Subjektive Benutzerzufriedenheit quantitativ erfassen: Erfahrungen mit dem User Experience Questionnaire UEQ

Bettina Laugwitz
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
bettina.laugwitz@sap.com

Ulf Schubert
DATEV eG
Südliche Fürther Str. 18-20
90329 Nürnberg
ulf.schubert@datev.de

Waltraud Ilmberger
Psychologisches Institut
Universität Heidelberg
Hauptstr. 47-51
69117 Heidelberg
Waltraud.Ilmberger@gmx.de

Nina Tamm
Psychologisches Institut
Universität Heidelberg
Hauptstr. 47-51
69117 Heidelberg
nina.tamm@yahoo.de

Theo Held
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
theo.held@sap.com

Martin Schrepp
SAP AG
Dietmar-Hopp-Allee 16
69190 Walldorf
martin.schrepp@sap.com

Abstract

Das Gebrauchstauglichkeitskriterium der Benutzerzufriedenstellung kann auf verschiedene Weise quantifiziert werden. Eine Möglichkeit sind Fragebögen, die auf eine spontane und möglichst unreflektierte Beurteilung eines Produkts abzielen können. Ein Beispiel für einen solchen Fragebogen ist der User Experience Questionnaire UEQ (Laugwitz et al. 2006). Berichtet wird über die Verwendung des Fragebogens in der Qualitätssicherung bei der Entwicklung von Business-Software bei DATEV. Zum anderen werden zwei wissenschaftliche Studien vorgestellt, in denen der UEQ ebenfalls eingesetzt wurde. Aus den dargestellten Erfahrungen werden zusammenfassend Vorzüge, Einschränkungen und Empfehlungen für den erfolgreichen Einsatz von Fragebögen in Usability-Praxis und -Forschung abgeleitet.

Keywords

User Experience, Fragebogen, Quantitative Methoden, Benutzerzufriedenstellung, Qualitätssicherung

1.0 Einleitung

Für die Quantifizierung des Kriteriums der Benutzerzufriedenstellung stehen verschiedene Methoden zur Verfügung. Die Messung erfolgt auf unterschiedlichen Ebenen, mit verschiedenen Methoden und mit unterschiedlicher Granularität. Sehr einfach einzusetzen sind Fragebögen. Auch hier gibt es eine große Variation möglicher Methoden, die ein breites Spektrum an Benutzerreaktionen einbeziehen. Fragebögen können beispielsweise einer umfassenden und detaillierten Erhebung spezifischer Gebrauchtauglichkeitsprobleme dienen, wie z. B. der IsoMetrics (Gediga & Hamborg 1999).

Der User Experience Questionnaire UEQ (Laugwitz et al. 2006; Laugwitz et al. 2008) zielt im Gegensatz dazu auf eine schnelle, unmittelbare und möglichst unreflektierte Beurteilung eines Produktes ab. Eine solche Beurteilung sollte nicht isoliert betrachtet werden, sondern muss sich in ein Untersuchungs- und Methodenkonzept einfügen, das der jeweiligen Fragestellung angepasst ist.

Im Folgenden werden verschiedene Einsatz-Szenarien des UEQ aus Praxis und Wissenschaft beschrieben und die Stärken und Schwächen in den jeweiligen Situationen ausgelotet. Zuvor wird der Fragebogen vorgestellt.

2.0 User Experience Questionnaire: Form, Zielsetzung und Entstehung

Ziel des User Experience Questionnaire UEQ ist eine effiziente Messung des Gesamteindrucks, den ein Benutzer in Bezug auf die Interaktion mit einem interaktiven Produkt entwickelt hat. Der UEQ besteht aus 26 bipolaren Items, die die Form eines 7-stufigen semantischen Differentials haben, z.B.:

kompliziert □ □ □ □ □ □ □ *einfach*. Die Items sind den folgenden sechs Skalen zugeordnet: *Effektivität, Durchschaubarkeit, Vorhersagbarkeit* (Benutzungsqualität; jeweils 4 Items), *Stimulation, Originalität* (Designqualität; jeweils 4 Items), *Attraktivität* (6 Items).

Zusätzlich zur deutschen Originalversion des UEQ ist auch eine englische Version verfügbar.

In mehreren Validierungs-Studien zur deutschen und englischen Version konnte gezeigt werden, dass die Skalen des UEQ eine hohe Reliabilität aufweisen. Eine Ausnahme ist hier die Skala Vorhersagbarkeit, für die in einigen Studien keine völlig zufriedenstellende Reliabilität (d.h. ein Wert < 0.7 für Cronbachs Alpha-Koeffizient) gefunden wurde. Weiterhin deuten die bisher durchgeführten Validierungs-Studien auf eine zufriedenstellende Konstruktvalidität hin.

Seit 2009 ist eine verkürzte Version des UEQ vorhanden (Short UEQ, S-UEQ). Diese Version wurde aus der Originalversion durch Eliminieren der vier Items

der Skala Vorhersagbarkeit sowie der beiden schwächsten Items der Skala Attraktivität erzeugt und enthält nur noch 20 Items.

3.0 Anwendung in der Praxis: Entwicklung von Businesssoftware

Bei der DATEV wird der UEQ im Rahmen des Entwicklungsprozesses für die Bewertung der Qualität von Software-Bedienoberflächen eingesetzt. Da die DATEV ein EDV-Dienstleister für Steuerberater, Wirtschaftsprüfer und Rechtsanwälte ist, handelt es sich in erster Linie um betriebswirtschaftliche und steuerliche Business-Software.

Ziel ist es zu erkennen, inwieweit die Bedienoberflächen der zahlreichen Software-Anwendungen die hohen Qualitätsziele hinsichtlich der verschiedenen in der Einleitung genannten Aspekte der User Experience erfüllen. Der UEQ spielt neben den technischen Evaluations- und Testmethoden eine wichtige Rolle in der Qualitätssicherung bei DATEV.

3.1 Einsatzszenarien

Der UEQ wird aktuell erfolgreich in zwei Einsatzszenarien eingesetzt.

Zum einen werden in regelmäßigen Abständen DATEV Software-Anwendungen gemessen, die bereits im Markt verfügbar sind. Es werden Anwender befragt, die erfahren im Umgang mit der betreffenden Anwendung sind. Die Stichprobe ist bei diesen Messungen in der Regel größer als n=30. Auf diese Weise wird gemessen, inwieweit sich Neuerungen und Änderungen in der Anwendung auf die wahrgenommene Qualität ausgewirkt haben.

Zum anderen wird der UEQ in Benutzerlaboren bzw. Usability Tests zur Bewertung von Gestaltungslösungen verwendet. Dabei geht es in erster Linie nicht darum eine möglichst genaue Bewertung der User Experience zu erheben. Die UEQ-Messung wird am Ende des Benutzerlabors durchgeführt und soll dem Entscheider eine Orientierung geben, ob die Gestaltungslösung eine deutliche Verbesserung bringen wird. Die UEQ-Ergebnisse werden dazu mit dem DATEV Benchmark und vorherigen Messungen zu der jeweiligen Anwendung vergleichend analysiert. Der DATEV Benchmark enthält die durchschnittlichen Skalenwerte von vergleichbaren Anwendungsgruppen.

Die Erfahrung hat gezeigt, dass die auf diese Weise im Benutzerlabor erhobene Bewertung ein guter Indikator ist und bei der Bewertung im Markt im Wesentlichen bestätigt wird. Mit Einschränkungen in der Übertragbarkeit der UEQ-Ergebnisse muss allerdings gerechnet werden, wenn die Neuerungen bzw. Änderungen nicht in den Hauptnutzungsszenarien der Anwendung vorgenommen wurden. In diesem Fall hat die Veränderung nur wenig Einfluss auf die Gesamtbewertung und ist mit dem UEQ schwer erfassbar.

Weiterhin wurde der UEQ testweise für die Messung von Einarbeitungsphasen verwendet. Hintergrund war die Frage, wie sich im Laufe der Einarbeitung die Bewertung der wahrgenommenen Qualität ändert und wann ein stabiler Zustand in der Bewertung erreicht wird. Es wurde angenommen, dass eine stabile Bewertung ein Indikator dafür ist, dass die Einarbeitungszeit abgeschlossen ist. Dazu wurde der UEQ nach Auslieferung einer neuen Anwendung im wöchentlichen Rhythmus zur Bewertung an eine gleichbleibende Stichprobe gesendet. Um zu vermeiden, dass die Befragten ihre Bewertungen quasi aus den Vorwochen abschreiben wurde die Reihenfolge der Eigenschaftspaare randomisiert.

3.2 Akzeptanz bei Anwendern

Insgesamt kann man sagen, dass die Akzeptanz des UEQ bei den Anwendern in den ersten beiden Einsatzszenarien (Kontinuierliche Messung und Benutzerlabor) positiv ist.

Bei der Messung der Einarbeitungsphase war die Akzeptanz seitens der Befragten leider so gering, dass keine verlässlichen Ergebnisse erhoben werden konnten. Die Gründe dafür waren zum einen der Zeitaufwand der häufigen Bewertungen und das kurze Zeitintervall. Um Zeit zu sparen hatten sich nach zwei bis drei Wochen viele Anwender den ausgefüllten Fragebogen ausgedruckt und schrieben die Werte dann trotz Randomisierung ab. Im Ergebnis war somit zwar eine frühzeitige aber trügerische Stabilisierung der Bewertungen zu beobachten. Da der UEQ begleitend zu einer Feldstudie eingesetzt wurde, konnte dieser Umstand schnell identifiziert und die Messung abgebrochen werden.

Um eine hohe Akzeptanz des UEQ zu erreichen, sollten nach unseren Erfahrungen folgende Faktoren berücksichtigt werden:

- Hintergrund und Nutzen der Methode sollten für die Befragten klar erkennbar sein.
- Beim Ausfüllen sollte den Befragten ein persönlicher Ansprechpartner zur Verfügung stehen.
- Die zeitlichen Abstände zwischen den einzelnen Messungen sollten ausreichend lang bemessen werden.

Anfänglich bestand bei den Befragten trotz guter Erläuterung auf dem Deckblatt des Fragebogens Skepsis über den Hintergrund und Nutzen der Methode. Für viele war es unverständlich, wie über ein Kreuz zwischen zwei Eigenschaften eine nützliche Aussage über die Qualität einer Bedienoberfläche zustande kommt.

Da diese Skepsis einen negativen Einfluss auf die Rücklaufquote hatte, wurde das Vorgehen zur Verteilung des Fragebogens geändert. Zu Beginn der Messungen wurden diese mittels einer Online-Befragung durchgeführt. Hier bestanden nur wenige Möglichkeiten Informationen über Hintergrund und Nutzen anschaulich zu vermitteln. Daher wurde das Vorgehen dahingehend geändert, dass die Fragebögen in ausgedruckter Form in einem persönlichen Gespräch vorgelegt wurden. Dieses Vorgehen hat sich bewährt und wird so fortgesetzt. Da die Vorlage im persönlichen Gespräch bei einer großen Stichprobe mit einem entsprechenden Zeitaufwand verbunden ist, werden wir in einer der nächsten Befragungswellen auf eine Online-Variante mit telefonischer Unterstützung verproben.

Eine gute Erläuterung sowie die Möglichkeit für Rückfragen sind auch dahingehend wichtig, da einige Eigenschaftspaare des UEQ, z.B. „langweilig-spannend“, im Zusammenhang mit Business-Software mit Befremden aufgenommen werden.

Aufgrund der Erfahrungen bei der Messung der Einarbeitungsphase wurden die Abstände zwischen den einzelnen Messungen auf einen halbjährigen Rhythmus umgestellt.

3.3 Akzeptanz bei Führungskräften und Produktverantwortlichen

Neben der Akzeptanz bei den Anwendern ist für einen erfolgreichen Einsatz des UEQ eine hohe Akzeptanz bei Führungskräften und Produktverantwortlichen notwendig.

Damit die UEQ-Ergebnisse als Grundlage für Designentscheidungen angewendet werden, wurde zu Beginn großen Wert darauf gelegt, die Zuverlässigkeit der UEQ-Ergebnisse zu belegen. Der Verweis auf wissenschaftliche Veröffentlichungen und Evaluierungen war dabei hilfreich. Zusätzlich wurden die Ergebnisse aus der ersten Messung im Rahmen einer Pilotierung des Verfahrens mit subjektivem Anwenderfeedback in Beziehung gesetzt, welches über andere Kanäle, z.B. eMail, eingegangen war. Im weiteren Verlauf wurde der Zusammenhang zwischen den in der Anwendung vorgenommenen Änderungen und der Veränderung der UEQ-Bewertungen aufgezeigt.

Um die Aussagekraft der Skalenwerte zu verstärken, werden diese immer im Vergleich zu den Skalenwerten aus vorherigen Messungen bzw. zu dem DATEV Benchmark dargestellt.

Es war weiterhin erfolgsentscheidend die Führungskräfte bei der Interpretation der UEQ-Auswertungen zu unterstützten. Insbesondere bei der Standard-Auswertungsgrafik, die das Ergebnis der Auswertung des UEQ mit Excel ist, bestanden anfänglich Unklarheiten darüber, wie die einzelnen Zahlenwerte zu interpretieren sind, d.h. was beispielsweise ein guter Wert ist. Daher wurde die Standard-Auswertungsgrafik gestalterisch etwas überarbeitet und die einzelnen Skalenwerte den firmeninternen Sprachgewohnheiten angepasst.

Zusätzlich wird zu den UEQ-Auswertungen immer ein Interpretationsvorschlag geliefert, welcher durch erfahrene User Experience Experten erstellt wird. Dieser Interpretationsvorschlag bildet die Grundlage für die Diskussion von eventuell notwendigen Maßnahmen.

3.4 Erweiterung für betriebswirtschaftliche und steuerliche Software

Bei betriebswirtschaftlicher bzw. steuerlicher Software wird die wahrgenommene Qualität neben den Skalen Effektivität, Durchschaubarkeit, Attraktivität, usw. auch maßgeblich durch die Aktualität der Anwendung bestimmt.

Damit ist beispielsweise gemeint, inwieweit eine Anwendung aktuelle Gesetzesänderungen berücksichtigt.

Daher wurde der UEQ bei DATEV um eine Skala für Aktualität ergänzt. Da für diese Skala noch keine evaluierten Eigenschaftspaare vorliegen, wird diese Skala gesondert von den UEQ-Eigenschaftspaaren abgefragt. Für die Weiterentwicklung des UEQ ist es aus DATEV-Sicht wünschenswert, dass er um diese Skala erweitert wird.

3.5 Ausblick: UEQ als Bestandteil der Business Intelligence

Aus DATEV-Sicht liefert der UEQ verlässliche und relativ leicht interpretierbare Zahlenwerte. Das Verfahren ist daher gut geeignet, Kennzahlen zur Qualität einer Software-Anwendung zu erheben und zu kommunizieren.

Produktverantwortliche und Führungskräfte können auf Basis der regelmäßig erhobenen Kennzahlen bzw. Skalenwerte schnell Handlungsbedarf erkennen.

Um diesen Mehrwert zukünftig noch stärker zu nutzen, wird bei der DATEV aktuell ein Konzept erarbeitet, wie die UEQ-Ergebnisse stärker in die Business Intelligence einfließen können.

4.0 Wissenschaftliche Anwendung: Zwei Beispiele

Auch für die Hypothesentestung in der wissenschaftlichen Arbeit können Fragebögen eingesetzt werden.

4.1 Beispiel 1: Eine interkulturelle Online-Studie

Die westlich geprägte Methode des Usability-Tests wird aufgrund des rapide wachsenden Informationstechnologiemarkts verstärkt in Asien eingesetzt. Einige Forschungsarbeiten weisen allerdings darauf hin, dass die etablierten

westlichen Usability-Test Methoden weniger effektiv sind, wenn sie mit Teilnehmern aus anderen Kulturen durchgeführt werden (z.B. Oyugi et al. 2008).

Damit Gebrauchstauglichkeitsmängel frühzeitig identifiziert werden können, ist es wichtig, dass negative Kritik seitens der Probanden direkt geäußert wird. Somit ist das Kommunikations- und Kritikverhalten der Probanden ausschlaggebend für die Aussagekraft eines Usability-Tests. Daher wurde die Frage untersucht, ob sich asiatische und westliche Probanden in ihrem Kommunikations- und Kritikverhalten im Rahmen von Usability-Tests unterscheiden.

Basierend auf Theorien, die interkulturelle Unterschiede im Kommunikationsverhalten (Hall 1976; Markus & Kitayama 1991) und Höflichkeitsniveau (Brown & Levinson 1987) postulieren, wurde die Hypothese aufgestellt, dass asiatische Probanden ihre Kritik an Mängeln indirekter ausdrücken als westliche Probanden (siehe Tamm, 2009). Dies sollte sich anhand einer positiveren Bewertung des untersuchten Prototyps seitens asiatischer Probanden zeigen.

Die Studie wurde als Onlinestudie durchgeführt, in der asiatische (N = 185) und westliche Probanden (N = 140) in Interaktion mit dem Prototyp eines Einkaufsportals eine Aufgabe lösen und den mit Usability-Mängeln behafteten Prototyp im Anschluss bewerten sollten. Das Kritikverhalten wurde anhand des Post-Test-Survey der SAP AG und des Short UEQ (S-UEQ, s. o.), der durch seine Skalen eine feine Differenzierung der Bewertung verschiedener Facetten der User Experience ermöglicht, erhoben. Durch die zusätzliche Messung der Bearbeitungszeit und die Bestimmung der Abbruchquote sollte überprüft werden, ob asiatische und westliche Probanden vergleichbare Schwierigkeiten in der Interaktion mit dem Prototyp erleben.

Die Cronbach's Alpha-Werte der Skalen des S-UEQ lagen für Durchschaubarkeit bei .85, für Effizienz bei .78, für Stimulation bei .88, für Originalität bei .81 und für Attraktivität bei .91.

Die Ergebnisse zeigen, dass asiatische Probanden den Prototyp anhand aller Skalen tendenziell positiver einstuften als westliche Probanden, obwohl sie signifikant mehr Zeit benötigten, um die Aufgabe zu bearbeiten und eine höhere Abbruchquote aufwiesen. Die Mittelwertsunterschiede für die Bewertung der Attraktivität ($t(323) = -2.19$, $p < .05$), Stimulation ($t(318;3) = -2.17$, $p < .05$) und Originalität ($t(323) = -2.75$, $p < .01$) waren signifikant, während die Unterschiede in den Bewertungen der Benutzungsqualität (Effizienz, Durchschaubarkeit) keine Signifikanzen erreichten.

Die Bewertungsrangfolge der unterschiedlichen Kriterien war für asiatische und westliche Probanden identisch. So bewerteten alle Probanden die Durchschaubarkeit des Prototyps am höchsten, während die Originalität am niedrigsten eingestuft wurde.

Die Skalen des S-UEQ korrelierten hoch mit der Gesamt-Usability-Bewertung anhand der Likertskalen des Post-Test-Survey der SAP AG ($r = .47$ bis $r = .75$).

Insgesamt weisen die Ergebnisse auf kulturelle Unterschiede im Kritikverhalten hin, die aber möglicherweise aufgrund der anonymen Durchführung als Onlinestudie, in der jegliche soziale Interaktion zwischen Testmoderator und Teilnehmer fehlt, gering sind.

Der S-UEQ diente einer differenzierteren Betrachtung kultureller Unterschiede im Bewertungsverhalten als die Likertskalen des Post-Test-Survey und konnte somit aufdecken, dass die Unterschiede in den Bewertungen lediglich für Attraktivität und die Skalen der Designqualität, nicht jedoch für die Skalen der Benutzungsqualität statistisch signifikant wurden. Darüber hinaus ermöglichte er durch seinen geringen Zeitaufwand einen schnellen Ablauf der Onlinestudie und minimierte somit die Gefahr von Abbrüchen seitens der Probanden.

4.2 Beispiel 2: Ein Laborexperiment zu Schönheit und Usability

Auf Grundlage von Studien, die einen engen Zusammenhang zwischen wahrgenommener Schönheit und wahrgenommener Gebrauchstauglichkeit annehmen lassen (Kurosu & Kashimura 1995, Tractinsky 1997), sollte dieser überraschend gefundene Zusammenhang näher beleuchtet werden (siehe Ilmberger 2008). Hierzu wurden zwei Theorien näher untersucht.

Gemäß dem Halo Effekt (Dion et al. 1972) würde man die hohen Korrelationen bisheriger Studien auf die nicht-interaktive Präsentation des Untersuchungsmaterials zurückführen. Demnach hätten die Teilnehmer die Usability-Bewertung aufgrund mangelnder Informationen auf Basis des salienten Attraktivitätsurteils vorgenommen, was dann wiederum zu den gefundenen hohen Korrelation führen würde. Ausgehend von der Halo-Theorie sollten sich die Usability-Bewertungen verschiedener Usability-Niveaus vor der Interaktion mit einem Produkt also nicht unterscheiden. Nach der Interaktion mit dem Produkt sollten die Teilnehmer dagegen ein differenzierteres Bild der Usability haben.
Außerdem wurde noch einem Erklärungsansatz von Norman (2003) nachgegangen, der die Stimmung des Nutzers als vermittelnde Variable vorschlägt. Demnach würde eine ästhetische Benutzungsschnittstelle zu einer positiven Stimmung des Nutzers führen und diese zu einer kreativeren, lösungsbezogenen Denkweise. Der Nutzer soll-

te demnach die Usability von ästhetischen Benutzungsschnittstellen aufgrund einer besseren Problemlösestrategie besser bewerten. Ein Indiz für den Ansatz wären geringe Korrelationen zwischen Usability- und Ästhetikurteilen vor und hohe Korrelationen nach der Interaktion mit dem System.

Zur Untersuchung der beiden Ansätze wurden vier verschiedene Versionen eines Onlineshops erstellt, die bezüglich Gebrauchstauglichkeit (gut vs. schlecht) und Farbgestaltung (ästhetisch ansprechend vs. ästhetisch nicht ansprechend) manipuliert wurden.
Betrachtet wurden zwei Messzeitpunkte. Zuerst wurde die Einschätzung der Shops mithilfe des UEQ nach der kurzen Präsentation eines Demonstrationsvideos erfragt. Dann sollten die Teilnehmer fünf charakteristische Aufgaben durchführen (z.B. einen Artikel suchen und ihn in den Warenkorb legen). Anschließend wurde ihnen nochmals der UEQ vorgelegt und sie sollten zusätzlich Probleme bei der Nutzung des Shops nennen. Zu beiden Messzeitpunkten schätzten die Teilnehmer die visuelle Ästhetik des bearbeiteten Shops anhand zweier zusätzlicher Items ein (z.B. Wie findest du die visuelle Gestaltung des Shops insgesamt?)

Es nahmen 72 Psychologiestudenten der Universität Heidelberg an dem Experiment im Labor teil. Die Bearbeitung dauerte ca. 45 Minuten.

Die Ergebnisse zeigten, dass sich die zwei Usability-Bedingungen nach einer sehr kurzen Demonstration - wie gemäß dem Halo-Effekt angenommen - nicht in Bezug auf deren Benutzungsqualität unterscheiden ließen (Effizienz $t(70) = 1,08$; $p > .05$, Vorhersagbarkeit $t(70) = 0,49$; $p > .05$, Durchschaubarkeit $t(70) = -0,64$; $p > .05$). Erst nach der Interaktionsphase zeigten sich signifikante Unterschiede in der erwarteten Richtung auf allen drei Skalen zur Benutzungsqualität (Effizienz $t(70) = 4,24$; $p < .01$, Vorhersagbarkeit $t(70) = 4,35$; $p < .01$ und Durchschaubarkeit $t(70) = 4,65$; $p < .01$). Besonders hervorzuheben ist hier, dass sich sehr deutliche Unterschiede zeigten, obwohl jeder Teilnehmer nur eine Version des Shops bearbeitete.

Eines der Hauptergebnisse bestand darin, dass die in vorhergehenden Studien gezeigte Höhe der Korrelationen zwischen Ästhetik und Usability-Urteilen nicht repliziert werden konnte. Darüber hinaus konnte der HALO-Ansatz nicht bestätigt werden, da - entgegen der Hypothese - zum zweiten Messzeitpunkt höhere Korrelationen zwischen wahrgenommener Usability und Ästhetik bestanden als zum ersten Messzeitpunkt.

Weiterführende Auswertungen zeigen eher in Richtung eines „what is usable is beautiful“-Zusammenhangs anstatt der von Norman (2003) vorgeschlagenen „what is beautiful is usable“-Richtung.

Der UEQ hat sich in der beschriebenen Studie für die differenzierte Erhebung von subjektiv wahrgenommenen Usability-Bewertungen als sehr nützlich erwiesen. Des Weiteren zeigten sich signifikant negative Korrelationen mit der Anzahl an berichteten Problemen und der Bearbeitungszeit für alle drei Skalen zur Benutzungsqualität (s. Tabelle 1).

	Effizienz	Vorhersagbarkeit	Durchschaubarkeit
Bearbeitungszeit	-0,42*	-0,43*	-0,49*
Anzahl berichteter Probleme	-0,45*	-0,40*	-0,41*

* $p < .01$

Tabelle 1: Korrelationen von UEQ-Skalen mit Bearbeitungszeit und Anzahl Probleme

Insgesamt wurde der UEQ von den Versuchsteilnehmern aufgrund der kurzen Bearbeitungszeit und der Prägnanz der Items positiv bewertet. Nur die Anzahl der Items und die Übersichtlichkeit der Darstellung wurden vereinzelt kritisiert.

5.0 Zusammenfassung

Fragebögen wie der UEQ sind in verschiedenen Untersuchungskontexten einsetzbar. Berichtet wurde über Erfahrungen mit dem UEQ in der Qualitätssicherung bei DATEV sowie in zwei wissenschaftlichen Untersuchungen.

Beide Kontexte ergaben Hinweise auf eine zufrieden stellende Validität, die sich in einem systematischen Zusammenhang zwischen den Werten auf den UEQ-Skalen und konzeptionell verwandten externen Kriterien zeigt.

Die Daten der berichteten interkulturellen Untersuchung deuten auf eine zufrieden stellende Reliabilität der Short-UEQ-Skalen hin. Diese muss allerdings noch in weiteren Untersuchungen nachgewiesen werden.

Quantitative Maße sind nur sinnvoll interpretierbar, wenn sie zueinander in Beziehung gesetzt werden. Das kann über einen Vergleich mit Benchmarks geschehen oder durch die Gegenüberstellung der Werte von verschiedenen Produktversionen oder Experimentalbedingungen. Die Standardisierung von Skalen, wie sie beispielsweise für den SUMI vorliegt (Kirakowski & Corbett 1993), ist prinzipiell erstrebenswert, weil sie eine Interpretation der Werte auch ohne konkrete Vergleichsbasis erlaubt, erfordert aber das Vorliegen einer immensen Datenmenge, die für frei verfügbare Fragebögen kaum erreichbar ist.

Fragebögen wie der UEQ und der S-UEQ können im persönlichen Kontakt in Labor und Feld verwendet werden, aber auch online, wenn es die Untersuchungssituation erfordert. Vor dem Einsatz in Online-Kontexten muss aber

sorgfältig abgewogen werden, ob mit systematischen Ausfällen oder Manipulationen zu rechnen ist und wie problematisch derartige Verzerrungen für das Untersuchungsziel sein können. Nach Möglichkeit ist die Untersuchung mit persönlichem Kontakt vorzuziehen. Dabei sollte darauf geachtet werden, dass die Teilnehmer sich der Anonymität ihrer Beurteilungen möglichst sicher sein können.

6.0 Literatur

Brown, P.; Levinson, S. C. (1987): Politeness: Some universals in language usage. New York: Cambridge University Press.

Dion, K. K.; Berscheid, E.; Walster, E. (1972): What is beautiful is good. Journal of Personality and Social Psychology, Vol. 24, S. 285–290.

Gediga, G.; Hamborg, K.-C. (1999): IsoMetrics: Ein Verfahren zur Evaluation von Software nach ISO 9241-10. In: Holling, H.; Gediga, G. (Hrsg.): Evaluationsforschung. Göttingen: Hogrefe, S. 195 - 234.

Hall, E. T. (1976): Beyond culture. Oxford: Anchor.

Ilmberger, W. (2008): Schön und gut? Über den Zusammenhang von Ästhetik und Usability bei der Bewertung von Benutzungsschnittstellen? Unveröffentlichte Diplomarbeit, Heidelberg: Psychologisches Institut der Ruprecht-Karls-Universität.

Kirakowski, J.; Corbett, M. (1993): SUMI: The Software Usability Measurement Inventory. British Journal of Educational Technology, Vol. 24, Nr. 3, S. 210–212.

Kurosu, M.; Kashimura, K. (1995): Apparent usability vs. inherent usability: experimental analysis of the determinants of the apparent usability. Denver, Colorado: Conference Companion of human factors in computing systems, S. 292–293.

Laugwitz, B.; Held, T.; Schrepp, M. (2008): Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (Hrsg.): USAB 2008. LNCS 5298, S. 63-76.

Laugwitz, B.; Schrepp, M.; Held, T. (2006): Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. In: A.M. Heinecke & H. Paul (Hrsg.): Mensch & Computer 2006, 125–134. München: Oldenbourg Verlag.

Markus, H. R.; Kitayama, S. (1991): Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, Vol. 98, Nr. 2, S. 224-253.

Norman, D. (2003): Emotional Design: Why We Love (Or Hate) Everyday Things. Boulder Colorado: Basic Books.

Oyugi, C.; Dunckley, L.; Smith, A. (2008): Evaluation methods and cultural differences: studies across three continents. Paper presented at the Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges.

Tamm, N. (2009) Der Einfluss von Kultur auf Kritikverhalten im Rahmen eines Usability-Test Szenarios- Ein Vergleich zwischen asiatischen und westlichen Probanden Unveröffentlichte Diplomarbeit, Heidelberg: Psychologisches Institut der Ruprecht-Karls-Universität.

Tractinsky, N. (1997): Aesthetics and Apparent Usability: Empirical Assessing Cultural and Methodological Issues. CHI'97 (http://www.acm.org/sigchi/chi97/proceedings/paper/nt.htm).