Neuland adé

Wie können wir virtuelle Realitäten evaluieren?

Marie Muhr
Psychologisches Institut
Universität Heidelberg
Heidelberg, Germany
muhr@stud.uni-heidelberg.de

Veronika Lerche
Psychologisches Institut
Universität Heidelberg
Heidelberg, Germany
veronika.lerche@psychologie.uni-heidelberg.de

Konstantin Knöß
UX Consultant
BLUPRNT GmbH
Offenbach, Germany
konstantin.knoess@bluprnt.de

Denis Villmen
UX Consultant
BLUPRNT GmbH
Offenbach, Germany
denis.villmen@bluprnt.de

ABSTRACT

Virtuelle Realitäten (VR) halten zunehmend Einzug in den privaten und professionellen Alltag. Im Vergleich zu traditionellen User Interfaces bringt die virtuelle Realität eine hohe Anzahl an Freiheitsgraden im Erleben mit sich. Damit wird die Fülle an Erfahrungen erweitert, die die Benutzer machen können. Gleichzeitig jedoch erreichen solche Neuerungen schnell ein hohes Maß an Komplexität in der Gestaltung von Design- und Evaluationsprozessen. Während es für die Usability-Evaluation von traditionellen User Interfaces erprobte Verfahren gibt, ist bislang unklar, ob diese Verfahren auch für VR-Applikationen geeignet sind. In dieser Studie wurden die Evaluationsmethoden Usability Test und Cognitive Walkthrough zur Evaluation einer VR-Applikation eingesetzt. Effektivität und Effizienz der Methoden wurden verglichen, indem die Anzahl der identifizierten leichten und schwerwiegenden Usability-Probleme erfasst wurde. Hieraus leiten sich Handlungsempfehlungen für die Praxis des VR-Testings ab.

KEYWORDS

Virtual Reality, Usability Evaluationsmethoden, UX-Test, Usability Test, Experten Evaluation

1 Herausforderungen bei der Evaluation von VR-Applikationen

Die scheinbar unbegrenzten Möglichkeiten von Virtual Reality (oder Virtual Environment (VE)) haben die Entwicklungsprozesse neuer Devices und Interaktionstechniken in den letzten Jahren stark vorangetrieben (LaViola, Kruijff, McMahan, Bowman & Poupyrev, 2017). Mit VR-Headsets halten virtuelle Realitäten zunehmend Einzug in den privaten und professionellen Alltag (z.B. Gaming, Ausbildung und Training, Produktentwicklung). Um eine hohe Akzeptanz solcher Anwendungen in der breiten Masse zu erreichen, müssen Usability-Professionals in der Lage sein, bestehende Usability-Probleme zu identifizieren und zu beheben (Dey, Billinghurst, Lindeman & Swan, 2018).

Neben den VR-spezifischen technologischen Merkmalen sind für Usability-Professionals vor allem die neuen Charakteristika aus der Erlebnisperspektive des Nutzers relevant, da diese Informationen über die Art und Beschaffenheit der Entscheidungs- und Nutzungsprozesse in der virtuellen Welt liefern. Die zentralen VR-spezifischen Charakteristika sind *Sense of Presence*, *Immersion* und *Involvement* (Witmer & Singer, 1998). *Sense of Presence* beschreibt das Gefühl voll und ganz in die virtuelle Welt einzutauchen. Das Erleben von Presence verzerrt die Wahrnehmung des Nutzers teilweise oder vollständig, sodass er den Anteil und die Rolle der Technologie während der virtuellen Erfahrung nicht adäquat einschätzen kann (International Society for Presence Research, 2000). Nach Slater (2009) beschreibt *Immersion* die objektiven und messbaren Eigenschaften und Methoden eines Systems, welche die Wahrnehmung und das Nutzerverständnis der virtuellen Welt fördern. *Involvement* bezeichnet die Bereitschaft des Nutzers, mit dem System durch die angebotenen Interaktionstechniken zu kommunizieren bzw. interagieren (Bouvier, Lavoué & Sehaba, 2014). Involvement und Immersion sind notwendige Bedingungen, um das Gefühl von Presence erleben zu können (Witmer & Singer, 1998). Auf Grund dieser neuen Eigenschaften von VR ist unklar, ob erprobte Verfahren zur Evaluation traditioneller User Interfaces (z.B. Heuristische Evaluation, Cognitive Walkthrough, Usability Tests, etc.) noch greifen (z.B. Dünser, Grasset & Billinghurst, 2008).

Bowman, Gabbard und Hix (2002) unterscheiden vier Kategorien von Herausforderungen, die sich bei der Evaluation von VR-Applikationen ergeben: *Physical Environment Issues*, *Evaluator Issues*, *User Issues* und *Evaluation Type Issues*.

Physical Environment Issues. Den größten offensichtlichen Unterschied zwischen einem traditionellen Interface und einem 3D User Interface stellen die physische Umwelt des Nutzers sowie die neuen Input- und Output-Devices dar. So kann der Evaluator beispielsweise die virtuelle Welt lediglich via 2D-Screen Übertragung aus der realen Welt verfolgen und die virtuelle Welt nicht so wie die Testperson erleben.

Evaluator Issues. Eine weitere Herausforderung für den Evaluator ergibt sich aus der VR-Charakteristik des Sense of

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

https://doi.org/10.18420/muc2019-up-0276

Presence. Der Testleiter könnte diesen Sense of Presence stören, falls die Testperson die reale Welt durch Geräusche oder Bewegungen des Testleiters wahrnimmt.

User Issues. Bisher liegen nur wenige Informationen dazu vor, welche Fähigkeiten und Erfahrungen (z.B. Domänenerfahrung, erworbene Gaming-Fähigkeiten, Hand-Augen-Koordination etc.) der Testpersonen einen Einfluss auf die Nutzung von VR-Applikationen haben. Ohne Berücksichtigung oder Erhebung VR-relevanter Eigenschaften der Testpersonen, sind die Evaluationsergebnisse schwer zu generalisieren. Deshalb werden große, diverse Stichproben empfohlen. Eine Unterscheidung zwischen Experten und Novizen kann daher zunächst ausschließlich in Bezug auf die Usability Expertise und nicht in Bezug auf VR-Expertise vorgenommen werden (LaViola, Kruijff, McMahan, Bowman, & Poupyrev, 2017).

Evaluation Type Issues. Bisher fehlen etablierte Heuristiken, Guidelines oder benutzerzentrierte Methoden zur Evaluation von VR-Applikationen.

Ziel dieser Studie ist die Weiterentwicklung und Adaption von Usability-Evaluationsmethoden für VR-Applikationen. Dabei analysieren wir, ob mit den Methoden *Usability Test* und *Cognitive Walkthrough* Usability-Probleme bei VR-Applikationen effektiv aufgedeckt werden können und ob der Sense of Presence trotz Testung mit Thinking-Aloud erhalten bleibt. Im Folgenden gehen wir zunächst auf die Effektivität von Usability-Evaluationsmethoden ein. Im Anschluss richten wir den Fokus auf Methoden zur Evaluation von VR-Applikationen.

2 Usability-Evaluationsmethoden

Zu den benutzerzentrierten Usability Evaluationsmethoden zählen *Usability Tests.* Während des Usability Tests bearbeiten „echte" Nutzer anwendertypische Aufgaben und berichten ihre Erfahrungen mit dem User Interface (z.B. bei Anwendung der *Thinking-Aloud-Methode*) (Nielsen, 1993; Sarodnick, & Brau, 2016). Zu inspektionsbasierten Verfahren zählt u.a. die *Heuristische Evaluation,* mit welcher Experten ein User Interface explorativ untersuchen und die gefundenen Usability-Probleme anhand von Heuristiken kategorisieren (Nielsen, 1993; Nielsen, 1994). Beim inspektionsbasierten Cognitive Walkthrough werden Usability-Experten anhand nutzertypischer Aufgaben durch das User Interface geführt (Sarodnick, & Brau, 2016).

Welche Evaluationsmethode den größten Nutzen verspricht, beschäftigt die Forschung und Praxis seit ca. 30 Jahren (z.B. Jeffries, Miller, Wharton & Uyeda, 1991; Nielsen, 1989; Molich & Nielsen, 1990; Schmettow, 2012; Virzi, 1992). Ein Faktor, der in diesem Zusammenhang untersucht wird, ist die nötige Anzahl an Personen (d.h. Nutzer bei der benutzerzentrierten Evaluierung und Evaluatoren bei der inspektionsbasierten Evaluierung) um einen Großteil der Usability-Probleme eines Produkts aufzudecken. Einen Meilenstein in der Forschungsdebatte stellen dabei die Studien von Nielsen (1992, 1993) dar, in welchen er drei Testgruppen mit unterschiedlicher Expertise verglich. Anhand einer Simulationsstudie zeigte Nielsen, dass bei der Heuristischen Evaluation des Prototypens eines Telefonbanking-Systems zwei bis drei Usability-Experten mit Domänenerfahrung oder drei bis fünf Usability-Experten ohne Domänenerfahrung nötig sind, um durchschnittlich ca. 80% der vorhandenen Usability-Probleme zu identifizieren. Um eine ähnliche Anzahl an Problemen zu identifizieren bedarf es dagegen 14 Usability-Novizen mit Domänenerfahrung.

Mit einem ähnlichen Verfahren wie Nielsen (1992, 1993) untersuchten Virzi (1992) und Faulkner (2003) die Anzahl der nötigen Personen bei der benutzerzentrieren Methode des Usability Tests. Beide konnten Nielsens Befunde (1992, 1993) replizieren. Der mittlere Prozentsatz der gefundenen Usability-Probleme betrug bei einer Stichprobengröße von fünf Testpersonen ca. 80%. Diese Zahl ist vergleichbar mit der Anzahl nötiger Evaluatoren bei der Heuristischen Evaluation nach den Befunden von Nielsen. Faulkner analysierte darüber hinaus die Variabilität der Problemidentifikation. Die Standardabweichung lag dabei bei 9.3% und der Prozentsatz der identifizierten Usability-Probleme einer zufällig ausgewählten Stichprobe von fünf Personen zwischen 55% und 100%. Es besteht also die Gefahr, dass bei einer aus fünf Testpersonen bestehenden Gruppe nur 55% der Usability-Probleme identifiziert werden. Die Studienergebnisse von Spool und Schroeder (2001) ergeben sogar, dass fünf Testpersonen im Rahmen von Usability Tests nur durchschnittlich 35% der Usability-Probleme identifizieren.

Basierend auf den Studien von Nielsen (1992, 1993) und Virzi (1992) hat sich die Annahme verbreitet, dass sowohl bei einer Heuristischen Evaluation, als auch bei einem Usability Test lediglich fünf Personen (Evaluatoren bzw. Testpersonen) nötig sind, um eine ausreichend hohe Anzahl von Usability-Problemen (ca. 80%) zu identifizieren. Diese Annahme wird bis heute kontrovers diskutiert. Die Empfehlungen für eine geeignete Anzahl an Testpersonen schwanken sowohl bei inspektionsbasierten, als auch bei benutzerzentrierten Methoden zwischen fünf und zehn Personen (z.B. Faulkner, 2003; Hwang, & Salvendy, 2010, Lewis, 1994; Schmettow, 2012; Spool & Schroeder, 2001). Diese Schwankungen können einerseits in den unterschiedlichen methodischen Herangehensweisen begründet liegen. Andererseits kann die Wahrscheinlichkeit, ein bestimmtes Usability-Problem zu finden auch von Eigenschaften der beteiligten Personen und des zu evaluierenden Systems abhängen. So hängt die Wahrscheinlichkeit der Identifikation u.a. von dem Schweregrad des Problems, den Erfahrungen und Fähigkeiten des Testers, der Stichprobengröße, dem Produkttyp und der Teststruktur ab (u.a. Caulton, 2001; Faulkner, 2003; Grosvenor, 1999; Hwang, & Salvendy, 2010, Lewis, 1994; Schmettow, 2012; Spool & Schroeder, 2001; Woolrych & Cockton, 2001). Falls z.B. ein Produkt bereits eine gute Usability vorweist, sind mehr Testpersonen nötig, um 80% der Usability-Probleme zu finden (Lewis, 1994).

Hinsichtlich des Schweregrads des Usability-Problems zeigte sich, dass anhand von Heuristischen Evaluationsmethoden—im Unterschied zu der Methode des Cognitive Walkthroughs, Guideline Evaluationen und Usability Tests—die meisten schwerwiegenden Usability-Probleme gefunden wurden (Jeffries, Miller, Wharton & Uyeda, 1991). Gleichzeitig wurden eine hohe Anzahl spezifischer und weniger schwerwiegender Usability-Probleme identifiziert, die lediglich ein einzelner Evaluator

entdeckt hat. Auch bei Nielsen (1992) fanden Experten einen höheren Anteil an schwerwiegenden Usability-Probleme als an weniger schwerwiegenden Usability-Problemen. Gleichzeitig ergibt sich aus der Analyse der absoluten Häufigkeiten, dass die Evaluatoren durchschnittlich mehr von den weniger schwerwiegenden, als von den schwerwiegenden Usability-Probleme entdecken. Virzis (1992) Forschungsbefunden zufolge finden die ersten fünf Testpersonen eines Usability Tests bereits alle Usability-Probleme, die als schwerwiegend klassifiziert wurden (siehe aber Lewis, 1994).

Bisher gibt es kaum Studien, die unterschiedliche Usability-Evaluationsmethoden direkt in einer Studie miteinander vergleichen. Dies ist jedoch essentiell, da sich alle Studien in verschiedenen Aspekten voneinander unterscheiden können. Eine Ausnahme stellt eine Studie von Jeffries und Kollegen (1991) dar. Nach den Ergebnissen dieser Methodenvergleichsstudie stellt die Heuristische Evaluation die effektivste Usability-Evaluationsmethode dar. Usability Tests kamen auf Platz zwei und anhand von Cognitive Walkthroughs ließen sich die wenigsten Usability-Probleme identifizieren. In einer aktuelleren Studie zur Untersuchung von inspektionsbasierten Methoden (Khajouei, Esfahani & Jahani, 2017) konnte dagegen kein signifikanter Unterschied hinsichtlich der Anzahl gefundener Usability-Probleme im Vergleich von Heuristischer Evaluation und Cognitive Walkthrough festgestellt werden. Karat, Campbell und Fliegel (1992) berichten in ihrer Studie, dass Testpersonen im Rahmen von Usability Tests eine signifikant höhere Anzahl an schwerwiegenden Usability Problemen gefunden haben im Vergleich zu Cognitive Walkthroughs.

Es lässt sich festhalten, dass es bisher noch keine eindeutige Antwort auf die Frage gibt, welche Usability-Evaluationsmethode am effektivsten ist und wie viele Personen für eine Usability-Evaluation notwendig sind. Dazu kommt, dass sich die User Interfaces in den letzten Jahren stark verändert haben und neue Produkte (z.B. VR-, Mixed Reality- oder Augmented Reality-Applikationen) entwickelt wurden. Für 3D-User Interfaces fehlen bislang noch Studien, in denen unterschiedliche Evaluationsmethoden im Hinblick auf ihre Effektivität und Effizienz verglichen werden. Im Nachfolgenden werden zunächst Evaluationsmethoden vorgestellt, die sich den neuen Herausforderungen von VR angenommen haben.

3 Usability-Evaluationsmethoden für VR

Gabbard (1997) war unter den ersten, die spezifische Evaluationsmethoden für VR-Applikationen entwarfen. Er entwickelte Usability-Klassifikationsschemata, die VR-spezifische Charakteristika berücksichtigen, wie z.B. die multidimensionale Objektselektion und Manipulationscharakteristika der virtuellen Welt sowie die Erfassung des Sense of Presence. Auf Grundlage seiner Taxonomie entwarf Gabbard 195 Design-Guidelines für die Praxis, deren Anwendung wenig ökonomisch erscheint. Stanney, Mollaghasemi, Reeves, Breaux und Graeber (2002) nutzten Gabbards Taxonomie zur Entwicklung ihres *Multicriteria Assessment of Usability for Virtual Environments* Systems (MAUVE). Dadurch kann ein breites Spektrum unterschiedlicher Usability- und VR-spezifischer Aspekte kategorisiert werden. Im Rahmen des computergestützten Bewertungsprozesses soll das zu evaluierende 3D User Interface mit einem Referenzmodell verglichen werden. Jedoch fehlt es an solchen etablierten Normen für VR, sodass die Aussagekraft der Ergebnisse und somit auch die Praxistauglichkeit limitiert sind (Domingues, Otmane, & Mallem, 2010).

Um die Nutzerperspektive zu erfassen, entwickelte Kalawsky (1999) einen Usability-Fragebogen: Der Fragebogen VRUSE ist ein Tool zur Bewertung der Usability eines VR-Systems basierend auf Feedback des Nutzers. Mit insgesamt 100 Items erscheint der Fragebogen jedoch nicht ökonomisch. Als effiziente Alternative entwickelten Sutcliffe & Gault (2004) 12 spezifische VR-Heuristiken. Die abstrakten Dimensionsbeschreibungen sollen Experten in der Kategorisierung der VR-Usability-Probleme unterstützen. Im Gegensatz zur Anwendung von Designprinzipien (z.B. Gabbard, 1997) soll die Nutzung der Heuristiken auch Auskunft über die Art und Ursache von vorliegenden Usability-Problemen geben. Sutcliffe & Gault (2004) empfehlen jedoch eine Überarbeitung der Heuristiken, da einige Dimensionen von Experten für VR nicht passend oder als nicht verständlich genug eingeschätzt wurden.

Ein globaler und jüngerer Ansatz zur Evaluation von 3D-User Interfaces wurde von LaViola und Kollegen (2017) entwickelt. Sie schlagen die Verwendung von drei Evaluationsmetriken vor. Zu den *System Performance Metrics* zählen Benchmarking Daten (z.B. durchschnittliche Bildrate oder Netzwerkverzögerungen). Die *Task Performance Metrics* erfassen die Qualität der Performance während einer Aufgabenbearbeitung durch den User (z.B. Navigationszeit, Fehlerrate). Schließlich sollen die *Subjective Response Metrics* die persönliche Wahrnehmung und Erfahrung des Nutzers während der Nutzung des Interfaces abbilden (z.B. User Experience, Usability, Motion Sickness, etc.).

Zwar liefert die Analyse von quantitativen Daten wie den Performanz-Metriken erste Informationen über mögliche Usability-Probleme (Bowman, Johnson, & Hodges, 1999; Gabbard, Hix, & Swan, 1999). Jedoch geben diese Daten keine direkte Auskunft über Art und Ursache von Usability-Problemen. Mit dem Einsatz von *Subjective Response Metrics* erhält der Evaluator zwar einen tieferen Einblick in das Nutzererleben, jedoch erfolgt die Bewertung der Erfahrung mit dem Interface retrospektiv. Das Berichten von Usability-Problemen und ihren Ursachen kann durch Gedächtniseffekte verzerrt werden. Bei traditionellen User Interfaces werden deshalb qualitative Methoden (z.B. *Thinking-Aloud)* eingesetzt. Damit können qualitativ hochwertige Informationen über die Ursachen von Usability-Problemen gewonnen werden (Nielsen, 1993). Bowman, Gabbard und Hix (2002) sprechen sich jedoch gegen das *Thinking-Aloud* bei der Evaluation von VR-Applikationen aus, da jede Interaktion zwischen Testleiter und Testperson den *Sense of Presence* stören könne.

Zusammenfassend halten wir fest, dass es den bisherigen Evaluationsmethoden für VR-Applikationen oftmals an ökonomischer Effizienz fehlt. Klassifikationsschemata wie die von Gabbard (1997) sind für den Einsatz in der Praxis zu zeit- und somit kostenaufwändig. Außerdem fehlen empirische

Validierungsstudien (Hale & Stanney, 2015). Bislang wurden zur Evaluation von VR-Applikationen quantitative Ansätze empfohlen, da die Annahme gemacht wurde, dass bei qualitativen Methoden wie z.B. dem Usability Test mit Thinking-Aloud der Sense of Presence—und somit ein zentrales Charakteristikum von VR-Applikationen—gestört wird. Diese Annahme ist jedoch bisher noch nicht empirisch auf ihre Gültigkeit hin untersucht worden. Darüber hinaus fehlen vergleichende Studien, welche die Eignung inspektionsbasierter gegenüber benutzerzentrierten Methoden im VR-Kontext untersuchen. Die Frage, wie viele Personen für die Evaluation von VR-Applikationen in Abhängigkeit von der Evaluationsmethode (inspektionsbasiert bzw. benutzerzentriert) nötig sind, wurde unseres Wissens nach noch nicht untersucht.

4 Methode der Studie

In der Studie, die in Kooperation mit der Beratung BLUPRNT GmbH durchgeführt wurde, untersuchen wir, wie viele Personen bei einer VR-Usability-Evaluation notwendig sind, um eine ausreichend hohe Anzahl an Usability-Problemen zu identifizieren. Wir haben dabei zwei verschiedene Evaluationsmethoden verglichen: die inspektionsbasierte Methode *Cognitive Walkthrough* und die benutzerzentrierte Methode *Usability Test*. Bei beiden Verfahren mussten die Personen während der Aufgabenbearbeitung laut denken. Wir untersuchen, wie viele Probleme (schwere und leichte) in Abhängigkeit von der Evaluationsmethode identifiziert werden können und ob sich Thinking-Aloud auf das Erleben von physischer und psychischer Präsenz im virtuellen Raum auswirkt.

4.1 Versuchspersonen

Insgesamt wurden 37 Versuchspersonen rekrutiert. Eine Person musste aufgrund technischer Probleme bei der Datenaufzeichnung ausgeschlossen werden. Die Gesamtstichprobe (N=36) besteht aus 20 Nutzern (M=32 Jahre, SD=6.32 Jahre, $Min.$=26, $Max.$=49, 56% weiblich) und 16 Experten (M=33 Jahre, SD=8.17 Jahre, $Min.$=22, $Max.$=48, 55% weiblich). Bis auf einen Experten haben alle Versuchspersonen angeben, die VR-Testanwendung noch nie zuvor verwendet zu haben.

4.2 Material

Für die Erhebung wurde das HMD Modell HTC Vive Pro inklusive HTC Vive Controller als Hardware eingesetzt. In der Trainingsphase wurde die Software StarWars verwendet und in der Testphase die Software Sharecare VR. Sharecare VR ist eine 3D Echtzeit-Simulation des menschlichen Körpers. Nutzer haben hier eine Auswahl an Organen und Funktionen exploriert. Die Nutzer konnten sich während der Interaktion mit der Software frei im Raum bewegen.

4.3 Evaluationsmethoden

Als Evaluationsmethoden kam eine benutzerzentrierte Methode (Usability Test) und eine inspektionsbasierte Methode (Cognitive Walkthrough) zum Einsatz. Vor der Testung haben wir einen Leitfaden mit nutzertypischen Aufgaben entworfen. Anhand dieser Aufgaben beschäftigte sich der Nutzer bzw. Experte mit der Software. Bei allen Erhebungen war ein Testleiter anwesend. Die Testpersonen der Usability Tests teilten ihre Erfahrungen mit der Software unter Einsatz der Thinking-Aloud-Methode dem Testleiter mit. Auch die Evaluatoren beim Cognitive Walkthrough, die in Einzeltestsessions die VR-Software evaluierten, wurden aufgefordert laut zu denken. Sie hatten die Aufgabe zu prüfen, inwiefern das System jeden durchgeführten Bearbeitungsschritt unterstützt. Die Aufgaben und der Ablauf der Testsession waren bei den beiden Evaluationsmethoden im Wesentlichen gleich. Der einzige Unterschied bestand in den Instruktionen des Testleiters, welcher die Experten explizit aufforderte, identifizierte Usability-Probleme zu benennen und zu beschreiben.

4.4 Ablauf

Der grundlegende Versuchsablauf war bei beiden Evaluationsmethoden identisch und teilt sich in drei Schritte: Zunächst wurden in einem Fragebogen (Fragebogen I) relevante Faktoren (u.a. Erfahrung mit VR) erfasst, die einen Einfluss auf die Bearbeitung der Testaufgaben haben könnten. Darauf folgte die Bearbeitung der Testaufgaben (Test-Session). Schließlich mussten die Versuchspersonen noch einen weiteren Fragebogen ausfüllen (Fragebogen II). Bei diesem Fragebogen ging es u.a. um die Erfassung einer möglichen Verletzung des Sense of Presence.

Die Dauer der Testsession der Nutzer betrug durchschnittlich 35 Minuten und die der Experten 38 Minuten. Davon verbrachten die Nutzer durchschnittlich 17 Minuten und die Experten 22 Minuten mit der VR-Applikation. Die Erhebungen wurden von drei erfahrenen Testleitern durchgeführt.

4.4.1 Fragebogen I

Die Mehrheit der Nutzer (n=20) gab an, Virtual Reality Anwendungen bisher noch nie (45.0%) oder ein- bis zweimal (30.0%) genutzt zu haben. Drei Nutzer gaben an, VR-Anwendungen schon mehr als zehn Mal genutzt zu haben. Zwei Nutzer hatten in ihrem beruflichen Kontext Kontakt zu VR-Anwendungen. Ihre Gaming-Erfahrung als Jugendliche schätzten die Nutzer auf einer fünfstufigen Likertskala durchschnittlich als hoch ein (M=3.75, SD=1.48), während die Gaming-Erfahrung im Erwachsenenalter als mittelhoch angegeben wurde (M=3.05, SD=1.5). Das medizinische Fachwissen wurde vom Durchschnitt der Nutzer als mittelhoch eingeschätzt (M=2.53, SD=0.74).

VR-Anwendungen wurden von 88.5% der Experten (n=16) mindestens einmal genutzt. Davon gaben 37.5% an, VR-Anwendungen zuvor ein- bis zweimal genutzt zu haben und 25.0% hatten VR Anwendungen drei- bis fünfmal zuvor ausprobiert. Über zehn Mal wurden VR-Anwendungen von 18.8% der Experten genutzt. Drei Personen hatten bereits in ihrem Arbeitsalltag mit VR-Anwendungen gearbeitet. Ihre Gaming-Erfahrung als Jugendliche schätzten die Experten auf einer fünfstufigen Likertskala durchschnittlich als mittelhoch ein (M=3.0, SD=1.2), auch die Gaming-Erfahrung im Erwachsenenalter (M=2.6, SD=0.9) und das medizinische Fachwissen (M=2.8, SD=1.0) wurde vom Durchschnitt der Experten als mittelhoch eingeschätzt.

4.4.2 Test-Session

Um den Erfahrungslevel der Testpersonen hinsichtlich VR-Applikationen anzugleichen, fand zunächst eine kurze Trainingsphase statt (Software StarWars). In der Testphase wurden die Versuchspersonen dann anhand eines Sets von nutzertypischen Aufgaben durch die VR-Anwendung geführt. Mit Hilfe des ersten Aufgabenblocks machten sich die Testteilnehmer vertraut mit den Inhalten der VR-Applikation, dem Hauptmenü und den Controller-Funktionen. Die Aufgabenschwierigkeit erhöhte sich im zweiten und dritten Aufgabenblock. Eine Aufgabe bestand z.B. darin herauszufinden, wie man sich mit der Software das Gehirn von Innen anschauen kann.

4.4.3 Fragebogen II

Angelehnt an die 12 VR-Heuristiken von Sutcliffe & Gault (2004) wurden 16 Fragebogenitems entworfen. Jede Dimension der 12 Heuristiken wurde durch ein bis zwei Items abgebildet. Beispielsweise wurde *Sense of Presence* durch folgendes Item erfasst: „Ich bin voll und ganz in die virtuelle Welt eingetaucht." Zusätzlich wurden zwei Items des *Motion Sickness Assessment Questionnaire* Fragebogens (MSAQ) erhoben, um das Erleben von Übelkeit während der VR-Erfahrung zu erfassen (Gianaros, Muth, Mordkoff, Levine, & Stern, 2001).

4.5 Design

Die unabhängigen Variablen sind die Evaluationsmethode (Cognitive Walkthrough vs. Usability Test) und die Anzahl der Personen (Evaluatoren bzw. Nutzer). Der Anteil identifizierter Usability-Probleme (in Abhängigkeit von der Schwere des Problems) ist die abhängige Variable. Die über die gesamte Stichprobe hinweg ermittelten Usability-Probleme bildeten dabei die Baseline. Dementsprechend wurde der Anteil identifizierter Probleme für jede Bedingung bestimmt, indem die Anzahl der gefundenen Probleme durch die Anzahl aller gefundenen Probleme geteilt wurde.

4.6 Datenerhebung und Kodierung

Zur Dokumentation wurden die Testsessions mit Screenrecording und Tonaufnahme aufgezeichnet. Durch Auswertung dieser Daten wurde für jede Versuchsperson erfasst, welche Usability-Probleme aufgetreten sind. In der Stichprobe der Experten wurden Usability-Probleme nur dann kodiert, wenn sie von dem Experten verbalisiert worden waren.

Im Anschluss an die Bestimmung aller aufgetretenen Usability-Probleme kodierte ein Usability Professional (der die Software kannte, selbst aber nicht an der Erhebung und Auswertung teilgenommen hatte) den Schweregrad der Probleme. Dabei wurde zwischen Usability-Problemen mit geringem Schweregrad (leichte Usability-Probleme) und hohem Schweregrad (schwerwiegende Usability-Probleme) unterschieden.

5 Ergebnisse

Die im Fragebogen I erhobenen Variablen VR-Erfahrung, Gaming-Erfahrung und medizinisches Fachwissen zeigten keine signifikanten Zusammenhänge mit der Anzahl gefundener Usability-Probleme. Diese Variablen wurden deshalb in die weiteren Analysen nicht aufgenommen.

5.1 Identifizierte Usability-Probleme

Bei den Erhebungen traten insgesamt 48 verschiedene Usability-Probleme auf. Davon wurden 24 als schwerwiegende und 24 als leichte Usability-Probleme kodiert. Bei den Usability Tests wurden 28 der 48 Usability-Probleme entdeckt (58.3%). Unter den 48 Usability-Problemen wurden zwei Usability-Probleme von nur einem Nutzer gefunden. Jeder Nutzer fand durchschnittlich sechs Usability-Probleme (M=5.7, SD=2.2). Die Experten fanden insgesamt 42 der 48 verschiedenen Usability-Probleme (87.5%). Unter den 48 Usability-Problemen wurden elf nur von einem Experten gefunden (22.9%). Jeder Experte fand durchschnittlich acht Usability-Probleme (M=7.7, SD=3.3).

Eine 2 (Issue-Schweregrad: gering vs. hoch) × 2 (Stichprobe: Nutzer vs. Experten) Varianzanalyse ergab sowohl einen Haupteffekt für den Schweregrad der Usability-Probleme (F[1,34]=16.794, p<.001, partielles η^2=.170) als auch einen Haupteffekt für die Stichprobe (F[1,34]=4.90, p=.034, partielles η^2=.078). Schwerwiegende Probleme wurden eher erkannt als weniger schwerwiegende Probleme. Außerdem wurde beim Cognitive Walkthrough ein größerer Anteil an Problemen identifiziert als beim Usability Test. Die Interaktion der beiden Variablen Evaluationsmethode und Issue-Schweregrad wurde nicht signifikant (F[1,43]=0.485, p=.491).

Im Weiteren haben wir den Zusammenhang zwischen der Stichprobengröße und dem Anteil identifizierter Probleme untersucht. Dazu haben wir in einer Simulationsstudie aus der Gruppe der Experten bzw. Nutzer Teilstichproben unterschiedlicher Größe gezogen (zwischen einer Person und 20 Personen; Ziehen mit Zurücklegen; 1000 Ziehungen). Abbildung 1 zeigt den durchschnittlichen Anteil identifizierter Usability-Probleme mit einem hohen und niedrigen Schweregrad in Abhängigkeit von der Stichprobengröße und der Gruppe (Experte vs. Nutzer). Aus der Abbildung geht hervor, dass die Kurven der Experten stärker ansteigen als die Kurven der Nutzer. Bis zu einer Stichprobengröße von ca. 10 Personen ist ein bedeutsamer Anstieg ersichtlich. Dann sind die Kurven schon relativ flach,

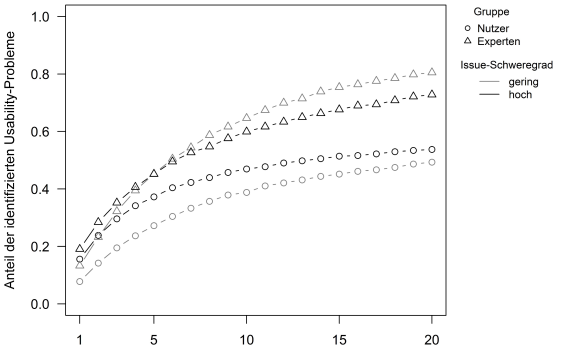

Abbildung 1: Anteil identifizierter Usability-Probleme in Abhängigkeit vom Schweregrad und von der Stichprobengröße.

wobei die Kurve der Experten noch länger ansteigt als die der Nutzer. Es zeigt sich außerdem die Tendenz, dass bei Gruppen von bis zu vier Experten ein höherer Anteil an schwerwiegenden, als weniger schwerwiegenden Usability-Probleme gefunden wird. Dieser Verlauf kehrt sich ab einer Stichprobengröße von fünf Experten um. Dann wird von Experten-Gruppen ein größerer Anteil weniger schwerwiegender als schwerwiegender Probleme identifiziert. In Usability Tests wird dagegen sowohl bei kleineren als auch bei größeren Stichproben ein größerer Anteil schwerwiegender als weniger schwerwiegender Probleme aufgedeckt.

Im Folgenden werden weitere Analysen berichtet, die separat für Usability-Probleme mit geringem bzw. hohem Schweregrad durchgeführt wurden (siehe auch Tabellen 1 bis 4).

5.2 Identifizierte Usability-Probleme mit geringem Schweregrad

Anhand einer Stichprobengröße von 20 Experten werden durchschnittlich 80.9% der leichten Usability-Probleme identifiziert (siehe Tabelle 1). Jedoch besteht das Risiko, dass aufgrund der breiten Spanne des Anteils der identifizierten Usability-Probleme (41.7% bis 91.7%) weniger als die Hälfte der leichten Usability-Probleme identifiziert werden. Mit einer gleich großen Stichprobe an Nutzern können durchschnittlich die Hälfte der leichten Usability-Probleme identifiziert werden und mindestens 29.2% (siehe Tabelle 2).

Ab einer Stichprobengröße von 16 Experten bzw. acht Nutzern steigt der durchschnittliche Anteil identifizierter Usability-Probleme mit jeder zusätzlichen Person lediglich um maximal zwei Prozent.

5.3 Identifizierte Usability-Probleme mit hohem Schweregrad

Bei einer Stichprobengröße von 20 Experten werden durchschnittlich 72.4% der schwerwiegenden Usability-Probleme identifiziert (siehe Tabelle 3). Mit einer ähnlich großen Stichprobengröße an Nutzern werden durchschnittlich 53.3% der schwerwiegenden Usability-Probleme identifiziert (siehe Tabelle 4). Die Range des Anteils der mindestens und maximal gefundenen schwerwiegenden Usability-Probleme ist bei den Experten deutlich größer (45.8% bis 83.3%), als bei den Nutzern (41.2% bis 58.3%). Ab einer Stichprobengröße von 13 Experten bzw. acht Nutzern steigt der durchschnittliche Anteil identifizierter Usability-Probleme mit jeder zusätzlichen Person lediglich um maximal zwei Prozent.

Tabelle 1: Anteil der identifizierten Usability-Probleme mit geringem Schweregrad in der Stichprobe der Experten

Stichprobengröße	*M*	*SD*	*Min*	*Max*
5	46.0%	1.4%	8.3%	75.0%
10	63.4%	2.5%	25.0%	91.7%
15	74.6%	10.2%	37.5%	91.7%
20	80.9%	8.3%	41.7%	91.7%

Tabelle 2: Anteil der identifizierten Usability-Probleme mit geringem Schweregrad in der Stichprobe der Nutzer

Stichprobengröße	*M*	*SD*	*Min*	*Max*
5	27.7%	8.7%	4.2%	54.2%
10	39.2%	8.0%	12.5%	58.3%
15	45.3%	6.8%	20.8%	58.3%
20	49.1%	6.4%	29.2%	58.3%

Tabelle 3: Anteil der identifizierten Usability-Probleme mit hohem Schweregrad in der Stichprobe der Experten

Stichprobengröße	*M*	*SD*	*Min*	*Max*
5	44.9%	9.3%	12.5%	66.7%
10	59.8%	2.5%	29.2%	83.3%
15	67.9%	8.1%	41.7%	83.3%
20	72.4%	7.4%	45.8%	83.3%

Tabelle 4: Anteil der identifizierten Usability-Probleme mit hohem Schweregrad in der Stichprobe der Nutzer

Stichprobengröße	*M*	*SD*	*Min*	*Max*
5	37.1%	6.4%	16.7%	54.2%
10	47.2%	9.3%	33.3%	58.3%
15	51.0%	4.7%	37.5%	58.3%
20	53.3%	4.0%	41.2%	58.3%

Tabelle 5: Fragebogen-Skalen *Natürliche Interaktion, Ausdruck natürlichen Handelns* und *Sense of Presence*

Skala	*M*	*SD*	*Min*	*Max*
Natürliche Interaktion	3.8	1.0	1	5
Ausdruck natürlichen Handelns	3.7	1.0	2	5
Sense of Presence	4.4	0.7	3	5

5.4 Fragebogen II

Nur eine Person (in der Usability Test Gruppe) gab an, unter Übelkeit zu leiden. Ein Ausschluss dieser Person änderte das Ergebnismuster nicht.

Außerdem untersuchten wir die Frage, ob VR-spezifische Heuristiken wie Sense of Presence durch die Methode des Thinking-Aloud negativ beeinflusst wurden. Dies scheint nicht der Fall zu sein, zumal die Versuchspersonen auf der Skala *Sense of Presence* eine hohe Ausprägung hatten (M=4.4, auf Likert-Skala von 1 bis 5; siehe auch Tabelle 5). Auch auf den Skalen *Natürliche Interaktion* und *Ausdruck natürlichen Handelns* wiesen die Personen Werte oberhalb des Skalenmittels von 3 auf. Im Hinblick auf die drei Skalen gab es auch keine signifikanten Unterschiede zwischen den beiden Evaluationsmethoden (alle p-Werte > .181).

6 Diskussion

Virtual Reality (VR) ist keine Zukunftsutopie mehr, doch fehlt es bis heute an geeigneten Usability-Evaluationsmethoden. Zwar wurden erste Ansätze der Evaluation entwickelt (z.B. Gabbard, 1997; Kalawsky, 1999; Stanney, Mollaghasemi, Reeves, Breaux & Graeber, 2002; Sutcliffe & Gault, 2004). Jedoch wurden diese Ansätze bisher nicht empirisch validiert. So ist es bisher auch eine offene Frage, wie viele Personen nötig sind, um einen Großteil der Usability-Probleme einer VR-Applikation zu identifizieren. Die Forschung zu traditionellen User Interfaces hat untersucht, wie viele Personen zur Identifikation von 80% der Usability-Probleme notwendig sind (z.B. Nielsen, 1992, 1993; Faulkner, 2003; Lewis, 1994; Spool & Schroeder, 2002; Virzi, 1992). Jedoch besteht wenig Konsistenz zwischen den verschiedenen Studienergebnissen und es ist unklar, ob die Zahlen auf VR-Applikationen zu übertragen sind. In unserer experimentellen Studie haben wir die praxisrelevante Frage untersucht, wie viele Personen für die Testung von VR-Applikationen notwendig sind, um einen Großteil der Usability-Probleme zu identifizieren.

Die Ergebnisse zeigen, dass ein größerer Anteil von Usability-Problemen mit der Methode des Cognitive Walkthroughs gefunden wurde als mit Usability Tests. Eine Simulationsstudie ergab, dass 20 Experten notwendig sind, um durchschnittlich 80.9% der Usability-Probleme mit geringem Schweregrad und 72.4% der Usability-Probleme mit hohem Schweregrad zu identifizieren. Es werden deutlich mehr Nutzer als Experten benötigt, um einen ähnlich hohen Anteil an Usability-Problemen zu finden. Mit einer Stichprobengröße von 20 Nutzern können durchschnittlich 49.1% der Usability-Probleme mit geringem bzw. 53.3% der Usability-Probleme mit hohem Schweregrad identifiziert werden. Die Ergebnisse stehen im Einklang mit der jüngeren Forschungsliteratur zur Evaluation von 2D-Interfaces (z.B. Faulkner, 2003; Lewis, 1994; Spool & Schroeder, 2001), welche die Annahme von Nielsen (1992, 1993) und Virzi (1992), es brauche lediglich fünf Evaluatoren bzw. Testpersonen, in Frage stellen. Ähnlich wie bei Spool & Schroeder (2001) wurden in unserer Studie bei einer Stichprobengröße von fünf Nutzern durchschnittlich 27.7% der leichten bzw. 37.1% der schweren Usability-Probleme gefunden. Fünf Experten finden durchschnittlich 46.0% der leichten bzw. 44.9% der schweren Usability-Probleme.

Die sich aus unserer Studie ergebenden notwendigen Stichprobengrößen übersteigen vielleicht auch deshalb die aus der älteren Literatur üblichen Empfehlungen, da VR eine neue Komplexitätsdimension mit sich bringt. Wir empfehlen somit, bei der Evaluation von VR-Applikationen zunächst größere Testgruppen einzuplanen. Jedoch ergab unsere Studie auch, dass selbst bei sehr großen Stichprobengrößen nicht alle Usability-Probleme identifiziert werden können. So zeigte unsere Simulationsstudie, dass es keinen sehr großen Unterschied macht, ob acht oder 20 Nutzer an einem Usability Test teilnehmen. Beim Cognitive Walkthrough kann man durch die Erhöhung der Zahl der Evaluatoren dagegen eher einen Anstieg in dem Anteil gefundener Probleme erreichen.

Ein weiteres Ziel unserer Studie bestand darin, die Anwendbarkeit der Thinking-Aloud-Methode im VR-Evaluationsprozess zu prüfen. Bowman, Gabbard und Hix (2002) haben sich gegen jegliche Interaktion zwischen Testleiter und Testperson während des Evaluationsprozesses ausgesprochen, um den Sense of Presence und das damit einhergehende Immersionsgefühl nicht zu stören. Unsere Versuchspersonen berichteten dagegen trotz der Nutzung der Thinking Aloud-Methode auf der Skala Sense of Presence hohe Ausprägungen (Mittelwert von 4.4 und Minimum von 3 auf der Skala von 1 bis 5). Es ist zwar möglich, dass der Sense of Presence bei einer Evaluationsmethode ohne Thinking-Aloud noch höher bewertet werden würde, aber unsere Studienergebnisse zeigen, dass das Erleben von Presence bei unseren Evaluationsmethoden nicht stark beeinträchtigt war.

6.1 Empfehlungen für die Usability-Praxis

Unsere Studie ergab, dass—anders als bisher angenommen—die Methode des Thinking-Aloud durchaus auch bei der Evaluation von VR-Applikationen eingesetzt werden kann, ohne dass dadurch der Sense of Presence verloren geht. Darüber hinaus zeigt sich, dass bei der Evaluation von VR-Applikationen durch die Anwendung des Cognitive Walkthrough-Verfahrens eine höhere Identifikationsrate von Usability-Problemen im Vergleich zu den Usability Tests erzielt werden kann. Wenn das Evaluationsziel darin besteht, in erster Linie Usability-Probleme eines hohen Schweregrads zu identifizieren, implizieren unsere Ergebnisse, dass Experten durchschnittlich ca. 10% mehr an schwerwiegenden Usability-Problemen identifizieren.

Aufgrund der großen Spanne zwischen dem Mindestanteil und dem Maximalanteil an identifizierten Usability-Problemen können nur vage Empfehlungen für geeignete Stichprobengrößen ausgesprochen werden. Die nachfolgenden Empfehlungen beruhen auf den Mittelwerten der Anteile identifizierter Usability-Probleme. Folgt man der verbreiteten Stichprobenempfehlung von fünf Personen pro Usability Evaluation (u.a. Nielsen, 1992, 1993; Virzi, 1992), dann würden fünf Nutzer unseren Ergebnissen zufolge durchschnittlich ca. 37% und fünf Experten durchschnittlich ca. 45% der Usability-Probleme identifizieren. Da ab einer Stichprobengröße von acht Nutzern lediglich ein geringer Anstieg in dem durchschnittlichen Anteil gefundener Usability-

Probleme zu erkennen ist (ca. 2% pro zusätzlicher Person), bringt eine Erhöhung der Nutzerzahl nur bedingt etwas. Beim Cognitive Walkthrough hingegen ist mit der Erhöhung der Anzahl der Evaluatoren bis einer Stichprobengröße von 13 Experten ein Anstieg im Prozentsatz gefundener Probleme zu erreichen (ab 13 Experten ca. 2% pro zusätzlicher Person). In der Praxis kommt aus Kosten-, Zeit- und Verfügbarkeitsgründen häufig nur ein Experte bei Produktevaluationen zum Einsatz. Die Ergebnisse zeigen jedoch, dass ein einzelner Experte zwischen 8.3% und 29.2% der schwerwiegenden Usability Probleme identifizieren kann. Diese breite Spanne legt nahe, dass die Auswahl des Evaluators entscheidend für die Qualität des Cognitive Walkthrough sein könnte.

6.2 Grenzen und Ausblick

Was die Vergleichbarkeit der Ergebnisse der Experten- und Nutzergruppe einschränkt ist, dass die Experten durchschnittlich fünf Minuten länger mit der Testsoftware interagierten. Mehr Zeit mit der Software könnte eine höhere Identifikationsrate von Usability-Probleme begünstigt haben. Eine Gewichtung des Anteils identifizierter Probleme an der aufgewendeten Zeit wäre denkbar. Jedoch ist unklar, welcher Zusammenhang zwischen den beiden Variablen Zeit und Anteil identifizierter Probleme besteht, so dass auch unklar ist, wie diese Gewichtung sinnvollerweise erfolgen müsste.

Eine Limitation unserer Studie besteht außerdem in der geringen Stichprobengröße (N=36). Schmettow und Vietze (2008) folgend möchten wir darüber hinaus auf die große Varianz in den Identifikationsraten innerhalb der Testgruppen und damit auf potentielle personenbezogene Einflussfaktoren hinweisen.

Ein Forschungsvorhaben zukünftiger Studien könnte die Analyse der Eigenschaften sein, die einen guten Evaluator ausmachen. Zusätzlich stellt sich die Frage, ob diese Fähigkeiten des Evaluators durch Trainings verbessert werden können. Ferner haben die Versuchspersonen unserer Studie eine Lernsoftware getestet, also nur eine spezifische Form von VR-Applikationen. Die Generalisierbarkeit der Ergebnisse auf andere Personengruppen und VR-Applikationen (z.B. Spiele, Simulationen, Trainings, etc.) ist somit eingeschränkt. In zukünftiger Forschung könnte der Einfluss der Fähigkeiten und Erfahrung der Usability-Experten (u.a. VR-Expertise) sowie der Einfluss der Art der VR-Software auf die Identifikationsraten von Usability-Problemen untersucht werden. In weiterer Forschung wäre es zudem sinnvoll, eine unabhängige Baseline-Erfassung der Usability-Probleme durch mehrere Usability-Experten (anders als beim Cognitive Walkthrough ohne Zeitbegrenzung) vorzunehmen.

Zusammenfassend lassen unsere Ergebnisse vermuten, dass acht bis 20 Testpersonen (Experten bzw. Nutzer) notwendig sind, um im Rahmen von VR Evaluationen eine zufriedenstellende Anzahl an schwerwiegenden Usability-Problemen zu identifizieren. Jedoch zeugen solche Stichprobengrößen in der Usability-Praxis von wenig wirtschaftlicher Attraktivität. Der Empfehlung von Nielsen (2000) folgend kann ein zukünftiges Forschungsvorhaben darin bestehen, den Einsatz von kleineren Stichprobengrößen zu mehreren Testzeitpunkten im Designprozess zu untersuchen. Wiederholtes Testen könnte auch den großen Schwankungen in der Anzahl identifizierter Usability-Probleme entgegenwirken. Alternativ könnte die Kombination verschiedener Evaluationsmethoden näher erforscht werden. Aufgrund des kontroversen Forschungsfeldes bedarf es mehr Forschung, die bestehende Ergebnisse zu einem Konsens zusammenführt, der valide und handlungsleitende Empfehlungen für die Praxis geben kann.

LITERATURVERZEICHNIS

[1] Bouvier, P., Lavoué, E. & Sehaba, K. (2014). Defining Engagement and Characterizing Engaged- Behaviors in Digital Gaming. Simulation and Gaming, *SAGE Publications*, 45 (4-5), 491-507.

[2] Bowman, D.A., Gabbard, J.L. & Hix, D. (2002). A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. *Teleoperators and Virtual Environment*, 11, 404-424.

[3] Bowman, D.A., Johnson, D.B., Hodges, L.F. (1999). Testbed evaluation of virtual environment interaction techniques. *Proceedings: ACM Symposium on Virtual Reality Software and Technology*, New York: ACM.

[4] Caulton, D.A. (2001). Relaxing the homogeneity assumption in usability testing. Behaviour & Information Technology, 20 (1), 1–7.

[5] Dey, A., Billinghurst, M., Lindeman, R. & Swan J.E. (2018). A Systematic Review of 10 Years of Augmented Reality Usability Studies: 2005 to 2014. *Frontiers in Robotics and AI*, 5 (37), 1-28.

[6] Domingues, C., Otmane, S. & Mallem, M. (2010). 3dui-ef: Towards a framework for easy empirical evaluation of 3d user interfaces and interaction techniques. *The International Journal of Virtual Reality, 9* (1), 73-80.

[7] Dünser, A., Grasset R. & Billinghurst, M. (2008). A Survey of Evaluation Techniques Used in Augmented Reality Studies. Technical Report TR-2008-02, HitLabNZ, University of Canterbury, New Zeland.

[8] Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35 (3), 379-383.

[9] Gabbard, J. L. (1997). A taxonomy of usability characteristics in virtual environments. Doctoral dissertation, Virginia Tech University.

[10] Gabbard, J.L., Hix, H. & Swan, J.E. (1999). User-centered design and evaluation of virtual environments. *IEEE Computer Graphics and Applications*, 19, 51-59.

[11] Gianaros, P.J., Muth, E.R., Mordkoff, T.J., Levine, M.E. & , Stern, R.M. (2001). A questionnaire for the assessment of the multiple dimensions of motion sickness. *Aviation, Space, and Environmental Medicine* 72, 2, 115.

[12] Grosvenor, L. (1999). Software usability: Challenging the myths and assumptions in an emerging field. Unpublished master's thesis, University of Texas.

[13] Hale, K.S. & Stanney, K.M. (2015). Handbook of Virtual Environments. Design, Implementation, and Applications. CRC Press, Traylor & Francis Group: Boca Raton.

[14] Hwang, W. & Salvendy, G. (2010). Number of People required for Usability Evaluation: The 10 +/- 2 Rule. *Communications of the ACM*, 53 (5), 130-133.

[15] International Society for Presence Research (2000). Presence defined. Abgerufen am 11. April 2019 von https://ispr.info/about-Presence-2/about-Presence/.

[16] Jeffries, R., Miller, J.R., Wharton, C., and Uyeda, K.M. (1991). User interface evaluation in the real world: A comparison of four techniques. Proceedings ACM CHI-91 Conference, 119-124.

[17] Karat C., Campbell R, Fiegel T (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Monterey: ACM.

[18] Khajouei, R., Esfahani, M.Z. & Jahani,Y. (2017). Comparison of heuristic and cognitive walkthrough usability evaluation methods for evaluating health information systems. *Journal of the American Medical Informatics Association*, 24 (1), 55-60.

[19] Kalawsky, R.S. (1999). VRUSE: a computerised diagnositc tool for usability evaluation of virtual/synthetic environment systems. *Applied Ergonomics*, 30 (1), 11–25.

[20] LaViola, J.J., Kruijff, E., McMahan, R.P. Bowman, D.A. & Poupyrev, I. (2017). 3D User Interfaces. Theory and Practice. Second Edition. Boston: Addison-Wesley.

[21] Lewis, J.R (1994). Sample sizes for usability studies: Additional considerations. *Human Factors* 36 (2), 368-378.

[22] Molich, R. & Nielsen, J. (1990). Heuristic evaluation of user interfaces. *ACM CHZ'90*, 249–256.

[23] Nielsen, J. (1989). Usability engineering at a discount. In G. Salvendy and M. J. Smith (Hrsg.), *Designing and using human-computer interfaces and knowledge-based systems* (S. 394-401). Amsterdam: Elsevier.

[24] Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Proceedings ACM CHI'92 Conference*, 373-380.

[25] Nielsen, J. (1993). Usability Engineering. London: LP Professional Ltd.

[26] Nielsen, J. (1994). Executive summary. In J. Nielsen & R. L. Mack (Hrsg.), *Usability Inspection Methods* (S. 1–23). New York: John Wiley & Sons.
[27] Nielsen, J. (2000). Why you only need to test with 5 Users. Abgerufen von https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/
Sarodnick, F. & Brau, H. (2016). Methoden der Usability Evaluation. Wissenschaftliche Grundlagen und praktische Anwendung. Bern: Hogrefe Verlag.
[28] Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1535), 3549–3557.
[29] Schmettow, M. & Vietze, W. (2008). Introducing Item Response Theory for measuring usability inspection processes. *Proceeding of SIGCHI conference on Human factors in computing systems, ACM*, 893–902.
[30] Schmettow, M. (2012). Sample size in usability studies. *Communications of the ACM* 55 (4), 64-70.
[31] Spool, J. & Schroeder, W (2001). Testing Web sites: Five users is nowhere near enough. *CHI Extended Abstracts on Human Factors in Computing Systems, ACM*, 285–286.
[32] Stanney, K.M., Mollaghasemi, M., Reeves, L., Breaux, R. & Graeber, D.A. (2002). Usability engineering of virtual environment (VEs): identifying multiple criteria that drive effective VE system design. *Int. J. Human-Computer Studies*, 58, 447-481.
[33] Sutcliffe, A. & Gault, B. (2004). Heuristic evaluation of virtual reality applications. *Interacting with Computers*, 16, 831-849.
[34] Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34, 457-468.
[35] Witmer, B. & Singer, M. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence*, 7 (3), 225–240.
[36] Woolrych,A., & Cockton,G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Hrsg.), *Proceedings of IHM-HCI 2001 Conference*, 2, (S.105- 108). Toulouse, France: Cépadèus.