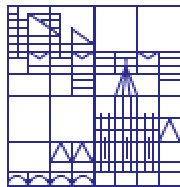


Jens Gerken

Validität und Aussagekraft von Usability Test Methoden

Ausarbeitung „Seminar zum Projektpraktikum“
Wintersemester 2002 / 2003



Universität Konstanz
FB Informatik und Informationswissenschaft
Arbeitsgruppe Mensch-Computer Interaktion

1. Einleitung	3
2. Methoden der Sozialwissenschaft	4
2.1 Das Quantitative Experiment	5
2.1.1 Das Laborexperiment	5
2.1.2 weitere Varianten des quantitativen Experiments	7
2.1.3 Validität und Aussagekraft quantitativer Experimente	7
2.2 Das Qualitative Experiment	12
2.2.1 Geschichte des qualitativen Experiments	12
2.2.2 Unterschiede zum quantitativen Experiment	13
2.3 Qualitative Methoden	14
2.3.1 Die Marienthalstudie	14
2.3.2 Teilnehmende Beobachtung	15
2.3.3 Das Interview	16
2.3.4 Gruppendiskussionen	17
2.3.5 Validität und Aussagekraft qualitativer Methoden	18
2.4 Methoden der Sozialwissenschaft - Fazit	20
3. Usability Test Methoden	21
3.1 Klassifizierung von Usability Test Methoden	22
3.1.1 Participatory Evaluation	22
3.1.2 Diagnostic Evaluation	25
3.2 Usability Test Methoden - Fazit	31
4. Schlussfolgerung	32
5. Referenzen	33

1. Einleitung

Validität und Aussagekraft von Usability Testmethoden – ein Thema das in der Software Entwicklung oft nur eine untergeordnete Rolle spielt. Der wohl ausschlaggebende Grund hierfür ist, dass sich Usability Tests noch immer nicht überall durchgesetzt haben und in vielen Unternehmen noch um einen festen Platz im Entwicklungsprozess gerungen werden muss. Um diesen zu erreichen, müssen oftmals Kompromisse geschlossen werden – die Erstellung eines nach wissenschaftlichen Kriterien validen Testsettings bleibt hier oftmals schon zu Beginn auf der Strecke. Usability Testing ist teuer und ob sich diese Investition jemals rechnet ist oftmals nicht sofort zu erkennen. Die Investition von zusätzlichen Mitteln in die Forschung nach validen Testsettings und dadurch die Möglichkeit die Aussagekraft eines Usability Tests auch nachweislich zu gestalten, ist zwar ein wünschenswertes Ziel – für die Entwicklung der Software an sich aber oftmals nur von zweitrangiger Bedeutung. Müssen Tests hier zum Beispiel wirklich reproduzierbar sein, wenn die Software doch ständig weiterentwickelt wird und sich somit das Objekt des Tests ständig ändert? Reicht es nicht aus, Usability Fehler aufzudecken und dann auszumerzen?

Andererseits muss, zumindest am Ende einer Entwicklung, auch ein quantitativer Vergleich mit Vorgänger- oder Konkurrenzprodukten möglich sein. Sei es für Marketingzwecke oder um überhaupt die Entwicklung einer neuen Version zu rechtfertigen. Hier spielt die Validität und auch die Aussagekraft, also die Übertragbarkeit in die wirkliche Welt, der Ergebnisse eine deutlich größere Rolle.

Gute Literatur zu diesem Thema aus dem HCI (Human Computer Interaction) Bereich ist nur schwer zu finden. Glücklicherweise sind die Methoden nach denen Usability Tests ablaufen keine Erfindung der letzten Jahre – was bedeuten würde, dass sie zwangsläufig noch unausgereift wären. Vielmehr werden hier Methoden adaptiert die man sehr ähnlich in den Sozialwissenschaften wieder findet. Der wohl größte Unterschied findet sich in dem Gegenstand des Experiments/Tests. Während die Sozialwissenschaften hier, zumindest meistens, klar den Menschen im Mittelpunkt der Untersuchung haben, tritt dieser bei Usability Tests einen Schritt zur Seite um dem Softwareprodukt Platz zu machen. Nichts desto trotz lassen sich viele Parallelen erkennen und auch in bezug auf Validität und Aussagekraft der Ergebnisse, liefert die Sozialwissenschaft bereits ausgereifte und wissenschaftlich fundierte Anleitungen.

Im folgenden werde ich aus diesem Grund zunächst eine Einführung in die Methoden der Sozialwissenschaft geben – mit besonderem Augenmerk auf das quantitative und das qualitative Experiment und anschließend die Brücke zu den Usability Testmethoden schlagen und versuchen die Gemeinsamkeiten und Unterschiede herauszustellen.

2. Methoden der Sozialwissenschaft

Nach Kleining (1986) gibt es in der Sozialwissenschaft drei verschiedene Formen der Methodik. Diese unterscheiden sich voneinander durch das Abstraktionsniveau. Angefangen bei den Alltagsmethoden, über die qualitativen Methoden, hin zu den quantitativen Methoden.

Die Methoden begründen sich auf Techniken, welche die Menschheit im Laufe der Geschichte entwickelt haben. Sie sind hierbei Formen der Auseinandersetzung des Forschers mit dem beobachteten Gegenstand – also Subjekt – Objekt Beziehungen. Unterschiede lassen sich vor allem in bezug auf das oben bereits angesprochene Abstraktionsniveau feststellen.

Die Alltagsmethoden sind grundsätzlich eigentlich keine Methoden. Sie umfassen das tägliche handeln (erlebe und tue) und sind demnach in ihrer ursprünglichen Form auch nicht wissenschaftlich. Diese Form erreichen sie erst, in dem man reflektiert handelt, also versucht sich gleichzeitig in eine Betrachterposition zu versetzen und damit den ersten Schritt zur Objektivität macht. Man erreicht dadurch bereits einen gewissen Abstraktionsgrad. Nach Kleining (1986) kann dieser Vorgang als Festlegung oder Definition von Alltagserfahrungen bezeichnet werden. Alltagserfahrungen werden hierdurch zu Methoden.

Ausgehend davon können in der Sozialwissenschaft zwei weitere Abstraktionsniveaus erreicht werden – wie bereits genannt unterscheidet man hier nun zwischen qualitativen Methoden und quantitativen Methoden. Hierbei erreichen die qualitativen Methoden ein niedrigeres Niveau als die quantitativen, was oftmals dazu führt, dass letztere als „wissenschaftlicher“ angesehen werden. Nach Kleining (1986) ist dieser Schluss aber falsch, da die Wissenschaftlichkeit nicht durch das Abstraktionsniveau bestimmt wird, sondern durch das Wahrheitskriterium, welches nicht direkt an eine der drei Formen gebunden ist. Alle drei Methoden lassen sich noch weiter untergliedern - jeweils in Beobachtung und Experiment.

In bezug auf die Form des Experiments, welches uns hinsichtlich der Usability Testing Methoden am meisten interessiert, bezeichnet Kleining das quantitative Experiment und das qualitative Experiment als wissenschaftlich, welche als Basis das Alltagsexperiment haben. Angewendet werden kann es prinzipiell auf alle Gegenstände, mit denen sich die Sozialforschung befasst. Hierbei sind mit Gegenständen natürlich auch, beziehungsweise vornehmlich, Personen(gruppen) gemeint.

Nach diesem kurzen Überblick möchte ich nun einen detaillierten Einblick in die Methodik des quantitativen Experiments und anschließend die des qualitativen Experiments geben. Weiterhin werde ich auch einige qualitative Methoden aufzeigen, die nicht direkt dem qualitativen Experiment untergeordnet werden können, jedoch in bezug auf ähnliche Usability Verfahren sehr interessant sind.

2.1 Das Quantitative Experiment

Das quantitative Experiment kann auf eine lange Tradition zurückblicken und gilt dementsprechend als weitestgehend erforscht. Grundlage des quantitativen Experiments ist die Hypothesenprüfung. Es wird also ähnlich wie bei statistischen Tests zunächst eine Hypothese aufgestellt. Durch das Experiment soll diese dann überprüft werden.

Prinzipiell geht es meistens darum, einen kausalen (Ursache-Wirkung) Zusammenhang festzustellen, wobei hier beachtet werden muss, nicht versehentlich eine Korrelation einem Kausalschluss gleichzusetzen. Beispielsweise ist es bekannt, dass in Gegenden, in denen viele Störche nisten, auch die Geburtenrate überproportional hoch ist. Es besteht also eine positive Korrelation zwischen den Störchen und der Geburtenrate. Dennoch wäre hier ein Kausalschluss unangebracht. Die Störche können nach wissenschaftlichen Gesichtspunkten kaum die Ursache für die erhöhte Geburtenrate sein. Vielmehr könnte es eine dritte Variable geben, die auf beide Einfluss nimmt. Eine Möglichkeit hierfür wäre der Urbanisierungsgrad, da in ländlichen Gebieten sowohl die Geburtenrate, als auch die Wahrscheinlichkeit, dass Störche nisten (in der Stadt quasi ja unmöglich) höher ist.

In diesem Fall wäre der Urbanisierungsgrad die unabhängige Variable und die Störche beziehungsweise die Geburtenrate die abhängige Variable. Ein quantitatives Experiment versucht im Idealfall einen kausalen Zusammenhang zwischen genau einer unabhängigen und einer abhängigen Variable zu zeigen. Dies wird dadurch erreicht, dass durch die Veränderung der unabhängigen Variablen ein Effekt nachgewiesen wird, sich die abhängige Variable also „von allein“ verändert. Um diese Veränderung dann eindeutig der unabhängigen Variablen zuschreiben zu können, müssen beide Variablen vorher isoliert werden. Es müssen also jegliche Dritt- oder Störvariablen (z.B. Telefon klingelt, etc.) entweder komplett aus dem Versuchsaufbau ausgeschlossen werden oder falls das nicht oder nur schwer möglich ist, in das Design integriert werden.

Ebenfalls muss bei der Erstellung der Hypothese darauf geachtet werden, dass die unabhängige Variable im Rahmen des Experiments überhaupt variiert werden kann.

Als letztes sei hier noch die Wiederholbarkeit genannt, die bei einem quantitativen Experiment auf der „Zwingend erforderlich“ - Liste steht.

Es gibt verschiedene Formen des quantitativen Experiments, die ich im folgenden zunächst beschreiben möchte. Anschließend werde ich dann die Probleme hinsichtlich der Validität und Aussagekraft herausstellen.

2.1.1 Das Laborexperiment

Das verbreitetste, quantitative Experiment ist wohl das Laborexperiment, da hierbei die Kontrolle von Drittvariablen und das Ziel der Wiederholbarkeit am einfachsten zu erreichen sind. Im Folgenden möchte ich an einem konkreten Beispiel die Vorgehensweise bei einem Laborexperiment schildern.

2.1.1.1 Das Laborexperiment - Ablauf an einem praktischen Beispiel

Ein Forscher stellt die Vermutung auf, dass Menschen die rauchen, in Stresssituationen einen erhöhten Zigarettenkonsum aufweisen. Als Hypothese formuliert:

„In Stresssituationen erhöht sich bei Rauchern der Zigarettenkonsum“

Um hier Missverständnisse zu vermeiden, müssen die verwendeten Begriffe eindeutig definiert und entweder als unabhängige oder abhängige Variable deklariert werden. In diesem Fall könnte man in Hinsicht auf das Testsetting (im Anschluss) die folgenden Definitionen wählen:

<i>Stresssituation:</i>	Eine Person, die unter zeitlichem Druck mathematische Aufgaben lösen muss.
<i>Zigarettenkonsum:</i>	Anzahl der gerauchten Zigaretten in einem festgelegten Zeitraum
<i>Raucher:</i>	Eine Person, die täglich (also regelmäßig) zwischen 15 und 20 Zigaretten raucht. Die genauen Zahlen sind hierbei nicht wichtig, solange sie vorher festgelegt werden.
<i>Abhängige Variable:</i>	Zigarettenkonsum
<i>Unabhängige Variable:</i>	Stresssituation

Es werden nun aus dem Pool der Versuchspersonen – dessen Auswahl im Optimalfall aus der Grundgesamtheit der Raucher stammt – 2 Gruppen gebildet. Die Experimentalgruppe und die Kontrollgruppe. An der zuerst genannten wird nachher die unabhängige Variable verändert, wohingegen die Kontrollgruppe diese Änderung nicht erfahren wird. Die beiden Gruppen werden in getrennte, aber ansonsten möglichst identische Räume geführt. Für jede Versuchsperson steht ein Tisch bereit. Auf diesem Tisch befinden sich neben Schreibmaterial auch Getränke, Aschenbecher, Feuerzeuge und die bevorzugte Zigarettenmarke in ausreichender Menge. Die Versuchspersonen werden nicht darüber informiert, dass der Zigarettenkonsum der eigentliche Gegenstand des Experiments ist.

Zunächst wird nun ein erster Testlauf gefahren. Hierbei werden beiden Gruppen mathematische Aufgaben ausgehändigt. Es wird kein zeitlicher Druck aufgebaut – jeder darf in der festgelegten Zeit so viele Aufgaben behandeln wie er möchte. Im Idealfall wird den Versuchspersonen auch dieses Limit verschwiegen, um den möglichen Ehrgeiz einiger Personen, alle Aufgaben in der Zeit zu lösen, zu bremsen. Das Limit sollte zudem so gewählt werden, dass es unmöglich ist, in der Zeit alle Aufgaben zu lösen. Nach Ablauf der Zeit werden in beiden Gruppen die gerauchten Zigaretten gezählt.

Nach einer kurzen Ruhepause wird der Test wiederholt. Nun jedoch mit einer Änderung: Während die Kontrollgruppe abgesehen von anderen Aufgaben genau den gleichen Test noch einmal macht, wird bei der Experimentalgruppe die unabhängige Variable Stress variiert. Bisher sollte diese, wenn möglich, nicht vorhanden sein. Es wird nun also eine künstliche Stresssituation aufgebaut. Die Versuchspersonen bekommen wie vorher auch mathematische Aufgaben ausgehändigt. Dieses Mal werden sie angewiesen, alle Aufgaben in einer festgelegten Zeit zu lösen. Eventuell muss hierbei noch ein Druck- beziehungsweise Motivationsmittel zusätzlich eingesetzt werden – beispielsweise einen Preis für denjenigen, der es als erstes schafft – wobei solche zusätzlichen Faktoren auch verfälschend wirken

können. Zum Beispiel wird nicht jeder Preis bei jeder Versuchsperson die gleiche Motivation wecken.

Nach Ablauf der Zeit werden nun in beiden Gruppen wieder die gerauchten Zigaretten gezählt. Wenn dabei in der Experimentalgruppe im Verhältnis zum ersten Testdurchlauf mehr Zigaretten geraucht wurden als bei der Kontrollgruppe, kann die Hypothese vorerst als bestätigt gelten.

Ein Laborexperiment zeichnet sich vor allem durch die Künstlichkeit der Versuchsanordnung aus. Durch diese künstliche Situation ist eine Kontrolle von Dritt- und Störvariablen optimal möglich. Dementsprechend dürfen sich die Kontroll- und Experimentalgruppe nur durch die Variation der unabhängigen Variable in der Experimentalgruppe unterscheiden. Dadurch lässt sich am Ende ein eindeutiger Kausalschluss ziehen.

2.1.2 weitere Varianten des quantitativen Experiments

A Feldexperiment: Der Ort der Untersuchung wird hierbei aus dem Labor hin zu dem angestammten Platz des zu untersuchenden Gegenstandes, also in dessen natürlicher Umgebung verlegt. Ansonsten wird auch hier versucht in dieser natürlichen Umgebung eine unabhängige Variable zu verändern um einen Kausalschluss zu erreichen und eine Hypothese zu prüfen. Aufgrund oftmals nur schwer zu kontrollierender äußerer Einflüsse, ist dies hier aber ungleich schwerer.

B Simultanexperiment: Es werden mehrere Gruppen gleichzeitig untersucht beziehungsweise beeinflusst.

C Sukzessives Experiment: Es existiert keine Kontrollgruppe – lediglich eine Gruppe wird sowohl vor als auch nach der Veränderung der unabhängigen Variable untersucht. In manchen Fällen kann eine Kontrollgruppe überflüssig oder nicht zwingend notwendig sein.

2.1.3 Validität und Aussagekraft quantitativer Experimente

Worin liegen nun die Schwierigkeiten bei quantitativen Experimenten, welche Probleme können auftreten und wie valide sind solche Tests?

2.1.3.1 Repräsentativität der Ergebnisse

Hier ist zunächst auf die Auswahl der Versuchspersonen von Bedeutung. Im optimalsten Fall stellt die Auswahl einen verkleinerten Ausschnitt der jeweiligen Grundgesamtheit dar. Mit Grundgesamtheit ist im allgemeinsten Fall die Menge aller Menschen gemeint. Da diese im Normalfall aber nur bedingt von Interesse ist, wird der Begriff Grundgesamtheit auch für Submengen verwendet, die für den Test relevant sind. In obigem Beispiel ist dies die Menge der Raucher, die durchschnittlich 15-20 Zigaretten am Tag rauchen. Unsere Auswahl an Versuchspersonen sollte somit, wenn möglich, ein repräsentativer Schnitt der 15-20 Zigaretten pro Tag rauchenden Bevölkerung sein. Die Definition der Grundgesamtheit entscheidet also, für welche Personengruppe das Ergebnis letztendlich signifikant ist. Oftmals ist es jedoch nicht möglich, allein durch ein Merkmal hier einschränkend vorzugehen. Selbst wenn wir die Auswahl noch lokal, auf zum Beispiel eine Stadt, begrenzen würden, wäre es

kaum möglich eine Gruppe von Rauchern zu wählen, die, bis auf die geringere Anzahl, der gewählten Grundgesamtheit in allen Facetten gleicht. Da letztendlich auch nur Freiwillige teilnehmen können, wird die Auswahl weiter erschwert. In der Praxis führt dies dazu, das oftmals Versuchspersonen dort gesucht werden, wo Gruppen von Menschen anzutreffen sind. Studenten, Soldaten und Arbeiter in großen Betrieben sind hierbei beliebte Pools. Die Verteilung innerhalb der Versuchspersonen entspricht dann jedoch nur noch bedingt der vorher definierten Grundgesamtheit.

Was heißt das nun für die Repräsentativität der Ergebnisse? Grundsätzlich sind ohne weitere Untersuchungen die Ergebnisse eines Experiments aus oben genannten Gründen meistens nicht repräsentativ für die entsprechende Grundgesamtheit, dürfen also nicht verallgemeinert werden. Da diese Situation natürlich das ganze Experiment ad absurdum führen würde, können wie bereits angedeutet weitere Überlegungen und Untersuchungen angestellt werden, um eine Übertragbarkeit doch zu ermöglichen. Zunächst muss untersucht werden, inwieweit sich die Versuchspersonen von einer wirklich repräsentativen Auswahl unterscheiden. Anschließend muss untersucht werden, ob diese Unterschiede Einfluss auf das Ergebnis haben. Hierbei kann unterschieden werden, ob der aufgetretene Effekt nur in seiner Höhe differieren könnte, oder ob er in der definierten Grundgesamtheit so eventuell gar nicht auftreten könnte.

In ersterem Fall können die Ergebnisse zwar auf die Grundgesamtheit ausgedehnt werden, jedoch nur als relative Aussage. In unserem Beispiel also: „Raucher die sich in Stresssituationen befinden, haben einen erhöhten Zigarettenkonsum.“ Was „erhöht“ bedeutet, darf und kann hierbei nicht absolut definiert werden!

Für den Fall, dass der Verdacht besteht, dass der entdeckte Effekt so bei der Grundgesamtheit gar nicht auftreten könnte, besteht die einzige Möglichkeit, die Ergebnisse doch noch verwertbar zu machen darin, die Grundgesamtheit weiter einzugrenzen. Es müssen hierbei die Merkmale klassifiziert werden, die dazu führten, dass die Ergebnisse so in einer wirklich repräsentativen Auswahl nicht auftreten. Anschließend kann dadurch eine neue, vermutlich aber kleinere Grundgesamtheit definiert werden, auf die die Ergebnisse übertragbar sind.

Letztendlich sind für die Übertragbarkeit aber noch weitere Faktoren ausschlaggebend wie zum Beispiel die externe Validität auf die ich unter anderem im Folgenden eingehen möchte.

2.1.3.2 Validität des Experiments

Die Validität bezeichnet die Gültigkeit eines Experiments. Misst die Untersuchung/das Instrument wirklich das, was es messen soll? Sind die Ergebnisse verwendbar? Um hier eine Antwort zu erhalten wird der Begriff zunächst weiter differenziert in interne Validität und externe Validität.

A interne Validität

Die interne Validität gibt Auskunft darüber, ob der beobachtete Effekt wirklich eindeutig der Veränderung der unabhängigen Variablen zugeschrieben werden kann. Sie ist also für die Kontrolle der internen Versuchssituation zuständig. Je künstlicher ein Versuchsaufbau ist (im Optimalfall Laborexperiment), desto einfacher lassen sich eventuelle Störeinflüsse kontrollieren und desto höher ist die interne Validität.

B externe Validität

Die externe Validität gibt Auskunft darüber, inwieweit sich der beobachtete Effekt auf die Grundgesamtheit generalisieren lässt. Eine hohe externe Validität stellt also sicher, dass die gemessenen Werte nicht nur innerhalb des Versuchsaufbaus zustande kommen, sondern auch in der realen Welt auftreten. Es geht hierbei vor allem darum, inwieweit der Versuchsaufbau die Ergebnisse beeinflusst und somit eine eins zu eins Übertragung in die wirkliche Welt verhindert.

C interne Validität contra externe Validität

Externe und interne Validität sind eng miteinander verknüpft. Zwischen ihnen herrscht eine Art Antagonismus. Eine hohe interne Validität ist zwar wünschenswert um einen eindeutigen Kausalschluss zu ermöglichen, jedoch wird dadurch die Situation auch unnatürlich, was dazu führt, dass die Ergebnisse nicht mehr ohne weiteres in die Realität außerhalb der Versuchsanordnung übertragen werden können. Kurz gesagt also: *Je höher die interne Validität desto niedriger die externe Validität und umgekehrt.*

D Stör- und Drittvariablen

Wie bereits mehrfach gesagt, können Störfaktoren das Ergebnis beeinflussen. Der Punkt dabei ist, dass es quasi unmöglich ist, Störvariablen vollkommen aus dem Testsetting zu verbannen. Vielmehr liegt oftmals das Ziel darin, die Störvariablen konstant zu halten und dadurch in das Testsetting zu integrieren (passiv). Eine aktive Integration ist in manchen Fällen auch möglich und fördert zumeist die Natürlichkeit des Experiments, erhöht also die externe Validität. Allerdings wird es dadurch schwerer die Ergebnisse zu interpretieren, da oftmals kein eindeutiger Kausalschluss mehr zulässig ist. Durch standardisierte Testabläufe und kleinere Testgruppen können Störvariablen beziehungsweise deren Einfluss weiter reduziert werden.

Das Auftreten von Drittvariablen ist äußerst ungünstig. Der Fehler liegt hierbei meistens in der Aufstellung der Hypothese, wenn bei der Identifizierung der unabhängigen und abhängigen Variablen eine vermutete Kausalität mit einer Korrelation verwechselt wird.

Allerdings sind monokausale Zusammenhänge auch äußerst selten. Somit existiert fast immer zumindest eine weitere Drittvariable, die Einfluss auf die abhängige Zielvariable nimmt. Die Hypothese muss aus diesem Grund mit sehr viel Bedacht gewählt werden, um möglichst schon von vorneherein Drittvariablen die nicht ausgeschlossen werden können, direkt zu integrieren. Ein gänzlicher Ausschluss dieser würde zwar die interne Validität erhöhen, eine Integration im Gegensatz dazu jedoch die externe. Es ist nicht immer leicht hier einen angemessenen Kompromiss zu finden!

E Demand Characteristics & Forced Exposure

“*Demand characteristics*“ bezeichnet eine Problematik, die die externe Validität beeinflussen kann. Es geht hierbei weniger um die Künstlichkeit beispielsweise einer

Laborsituation sondern vielmehr darum, dass die Versuchspersonen im Allgemeinen wissen, dass sie an einem Experiment teilnehmen. Aus ethischen Gesichtspunkten sollten sie das auch, jedoch kann es ihr Verhalten beeinflussen. Oftmals verhalten sich Versuchspersonen in Testsituationen anders als in ihrem normalen Umfeld. Sie versuchen eventuell das „wahre“ Versuchsziel zu erkennen – selbst wenn dieses ihnen mitgeteilt wurde.

Ebenfalls wichtig für die externe Validität ist das sogenannte *“forced exposure“*. Diese Problematik wird in der Literatur nur sehr selten behandelt, stellt sie doch das Grundprinzip des Experiments in Frage. Während eines Versuchs können die Versuchspersonen letztendlich nur den Anweisungen des Versuchsleiters Folge leisten. In unserem Beispiel müssen sie zum Beispiel mathematische Aufgaben lösen, auch wenn sie dazu vielleicht gar keine Lust haben und im Moment lieber was anderes tun würden. Der einzige Ausweg besteht im Abbruch des Experiments – allerdings scheuen sich viele diesen Weg zu gehen um zum Beispiel vor einer Gruppe nicht als Versager dazustehen. In gewissem Sinne kann zumindest bei einigen Versuchspersonen die Situation eintreten, dass sie sich zu dem Experiment gezwungen fühlen obwohl sie sich freiwillig gemeldet haben. In unserem Beispiel tritt dieser Punkt nicht so stark auf – in einer realistischen Stresssituation kann man meistens auch nicht mehr frei entscheiden, was man machen möchte. Nichts desto trotz ist es sehr schwierig, dieser Problematik gerecht zu werden, da zum Beispiel eine Lockerung des Versuchsablaufs unkontrollierbare Störvariablen mit ins Spiel bringen würde.

F Zufällige und Systematische Fehler

Zufällige und systematische Fehler sind eng mit dem Auftreten von Störvariablen verbunden. Die Gefahr besteht darin, dass eine Störvariable eventuell unentdeckt bleibt oder einfach nicht zu verhindern ist. Falls es sich um Einzelversuche handelt (die Versuchspersonen also Experimental und Kontrollgruppe zugeordnet werden, aber einzeln/isoliert das Experiment absolvieren), kommt es im Normalfall zu zufälligen Fehlern. Das heißt, dass zwar ein solcher Fehler das Ergebnis in seiner Genauigkeit beeinflusst, es aber nicht in eine bestimmte Richtung verändert. Je mehr zufällige Fehler auftreten, desto ungenauer werden die Ergebnisse, eventuell kann dann sogar ein vorhandener Unterschied zwischen Kontroll- und Experimentalgruppe verdeckt werden.

Zu systematischen Fehlern kommt es, wenn die Versuchspersonen in Gruppen getestet werden. Tritt hier eine Störung auf, beeinflusst das unter Umständen sofort die gesamte Gruppe. Die Ergebnisse verändern sich dementsprechend nicht nur in der Genauigkeit sondern sie werden verzerrt. Somit verringern systematische Fehler die interne Validität eines Experiments. Je kleiner die Gruppen sind, desto geringer wird die Auswirkung des Störfaktors auf das Ergebnis. Im besten Fall ergibt dies dann also wieder Einzelversuche und die damit verbundenen zufälligen Fehler. Allerdings sind Einzelversuche oftmals aus wirtschaftlichen Gründen nicht realisierbar, da sie zu viel Zeit in Anspruch nehmen.

G Lern- und Reifungseffekte

Selbst wenn die finanziellen Mittel und die Zeit für Einzeltests zur Verfügung steht, können diese nicht vorbehaltlos eingesetzt werden. Da sie zwangsläufig über einen längeren Zeitraum stattfinden müssen, sind sie anfällig für sogenannte Lern- und Reifungseffekte. Diese können sowohl beim Versuchsleiter als auch bei den Versuchspersonen auftreten. Der Versuchsleiter bekommt über die Zeit mehr Routine und passt sein am Anfang vielleicht noch etwas unsicheres Auftreten mit der Zeit an. Versuchspersonen die an einem späteren Test teilnehmen, fühlen sich deshalb vielleicht besser aufgehoben oder einfach nicht so unwohl, wenn der Versuchsleiter ein sympathisches Auftreten an den Tag legt. Die Versuchspersonen gehen demgegenüber vielleicht nicht mehr unvorbelastet in das Experiment. Eventuell haben sie von Bekannten, die bereits daran teilgenommen haben, Details erfahren und wissen somit schon, was auf sie zu kommt. Hier kommen dann beispielsweise die oben angesprochenen „demand characteristics“ zum tragen.

H Versuchsleitereffekte

Wie eben bereits angemerkt können die Versuchsleiter ebenfalls das Experiment beziehungsweise die Ergebnisse beeinflussen. Je nach Art des Experimentes hat der Versuchsleiter die Aufgabe den Versuchspersonen den Test zu erklären, die Messungen durchzuführen, Störvariablen auszugrenzen, den Versuchspersonen die Aufgabe zu präsentieren und für eventuelle Hilfestellungen oder Rückfragen zur Verfügung zu stehen. Sie sind somit das einzige Bindeglied zwischen Experiment und Versuchsperson. Dadurch besteht aber auch die Gefahr, dass die Versuchsleiter die Versuchspersonen in irgendeiner Hinsicht beeinflussen und somit das Ergebnis systematisch verzerren. Versuchsleitereffekte lassen sich grundsätzlich in drei Kategorien unterteilen:

- Effekte die auf physischen oder sozialen Merkmalen des Versuchsleiters basieren. Hierzu können beispielsweise ein ausgeprägter Dialekt oder aber auch physische Attraktivität (oder auch das Gegenteil) gehören.
- Lern und Reifeffekte wie bereits im vorigen Abschnitt beschrieben
- Effekte die sich aus den persönlichen Erwartungen des Versuchsleiters an das Experiment ergeben. Eventuell vermutet oder erhofft der Versuchsleiter bereits ein bestimmtes Ergebnis der Messungen und versucht unbewusst die Versuchspersonen dahingehend zu beeinflussen, dass dieses auch eintritt.

Die Lösung dieser Probleme ist nicht ganz einfach und letztendlich nicht ohne Kompromiss möglich. Eine Möglichkeit besteht darin, den Ablauf des Experiments weitgehend zu standardisieren, zum Beispiel auch dahingehend, dass die Instruktionen für die Versuchspersonen schriftlich ausgehändigt werden. Kurz gesagt, dass der Kontakt zwischen Versuchsleiter und Versuchsperson auf ein Minimum reduziert wird. Hierdurch steigt jedoch wieder die Künstlichkeit der Situation und die Teilnehmer fühlen sich eventuell unwohl.

Lern- und Reifeffekte können durch intensive Schulungen der Versuchsleiter minimiert werden. Dadurch gewinnen diese die nötige Erfahrung um von Anfang an, zum Beispiel auf Nachfragen der Teilnehmer, richtig zu reagieren.

Um die persönliche Bindung des Versuchsleiters mit dem Ziel des Experiments zu verhindern, sollte dieser im Idealfall dieses gar nicht vollständig kennen. Er sollte

somit auch mit der Forschung an sich nichts zu tun haben sondern im Idealfall ein externer Experte sein.

Versuchsleitereffekte sind aber allgemein ein sehr großes Problem, dem vor allem zumeist zu wenig Beachtung geschenkt wird.

1 Ethische Bedenken

Ethische Bedenken bei quantitativen Experimenten sind keinesfalls von der Hand zu weisen. Es gibt einige unrühmliche Beispiele in der Geschichte, die deutlich machen, dass es oftmals ein Drahtseilakt sein kann, was moralisch vertretbar ist und was nicht (Beispiel „Stanford Prison Experiment“, verschiedene Gehorsamkeitsexperimente). Letztendlich muss man sich immer bewusst sein, dass Menschen keine Laborratten sind und man dementsprechend auch nicht ihre persönliche Freiheit und Selbstbestimmung einschränken darf. Fragwürdig werden Experimente beispielsweise immer dann, wenn die Versuchspersonen über den wahren Zweck hinweggetäuscht werden. Hier sollte spätestens am Ende ein sogenanntes „debriefing“ erfolgen, das den Teilnehmern das wahre Ziel enthüllt.

Abschließend ist zu sagen, dass die Validitätsprobleme weitestgehend erforscht sind und auch Lösungsmöglichkeiten existieren. Jedoch muss jede Lösung teuer erkaufte werden – es gibt kein Geheimrezept für das perfekte Experiment, welches absolut repräsentative Ergebnisse liefert und nicht angezweifelt werden kann.

2.2 Das Qualitative Experiment

Das qualitative Experiment ist in vielem der genaue Gegensatz des quantitativen Experiments. Bevor ich die Unterschiede aber näher beleuchte, möchte ich einen kurzen Einblick in die Geschichte geben.

2.2.1 Geschichte des qualitativen Experiments

Zunächst muss gesagt werden, dass qualitative Methoden und insbesondere das qualitative Experiment in den letzten drei Jahrhunderten keinen allzu großen Stellenwert hatten. Dabei kann man die Anfänge des qualitativen Denkens bis auf Aristoteles zurückführen, wobei damals jedoch keine strikte Trennung zwischen den Methoden erfolgte. Es ist also falsch zu sagen, Aristoteles sei ein Vertreter der rein qualitativen Forschung gewesen. Vielmehr hat er sowohl explorative (qualitative) als auch demonstrative (quantitative) Experimente geführt. Der letztendliche Wandel dieses Forschens wird oftmals an Galilei festgemacht, der das neue Denken vertrat, dass sich alles auf allgemeine Naturgesetze reduzieren ließe. Auch hier muss aber ganz klar gesagt werden, dass trotz dieser Ansicht viele Experimente dieser Zeit auch stark explorativen Charakter hatten. Als Beispiel sei hier die Entdeckung der Fallgesetze durch Newtons Experimente mit Pendel und schiefer Ebene genannt. Auch die im 17. Jahrhundert vorangetriebene Suche nach dem „Stein der Weisen“ in der Alchemie förderte durch ihren explorativen Charakter bedeutende Zufallsentdeckungen zu Tage. Der Bruch zwischen qualitativer und quantitativer Forschung kann vielmehr zu Beginn des 19. Jahrhunderts festgemacht werden. Anlass war hierzu die strikte Trennung von Physik und Anschauung. Goethe stellte beispielsweise seine nicht naturwissenschaftlichen

Experimentierformen den physikalischen Experimenten Newtons klar gegenüber. Somit muss man seit der Romantik qualitative und quantitative Experimente gesondert und in ihrem Verhältnis zueinander betrachten (Kleining 1986).

Wieder aufgegriffen wurde die qualitative Methodik letztendlich von Wundt und ausschlaggebend für das qualitative Experiment von Ernst Mach. Dieser hat den Begriff 1905 auch geprägt. Bis hin zum zweiten Weltkrieg bauten auf seinen Methoden unter anderem die Würzburger Denkexperimente und die Gestaltpsychologie auf. Nach dem Krieg nahm der Umfang des quantitativen Experiments in der Sozialwissenschaft derart zu, dass kein Platz mehr für andere Formen war. Auch wenn das quantitative Experiment oft in der Kritik stand, verschwand der Begriff des qualitativen Experiments nahezu vollständig aus der Literatur. Erst der radikale Gesinnungswandel weg vom Rationalen hin zu dem Menschen an sich brachte qualitative Methoden wieder ins Gespräch. Letztendlich gilt Gerhard Kleining als derjenige, der 1986 das qualitative Experiment wieder zum Leben erweckte. Seitdem schwinden die Konflikte zwischen qualitativer und quantitativer Forschung zusehends dahin. Auch wenn das qualitative Experiment noch immer einen schweren Stand hat, so konnten sich doch zumindest qualitative Methoden weithin etablieren und in eine weitgehend friedliche Koexistenz mit ihren quantitativen Konterparten treten.

2.2.2 Unterschiede zum quantitativen Experiment

Kleining definiert das qualitative Experiment als den nach wissenschaftlichen Regeln vorgenommenen Eingriff in einen sozialwissenschaftlichen Gegenstand, mit dem Ziel, dessen Struktur zu erforschen. Es ist die explorative, heuristische Form des Experiments (Kleining 1994). Im Gegensatz zum quantitativen Experiment soll kein Kausalschluss zwischen einer unabhängigen und einer abhängigen Variable festgestellt werden und die dort geforderte Wiederholbarkeit widerspricht seinem Charakter. Vielmehr zielt es auf die Aufklärung und Erforschung von Strukturen. Es werden hierbei nicht nur kausale Zusammenhänge erforscht sondern jegliche Art von Abhängigkeiten, Beziehungen und Relationen können von Bedeutung sein. Ein standardisierter Versuchsaufbau ist hierbei hinderlich, da dieser den zu erforschenden Gegenstand nicht als Ganzes erfassen kann. Letztendlich ist nicht das Überprüfen von Hypothesen der Zweck des qualitativen Experiments, sondern vielmehr können Hypothesen aus den Ergebnissen, die es liefert, formuliert werden.

Das qualitative Experiment stellt also einen aktiven Eingriff in einen Gegenstand dar. Dabei wird darauf geachtet, dass dieser Eingriff nicht beliebig erfolgt sondern möglichst gegenstandsadäquat vorgenommen wird. Es existieren verschiedene Eingriffstechniken auf die ich hier aber nicht näher eingehen möchte (Separation, Kombination, Reduktion, Transformation um einige Beispiele zu nennen). Wichtig ist, dass nach jedem Eingriff der Gegenstand möglichst genau beschrieben wird, um Gemeinsamkeiten und Veränderungen festzustellen und somit auf seine Struktur Rückschlüsse zu erhalten.

Der grobe Ablauf eines qualitativen Experiments sieht demnach wie folgt aus:

- Beschreibung des Gegenstands
- Experimenteller Eingriff
- Beschreibung des Gegenstands
- Schlussfolgerung auf die Struktur

In der Regel ist der Ablauf iterativ aufgebaut, das heißt die Schritte 2 und 3 werden mehrfach durchlaufen.

Bei der Vorbereitung und Durchführung des Experiments müssen noch einige grundlegende Dinge beachtet werden:

- Der Forscher muss sein eigenes Vorverständnis über das Experiment als vorläufig und revidierbar ansehen.
- Dies gilt auch in Bezug auf den Gegenstand, der untersucht wird. Seine Struktur ist erst am Ende eindeutig bestimmt
- Die Eingriffe in den Gegenstand müssen bei den iterativen Durchläufen maximal variiert werden

2.3 Qualitative Methoden

Auch wenn das qualitative Experiment in dieser, sehr theoretisch definierten Reinform immer noch kaum Verwendung findet, gibt es doch eine große Anzahl qualitativer Methoden, die zum Teil die hier beschriebene Vorgehensweise aufgreifen. Zu diesen konkreten Verfahren existieren auch in Bezug auf Validität und Aussagekraft genauere Untersuchungen. Im folgenden möchte ich einige Verfahren genauer vorstellen.

2.3.1 Die Marienthalstudie

Die Marienthalstudie ist ein Beispiel qualitativer Forschung aus den 30er Jahren. Explizit beschrieben wurde sie unter anderem 1975 von Jahoda, Lazarsfeld und Zeisel und 1990 von Mayring. Da in dieser Studie eine Vielzahl von qualitativer Methoden Verwendung fand, möchte ich sie etwas ausführlicher beschreiben.

Marienthal war der Name einer kleinen österreichischen Ortschaft, dessen Bevölkerung fast ausschließlich in einer ansässigen Textilfirma beschäftigt war. Diese nahm ab 1930 jedoch Massenentlassungen vor, was die Situation im Ort bezüglich Arbeitslosigkeit deutlich verschlechterte. Mehrere Forscher begaben sich in dieser Zeit nach Marienthal um Materialien zur psychosozialen Lage der Bevölkerung zu sammeln. Sie überschritten dabei jedoch von Anfang an die Grenze reiner Beobachtung – sie griffen aktiv in das Leben der Menschen ein und versuchten dieses durch verschiedene Aktionen zu verbessern. Die Forscher gingen hierzu sehr sorgsam vor und versuchten sich langsam in das Gemeinschaftsleben zu integrieren. Insgesamt verbrachten sie 120 Arbeitstage in Marienthal. Die Aktionen dienten, neben der Hilfe die sie beinhaltenen, vornehmlich dazu, den Kontakt zu den Menschen herzustellen. Einige Beispiele für durchgeführte Aktionen:

- Es wurde eine Kleidersammelaktion durchgeführt. In Wien wurden etwa 200 Kleidungsstücke gesammelt, ausgebessert und an die Marienthaler verteilt
- Die Forscher engagierten sich in politischer Mitarbeit aktiv in den ansässigen Vereinen und Verbänden
- Ein Schnittzeichenkurs wurde kostenlos angeboten – etwa 50 Frauen besuchten diesen Kurs 2 mal wöchentlich
- Eine Frauen- und Kinderärztin hielt einmal in der Woche eine Sprechstunde ab, in die jeder kommen konnte.
- Für junge Mädchen wurde ein Turnkurs angeboten
- Eine Beratungsstelle wurde eingerichtet die bei Problemen mit der Erziehung und des häuslichen Lebens helfen sollte

Durch die vielen Gespräche konnte in dieser Zeit sehr viel Material gesammelt werden. Angefangen bei ausführlichen Lebensgeschichten der Bewohner, über den Tagesablauf von mehr als 80 Personen, hin zu der genauen Beschreibung der Mahlzeiten von 40 Familien und sogar die Beschreibung der Weihnachtsgeschenke von 80 Kindern. Es wurde weiterhin untersucht welche Bücher in der öffentlichen Bibliothek ausgeliehen wurden und was die vornehmlichen Gesprächsthemen und Beschäftigungen in öffentlichen Lokalen war.

Letztendlich wurde dieses Material auf Gemeinsamkeiten analysiert, um eine Klassifizierung zu ermöglichen hinsichtlich des Verhaltens in Bezug auf die Arbeitslosigkeit. Diese ergab folgendes Bild:

- *Ungebrochene Haltung*: Aufrechterhalten des Haushaltes, Pflege der Kinder, subjektives Wohlbefinden, Aktivität, Pläne und Hoffnungen für die Zukunft, aufrechterhaltende Lebenslust, immer wieder Versuche zur Arbeitsbeschaffung
- *Resignierte Grundhaltung*: gleichgültig, erwartungsloses Dahinleben, überzeugt dass man ja doch nichts gegen die Arbeitslosigkeit machen kann, aber relativ ruhige Grundstimmung
- *Verzweifelte Grundhaltung*: Verzweiflung, Depression, Hoffnungslosigkeit, jede Bemühung vergeblich → die Suche nach Arbeit bereits aufgegeben, letztendlich auch keine Versuche mehr die Situation zu verbessern, Schwelgen in den besseren Tagen der Vergangenheit
- *Apathische Grundhaltung*: Energie- und tatenlos, ungepflegte Wohnung und Kinder, häufig wechselnde Stimmung, keine Zukunftspläne, schlechte Hauswirtschaft, oftmals verbunden mit Alkoholismus, Zerfallserscheinungen innerhalb der Familie

2.3.2 Teilnehmende Beobachtung

Die zugrundeliegende Methodik der Marienthalstudie ist wohl die teilnehmende Beobachtung. Hierbei sieht sich der Forscher nicht in einer rein beobachtenden Position sondern nimmt selbst aktiv an der sozialen Situation teil, in die der zu untersuchende Gegenstand eingebettet ist. Es wird also eine persönliche Beziehung zu den Beobachteten aufgebaut – man nimmt an deren Leben teil, während man im Hintergrund Daten sammelt. Dadurch erhofft man sich genauere Einblicke in die Struktur des Gegenstands zu bekommen, die Innenperspektive erleben zu können. In manchen Fällen ist eine strukturelle Erschließung nur über diese Technik möglich.

Die Methodik der teilnehmenden Beobachtung ist nur teilweise standardisiert. Es sollte zwar ein Beobachtungsleitfaden erstellt werden und dieser sollte auch theoriegeleitet sein, der Beobachter/Forscher muss diesen Leitfaden aber weder immer parat haben, noch muss er ihn strikt befolgen. Vielmehr sollte er ihn verinnerlicht haben. Er dient dann vor allem für die Erstellung der Protokolle als Leitfaden um die Daten mehrerer Forscher leichter vergleichen zu können. Die Schwierigkeit einer teilnehmenden Beobachtung liegt darin, dass die Forscher einen Weg finden müssen, um sich in die soziale Umgebung integrieren zu können ohne abgelehnt oder als Störung empfunden zu werden. Nach Mayring (1990) ist diese Methode auch sehr gut geeignet um durch ihre explorative Form Hypothesen zu generieren – ähnlich dem qualitativen Experiment. Nach Kleining sind in der Marienthalstudie auch Formen qualitativen Experimentierens vorhanden.

2.3.3 Das Interview

Da die reine Beobachtung, sei sie auch teilnehmend, allein oft nicht ausreicht, wird zusätzlich auf eine weitere Form qualitativer Forschung zurückgegriffen – das Interview. Im Gegensatz zu Befragungen in der quantitativen Forschung, in der die Befragten oftmals nur die Auswahl aus mehreren möglichen Antworten haben, sind qualitative Interviews deutlich offener und unstrukturierter gestaltet. Sie sollen dem Befragten die Möglichkeit geben zu sagen, was er persönlich von einem Sachverhalt denkt. Der Unterschied zeigt sich auch in der Art der Befragung – während quantitative Befragungen meist schriftlich anhand von Fragebögen ablaufen, vertraut die qualitative Forschung auf die verbale Form.

Mittlerweile gibt es eine Vielzahl von Interviewtechniken, die meisten unterscheiden sich jedoch nur in dem Grad der Standardisierung. Somit können hier erst mal einige grundlegende Techniken und Verhaltensweisen vorgestellt werden. Prinzipiell wird in der Vorbereitungsphase zunächst der interessierende Gegenstandsbereich unter theoretischen Gesichtspunkten aufgearbeitet. Anschließend werden die Fragen formuliert, je nach verwendeter Methode mehr oder weniger konkret. Diese werden in einem sogenannten Interviewleitfaden zusammengestellt.

Die nun folgende Pilotphase, in der einige Testpersonen interviewt werden, soll neben der Sicherstellung Qualität der Fragen vor allem dazu dienen, dem Interviewer die notwendige Erfahrung für die anschließenden „richtigen“ Befragungen geben. Für die Erstellung der Fragen und des Leitfadens können einige grundsätzliche Dinge beachtet werden:

- Die Fragen sollten verständlich, möglichst einfach und nicht zu lang formuliert sein.
- Die Fragen dürfen die befragte Person nicht überfordern, dementsprechend muss darauf geachtet werden, dass kein Wissen implizit vorausgesetzt wird, welches der Befragte eventuell gar nicht aufweist.
- Die Fragen dürfen nicht suggestiv sein, also nicht von vorneherein eine bestimmte Antwort nahe legen.
- Die Eingangsfragen sollten möglichst leicht und unkontrovers zu beantworten sein, um dem Befragten den Einstieg zu erleichtern
- Fragen haben oftmals einen Ausstrahlungseffekt auf die nachfolgenden Themen, zum Teil können diese überflüssig werden oder müssen während des Interviews abgeändert werden – je nach Technik kann hier mehr oder weniger flexibel gehandelt werden.

Im folgenden möchte ich zwei Interviewtechniken, das strukturierte Interview und das problemzentrierte Interview näher vorstellen

2.3.3.1 Das strukturierte Interview

das strukturierte Interview, auch zum Teil als standardisiertes Interview bekannt, versucht eine möglichst hohe Standardisierung zu erreichen. Dementsprechend müssen nicht alle Fragen offen gehalten werden sondern können durchaus auch geschlossen sein. Um das strukturierte Interview einsetzen zu können, müssen bereits Informationen über den interessierenden Gegenstandsbereich vorliegen. Durch seine Verwendung können oftmals Anhaltspunkte für das Vorhandensein bestimmter Variablen erlangt werden.

2.3.3.2 Das problemzentrierte Interview

Das problemzentrierte Interview stellt ein offenes, halbstrukturiertes Verfahren dar. Es soll einem offenen Gespräch möglichst nahe kommen, soll aber zugleich auf ein Bestimmtes Problem (meist gesellschaftlich begründet) fokussiert/zentriert sein. Der Interviewer und der Befragte sollten möglichst eine offene, gleichberechtigte Beziehung aufbauen.

Es existieren drei unterschiedliche Fragetypen die in einem problemzentrierten Interview verwendet werden:

- *Sondierungsfragen* dienen dazu, einen Einstieg in die Thematik zu geben. Sie sollten dem Interviewer Informationen darüber liefern, inwieweit den Befragten das Thema überhaupt interessiert und welche subjektive Bedeutung es für ihn hat.
- *Leitfadenfragen* dienen dazu, die wichtigsten Themengebiete vorher bereits abzustecken. Sie sollten möglichst offen gestellt werden um ein normales Gespräch zu ermöglichen. Im Optimalfall sollte der Befragte gar nicht merken, dass gerade eine vorher bereits festgelegte „Frage“ gestellt wurde.
- *Ad hoc Fragen* sind vorher nicht festgelegte Fragen oder Themen, die aber aufgrund der Gesprächsentwicklung von Bedeutung sein können. Hier muss der Interviewer spontan reagieren können, wenn er auf vorher zwar nicht bedachte, aber doch interessante Aspekte trifft.

2.3.4 Gruppendiskussionen

Interviews bieten die Möglichkeit, sehr genau auf einzelne Personen eingehen zu können. Allerdings sind viele Meinungen und Einstellungen stark an soziale Zusammenhänge gebunden. Diese können am besten in einer Gruppenbefragung erforscht werden. Besonders geeignet sind sie bei der Untersuchung von Vorurteilen und Ideologien. Im Normalfall bringt eine direkte Frage nach beispielsweise antisemitischen Vorurteilen kaum Ergebnisse, da die Befragten hier nicht offen antworten. Durch die Gruppendynamik können aber Diskussionen entstehen, in denen eventuell vorhandene Vorurteile oder Einstellungen zu diesem Thema viel offener zu Tage treten. Eine gut geführte Gruppendiskussion vermag die psychischen Sperrn zu schwierigen Themen zu überwinden.

Zu Beginn wird der Gruppe meistens ein sogenannter Grundreiz, beispielsweise eine besonders kontroverse These präsentiert. Darauf aufbauend entwickelt sich dann die weitestgehend frei ablaufende Diskussion. In bestimmten Fällen kann es sinnvoll sein, am Ende eine sogenannte Metadiskussion durchzuführen. Hier kann der Diskussionsleiter Fragen stellen, die darauf abzielen, ob die Teilnehmer ihre Ansichten auch wirklich frei äußern konnten und wie sie sich dabei gefühlt haben.

In Bezug auf die Planung einer solchen Gruppendiskussion gibt es einige Punkte die es zu beachten gilt:

- *Diskussionsthema*: Hier kann unterschieden werden zwischen eng umschrieben und wenig strukturierten Themen. Weiterhin ist von Belang, inwieweit die Teilnehmer persönlich von dem Thema betroffen sind, sprich inwieweit sie motiviert sind, darüber zu diskutieren.

- *Gruppengröße*: Optimal gilt eine Gruppengröße von 5-15 Personen. Je größer die Gruppe, desto weniger Sprechzeit hat ein Einzelner, jedoch um so mehr verschiedene Meinungen existieren auch.
- *Zusammensetzung der Gruppe*: Um eine Diskussion zu ermöglichen, in der jeder Teilnehmer das Gefühl hat den Diskussionspartnern ebenbürtig zu sein, sollten die Personen in Bezug auf sogenannte soziodemographische Merkmale möglichst homogen ausgewählt werden. Zu den interessanten Aspekten gehören hier zum Beispiel Art der Ausbildung, Sachkompetenz in Bezug auf das Diskussionsthema, etc.
- *Bekanntheit der Mitglieder der Gruppe*: Hier kann unterschieden werden, ob die Diskussionsteilnehmer sich aus ihrem sozialen Umfeld bereits kennen, oder ob die Gruppe nur aufgrund der Untersuchung zusammengeführt wurde – eine sogenannte ad-hoc Gruppe. Prinzipiell haben beide Vor- und Nachteile, je nach Themengebiet lässt sich hier aber keine generalisierende Empfehlung geben.
- *Meinungsverteilung*: Damit überhaupt eine Diskussion zustande kommt, sollten die Meinungen zu dem Themengebiet möglichst vielfältig sein
- *Schweiger*: Damit sind Teilnehmer gemeint, die sich gar nicht oder nur sehr selten zu Wort melden. Die Gründe können hier vielfältig sein, angefangen bei der Persönlichkeit des Teilnehmers hin zu dem Problem, dass er mit dem Thema vielleicht nicht so viel anfangen kann. In kleineren Gruppen und in solchen, in denen sich die Teilnehmer kennen, ist die Anzahl der Schweiger im Allgemeinen kleiner.
- *Verhalten des Diskussionsleiters*: Der Diskussionsleiter hat mehrere Möglichkeiten die Diskussion zu führen. Er kann zum einen die sogenannte formale Gesprächsleitung übernehmen, also beispielsweise die Steuerung, wann wer mit Reden an der Reihe ist, um hier ein Durcheinander bei hitzigen Diskussionen zu vermeiden. Er kann aber auch aktiver in die Diskussion eingreifen, sei es nur durch das Einbringen von Themenrelevanten Stichworten um bisher nicht angesprochene Themenbereiche zu erfassen oder aber durch die aktive Teilnahme an der Diskussion. Hierzu ist es ratsam nicht direkt mit zudiskutieren, jedoch von Zeit zu Zeit weitere Reizargumente einzubringen um die Diskussion aktiv zu steuern.

Wenn möglich sollten die Diskussionen auf Video und/oder Tonband aufgezeichnet werden um die Auswertung zu erleichtern. Ebenfalls vorteilhaft kann das Einbringen eines sogenannten „stillen Beobachters“ sein. Dieser beteiligt sich nicht an der Diskussion sondern achtet vielmehr auf die Gestik und Mimik der Teilnehmer sowie sonstige Auffälligkeiten.

2.3.5 Validität und Aussagekraft qualitativer Methoden

In Bezug auf die Gütekriterien herrscht bei der qualitativen Forschung Uneinigkeit darüber, ob diese in Relation zu denen der quantitativen Forschung gesehen werden müssen oder ob die qualitative Forschung nicht besser eigene definieren sollte. Letztendlich existieren beide Ansätze die ich im Folgenden auch näher beschreiben möchte.

2.3.5.1 Validität

Die Validität besitzt in der qualitativen Forschung einen besonders hohen Stellenwert, da sie selbst an sich den Anspruch stellt, besonders gegenstandsangemessen vorzugehen. Allerdings gibt es auch hier einige Aspekte die es zu beachten gilt. Beispielsweise ist es fraglich, inwieweit die Äußerungen in einem Interview wirklich authentisch sind, oder ob der Interviewer vielleicht unbewusst auf den Befragten eingewirkt und so die Antworten verzerrt

hat. Diese Verzerrungen lassen sich letztendlich sowohl an dem Interviewer als auch an dem Befragten festmachen:

- *Verzerrung durch den Interviewer:* Probleme entstehen hier, wenn der Interviewer an entscheidenden Stellen nicht nachhakt, nicht auf den Befragten eingeht, es eventuell versäumt eine für den Befragten angenehme Atmosphäre zu schaffen. Verhindert werden können diese Fehler größtenteils durch intensive Schulung der Interviewer.
- *Verzerrung durch den Befragten:* Diese Art der Verzerrung kann entstehen, wenn der Befragte aus welchen Gründen auch immer, nicht bereit ist, wahrheitsgemäß zu antworten. Es ist nicht leicht eine solche Situation als Interviewer zu erkennen und angemessen darauf zu reagieren. Wie oben bereits erwähnt können die Schaffung einer vertrauensvollen Atmosphäre oder auch die einer möglichst transparenten Untersuchungssituation - damit der Befragte sich wirklich sicher ist, dass er den wahren Grund der Befragung kennt – mögliche Gegenmaßnahmen sein.

2.3.5.2 Interne und Externe Validität

- *Interne Validität:* Als wichtigste Maßnahme zur Sicherung der internen Validität gilt in der qualitativen Forschung die Konsensbildung. Sobald mehrere Personen sich auf die Glaubwürdigkeit und den Bedeutungsgehalt der Ergebnisse einigen können, kann dies als Hinweis auf seine Validität aufgefasst werden. Die Konsensbildung kann hierbei nicht nur zwischen mehreren Forschern stattfinden sondern auch zwischen Forscher und erforschter Person (kommunikative Validierung) oder zwischen Forschern und sogenannten Laien (argumentative Validierung).
- *Externe Validität:* Wie bei der quantitativen Methodik geht es bei der Frage der externen Validität darum, inwieweit die Ergebnisse sich auf die Wirklichkeit übertragen lassen, inwieweit sie verallgemeinerbar sind. Hier muss in der qualitativen Methodik unterschieden werden zwischen den hypothesengenerierenden Verfahren (z.B. qualitatives Experiment) und den hypothesenprüfenden Verfahren (z.B. Grounded Theory). In ersterem Fall kann die externe Validität weitestgehend vernachlässigt werden, da es dem Verfahren nicht darauf ankommt, dass die Ergebnisse verallgemeinerbar sind. Es werden ja keine Hypothesen überprüft, vielmehr soll ein Gegenstand erst erkundet, exploriert werden um seine Struktur zu erfassen. Es erfolgt also keine Prüfung auf Verallgemeinerbarkeit – somit ist die externe Validität auch nicht von Bedeutung.

Hypothesengenerierende Verfahren sind eher selten in der qualitativen Methodik. Als Beispiel könnte man hier die des „theoretical sampling“ in der „Grounded Theory“ nennen auf die ich jedoch nicht näher eingehen möchte. Letztendlich muss in diesem Fall eingestanden werden, dass die Ansprüche der quantitativen Forschung an die externe Validität von diesen Verfahren nur unzureichend erfüllt werden.

2.3.5.3 Ethik

Ähnlich wie bei der quantitativen Forschung existieren auch bei der qualitativen einige Bedenken bezüglich der ethischen Angemessenheit. Beispielsweise kommt es oft vor, dass bei einer teilnehmenden Beobachtung die beobachteten Personen über die wahren Absichten der Forscher getäuscht werden. In Interviewsituationen die nicht als solche zu erkennen sind, gibt der Befragte eventuell sehr persönliche Dinge preis.

Selbst wenn ein reguläres Interview stattfindet kann beispielsweise ein missbilligender Blick des Interviewers zu einem massiven Vertrauensverlust bei der befragten Person führen. Oftmals kann das Eintreten von negativen Konsequenzen nicht von vorneherein

ausgeschlossen werden. Gerhard Kleining ist zwar beispielsweise davon überzeugt, dass das qualitative Experiment aus ethischer Sicht keine Bedenken hervorruft, da sein explorativer Charakter den Eingriff in den Gegenstand soweit lenkt, dass dieser dabei nicht zerstört werden kann. Bei sachgemäßer Handhabung mag das auch zutreffen, jedoch besteht beim Eingriff in einen Gegenstand immer die Gefahr, dass dieser dabei zu Schaden kommt.

2.3.5.4 eigene Gütekriterien der qualitativen Forschung

Auch wenn die Gütekriterien quantitativer Forschung weitestgehend auf die qualitative übertragbar sind, gibt es Bewegungen innerhalb der qualitativen Forschung, die auf die Definition von eigenen Gütekriterien drängen. Dies hängt vor allen Dingen damit zusammen, dass beispielsweise das „schlechte“ Abschneiden qualitativer Verfahren bei Gütekriterien, wie Reliabilität oder externer Validität, diese in ein schlechtes Licht rückt und für Außenstehende schnell den Eindruck der nicht oder unzureichenden Wissenschaftlichkeit erweckt wird. Mayring (2000) schlägt zum Beispiel eine Unterteilung in 6 eigenständige Gütekriterien vor. Diese lassen sich jedoch letztendlich zum Großteil wieder den bereits bekannten der quantitativen Forschung zuordnen, weswegen der Sinn dieser durchaus bezweifelt werden darf. Letztendlich führen sie eher zu einer noch größeren Spaltung zwischen qualitativer und quantitativer Forschung anstelle der wünschenswerten Annäherung.

2.4 Methoden der Sozialwissenschaft - Fazit

Abschließend bleibt zu sagen, dass sowohl qualitative als auch quantitative Methoden Vor- und Nachteile haben. Die qualitativen Methoden schwächeln etwas in Bezug auf die Gütekriterien – insbesondere eignen sie sich nicht zur Hypothesenprüfung. Das müssen sie allerdings auch gar nicht, denn hier liegt das Spezialgebiet der quantitativen Methoden. Die qualitativen Methoden eignen sich vielmehr um überhaupt erst einmal Hypothesen zu entwickeln. Durch ihre größere Gegenstandsnahe erlauben sie zudem einen besseren Einbezug der gesamten Situation – es werden keine „Störfaktoren“ ausgeblendet, die vielleicht für die Struktur des Gegenstands wichtig sind.

In Bezug auf ethische Bedenken haben beide Richtungen ihre Probleme und sollten bei der Planung des Verfahrens bedacht werden.

Es spricht letztendlich nichts gegen eine friedliche Koexistenz beider Richtungen, von der die gesamte Forschung profitieren würde. Bei der Auswahl von qualitativen und quantitativen Methoden sollte diese, wenn möglich, so gegenstandsbezogen wie nur möglich erfolgen. Auch die Entwicklung von Verfahren die sowohl qualitative als auch quantitative Aspekte beinhalten, ist nur positiv entgegenzusehen. Dadurch können eventuell die jeweiligen Schwächen kompensiert und die Stärken gebündelt werden.

3. Usability Test Methoden

In Hinsicht auf Usability Test Methoden stellt sich nun die Frage, wie man von dem Wissen, das die Sozialwissenschaften bereits erlangt haben, profitieren kann, um Usability Tests valider und aussagekräftiger zu machen. Dazu müssen zunächst einmal bekannte Usability Test Verfahren analysiert werden und nach Gemeinsamkeiten zu Verfahren in der Sozialwissenschaft gesucht werden. Auch wenn vieles direkt übertragbar ist, so darf doch eins nicht vergessen werden: Nicht der Mensch, sondern das Softwareprodukt steht im Mittelpunkt eines Usability Tests. Ein blindes Verweisen auf sozialwissenschaftliche Methoden ist also keineswegs angebracht.

Wenn man existierende Usability Methoden betrachtet fällt schnell die erste Gemeinsamkeit auf – auch hier existieren qualitative und quantitative Methoden. Allerdings treffen hier nicht 2 verschiedene Forschungsrichtungen aufeinander, vielmehr hat man bereits früh erkannt, dass je nach Anwendung beide Methodenrichtungen ihre Vor und Nachteile haben. Viele Verfahren sind sogar sowohl qualitativer als auch quantitativer Natur. Als Beispiel möchte ich hier nur mal die „heuristische Evaluation“ nennen. In Bezug auf den Ablauf des Tests ist es in jedem Fall ein qualitatives Verfahren, wenn auch ein verhältnismäßig stark standardisiertes. Die Auswertung liefert jedoch auch quantitative Daten.

Der in den Sozialwissenschaften oftmals vorhandene Konflikt zwischen qualitativer und quantitativer Forschung existiert also nicht, was Usability Verfahren in dieser Hinsicht ihren Vorbildern aus der Sozialwissenschaft sogar überlegen erscheinen lässt.

Grundsätzlich lassen sich Usability Tests in zwei Kategorien einordnen. Da wären zum einen die sogenannten „Gestaltenden Tests“ (*formative Evaluation*) und zum anderen die „Kontrolltests“ (*summative Evaluation*). Letztendlich ist diese Unterteilung rein zeitlich zu sehen.

Die *formative Evaluation* umfasst die Entwicklung des Produkts und jegliche Tests die in dieser Zeit stattfinden, seien sie nun qualitativer oder quantitativer Natur. Jedoch sind die meisten Tests darauf ausgelegt, Usability Schwächen in dem Produkt zu finden. Da die Ergebnisse solcher Tests im besten Fall schon Redesign-Vorschläge darstellen, sind rein quantitative Tests wie beispielsweise „Performance Testing“ an diesem Punkt eher selten anzutreffen.

Die *summative Evaluation* hingegen steht am Ende des Entwicklungsprozess. Im Allgemeinen sind die Testmethoden, die hier Anwendung finden, eher quantitativer Natur. Der Grund liegt darin, dass die Entwicklung abgeschlossen ist und das Ziel einer abschließenden Evaluation darin liegt, die Güte des Systems als Ganzes zu überprüfen. Da qualitative Tests oftmals mehr fehlersuchender Natur sind, eignen sie sich dafür meist nicht. Vielmehr sollen nun Anhand dieser Tests quantitative Daten erzeugt werden, die beispielsweise einen Vergleich zu den zu Beginn definierten Usability Goals zulassen oder auch um das fertige Produkt mit Konkurrenzprodukten zu vergleichen. Diese Tests sind ungemein wichtig, ist es doch meist die letzte Chance noch zu erkennen, dass das Produkt eventuell gar nicht konkurrenzfähig ist.

Hinsichtlich des Entwicklungsprozesses ist diese Unterteilung sicherlich sinnvoll, da es jedoch eine Reihe von Usability Tests gibt, die sich beiden Kategorien zuordnen lassen, habe ich im Folgenden eine Klassifizierung nach Art des Tests gewählt.

3.1 Klassifizierung von Usability Test Methoden

Es gibt sicherlich unzählige Möglichkeiten, Usability Tests in Kategorien einzuordnen und zu klassifizieren. Die Unterscheidung nach qualitativen und quantitativen Methoden halte ich wie oben bereits beschrieben nicht für sonderlich sinnvoll. Vielmehr sollten die Tests nach der dahinter steckenden Methodik klassifiziert werden.

Es lassen sich letztendlich 3 verschiedene Arten von Usability Tests identifizieren. Zum einen wäre da der klassische Usability Test im Labor, bei der die Versuchsperson verschiedene Aufgaben am System bewältigen muss. Sowohl qualitative Techniken wie „Thinking Aloud“ als auch quantitative wie „Performance Testing“ lassen sich unter dem Begriff „Diagnostic Evaluation“ zusammenfassen. Weiterhin kann man die Gruppe der, weitestgehend qualitativen, Methoden identifizieren, die jedoch nicht an einen Test im Labor gebunden sind. Hierzu gehören Tests zu Beginn der Entwicklung, wenn noch gar kein Prototyp existiert (z.B. Card Sorting Techniken) aber auch Tests zu späteren Zeitpunkten, die hauptsächlich an Gruppendiskussionen und Interviewtechniken angelehnt sind. Sie fasst man unter „Participatory Evaluation“ zusammen. Abschließend kann noch die Gruppe der Tests identifiziert werden, die beispielsweise Onlineumfragen oder aber auch Post Test Fragebögen umfasst – genannt „Subjective Evaluation“. Da dieses Thema bei angemessener Betrachtung eine eigene Seminararbeit verdient, möchte ich es in dieser Arbeit ausklammern.

Die Methoden der Sozialwissenschaft haben gezeigt, dass die Auswahl der Versuchspersonen oftmals entscheidenden Einfluss auf die Repräsentativität der Ergebnisse haben kann. Da in der Softwareentwicklung sowieso zu Beginn eine Benutzeranalyse gemacht werden sollte, lohnt es sich in Hinblick auf zukünftige Usability Tests, diese auch möglichst gründlich durch zu führen. Deborah J. Mayhew beschreibt in ihrem Usability Lifecycle einige Methoden hierzu sehr ausführlich, weswegen ich auch nicht näher darauf eingehen möchte. Anzumerken sei jedoch, dass sich auch hier wieder klar die Verwandtschaft zu den Methoden der Sozialwissenschaft zeigt. Angefangen bei Interviewtechniken bis hin zu teilnehmenden Beobachtungen können diese, zumeist qualitativen Verfahren, quasi eins zu eins übernommen werden.

3.1.1 Participatory Evaluation

In der Literatur wird dieser Begriff meistens recht einschränkend auf die sehr frühe Entwicklungsphase beschränkt. Im Rahmen der von mir gewählten Klassifizierung, gehören aber durchaus auch Methoden, die erst oder auch zu fortgeschrittener Entwicklungszeit angewendet werden können dazu. Es geht hierbei weniger darum, dass die Versuchsperson bestimmte Aufgaben am System tätigt und dabei beobachtet wird um beispielsweise die Fehlerrate zu messen. Vielmehr steht der Versuchsleiter bei dieser Art von Tests in aktivem Kontakt mit der Versuchsperson. Bei Verfahren, die dem ersten Anschein nach an einen klassischen Usability Test erinnern, besteht der Unterschied darin, dass die Versuchsperson aktiv mit dem Versuchsleiter diskutiert und auf Probleme bei der Anwendung hinweist. In früheren Phasen der Entwicklung besteht das Ziel darin, den potentiellen Benutzer besser kennen zu lernen und seine Bedürfnisse, Wünsche und Erwartungen zu verstehen.

Es eignen sich hierfür im speziellen qualitative Verfahren, wobei trotz allem auch quantitative Daten gesammelt werden können.

3.1.1.1 Verfahren der Participatory Evaluation

Da es kaum möglich ist, einen umfassenden Überblick über alle möglichen Verfahren zu geben, werde ich einige ausgewählte näher beschreiben. Im Allgemeinen sind andere Verfahren zumindest teilweise an diese angelehnt beziehungsweise Variationen.

A Card Sorting Techniken

Card Sorting findet besonders im frühen Entwicklungsstadium Anwendung. Der Vorteil liegt darin, dass kein funktionierender Prototyp existieren muss. Ein beliebtes Anwendungsbeispiel ist das Testen der Menüstruktur auf Erwartungskonformität. Hierbei wird die geplante oder auch bereits implementierte Menüstruktur einzeln auf Kärtchen geschrieben – jeder Menüpunkt steht auf einer Karte. Die Versuchsperson bekommt nun einen Haufen Karten vorgelegt, und erhält die Aufgabe, daraus wieder eine Menüstruktur zu bauen, die Karten also zu ordnen. Begriffe die für einen Entwickler klar zusammengehören, werden vom eigentlichen Benutzer oftmals völlig anders eingeordnet. Die Schwierigkeit ist, nun zu entscheiden, wer von beiden Recht hat. Es ist für die Versuchsperson ungleich schwerer ein Menü aus einzelnen Punkte aufzubauen, ohne beispielsweise die Anwendung zu kennen. Oftmals werden Begriffe von den Entwicklern nicht nur aufgrund der thematischen Verbindung gemeinsam platziert, sondern vielmehr, weil sie in ihrer Anwendung häufig zusammen benutzt werden. Dies kann die Versuchsperson aber nur schwer wissen, weswegen den Ergebnissen dieses Tests nicht blind vertraut werden darf.

B Interviewtechniken

Userbefragungen haben auch im Usability Bereich eine lange Tradition. Prinzipiell ist auch hier das Ziel der Befragung eine genauere Vorstellung über die Zielgruppe zu erhalten. Weiterhin können die befragten Personen Informationen über ihre Wünsche und Vorstellungen hinsichtlich der Software nennen. Problematisch ist hierbei, dass oftmals die Personen selber nicht so richtig wissen was sie eigentlich für die tägliche Arbeit mit dem System benötigen und wirklich brauchen. Interviews können an fast jeden Test angeschlossen werden. Hierbei interessiert dann zumeist, was die Person über den gerade absolvierten Test denkt. Falls dies beispielsweise ein Performance Test war, kann über ein anschließendes Interview noch qualitatives Material gesammelt werden, um die eventuell aufgetretenen Probleme besser verstehen zu können.

C Focus Groups

Bei Focus Groups stellt der Versuchsleiter eine Gruppe Personen zusammen, welche im Optimalfall nicht nur aus möglichen Endnutzern sondern auch aus einigen Programmierern und weiteren Usability Experten besteht. Der Test ist letztendlich eine Gruppendiskussion zwischen diesen Personen. Wenn möglich wird an einem aktuellen Prototyp gemeinsam das Produkt beziehungsweise ein bestimmter Teil des Produktes (Focus!) genauer betrachtet. Der Versuchsleiter hat hierbei die Aufgabe der Moderation und des Einbringens von möglichen Fragestellungen. In die Diskussion an sich sollte er wenn möglich nur eingreifen, um das Gespräch auf den gewählten Focus zu richten und nicht zu sehr abdriften zu lassen.

D Heuristische Evaluation

Die Heuristische Evaluation wird in den meisten Fällen nirgends untergliedert sondern extra behandelt. Der Grund liegt darin, dass hier keine Endnutzer das Produkt testen, sondern in den meisten Fällen Usability Experten. Allerdings ist es auch möglich Experten anderer Fachrichtungen teilnehmen zu lassen, beispielsweise im Fall der Evaluation einer Benutzeroberfläche können Experten die sich mit den kognitiven Fähigkeiten der Menschen beschäftigen ebenfalls interessante und wichtige Ergebnisse liefern.

Der Test selbst läuft so weit wie möglich standardisiert ab. Das heißt die Experten untersuchen das System anhand einer festgelegten Liste – auch Fragebogen genannt (XEROX hat beispielsweise eine solche Liste entwickelt). Es ist wichtig, dass diese Liste nicht von den Entwicklern erstellt wurde sondern möglichst allgemein ist, da ansonsten bereits eine Vorauswahl getroffen wird. Stoßen die Experten auf ein Usability Problem, ist es ihre Aufgabe dieses in seiner Schwere zu klassifizieren und eventuell noch Re-design Vorschläge zu geben. Am Ende der heuristischen Evaluation kann noch eine Nachbesprechung, im Falle von einzelnen Versuchen eine Art Interview, ansonsten eine Gruppendiskussion, stattfinden. Hier können besonders auffällige Probleme direkt angesprochen werden um eventuell vorhandene Missverständnisse auszubügeln.

3.1.1.2 Vergleich zu Sozialwissenschaften & Validität

Die Zusammenhänge mit den Methoden der Sozialwissenschaften sind auch hier wieder schnell zu erkennen. Vor allem Focus Groups (Gruppendiskussion) und Interviewtechniken sind sehr stark an die entsprechenden qualitativen Verfahren der Sozialwissenschaft angelehnt. Die dort genannten Richtlinien und Hinweise in Bezug auf die Durchführung, sind allesamt auch in diesem Anwendungskontext gültig.

In Bezug auf die Validität dieser Tests kann somit auch die Sozialwissenschaft als Grundlage dienen. Der Interviewer muss auch hier darauf achten, sich der Person gegenüber angemessen zu verhalten und eine angenehme Atmosphäre zu schaffen. Auch wenn die befragten Personen im Allgemeinen hier keine persönlichen Dinge preisgeben müssen, kann ein falsches Auftreten trotzdem zu Falschaussagen führen. Beispielsweise könnte sich eine Versuchsperson nicht trauen, ihre Probleme mit einer bestimmten Software zuzugeben, da sie vor dem Interviewer nicht „dumm“ erscheinen möchte. Ebenfalls könnte sie sich im Anschluss an einen Benutzertest nicht trauen, offen Kritik an dem System zu üben, sei es weil sie die Schuld für Probleme sich selbst gibt (oft zu beobachten bei Menschen die wenig mit Computern zu tun haben) oder denkt durch Kritik den Versuchsleiter persönlich anzugreifen. Intensive Schulung von Interviewer beziehungsweise Versuchsleiter ist somit auch hier unabdingbar.

In Focus Groups ist ebenfalls auf eine ausreichende Schulung zu achten. In Bezug auf die Gruppengröße sollte diese 10 Personen nicht überschreiten, da ansonsten eine Konsensbildung schwierig wird. Schweiger sind hier jedoch relativ problemlos, wiederum aus dem Grund, da ja nicht die Menschen Gegenstand der Untersuchung sind. Ein Schweiger entspricht somit einem Test mit einer Person weniger, verzerrt das Ergebnis letztendlich aber nicht. Das Einbringen eines stillen Beobachters ist hier ebenfalls eher unnötig, da keine soziale Problematik erörtert wird und somit Reaktionen oder Verhaltensmuster der Teilnehmer irrelevant sind.

Die Sicherung der internen Validität funktioniert zumeist über Konsensbildung, in der Regel zwischen Usability Experten und Entwicklern. Im Falle eines Focus Groups Test prinzipiell

schon während der Diskussion (wodurch auch Nicht-Entwickler beteiligt sind), an deren Ende wenn möglich ein Konsens stehen sollte.

Ethische Bedenken bestehen eigentlich nur sehr bedingt, da ja die Software zumeist Ziel der Untersuchung ist. Das lässt sich prinzipiell auch auf jede Art von Usability Test übertragen.

Für Card Sorting und heuristische Evaluation gibt es keine direkte Entsprechung in der Sozialwissenschaft. Da diese Techniken sehr in das Umfeld der Softwareentwicklung eingebettet sind, ist das auch nachvollziehbar. In Bezug auf die Validität des Card Sortings kann gesagt werden, dass hier wiederum die Auswahl der Versuchspersonen relativ wichtig ist. Da die Darstellung, beispielsweise einer Menüstruktur, anhand von Kärtchen sehr abstrakt ist, sollten sie über genügend Vorstellungskraft verfügen, wie das Menü in einem real existierenden System aussehen könnte. Anwender die diesen Sprung nicht schaffen, ordnen die Karten eventuell mehr oder weniger nach dem Zufallsprinzip. Ein Interview am Ende eines solchen Tests kann diese Schwierigkeiten eventuell aufdecken.

Heuristische Tests sollten wie schon gesagt möglichst standardisiert ablaufen. Vor allen Dingen sollte der Fragebogen nicht von den Entwicklern selbst zusammengestellt worden sein, da sie hier bereits unbewusst eine Vorauswahl an kritischen Bereichen treffen könnten. Andererseits haben standardisierte Bögen den großen Nachteil, oftmals Punkte zu enthalten, die auf die getestete Software überhaupt nicht anwendbar sind (Beispielsweise Fragen nach einer Konsole – in dem Produkt existiert nur keine). Je nach Erfahrung der Experten kann das durchaus zu Verwirrung führen und dadurch auch das Ergebnis beeinflussen.

3.1.2 Diagnostic Evaluation

Der Usability Test der in dieser Kategorie angesiedelt ist, entspricht am ehesten vom Ablauf her dem eines quantitativen Experiments der Sozialwissenschaft. Im Gegensatz zu diesem kann das Verfahren sowohl qualitativ als auch quantitative Ergebnisse liefern. Allen Varianten gemeinsam ist der grundsätzliche Ablauf eines solchen Tests. Eine Versuchsperson muss an einem Prototypen oder fertigen System mehrere Aufgaben bewältigen. Der Versuch findet zumeist in einer laborähnlichen Umgebung statt um allen Teilnehmern gleiche Bedingungen zu garantieren. Der Versuchsleiter kann den Test entweder qualitativ oder quantitativ ausrichten und dementsprechend gestaltet sich die Art der Testaufgaben und deren Auswertung. Für gewöhnlich darf die Versuchsperson am Ende noch einen sogenannten Post Test Questionnaire ausfüllen.

3.1.2.1. Verfahren der Diagnostic Evaluation

Im Detail sieht das folgendermaßen aus: Zunächst sollte ein Prototyp des zu testenden Systems zur Verfügung stehen. Aus diesem Grund eignet sich das Verfahren zumeist nicht in ganz frühen Entwicklungsstadien. Es sollte weiterhin definiert werden, welche Teile des Systems genau getestet werden sollen, da ein Test des kompletten Systems meistens den Umfang sprengen würde. Danach sollte entschieden werden, ob der Test qualitativer oder quantitativer Natur sein soll. Dementsprechend müssen die Testaufgaben erarbeitet werden. Qualitative Tests beschäftigen sich mehr mit der Interaktion und dem System an sich, wollen den gewählten Bereich möglichst umfassend abdecken. Um zusätzliche Informationen zu erhalten können hier auch schon mal Fragen eingestreut werden, die die Versuchsperson dazu auffordern, zu einem bestimmten Designelement Stellung zu nehmen. Wie schon angedeutet sollten die Fragen möglichst den gewählten Systembereich komplett in Punkto

Interaktionsmöglichkeiten abdecken um so viele Fehler wie möglich aufzudecken. Während des Tests wird die Versuchsperson meistens gebeten laut zu denken, also ihre Handlung zu kommentieren. Weiterhin sollte der Test auf Video aufgezeichnet werden um die Auswertung zu erleichtern. Die eigentliche Auswertung versucht dann die gefundenen Usability Probleme zu klassifizieren und zu ordnen. Je nachdem wie viele Nutzer teilgenommen haben (bei Tests qualitativer Natur zumeist unter 10), können auch noch begrenzt statistische, also quantitative Daten, wie zum Beispiel Fehlerrate, erfasst werden.

Ein rein quantitativer Testablauf versucht vergleichbare Ergebnisse zu liefern, um beispielsweise die Marktchancen gegenüber Konkurrenzprodukten besser abschätzen zu können oder um die festgelegten Usability Goals zu überprüfen. Ein solcher Test findet meistens am Ende der Entwicklung im Rahmen der *summative Evaluation* statt.

In Hinblick auf unterschiedliche qualitative Verfahren sind besonders die verschiedenen *Thinking aloud* Varianten interessant. Als Beispiel für ein quantitatives Verfahren werde ich das sogenannten *Performance Testing* genauer erläutern.

A Thinking aloud – Standard Version

Während die Versuchsperson versucht die Testaufgaben zu lösen, wird sie gebeten, durchgängig laut zu denken, das heißt, die eigene Handlung zu kommentieren und gegebenenfalls auch zu erklären. Die Vorteile liegen auf der Hand – wenn die Versuchsperson nicht weiterkommt oder eine Aufgabe falsch löst, kann man auch den Grund dafür feststellen. Gleichzeitig bekommt man ein deutlicheres Bild von den wirklichen Erwartungen des Benutzers an die Software. Informelle Befragungen zielen zwar auch auf diesen Sachverhalt ab, oftmals erkennen Benutzer Probleme aber erst, wenn diese auch auftreten, also in der Anwendung an sich. Problematisch an solchen Tests ist die Unnatürlichkeit der Situation. Die Versuchsperson ist es eigentlich nicht gewohnt, die ganze Zeit die eigene Handlung zu kommentieren und muss dementsprechend öfters daran erinnert werden. Dadurch kann sich bei ihr das Gefühl einstellen, etwas falsch zu machen und sie wird eventuell verunsichert. Die Leistungsfähigkeit der Person kann also absinken. Es hängt somit viel von der Erfahrung des Versuchsleiters, aber auch der Versuchsperson selber ab, inwieweit die Ergebnisse aussagekräftig sind.

B Thinking aloud – Constructive Interaction

Der Unterschied zur Standard Version liegt an dem Umstand, dass nun 2 Versuchspersonen gleichzeitig das System testen. Diese sollten wenn möglich den gleichen Kenntnisstand besitzen. Während dem Test muss nun nicht mehr die Versuchsperson ihre Gedanken laut äußern, vielmehr sollen die beiden Teilnehmer über ihre Lösungswege diskutieren und somit in dieser Diskussion ihre Handlungsüberlegungen laut äußern. Der Vorteil liegt darin, dass die Versuchssituation deutlich natürlicher wird und es den Teilnehmern leichter fällt, ihre Gedanken zu äußern. Insbesondere falls die Zielgruppe des Produktes Kinder und somit auch Kinder die Versuchspersonen sind, bietet sich diese Technik an. Kinder haben ansonsten oftmals enorme Probleme sich richtig zu artikulieren, was ein Standard Thinking Aloud fast unbrauchbar macht. Nachteile sind jedoch auch vorhanden. Dadurch dass zwei Personen an dem System arbeiten, verschwinden eventuell einige Usability Probleme, auf die nur ein geringerer Teil der Versuchspersonen in Einzeltests getroffen wäre, vollständig. Zum anderen kommen nicht alle Teilnehmer immer gut miteinander aus. Oftmals hängt das mit völlig unterschiedlichen Lösungsansätzen zusammen – je nach Charakter können sich die Teilnehmer dann nicht mehr auf einen gemeinsamen Weg einigen.

C Retrospective Testing

Retrospective Testing verfolgt einen anderen Ansatz. Um die Versuchsperson während dem eigentlichen Test nicht mit thinking aloud zusätzlich zu belasten, wird der Versuch auf Video aufgezeichnet – was allerdings sowieso der Fall sein sollte. Während im Normalfall der Versuchsleiter das Video auswertet tritt nun die Versuchsperson in Aktion. Am Ende des eigentlichen Tests schauen sich Versuchsleiter und Versuchsperson gemeinsam das Video an. Um den Zeitaufwand zu reduzieren kann sich der Versuchsleiter während dem Test interessante Stellen bereits notieren. Aufgrund der technischen Möglichkeiten eines Videos, sollte es kein Problem sein, beim gemeinsamen Anschauen sich auf diese Punkte zu konzentrieren. Die Versuchsperson hat nun die Aufgabe zu den interessanten Stellen im Video zu erklären, wieso sie dieses oder jenes getan hat und was ihr dabei durch den Kopf ging.

In meinen Augen ist ein derartiger Test jedoch wenig sinnvoll. Auch wenn die kognitive Belastung während dem Test sinkt, was auch sicherlich zu begrüßen ist, wird dieser Vorteil jedoch zu teuer erkaufte. Rein wirtschaftlich gesehen steigt die Testzeit und damit auch die Kosten für den Test beträchtlich. Weiterhin besteht die Gefahr, dass die Versuchsperson beim wiederanschauen gar nicht mehr genau weiß, wieso sie einzelne Aktionen so und nicht anders ausgeführt hat. Sie fühlt sich eventuell aber unter Druck gesetzt die Aktion zu kommentieren und gibt dementsprechend falsche Aussagen von sich.

D Performance Testing

Performance Testing ist die Standard Testmethode, wenn es darum geht, quantitative, vergleichbare Ergebnisse zu erzeugen. Angewendet werden kann es zumeist am Ende der Entwicklung wenn das Produkt nur noch die *summative Evaluation* durchlaufen muss. In diesem Fall dienen die Ergebnisse zur Verifizierung der Usability Goals. Weiterhin kann so auch ein Vergleich zu Konkurrenzprodukten gezogen werden. Das ist zum einen für Marketingzwecke wichtig und zum anderen können dadurch auch die Marktchancen besser abgeschätzt werden. Im Notfall könnte hier auch noch die Handbremse gezogen werden, um das Produkt einem weiteren Entwicklungszyklus zu unterwerfen, falls es noch nicht konkurrenzfähig ist. In frühen Phasen dient Performance Testing vor allem dazu, mehrere Prototypen zu vergleichen.

Die Testaufgaben sollten möglichst so gestellt sein, dass quantitative Daten überhaupt gemessen werden können. Im Gegensatz zu qualitativen Methoden sollte der Versuchsleiter wenn möglich gar nicht eingreifen, da ansonsten die Daten nicht mehr vergleichbar sind. Weiterhin werden die erhobenen Daten statistisch ausgewertet. Grundsätzlich muss hier zwischen *Effizienz* und *Effektivität* unterschieden werden. Zu Effizienz gehören:

- *Task Time*
Dies umfasst zum einen die durchschnittliche Zeit die zur Bewältigung der Aufgaben benötigt wurde für jede Versuchsperson, als auch arithmetisches Mittel, Standardabweichung, Minimum und Maximum über alle Versuchspersonen.
- *Completion Rate / Mean Time on Task*
gibt die Prozentzahl der erfolgreich absolvierten Aufgaben in Bezug zur durchschnittlich benötigten Zeit an. Auch hier werden über alle Versuchspersonen noch arithmetisches Mittel, Standardabweichung, Minimum und Maximum bestimmt. Dies ist die wichtigste Maßzahl der Effizienzmessung. Wenn verschiedene Systeme verglichen werden und für jedes die gleichen Testaufgaben Verwendung finden, kann

hier sehr schnell erkannt werden, welches System effizienter ist, da der Wert nicht nur angibt wie schnell jemand die Aufgaben gelöst hat, sondern auch wie richtig.

Die Effektivität lässt sich folgendermaßen ausdrücken:

- *Completion Rate*
Sie gibt für jede Versuchsperson die Prozentzahl an erfolgreich absolvierten Aufgaben an. Dieser Wert kann noch unterteilt werden in erfolgreich absolvierte Aufgaben ohne Hilfe und mit Hilfe.
- *Errors*
Die Fehleranzahl ist ebenfalls ein entscheidender Faktor. Eine zusätzliche Klassifizierung der Fehler ist zu empfehlen.
- *Assists*
Die Anzahl an Hilfen die gegeben wurden. Falls Hilfen gewährt werden, muss bei der Completion Rate unterschieden werden zwischen Aufgaben mit Hilfe und ohne Hilfe.

Wie auch bereits bei der Effizienz wird für jeden Wert noch die Standardabweichung, arithmetisches Mittel, Minimum und Maximum über alle Teilnehmer bestimmt. Je nach Ausführung können die Ergebnisse nur über alle Aufgaben angegeben werden oder aber auch für jede einzelne Aufgabe. Die Abbildung zeigt eine Beispieltabelle in welche die oben genannten Messwerte eingetragen werden können.

User #	Unassisted Task Effectiveness [(%)Complete]	Assisted Task Effectiveness [(%)Complete]	Task Time (min)	Effectiveness / Mean Time-On-Task	Errors	Assists
1						
2						
N						
Mean						
Standard Deviation						
Min						
Max						

Performance Testing ermöglicht das Aufspüren von Usability Schwächen in einem Produkt, welche weniger formalen Methoden oftmals verborgen bleiben. Insbesondere auch was die Leistungsfähigkeit der Versuchspersonen und damit auch der Zielgruppe betrifft. Für Produkte die sehr zeitkritisch sind, beispielsweise eine Software für Fluglotsen, sind Performance Tests unabdingbar.

Wie bereits einleitend erwähnt, ist das Performance Testing von seiner Methodik stark an das quantitative Experiment der Sozialwissenschaft angelehnt. Insofern liegt der Schluss nahe, dass es sich bei Betrachtung von Validität und Aussagekraft auch an diesem messen muss.

3.1.2.2 Validität und Aussagekraft – Performance Testing

Ein bedeutender Unterschied zwischen Sozialwissenschaften und Usability Testing liegt bei dieser Testform darin, dass beim Usability Testing normalerweise keine Kontrollgruppe existiert. Dies hängt damit zusammen, dass zum Beispiel bei Vergleichsmessungen mit Konkurrenzprodukten die unabhängige Variable die Software selbst ist, die variiert wird. Würde man nun zunächst beide Gruppen mit Produkt A arbeiten lassen und anschließend die eine Gruppe mit Produkt B und die andere Gruppe weiterhin mit Produkt A, wäre der Lerneffekt den die 2. Gruppe bei der erneuten Benutzung von Produkt A derart groß, dass die Ergebnisse völlig verzerrt würden. Um hier einigermaßen Nutzen aus 2 Gruppen zu ziehen, müssten beide sowohl auf Produkt A als auch auf Produkt B trainiert werden, um Lerneffekte auszuschließen. Der Nutzen würde den zusätzlichen Aufwand allerdings kaum rechtfertigen.

A Repräsentativität der Ergebnisse

Es gilt hier prinzipiell das gleiche wie bereits beim quantitativen Experiment der Sozialwissenschaft (im Folgenden mit q.E.S. abgekürzt). Die Auswahl der Versuchspersonen ist auch hier entscheidend wenn die Ergebnisse auf eine Grundgesamtheit verallgemeinert werden sollen. Falls die Versuchspersonen beispielsweise alle gute bis sehr gute Computerkenntnisse hatten, die Software aber auch von Einsteigern benutzt werden soll, können völlig falsche Ergebnisse entstehen. Einsteiger haben oftmals gänzlich andere Probleme als erfahrene Computer Benutzer. Marketingaussagen unterscheiden hier oftmals jedoch nicht, denn diese sollen so allgemein wie nur möglich formuliert sein („mit unserem Produkt kommen sie doppelt so schnell ans Ziel wie mit dem der Firma XY“). Weiterhin werden fast immer absolute Werte verwendet – dass dies eigentlich fast nie möglich ist, zeigt das q.E.S. Auch Usability Goals werden oftmals absolut formuliert und verlangen somit auch von den Ergebnissen des Performance Testings absolute Werte. Es muss in jedem Fall klar darauf geachtet werden, dass dies für die definierte Grundgesamtheit, also die Zielgruppe wirklich zutrifft.

B interne Validität & externe Validität

Um die Ergebnisse vergleichbar zu machen, ist die interne Validität auch hier ein wichtiger Faktor. Wie beim q.E.S. herrscht zwischen interner und externer jedoch auch hier ein Antagonismus. Es muss je nach Softwareprodukt sehr sorgfältig überlegt werden, welche der beiden wichtiger ist. Beispielsweise würde eine hohe interne Validität bei der oben angesprochenen Lotsensoftware eventuell dafür sorgen, perfekte Ergebnisse zu erhalten – allerdings muss man sich doch fragen, ob diese Ergebnisse in der Realität erzielt werden können, wenn der Lotse in einer extremen Stresssituation ist und dazu noch viel Aufruhr um ihn herum herrscht. In diesem Fall wäre also eine möglichst realitätsnahe Situation besser geeignet, die dann aber gegebenenfalls schlechtere Zahlen liefern würde – was sich wiederum schlechter vermarkten lässt.

C Stör und Drittvariablen, zufällige Fehler

Stör und Drittvariablen haben hier eine etwas andere Bedeutung. Da grundsätzlich eigentlich Einzeltests Anwendung finden, ist die Gefahr durch solche ungewollten Einflüsse das Ergebnis zu verzerren eher gering. Trotzdem sollten sie natürlich Beachtung finden, da die Anzahl an Usern meistens nicht derart groß ist, dass hier viele Fehler abgefangen werden

können. Während bei einem q.E.S. solche Störvariablen meistens komplett ausgeschlossen werden, ist es beim Performance Testing am wichtigsten, dass alle Teilnehmer die gleichen Bedingungen haben. Es kommt also darauf an, mit was die Ergebnisse verglichen werden sollen. Wenn in dem gleichen Testlauf auch das Konkurrenzprodukt getestet wird, sind vorhandene aber konstante Störeinflüsse verkraftbar. Wenn allerdings die absolute Höhe der Ergebnisse wichtig ist, sollten solche unerwünschten Einflüsse naturgemäß ebenfalls weitestgehend unterbunden werden.

D Demand Characteristics & Forced Exposure

Der erste Punkt, „demand characteristics“, ist beim Performance Testing weniger wichtig, da den Versuchsteilnehmern das Versuchsziel vorher klar dargelegt wird und die Versuchsperson auch nicht Gegenstand des Tests ist. Trotzdem ist es natürlich eine unnatürliche Situation und die Personen verhalten sich dementsprechend eventuell anders. Solange dadurch nicht ihre Leistungsfähigkeit beeinflusst wird, ist das jedoch vernachlässigbar.

Das Problem des „forced exposure“ ist naturgemäß auch hier vorhanden, jedoch gilt auch hier, dass es erst kritisch wird, wenn die Leistungsfähigkeit der Versuchsperson beeinträchtigt wird. Man sollte aus diesem Grund der Versuchsperson zu Beginn des Tests eindeutig erklären, dass sie jede Aufgabe und auch den ganzen Test jederzeit abbrechen kann. Zumindest das Abbrechen einzelner Aufgaben wird in der Praxis auch des öfteren wahrgenommen.

E Versuchsleiter Effekte

Dem Versuchsleiter kommt die gleiche Aufgabe zu, wie beim q.E.S.. Prinzipiell können die dort genannten Probleme auch eins zu eins übernommen werden, allerdings mit einigen Einschränkungen.

Da nicht das Verhalten der Versuchspersonen analysiert wird sondern ihre Leistungsfähigkeit mit der Software, sind Effekte aufgrund physischer oder sozialer Merkmale eher unbedeutend. Wichtiger ist, dass der Versuchsleiter für eine angenehme Atmosphäre sorgt und der Versuchsperson glaubhaft vermitteln kann, dass nicht sie getestet wird sondern die Software. Das ist in der Praxis oftmals gar nicht so einfach, da viele Teilnehmer sich trotzdem beobachtet und überwacht fühlen, wenn neben einer Videokamera noch ein Protokollant und der Versuchsleiter im Raum sitzt – ob das durch ein Labor in dem die Versuchsperson allein im Raum sitzt und durch einen Spiegel oder auch nur durch Kameras beobachtet wird besser ist, wage ich zu bezweifeln.

Ein Teilnehmer der sich physisch zu dem Versuchsleiter hingezogen fühlt, ist eventuell in seiner Leistungsfähigkeit auch etwas eingeschränkt und verunsichert, jedoch dürfte sich dies hier im Rahmen halten.

Lern und Reifeeffekte sind hier ähnlich problematisch wie beim q.E.S. Es hängt hier von dem Grad der Standardisierung ab, wie weit sie von Bedeutung sein können. Da der Versuchsleiter aber unter anderem auch für Hilfen zuständig ist und im Allgemeinen allein entscheidet, wann eine solche zu geben ist, hat er entscheidenden Einfluss auf die Ergebnisse. Eine Schulung ist hier in jedem Fall auch wünschenswert.

Wenn möglich sollten ebenfalls externe Versuchsleiter verwendet werden, die nicht aktiv an der Entwicklung des getesteten Produkts beteiligt sind. Da der Versuchsleiter wie oben bereits erwähnt unter anderem für die Hilfen verantwortlich ist, könnte er auch hier unterbewusst das

Ergebnis verfälschen, wenn er persönlich an einem bestimmten Ergebnis interessiert ist. Die Entwickler selbst haben prinzipiell nichts im Labor zu suchen. Es fällt ihnen naturgemäß sehr schwer ruhig zu bleiben, wenn die Versuchsperson bei einfachen Aufgaben verzweifelt oder wenn ein Feature nicht gefunden oder missbilligt wird.

3.2 Usability Test Methoden - Fazit

Insbesondere *Diagnostic Evaluation* zeigt sehr schön die Verwandtschaft zu den Methoden der Sozialwissenschaft. Methoden wie Performance Testing haben ihren Ursprung klar im quantitativen Experiment der Sozialwissenschaft und sollten dementsprechend auch die gezogenen Lehren bezüglich der Validität übernehmen. In vielen Fällen wird hier leider nicht ein sonderlich großer Aufwand betrieben, auch wenn das für die Akzeptanz von Usability Tests ungemein wichtig wäre.

Wichtig wäre auch eine Klassifizierung der angewendeten Methoden, um überhaupt einen Vergleich zu den Sozialwissenschaften ziehen zu können. Wenn dies konsequent gemacht werden würde und gleichzeitig noch die Erfahrungen des Usability Testings selbst integriert würden, könnte sich in kurzer Zeit eine wirklich eigenständige Methodologie entwickeln. Momentan wird man jedoch fast noch überwältigt von den Instant Usability Anleitungen, die die jeweilige Testmethode auf erklären, sich aber meistens auf eine Schritt für Schritt Anleitung beschränken. Dies ist zwar wichtig um Usability Testing zu verbreiten und Publik zu machen. Wenn die Hintergründe aber nicht vermittelt werden, werden sie irgendwann in Vergessenheit geraten. Ein Test kann nur dann richtig durchgeführt werden, wenn man nicht nur weiß, was man machen muss, sondern auch wieso man es tun sollte. Mir persönlich ist kein Buch oder Paper bekannt, dass eine umfassende Übersicht über Usability Testmethoden gibt und diese nicht nur erklärt, sondern auch die Hintergründe insbesondere bezüglich der Validität liefert.

4. Schlussfolgerung

Die Gemeinsamkeiten zwischen Usability Test Methoden und jenen der Sozialwissenschaften sind unübersehbar vorhanden. Trotzdem dürfen Überlegungen bezüglich des Testsettings und der Validität und Aussagekraft nicht blindlings übernommen werden. Der entscheidende Unterschied, dass eben ein Softwareprodukt und nicht der Mensch im Mittelpunkt des Tests steht muss in jedem Fall Beachtung geschenkt werden. Auch eine Klassifizierung zwischen qualitativen und quantitativen ist sinnvoll – nicht aber eine Unterscheidung. Beide Richtungen können problemlos auch gemeinsam angewendet werden. Um Rückschlüsse auf die geforderte Validität und Aussagekraft treffen zu können ist eine Einstufung aber nützlich. In dieser Hinsicht haben die Usability Testing Methoden den Konflikt, der innerhalb der Sozialwissenschaften herrscht, erfolgreich umschifft. Auch wenn es Präferenzen gegenüber einzelnen Methoden gibt, so hat doch jede Form ihren Platz im Entwicklungsprozess.

In Hinsicht auf die Validitätsprobleme sollte man sich immer bewusst machen, dass der perfekte Test nicht existiert. Wie am Problem der externen und internen Validität schön zu sehen ist, hat jede positive Änderung beim einen, negative Auswirkungen beim anderen. Es muss also in jedem Fall der goldene Mittelweg gefunden werden und das ist nicht immer einfach. Ein großes Problem liegt hierbei sicherlich darin, dass es im Usability Bereich um Software geht und im Speziellen um den Verkauf solcher. Vor der Marketingabteilung zählen Kompromisse aber oftmals wenig, Ergebnisse, denen zusätzlich eine Einschränkung „aber nur unter der und der Bedingung“ folgt, sind unerwünscht. Eventuell liegt hier auch ein Teil der Problematik begraben - Usability Tests sind kein Forschungsgebiet, sie dienen der Qualitätssicherung eines Produktes, welches an den Mann gebracht werden muss. Dementsprechend effektiv müssen sie sein was verwertbare Ergebnisse angeht, ansonsten sind sie den Aufwand ganz einfach nicht wert.

5. Referenzen

Gerhard Kleining – Das qualitative Experiment (1986)

Gerhard Kleining – Qualitativ-heuristische Sozialforschung: Schriften zur Theorie und Praxis (1994a)

Raphael Rossmann – Das sozialwissenschaftliche Experiment (Skript WS 2001/2002)
<http://www.rafael-rossmann.de/exp.htm>

Peter Atteslander - Methoden der empirischen Sozialforschung (1995)

Prof. Dr. W. Hussy – Einführung in die psychologische Methodenlehre, qualitative Methoden (Skript WS 2001/2002)

Jutta Schäfer – Glossar qualitativer Verfahren, Berliner Public Health Zentrum (1995)

Joseph S. Dumas, Janice A. Redish – A practical guide to usability testing (1999)

Industry USability Reporting Group – Common Industry Format for Usability Test Reports
<http://zing.ncsl.nist.gov/iusr/>

Jakob Nielsen - <http://www.useit.com>, several papers

UsabilityNet - <http://www.usabilitynet.org>

Alexander Reifinger – Seminararbeit “Usability”, Technische Universität München