Definition of Done Done Done?

Warum Akzeptanzkriterien auch auf realem Nutzungsverhalten fußen sollten

Britta Karn
UX Researcher
Centigrade GmbH
Saarbrücken, Saarland, Deutschland
britta.karn@centigrade.de

ABSTRACT

Product Owner (POs) haben kaum Zeit, Backlog-Items zu priorisieren und eindeutige Akzeptanzkriterien zu schreiben. Mit der in der Praxis vorherrschenden „Definition of Done“ lügen sie sich zudem oft in die Tasche: eine User Story kann vom Team als abgeschlossen und qualitativ gut eingeschätzt werden („Done“) und auch auslieferfertig in den Code integriert worden sein („Done Done“), wird jedoch im Feld trotzdem nicht von den Nutzern akzeptiert. Wirklich fertig ist aus UX Sicht nur eine User Story, die bewiesenermaßen die Akzeptanz des Nutzers findet und somit im Feld performt („Done Done Done“). Doch wie kann das dritte „Done" in agilen Prozessen erreicht werden? Im Forschungsprojekt Opti4Apps wurde ein neuartiger Ansatz der Nutzungsdatenanalyse entwickelt, der Daten auf der Basis von User Stories sammelt und aggregiert. Durch reale und messbare Nutzerakzeptanzkriterien, wird es jedem PO möglich, Entscheidungen bezüglich seiner Backlog-Priorisierung zu treffen. Die Nutzungsdatenanalyse ergänzt den "Continuous UX" Baukasten, um in agile Prozesse regelmäßig und automatisiert Nutzerfeedback einzubinden.

KEYWORDS

Agile, Definition of Done, User Experience, UX Research

1 Einleitung

Digitale Produkte werden heutzutage größtenteils agil entwickelt [1]. Das ist auch notwendig, da die time-to-market vor allem von mobilen Applikationen immer kürzer wird und Produkte schnell veraltet sind. Doch durch diese kurzen Entwicklungszyklen, meist in Form von kleinen Produktpaketen, die dem Nutzer einen Mehrwert bringen, den so genannten Minimum Viable Products (MVPs), wird häufig die eher zeitintensive Qualitätssicherung außer Acht gelassen. Nichtsdestotrotz ist bei der Entwicklung mobiler Apps eine nutzerzentrierte Herangehensweise erstrebenswert, entscheidet doch letztlich auch die User Experience über den späteren Erfolg einer Anwendung. Das vorherrschende Vorgehen in agilen Prozessen besteht daraus, eine Funktion bzw. eine User Story anhand einer Checkliste von einem Team während des Sprints als abgeschlossen und qualitativ angemessen bewerten zu lassen („Definition of Done“, [2]) und schließlich auslieferfertig in den Code zu integrieren („Definition of Done Done“ [3]). Wenn jede User Story dann noch ihre ganz spezifischen, Feature-getriebenen Akzeptanzkriterien erfüllt, gilt die Story als „fertig“. Was hier vergessen wird ist, dass dieses „Done“ noch nicht heißt, dass die Nutzer das entwickelte Feature auch nutzen bzw. es brauchen. Es fehlt an Kriterien, die die Akzeptanz bei den Nutzern sowie die Performance im Feld bewerten. Die Entwicklung und Iteration auf Basis von Nutzerbedürfnissen entscheidet am Ende darüber, wie erfolgreich eine App ist und dies manifestiert sich letzten Endes in Downloadzahlen oder Sterne-Bewertungen. Die Nutzerakzeptanz zum ultimativen Erfolgskriterium zu küren („Definition of Done Done Done“) ist für Product Owner mit Sinn für nutzerzentriertes Design daher unerlässlich.

Besonders in agilen Projekten ist die manuelle Berücksichtigung von Nutzerfeedback aber schwer realisierbar, da sie sehr viel Zeit kostet. Die bekannten UX Research Methoden sind größtenteils zu behäbig für den schnelllebigen Alltag bei der agilen Entwicklung von Software-Produkten. Dadurch geht den Projekten jedoch ein wichtiger Maßstab verloren, um einschätzen zu können, ob ein Produkt später erfolgreich und akzeptiert sein wird oder nicht. Es stellt sich also die Frage, mit welchen Methoden Nutzerfeedback in die agile Entwicklung eingebunden werden kann, ohne den agilen Prozess zu behindern oder zu verlangsamen.

Eine bereits bekannte Methode, anhand derer schnell und kostensparend Nutzerrückmeldungen erfasst werden kann, ist die Analyse von Nutzungsdaten. Wie diese Methode konkret im UX Kontext genutzt werden kann, war eine Fragestellung des Forschungsprojektes Opti4Apps.

Veröffentlicht durch die Gesellschaft für Informatik e.V. und die German UPA e.V. 2019 in S. Hess & H. Fischer (Hrsg.):
Mensch und Computer 2019 – Usability Professionals, 08.-11. September 2019, Hamburg

https://doi.org/10.18420/muc2019-up-0263

2 Entwicklung der Idee im Projektkontext

Im vom BMBF geförderten Forschungsprojekt Opti4Apps wurde ein Qualitätssicherungsansatz (siehe Abbildung 1) entwickelt, der die oben beschriebene Forschungsfrage betrachtet: Gibt es einen praktikablen Weg, die Qualität mobiler Apps zu verbessern, indem Nutzungsdaten und Nutzerfeedback (semi-) automatisiert gesammelt und auswertet werden? Dabei sollte sowohl implizites (Nutzerdaten) als auch explizites (textuelles Feedback z.B. in App Stores) Nutzerfeedback einbezogen werden.

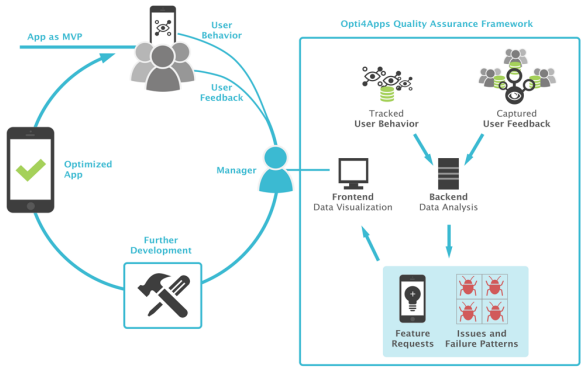

Abbildung 1: Der Opti4Apps Ansatz

In dem Forschungsprojekt Opti4Apps arbeiteten vier Projektpartner zusammen (Bridging IT, Fraunhofer IESE, Hochschule Heilbronn und Centigrade), die jeweils ihre Expertise in das Projekt einbrachten. Die Rolle von Centigrade bestand vor allem darin, ein Prozessmodell zu entwickeln, anhand dessen Nutzerfeedback in den Entwicklungsprozess integriert werden kann sowie die Nutzerstudien zu leiten und auszuwerten. Die Projektpartner beschäftigten sich darüber hinaus vor allem mit der Entwicklung der Tracking Library zur Aufzeichnung der Nutzungsdaten, mit verschiedenen Feedbackkanälen (App Store Bewertungen etc.) sowie der Automatisierung der Interpretation von Fehlermustern.

Schon zum Zeitpunkt des Projektantrages von Opti4Apps stand fest, dass die Steigerung der vom Nutzer wahrgenommenen App-Qualität vor allem durch automatisiert erfasste Nutzungsdaten erreicht werden sollte, da diese schnell, kostengünstig und in großem Umfang gewonnen werden können.

Die ersten Schritte im Projekt bestanden aus Recherchen und Befragungen zu agilen Methoden und Vorgehensweisen. Dadurch wurde eine Wissensbasis in diesem Themenbereich aufgebaut und erste Rückschlüsse auf ein Prozessmodell zur Integration von Nutzerfeedback gezogen. Centigrade konnte außerdem auf die Erkenntnisse aus dem Forschungsprojekt PEBeMA zurückgreifen, wo ein Tool zur Prozessbegleitung von UX Managern entwickelt wurde. Dieses sollte bei Opti4Apps auf den mobilen Kontext übertragen werden und wenn möglich zur Unterstützung der Erarbeitung der Forschungsfrage dienen. Das Tool wurde basierend auf dem Continuous UX Baukasten [4] entwickelt, der auch bei Opti4Apps die Grundlage für die weiteren Arbeiten bildete. Anhand eines Beispielprojektes mit dem Projektpartner Bridging IT konnte die Anwendbarkeit des Continuous UX Baukastens im mobilen Kontext bestätigt werden.

Nachdem diese theoretischen und evaluativen Schritte getan waren, sollten erste Daten generiert werden. Hierzu diente eine Kooperation mit der Deutschen Messe AG, die die App für eine ihrer Messen für die Zwecke des Projektes zur Verfügung stellte. Hierbei wurden erste Versuche mit der Tracking Library, die in die Messe-App integriert wurde, unternommen sowie eine qualitative Studie vor Ort durchgeführt. Die Tracking Library zeichnete dabei zunächst alle Interaktionen der Nutzer mit der App auf, egal ob auf aktiven oder inaktiven Bereichen des Interfaces. Trotz einiger technischer Probleme und recht weniger Nutzer, konnten dennoch einige Daten gesammelt und ausgewertet werden.

Diese ersten Datenanalysen offenbarten, dass „Big Data“ ohne konkrete Fragestellungen in Bezug auf UX Metriken keine Mehrwerte für Product Owner erzielen wird bzw. sich sinnvolle Interpretationen als viel zu zeitaufwändig für den Praxisalltag erweisen. Trotz der verhältnismäßig wenigen Nutzer auf der Messe entstanden sehr große Datenmengen. Um sinnvolle Entscheidungen auf Basis dieser Nutzungsdaten abzuleiten, benötigt es ein Mindestmaß an Kontextinformation, wie z.B. “Welcher Nutzer agiert in welcher Rolle mit welchem Ziel?“ Letztere Information sollte im agilen Entwicklungsprozess eigentlich bereits vorliegen – nämlich in Form einzelner User Stories –, geht jedoch kurz nach der Umsetzung verloren, da implementierten User Stories in der Praxis oft keine weitere Relevanz mehr beigemessen wird.

Es entstand daraus die Idee, die Datenauswertung auf Basis von User Stories und deren Umsetzungen vorzunehmen, weil eine App im besten Fall sowieso sukzessive aus diesen Organisationseinheiten hervorgegangen sein sollte. Da die Umsetzung von User Stories zudem ohnehin mit Akzeptanzkriterien versehen wird, wurden spezielle Nutzer-Akzeptanzkriterien eingeführt, die auf gängigen UX Metriken fußen. Mit dieser Methode wurde das kontinuierliche Einbeziehen von Nutzerbedürfnissen über den gesamten Produktentwicklungszyklus (also auch nach Auslieferung) adressiert.

Leider konnte die Messeapp im weiteren Verlauf nicht für die Weiterentwicklung der ersten Idee genutzt werden. Mit Hilfe einer App zur Arbeitszeiterfassung von wissenschaftlichen Mitarbeitern, die die Hochschule Heilbronn während des Projektes entwickelte, wurde daher die neue Methode entworfen, getestet und iterativ verbessert (siehe Kapitel 3).

Die Zeiterfassungsapp wurde zunächst in ihre wichtigsten User Stories aufgeteilt. Hierzu wurde eine Persona (Wladimir Michels, wissenschaftlicher Mitarbeiter) und ein Szenario entwickelt. Die relevanten User Needs und damit User Stories wurde definiert und ihren Umsetzungen zugeordnet. Hierbei wurde der Prozess sozusagen rückwärtsgegangen, da die App ursprünglich nicht auf Basis des nutzerzentrierten Continuous UX Prozesses entwickelt wurde. Diese Vorgehensweise zeigte jedoch, dass die Methode auch dann anwendbar ist, wenn zunächst nicht nach einem nutzerzentrierten Prozess entwickelt wurde. Nach der Priorisierung der User Stories wurde definiert, anhand welcher

Key Performance Indicators (KPIs) die Qualität der App bzw. die Akzeptanz der einzelnen User Stories und deren Umsetzung bewertet werden. Als KPI im Beispielfall wurde zumeist die UX Metrik „time on task“ der Nutzer gewählt [5] [6]. Etwas spezifischer handelt es sich in diesem Fall um die „time on story“, also die Zeit, die zwischen der ersten und letzten Interaktionsmöglichkeit liegt, die im Rahmen einer User Story umgesetzt wurde. Da die „time on story“ automatisiert nur während der Laufzeit einer App bestimmt werden kann, wurden im Zuge der Hauptstudie die zur Umsetzung der User Story gehörenden Ein- und Ausstiegspunkte (Entries und Exits) mit Hilfe der Tracking Library im Quellcode markiert. Jede Nutzer-Interaktion, die zwischen Ein- und Ausstiegspunktes der Umsetzung der Story durchgeführt wurde, wurde von der Tracking Library auch entsprechend dieser User Story zugeordnet. Die „time on story“ wurde anschließend in Relation zu einem Referenzwert gesetzt, der die optimale Durchführung abbildete. Mehr dazu in Kapitel 3.

In Abbildung 2 ist anhand eines Zustandsdiagrammes dargestellt, wie die Übergänge zwischen den Repräsentationen der User Stories (blaue Kästen) definiert wurden. Konkrete Interaktionen (Pfeile) triggern die Ausstiegspunkte der Repräsentationen einer Story und meist gleichzeitig den Einstieg in die nächste. Der Ausstieg kann dabei „constructive“ (wie erwartet), „neutral“ (anderer Weg, aber trotzdem in Ordnung) oder „destructive“ (unerwartet, negativ für die Journey) sein. Beispiel: der Nutzer startet auf der Monatsübersicht der Zeiterfassungsapp (siehe Abbildung 2). Entsprechend der Journey wählt er den Plus-Button, um die Zeit des aktuellen Tages zu erfassen. Am Ende der Eingabe der Zeitdaten speichert er jedoch nicht (da z.B. der Speichern-Button missverständlich ist), kehrt zur Monatsübersicht zurück und verliert so seine eingegebenen Daten. Hier wäre der Übergang zwischen Monatsübersicht und Zeiterfassung „constructive“, der zwischen Zeiterfassung und Monatsübersicht aber „destructive“ und müsste später näher betrachtet werden. Die Auswertung der Ausstiegspunkte sollte somit über die Metriken hinaus Hinweise darauf liefern, welche Umsetzungen der User Stories akzeptiert wurden und welche nicht.

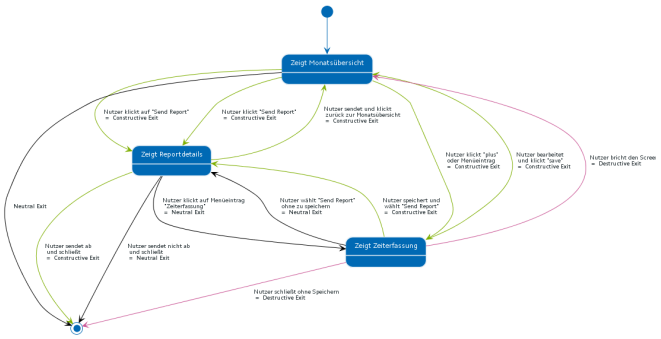

Abbildung 2: Zustandsdiagramm mit den Übergängen zwischen User Stories

Im weiteren Verlauf werden die Vor- und die Hauptstudie im Projekt im Detail vorgestellt, anhand derer die erste Idee der Methode weiterentwickelt und evaluiert wurde.

3 Studien im Rahmen des Projektes

In den beiden Studien im Labor wurde die neue Datenauswertungsmethode evaluiert und weiterentwickelt. In der Vorstudie lag der Fokus dabei vor allem auf der Evaluation und Weiterentwicklung der Tracking Library, während in der Hauptstudie die Frage beantwortet werden sollten, inwieweit es möglich ist, Aussagen über die Qualität einer mobilen Applikation nur anhand von automatisiert gesammelten Nutzungsdaten zu treffen.

Bei der prototypischen Anwendung, die in den Studien genutzt wurde, handelte es sich um die sogenannte „Timely App“, eine Zeiterfassungsapplikation für wissenschaftliche Mitarbeiter (siehe Abbildung 3). Mit dieser App ist es einem wissenschaftlichen Mitarbeiter möglich, seine Arbeitszeiten auf dem Mobilgerät zu erfassen und zudem Krankheiten sowie Urlaubszeiten mitzuteilen, was bisher nur in Excel oder auf Papier möglich war.

Die beiden Studien wurden mit wissenschaftlichen Mitarbeitern von zwei deutschen Hochschulen durchgeführt (entsprechend der Persona), die anhand eines Testszenarios mit der Zeiterfassungsapplikation auf einem Smartphone interagierten. Dabei wurden Videoaufzeichnungen der Personen und der Interaktionen gemacht, um die Auswertung zu erleichtern (Vorstudie). In der Hauptstudie haben die Probanden die App ohne Videoaufzeichnung genutzt, jedoch weiterhin mit Moderation der Aufgaben.

Die Daten wurden im Anschluss unter Berücksichtigung von UX KPIs ausgewertet.

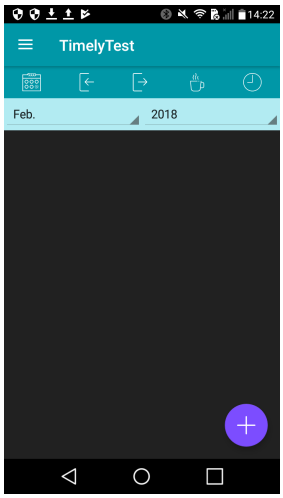

Abbildung 3: Monatsübersicht der Timely App

3.1 Vorstudie

Die erste Studie wurde mit vier wissenschaftlichen Mitarbeitern der Hochschule Heilbronn durchgeführt. Hierbei bekamen die Probanden ein Handy mit der Timely App zur Verfügung gestellt und wurden durch die Moderatorin anhand von Aufgaben durch die Applikation geleitet. Ziel war es, Fehlermuster in der Interaktion zu erkennen, sowie die Auswertungsmethode zur Vorbereitung der Hauptstudie zu evaluieren. Die Teilnehmer trugen eine GoPro Kamera auf dem Kopf, damit Bewegungen und Blickrichtungen aufgezeichnet werden konnten. Zudem wurde der Bildschirm des Handys mit einem Screencapturing-Tool aufgezeichnet, um später die Auswertung zu erleichtern. Die Teilnehmer arbeiteten circa 15 Minuten mit der App und sollten unter anderem ihre Arbeitszeiten für die zurückliegende Woche erfassen. Dabei wurden Kommentare und beobachtete Probleme von der Moderatorin mitnotiert.

Im Nachgang der Durchführung wurden die automatisiert erhobenen Daten im Kontext der live beobachteten Daten ausgewertet. Der Fokus der Auswertung lag auf der Analyse der Güte der elf definierten User Stories sowie der Optimierungspunkte für die Tracking Library. Die Start- und Endpunkte der Umsetzung der User Stories mussten in der Vorstudie noch manuell aus den Videos ausgelesen werden. In der Hauptstudie sollten dementsprechend Start- und Endpunkte der Repräsentationen der User Stories im Vorfeld definiert und von der Tracking Library automatisiert erfasst werden, damit eine Auswertung der Daten ohne Hilfsmittel bzw. Vor-Ort-Tests möglich ist.

Es fiel auf, dass einige der User Stories in dem vorliegenden Testkontext nicht erfasst und analysiert werden konnten, wie z.B. der Nutzungszeitraum der App pro Tag.

Bei vielen anderen Stories wurde der Parameter „time on task“ im Vergleich zu einem Referenzwert als KPI für die Güte verwendet. Die „time on task“ wurde anhand der aufgezeichneten Videos gestoppt. Hier zeigte sich als Optimierungspunkt für die Hauptstudie, dass die Zeit der einzelnen Umsetzungen der User Stories automatisch mitgetrackt werden sollte.

Trotz einiger Einschränkungen bei der Datenauswertung, konnte anhand der vorliegenden Daten bereits erkannt werden, dass bei fünf der elf umgesetzten User Stories das geometrische Mittel der Testdurchläufe nicht dem geforderten Referenzwert entsprach und signifikant darüber lag. Hier kann die Akzeptanz durch die Nutzer nicht bestätigt werden.

Kurzer Exkurs: Es wurde für die Datenauswertung das geometrische und nicht das arithmetische Mittel verwendet, da die „time on task“ die Tendenz hat, positiv schief verteilt zu sein (linksschief). Der Grund dafür ist, dass in dieser Art Studien immer Teilnehmer zu finden sind, die eine vergleichsweise lange Durchführungszeit benötigen, um ihre Aufgaben abzuschließen. Dies lässt sich dadurch erklären, dass einige Menschen mit Smartphones, Betriebssystemen oder Gesten weniger vertraut sind. Andererseits ist es für einen Nutzer eher schwer möglich deutlich schneller als ein Minimalmaß an Zeit zu werden. Diese wenigen, aber deutlich längeren Zeiten erhöhen fälschlicherweise die mittlere „time on task“. Der Median wäre in diesem Fall ebenso nützlich gewesen, um mit Ausreißern umzugehen. Aber hier gibt es ein Problem mit kleinen Stichprobengrößen, bei denen der Stichprobenmedian den Bevölkerungsmedian recht schlecht schätzt (insbesondere bei Stichproben von weniger als 25 Personen). Im UX-Bereich sind solche Stichprobengrößen durchaus üblich (wie auch in der vorliegenden Studie), sodass der Median nicht die Methode der Wahl sein konnte. In der Literatur wird für solche Fälle das geometrische Mittel empfohlen [7].

Die Umsetzungen dreier weiterer User Stories waren erfolgreich (Referenzwert lag im Konfidenzintervall der „time on task“) und drei User Stories konnten im Rahmen der Studie nicht ausgewertet werden. Die Videos und teilweise auch die Daten boten außerdem erste Hinweise auf Fehlermuster in der Interaktion, wie beispielsweise der fehlende Aufforderungscharakter von Interaktionselementen. Auch hier war der Wunsch, diese zukünftig automatisch in den Daten zu erkennen.

Anhand der bereits erwähnten Optimierungsanforderungen an die Tracking Library, wurde diese überarbeitet und für die Hauptstudie vorbereitet.

3.2 Hauptstudie

Die Hauptstudie wurde mit zehn wissenschaftlichen Mitarbeitern der Universität des Saarlandes aus den Fachbereichen Linguistik und Psychologie durchgeführt. Die Teilnehmer bekamen erneut ein Smartphone mit derselben Version der Timely App zur Verfügung gestellt wie in der Vorstudie und wurden ebenfalls durch die Moderatorin anhand eines Testszenarios durch die Applikation geleitet. Im Gegensatz zur Vorstudie bekamen die Teilnehmer nun keine Kamera mehr aufgesetzt, da der Fokus auf der Auswertung der automatisierten Daten und deren alleiniger Aussagekraft lag. Die Durchführung erfolgte ansonsten wie in der Vorstudie.

Im Anschluss an die Studiendurchführung wurden ausschließlich die Daten der automatisierten Erhebung betrachtet und analysiert. In den Daten waren nun die vorher definierten Start- und Endpunkte der umgesetzten User Stories ersichtlich sowie ein Zeitstempel für jedes Ereignis in den Daten gegeben. Es wurden nur noch acht User Stories evaluiert, da die drei User Stories, die schon in der Vorstudie nicht auswertbar waren, hier weggelassen wurden.

Es konnte anhand der Daten erkannt werden, dass sieben der acht User Stories nicht erfolgreich umgesetzt waren. Als Beispiel sei hier das Anlegen des Working Profiles genannt (siehe Abbildung 3 und 4). Das Working Profile dient in der App dazu, die eigene Arbeitszeit, die laut Arbeitsvertrag pro Tag gearbeitet werden muss, zu hinterlegen. Bevor der Nutzer dies nicht getan hat, kann er auch keine Arbeitszeiten erfassen. Mit dieser Aufgabe hatten die Teilnehmer große Probleme. Zunächst bestand die Herausforderung darin, überhaupt zu verstehen, dass das Working Profile benötigt wird und dann den Menüpunkt zu erreichen, im Weiteren sollten dann die Zeiten eingetragen werden.

In den Diagrammen (Abbildung 4 und 5) ist die „time on task“ für das Anlegen des Working Profiles aufgezeigt, wenn der Nutzer

den Menüpunkt bereits gefunden hat. Hierbei zeigte sich, dass alle Teilnehmer für das Eintragen der Zeiten im Mittel, aber auch jeder einzelne (siehe Abbildung 4 deutlich länger als den Sollwert von 20 Sekunden brauchten, der für die erfolgreiche Ausführung dieser Aufgabe erwartet werden konnte. Abbildung 4 zeigt die „time on task“ für jeden einzelnen Teilnehmer der Studie. Die blaue Linie zeigt den Referenzwert von 20 Sekunden, die orange Linie den Durchschnittswert über alle Teilnehmer. Hierbei ist bereits ersichtlich, dass beide Linien recht weit auseinanderliegen. Teilnehmer 6 und Teilnehmer 9 stechen dabei mit besonders langen Zeiten hervor. Generell ist keiner der Teilnehmer genau beim Referenzwert gelandet, einige sind jedoch in der Nähe davon. Anhand dieser Abbildung kann jedoch noch keine Aussage über die Signifikanz des Ergebnisses getroffen werden. Die zweite Grafik zeigt den gemittelten Wert der „time on task“ über alle Teilnehmer mit dem entsprechenden Konfidenzintervall. Hier zeigt sich, dass der Sollwert (grüne Linie) nicht im Konfidenzintervall der „time on task“ der Teilnehmer lag und somit die Umsetzung der User Story nicht erfolgreich war, da der das geometrische Mittel signifikant vom Sollwert abweicht.

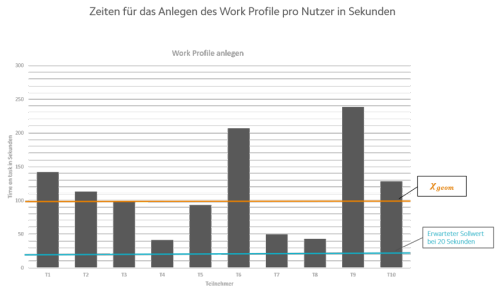

Abbildung 4: Zeiten pro Teilnehmer für das Anlegen des Working Profiles in Sekunden

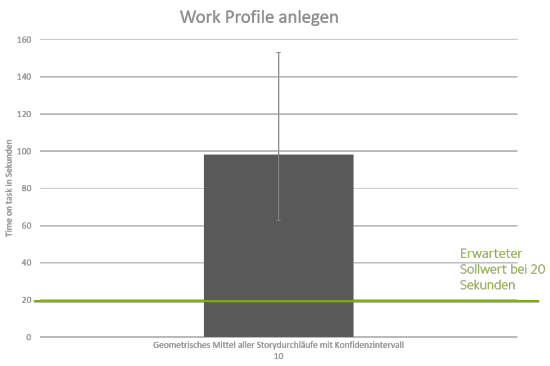

Abbildung 5: Konfidenzintervall um das geometrische Mittel der „time on task“ für das Anlegen des Working Profiles, gemittelt über alle Probanden

Obwohl in dieser Studie bereits die Zeitstempel zu den Interaktionen erfasst wurden, mussten die Zeiten pro User Story in einer Excel Tabelle mit den erfassten Daten noch händisch berechnet werden (Endzeitpunkt minus Startzeitpunkt). Hierbei ist es aus Analysesicht wünschenswert, in Zukunft den Timer beim Eintreten in eine neue User Story immer auf null zurückzusetzen und bis zum Austritt aus der User Story hochzuzählen, sodass die „time on task“ sofort ablesbar ist. Weiterhin wurde erkannt, dass User Stories, die eine Abhängigkeit voneinander haben noch nicht ausreichend ausgewertet werden können (z.B. Endpunkt von User Story 1 ist Startpunkt von User Story 2). Eine solche Transition zwischen User Stories konnte mit der aktuellen Tracking Library technisch noch nicht erfasst werden und sollte für folgende Studien definiert und eingebaut werden.

Interessant war bei der Datenauswertung darüber hinaus, dass in den Daten erkennbar war, wenn in User Stories eingetreten, diese aber nicht beendet wurden („destructive exit“). An den entsprechenden Stellen konnten im Kontext der weiteren Interaktionen erste Vermutungen darüber aufgestellt werden, warum der Nutzer die User Story betreten, aber nicht beendet hat (sucht z.B. eine Funktion, die er nicht findet und klickt sich dabei durch alle Bereiche).

Über die Auswertung der User Stories anhand der „time on task“ hinaus, konnten in den Daten weitere Muster erkannt werden, die auf Fehler in der Interaktion hindeuteten. Beispielsweise wurde ein Muster erkannt, bei dem die Teilnehmer nach dem Speichern der eingetragenen Zeiten erneut oder mehrfach den Speicherbutton betätigten. Hier ist die Vermutung, dass das Feedback auf das Speichern unzureichend bzw. die Erwartung der Nutzer eine andere war, was nach dem Speichern passiert („Lande auf der Übersicht“). Eine weitere Erkenntnis war, dass die Teilnehmer insgesamt 117-mal auf nicht klickbare Elemente des Interfaces drückten. Hierbei scheint der Aufforderungscharakter der Elemente die Nutzer irrezuführen, sodass sie hinter Bildern Funktionen vermuteten. Ähnliche „Tap“-Muster könnten in Zukunft direkt vom Agenten erkannt und als Fehler klassifiziert werden.

4 Diskussion

Die oben beschriebene Auswertung der automatisierten Daten deutet darauf hin, dass die Forschungsfrage hinsichtlich der Ableitung von Verbesserungsempfehlungen aus automatisierten Daten bejaht werden kann. Es konnten aus den Daten bereits Handlungsempfehlungen abgeleitet werden, ohne dass der Produktentwickler bei der Interaktion anwesend sein musste. Die Erhebung von großen Mengen von Nutzungsdaten unter dem Schirm sinnvoller Fragestellungen gibt dem Entwicklungsteam Hinweise drauf, wo optimiert werden sollte. Über die User Story-basierte Gruppierung hinaus, zeigten sich weitere Muster in den Daten, die auf potenzielle Optimierungspunkte hinweisen.

Trotz aller positiven Ergebnisse haben sich einige Einschränkungen ergeben. Diese bezogen sich vor allem auf Informationen zu Nutzergruppen und Nutzungskontext, die ohne den direkten Nutzerkontakt nur schwer zu erfassen sind (z.B.

durch Fragebögen möglich). Sind Zielgruppe und Nutzungskontext wenig bekannt, kann es daher notwendig sein, die automatisierte Datenanalyse mit klassischen Methoden aus dem UX Research zu kombinieren.

Zudem konnten einige Metriken in der Laborstudie nicht gemessen werden, da die zugehörigen KPIs nur in einem realen Nutzungsszenario und über einen längeren Zeitraum hinweg erfasst werden können. Solche Feld-abhängigen User Stories sollen in einer weiteren Studie untersucht werden.

Auch sind die Interpretationen der in den Daten beobachteten Muster Annahmen. Der Datenanalytiker sieht die Teilnehmer nicht, sodass er nur annehmen kann, warum bestimmte Muster und Probleme auftreten. Benötigen einige Teilnehmer mehr Zeit, mehr Klicks oder nutzt die App anders als andere Benutzer, sind mehrere Interpretationen denkbar: Sie werden durch etwas in ihrer Umgebung abgelenkt, sie sind nicht mit den verwendeten Gesten oder Interaktionsmustern vertraut, sie werden durch Benachrichtigungen von anderen Anwendungen abgelenkt, sie suchen nach fehlenden Interaktionselementen, sie werden verwirrt durch Elemente, die sie nicht erwartet haben, sie verstehen Teile der Oberfläche nicht, etc. Wie wir sehen können, gibt es für solche Probleme viele Interpretationsmöglichkeiten, und wir können nur sehr wenige davon durch die Analyse der Daten ausschließen. Dennoch kann ein Analyst mit Hilfe des Tools erste Hinweise zur Optimierung geben und der Product Owner kann dahingehend sein Backlog priorisieren. Dies verschafft ihm einen Wettbewerbsvorteil in der agilen Welt der Anwendungsentwicklung, auch wenn er die Ursache für die Fehler vielleicht noch nicht kennt. Im Falle schwerwiegender Probleme können Benutzertests vor Ort durchgeführt werden, um diese Probleme genauer zu untersuchen.

Die Studie wurde mit zehn Teilnehmern durchgeführt. Dies ist eine angemessene Stichprobengröße für qualitative UX-Studien [8], aber eine Bedrohung für die Validität im Hinblick auf die Generalisierung. Wenn quantitative Daten analysiert werden sollen, ist eine größere Datenmenge erforderlich, um den Ansatz zu bewerten. In der Vorstudie und in der Hauptstudie konnten wir sehen, dass auch in diesen sehr ähnlichen Testsituationen die Ergebnisse zu den Umsetzungen der User Stories hinsichtlich ihrer Qualitätsbewertung sehr unterschiedlich waren. Darüber hinaus handelte es sich um eine Labortestsituation, bei der der Ansatz in einem realen Nutzungskontext angewendet werden sollte. Zu diesem Zweck wäre es sinnvoll, einen Feldtest mit dieser oder einer anderen App durchzuführen, um den Prozess mit realen Daten zu bewerten und einige der KPIs zu sammeln, die wir aus der Laborstudie nicht erhalten konnten (z.B. Nutzungszeit, Nutzungshäufigkeit).

Wie gerade beschrieben, wurde das Tool bisher von UX-Analysten, Experten für Statistik und Daten, verwendet. Eines der Ziele dieses Forschungsprojektes war es jedoch, ein zugängliches Werkzeug zu entwickeln, das jeder Product Owner selbst nutzen kann. Um dieses Ziel zu erreichen, war es notwendig, eine Studie mit Product Ownern durchzuführen, um die Verständlichkeit und den Nutzen der aufbereiteten Daten für ihre tägliche Arbeit zu bewerten. Zu diesem Zweck müssen auch die Daten besser zugänglich sein. Hierbei wurden zwei Product Ownern verschiedene Darstellungsformen der Daten präsentiert. Sie wurden dazu aufgefordert, die Daten zu interpretieren und die Visualisierungen hinsichtlich ihrer Nützlichkeit zu sortieren. Es konnte festgestellt werden, dass eine Kombination aus Informationen zur User Journey und Hinweise auf Probleme bei der Durchführung der einzelnen User Stories als besonders hilfreich angesehen wurde.

5 Fazit und Ausblick

Im Forschungsprojekt Opti4Apps wurde erneut die Notwendigkeit erkannt, leichtgewichtige Feedbackanalysemethoden zu entwickeln, um nutzerzentriert in agilen Projekten arbeiten zu können. Mit Hilfe des entwickelten Datenauswertungsansatzes, wird Product Ownern eine Methode an die Hand gegeben, mit der sie einfach und schnell ihre Backlog-Items priorisieren können und gleichzeitig etwas über die Qualität ihrer Anwendung und über die Akzeptanz bei den Nutzern lernen. Die „Definition of Done Done Done“ hat für Product Owner also den entscheidenden Vorteil, dass sie nicht nur *glauben* müssen, dass die Umsetzung einer User Story ausreichend reif beim Nutzer aufschlägt, sondern sich dies auch tatsächlich nachweislich im Feld bestätigen lässt. Nach der oben beschriebenen „Definition of Done“ und „Definition of Done Done“, mit denen das Team nach einem Sprint die Auslieferbarkeit des Produktes bestätigt, sollte anhand der „Definition of Done Done Done“ festgelegt werden, welche Kriterien das Produkt im Feld erfüllen muss, damit es wirklich als fertig, also beim Nutzer akzeptiert gilt. Diese könnte zum Beispiel heißen „Es wurden Nutzungsdaten erhoben und analysiert.“ Weitere Akzeptanzkriterien definieren dann pro User Story anhand konkreter Metriken und KPIs, ob jede einzelne User Story performt oder nicht. Diese Informationen können Entwicklungsteams in Zukunft größtenteils automatisiert erhalten und somit das Nutzerfeedback in ihren zeitkritischen Prozess integrieren.

Natürlich ist vor allem die automatisierte Auswertung der Daten noch in einem frühen Stadium der Entwicklung. Da sich das Projekt vor allem auf die Evaluation der Methode konzentrierte, sollte in Zukunft weiter an der Automatisierung der Datenanalyse und Fehlererkennung gearbeitet werden.

Ein weiterer Punkt ist, die Auswertung von textuellem und automatisch aufgezeichnetem Feedback zu verbinden. Bereits im Projekt beschäftigten sich die Projektpartner mit der Erfassung von textuellem Feedback, z.B. in App-Stores sowie mit der Bedeutung von Emojis im Rahmen von Bewertungen. Offen ist hierbei die Frage, inwieweit sich beide Feedbackarten ergänzen und wie eine ganzheitliche Visualisierung erreicht werden kann, da gerade das textuelle Feedback interessante Informationen liefert, die bereits in der digitalen Welt vorhanden sind.

Der nächste Schritt ist nun, möglichst unabhängig von der eingesetzten Frontend-Technologie individuelle Integrationsmöglichkeiten für die User Story basierte Nutzungsdatenaufzeichnung und -auswertung zu schaffen. Damit wird das Thema außerdem aus dem App Kontext auf weitere Anwendungsbereiche erweitert.

ACKNOWLEDGMENTS

Die in diesem Beitrag beschriebene Forschung wurde im Rahmen des Projekts Opti4Apps (Förderkennzeichen 02K14A182) des Bundesministeriums für Bildung und Forschung (BMBF) durchgeführt.

REFERENCES

[1] State of Agile, 2018

[2] M. Levison, Definition of Done vs. User Stories vs. Acceptance Criteria, https://agilepainrelief.com/notesfromatooluser/2017/05/definition-of-done-vs-user-stories-vs-acceptance-criteria.html#.XSRysOgzbmE

[3] *J. Shore and S. Warden, The Art of Agile Development*, O'Reilly, 2008.

[4] T. Immich, „Continuous UX - Lean und Large unter einem Dach," in Tagungsband der Mensch & Computer 2018, Dresden, 2018.

[5] J. Sauro, 10 Thing to know about task times, https://measuringu.com/task-times/

[6] T. Tullis and B. Albert, Measuring the user Experience, 2nd ed., Morgan Kaufmann, 2013, pp. 74-82.

[7] J. Sauro and J. R. Lewis, Quantifying the user experience, 2nd ed., Morgan Kaufmann, 2016, pp.30-33.

[8] T. Tullis and B. Albert, Measuring the user Experience, 2nd ed., Morgan Kaufmann, 2013, p. 14.