Ergebnisdokumentation

Autor(en): Backhaus, Claus

Titel: Usability-Engineering in der Medizintechnik: Grundlagen- Methoden- Beispiele

Jahr: 2010

Kapitel: 5.3 Qualitätskriterien beim Einsatz von Usability-Methoden

Laufende Nummer im Literaturverzeichnis: (01)

Schlagwörter: Reliabilität, Usability- Methoden, Testgüte

Kurzbeschreibung:

Das Kapitel „Qualitätskriterien beim Einsatz von Usability-Methoden“ aus dem Buch „Usability-Engineering in der Medizintechnik: Grundlagen- Methoden- Beispiele“ beschreibt die unterschiedlichen Hauptkriterien der Testgüte (Objektivität, Validität und Reliabilität), anhand welcher Usability-Methoden hinsichtlich ihrer Qualität beurteilt werden können. Da es bei dieser Recherche vorrangig um das Gütekriterium „Reliabilität“ gehen soll, werden die anderen Punkte nur kurz abgehandelt.

Unter Objektivität wird die Fähigkeit verstanden, dass, obwohl unterschiedliche Versuchsleiter eine Usability-Methode durchführen können, stets konstante Ergebnisse erzielt werden. Es herrscht somit eine Unabhängigkeit vom Anwender.

Validität ist das Ausmaß dafür, ob eine Methode tatsächlich das Merkmal misst, welches erfasst werden sollte.

Die Reliabilität wiederum gibt an, mit welcher Genauigkeit ein geprüftes Merkmal erhoben wurde. Ausschlaggebend ist hierfür die Fähigkeit einer Usability-Methode, dieselben Ergebnisse bei wiederholter Durchführung zu erzielen. Es kann festgehalten werden: Je geringer der Fehleranteil der Messwerte, desto besser die Reliabilität. Des Weiteren kann die Reliabilität durch eine Standardisierung des Prozesses gesteigert werden. Unerwünschte Variationen in der Durchführung einer Methode werden somit kontrolliert. Ebenso verhindert die Erhebung von großen Datenmengen eine Verzerrung des Gesamteffektes durch Ausreißer oder andere atypische Ergebnisse.

Es werden weiterhin zwei Arten von Reliabilität bei Usability-Methoden unterschieden: Die Retest-Reliabilität und die Paralleltest-Reliabilität. Ersteres gibt die Genauigkeit an, mit welcher zwei in definiertem, zeitlichem Abstand folgende Messungen konstante Ergebnisse erzielen. Die Paralleltest-Reliabilität zeigt an, inwiefern zwei Usability-Methoden bei der Messung von identischen Indikatoren übereinstimmen.

Alle drei Kriterien stehen in engem Zusammenhang und bedingen sich untereinander, sodass eine Verminderung eines Kriteriums zu einer Verschlechterung eines anderen führen kann, was sich somit nachteilig auf die Testgüte einer Usability-Methode auswirken würde.

Erstellt von: Anja Marckwardt, Matrikelnr.: 333029

Autor(en): Backhaus, Claus

Titel: Usability-Engineering in der Medizintechnik: Grundlagen- Methoden- Beispiele

Jahr: 2010

Kapitel: 5.2.3. Expertenbefragung- Usability-Inspection-Methods

Laufende Nummer im Literaturverzeichnis: (02)

Schlagwörter: Usability-Methoden, Expertenbefragung, Cognitive-Walkthrough

Kurzbeschreibung:

Das vorliegende Kapitel handelt von Expertenbefragungen, mit Hilfe welcher Bedienmängel eines Systems oder Produktes frühzeitig identifiziert und reduziert werden sollen. Die Expertenbefragungen stellen allerdings keinen vollständigen Ersatz für Benutzerbefragungen dar, da sie eher einfache Bedienprobleme aufdecken und sollten daher als ergänzende Komponente für ein effizientes Usability-Engineering gesehen werden. Die beiden wichtigsten Methoden in der Expertenbefragung sind die Heuristische Evaluation und der Cognitive-Walkthrough, wobei der Fokus dieser Recherche auf Zweiterer liegt.

Bei der Heuristischen Evaluation geht es darum, dass eine bestimmte Anzahl von Experten an der Benutzerschnittstelle eines Produktes überprüft, inwiefern die Usability-Regeln, also die Heuristiken, übereinstimmen. Vorteilhaft bei dieser Methode sind deren gute Erlernbarkeit sowie die einfache Anwendung. Weiterhin sind die Ergebnisse schnell verfügbar. Allerdings wird der Anwendungskontext bei der Methode zu großen Teilen vernachlässigt und es findet keine Priorisierung der gefundenen Defizite statt.

Beim Cognitive-Walkthrough werden den Experten typische Handlungsabläufe vorgegeben, die im Rahmen der Evaluation untersucht werden sollen. Diese werden im Laufe der Untersuchung mental durchlaufen. In der Vorbereitungsphase bekommen die Experten detaillierte Informationen über das Produkt, was in Form eines Prototyps sein kann. Dadurch soll überprüft werden, ob die vorgegebenen Abläufe vom Nutzer ausgeführt werden können und an welchen Stellen potentielle Probleme auftauchen. Weiterhin werden Angaben über die Nutzergruppe und typische Einsatzbedingungen gegeben. Während der eigentlichen Evaluation, der Analysephase, gehen die Experten die festgelegten Handlungen schrittweise gedanklich durch und notieren sich alle relevanten Informationen. Des Weiteren werden mögliche Quellen der Fehlbedienung aufgenommen.

Besonders vorteilhaft bei dieser Methode sind ihre Einfachheit, ein geringer Aufwand und die eingeschränkte Zahl an ermittelbaren Bedienproblemen. Aus diesem Grund ist der Cognitive-Walkthrough bereits in frühen Entwicklungsphasen einsetzbar, wo eine Benutzerbefragung noch undenkbar ist. Allerdings werden so auch keine Defizite der Durchführung vom tatsächlichen Anwender mit aufgenommen.

Erstellt von: Anja Marckwardt, Matrikelnr.: 333029

Autor(en): Hocko, Jennifer M.

Titel: Reliability of Usability Evaluation Methods

Jahr: 2009

Laufende Nummer im Literaturverzeichnis: (03)

Schlagwörter: Usability-Methoden, Reliabilität, Cognitive-Walkthrough, Evaluator Effekt

Kurzbeschreibung:

Reliabilität ist eines der Hauptkriterien, um Usability-Methoden zu untersuchen. Die untersuchten Merkmale sind unter diesen Umständen konstant und unabhängig von der individuellen Durchführung einer Usability Evaluation. Die Autorin verweist auf Muller, et al. und deren Definition zweier Begriffe, die die Reliabilität weiter beschreiben sollen: Die „Rohausbeute“ (raw yield) und die „verfeinerte Ausbeute“ (refined yield). Ersteres weist auf die Anzahl von spezifischen Klassen von Usability-Problemen hin, die durch ein Bewertungsverfahren aufgedeckt werden. Der zweite Begriff beschreibt den Anteil der schweren Probleme, die durch eine Methode aufgedeckt werden. Eine Usability-Methode gilt also dann als zuverlässig, wenn beide Bedingungen ausreichend erfüllt sind.

Dennoch weisen einige der am häufigsten genutzten Usability-Methoden keine gute Reliabilität auf. In dem Paper werden drei dieser Methoden näher diskutiert: Die Heuristische Evaluation, Usability Tests und der Cognitive-Walkthrough, wobei sich die Recherche auf letzere beschränkt.

Während des Cognitive-Walkthroughs interagiert ein Experte systematisch mit einem Produkt oder einer Schnittstelle mit vorgegebenen Aufgaben-Szenarien. Er versucht dabei, den Problemlösungsprozess eines Nutzers zu simulieren. Im Sinne der Erfinder soll die Methode individuell oder in Gruppen durchgeführt werden, die während des kompletten Evaluationsprozesses zusammenarbeiten.

Die Rohausbeute des Cognitive-Walkthroughs ist schwer zu bestimmen. So zeigen mehrere Studien, dass die Anzahl der gefundenen Probleme stark variieren, abhängig von der Art der Studie. Hinsichtlich der verfeinerten Ausbeute liefert die Methode ähnlich schlechte Ergebnisse, da viele der schweren Probleme nicht aufgefunden werden können.

Eine Erklärung für diese Variation der Zuverlässigkeit ist der sogenannte Evaluator Effekt, ein Faktor, der beschreibt, dass verschiedene Gutachter, die ein und dieselbe Usability-Methode nutzen, nicht fähig sind, dieselbe Anzahl von Problemen zu ermitteln. Hauptgründe für das Auftreten des Effektes sind Unterschiede in der Erfahrung des Gutachters und in der Anwendung der Methode. Auch wenn gezeigt werden konnte, dass Experten, die sowohl Erfahrungen im Bereich Usability als auch mit dem Produkt selber haben, einen geringeres Auftreten des Evaluator Effektes aufzeigen, kann dieser dennoch nicht vollständig vermieden werden. Dies deutet allgemein auf eine geringere Reliabilität des Cognitive-Walkthroughs als bei anderen Usability-Methoden hin. Wegen des möglichen Einsatzes der Methode zu Beginn der Entwicklungsphase, wird diese dennoch oft und gerne verwendet.

Erstellt von: Anja Marckwardt, Matrikelnr.: 333029

Autor(en): Kessner, M.; Wood, J.; Dillon, R. F. & West, R. L.

Titel: On the Reliability of Usability Testing

Jahr: 2001

Laufende Nummer im Literaturverzeichnis: (04)

Schlagwörter: Usability-Methoden, Reliabilität

Kurzbeschreibung:

Laut der Autoren haben bisherige Studien deutlich gemacht, dass unterschiedliche Teams, die bei demselben Produkt einen Usability Test anwenden, auf unterschiedliche Ergebnisse kommen. Gründe dafür könnten unter anderem sein, dass etwa das Produkt bereits lange auf dem Markt war, bevor der Test durchgeführt wurde oder dass die Versuchsleiter voneinander abweichende Erfahrung im Umgang mit den Methoden haben.

In ihrer Studie legen die Autoren diese Faktoren derart fest, sodass ein Auftreten der Problematik unterschiedlicher Ergebnisse vermieden werden kann. Sechs professionelle Usability-Teams wurden zu der Usability-Untersuchung eines neuen Öffnen-Dialog-Fensters (PopUp-Menü zur Schaltfläche Öffnen) einer Softwarefirma gebeten. Dabei ergab sich, dass spezifischere und fokussiertere Aufforderungen zu einer größeren Übereinstimmung der Ergebnisse zwischen den einzelnen Test-Teams führen. Die Reliabilität wurde folglich erhöht. Ausschlaggebend dafür ist, dass der Klient und die Versuchsleiter eng zusammenarbeiten, um ein optimales Untersuchungsdesign zu schaffen und welche Usability-Ziele es gibt. Zwar konnten immer noch nicht alle Usability-Probleme aufgedeckt werden, dennoch tragen die Methoden laut der Autoren zu wichtigen und wertvollen Erkenntnissen bei. Dennoch ist es als nachteilig anzusehen, dass weiterhin die Notwendigkeit besteht, Usability-Tests hinsichtlich ihrer Gütekriterien zu verbessern.

Erstellt von: Anja Marckwardt, Matrikelnr.: 333029

Autor(en): Sarodnik, Florian & Brau, Henning

Titel: Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung

Jahr: 2011

Kapitel: 4.4.5 Reliabilität und Validität

Laufende Nummer im Literaturverzeichnis: (05)

Schlagwörter: Usability-Methoden, Reliabilität

Kurzbeschreibung:

In dem Kapitel zu den Gütekriterien „Reliabilität und Validität“ wird beschrieben, dass individuelle Unterschiede der Testpersonen bei einem Usability-Test für das Auffinden möglichst vieler und unterschiedlicher Usability-Probleme in einem System wünschenswert sind. Allerdings kann es dadurch zu Einbußen in der Reliabilität kommen. Aus diesem Grund ist es wichtig, dass eine angemessen große Anzahl an Probanden an einem Test teilnehmen, insbesondere bei der Erhebung von quantitativen Daten oder bei dem Vergleich zweier Systeme. Einige Forscher geben an, dass es möglich ist, 80% aller Usability-Probleme mit einer Anzahl von 5 Testpersonen auffinden zu können. Da sich daraus jedoch ein großes Konfidenzintervall ergibt, verweisen die Autoren auf die Analysen anderer Studien hin, deren Ergebnisse zeigen, dass beispielsweise die durchschnittlich gefundenen Probleme bei einer Anzahl von 5 Personen 85,55% (SD=9,30) beträgt.

Der tatsächliche Bedarf einer Testgruppe ergibt sich aus einer Vielzahl von projektspezifischen Faktoren. Darunter fallen unter anderem die Zielgruppe des Systems, der Einsatzbereich oder das Budget. Ebenso muss zwischen Experten und Novizen unterschieden werden, da diese unterschiedlich viele und voneinander verschiedene Probleme aufdecken können. Experten sind in ihrer Herangehensweise sehr geschult und können auf ein breites Wissen an Vorerfahrung bei der Durchführung einer Usability-Methode zurückgreifen. Novizen haben diese Vorerfahrung nicht, gehen allerdings oft ungezwungener an eine Methode heran und können eventuell neue Probleme aufdecken. Unabhängig davon sei darauf hinzuweisen, dass es in jedem System immer Usability-Probleme geben wird, die nicht so offensichtlich und daher schwerer aufzufinden sind als andere.

Erstellt von: Anja Marckwardt, Matrikelnr.: 333029

Literaturverzeichnis

(01) Backhaus, Claus (2010a). Qualitätskriterien beim Einsatz von Usability-Methoden. In *Usability-Engineering in der Medizintechnik: Grundlagen- Methoden- Beispiele* (pp. 56-57). Heidelberg: Springer.

(02) Backhaus, Claus (2010b). Expertenbefragung- Usability-Inspection-Methods. In *Usability-Engineering in der Medizintechnik: Grundlagen- Methoden- Beispiele* (pp. 51-53). Heidelberg: Springer.

(03) Hocko, Jennifer M. (2009). Reliability of Usability Evaluation Methods. Internetressource: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=A77D3A3FF352548BB31837F9897B9F77?doi=10.1.1.131.243&rep=rep1&type=pdf> [14.11.2016].

(04) Kessner, M.; Wood, J.; Dillon, R. F. & West, R. L. (2001). On the Reliability of Usability Testing. In *Proceedings of CHI 2001* (pp. 97-98). Seattle.

(05) Sarodnik, Florian & Brau, Henning (2011). Reliabilität und Validität. In *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung* (pp. 173-174). Bern: Verlag Hans Huber.

Literaturrecherche für

Methoden des Usability Engineering und Testing WiSe 16/17

Gruppe: C1

Thema: Testgüte von Usability Methoden

Erstellt von: Ian-Christopher Jung

Matr.-Nummer: 337560

Autor(en): Pospeschill, M
Titel(Buch): Testkonstruktion, Testtheorie Testevaluation
Titel (Kapitel): Kriterien wissenschaftlicher Testgüte
Jahr: 2010
Schlagwörter 3-5: Kriterien wissenschaftlicher Testgüte, Hauptgütekriterien, Nebengütekriterien

Laufende Nummer im Literaturverzeichnis: 3

Erstellt von: Ian-Christopher Jung , Matr. Nr.: 337560

Kurzbeschreibung:

Das Lehrbuch behandelt die Themen der Testtheorie, Testkonstruktion und Testevaluation in Bezugnahme auf wissenschaftliche psychologische Tests. Im betrachteten 1. Kapitel des Lehrbuches werden die Kriterien wissenschaftlicher Testgüte in Hauptgütekriterien und Nebengütekriterien Unterteilt.

Als Hauptgütekriterien werden die Kriterien Objektivität, Reliabilität und Validität angegeben. Die Objektivität eines Testes sei dann gewährleistet, wenn keine Abhängigkeiten zwischen dem Testergebnis und dem Versuchsleiter, dem Testauswerter und der Ergebnisinterpretation bestünden. Hieraus ergäben sich die Unterpunkte: Durchführungsobjektivität, Auswertungsobjektivität, Interpretationsobjektivität. (Würde ein Test von unterschiedlichen Personen am gleichen Probanden ausgeführt müsste ein objektiver Test gleiche Ergebnis liefern). Laut des Autors beschreibt Reliabilität die Genauigkeit bzw. Zuverlässigkeit eines Tests. Ein Test mit hoher Reliabilität zeichne sich dadurch aus, dass möglichst keine Messfehler aufträten. Um Reliabilität genauer zu untersuchen könne man die folgenden Aspekte betrachten: Test-Retest-Reliabilität (Wiederholung des Test zu unterschiedlichen Zeitpunkt), Paralleltest-Reliabilität(Parallel Erhebung in einem Test mittels ähnlichen Items aus unterschiedlichen Tests), Splithalf- Reliabilität(Aufteilung eines Test in zwei Hälften mit äquivalenten Testhälften). Die Validität eines Tests zeige, in wie weit der Test misst was mit ihm gemessen werden solle. Objektivität und Reliabilität werden als notwendige Bedingung für die Validität betrachtet, sie alleine sind jedoch keine hinreichenden Bedingungen. Erst ein Valider Test ermöglicht die Generalisierung der Stichprobentestergebnisse auf die Grundgesamtheit.

Die Gruppe der Nebengütekriterien umfasse die Aspekte: Normierung (Eichung), Skalierung, Ökonomie (Wirtschaftlichkeit), Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit, Fairness, Vergleichbarkeit und Transparenz. Die Normierung würde durchgeführt, in dem Normergebnisse bereit gestellt würden, mit denen man spezifische Testergebnisse vergleichen könnte. Dem Autor zufolge beschreibt die Skalierung, in wie weit die Ergebnisabstände den tatsächlichen vorhandenen Abständen entsprechen. Ein Test sei ökonomisch, wenn der notwendige Ressourcen-Einsatz in einem akzeptablen Verhältnis zum Erkenntnisgewinn stünde, welcher durch den Test erreicht wird. Der Autor erklärt, dass ein Test das Kriterium der Nützlichkeit erfüllt, wenn es keinen anderen Test gibt mit dem das betrachtete Merkmal gleich gut oder besser untersucht werden kann. Zudem solle sichergestellt sein, dass aus dem Test möglicherweise abgeleitete Maßnahmen ein positives Verhältnis von Nutzen und Schaden besitzen. Unter Kriterium Zumutbarkeit würde geprüft ob der Test für die Testperson/den Probanden zumutbar wäre. Laut dem Lehrbuch erfüllt ein Test die Unverfälschbarkeit, wenn die Versuchsperson nicht in der Lage ist, durch gezieltes Testverhalten (möglicherweise gewünschte) Testergebnis zu erzielen. Testfairness sei erfüllt, wenn keine Testperson durch den Test wegen der Zugehörigkeit zu ethnischen, soziokulturellen, oder geschlechtsspezifischen Gruppen systematisch benachteiligt wird. Dem Autor zufolge kann von Vergleichbarkeit gesprochen werden, wenn ein Test mit einem anderen Test verglichen werden kann, der ähnliche Aspekte betrachtet. Dies sollte bei gleichen Testpersonen zu ähnlichen Testergebnissen führen. Ein Test könnte zu dem als Transparent angesehen werden, wenn der Test

verständlich für die Testperson gestaltet sei und der Testperson nach Testdurchführung die Intention und die Ergebnisse des Testes mitgeteilt würden.

Die in diesem Lehrbuch angegebenen wissenschaftlichen Gütekriterien ermöglichen es, unterschiedliche Tests hinsichtlich ihrer Qualität zu bewerten. Zudem schaffen sie Vergleichsmöglichkeiten zwischen verschiedenen Tests. Eine Anwendung dieser wenn auch leicht angepassten Kriterien, auf Usability Methoden dürften die Bewertung und den Vergleich von eben diesen ermöglichen

Autor(en): Stanton N. A. , Young M. S.
Titel: What price ergonomics ?
Jahr: 1999
Schlagwörter 3-5: Reliabilität, Validität, Kosten-Nutzen-Analyse

Laufende Nummer im Literaturverzeichnis: 4

Erstellt von: Ian-Christopher Jung , Matr. Nr.: 337560

Kurzbeschreibung:

Dieses Paper beschreibt zunächst die Vielfältigkeit der Arbeitsbereiche von Human Factors Experten. Darauf aufbauend wird der Schluss gezogen, dass ein Human Factors Experte sich in seiner Analyse nicht nur auf „sichere“ Daten, wie sie in der Anthropometrie vorkommen, stützen könnte, sondern auch Methoden nutzen müsste um menschlich kognitive Funktionen zu analysieren.

Die Autoren zweifeln an, dass diese Methoden hinsichtlich ihrer Reliabilität und Validität ausreichend geprüft werden, bevor sie angewendet werden. Für die Autoren dienen die Güte Kriterien Reliabilität und Validität dazu herauszufinden, ob eine Methode funktioniert oder nicht. Zudem könnten ihrer Meinung nach Methoden mit nachgewiesener Reliabilität und Validität mit erhöhtem Vertrauen angewendet werden.

In mehreren Studien haben die Autoren unterschiedliche Human Factors Methoden, welche jeweils unterschiedliche Aspekte von Usability betrachten, hinsichtlich ihrer Reliabilität und Validität geprüft. Hierfür wurden Personen in der Anwendung der zu prüfenden Methoden trainiert und ihre Untersuchungsergebnisse eines „Usability Experiments“ mit 30 Versuchspersonen mit einander verglichen. Die Studie liefert keine eindeutigen Ergebnisse für die Validität und Reliabilität der untersuchten Methoden (einzelne Ergebnisse für Methoden siehe Studie). Hieraus ziehen die Autoren den Schluss, dass es sich nicht sagen lässt, ob die Anwendung der Methoden lohnenswert ist oder nicht. Da die Studie nur mit Personen durchgeführt wurde, welche die Anwendung der Methode erst kurz zuvor erlernt hatten, ist es als fraglich anzusehen, ob die Ergebnisse für eine allgemeine Bewertung einer Methode generalisiert werden können.

Trotzdem wird herausgehoben, dass durch die Anwendung von Methoden mit niedriger Validität in der Praxis ein finanzieller Vorteil für ein Unternehmen geschaffen werden könnte z. B. durch frühzeitige Aufdeckung von Fehlern und dem Vermeiden später Kosten, durch erhöhte Verkaufszahlen oder durch verminderte Trainings- und Supportkosten. Die Entscheidung über den Einsatz solcher Methoden würde in der Praxis oft Anhand von Kosten-Nutzen-Analysen geführt, mit der Formel: geschätzte Einsparung durch Anwendung der Methode – Kosten der Anwendung der Methode = Vorteil der Anwendung der Methode. Die Autoren schlagen vor, dass es besser sei, in die Kosten-Nutzen-Analyse die Genauigkeit der Methode miteinzubeziehen, da fehlerhafte Ergebnisse der angewendeten Methoden hohe Kosten verursachen könnten. Die Autoren geben mit der Formel: (Grad der Validität der Methode* Voraussichtlicher finanzieller Nutzen der Methode) - Kosten der Anwendung der Methode = Nutzen der Anwendung der Methode, eine alternative Formel für die Kosten-Nutzen-Analyse an.

Das Paper zeigt, dass zum einen Hauptgüte Kriterien wie Reliabilität und Validität Einfluss auf Nebengüte Kriterium wie die Wirtschaftlichkeit haben können. Zum andern wird angedeutet, dass außerhalb der wissenschaftlichen Anwendung Aspekte der Nebengüte Kriterien, z.B. die Wirtschaftlichkeit, eine entscheidende Rolle dabei spielen können, welche Methode eingesetzt wird. Und dass wiederum die Entscheidungsfindung unterstützt werden kann, wenn für eine Methode klare Informationen zu der Reliabilität und Validität vorhanden sind.

Autor(en): Stanton N. A.
Titel: On the reliability and validity of, and training in, ergonomics methods: a challenge revisited
Jahr: 2016
Schlagwörter 3-5: Reliabilität, Validität, Human Factor Methoden

Laufende Nummer im Literaturverzeichnis: 5

Erstellt von: Ian-Christopher Jung , Matr. Nr.: 337560

Kurzbeschreibung:

Im Paper wird beschrieben, dass der Autor und einige seiner Kollegen bereits vor über 15 Jahren gefordert haben, dass Forscher im Bereich Human Factors (HF) nachweisen sollten, dass ihre Methoden die Kriterien der Validität und Reliabilität einhalten und diese Ergebnisse veröffentlicht werden sollten. Laut dem Autor gab es in 15 Jahren allerdings relativ wenig Veröffentlichungen, die sich mit diesen Aspekten beschäftigten. So spiele der Nachweis von Reliabilität und Validität von HF Methoden aber eine wichtige Rolle für die Aufrechterhaltung des wissenschaftlichen Anspruches und der praktischen Professionalität im Bereich einer Ingenieurdisziplin.

Des Weiteren werden die Begriffe Reliabilität und Validität vom Autor definiert. Reliabilität wird als Stabilität einer Methode beschrieben, zum einen der Stabilität zu unterschiedlichen Zeitpunkten als auch der Stabilität, wenn sie von unterschiedlichen Personen angewendet wird. So sollte eine Methode die gleichen Ergebnisse liefern, wenn sie von Unterschiedlichen Personen oder zu unterschiedlichen Zeitpunkten von der gleichen Person auf dieselbe Ausgangssituation angewendet wird. Eine minimale akzeptable Reliabilität einer Methode sei geben, wenn der Methodenerfinder an zwei unterschiedlichen Zeitpunkten durch die Anwendung der Methode auf das gleiche Ergebnis käme. Wann welcher Grad von Reliabilität einer Methode für die Anwendung ausreichend sei, hänge von einer Vielzahl unterschiedlicher Faktoren ab, wie zum Beispiel: des Untersuchungsgegenstandes, der Art des Projekts, den zu Verfügung stehenden Ressourcen z.B.: Zeit oder Geld und der Expertise des Anwenders.

Es gäbe vier unterschiedliche Typen von Validität: Konstrukt-Validität, Inhalts-Validität, Übereinstimmungsvalidität, Vorhersagevalidität. Das theoretische Grundgerüst werde durch die Konstrukt-Validität betrachtet. Inhalts-Validität befasse sich mit der Glaubwürdigkeit, welche die Methode bei ihren Anwendern erwerben könne. Übereinstimmungsvalidität und Vorhersagevalidität sollen beschreiben inwieweit das Messergebnis die zu Messende Größe widerspiegelt. Ihre Unterschiede lägen in der zeitlichen Komponente. Übereinstimmungsvalidität betreffe einen gegenwärtigen Zeitpunkt dieses Aspektes, wohin gegen Vorhersagevalidität sich auf einen Zukünftigen Zeitpunkt beziehe.

Der Autor erklärt, dass die Validität und Reliabilität einer Methode Aussage darüber gibt ob eine Methode funktioniert oder nicht. Zudem sei Reliabilität eine notwendige, allerdings keine hinreichende Bedingung für Validität.

Im Paper wird eine durchgeführte Studie beschrieben, in der HF-Methoden auf ihre Validität und Reliabilität geprüft wurden. Hierfür wurde Intra-Rater Reliabilität getestet als auch die Inter-rater Reliabilität. Bei der Intra-Rater Reliabilität wurden jeweils Ergebnisse von einzelnen Methoden Anwendern zu zwei unterschiedlichen Zeitpunkten miteinander verglichen. Bei der Inter-Rater Reliabilität wurden die Ergebnisse einer Methode von allen Anwendern mit einander verglichen. Zudem wurde die Validität der Methoden bzgl. der Korrektheit ihrer Vorhersagen anhand ihrer „hit rate“ und „false alarm rate“ gemessen. Die Ergebnisse der Methoden unterscheiden sich in allen drei Punkten stark voneinander als auch die Werte für Inter-rater und Intra-rater der einzelnen Methode. Leider zeigt sich

der vom Autor erwähnte Zusammenhang zwischen Reliabilität und Validität in den Ergebnissen nicht sonderlich stark.

Zudem wird in diesem Paper eine Studie beschrieben, in der untersucht wird wie hoch der Zeitaufwand ist, um unterschiedliche HF Methoden zu erlernen und zu trainieren, um diese anzuwenden. Das Ergebnis dieser Studie zeigt deutliche Unterschiede zwischen den Methoden auf. Der Autor stellt die These auf, dass populäre Methoden wie Heuristiken, Checklisten, Interviews oder Fragebögen so beliebt sind, weil der Zeitaufwand der mit ihrer Anwendung einhergeht, im Vergleich zu komplexeren Methoden wie der Hierarchical task analysis, relativ gering ist. Auch wenn es im Paper nicht beschrieben ist, so können diese Aspekte Einfluss auf die Nebengüte Kriterien einer Methode haben, wie z.B.: der Ökonomie.

Autor(en): Kains, H.
Titel: Reliability and validity of findings in ergonomics research
Jahr: 2014
Schlagwörter 3-5: Validität, Reliabilität, Menschlicher Einfluss

Laufende Nummer im Literaturverzeichnis: 2

Erstellt von: Ian-Christopher Jung , Matr. Nr.: 337560

Kurzbeschreibung:

Der Autor hat mittels einer Studie andere veröffentlichte wissenschaftliche Artikel aus dem Bereich Human Factors (HF) auf ihren Umgang mit den Güte Kriterien Reliabilität und Validität untersucht. Laut dem Autor gibt es in diesem Bereich der HF Forschung noch viel Verbesserungspotential und es sei wenig Fortschritt diesbezüglich in den letzten 20 Jahren zuerkennen.

Zu nächst werden in dem Paper wichtige Evaluationsprozeduren aus den Bereichen der technischen Wissenschaften und der quantitativen Sozial- /Geisteswissenschaften vorgestellt, da diese den größten Einfluss auf den Bereich der Human Factors Forschung hätten. Wichtige Bestandteile dieser Prozeduren werden in einem Flowchart-diagramm erfasst, welches zum einen dazu dient, die betrachteten Paper aus dem HF-bereich zu untersuchen und begangene Fehler bezüglich der Evaluation von Forschungsergebnissen zu verdeutlichen. Zum anderen solle das Flowchart-diagramm als Hilfe dienen können, um die Qualität der Forschung im HF-bereich in Bezug auf die Aspekte Reliabilität und Validität zu verbessern.

Die zwei vorgestellten Evaluationsmethoden aus den technischen Wissenschaften sind die Wiederholung von Messungen und der Vergleich von Untersuchungsergebnissen mit unabhängigen Kriterien. Hieraus ergäben sich Reliabilität als Freiheit der Ergebnisse von Streuung und Validität als Freiheit der Ergebnisse von Abweichungen. Da in HF Forschung so gut wie immer ein menschlicher Aspekt in den Untersuchungen beteiligt ist, müsse betrachtet werden, in wie weit ein Menschlicher Einfluss auf die Messergebnisse ausgeübt wird. So könne man unterteilen in die Bereiche des passiven und aktiven menschlichen Einflusses. Bei einem passiven menschlichen Einfluss könnten die Testpersonen die Konstanz des Messergebnisses nicht beeinflussen, somit wären hier mehrfache Messungen durchführbar. Bei dem aktiven menschlichen Einfluss könnte die Konstanz der Messeergebnisse durch aus dem Messvorgang ausgelöste „Veränderungen“ beeinflusst werden, insbesondere bei der (mehrfachen) Wiederholung des Messvorganges an derselben Testperson.

Dies sei auch der Grund warum es wichtig, sei Evaluations-Prozeduren der quantitativen Sozial-/Geisteswissenschaften zu betrachten. Daher werden die folgenden Prozeduren der Evaluation aus diesem Bereich in diesem Paper betrachtet: die Mehr-Teilnehmer Prozedur, der Reliabilitäts- Koeffizient und die angestrebte/angenommene Messung. Die Anwendung der Mehr-Teilnehmer Prozedur sei anzuwenden, wenn zu es einem aktiven menschlichen Einfluss auf die Konsistenz der Messergebnisse kommen könnte. Hierbei werden bei mehreren Teilnehmern eine begrenzte Anzahl von Messungen vorgenommen und diese Verglichen. Allerdings müsse für das zu messende Phänomen die inter-individuelle Gleichwertigkeit als notwendige Bedingung erfüllt werden. Inter-individuelle Gleichwertigkeit könnte nicht einfach angenommen werden, sondern müsste zunächst überprüft werden. Es wird beschrieben, dass sich der Punkt der angestrebten/angenommenen Messung auf die Validität von Messergebnissen bezieht. Diese sei in der quantitativen Sozial- /Geisteswissenschaften besonders schwer nachzuweisen, wie die Definition schon erahnen lasse. Validität als Grad welcher beschreibe, zu welchem Ausmaß das gemessen wird was gemessen werden soll. Eine von Experten bewertete Annahme von Validität würde zudem den empirischen Anspruch an Validität nicht gerecht werden und könnte somit nicht als Alternative in Betracht gezogen werden.

Der Autor hat mit Hilfe des Flowcharts Aspekte der Evaluation aus beiden Bereichen strukturiert und zu den Punkten die wesentlichen Fallstricke erarbeitet. Dieses Hilfsmittel wurde, benutzt um zu zeigen, wie viele Schwächen und Fehler bzgl. der Aspekte Reliabilität und Validität in der dem Bereich der Human Factors Forschung aufgedeckt werden können. So werde in vielen Papern mehrere Bestandteile der Evaluations Prozeduren ignoriert, falsch interpretiert oder ihnen falsch begegnet. Zudem werde in vielen Studien einfach angenommen, dass die Evaluationsprozeduren erfüllt werden. Ziel dieses Papers und der Studie ist es laut dem Autor, eine Diskussion innerhalb der HF Forschungsgemeinschaft über die Reliabilität und Validität der veröffentlichten Arbeiten auszulösen und durch diese Diskussion die Qualität der Forschung zu verbessern. So wäre es zudem wünschenswert, wenn diese Punkte im Rahmen von Konferenz verstärkt behandelt würden.

Autor(en): Annett J.
Titel: A note on the validity and reliability of ergonomics methods
Jahr: 2002
Schlagwörter 3-5: Validität, Reliabilität, Human Factors Methoden

Laufende Nummer im Literaturverzeichnis: 1

Erstellt von: Ian-Christopher Jung , Matr. Nr.: 337560

Kurzbeschreibung:

Das Paper betrachtet die Aspekte der Validität und Reliabilität in der Human Factors (HF) Forschung. Hierbei wird der These nachgegangen, dass es zu stark vereinfacht wäre, wenn man an HF Methoden den selben Anspruch hinsichtlich Reliabilität und Validität richten würde wie zum Beispiel im Bereich der Psychometrie.

So müssten zunächst die Methoden der HF Forschung in zwei Kategorien unterteilt werden. Demnach gäbe es Analytische Methoden und Evaluation Methoden, welche sich hinsichtlich ihrer Zielsetzung unterscheiden sollen. Analytische Methoden hätten es zum Ziel, einen Erkenntnisgewinn über die einer Interaktion zwischen Mensch und Maschine, zu grundlegenden Mechanismen zu generieren. Beispiele für so eine Methode seien die task analysis oder die error prediction. Hingegen sei das Ziel von Evaluations-Methoden eine Beurteilung von bestimmten Aspekten einer Interaktion zwischen Menschen und Maschinen. Ein Beispiel sei die Bestimmung vom Komfort einer Arbeitsumgebung.

Laut dem Autor lässt sich Validität in Konstrukt-Validität und Vorhersage-Validität unterscheiden. So sei die Konstrukt-Validität von besonderer Bedeutung bei Analytischen Methoden, da sich diese mit der Genauigkeit der Methode der zugrundeliegenden Theorie befasst welche genau der Hintergrund sei mit dem Analytische Methoden zu einem Erkenntnisgewinn gelangen. Hingegen sei die Vorhersage-Validität von höherem Interesse bei Evaluations-Methoden. Zudem seien in diesem Zusammenhang die Ähnlichkeiten von evaluativen Methoden und Methoden der Psychometrie hoch. Es lasse sich deshalb ein ähnlicher Anspruch an die Validität in beiden Bereichen stellen, jedoch nicht für Analytische Methoden. Die Tatsache, dass an dieser Stelle keinerlei Überlegungen über die Art des menschlichen Einflusses und auf welche Weise diese die Konstanz möglicher Testergebnisse beeinflussen könnte angestellt werden, dürften einen Kritikpunkt darstellen. Im Bereich der Psychometrie dürfte häufig der menschliche Einfluss passiver Natur sein, welches im HF-Bereich etwas fraglich erscheinen dürfte.

Bezüglich der Reliabilität sieht der Autor eher Probleme bei analytischen Methoden und nicht bei evaluativen Methoden. Diese Annahme impliziert, dass evaluative Methoden aufgrund ihres Aufbaus immer eine inter-individuelle Gleichwertigkeit bei den Testpersonen einhalten, diese Annahme sollte jedoch für jede einzelne Methode überprüft werden und nicht einfach generell angenommen werden. Probleme der Reliabilität von analytischen Methoden sind dem Autor zufolge nicht ausschließlich auf die Methode selbst zurückzuführen, sondern seien häufig Folgen der Erfahrungheit des Anwenders. Ein solcher Rückschluss aus Studien, die nicht ausschließen konnten, dass die Erfahrung des Anwenders einen Einfluss auf die Reliabilität von Ergebnissen der Methoden hat, scheint ohne weitere Studien diesen Aspekt untersuchen etwas weit gegriffen.

Quellenverzeichnis:

(1) Annett, J. (2002). A note on the validity and reliability of ergonomics Methods. *Theoretical Issues in Ergonomics Science*, 3(2) ,228-232.

(2) Kanis, H. (2014). Reliability and validity of findings in ergonomics research. *Theoretical Issues in Ergonomics Science*, 15(1), 1-46.

(3) Pospeschill, M., (2010). Kriterien wissenschaftlicher Testgüte. In *Testkonstruktion, Testtheorie Testevaluation* (pp.16-32) München : Reinhardt Verlag

(4) Stanton, N. A., & Young, M. S. (1999). What price ergonomics?. *Nature*, 399(6733), 197-198.

(5)Stanton, N. A. (2016). On the reliability and validity of, and training in, ergonomics methods: a challenge revisited. *Theoretical Issues in Ergonomics Science*, 17(4), 345-353.

Autor(en): Sarodnik, Florian; Brau, Henning

Titel: Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung

Jahr: 2011

Kapitel: 4.5 und 4.6

Laufende Nummer im Literaturverzeichnis: ()

Schlagwörter: Usability-Methoden, Objektivität, Fragebögen

Kurzbeschreibung:

In Kapitel 4.5 "Methoden der Usability Evaluation: Fragebogen" befassen sich Sarodnik & Brau mit Fragebogen als Instrument des Usability Testings. Sie weisen darauf hin dass die Konstruktion eines Fragebogens der den Gütekriterien, darunte auch Objektivität, genügt sehr zeit- und kostenaufwendig ist und Standardisierung und Eichung anhand großer Stichproben bedarf. Fragebogen die nach einer Norm (z.B ISO Norm 9241/10) entwickelt wurden, erfüllen die Gütekriterien meist hinreichend, da die Norm bereits ausgiebiger Prüfung hinsichtlich der Kriterien unterzogen wurde. Insbesondere die Objektivität sei bei nach Normen entwickelten Fragebogen hoch, unter anderem da der Interviewer wenig Raum zur Beeinflussung hat.

Im Kapitel 4.6 "Methoden der Usability Evaluation: Vergleich und Beurteilung" ordnen die Autoren die verschiedenen Evaluationsmethoden in Bezug auf verschiedene Bewertungskriterien eine Beurteilung in Form von ein bis drei Punkten zu. Dies sei jedoch keinesfalls ein Maß für die Qualität, lediglich für die Quantität. Dabei schneiden die Fragebogen und die GOMS-Methode (Goals, Operators, Methods, Selections Rules) besonder hoch ab.

Autor(en): Amelang, Manfred; Schmidt-Atzert, Lothar

Titel: Psychologische Diagnostik

Jahr: 2012

Kapitel: 2.3.1.1

Laufende Nummer im Literaturverzeichnis: ()

Schlagwörter: Gütekriterien, Objektivität, psychometrische Tests

Kurzbeschreibung:

In dem Kapitel "Grundlagen diagnostischer Verfahren: Gütekriterien diagnostischer Verfahren: Hauptkriterien: Objektivität" definieren Amelang & Schmidt-Atzert Objektivität als "[...] das Ausmaß, in dem die Ergebnisse eines Tests unabhängig von der Person des Versuchsleiters sind" (Seite 133). Die Objektivität lasse sich dabei in drei Aspekte unterteilen. Der erste Aspekt ist die Durchführungsobjektivität. Diese bezeichnet den Grad, den das Verhalten des Versuchsleiters während der Testdurchführung aber auch bei der Bestimmung der Durchführungsbedingungen das Verhalten der Versuchsperson beeinflusst. Erstrebenswert zur Gewährleistung einer möglichst hohen Durchführungsobjektivität seien eine maximale Standardisierung der Testsituation und Minimalisierung der sozialen Interaktion. Eine quantitative Messung sei dabei nur schwer möglich.

Der zweite Aspekt ist die Auswertungsobjektivität. Diese beschreibt, ob gleiches Verhalten von Testpersonen auf die gleiche Weise ausgewertet wird. So kann z.B. bei Leistungsfragebogen bei denen die richtige Lösung festgelegt ist, die Auswertungsobjektivität als absolut eingeschätzt werden. Bei anderen Methoden wie z.B. Fragebogen mit ungebundenen Antworten ist die Objektivität gefährdet. Hier kann die Auswertungsobjektivität empirisch und quantitativ bestimmt werden, indem man mehr als eine Person die Auswertung durchführen lässt, und die Intraklassenkorrelation berechnet. Des Weiteren kann selbst bei ungebundenen Fragebogen die Auswertungsobjektivität gefährdet sein, wenn der Versuchsleiter die Fragebögen unter Zuhilfenahme einer Schablone von Hand ausgewertet, da dieser Prozess fehlerhaft durchgeführt werden kann. Die Autoren empfehlen in diesen Fällen die Nutzung maschineller Auswertungsmethoden.

Der dritte und letzte Aspekt ist die Interpretationsobjektivität. Diese gilt als gegeben, wenn unterschiedliche Versuchsleiter aus den erhobenen Daten die selben Aussagen über die Testperson ableiten. Hierbei muss dem Versuchsleiter Material vorliegen, das die genaue Interpretation der Rohwerte vorgeben. Auch eine Standardisierung der Merkmale ist von Nöten.

Autor(en): Fisseni, Hermann J.

Titel: Lehrbuch der psychologischen Diagnostik: Mit Hinweisen zur Intervention

Jahr: 2004

Kapitel: 4.3.1

Laufende Nummer im Literaturverzeichnis: ()

Schlagwörter: Gütekriterien, Objektivität, Testtheorie

Kurzbeschreibung:

In Kapitel 4.3.1 "Abriss der klassischen Testtheorie: Ermittlung der Testgüte: Objektivität" definiert Fisseni die Objektivität als das Maß "[...] wie weit das Verhalten [der Versuchsperson] als empirisches Relativ eindeutig eindeutig quantifiziert wird in Item- und Test-Scores als numerischem Relativ und wie weit diese Quantifizierung sich eindeutig interpretieren lässt." Das Verhalten einer Versuchsperson soll in einem Versuch immer gleich quantifiziert und auf Basis dieser Quantifizierung auch immer gleich interpretiert werden. Insbesondere dürfen Quantifizierung und Interpretation nicht vom Versuchsleiter abhängen. Fisseni unterscheidet drei Arten von Objektivität. Die Erste ist die Durchführungsobjektivität. Bei dieser geht es um Raum und Zeit des Versuches, aber auch die kognitiv-emotionale Verfassung der Versuchsperson. Eine Standardisierung dieser Aspekte ist für eine hohe Durchführungsobjektivität nötig.

Die Auswertungsobjektivität beschreibt inwieweit die gegebenen Antworten der Versuchsperson auch immer auf gleichen numerischen Wert abgebildet werden. Prüfen lässt sich die Auswertungsobjektivität indem man mehrere Versuchsleiter den Versuch auswerten lässt. Die zwei oder mehr Auswertungen lassen sich nun statistisch geschätzt werden.

Die Interpretationsobjektivität ist ein Maß dafür wie eindeutig ein numerischer Wert aus der Auswertung von dem Versuchsleiter einer Merkmalsausprägung zugeordnet. Hierbei ist die Objektivität eines Testverfahrens gegeben, wenn vom Autor ein Interpretationsschlüssel beiliegt.

Fisseni merkt an dass die Objektivität als Ganzes praktisch eine interaktionsfreie Interaktion des Versuchsleiters mit der Versuchsperson erfordert. Alternative dazu besteht die Möglichkeit das Verhalten des Versuchsleiters und die Interaktion mit der Versuchsperson als festen eingeplanten Bestandteil des Versuches zu sehen und in Durchführung, Auswertung und Interpretation mit einzubeziehen.

Autor(en): Moser, Christian

Titel: User Experience Design

Jahr: 2012

Kapitel: 10

Laufende Nummer im Literaturverzeichnis: (4)

Schlagwörter: Gütekriterien, Objektivität, Usability Testing

Kurzbeschreibung:

In Kapitel 10 "Usability Testing" geht Moser auf verschiedene Usability Test Methoden ein: Das Hallway-Testing, den Pluralistic Walkthrough, den Formalen Usability-Test, die Heuristische Evaluation, den Cognitive Walkthrough, die Usability-Befragung, die GOMS Methode und A/B-Tests. In diesem Kapitel erwähnt er auch die Gütekriterien und damit auch die Objektivität, hier auch 'Unbeeinflusstheit' genannt. In einer Tabelle bewertet er die genannten Usability Methoden in Hinblick auf die Gütekriterien. Hierbei schneiden die GOMS Methode, die A/B-Tests und die Usability-Befragung sehr gut ab, Hallway Testing, die heuristische Evaluation und der Usability Walkthrough hingegen sehr schlecht. Diese Bewertungen sind jedoch als recht subjektiv zu betrachten, da Moser keine Referenzen oder Ähnliches für diese Tabelle angibt, und auch keine Erklärung wie diese Ergebnisse zustande kamen.

Autor(en): Lin, Han X.; Choong, Yee-Yin; Salvendy, Garvriel

Titel: A proposed index of usability: a method for comparing the relative usability of different software systems

Jahr: 2010

Laufende Nummer im Literaturverzeichnis: ()

Schlagwörter: Gütekriterien, Objektivität, Usability Testing, Fragebogen

Kurzbeschreibung:

In ihrem paper präsentieren Lin et al den Purdue Usability Testing Questionnaire (PUTQ), ein Fragebogen Methode zum Vergleich der Usability verschiedener Software Systeme. Sie korrelieren dabei die performance des PUTQ mit dem QUIS von Norman und Shneiderman, und kommen zu dem Ergebnis dass der PUTQ in der Lage ist weit mehr Unterschiede zu entdecken als der QUIS. Auf Seite 268 befassen Lin et al sich mit Usability Tests. Im Zuge der Diskussion über die Validität, der Usability Tests, merken Lin et al an dass diese auch eine hohe Objektivität haben. Sie geben hierfür jedoch leider keine Referenzen oder anderweitige Beweise. Eine weitere Usability Methode der sie eine hohe Objektivität beimessen ist die Formale Modellierung, insbesondere das GOMS Modell von Card et al. Auch hierfür geben sie jedoch keine weiteren Referenzen.

[1] Amelang, Manfred; Schmidt-Atzert, Lothar (2012). *Psychologische Diagnostik*. Springer Verlag.

[2] Fisseni, H. J. (2004). *Lehrbuch der psychologischen Diagnostik: mit Hinweisen zur Intervention*. Hogrefe Verlag.

[3] Han X. Lin , Yee-Yin Choong & Gavriel Salvendy (1997) *A proposed index of usability: A method for comparing the relative usability of different software systems*, *Behaviour & Information Technology*, 16:4-5, 267-277, DOI: 10.1080/01449299711983

[4] Moser, Christian (2012). *User Experience Design*. Springer Vieweg.

[5] Sarodnik, Florian; Brau, Henning (2011). *Methoden der Usability Evaluation: Wissenschaftliche Grundlagen und praktische Anwendung*. Huber Verlag.

Autor(en): Wayne D. Gray & Marilyn C. Salzman

Titel: DAMAGED MERCHANDISE? A REVIEW OF EXPERIMENTSTHAT COMPARE USABILITY EVALUATION METHODS

Jahr: 1998

Laufende Nummer im Literaturverzeichnis: 3

Schlagwörter (3 bis 5): Usabilitymethodenvergleich, Validität, Bedrohungen (threats) der Validität, Zusammenfassung

Erstellt von: Noah Harnisch-Diamond, 387057

Kurzbeschreibung

Der Text untersucht Schwächen, und irreführende Praktiken von fünf für das Themengebiet der Usability ausschlaggebenden Publizierungen. Der Text beginnt mit einer ausführlichen Zusammenfassung der verschiedenen Aspekte der Validität. Validität kann man grob in folgende Teilgebiete untergliedern: Statistische (Schlussfolgerung) Validität, Internevalidität, Externevalidität, Augenscheinvalidität, wie auch die Konstruktvalidität. Diese kann man jeweils auch noch untergliedern. Dies sollte hilfreich für die Erstellung des Wikis sein.

Weiterhing wird der Unterschied zwischen Usertests und Experimenteller Tests deutlich gemacht. Die Hauptaussage hier ist, dass man mit Experimentellen Versuchsanordnungen leichter die Frage der Kausalität beantwortet werden kann. Mann aber mit Usertests oft schon mit nur wenigen Versuchs Teilnehmern schwerwiegende Fehler findet.

Abschließen wird auf das Problem vieler Usability Studien eingegangen über die Resultate hinaus Aussagen zu treffen. Diese Praktik nennt sich Extrapolation und ist methodisch problematisch.

Autor(en): Dumas, J. S., & Fox, J. E.

Titel: Usability testing: Current practice and future directions

Jahr: 2009

Laufende Nummer im Literaturverzeichnis: 2

Schlagwörter (3 bis 5): Usabilitymethodenvergleich, Validität, Bedrohungen (threats) der Validität, Zusammenfassung

Erstellt von: Noah Harnisch-Diamond, 387057

Kurzbeschreibung

Seiten 232-244

Der Text versucht festzulegen was Valide Usability Methoden sind. Anstelle dessen werden aber einfach einige negativ Beispiele und einige wenige Verbesserungsvorschläge genannt. Es wird zwar zuerst eine Liste erstellt welche Punkte bei Usability ausschlaggebend sind:

Fokus auf Usability,

Teilnehmer der Studie sind EndNutzer,

es existiert ein System oder Produkt zum Evaluieren,

Teilnehmer führen einen Task durch während sie laut sagen was sie tun,

Daten werden aufgenommen,

die Resultate werden an ein passendes Publikum vermittelt (Klingt etwas schwammig).

Zu diesen Punkten werden dann auch einige Vorschläge gemacht wie man sie besser umsetzen kann. Für das Think out loud wird z.B. empfohlen das die User die core tasks machen. Das sind Tasks welche im Kern der Aufgabe liegen. Weiterhin wird darauf hingewiesen das viele Developer sich nicht bewusst sind, und es auch ungern gesagt bekommen, dass sie einen Tunnel blick haben. Es wird leider nicht gesagt wie man damit umgehen soll. Der Text weicht aber schnell von einer Operationalisierung ab, und ähnelt eher einer Sammlung von dos and don'ts welche von einem „Praktitioner“ erstellt wurde. Im Großen und Ganzen scheint der Text viele Probleme zu nennen. Die Lösungen, wenn sie denn genannt werden scheinen aber eher aus dem „Commonsense“ Bereich zu rühren als aus empirischer Forschung.

Author: MICHAEL J. WENGER & JAN H. SPYRIDAKIS

Title: The Relevance of Reliability and Validity to Usability Testing

Jahr: 1989

Laufende Nummer im Literaturverzeichnis: 5

Schlagwörter (3 bis 5): Validitätsprobleme, Verbesserungsvorschläge, Schwächen von Usability Methoden, ~Guide lines~

Erstellt von: Noah Harnisch-Diamond, 387057

Kurzbeschreibung

Michael Wenger und Jan Spyridakis merken an dass viele Usability Methoden einfach aus "common sense" Praktiken übernommen wurden (meine Kritik an dem vorherigen Text). Hierbei wurde aber nicht viel auf die Validität der Methoden geachtet. Sie sagen dass viele Autoren Verbale Protokolle als die Methode mit der geringsten Validität und Reliabilität sehen. Sie sagen dass die Beobachter des verbalen Protokolls trotz hoher Urteilerübereinstimmung (Interrater-Reliabilität) eine schlechte Übereinstimmung mit dem zu verwendenden Schlüssel haben. Hinzu kommt dass die Probanden welche über sich selbst berichten, durch eben dieses Verhalten, dass zu beobachtende Verhalten verfälschen. Sie sagen dass es zwar von Ericsson und Simon Vorschläge gibt wie man die Schwächen des Verbalen Protokolls verbessern könnte, dass die meisten Studien diese aber nicht befolgen.

Als nächstes werden Umfragen angegangen. Diese werden als gut angesehen, haben aber die Gefahr schlecht gestaltet zu sein. Dies kommt überraschend oft vor.

Interviews haben die größte Schwäche in der Person welche die Interviews führt.

Es werden noch weitere Methoden untersucht. Dann werden Richtlinien angegeben an welche sich Forscher halten sollten, wenn sie aus Usabilitytesting eine Reliable und Valide resultate haben wollen.

1. Sehr penibel definierte Konstrukte
2. Multiple Messungen und Beobachter
3. Peer Review
4. Verhalten Technik aufzeichnen (nicht durch Menschen)

Außerdem müssen die beteiligten die Schwächen ihrer Methoden eingestehen und ihre Vorschläge im Rahmen dieser Schwächen/Limitationen machen.

Autor(en): Robert Rosenthal; Ralph L. Rosnow

Titel: Essentials of Behavioral Research : Methods and Data Analysis 3rd (Kapitel 4)

Jahr: 2007

Laufende Nummer im Literaturverzeichnis:4

Schlagwörter (3 bis 5): Validitäts Kriterien, Criterion, Content, Concurrent, Convergent, Discriminant

Erstellt von: Noah Harnisch-Diamond, 387057

Kurzbeschreibung

Seite: 113-122

Der Text nennt die wichtigsten Kriterien um einen Test valide nennen zu können. Als erstes Content Validity. In dieser Form von Validität schaut man inwiefern der Test das erfragt was man möchte. Als zweites wird Criterion validity genannt. Das ist die Übereinstimmung dieses Tests mit einem oder mehreren relevanten Kriterien. Im Falle von Usability Studien wäre das dann ob die vorausgesagte Usability denn auch so von dem Endverbraucher empfunden wird. Man kann sich auch die concurrent validity angucken in dem man die Übereinstimmung des Tests, mit der kleinerer Tests, der gleichen Art überprüft. Weiterhin gibt es convergent validity. Bei usability wäre das z.B. ob verschiedene Usabilitylabs die gleichen Ergebnisse liefern oder nicht. Zu guter Letzt gibt es dann noch discriminant validity. Hier wird geguckt inwiefern der Test das heraus kristallisiert was man möchte. Anders gesagt, kann man sich sicher sein das die gefundenen Probleme durch die Ursachen ausgelöst werden die man denkt und nicht andere ähnliche Ursachen. Die verschiedenen Arten der validity sollten als Grundlage dienen um zu entscheiden ob man einen validen test gemacht hat. Dieser Text wird für die Erstellung des Wikis sehr hilfreich. Da er klare und messbare Ansprüche an validity stellt.

Author: R. Stanley Dicks

Title: Mis-Usability: On the Uses and Misuses of Usability Testing

Jahr: 2002

Laufende Nummer im Literaturverzeichnis: 1

Schlagwörter (3 bis 5): Verbesserung, Usability, Pitfalls, weiterführend

Erstellt von: Noah Harnisch-Diamond, 387057

Kurzbeschreibung

Der Autor nennt am Anfang 5 Kernfehler welche schlechte Usability Tests ausmachen (siehe Liste rechts.) Die Publikation dient der besseren Verwendung von Usability Methoden, und verwendet qualitative Daten. Der Autor definiert Usability ("easy to learn, useful, easy to use, pleasant to use"). Er sagt das viele Menschen Dinge "usability testen" obwohl man sie nicht testen kann. Z.B. wenn Studenten ein „talk out loud“ über ihren Kurs machen, dann ist das eine mehr eine markt Analyse als wirkliches Usability testing. Ein weiteres Problem ist der Unterschied zwischen Usability Daten sammeln und eine empirische Usability Studie durchführen.

1. Misunderstanding of the concept of usability itself and of the distinctions between usability studies and empirical tests.
2. Two types of problems with statistics: assuming that a set of quantitative statistics equals a usability test, and misusing statistical results, especially from tests performed without large enough user samples.
3. Using usability tests for verification rather than usability.
4. Lack of knowledge of the limitations of and the proper methods of usability testing to ensure valid and reliable results.
5. Testing for ease of use but not usefulness.

Der Autor kommt zu dem Schluss das Usability Analysen weiterhin durchgeführt werden sollen. Diese aber nüchterner, ohne zu große Versprechungen zu machen. Am wichtigsten ist es dem Autor das Usability ernst genommen wird, und das nicht jeder der sich Gedanken zu einem Produkt macht dies gleich eine Usability Studie nennt. Er befürchtet dies könne das Feld in ein schlechtes Licht werfen, und es Praktikern erschweren einen Manager von der Notwendigkeit einer Analyse zu überzeugen.

Literaturverzeichnis

(01) Dicks, R. S. (2002, October). Mis-usability: on the uses and misuses of usability testing. In Proceedings of the 20th annual international conference on Computer documentation (pp. 26-30). ACM.

(02) Dumas, J. S., & Fox, J. E. (2009). Usability testing: Current practice and future directions. Human-Computer Interaction: Development Process, 231. (03) Izard, C. E. (1977). Human emotions. New York: Plenum Press.

(03) Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. Human-Computer Interaction, 13(3), 203-261.

(04) Rosenthal, R., & Rosnow, R. L. (1991). Essentials of behavioral research: Methods and data analysis. McGraw-Hill Humanities Social.

(05) Wenger, M. J., & Spyridakis, J. H. (1989). The relevance of reliability and validity to usability testing. IEEE transactions on professional communication, 32(4), 265-271.

Erläuterung

(01) Journal Artikel

(02) Buch

(03) Journal Artikel

(04) Kapitel (mit dem Titel „Reliability and Validity of Measurements“) in einem Sammelband. Das Original erschien in einem Buch.

(05) Journal Artikel