J. Ziegler & A. Schmidt (Hrsg.): Mensch & Computer 2010
München: Oldenbourg Verlag, 2010, S. 271-280

Effizientere Usability Evaluationen mit gemischten Prozessen

Martin Schmettow, Cédric Bach, Dominique Scapin

Zusammenfassung

Viele Arbeiten im Usability Engineering befassen sich damit, neue Methoden zur Aufdeckung von Usability Schwachstellen zu entwickeln und ihre Effizienz mit etablierten Methoden zu vergleichen. Hier wird die Perspektive eingenommen, dass Evaluationsmethoden grundsätzlich selektiv in Bezug auf Typen von Schwachstellen sind. Diese Selektivität ist teilweise dafür verantwortlich, dass es bisher nicht gelungen ist die Effizienz von Usability Evaluationen nennenswert zu steigern. Wir zeigen einen einfachen Ausweg auf, der darin besteht, in einem Evaluationsprozess Methoden mit komplementären Profilen zu mischen. Am Fall der Evaluation von Virtual Environment Anwendungen wird gezeigt, dass auf diese Weise Effizienzsteigerungen von 20% bzw. Kostensenkungen von 30% möglich sind.

1 Einleitung

Ein wesentlicher Beitrag der Usability Forschung zur Förderung benutzerorientierter Softwareentwicklung ist die Bereitstellung effizienter Methoden zur Aufdeckung von Usability Schwachstellen. Zwei Meilensteine in dieser Forschungslinie sind der Usability Test (UT), der auf direkter Verhaltensbeobachtung beruht, und die Heuristische Evaluation (HE) als expertenbasierte Inspektionsmethode (Nielsen, 1994). Beide Methoden haben in der industriellen Praxis weite Verbreitung gefunden. Die Effizienz dieser und weiterer Usability Methoden ist Gegenstand vieler Studien, wobei im Allgemeinen zwei unterschiedliche Perspektiven eingenommen werden: die quantitative Steuerung des Evaluationsprozesses und der Vergleich der Effizienz von Evaluationsmethoden.

In zahlreichen Studien wurden neue Prozeduren vorgestellt, um die Effizienz der Schwachstellenentdeckung zu steigern. Typischerweise werden diese vorgeschlagenen Prozeduren mit etablierten Methoden (meistens HE oder UT) in experimentellen Studien verglichen. Einen Überblick über diese Studien im Bereich der Inspektionsmethoden geben Cockton et. al. (2003). Gray & Salzman (1998) kritisierten jedoch viele der frühen (und einflussreichen) Studien wegen ihrer erheblichen methodischen Mängel, was die sogenannte „Damaged Merchandise“ Debatte auslöste. Schmettow & Vietze (2008) wiesen zudem auf die unzureichende statistische Behandlung der Effizienzvergleiche hin, da die meisten dieser Studien die Effizienz in einer einzelnen Statistik zusammenfassen: der mittleren Rate erfolg-

reich aufgedeckter Schwachstellen. Sie heben hervor, dass dieses Maß die Varianz in der Entdeckbarkeit von Schwachstellen und der Entdeckungsfähigkeit der Experten vernachlässigt.

Nur sehr wenige Studien sind über den Ansatz der mittleren Entdeckungsrate hinausgegangen und haben qualitative Unterschiede zwischen Evaluationsmethoden betrachtet. So stellen Frokjaer & Hornbaek (2008) eine neue Inspektionsmethode vor, die auf psychologischen Metaphern beruht, der Heuristischen Evaluation in prozeduraler Hinsicht jedoch stark ähnelt. Allerdings konnten diese Autoren im direkten Vergleich zur HE auf der Ebene reiner Mittelwerte keine nennenswerten Vorteile nachweisen. Möglicherweise in Anbetracht dessen wurde eine tiefer gehender Vergleich durchgeführt. Dabei wurden die Schwachstellen in mehrerer Hinsicht klassifiziert und die Effizienz innerhalb dieser Klassen verglichen. Dabei traten tatsächlich qualitative Unterschiede zu Tage: Bestimmte Schwachstellen ließen sich mit der neuen Methode effizienter aufdecken, andere mit der HE.

In ähnlicher Weise haben Fu et. al. (2002) qualitative Unterschiede zwischen dem empirischen Usability Test und der expertenbasierten HE nachgewiesen, und zwar – bemerkenswerterweise – in einer theoriegeleiteten Studie: Auf Basis des Handlungssteuerungsmodells von Rasmussen (1986) argumentieren sie, dass expertenbasierte Evaluationsmethoden sich eher zur Aufdeckung von *skill-* und *rule*-basierten Schwachstellen eignen, während *knowledge*-basierte Schwachstellen besser in empirischen Usability Tests entdeckt werden. Diese Vorhersage erwies sich als zutreffend und ist ein klares Argument für derartig qualitative Vergleiche von Evaluationsmethoden.

Letzlich ziehen Frokjaer & Hornbaek (2008) aus den Ergebnissen des qualitativen Vergleichs den Schluss, ihre neue Methode sei vorteilhaft, weil damit zwar nicht mehr, aber gravierendere Schwachstellen aufgedeckt würden. Anbetracht des unzureichend verstandenen und operationalisierten Konzepts des Schweregrades von Schwachstellen ist das eher fragwürdig (Hertzum & Jacobsen, 2001). Insbesondere ist die Annahme nicht zulässig, Schwachstellen auf niedrigeren Ebenen der Handlungssteuerung seien weniger schwerwiegend. Dies sei an folgendem Beispiel veranschaulicht: Bei PCs mit deutscher Tastenbelegung liegt das Zeichen '@' auf ALT GR – Q. Dieses Zeichen ist vor allem für die Eingabe von Emailadressen relevant und dürfte bei den meisten Benutzern hoch überlernt sein; das heißt, es liegt eine Handlungssteuerung auf *skill*-Ebene vor. Auf Rechnern der Marke Apple hingegen führt diese Tastenkombination zum sofortigen Schließen der Anwendung ohne Rückfrage. Für Umsteiger von anderen Betriebssystemen führt dies unweigerlich zu Fehlbedienungen mit schweren Konsequenzen. Wegen der Unbewussheit der Handlungsteuerung auf *skill*-Ebene sind diese Fehlbedienungen für den Benutzer besonders schwer zu vermeiden oder zu „verlernen“.

Im Gegensatz dazu betonen Fu et. al. (2002), dass die Stärke beider Methoden gerade in ihrer Unterschiedlichkeit liegt, indem sie den Nutzen in unterschiedlichen Phasen des Entwicklungsprozesses herausstellen. Sie argumentieren, dass sich expertenbasierte Methoden gut zur Evaluation früher Prototypen im Designprozess eignen, um in anschließenden Usability Tests die Schwachstellen auf höheren Ebenen der Handlungssteuerung aufzudecken. Die Autoren kommen jedoch auch zu dem Schluss, dass ein derartiges Vorgehen höhere Kosten nach sich zieht und deswegen in der Praxis wenig Verbreitung finden könnte.

2 Forschungsfragen

Wir stimmen mit der Ansicht von Fu et. al. (2002) überein, dass die Stärke von Evaluationsmethoden in ihren unterschiedlichen Profilen liegt. Jedoch werden wir im Folgenden nachweisen, dass die Mischung von Methoden im Evaluationsprozess zu einer erheblichen Kosten*ersparnis* führen kann. Dazu dienen die Daten einer jüngeren Studie, in der die Effizienz von drei Evaluationsmethoden im Bereich der *Virtual Environments* Anwendungen verglichen wurde (Bach & Scapin, 2010).

Eine Voraussetzung für die hier aufgestellten Forschungsfragen ist, dass sich Schwachstellen grundsätzlich darin unterscheiden, wie gut sie (mit einer bestimmten Methode) aufgedeckt werden können. Dieser Sachverhalt wurde bereits in zwei früheren Studien nachgegangen: Schmettow (2008) wies anhand von Zähldatenmodellen nach, dass die Sichtbarkeit von Schwachstellen in der Regel deutlich variiert (Schwachstellenheterogenität). In einer Folgestudie konnte außerdem gezeigt werden, dass Schwachstellenheterogenität Auswirkungen auf die Performanz des Evaluationsprozesses hat (Schmettow, 2009). Schwachstellenheterogenität kann auch als Selektivität der eingesetzten Evaluationsmethode interpretiert werden: Zunächst werden sehr schnell diejenigen Schwachstellen aufgedeckt für die die Methode besonders sensitiv ist; das erschöpft sich jedoch mit zunehmender Anzahl der unabhängigen Testdurchläufe (im folgenden als *Prozessgröße* bezeichnet). Es werden dann um so mehr Durchläufe benötigt, um solche Schwachstellen zu finden, für die die Methode eigentlich nicht geeignet ist. In letztgenannter Studie wurde Schwachstellenheterogenität für die hier betrachteten Datensätze bereits nachgewiesen, weshalb hier auf diesen Schritt verzichtet werden kann.

Zunächst gehen wir der Frage nach, ob sich eine spezifische Selektivität von Evaluationsmethoden nachweisen lässt. Dazu bedienen wir uns eines grafischen Verfahrens und eines eigens entwickelten statistischen Tests. Es wird sich zeigen, dass insbesondere der Usability Test ein deutlich anderes Profil aufweist als expertenbasierte Methoden.

Dann gehen wir von der Annahme aus, dass sich selektive Methoden erschöpfen *bevor* alle Schwachstellen aufgedeckt wurden. Es wird untersucht, ob sich ein Effizienzgewinn erzielen lässt, indem man von vorne herein zwei Methoden mit komplementärer Selektivität zur Evaluation verwendet. Dazu zeigen wir, dass derart „gemischte“ Prozessen stets zu einer höheren Rate *neu* entdeckter Schwachstellen führen als „reine“ Prozesse.

3 Studiendesign und Datensätze

Im Folgenden soll kurz das empirische Design der Studie vorgestellt werden, das dem typischen Aufbau vergleichender Experimente an Usability Evaluationsmethoden folgt. Eine detaillierte Beschreibung der Studie und eine elaborierte Diskussion des Anwendungsbereiches virtueller Umgebungen finden sich bei Bach & Scapin (2010).

Drei Evaluationsmethoden wurden in unabhängigen Bedingungen verglichen: empirisches Usability Testing (UT), eine dokumentenbasierte Inspektionsmethode (DI) und eine Experteninspektion (EI). In der Bedingung DI lasen und benutzten die Probanden einen umfassenden Katalog ergonomischer Kriterien für virtuelle Umgebungen, während sie in der EI Bedingung allein auf ihr Vorwissen in Softwareergonomie angewiesen blieben (s.u.).

Gegenstand der Evaluation waren zwei virtuelle Umgebungen: eine Lernsoftware zu einem 3D Videospiel (EDU) und eine dreidimensionale virtuelle Karte des Chamonix Tal in den französischen Alpen (MAP). EDU folgt einem klar strukturierten Benutzungsszenario, indem der Benutzer nacheinander 35 Lernaufgaben auf unterschiedlichem Schwierigkeitsniveau löst. In MAP kann der Benutzer relativ frei die Umgebung des Chamonix Tales erkunden und auf touristische Informationen zugreifen.

An dem Usability Test nahmen zehn Personen im Alter zwischen 19 und 24 Jahren teil. Die Teilnehmer verfügten über normale Seh- und Hörfähigkeit, hatten normale Kenntnisse im Umgang mit Computern, jedoch keine ausgewiesenen Vorerfahrungen mit virtuellen Umgebungen. An den beiden Inspektionsbedingungen nahmen 19 Studenten des Faches Arbeitspsychologie im fünften Jahr ihres Studiums teil. Diese hatten wenigstens einen Kurs in Softwareergonomie absolviert, jedoch keine praktische Erfahrung in der Evaluation von virtuellen Umgebungen. Die Zuordnung zu den beiden Inspektionsbedingungen erfolgte randomisiert. Jeder Teilnehmer wurde mit beiden Anwendungen konfrontiert, wobei jede Evaluation pro Anwendung genau 30 Minuten dauerte. Jeder Durchlauf wurde auf Video aufgezeichnet, zusätzlich wurden die Teilnehmer an den Inspektionsbedingungen aufgefordert, die gefundenen Schwachstellen schriftlich zu dokumentieren.

Tabelle 1: Überblick experimentelle Bedinungen

Methode	N	Anzahl Schwachstellen	
		EDU	MAP
DI - Dokumenteninspektion	10	79	88
EI - Experteninspektion	9	39	52
UT - Usability Test	10	76	84
Gesamt	29	127	147

Auf Basis dieses Materials wurden anschließend die einzelnen Schwachstellenberichte validiert und normalisiert. Dabei identifizierten mehrere Experten durch Konsensbildung mögliche falsche Alarme und fassten die validen Schwachstellenereignisse zu Schwachstellen zusammen. Diese Vereinheitlichung *(Matching)* folgte den Empfehlungen von Cockton & Lavery (1999). Tabelle 1 gibt einen Überblick über die Anzahl der vereinheitlichten Schwachstellen in den sechs experimentellen Bedingungen.

4 Selektivität der Evaluationsmethoden

Im Folgenden wird untersucht, inwieweit man bei den drei Evaluationsmethoden von einer Selektivität gegenüber bestimmten Schwachstellen ausgehen kann. Einen deutlichen Hinweis auf Methodenselektivität gibt jedoch schon Tabelle 1, indem ersichtlich wird, dass jede der drei Methoden zwischen ein und zwei Drittel aller bekannten Schwachstellen völlig „übersieht". Ebenfalls bemerkenswert ist, dass die beiden Methoden UT und DI bei einer Prozessgröße von jeweils zehn Durchläufen in etwa gleich viele Schwachstellen aufdecken. Die Methode EI hingegen erscheint deutlich weniger effizient.

Für eine genauere Bestimmung der Selektivität wird im Folgenden die Effizienz je zweier Methoden auf der Ebene individueller Schwachstellen verglichen. Die Frage ist jeweils, wieviele der Schwachstellen signifikant häufiger mit einer der beiden Evaluationsmethoden aufgedeckt werden. Dazu haben wir einen Test entwickelt, der dem Binomialtest ähnelt, jedoch anders als dieser nicht gegen eine a priori festgesetzte Erfolgswahrscheinlichkeit prüft. Stattdessen wird geprüft, wie hoch die Wahrscheinlichkeit ist, dass in zwei Ziehungsreihen dieselbe Grundwahrscheinlichkeit gegeben ist.

Mit diesem Test lässt sich für jede Schwachstelle bestimmen, ob sie eine signifikante „Präferenz" für eine der beiden Methoden hat oder indifferent ist. In Abbildung 1 ist der paarweise Vergleich als *flower plot* dargestellt, wobei die Anzahl der Blätter der Anzahl der Schwachstellen auf der Koordinate entspricht. Da diese Analyse einem vornehmlich explorativen Zweck dient, wurde ein großzügiges Konfidenzintervall von 90% (zweiseitig) gewählt. Damit ist zu erwarten, dass 10% aller signifikanten Ergebnisse rein zufällig zustande gekommen ist.

In beiden Vergleichen von DI und EI (links) wird eine nahezu zusammenhängende Punktwolke sichtbar, die sich jedoch deutlich zur X-Achse neigt. Nur für relative wenige Schwachstellen wird der Unterschied signifikant. Darin wird deutlich, dass die beiden Inspektionsmethoden ein sehr ähnliches Entdeckungsprofil bieten, wobei DI eindeutig effizienter ist. Ein anderes Bild ergibt der Vergleich zwischen DI und UT (Mitte): Die Punktwolke ist wesentlich breiter; in der EDU Bedingung lassen sich sogar visuell drei Partitionen ausmachen, wobei etwa ein Drittel aller Schwachstellen eine signifikante Präferenz aufweist; in der MAP Bedingung sind dies immerhin noch etwa ein Viertel aller Schwachstellen. Eine ähnliche Anzahl signifikanter Präferenzen wird auch im Vergleich von EI und UT beobachtet, wobei erneut die geringe Effizienz von EI deutlich wird. Es sei nur am Rande angemerkt, dass der Effizienzunterschied zwischen EI und DI nicht weiter verwunderlich ist, hatten doch die Probanden in der DI Bedingung einen (offenbar sinnvollen) Kriterienkatalog zur Verfügung.

Diese Ergebnisse illustrieren deutlich den qualitativen Unterschied zwischen dem Usability Test einerseits und den Inspektionsmethoden andererseits. Das lässt sich wie folgt zusammenfassen: Die Dokumenteninspektion evaluiert dasselbe wie die Experteninspektion, nur besser. Der Usability Test evaluiert genauso effizient wie die Dokumenteninspektion, aber etwas anderes.

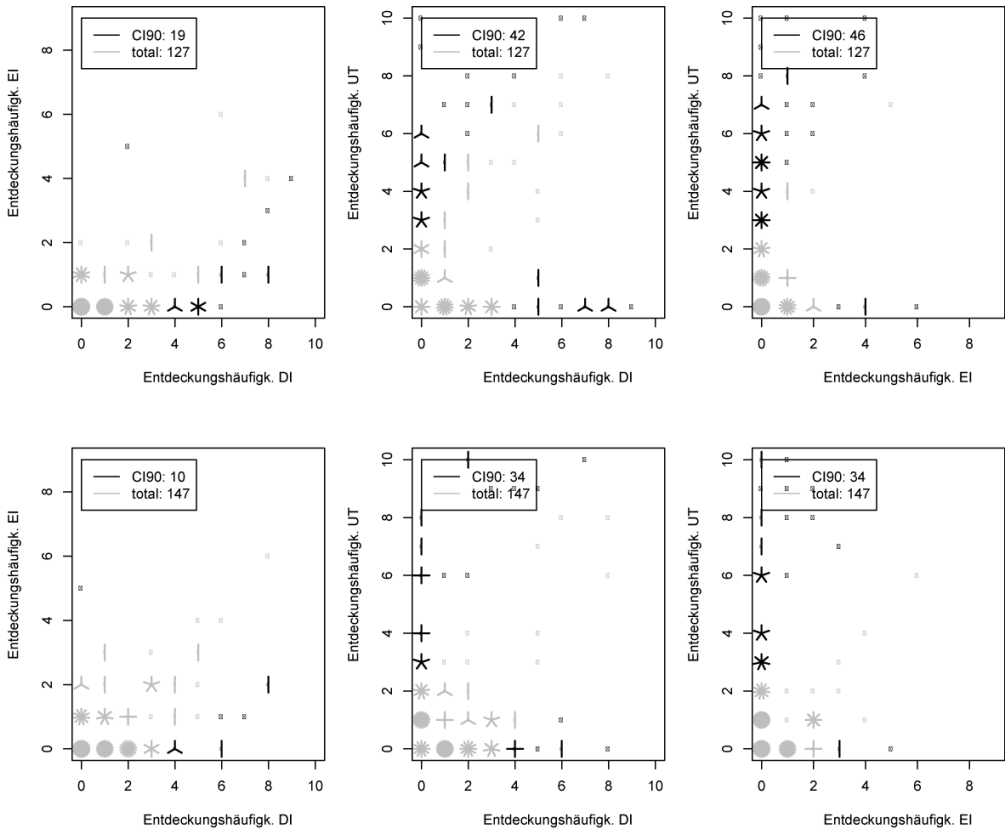

Abb. 1: Bivariater Vergleich der Methodeneffizienz auf Ebene einzelner Schwachstellen

5 Effizienz gemischter Evaluationsprozesse

Ähnlich wie in der Studie von Fu et. al. (2002) sind also Usability Test und Inspektion komplementär, d.h. sie unterscheiden sich in ihren Stärken und Schwächen. Damit ist die Voraussetzung für die zweite Forschungsfrage erfüllt: Welcher Nutzen ergibt sich aus einem Evaluationsprozess, in dem zwei komplementäre Methoden in einem optimalen Verhältnis gemischt werden? Dieser Frage wird im Folgenden nachgegangen, wobei die wenig empfehlenswerte Methode EI keine weitere Berücksichtigung finden soll. Außerdem werden die Bedingungen MAP und EDU der Einfachheit halber vereinigt. Das ist möglich, da auf diesem Faktor ein *within-subject* Design vorliegt.

Um das optimale Mischungsverhältnis von DI und UT und den Zusatznutzen gegenüber den jeweils reinen Prozessen zu ermitteln, wird auf das Verfahren des Monte-Carlo Samplings zurückgegriffen, das in ähnlicher Weise bereits von Schmettow & Niebuhr (2007) verwendet wurde. Dazu werden für eine betrachtete Prozessgröße UT-Durchläufe und DI- Durchläufe

in einem bestimmten Mischungsverhältnisses zufällig gezogen und dann die Anzahl mindestens einmal entdeckter Schwachstellen bestimmt. Für jede betrachtete Prozessgröße und jedes mögliche Mischungsverhältnis wird diese Ziehung 1000 Male wiederholt, um eine Verteilung der Effizienz zu ermitteln. Anhand dieser Verteilungen kann die Effizienz der Mischungsverhältnisse, einschließlich der reinen Prozesse, verglichen werden.

Die Ergebnisse für die Prozessgrößen 6,8 und 10 sind in Abbildung 2 dargestellt. Es wird deutlich, dass das Mischen von Methoden in einem Evaluationsprozess die Effektivität der Schwachstellenerkennung erhöht. Zum Beispiel genügt bereits ein einziger UT Durchlauf, um in mehr als 50% der Fälle eine höhere Effektivität zu erzielen als mit einem reinen DI Prozess derselben Größe *jemals* möglich erschien. Bei jeder der drei Prozessgrößen werden im optimalen Mischungsverhältnis etwa 20% mehr Schwachstellen aufgedeckt, als mit dem effizienteren der beiden reinen Prozesse. Dieser Zusatznutzen lässt sich im Vergleich *zwischen* den Prozessgrößen auch als Kostenersparnis ausdrücken: So entdeckt man mit einem optimal gemischten Prozess der Größe 6 deutlich mehr Schwachstellen als mit einem reinen DI Prozess der Größe 8 und praktisch genauso viele wie mit einem reinen UT Prozess der Größe 10.

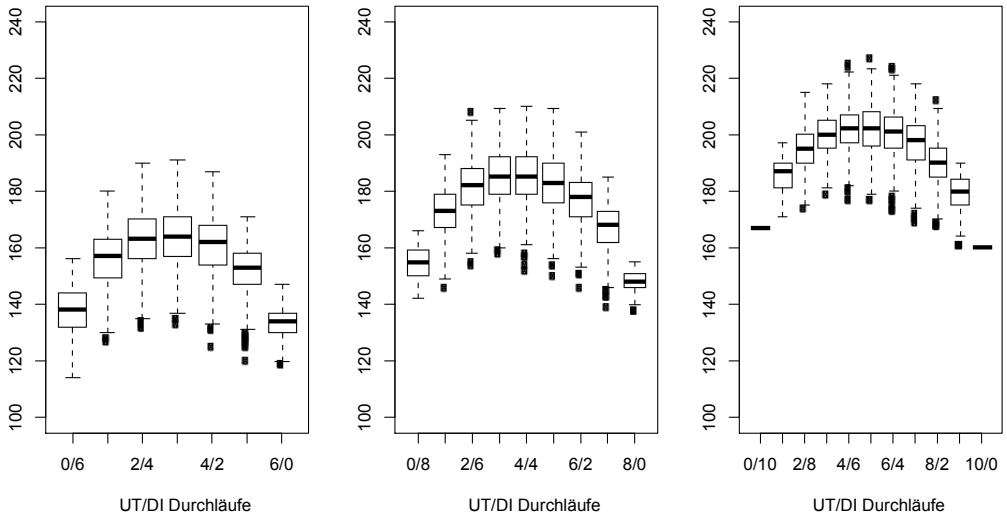

Abb.2: Ergebnis des Monte-Carlo Samplings. Verteilung der Effizienz in variierenden Mischungsverhältnissen bei Prozessgrößen 6, 8 und 10.

Einen Überblick über die Ergebnisse bei den Prozessgrößen 2 bis 10 gibt die Tabelle 2. So ist ein merklicher Zusatznutzen von Mischungen bereits bei sehr kleinen Prozessgrößen gegeben, und steigt dann auf über 20% an. Mit sechs gemischten Durchläufen wird bereits ein besseres Ergebnis erzielt als mit neun DI Durchläufen, was einer Ersparnis von einem Drittel entspricht. Aufgrund der limitierenden Stichprobengröße von 10 konnten wir die tatsächliche Ersparnis bei den Prozessgrößen 7-10 nicht ermitteln. Nach unserer Einschätzung dürfte sie sich aber in Anbetracht des asymptotischen Verhaltens auch für größere Prozesse im Bereich von 30% und möglicherweise mehr bewegen.

Tabelle 2: Mittlere Effizienz sowie Zusatznutzen und Kostenersparnis von optimal gemischten Prozessen

Prozessgröße	2	3	4	5	6	7	8	9	10
Mean(DI)	78	98	114	127	138	147	154	161	167
Mean(UT)	83	101	114	125	133	141	148	154	160
Optimaler Mix DI/UT	1/1	1/2	2/2	2/3	3/3	3/4	4/4	4/5	5/5
Mean(Optim. Mix)	91	115	135	150	164	176	186	195	202
Zusatznutzen (% Schwachst.)	9,6%	13,9%	18,4%	18,1%	18,8%	19,7%	20,8%	21,1%	21,0%
Ersparnis (Anz. Durchläufe)	0	1	1	2	3	≥ 3	≥ 2	≥ 1	-

Weiterhin fällt auf, dass die optimale Mischung unabhängig von der Prozessgröße in einem Gleichverhältnis zu liegen scheint. Das ist hier durch die ähnliche Effektivität der beiden Methoden gegeben und muss keineswegs immer der Fall sein. So zeigte eine Analyse der Mischung von UT und EI, dass ein Verhältnis von 5/1 ebenfalls einen (zugegeben geringen) Zusatznutzen von 2,3% gegenüber einem reinen UT Prozess ergibt. Demzufolge kann auch eine vergleichsweise ineffiziente Methode einen gewissen Nutzen entfalten, sofern sie ein komplementäres Profil aufweist, also bestimmte Schwächen der effizienteren Methode ausgleicht.

6 Diskussion und Fazit

Es gibt Situationen, in denen Usability eine derart geschäftskritische Rolle spielt (etwa im E-Commerce), dass eine möglichst umfassende Usability Evaluation notwendig ist. In diesem Falle schlagen wir den Einsatz gemischter Evaluationsprozesse vor. Dabei wird das spezifische Profil einzelner Methoden ausgeglichen, so dass sich der Evaluationsprozess weniger schnell „erschöpft“. Unsere Ergebnisse zeigen, dass dadurch mehr Schwachstellen mit erheblich weniger Aufwand aufgedeckt werden können. Der Nutzen ist in unserer Studie deutlich größer als bei den meisten Versuchen, die Effizienz von Inspektionsmethoden durch Modifikationen zu steigern.

Dazu gilt es jedoch ebenfalls, die spezifischen Stärken und Schwächen von gängigen Methoden zu ermitteln, da der Effizienzgewinn von der Komplementarität der Methoden abhängt. Einen einfachen Ansatz dazu haben wir hier bereits vorgestellt: Die bivariaten Plots eignen sich zunächst dazu, das Ausmaß von Komplementarität zu bestimmen; sie können jedoch auch als Ausgangsbasis zur inhaltlichen Klassifikation von Schwachstellen dienen, was eine Voraussetzung für die inhaltlichen Bestimmung von Methodenprofilen ist. Mächtigere statistische Verfahren sind möglicherweise in der probabilistischen Testtheorie zu suchen, wie in einer früheren Arbeit bereits vorgeschlagen wurde (Schmettow & Vietze, 2008). Anzumerken sei noch, dass sich derartige Analysen ohne weiteres mit bestehenden Datensätzen durchführen lassen. Das Paradigma der experimentell vergleichenden Evaluationsstudie bleibt von unseren Vorschlägen unberührt.

Für eine realistische Kosten-Nutzen-Beurteilung ist unter anderem auch der Schweregrad der Schwachstellen von Bedeutung. Schweregradeinschätzungen lagen in dieser Studie nicht vor. Es könnte also durchaus sein, dass etwa der Usability Test die schwerwiegenderen Probleme aufdeckt und aus diesem Grunde die bevorzugte Methode sein müsste. Diese Frage lässt sich jedoch ausschließlich mit einem validen Konzept zur Schweregradeinschätzung beantworten, das derzeit nicht verfügbar ist (Hertzum & Jacobsen, 2001). Und, wie wir oben an einem Beispiel gezeigt haben, ein einfacher Zusammenhang zwischen psychologischen Modellen und der Relevanz einer Schwachstelle besteht nicht.

Das Augenmerk sollte in Zukunft in der Frage liegen, welche der existierenden Methoden sich wie und wann effizient in den Entwicklungsprozess integrieren lässt. Dazu muss zunächst untersucht werden, was eine Evaluationsmethode eigentlich genau leistet, anstatt zu messen wie gut sie *irgendetwas* tut. Der Ansatz von Fu et. al. (2002) ist nicht nur wissenschaftlich fundiert, er zeigt auch praktisch diese Richtung auf. Allerdings ist deren Schlussfolgerung nicht ganz eindeutig, dass man *skill*-basierte Schwachstellen erst durch Inspektionen ausmerzen müsse, bevor man sich den *knowledge*-basierten in Usability Tests zuwende. Ebenso ist in einem konkreten Entwicklungsprojekt von Belang, welche Art von Schwachstellen man in der anstehenden Iteration zu beseitigen gewillt ist. Unter anderem kommt hier die Erkenntnis des Software Engineerings zum Tragen, dass früh eingeführte Schwachstellen die höchsten Kosten in der Beseitigung nach sich ziehen (Boehm & Basili, 2001). Beispielsweise ist denkbar, dass gerade die *knowledge*-basierten Schwachstellen mit grundlegenden Benutzeranforderungen im Zusammenhang stehen, etwa dem Ablauf von Arbeitsprozessen. Diese sind oft tief in der Architektur des Systems verankert (und damit teuer in der Schwachstellenbeseitigung), was dann für einen frühen Einsatz empirischer Evaluationsmethoden spräche.

Abschließend sei noch angemerkt, dass die Inspektionsforschung im Software Engineering das verwandte Prinzip der perspektivenbasierten Inspektion entwickelt und positiv evaluiert hat. Es wurde außerdem von Zhang, et. al. (1998) mit Erfolg auf Usability Inspektionen übertragen, was aber wenig Beachtung gefunden hat.

Kontaktinformationen

Martin Schmettow	Cédric Bach	Dominique Scapin
m.schmettow@utwente.nl	cedric.bach@irit.fr	dominique.scapin@inria.fr
University of Twente	University of Toulouse	INRIA
7500AE Enschede, Niederlande	31062 Toulouse Frankreich	78153 Le Chesnay Frankreich

Literatur

Bach, C. & Scapin, D.L. (2010). Comparing inspections and user testing for the evaluation of virtual environments. *Intern. Journal of Human-Computer Interaction*, In press.

Boehm, B.W. & Basili, V.R. (2001). Software defect reduction top 10 list. *IEEE Computer 34*(1):135–137.

Cockton, G. & Lavery, D. (1999). A framework for usability problem extraction. In: *Proceedings of Interact 99*. IOS Press: Amsterdam, 344–352.

Cockton, G., Lavery, D & Woolrych, A. (2003). Inspection-based evaluations. In: *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. Lawrence Erlbaum Associates, 1118–1138.

Frøkjær, E. & Hornbæk, K. (2008). Metaphors of human thinking for usability inspection and design. *ACM Transactions on Computer-Human Interaction*, *14*(4), ACM Press, 1–33.

Fu, L. Salvendy, G. & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology, 21*(2), 137–143.

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.

Hertzum, M. (2006). Problem prioritization in usability evaluation: From severity assessments toward impact on design. *International Journal of Human-Computer Interaction, 21*(2), 125–146.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press: New York, 152–158.

Schmettow, M. (2008). Heterogeneity in the usability evaluation process. In David England & Russell Beale: *Proceedings of the HCI 2008*, Band 1. British Computing Society, 89–98.

Schmettow, M. (2009). Controlling the usability evaluation process under varying defect visibility. In Blackwell, A.F.: *Proceedings of the HCI 2009*. British Computing Society, 188-197.

Schmettow, M. & Niebuhr, S. (2007). A pattern-based usability inspection method: First empirical performance measures and future issues. In Ramduny-Ellis, D. & Rachovides, D.: *Proceedings of the HCI 2007*, Band 2. British Computing Society, 99–102.

Schmettow, M. & Vietze, W. (2008). Introducing Item Response Theory for measuring usability inspection processes. In: *Proceeding of SIGCHI conference on Human factors in computing systems*. ACM Press: New York, 893–902.

Zhijun Zhang, Victor Basili, and Ben Shneiderman (1999). Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Softw. Engineering 4*(1), 43–69.