wie der Fragebogen gestellt wird, andere Ergebnisse. Wird der TLX am Computer durchgeführt, sind die Ergebnisse im Durchschnitt um zwei Punkte höher als bei einer verbalen Befragung. Computerergebnisse sind im Vergleich zum Ausfüllen des TLX auf dem Papier sogar um sieben Punkte höher [2]. Die Verteilung der Ergebnisse bleibt dabei aber gleich, es ändert sich nur der durchschnittliche Wert der Kenngrößen und Gesamtbeanspruchung. Allerdings erschwert dies den Vergleich von TLX Werten, die mit unterschiedlichen Werkzeugen aufgenommen wurden. Wird der TLX mehrmals erfasst, sollte darauf geachtet werden, sich bei allen Wiederholungen für eine einzige Aufnahmemethode zu entscheiden, um die Auswertung einfacher zu halten. Ansonsten muss die Abweichung zwischen den Aufnahmemethoden hinzugezogen werden, um die Werte vergleichbar zu halten.

Für den höheren Aufwand gibt es auch mehr verwertbare Informationen. Die Gesamtbeanspruchung erlaubt es, alle Aufgaben der Software direkt miteinander zu vergleichen, wenn mit ihr mehrere TLX bestimmt wurden. Die Beanspruchung wird außerdem noch genauer in Parameter aufgeschlüsselt, mit denen Hinweise geliefert werden, welche Faktoren die Arbeitsbelastung erhöhen und wie sie verbessert werden kann. Mit Wichtung und Wertung erhält der Verwender der Methode einen genauen Überblick und kann im Anschluss zielgerichtete Problemlösungen erarbeiten.

C. Eigenschaften von UEQ

Der User Experience Questionnaire stellt auf den ersten Blick einen Mittelwert zwischen der Methodik von SUS und TLX dar. Zwar ist hier die Fragenanzahl höher, dafür sind sie einfach formuliert und über Likert Skalen schnell zu beantworten. Die Klassifizierung der Fragen ermöglicht eine genauere Auswertung als die SUS Methode. Zudem gibt es innerhalb einer Dimension keine Wiederholungen von Aussagen wie bei SUS. Die Kontrolle der Ergebnisse wird durch die höhere Fragenanzahl zum jeweiligen Themenkomplex erzielt. Im Vergleich zum TLX fehlt eine Wichtung der einzelnen Gruppen. Damit wird im UEQ davon ausgegangen, dass alle Dimensionen einen ähnlichen Einfluss auf die User Experience haben. Würde man eine Wichtung vornehmen, wäre der UEQ ähnlich komplex wie die Durchführung des TLX.

Die „User Experience“ ist dabei per Definition subjektiv und benutzerabhängig, benötigt also eine größere Menge an Probanden und eine klarere Auswahl dieser, zugeschnitten auf die zu testende Software. Zwar zeigt der Fragebogen an, an welchen Stellen die User Experience beeinträchtigt ist, bietet aber kaum Lösungen dafür an, mit Ausnahme der Dimensionen. Es wird empfohlen, Die User Experience auch mit anderen Methoden zu messen oder mit Usability Tests zu kombinieren [6]. eine Möglichkeit dazu wäre etwa das Kombinieren von UEQ mit SUS. Beide Tests sind einfach gehalten. Damit kann man sie leicht gleichzeitig einem Probanden vorlegen, um sowohl Usability als auch User Experience zu ermitteln. Sie bieten einzeln wenige Informationen, könnten aber in Wiederholung und Kombinationen bessere Ergebnisse erzielen.

D. Eigenschaften von UTAUT

UTAUT ist der Fragebogen mit den wenigsten Gemeinsamkeiten zu den anderen hier vorgestellten. Während die anderen Größen zwar ebenfalls psychologische Hintergründe haben, so beschäftigt sich die UTAUT ausschließlich mit den subjektiven menschlichen Wahrnehmungen. Während TLX, SUS und UEQ am praktischen Beispiel durchgeführt werden, ist die UTAUT theoretischer. Sie gibt einen allgemeinen Ausblick darauf, wie Menschen mit der getesteten Technologie umgehen.

Während die anderen Methoden in der Softwareentwicklung übliche Größen abdecken und zum Teil Überschneidungen aufweisen, behandelt die UTAUT den grundlegenderen Aspekt der Akzeptanz. Ist eine Softwarelösung zwar funktional und nutzbar, wird aber am Ende vom Nutzer nicht angenommen, ist dies ebenfalls ein Problem. Damit bietet sich die UTAUT eher zu Beginn der Entwicklung an. Sie hat wenig Einfluss auf den Entwicklungsprozess, davon abgesehen, dass bei Nichtakzeptanz der Software die Entwicklung eingestellt werden könnte.

Sie ist zudem nicht in erster Linie an die Entwickler einer Software gerichtet, sondern an Manager, die die Idee zu einem Produkt bewerten müssen. Die UTAUT soll dazu dienen, ihnen ein Werkzeug zu geben, um zu prüfen, ob ein neues Produkt oder eine Software akzeptiert wird. Aufgrund dessen ist die Verwendung von UTAUT zu Beginn der Entwicklung empfohlen, bevor Ressourcen in ein Projekt investiert werden, dass unter Umständen nicht von den Nutzern benutzt wird und damit von Anfang an zum Scheitern verurteilt ist.

Damit ist es schwieriger als bei den anderen Fragebögen, die UTAUT durchzuführen. Während die anderen Methoden direkt nach Erfüllung einer konkreten Aufgabe verwendet werden, ist dies bei UTAUT kaum möglich. Zu dem Zeitpunkt, bei dem sich UTAUT anbietet, fehlen die Aufgaben, die der Proband erfüllen kann. Damit ist UTAUT theoretischer und weniger aufgabenbezogen und konkret als die anderen Fragebögen.

Von den vorgestellten Methoden hat diese den für den Anwender höchsten Aufwand, um sie zur Verwendung aufzubereiten. Die Aussagen zum abgefragten „System“ sind bewusst allgemein gehalten, um für verschiedenste Anwendungen verwendbar zu sein, und müssen vor Verwendung erst auf das tatsächliche Problem angepasst und heruntergebrochen werden. Die Zusammensetzung der Themen ist aufgrund der Entstehung der Methode komplexer und unübersichtlicher. Auch die Skalierung und Auswertung muss der Software

angepasst werden. Dementsprechend ist die Durchführung und Beantwortung der Fragen ähnlich schwierig.

IV. Auswertung

Zusammenfassend wird klar, dass die Fragebögen alle eine Daseinsberechtigung haben, aber entsprechend ihrer Aufmachung unterschiedlich verwendet werden sollten.

Die Usability ist mit SUS schnell zu testen. Frühzeitig im Projekt angewendet, kann der Test wiederholt werden, um im Vergleich eine Variante mit guter Usability zu finden.

Sind einzelne Aufgaben der Software fertig implementiert, kann TLX zu Rate gezogen werden, um die Umsetzung in ihrem Nutzungsaufwand zu verbessern und zu optimieren.

UEQ kann wie ein gründlicheres SUS verwendet werden, um die Verwendbarkeit der Software zu überprüfen.

Die UTAUT ist mächtiger als die anderen Methoden, benötigt aber entsprechende Aufbereitung. Die Akzeptanz der Software ist ein eigener Komplex, den die anderen Methoden kaum abdecken. Zu Beginn eines Softwareprojekts eingesetzt, stellt sie sicher, dass das Produkt vom Nutzer akzeptiert wird.

Damit haben alle Fragebögen einen anderen Aufgabenbereich und decken verschiedene Methoden und Aspekte einer Software und ihrer Nutzer ab. Wie sie in der Praxis eingesetzt werden können, hängt vom konkreten Projekt ab, das getestet werden soll.

V. Anwendungsstrategien

Wie und welche der Fragebögen verwendet werden sollten, hängt vom Kontext der Software ab. Je nach Entwicklungsstand und Vorgehen ergeben sich andere Fragen im Entwicklungsprozess, zu deren Lösung einzelne Fragebögen oder mehrere in Kombination eingesetzt werden können.

A. allgemeine Hinweise

Einige allgemeine Fakten gelten dabei für alle Fragebögen und sollten bei jeder Durchführung eines Fragebogens beachtet werden. So sollten Fragebögen idealerweise unmittelbar nach Durchführung der Aufgabe, die in ihnen beschrieben werden soll, ausgefüllt werden. Wie der Fragebogen verbreitet wird, sollte einheitlich auf eine Plattform (Stift und Papier, Computer, Befragend...) festgelegt sein, um Abweichungen zwischen ihnen,wie sie etwa bei TLX möglich sind, zu vermeiden. Auch sollte streng kontrolliert werden, dass die Fragebögen ernsthaft und unbeeinflusst ausgefüllt werden. Bestimmte Begleitumstände können die Probanden zu bestimmten Antworttendenzen verleiten oder sie vom eigentlichen Test ablenken. Beispielsweise könnten Belohnungen nach Ausfüllen des Fragebogens, wie die Teilnahme an einem Gewinnspiel, dafür sorgen, dass die Probanden den Fragebogen überhastet ausfüllen. Besteht ein Abhängigkeitsverhältnis, stellt etwa ein Professor einen Fragebogen an seine Studenten, kann dies die Probanden einschüchtern, was wiederum die erfassten Ergebnisse verändert [3]. Fallen bei der Auswertung Ausreißer bei den Antworten einzelner Probanden auf, sollten diese kritisch betrachtet werden.

B. Anwendungszeitpunkt

Jeder Fragebogen bietet sich zu einem anderen Zeitpunkt in der Entwicklung an. Auch die Wiederholung, also das erneute Erfassen eines Fragebogens zu einem späteren Zeitpunkt, ist nur bei bestimmten Fragebögen sinnvoll.

Zu Beginn der Entwicklung kann sich zuerst die Anwendung einer UTAUT lohnen, vor allem dann, wenn es sich um neuartige Softwareanwendungen handelt, deren Akzeptanz beim Nutzer gefährdet sein könnte. So können rechtzeitig Maßnahmen getroffen werden, sollte die Anwendung von der Zielgruppe nicht gewünscht sein.

Zu einem späteren Zeitpunkt eingesetzt, bietet die UTAUT zwar aktuellere Akzeptanzwerte, aber kaum Lösungsmöglichkeiten für die ermittelten Probleme, weshalb stattdessen besser andere Evaluationen gewählt werden sollten, die auch Hinweise zum Lösen der Herausforderungen bieten. Da es sich bei den Fragen um grundlegende zum Zweck der Software handeln, die sich während der Entwicklung kaum ändern, bietet sich keine Wiederholung an.

Ebenfalls früh kann mit der Anwendung von SUS begonnen werden. Sie bietet einen schnellen Überblick über die Usability. Während des Projekts regelmäßig wiederholt, kann die Entwicklung der Usability dokumentiert werden. Die Wiederholung ist dabei äußerst wichtig, da nur so vergleichbare SUS Werte entstehen. Auch bei kleineren Veränderungen kann sich die subjektiv wahrgenommene Usability verändern, weshalb auch kurz aufeinanderfolgende Wiederholungen sinnvoll sind.

Später kann der TLX hinzugezogen werden, sobald kritische Aufgaben vollständig implementiert sind. Allerdings sollte TLX auch nicht zu spät hinzugezogen werden. Die umfassendere Auswertung benötigt ausreichend Zeit, ansonsten können die aus ihm gewonnen Erkenntnisse nicht mehr in der Entwicklung eingesetzt werden. Die durch ihn ermöglichte umfassende Auswertung sorgt dafür, dass er weniger häufig wiederholt werden muss, außer die Technik zum Lösen der Aufgabe hat sich stark verändert.

Das UEQ kann ebenfalls später eingesetzt werden, wenn vollständige Aufgaben eine genauere Bewertung der Nutzererfahrung ermöglichen. Wird es eingesetzt, kann eine Kombination mit SUS verwendet werden, um die Gemeinsamkeiten beider Fragebögen auszunutzen. Auch bei UEQ gilt, dass eine Wiederholung nur bei tiefgreifenden Veränderungen nötig ist. Ansonsten bietet sich eher eine weitere SUS Wiederholung an.

C. getestete Software

Welche Fragebögen anwendbar sind, hängt auch vom Typ und Charakter der Software ab.

UTAUT ist bei neuartigen Lösungen relevant, aber auch bei gewünschten Anwendergruppen, für die besondere Hindernisse gelten, etwa Kinder, Rentner oder Menschen mit Behinderungen. Bei ihnen ist die Akzeptanz unter Umständen niedrig, sollte also mit UTAUT geprüft werden. Bei geringen Akzeptanzwerten kann so geprüft werden, ob die weitere Entwicklung sinnvoll ist oder ob an die Grundbedingungen, die die Akzeptanz verringern, verändert werden müssen.

Kommen in der Software komplexe Aufgabenstellungen vor, sollte der TLX eingesetzt werden, um die Belastung für den Nutzer gering zu halten. Auch bei kleineren Aufgaben sind oftmals Optimierungen möglich, allerdings muss dann abgewogen werden, ob der Aufwand zur Erhebung des TLX in Relation zum möglichen Gewinn an verringerter Arbeitsbelastung steht.

SUS und UEQ sind dagegen bei allen Programmen relevant, bei denen Nutzerinteraktion möglich ist. Die Interaktivität hängt direkt davon ab, wie gut die Usability und damit die User Experience ist. Damit ist ihre Verwendung grundsätzlich immer nützlich. Bei kleineren Projekten kann dabei die Verwendung von SUS ausreichen, am besten iterativ. Sind die zu testenden Funktionen umfangreicher, kann auch das UEQ hinzugezogen werden, unter Umständen auch in Kombination mit SUS.

D. Kombinationen

Die Wirkung der Fragebögen kann verbessert werden, wenn während der Entwicklung verschiedene in Kombination eingesetzt werden. Dabei sind aber nur bestimmte Paare von Interesse.

Die SUS lässt sich ausgezeichnet mit der UEQ kombinieren. Beide messen ähnliche Werte und sind sich in Umfang und Verwendung ähnlich. Die Kombination sorgt dafür, dass die wenig interpretierbaren SUS Werte höhere Aussagekraft erhalten, weil die SUS Fragen mit ähnlichen aus dem UEQ verglichen werden können.

SUS und TLX haben ebenfalls Überschneidungen. Verbessert sich die Usability, so wird die Effizienz besser, mit der eine Aufgabe bearbeitet werden kann, was die gefühlte Arbeitslast verringert. Im Vergleich zum TLX ist der Aufwand für das Durchführen eines SUS Tests gering genug, um nicht weiter ins Gewicht zu fallen.

SUS und die UTAUT sind schwerer zu kombinieren. Es fehlt ein Zeitpunkt, an dem ein SUS Test möglich ist, weil Aufgaben abgearbeitet werden können, aber der gleichzeitig so früh in der Entwicklung liegt, dass eine UTAUT Betrachtung in Frage kommt. Dabei gilt aber, dass die Usability ein guter Anhaltspunkt dafür sein kann, wie der Nutzer eine Software akzeptiert. Ist die allgemeine Akzeptanz zu Beginn bereits mit UTAUT getestet worden, kann ein späterer SUS Test aber durchaus gut geeignet sein, um die Akzeptanzentwicklung anhand der Usability weiter kontrollieren zu können.

UEQ und TLX haben wenige Überschneidungen. Während beide zwar subjektiv wahrgenommene Größen in Form von Arbeitsbelastung und User Experience betrachten, so versucht der TLX doch mit Gruppierung und Wichtung, die Arbeitsbelastung möglichst objektiv aus den einzelnen Faktoren zu ermitteln, während das UEQ vor allem die User Experience des einzelnen Nutzers erfasst. Damit bietet die Kombination beider einen umfassenderen Überblick über die Software und ihre Entwicklung, als es die anderen Kombinationen ermöglichen. Dabei steigt allerdings auch der Auswertungsaufwand. Sowohl TLX und UEQ bestehen aus mehreren Dimensionen und Werten. Die Auswertung dieser muss in der Entwicklung eingeplant werden.

UEQ und UTAUT sind beides Methoden, die einen stärkeren psychologischen Hintergrund haben. Dabei haben sie jeweils einen anderen Blickwinkel. Der UTAUT bietet den theoretischen Aspekt zum Nutzerverhalten, die UEQ gibt den praktischeren Teil der User Experience an konkreten Aufgaben an. Damit ist ein umfassender Überblick über die Art und Weise möglich, wie die Software verwendet wird.

VI. Ausblick

Es wird klar, dass es für die verschiedensten Softwarearten und Aspekte der Entwicklung passende Fragebögen gibt. Werden diese korrekt ausgewählt, durchgeführt und ausgewertet, kann die Entwicklung davon profitieren. Man sollte kaum in die Situation kommen, einen eigenen Fragebogen erstellen zu müssen, auch wenn die Option besteht, einen bestehenden und validierten auf das eigene Programm zuzuschneiden.

Das einzige dabei weiterhin bestehende Problem ist die Auswahl und das Finden eines passenden Fragebogens. Unter einer Vielzahl von einfachen und nur für einzelne Anwendungen erstellten Fragebögen ist es schwierig, einen standardisierten und validierten für das eigene Problem zu finden. Außerdem muss man wissen, wie man nach den passenden Fragebögen sucht. Beispielsweise bietet die UTAUT einen interessanten anderen Blickwinkel, ist aber an sich eher wenig bekannt in der Softwareentwicklung.

Sollten trotz allen Vorteilen und Hinweisen zur optimalen Verwendung von Fragebögen noch Schwierigkeiten bestehen, so gibt es immer noch andere Evaluationsmethoden, die herangezogen werden können. Auch ein eigener Fragebogen kann erstellt werden, angelehnt an die standardisierten und ihre Funktionsweise.

Unter den existierenden Methoden stellen Fragebögen einen beliebten, ersten Start in die Evaluation dar. Entsprechend aufbereitet, lassen sie sich auch für mehr einsetzen. Es spricht nichts dagegen, sie zu Beginn der Evaluation einzusetzen und zu sehen, bis zu welchem Stand sie einen bringen können.

Literatur

[1] J. Brooke, "Sus: A quick and dirty usability scale," vol. 189, 11 1995.

[2] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, 1988, vol. 52, pp. 139 – 183.

[3] M. Rauschenberger, M. Schrepp, and J. Thomaschewski, "User experience mit fragebögen messen – durchführung und auswertung am beispiel des ueq." 09 2013.

[4] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.

[5] K. Orfanou, N. Tselios, and C. Katsanos, "Perceived usability evaluation of learning management systems: Empirical evaluation of the system usability scale," *The International Review of Research in Open and Distributed Learning*, vol. 16, no. 2, 2015.

[6] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *HCI and Usability for Education and Work*, A. Holzinger, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 63–76.