OTTO-VON-GUERICKE-UNIVERSITÄT MAGDEBURG

Fakultät für Informatik
Institut für Simulation und Graphik

Diplomarbeit

Konzeption und Durchführung einer experimentellen Evaluierung von hervorgehobenen Fokus-Strukturen in der medizinischen Visualisierung

Friederike Adler

Friederike Adler:
Matrikelnummer: 16 55 17
Konzeption und Durchführung einer experimentellen Evaluierung von hervorgehobenen Fokus-Strukturen in der medizinischen Visualisierung
Diplomarbeit, Otto-von-Guericke-Universität
Magdeburg, 2009.

Konzeption und Durchführung einer experimentellen Evaluierung von hervorgehobenen Fokus-Strukturen in der medizinischen Visualisierung

Diplomarbeit

an der
Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von:	Friederike Adler
geb. am:	16. April 1982
in:	Stendal
Matrikelnummer:	16 55 17

Gutachter:	Jun. -Prof. Dr.-Ing. Raimund Dachselt
	Dr.-Ing. Zein Salah
Betreuer:	Prof. Dr.-Ing. Bernhard Preim
	Dipl. Ing. Alexandra Baer

Zeit der Diplomarbeit:	14.03.2009 - 14.08.2009

Selbstständigkeitserklärung

Hiermit versichere ich, Friederike Adler (Matrikel-Nr. 165517) die vorliegende Arbeit allein und nur unter Verwendung der angegebenen Quellen angefertigt zu haben.

Friederike Adler

Magdeburg, den 14. August 2009

Danksagung

Mein größter Dank gilt meinen lieben Eltern und meiner Großmutter Lisa Adler, ohne die dieses Studium und mein Auslandsaufenthalt in Australien nie möglich gewesen wären. Bedanken möchte ich mich auch ganz herzlich bei meinen Betreuern Bernhard Preim und Alexandra Baer, die mich während meiner Diplomzeit hervorragend fachlich unterstützt haben und stets motivierende Worte fanden. Weiterhin möchte ich meinen Korrekturlesern, insbesondere Antje Buttkus und Alexander Kuhn und den Diplom-Psychologen Daniel Lenz und Sascha Tyll für ihre Zeit und hilfreichen Anmerkungen danken. Ganz besonderen Dank gilt auch meinem lieben, geduldigen Freund Florian.

Inhaltsverzeichnis

1 Einleitung

Wird das Immunsystem des Menschen angegriffen, kommt es zur Produktion von Lymphozyten, was ein Anschwellen der Lymphknoten zur Folge hat. Überschreitet die Größe der Lymphknoten einen Durchmesser von 3 *cm*, könnte dies ein Anzeichen für eine schwerwiegende Infektion oder eine Krebserkrankung sein. Um tiefer an den Lymphbahnen liegende vergrößerte Lymphknoten zu erkennen, die sich nicht ertasten lassen, sind Ärzte auf bildgebende Verfahren wie beispielsweise CT, MRT oder Ultraschall angewiesen. Ein Nachteil dieser Verfahren besteht bisher darin, dass die Detektion von pathologischen Lymphknoten oft ungenau und relativ zeitaufwendig ist. Daher wird am Lehrstuhl für Visualisierung an der Otto-von-Guericke-Universität in Magdeburg an der Entwicklung automatischer Detektions- und Segmentierungsverfahren gearbeitet, die Ärzte bei der Diagnose pathologischer Strukturen besser unterstützen sollen.

Neben der Kennzeichnung auffälliger Strukturen in den Bildern bildgebender Verfahren ist die dreidimensionale Darstellung anatomischer Bereiche im Vordringen begriffen. Hierfür werden Volumendaten aus den zur Verfügung stehenden Schnittbildern generiert. Als ein Problem hat sich dabei herauskristallisiert, dass Verfahren zum Hervorheben oder Abschwächen von Fokus- und Kontextstrukturen, wie sie auf Schnittbildern angewendet werden, sich nicht oder nur ungenügend auf dreidimensionale Darstellungen übertragen lassen. Diesbezügliche Untersuchungen in der Computervisualistik, wie sie in Kapitel 2.4 besprochen werden, genügen wissenschaftlichen Evaluierungskriterien nur bedingt.

In der experimentellen Psychologie wird unter anderem mit Hilfe von Reaktionszeittests die Sensitivität des Menschen auf bestimmte Reize untersucht. Bei fachgerechter Durchführung und Auswertung eines solchen Tests ist es unabhängig von der Reaktionsneigung der Befragten möglich, ein Ergebnis zu erhalten, welches der physiologischen Leistungsfähigkeit dieser entspricht.

1.1 Zielsetzung

Durch die Adaption experimenteller Designs von Reaktionszeittests soll eine neue, zuverlässige Methode zur Evaluierung wahrnehmungsunterstützender Visualisierungen entwickelt werden. Ob eine solche Übertragung auf den Bereich der Visualisierung möglich bzw. sinnvoll ist, soll in dieser Arbeit untersucht und diskutiert werden. Hierfür wird

ein experimentelles Design zur Untersuchung von hervorgehobenen Fokusstrukturen bezüglich ihrer Wahrnehmungsunterstützung in medizinischen Visualisierungen konzipiert.

In dieser Arbeit werden die Hervorhebungstechniken *CutAway, Stippling* und *rote Einfärbung* im Hinblick auf die signifikante Unterstützung der visuellen Erfassung vergrößerter Lymphknoten in dreidimensionalen Halsrenderings untersucht und verglichen. Das dafür entwickelte und angewendete Konzept zur experimentellen Evaluierung soll beispielhaft für weitere empirische Untersuchungen in ähnlichen Bereichen sein. Dabei soll es den Evaluierungskriterien der Objektivität, Validität und Reliabilität genügen. Zu diesem Zweck wird das in der Statistik verwendete Konzept zur Hypothesenüberprüfung auf visualisierungsspezifische Hypothesen angewendet. Eine Voraussetzung hierfür ist zum einen die Beachtung verschiedener Validitätskriterien bezüglich der Stichproben, Messinstrumente sowie der Durchführung der Messung. Zum anderen ist das Verständnis der spezifischen statistischen Auswertungsverfahren von großer Bedeutung für deren richtige Anwendung.

Ausgehend von der Fragestellung und und unter Berücksichtigung der Erkenntnisse und Theorien aus der Wahrnehmungspsychologie, sind entsprechende Forschungshypothesen bezüglich der zu untersuchenden Hervorhebungstechniken zu postulieren. Diese werden mit Hilfe der aus dem Reaktionszeittest ermittelten Daten überprüft. Die hieraus gewonnenen Ergebnisse sollen den subjektiven Antworten aus Fragebögen gegenübergestellt und ausgewertet werden.

Zusammengefasst ergeben sich folgende Teilschritte:

- Evaluierung ausgewählter Hervorhebungstechniken für vergrößerte Lymphknoten im Anwendungsbereich der Halsvisualisierungen,
  - ▷ Postulierung überprüfbarer Hypothesen,
  - ▷ Entwurf eines experimentellen Designs für eine valide Überprüfung der aufgestellten Hypothesen,
  - ▷ Implementierung und Durchführung des Experimentes mit einer zuverlässigen Stichprobengröße,
  - ▷ Anwendung inferenz-statistischer und deskriptiver Tests auf den gewonnenen Messdaten mit anschließender Auswertung.
- Gegenüberstellung der Evaluierungsergebnisse und der subjektiven Einschätzungen der Probanden,
- Bewertung des Untersuchungsergebnisses.

1.2 Gliederung der Arbeit

In den in dieser Arbeit zugrunde liegenden Kapiteln werden neben den theoretischen Grundlagen, der Entwurf sowie die Realisierung des durchgeführten Experimentes dargelegt.

In **Kapitel 2** und **3** wird zunächst ein Überblick über einige für das Thema relevante theoretische Grundlagen der medizinischen Visualisierung, der Evaluierung sowie experimenteller Versuchspläne gegeben. Des Weiteren werden wichtige Theorien der visuellen Suche in Bildern sowie für diese Studie wichtigen statistischen Analyseverfahren vorgestellt.

Auf dieser Grundlage werden in **Kapitel 4** dem Untersuchungsgegenstand entsprechende Hypothesen aufgestellt und darauf aufbauend das Konzept des experimentellen Designs vorgestellt. Zudem wird ein Fragebogen erstellt, mit dem die subjektiven Meinungen der Versuchspersonen mit den objektiv erfassten Daten gegenübergestellt werden sollen.

In **Kapitel 5** werden sowohl die Implementierung als auch die Durchführung des im dritten Kapitel entwickelten Konzepts beschrieben.

Die Auswertung der mit dem Experiment erfassten Messdaten sowie die aus den Fragebögen entnommenen Antworten wird mit Hilfe geeigneter statistischer Verfahren in **Kapitel 6** dargestellt.

Abschließend werden in **Kapitel 7** die wichtigsten Erkenntnisse und Evaluierungsergebnisse zusammengefasst und ein Ausblick auf mögliche weiterführende Studien gegeben.

2 Theoretische Grundlagen und verwandte Arbeiten

Im Folgenden soll ein kurzer Überblick über die Entstehung von Visualisierungen von Halsdatensätzen und die in ihnen enthaltenen Strukturen gegeben werden. Auf eine detaillierte Darstellung der einzelnen Entstehungsprozesse oder der Funktionsweise dieser Strukturen wird verzichtet, da sie für das Verständnis der in dieser Arbeit beschriebenen Studie nicht erforderlich sind. Zu bestimmten Sachverhalten wird daher lediglich auf weiterführende Arbeiten verwiesen.

Weiter sollen verschiedene Arbeiten vorgestellt werden, die die Grundlage der gegenwärtig durchgeführten Experimente zur visuellen Wahrnehmung bilden. Abschließend soll auf Arbeiten eingegangen werden, die sich zum einen mit der visuellen Suche in Bildern und zum anderen mit medizinischen Visualisierungen beschäftigen.

2.1 Medizinische Visualisierung von Halsdatensätzen

Die dieser Arbeit zugrunde liegenden Bilder wurden aus dreidimensionalen Volumenvisualisierungen von computertomografischen Halsschichtbildern des Menschen generiert. Dargestellte Strukturen wie Weich- und Knochengewebe in diesen anatomischen Szenen entstammen somit realen anonymisierten Halsdatensätzen. Abbildung 2.1 zeigt die Überführung einzelner zweidimensionaler CT-Schichtbilder in eine solche räumliche Darstellung. Eine derartige Darstellung unterstützt neben der Orientierung innerhalb der Szene auch die räumliche Beurteilung von Lagebeziehungen sowie der Größe von Strukturen. Die für den Beobachter irrelevanten Strukturen können ausgeblendet oder abgeschwächt dargestellt werden. Eine genaue Beschreibung der einzelnen Verfahren zur Volumenvisualisierung findet man in [Preim und Bartz, 2007].

2.1.1 Lymphknoten in medizinischen Visualisierungen

Anhand der Größe eines Lymphknotens kann sich ein Arzt ein Bild darüber machen, inwieweit das Immunsystem eines Menschen angegriffen ist. Gesunde, inaktive Lymphknoten sind in der Regel nur einige Millimeter groß. Bei der Auseinandersetzung mit

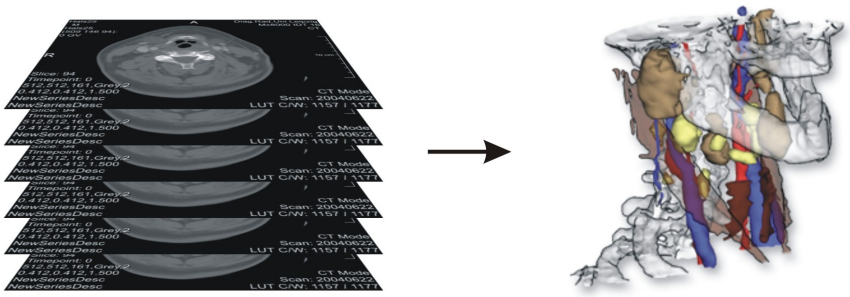
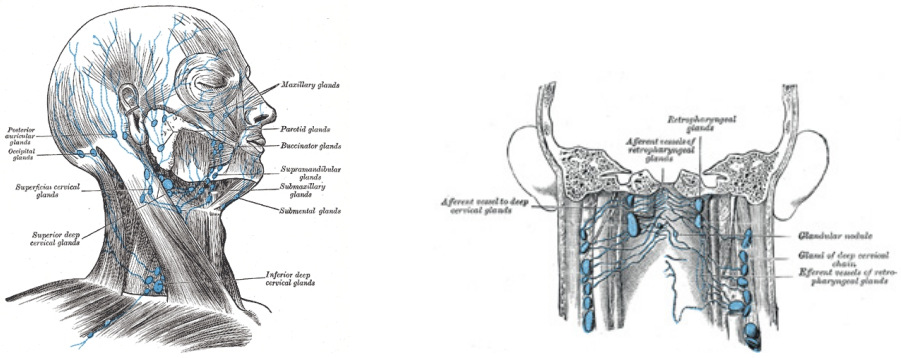


Abbildung 2.1: Bilderstapel eines Halsdatensatzes (links), aus dem relevante Strukturen zunächst segmentiert und anschließend als Oberflächenmodelle (rechts) visualisiert werden.

Krankheitserregern werden Lymphozyten gebildet und die Lymphknoten schwellen an. Wenn deren Größe 30 *mm* überschreitet, könnte dies auf eine schwerwiegende Krankheit wie Leukämie oder maligne Tumoren hindeuten. Im Fall einer Tumorerkrankung können diese Lymphknoten metastieren und andere Strukturen infiltrieren. Die Klassifizierung eines solchen Tumors erfolgt dann anhand seiner anatomischen Ausbreitung und wird TNM-Klassifikation bezeichnet. Diese beschreibt zum einen die lokale Ausbreitung des Tumors (T), dessen Metastasierung in regionäre Lymphknoten (N) und zum anderen das Vorhandensein von Fernmetastasen (M) [Wittekind et al., 2005]. Je früher von Metastasen befallende Knoten erkannt werden, desto größer ist die Chance, die Erkrankung einzudämmen oder zu heilen. Daher sind für die Diagnostik geeignete Visualisierungen von Halsdatensätzen und Hervorhebungen pathologischer Strukturen in diesen von großer Bedeutung.

Die grundlegende Frage bei Visualisierungen ist generell die, worauf der Fokus des Betrachters liegt. Ausgehend von dem Ziel, pathologische Lymphknoten sicher zu erkennen, ist bei der Visualisierung zu bedenken, inwieweit auf die Darstellung des umliegenden Weich- und Knochengewebes verzichtet werden kann. Die alleinige Darstellung eines einzelnen Knotens ist nicht ausreichend, da für seine Beurteilung sowohl Informationen über die Größe als auch die Lage und Infiltration vorliegen müssen. Es ist deshalb sinnvoll, jene Strukturen in die Visualisierung einzubeziehen, mit denen das untersuchte Objekt in Beziehung steht und die der Orientierung innerhalb der Visualisierung dienen. Diese Strukturen werden als Kontextstrukturen bezeichnet. Mit Fragen ihrer Auswahl anhand semantischer Informationen befasst sich [Kellermann, 2009]. Eine ausführliche Beschreibung der Lage und Funktion von Halsstrukturen wird in [Tietjen, 2009] gegeben.

Die Diagnose pathologischer Lymphknoten erfolgt in der ärztlichen Praxis anhand von CT- oder MRT-Schnittbildern. Weitere bildgebende Verfahren zur Lymphknotenlokalisierung sind beispielsweise Sono- und Szintigrafie. Die Detektion vergrößerter Lymphknoten in den bei diesen Verfahren erzeugten Bildern ist aufwendig und erfordert ein gutes räumliches Vorstellungsvermögen. Eine Möglichkeit der Orientierung bietet da-



(a) Lymphbahnen im Hals des Menschen.

(b) Symmetrische Anordnung der Lymphbahnen im Hals.

Abbildung 2.2: Lymphknoten im Halsbereich des Menschen [Gray, 1918, Fig.602, Fig.603].

bei die symmetrische Anordnung der Lymphknoten entlang der Gefäße. Abbildung 2.2 (a) zeigt Lymphknoten und deren Lymphbahnen im Halsbereich. Deren symmetrische Anordnung ist in Bild (b) dargestellt.

2.1.2 Hervorhebungstechniken

Im Ergebnis der bisherigen Forschung gibt es einige gute automatisierte Detektions- und Segmentierungsverfahren für Lymphknoten. Die Hervorhebung dieser Knoten gestaltet sich in Grauwertbildern verhältnismäßig einfach. Generell lassen sich sehr gut Farben zur Strukturfüllung oder Konturenzeichnung verwenden. Weiterhin sind runde oder rechteckige Bounding Boxen üblich, die den fokussierten Bereich eingrenzen. Rein technisch lassen sich genannte Methoden relativ einfach in entsprechenden Volumenvisualisierungen umsetzen. Deren Praktikabilität in 3D-Visualisierungen ist jedoch fraglich. Dies ist vor allem darin begründet, dass die in der Volumenvisualisierung dargestellten Strukturen farbig sind und Strukturen sich gegenseitig verdecken können. Eine farbige Hervorhebung der Fokusstruktur hätte daher nicht die gleiche präattentive Wirkung wie auf Grauwertbildern.

Aus diesem Grund sind Hervorhebungstechniken für Fokusstrukturen in dreidimensionalen medizinischen Visualisierungen erforderlich, die den Blick des Betrachters möglichst schnell auf sich ziehen. Im Folgenden sollen Methoden zur Hervorhebung von Fokusstrukturen vorgestellt werden.

Hervorhebung von Fokusstrukuren

Durch die Hervorhebung von Fokusstrukturen soll erreicht werden, dass die relevanten Strukturen sich deutlich von umliegenden Strukturen abheben und dadurch schneller

vom Betrachter wahrgenommen werden. Generell unterscheiden sich Hervorhebungstechniken hinsichtlich ihrer Einflussgröße bezüglich der involvierten Szene. Mit form- und größenverzerrenden Techniken sollte in medizinischen Visualisierungen vorsichtig umgegangen werden, da durch sie möglicherweise eine falsche Diagnose provoziert werden könnte. In [Preim und Ritter, 2002] wird zwischen lokalen, regionalen und globalen Hervorhebungstechniken differenziert.

Lokale Hervorhebungstechniken verändern bestimmte Merkmale wie Form, Farbe, Größe und Textur des zu fokussierenden Objekts. Beispiele hierfür sind Einfärbungsmethoden, Illustrationen oder Kennzeichnung relevanter Strukturen über Symbole. Da hier benachbarte Strukturen unangetastet bleiben, bleibt die Szene relativ ungestört und unterstützt somit eine weitere Exploration. Lediglich größenverändernde Methoden wirken dem entgegen. Voraussetzung für die Anwendbarkeit dieser Techniken ist, dass die hervorzuhebenden Strukturen nicht durch andere Strukturen verdeckt werden. In diesem Fall sind Verfahren zur Sichtbarmachung verdeckter Strukturen erforderlich, wie sie bei regionalen Hervorhebungstechniken angewendet werden.

Regionale Hervorhebungstechniken beinhalten zumeist Verfahren zur Sichtbarmachung verdeckter Strukturen, wobei der Veränderungsbereich über den der Fokusstrukturen selbst hinausgehen kann. Regionale Hervorhebungstechniken eignen sich daher vor allem in Volumenvisualisierungen. Beispielhaft hierfür sind sogenannte *CutAways*. Eine weitere Möglichkeit besteht darin, verdeckende Strukturen transparent zu gestalten oder den Kontrast benachbarter Strukturen abzuschwächen.

Globale Hervorhebungstechniken tragen zur Änderung der gesamten Szene bei. Rotation und verschiedene Tiefeneinstellungen ermöglichen verschiedene Sichten auf die Visualisierung und innerhalb der Visualisierung. Weitere globale Veränderungen können beispielsweise durch Verzerrungen oder Unschärfen erreicht werden.

Eine ähnliche Unterteilung der verschiedenen Hervorhebungstechniken findet sich auch in [Kosara et al., 2002].

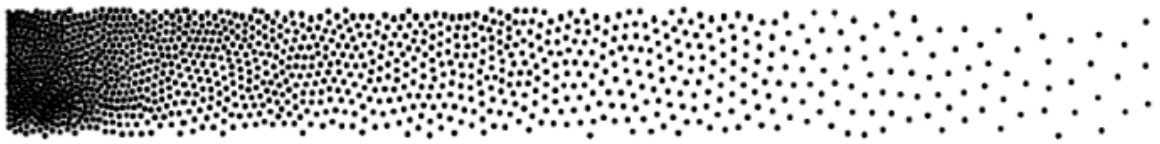
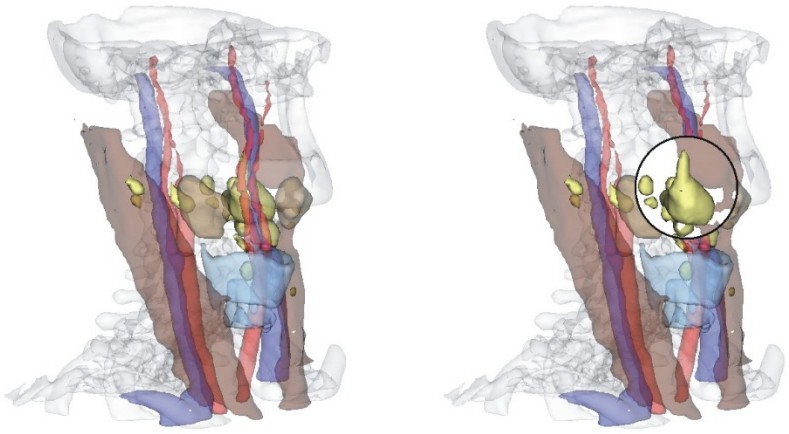


Abbildung 2.3: Die unterschiedliche Punktdichte- und anzahl beim *Stippling* führt zu verschiedenen Helligkeiten (Quelle: [Strothotte und Schlechtweg, 2002]).

Gegenstand dieser Arbeit ist die Evaluierung der lokalen Hervorhebungstechniken *rote Einfärbung* und *Stippling* sowie der regionalen Hervorhebungstechnik *CutAway*, angewandt auf vergrößerte Lymphknoten in 3D-Visualisierungen von Halsdatensätzen. Beim



(a) (b)

Abbildung 2.4: Lymphknoten im Halsbereich des Menschen ohne Hervorhebung (a) und mit *Cut-Away*-Hervorhebung (b).

Stippling werden Form und Helligkeit der Strukturen durch die Punktdichte und -größe charakterisiert, siehe Abbildung 2.3. Die Methode eignet sich vor allem zur Darstellung einfacher Strukturen, über die keine Richtungs- und Krümmungsinformationen vorliegen oder bei denen diese Informationen nicht von Bedeutung sind. Sie ist eine häufig angewendete Methode zur Darstellung anatomischer Strukturen in Anatomie-Atlanten. Bei der *roten Einfärbung* macht man sich die besondere Eigenschaft der roten Farbe zu nutze, als Signalfarbe die Aufmerksamkeit des Betrachters auf sich zu ziehen. Mit *Cut-Away* wird dem Betrachter mit Hilfe eines Sichtkanals Einblick auf Strukturen gegeben, die sonst durch andere Strukturen verdeckt wären, siehe Abbildung 2.4.

2.2 Evaluierungsmethoden

In einer Studie im Jahr 1994 von Lukowicz et al. [1994] wurden wissenschaftliche Artikel im Bereich der Informatik hinsichtlich ihrer Evaluierungsmethoden begutachtet. Dabei wurde festgestellt, dass trotz der ständigen Entwicklungen in den verschiedensten Forschungsgebieten, nur relativ wenige Wissenschaftler gewillt sind diese auf ihre Anwendbarkeit hin zu prüfen.

> „The low ratio of validated results appears to be a serious weakness in computer science research. This weakness should be rectified for the long-term health of the field.“[Lukowicz et al., 1994]

Denning [1980] geht sogar soweit, zu behaupten, dass wenn sich die Informatik weiterhin vor Evaluierungen streube, diese von anderen Wissenschaften nicht mehr Ernst genommen würde.

> „If we do not live up to the traditional standards of science, there will come a time when no one takes us seriously.“[Denning, 1980]

Mittlerweile bemühen sich immer mehr Wissenschaftler ihre Arbeiten zu evaluieren und deren Ergebnisse öffentlich zu präsentieren [Ritter et al., 2006], Tory und Möller [2005]. Welche Kriterien hierbei zu beachten sind, soll im Folgenden geklärt werden.

Die zu entwickelnde Evaluierungsmethodik muss daher folgenden, auch von der DeGEval [2009] und proEval [2009] anerkannten Kriterien für die Erfassung und Auswertung von Daten genügen:

- ***Objektivität***
  Objektivität erfordert strenge Sachlichkeit und Darstellung unter größtmöglicher Ausschaltung des Subjektiven. Objektivität ist dann gegeben, wenn die Datenerfassung und -verwertung von einer Person durchgeführt wird, die selbst nicht Subjekt der Studie und somit nicht in der Menge der erfassten Daten enthalten ist. Der Versuchsleiter sowie der Versuchsauswerter haben lediglich die Position des Beobachters.

- ***Reliabilität***
  Der Reliabilität wird die Objektivität vorausgesetzt. Sie bestimmt die Zuverlässigkeit und Messgenauigkeit einer Evaluation. Eine Untersuchung ist genau dann reliabel, wenn sie durch mehrmalige Wiederholung zu verschiedenen Zeiten und mit unterschiedlichen Personen immer die gleichen Resultate liefert.

- ***Validität***
  Unter Validität wird zunächst die Gültigkeit des Messverfahrens in Bezug auf das zu untersuchende Merkmal verstanden. Sie setzt eine ausreichende Kontrolle der Störvariablen (siehe Begriffsklärung in Abschnitt 3.1) sowie geeignete Analyseverfahren voraus, was der internen Validität entspricht. Validität in Bezug auf andere, ähnliche Bereiche liegt dann vor, wenn der Versuchsaufbau und die Ergebnisse einer Untersuchung auf diese Bereiche übertragbar sind (externe Validität).

Im Folgenden sollen Evaluierungsmethoden vorgestellt werden, die für die Evaluierung medizinischer Anwendungen in Frage kommen können.

2.2.1 Formale und Heuristische Evaluierung

Für die Entwicklung eines konstruktiven Interaktionssystems innerhalb einer dreidimensionalen anatomischen Szene ist es sinnvoll eine Evaluierungsmethode zu wählen, die sich vor allem mit den möglichen Aufgaben und Zielen eines solchen Systems auseinandersetzt. Hierfür sind die dafür vorgesehenen aufeinanderfolgenden Aktionen sowie

die damit einhergehenden Methoden und empirisch gemessenen, aufgebrachten Zeiten zu untersuchen. Gibt es verschiedene Methoden zum Erreichen eines Ziels, kann eine entsprechende Selektion dieser erfolgen. Diese formale Evaluierung durch Ziel-, Operatoren,- Methoden- und Selektionregelbetrachtung basiert auf dem GOMS-Modell (goals, operators, methods, selection rules)[Card et al., 1983] und wird auch von Preim [1999] in Bezug auf die Entwicklung interaktiver Systeme erläutert.

Eine weitere Methode zur Überprüfung der Usability eines Systems ist die informale Evaluierung durch Experten beziehungsweise geübten Nutzern mittels zuvor definierter Heuristiken und unter Berücksichtigung bekannter Zusammenhänge. Vorteil hierbei ist, dass die heuristische Evaluierung vom Entwurf bis hin zur Umsetzung eines zu evaluierenden Systems durchgeführt werden kann. In [Tory und Möller, 2004] und [Kobsa, 2004] wird deutlich wie wichtig diese Formen der Evaluierung für die Entwicklung interaktiver Systeme sind. Zudem wird auf Arbeiten verwiesen in denen formale und heuristische Evaluierungen erläutert und durchgeführt wurden.

Der Untersuchungsgegenstand dieser Arbeit schließt diese Formen der Evaluierung jedoch aus, da kein zu evaluierendes Interaktions-System vorliegt.

2.2.2 Empirische Evaluierung

In der empirischen Evaluierung können aufkommende Fragestellungen entweder qualitativ oder quantitativ untersucht werden. Unter qualitativen Forschungsmethoden werden Methoden der subjektiven Datenerfassung aus beispielsweise Interviews und Protokollen bezeichnet. Dieses Verfahren findet häufig Anwendung in den Sozialwissenschaften, wird aber von vielen anderen Wissenschaften aufgrund der Verletzung der zuvor genannten Gütekriterien abgelehnt. Um diesen Kriterien nach Möglichkeit gerecht zu werden sind quantitative Methoden, die zumeist nach bestimmten Vorgaben durchzuführen sind, vorzugswürdig. Die Datenerhebung erfolgt in der Regel in Form von Experimenten, aber auch subjektive Methoden der Datenerhebung wie standardisierte Fragebögen und Interviews sind möglich. Wichtig ist hierbei, dass diese durchgeführten Methoden jederzeit reproduzierbar sind und im Fall einer Wiederholung zu den gleichen Erkenntnissen gelangt werden kann. Deskriptive Darstellungen in Tabellen oder Grafiken sind dabei eine beliebte aber unzureichende Form der Ergebnispräsentation. Darüber hinaus sind verschiedene inferenzstatistische Signifikanztests anzuwenden und deren Ergebnisse darzulegen.

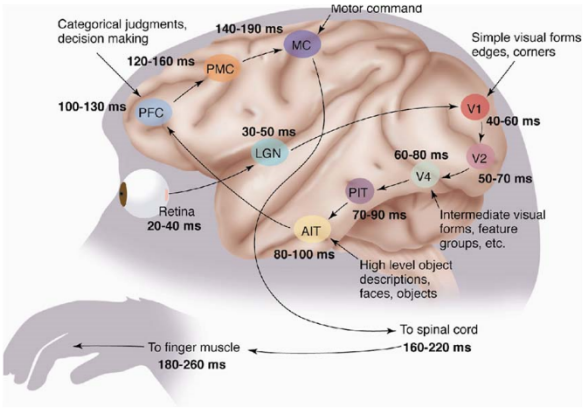
Bisherige medizinische Visualisierungsverfahren wurden häufig durch Befragungen der Anwender in Form von standardisierten Fragebögen evaluiert [Tietjen, 2004], [Mirschel, 2004], [Neugebauer, 2006], [Hansen, 2006]. Diese Form der Evaluierung ist einfach zu wiederholen und wird sehr wahrscheinlich immer zu den gleichen Ergebnissen kommen, sofern Personen des selben Anwendungsbereichs befragt werden, so dass mit reliablen Ergebnissen gerechnet werden kann. Die in einigen dieser Arbeiten

deskriptiven Darstellungen der Ergebnisse würden zudem mit Hilfe statistischer Analyseverfahren eine höhere Validität erreichen wie sie von Hansen [2006] und Ritter et al. [2006] bei der Evaluierung von Gefäßvisualisierungen erzielt wurden.

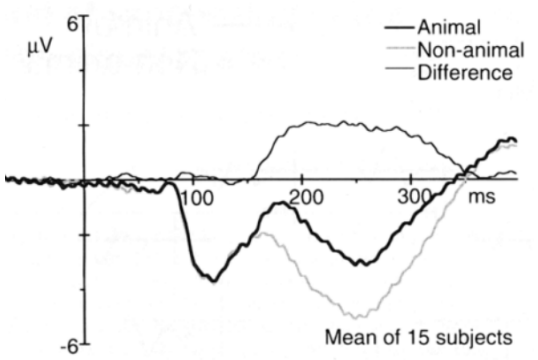
2.3 Visuelle Suche in Bildern

In der Wahrnehmungspsychologie werden häufig Reaktionszeittests, teilweise auch in Kombination mit MEG-, EEG- oder MRT-Messungen, durchgeführt. Die Probanden werden hierbei einer Folge von auditiven oder visuellen Reizen, sogenannte Stimuli, ausgesetzt. Die Reaktionen der Versuchspersonen auf diese Reize sind Untersuchungsgegenstand der Studie. Die in der experimentellen Psychologie verwendeten visuellen Stimuli sind häufig relativ einfach aufgebaut und entsprechen vom Komplexitätsgrad den in Abbildung 2.6 dargestellten Bildern.

Thorpe et al. [1996] wiesen in einem psychophysischen EEG-Experiment nach, dass Menschen in der Lage sind innerhalb von 150 ms Objekte wahrzunehmen. Die Aufgabe der Versuchspersonen bestand in seinem Experiment darin zu entscheiden, ob ein ihnen präsentiertes Bild ein Tier enthielt oder nicht. Während des Experimentes waren die Versuchspersonen an ein EEG angeschlossen, mit dem die verschiedenen Hirnaktivitäten während der Entscheidungsfindung gemessen wurden. Abbildung 2.5 (a) zeigt die



(a) Verarbeitungsprozesse im Gehirn des Menschen bei der Objektwahrnehmung.

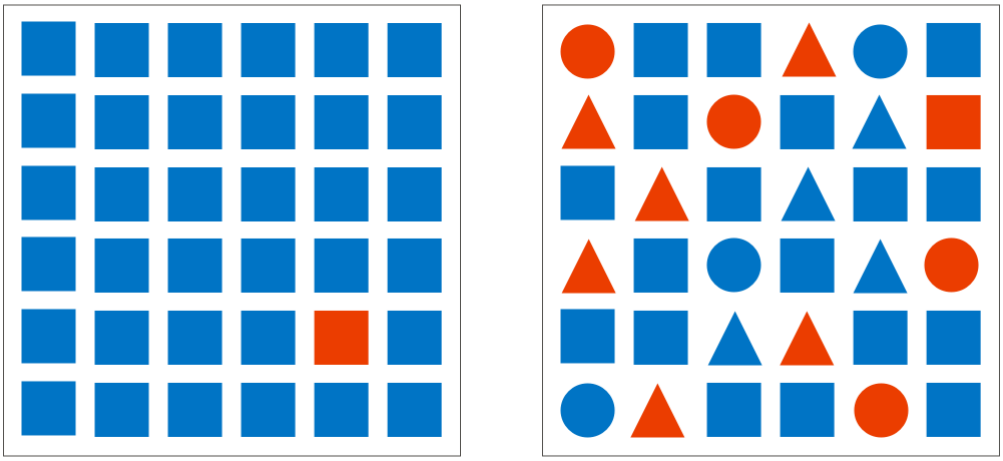


(b) Ereigniskorrelierte Potentiale der Vpn für Bilder mit Zielobjekt und ohne Zielobjekt.

Abbildung 2.5: Bild (a) zeigt die einzelnen Verarbeitungsschritte sowie die dafür benötigten Zeiten bei der Objektwahrnehmung. Diese Prozesse können anhand der in einem EEG aufgezeichneten ereigniskorrelierten Potentiale den entsprechenden Gehirnregionen zugeordnet werden [Thorpe et al., 1996].

nacheinander ablaufenden Prozesse im Gehirn einer Vpn bei der Wahrnehmung eines Objektes. In den ersten Prozessen ($20 - 100$ ms) wird ein Objekt zunächst visuell erfasst und anhand bestimmter Eigenschaften kategorisiert. Bereits nach 150 ms kann

eine Entscheidung dahingehend getroffen werden, ob das wahrgenommene Objekt dem Zielobjekt entspricht oder nicht. Eine der Aufgabenstellung entsprechende Reaktion kann dann innerhalb von weiteren 100 *ms* ausgeführt werden. In Abbildung 2.5 (b) sind die im EEG erfassten ereigniskorrelierten Potentiale (EKP) innerhalb der ersten 350 *ms* nach Stimulus Darbietung dargestellt. Hier ist deutlich zu erkennen, dass der Erkennungsprozess beider Stimuli anfangs nahezu gleich verläuft. Nach 150 *ms* ist ein Unterschied zwischen den Potenzialen zu erkennen, die Versuchspersonen hat nun eine Entscheidung hinsichtlich des Stimulus getroffen und wird dementsprechend reagieren. Das Experiment und die daraus gezogenen Schlüsse können jedoch lediglich die einzelnen Verarbeitungsprozesse des Gehirns eines Menschen bei direkter Reizdarbietung erklären. Bei der visuellen Suche von Zielobjekten unter einer Vielzahl von Objekten laufen diese Verarbeitungsschritte wesentlich langsamer ab, da vom Gehirn mehrere Reize und Informationen verarbeitet werden müssen.



(a) Parallele Suchaufgabe (b) Serielle Suchaufgabe

Abbildung 2.6: Das in Bild (a) zu suchende Zielobjekt *rotes Quadrat* unterscheidet sich von den umliegenden Objekten nur hinsichtlich der Eigenschaft Farbe und kann ohne besondere Aufmerksamkeit des Betrachters wahrgenommen werden. Bei der seriellen Suchaufgabe in Bild (b) hingegen muss der Betrachter aufmerksam das Bild nach dem Zielobjekt *rotes Quadrat* absuchen, da mehrere Distraktoren mit derselben Farbe und Form auftauchen.

Eine der etabliertesten Theorien der visuellen Suche ist die Merkmalsintegrationstheorie (FIT = Feature Intergration Theory) von Treisman und Gelade [1980]. Nach dieser Theorie unterliegt die visuelle Suche zwei verschiedenen Wahrnehmungsprozessen. Der erste Prozess ist die Wahrnehmung und Lokalisation einer bestimmten Eigenschaft und wird als parallele Suche bezeichnet. Hebt sich das gesuchte Objekt lediglich in einer Eigenschaft von anderen umliegenden Objekten ab, kann dieses präattentiv und unter 250 *ms* wahrgenommen werden, da es in der Regel sofort ins Auge sticht und keine gerichtete Aufmerksamkeit des Betrachters fordert. Die präattentive Wahrnehmung einfacher Eigenschaften ist laut *Treisman* unabhängig von der Anzahl der umliegenden

Distraktoren sowie von der Größe der dargestellten Szene. Ist das gesuchte Objekt hingegen durch mehrere Eigenschaften charakterisiert und gleicht es in einem dieser Eigenschaften einem anderen Objekt, muss das Bild seriell nach diesen Eigenschaften abgesucht werden bis das gesuchte Objekt gefunden wurde. Bei diesem Prozess ist die Aufmerksamkeit gerichtet, weil eine Beurteilung aller betrachteten Objekte notwendig ist. Da bei dieser Suche mehrere Eigenschaften eines Objektes betrachtet werden, wird diese als Konjunktionssuche bezeichnet. Abbildung 2.6 zeigt jeweils ein Beispielbild einer parallelen Suche (a) und einer seriellen Suchaufgabe (b). Treisman und Gelade [1980] zeigten in einer Vielzahl von Experimenten, dass bei einer konjunktiven Suche mehr Zeit benötigt wird als bei der parallelen Suche. Darüber hinaus stellten sie in ihren Experimenten zur seriellen Suche fest, dass mit zunehmender Größe des Suchfeldes der Suchaufwand steigt.

Wolfe et al. [1989] und Kosara et al. [2003] konnten jedoch in einer Reihe von experimentellen Versuchen nachweisen, dass nicht jede Konjunktionssuche mehr Zeit beansprucht als eine vergleichbare parallele Suche. Insbesondere zeigten Wolfe et al. [1989] dass Objekte, welche durch drei verschiedene Eigenschaften charakterisiert sind schneller wahrgenommen werden können, als einfache Konjunktionen zweier Eigenschaften. Zudem zeigten die Konjunktionen von Farbe und Form, Farbe und Orientierung sowie Farbe und Größe einen ähnlichen Suchaufwand, wie die der parallelen Suche.

2.4 Evaluierungen in der Visualisierung

Der heutigen medizinischen Diagnostik liegen zumeist zweidimensionale Bilder zugrunde. Die dreidimensionale Darstellung dieser Bilder sollen den Betrachter das Verständnis der in diesen Bildern dargestellten Strukturen und deren Beziehungen zueinander erleichtern.

Tory [2003] entwickelte ein Verfahren (ExoVis) zur Darstellung globaler und lokaler Informationen in 3D-Visualisierungen mittels Kombination von 2D- und 3D-Ansichten. In einer experimentellen Benutzerstudie wurde dieses Verfahren mit Verfahren verglichen, in denen ebenfalls 2D- und 3D-Ansichten miteinander kombiniert sind. Abbildung 2.7 zeigt mögliche Anwendungsbereiche der in der Studie von Tory [2003] verglichenen Verfahren (oben), repräsentiert durch die unten dargestellten Testbilder. Die zu vergleichenden Verfahren waren zum einen die *in-place*-Methode *Cutting Plane* und zum anderen eine Methode in der 2D- und 3D- Ansichten mit *Orientation Icons* getrennt dargestellt werden und dem von *Tory* entwickelten *out-of-place*-Verfahren *ExoVis*. Ziel dieser Studie war es herauszufinden, welches dieser Verfahren am besten geeignet ist, korrespondierende Punkte zwischen den verschiedenen Ansichten zu lokalisieren. In *Tory´s* Studie wurden mehreren Versuchspersonen jeweils eine $3 \times 3 \times 3$ Würfelkombination mit 22, 25 oder 19 unterschiedlich angeordneten Würfeln gezeigt, wobei immer

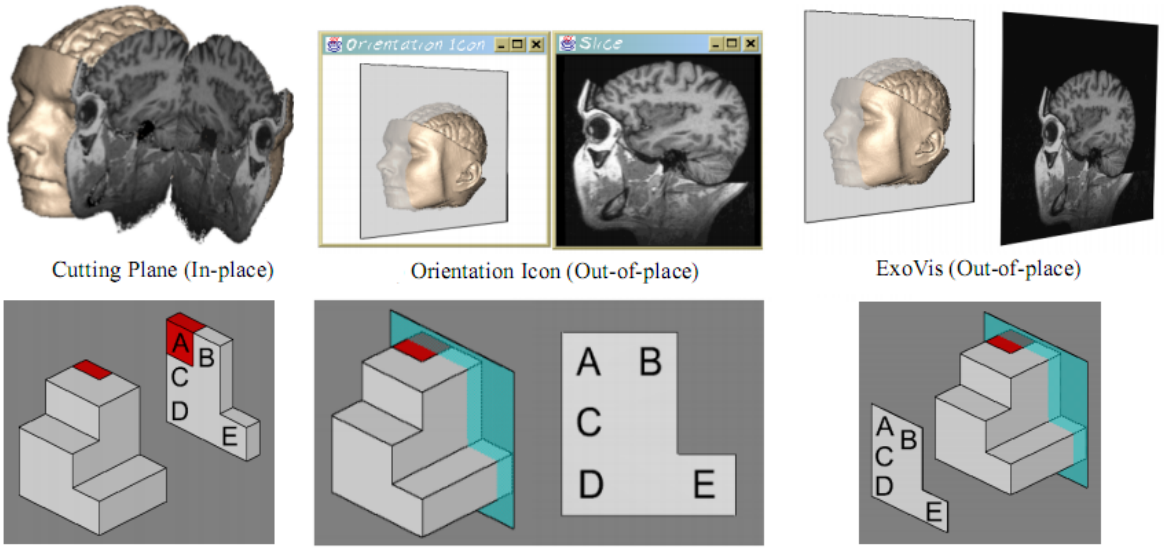
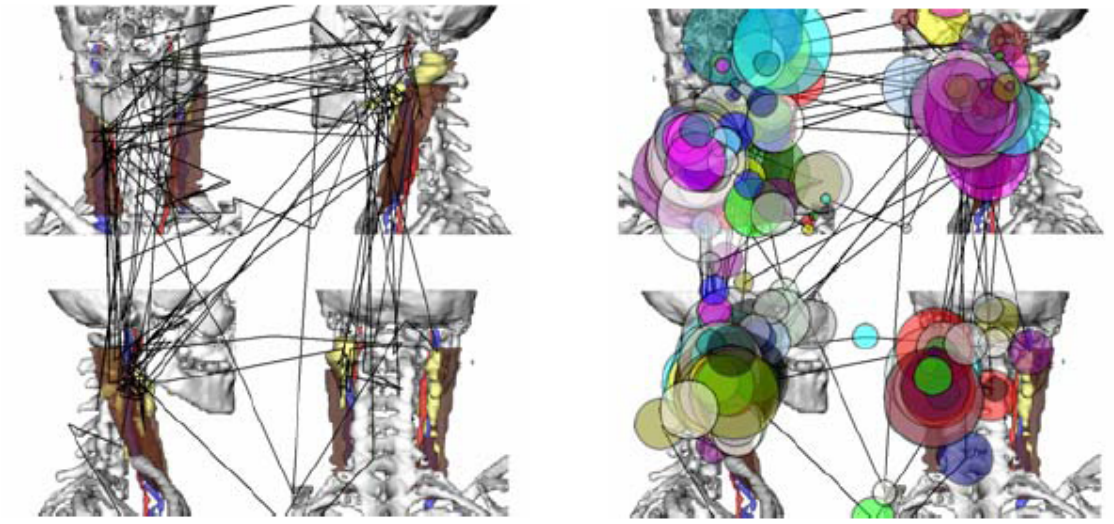


Abbildung 2.7: Medizinische Anwendungsbilder mit den Verfahren *Cutting Plane*, *Orientation Icon* und *ExoVis* und den entsprechenden Testbildern mit isometrisch projezierten Würfelkombinationen (3D-Ansicht) und korrespondierenden 2D-Ansichten mit Buchstaben [Tory, 2003].

ein Würfel rot dargestellt war. Zusätzlich wurde entsprechend dem verwendeten Verfahren eine 2D-Ansicht dieser Würfelkombination gezeigt, in der die Buchstaben von A bis E enthalten waren. Aufgabe der Versuchspersonen war es dem rot hervorgehobenen Würfel der 3D-Ansicht dem in der 2D-Ansicht entsprechenden Buchstaben zu zuordnen. Bewertet wurden die untersuchten Verfahren nach Genauigkeit, der für die Aufgabenbewältigung benötigte Zeit und der subjektiven Meinungen der Versuchspersonen, welche in einem Fragebogen aufgenommen wurden.

Alle drei Visualisierungsverfahren zeigten statistisch signifikante Unterschiede hinsichtlich der gemessenen Zeiten zur Aufgabenbewältigung, der Zuordnungsgenauigkeit der roten Würfel zu den Buchstaben und der subjektiven Bewertungen der Verfahren. Dabei erwies sich *Cutting Plane* als die am besten geeignete und *Orientation Icon* als die ungeeignetste Methode zur parallelen Darstellung von 2D- und 3D-Informationen. *Tory's* Arbeit besticht durch eine hohe Transparenz bezüglich der von ihr angewendeten statistischen Verfahren und sie gibt diesbezüglich kurze Erläuterungen und Begründungen für die Auswahl dieser. Die von ihr durchgeführte Studie entsprach den objektiven, den zuverlässigen und den validen Gütekriterien.

Darüber hinaus wurde nachgewiesen, dass die Verwendung unterschiedlicher 2D-Ansichten (frontal, seitlich, Sicht von oben) sich ebenfalls auf die Aufgabenbewältigung auswirkt. Dies wurde zum einen durch eine subjektive Umfrage und zum anderen mit Hilfe der experimentellen Benutzerstudie am Beispiel der drei Verfahren nachgewiesen. Würfelkombinationen deren korrespondierende 2D-Ansicht von oben dargestellt wurde verursachte die meisten Fehlzuweisungen und beanspruchte zudem mehr Zeit.



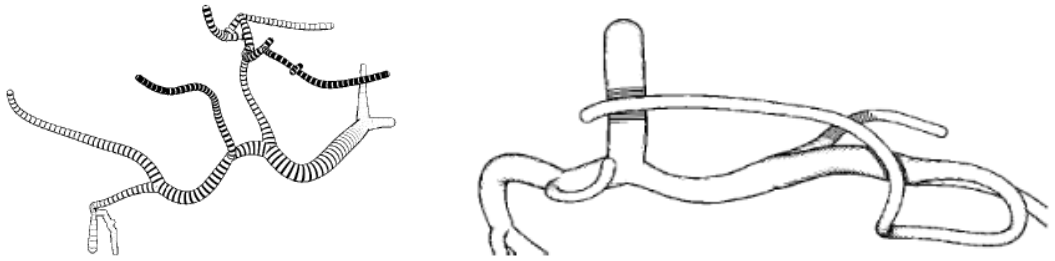
(a) Betrachtungs-Pfad (b) Fixierungsdauer

Abbildung 2.8: Der Betrachtungs-Pfad der Versuchspersonen ist in Bild (a) dargestellt. Wie lange die Versuchspersonen einen bestimmten Bereich fixiert haben, ist in Bild (b) zu sehen [Burgert et al., 2007]. Dabei stellen die verschiedenen Farben die unterschiedlichen Fokussierungsbereiche und die Kreisdurchmesser die Dauer der Fixierung dar.

Im Gegensatz zu *Tory* versuchten Ritter et al. [2006] und Burgert et al. [2007] Evaluierungen mit Bildern, wie sie auch in der medizinischen Praxis, verwendet werden, durchzuführen. Diese Art von Evaluierung ist aufgrund der Komplexität und Spezifität der Bilder schwer mit unerfahrenen Personen durchzuführen. Daher bestand deren Stichprobe überwiegend aus erfahrenen Ärzten.

Burgert et al. [2007] untersuchten 13 Ärzte auf deren serielles Betrachtungsverhalten von 3D-Halsvisualisierungen. Mit Hilfe eines Eye-Tracking-Systems war es ihnen möglich die Blickerichtungen sowie die Fixierungsdauer bestimmter Bereiche der Ärzte zubeobachten und auszuwerten. Den Versuchspersonen wurden vier verschiedene Halsvisualisierungen mit unterschiedlichen Farb- und Transparenzdarstellungen der enthaltenen Strukturen. Jede diese Visualisierung wurde in vier verschiedenen Ansichten (frontal, von der linken und rechten Seite, dorsal) in einem Bild gezeigt. Während die Ärzte Fragen bezüglich der ihnen präsentierten Bilder beantworten mussten, wurde deren Blickverhalten aufgezeichnet. Abbildung 2.8 zeigt eine der vier gezeigten Halsvisualisierungen und den vom Eye-Tracking-System aufgezeichneten Pfad der Blickrichtungen der Versuchspersonen (a). In Bild (b) sind die am häufigsten und längsten fixierten Bereiche der Halsvisualisierungen dargestellt. Auch wenn keine statistischen Analyseverfahren auf die Ergebnisse angewendet wurden, wahrscheinlich aufgrund der geringen Stichprobengröße, ist ein deutlicher Unterschied anhand der durchschnittlichen Fixierungsdauer für die verschiedenen Ansichten erkennbar. Am längsten verweilten die Ärzte bei der ‚links oben dargestellten, frontalen Ansicht der Halsvisualisierung.

Dies kann jedoch auch darauf zurückzuführen sein, dass die Exploration beim Menschen häufig in Leserichtung erfolgt.



(a) Variierende Strich-Stärken zur Tiefenkodierung

(b) Tiefenkodierte Abstände zwischen Überschneidungen

Abbildung 2.9: Bilder zur Tiefenwahrnehmung unter Anwendung von Schraffur- und Schattierungsverfahren auf Gefäßstrukturen [Ritter et al., 2006].

Ritter et al. [2006] führten ein web-basierte Nutzerstudie durch, in der Visualisierungstechniken zur Tiefenkodierung von Gefäßstrukturen evaluiert wurden. Den 160 teilnehmenden Versuchspersonen (darunter waren 38 Mediziner) wurden Bilder von Gefäßbäumen gezeigt, in denen sie entweder die Tiefe der einzelnen Gefäßteilbäume oder von Gefäßüberschneidungen abschätzen mussten, siehe Abbildung 2.9. In ihrem Paper stellen die Autoren neben den evaluierten Techniken auch den Aufbau, die verwendeten Bilder sowie den Ablauf des Experimentes vor. Die in dieser Studie objektiv und subjektiv erfassten Messdaten und Auswertungsergebnisse wurden sowohl deskriptiv als auch mit statistischen Kenngrößen, wie der verwendeten Prüfgröße und der Signifikanz des Evaluierungsergebnisses, dargestellt.

Inwieweit die Studie von Ritter et al. [2006] den Gütekriterien genügt, kann nur schwer beurteilt werden, da eine Kontrolle der Störvariablen in einer web-basierten Nutzerstudie kaum möglich ist. Zudem kann allein durch die Verwendung verschiedener Computer und Bildschirme das Ergebnis beeinflusst worden sein.

2.5 Zusammenfassung und Schlussfolgerung

In diesem Kapitel wurde auf die Entstehung dreidimensionaler anatomischer Visualisierungen aus Schnittbildern sowie die Problematik der Lokalisation pathologischer Lymphknoten innerhalb von Halsvisualisierungen eingegangen. Weiterhin wurden Techniken vorgestellt, die unter anderem zur Hervorhebung von Fokusstrukturen entwickelt wurden. Ob diese die Wahrnehmung vergrößerter Lymphknoten tatsächlich unterstützen, soll anhand einer experimentellen Benutzerstudie herausgefunden werden, wie sie

in der Psychologie üblich ist und in einigen Bereichen computergrafischer Visualisierungen bereits angewendet wird. Zwei ausgewählte lokale und eine regionale Hervorhebungstechnik sollen auf ihre Eignung als Hervorhebungstechnik für Fokusstrukturen untersucht werden. Eine Vielzahl an Probanden soll die Zuverlässigkeit der erhobenen Daten gewährleisten. Die Objektivität soll im Rahmen eines psychophysischen Experimentes garantiert werden. Die geeignete Auswahl von Auswertungsverfahren sowie die Unterdrückung von ergebnisverzerrenden Störeinflüssen soll der Studie eine hohe Validität verleihen.

Für die Entwicklung einer zuverlässigen Versuchsanordnung sind die in Abschnitt 2.3 gewonnen Erkenntnisse zur visuellen Reizwahrnehmung und -suche bei Menschen zu berücksichtigen. So entspricht die Lokalisation von vergrößerten Lymphknoten innerhalb einer dreidimensionalen Halsvisualisierung mit vielen ähnlichen Strukturen einer Konjunktionssuche, da jeder Lymphknoten nach seiner Form und Größe beurteilt werden muss. Inwiefern diese Suche mit Hilfe von Hervorhebungstechniken unterstützt wird, ist Gegenstand dieser Arbeit.

Nach dem Vorbild der in Ritter et al. [2006] und Burgert et al. [2007] durchgeführten Studien werden die zu untersuchenden Hervorhebungstechniken ebenfalls in anwendungsrelevanten Bereichen, in anatomischen Visualisierungen evaluiert.

3 Grundlagen der experimentellen Evaluierung

Die experimentelle Hypothesenprüfung ist eine Forschungsmethode, welche heute in den verschiedensten Wissenschaftszweigen angewendet wird, wie zum Beispiel in der experimentellen Archäologie, der experimentellen Biologie und der experimentellen Physik sowie in den Geistes- und Sozialwissenschaften. Dabei unterliegen alle denselben Anforderungskriterien. Dazu zählt, dass die Experimente unabhängig vom Versuchsleiter sowie von Zeit und Ort reproduzierbar sein müssen. Weiterhin sind Objektivität und Messbarkeit der zu untersuchenden Größe gefordert. Dennoch unterscheiden sich die jeweiligen Disziplinen in der Durchführung ihrer Experimente. Während Otto von Guericke 1654 die auf Aristoteles und Galilei zurückgehende Hypothese „Horror vacui“ [1] anhand von öffentlichen Demonstrationen widerlegte, sind beispielsweise Experimentalpsychologen und Sozialwissenschaftler auf quantitative Methoden angewiesen, um zuverlässige Aussagen bezüglich aufgestellter Hypothesen treffen zu können.

Das Konzept einer solchen Forschungsmethode wird als experimentelles Design bezeichnet. Um in Abschnitt 4 dieser Arbeit den Entwurf des Versuchsdesigns verständlich darstellen zu können, werden im Folgenden zunächst grundlegende Begriffe definiert und erläutert sowie die dafür verwendeten Abkürzungen benannt.

3.1 Begriffsdefinitionen

In den meisten Fällen liegt einem Experiment eine bestimmte Fragestellung oder Hypothese zugrunde. Was bei der Aufstellung einer Hypothese zu beachten ist, soll anhand der folgenden Begriffsklärungen verdeutlicht werden [Huber, 2005].

Hypothesen: Im Unterschied zur einfachen Fragestellung beinhaltet eine Hypothese eine bestimmte Annahme über den Ausgang eines durchzuführenden Experimentes.

[1] Horror vacui ist lateinisch und bedeutet „Abscheu vor der Leere“. Der Philosoph Aristoteles (†322 v.Chr.) nahm an, dass die Natur vor leeren Räumen zurückschrecke und deshalb Gase und Flüssigkeiten von der Leere angezogen werden, um diese zu füllen. Der Physiker Galileo Galilei (†1642) vertrat dessen Meinung. Der Magdeburger Otto von Guericke(†1686) widerlegte dies mit seinem Magdeburger Halbkugel-Experiment.

Diese Annahme ergibt sich häufig aus ähnlichen Forschungsarbeiten, fachspezifischen Recherchen, Beobachtungen oder Erfahrungen. Die Untersuchungsobjekte des Experimentes bilden die Hypothesenmerkmale.

Operationalisierung: Unter Operationalisierung wird die Deklaration der Hypothesenmerkmale in unabhängige und abhängige Variablen verstanden. Im Rahmen der Operationalisierung werden die zu verwendenden Messgrößen, Messinstrumente und Prüfparameter festgelegt.

Unabhängige Variablen (UV): Untersuchungsobjekte in einem Experiment, die nicht auf definierte Einflussfaktoren (Stimuli) reagieren, sondern von diesen unabhängig sind, werden als unabhängige Variablen oder Faktoren bezeichnet. Liegen diese in verschiedenen Variationen (Faktorstufen) vor, sind in einem Experiment unterschiedliche Beobachtungen und Messergebnisse zu erwarten.

Abhängige Variablen (AV): Die Variation einer unabhängigen Variable, beispielsweise eines Stimulus, bestimmt die Reaktion auf diesen, die abhängige Variable. Die AV können beispielsweise durch die Reaktionsgenauigkeit, -latenz, -dauer oder -häufigkeit repräsentiert werden.

Nicht nur menschliche Reaktionen können experimentell erfasst werden. Sollen beispielsweise die Laufzeiten unterschiedlicher Algorithmen miteinander verglichen werden, entsprechen die Algorithmen den UV und die gemessenen Laufzeiten den AV.

Stichprobe: Mit der Stichprobe wird die Auswahl der Untersuchungsobjekte für das Experiment definiert. Handelt es sich dabei um Versuchspersonen (Vpn), so ist oft deren Verhalten Gegenstand der Untersuchung. Um ein wissenschaftlich fundiertes Ergebnis eines Experimentes zu erzielen, ist es erforderlich, dass die Stichprobe einer von den statistischen Analyseverfahren geforderten Mindestgröße entspricht. Können keine Personen aus der Zielpopulation gewonnen werden, sind Personen zufällig auszuwählen und als Zufallsstichprobe zu bezeichnen. Grundsätzlich wird bei Stichprobenvergleichen zwischen unabhängigen und abhängigen beziehungsweise ungepaarten und gepaarten Stichproben unterschieden. Vpn der einen Stichprobe sind entweder unabhängig von den Vpn der zu vergleichenden Stichprobe oder werden genau einer Vpn zugeordnet. Die häufigsten abhängigen Stichprobengruppen ergeben sich aus Experimenten mit Messwiederholungen.

Between-subject- und within-subject-Design: Je nach Versuchsaufbau können Vpn in verschiedene Untersuchungsgruppen eingeteilt werden. Handelt es sich um ein *between-subject*-Design, werden die Vpn entweder einer Experimentalgruppe EG oder einer Kontrollgruppe KG zugewiesen. Lediglich die Vpn der Experimentalgruppen werden mit jeweils einer Faktorstufe der UV konfrontiert und anschließend die AV der Gruppen miteinander verglichen. Im Unterschied dazu werden beim *within-subject*-Design sämtliche Vpn allen Variationen der UV ausgesetzt.

Das dieser Arbeit zugrunde liegende Experiment wird entsprechend dem *within-subject*-Design durchgeführt. Das bedeutet, dass sämtliche Vpn allen Variationen der UV ausgesetzt werden und somit in jeder Experimental- und Kontrollgruppe enthalten sind. Es ergeben sich hierdurch abhängige Stichprobengruppen.

Störvariablen: Alle Einflüsse, die das Ergebnis einer Versuchsreihe verzerren können, werden Störvariablen genannt. Darunter fallen sowohl Zustände bestimmter Umgebungskomponenten als auch Motivation und Eignung der Vpn. Positionseffekte bezeichnen Störeinflüsse auf die AV, welche durch Ermüdungserscheinungen bei langen Experimentzeiten aufkommen können. *Carry-Over*-Effekte hingegen entstehen bei fester Abfolge der Faktorstufen. Mit Hilfe von Parallelisierung, Randomisierung sowie Eliminierung ist eine Kontrolle der Störvariablen möglich. Unter Parallelisierung versteht man die Gleichverteilung der Vpn beispielsweise nach Alter, Geschlecht und Beruf auf die Untersuchungsgruppen. Positions- und *Carry-Over*-Effekte können durch Randomisierung der Variationen einer UV unterdrückt werden. Umgebungseinflüsse können, soweit möglich, eliminiert werden.

Analyseverfahren: Wurde eine Hypothese postuliert, ist für deren Überprüfung ein entsprechendes Analyseverfahren anzuwenden. Diese Verfahren überprüfen nur die für die Hypothese relevanten Variablen. Liegt dem Experiment keine Hypothese zugrunde, können verschiedene explorative Analyseverfahren angewendet werden, deren Ergebnisse die Grundlage für eine Hypothese schaffen. Jedes Analyseverfahren ist an bestimmte Voraussetzungen der Messdaten geknüpft. Deren Vorliegen ist vor jeder Anwendung eines Verfahrens zu prüfen, um die Anwendbarkeit dieser zu gewährleisten.

Skalenniveau der Messdaten: Skalenniveaus beschreiben die Eigenschaften der zugrunde liegenden Messdaten. Die einfachste Skalierung ist die der Nominalskala. Auf diesem Niveau können Variablen lediglich beschrieben oder kategorisiert werden und unterliegen zumeist keiner bestimmten Ordnung oder Wertung. Nominalskalierte Variablen können zudem nur über eine Häufigkeitsauszählung ausgewertet werden. Unterliegen die Messdaten einer bestimmten Rangordnung und ist der Abstand zwischen den Rängen unbekannt sind sie ordinalskaliert, andernfalls intervallskaliert. Intervallskalierte Messdaten unterliegen dem metrischen Messniveau. Die Auswahl von statistischen Auswertungsverfahren ist ebenfalls abhängig vom Skalenniveau der Messdaten.

In den folgenden Abschnitten werden die ersten Schritte für eine erfolgreiche Durchführung eines Experimentes aufgezeigt. Im Anschluss daran werden Verfahren vorgestellt, die die gewonnenen Messdaten auf Eigenschaften untersuchen, die für die Auswahl hypothesenprüfender Analyseverfahren von Bedeutung sind.

3.2 Die Hypothesen

Bevor ein Experiment durchgeführt werden kann, sind die zu untersuchenden Merkmale festzulegen. Wenn es sich um eine rein explorative Untersuchung handelt, können diese Merkmale direkt operationalisiert werden. In einem Prüfexperiment hingegen müssen diese Merkmale sowie entsprechende Annahmen über ihren Einfluss zunächst in einer Forschungshypothese formuliert werden. Aus dieser Hypothese heraus werden die genannten Merkmale in UV und deren Kennwerte in AV unterteilt und geeignete Messverfahren ausgewählt.

3.2.1 Hypothesenbildung

In einer Forschungshypothese können angenommene Beziehungen wie Unterschiede und Zusammenhänge zwischen den Faktorstufen einer oder mehrerer UV bezüglich ihrer Messdaten (AV) formuliert werden. Des Weiteren können Vermutungen in Hinblick auf Veränderungen über einen bestimmten Zeitraum aufgestellt werden. Hieraus ergeben sich drei Klassifikationen für Forschungshypothesen:

- ***Zusammenhangshypothesen***
- ***Unterschiedshypothesen***
- ***Veränderunsghypothesen***

Forschungshypothesen können unterschiedlich präzise formuliert werden [Bortz und Döring, 2006, S.492-494]. Die Genauigkeit eines vermuteten Zusammenhangs, eines Unterschieds oder einer Veränderung wirkt sich auf die Wahl der Prüfparameter und -verfahren sowie auf die Eindeutigkeit der daraus resultierenden Untersuchungsergebnisse aus.

Gerichtete und ungerichtete Hypothesen: Wird eine Aussage über die Richtung eines Zusammenhangs, eines Unterschieds oder einer Veränderung gemacht, handelt es sich um eine gerichtete Hypothese, andernfalls wird von einer ungerichteten Hypothese gesprochen. Im Fall einer Zusammenhangshypothese wird die Stärke eines Zusammenhangs mit r angegeben.

- **Ungerichtet:** $AV_x \neq AV_y$ für Unterschieds- und Veränderungshypothesen und $r \neq 0$ für Zusammenhangshypothesen.

- **Gerichtet:** $AV_x > AV_y$ (rechts-gerichtet), $AV_x < AV_y$ (links-gerichtet) oder $AV_x = AV_y$ für Unterschieds- und Veränderungshypothesen und $r > 0$, $r < 0$ oder $r = 0$ für Zusammenhangshypothesen.

AV_x und AV_y stellen die abhängigen Variablen der Stichprobengruppen x und y dar.

Spezifische und unspezifische Hypothesen: Von einer spezifischen Hypothese wird dann gesprochen, wenn in einer gerichteten Hypothese die relevante Stärke S eines Unterschiedes, einer Veränderung oder eines Zusammenhangs definiert ist. Andernfalls handelt es sich um eine unspezifische Hypothese. Für die Annahme einer spezifischen Hypothese ist das Vorliegen eines signifikanten Ergebnisses allein nicht ausreichend. Darüber hinaus ist es erforderlich, dass die Relation von relevanter Stärke S und tatsächlicher Stärke mit der in der Hypothese spezifizierten Relation übereinstimmt. Ist dies nicht der Fall, kann trotz signifikantem Ergebnis die Hypothese nicht angenommen werden.

▶ **Unspezifisch:** $AV_x > AV_y$, $AV_x < AV_y$ für Unterschieds- und Veränderungshypothesen und $r > 0$, $r < 0$ für Zusammenhangshypothesen.

▶ **Spezifisch:** $AV_x \geq AV_y + S$, $AV_x \leq AV_y - S$ für Unterschieds- und Veränderungshypothesen und $r \geq S$, $r \leq S$ oder $r = S$ für Zusammenhangshypothesen.

Neben der Hypothese über zwei Merkmale kann auch eine Vermutung über den Kennwert (AV) eines Merkmals (UV) bezüglich eines konstanten Wertes, wie $AV_x \geq c$, angestellt werden.

3.2.2 Hypothesenprüfung

Um die Stärke eines Zusammenhangs zwischen zwei Kennwerten (AV) zu bestimmen, werden Korrelationskoeffizienten berechnet. Welcher Koeffizient berechnet wird, hängt dabei von den Verteilungen der miteinander zu vergleichenden abhängigen Variablen ab. Liegt eine Normalverteilung zweier Kennwerte AV_x und AV_y vor, wird der Produkt-Moment-Korrelationskoeffizient nach *Pearson* berechnet [Bulmer, 2003, S.191-196]:

$$r = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{(n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2) \cdot (n \cdot \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2)}} \tag{3.1}$$

Dabei entspricht n der Anzahl der in AV_x und AV_y enthaltenen Messwerte x_i und y_i.

Falls die vorliegende Verteilung unbekannt oder nicht normalverteilt ist, muss der Rangkorrelationskoeffizient nach Spearman [1907] berechnet werden. Eine gängige Berechnungsmethode ist in Gleichung 3.2 zu sehen, indem x_i und y_i durch Rangplätze ersetzt werden, die sie innerhalb ihrer Messreihen belegen. Anschließend wird mit der Differenz

d_i zwischen diesen konvertierten Messwerten $Rang(x_i)$ und $Rang(y_i)$ der Korrelationskoeffizient berechnet.

$$r \;=\; 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},\ mit\ d_i = Rang(x_i) - Rang(y_i) \tag{3.2}$$

Der Koeffizient r beider Gleichungen 3.1 und 3.2 liegt dabei im Intervall von $[-1; 1]$ mit $r \in \mathbb{R}$, wobei $|r|$ die Stärke des Zusammenhangs beschreibt. In Tabelle 3.1 wird die Aussage der Korrelationskoeffizienten bezüglich der Stärke des Zusammenhangs dargestellt. Viele Statistikprogramme wie *SPSS*, *PSPP* oder auch *Microsoft Excel* stellen

Korrelationskoeffizient	**Stärke des Zusammenhangs**
$\|r\| \leq 0,2$	sehr gering
$0,2 < \|r\| \leq 0,5$	gering
$0,5 < \|r\| \leq 0,7$	mittel stark
$0,7 < \|r\| \leq 0,9$	stark
$0,9 < \|r\| \leq 1,0$	sehr stark

Tabelle 3.1: Einstufung der Korrelationskoeffizienten nach ihrer Bedeutung.

Funktionen zur Berechnung dieser Koeffizienten bereit.

Neben der Fragestellung, ob zwei Merkmale zusammenhängen, ist häufig auch die Art der Beziehung zwischen ihnen von Interesse. Hierfür werden sogenannte Unterschiedshypothesen postuliert, in denen ein bestimmter Unterschied angenommen wird. Um einen signifikanten Unterschied feststellen zu können, sind geeignete Tests anzuwenden, deren Auswahl sowohl von der Verteilung als auch von dem Varianzverhalten der AV abhängt. Gleiches gilt für die Überprüfung auf Veränderungen über einen bestimmten Zeitraum.

Die Überprüfung der Hypothesen erfolgt mit Hilfe statistischer Signifikanztests, die auf die entsprechenden komplementären Hypothesen, die Nullhypothesen, angewendet werden. Im weiteren Verlauf dieser Arbeit wird deshalb zwischen den Forschungshypothesen und den Nullhypothesen, welche im Folgenden mit H_1 und H_0 bezeichnet werden, unterschieden. In der Regel stellt die H_0 die Behauptung auf, dass es keinen Zusammenhang, keinen Unterschied oder keine Veränderung zwischen den AV der Hypothesenmerkmale gibt.

Das Ergebnis eines Signifikanztests gibt diejenige Wahrscheinlichkeit an, mit der eine H_0 zutreffend ist. Hierfür ist eine Grenzwahrscheinlichkeit α festzulegen, welche im Folgenden als Signifikanz- oder α-Niveau bezeichnet wird. Die in der Praxis am häufigsten verwendeten Werte für α sind $0,05$ und $0,01$ beziehungsweise 5% und 1% [Bortz und Döring, 2006, S.494]. Ein Signifikanztest geht von der Gültigkeit der Nullhypothese aus. Es wird angenommen, dass sich $1 - \alpha$ der Kennwerte nicht unterscheiden, nicht

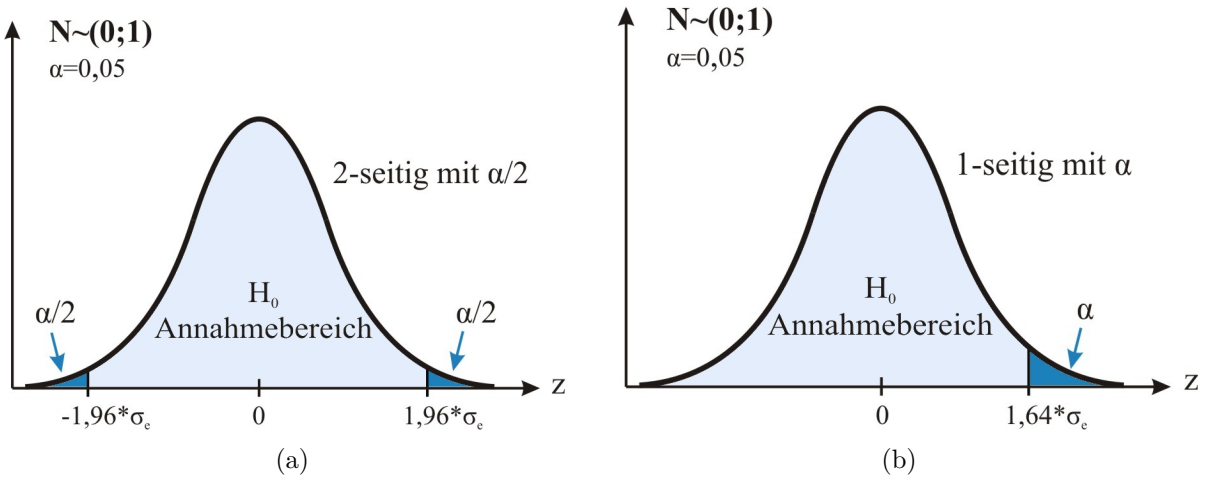


Abbildung 3.1: Hypothesenprüfung zwei- und einseitig (rechts-gerichtet) mit Signifikanzniveau $\alpha = 0,05$. Die kritischen z-Werte $z_{krit} \in \left\{z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right\} = \pm 1,96$ in (a) und $z_{krit} = z_{1-\alpha} = 1,64$ in (b) ergeben sich aus der Tabelle für die Standardnormalverteilung, siehe Anhang A.

zusammenhängen oder sich nicht verändern werden. Um dies zu überprüfen, stehen unterschiedliche statistische Verfahren zur Verfügung. Abbildung 3.1 zeigt am Beispiel einer Unterschiedshypothese anhand der Standardnormalverteilung mit der Standardabweichung $\sigma = 1$ den Gültigkeitsbereich der Nullhypothese um den Mittelwert $\mu = 0$. Mit dem Signifikanztest wird überprüft, ob sich die Kennwerte innerhalb des durch α begrenzten Annahmebereichs (hellblau) für H_0 befinden.

Handelt es sich bei der zu prüfenden H_1 um eine ungerichtete Hypothese, werden beide Seiten der Verteilung auf Kennwerte untersucht. In diesem Fall teilt sich das geforderte Signifikanzniveau gleichmäßig auf beide Seiten auf. Ist die Richtung bekannt, ist die einseitige Prüfung ausreichend.

Um zu bestimmen, ob sich die Kennwerte innerhalb des Annahmebereichs befinden, wird dem α-Niveau entsprechend ein kritischer z_{krit}-Wert bestimmt. Bei einem zweiseitigen Test kann der Annahmebereich somit durch ein von z_{krit} begrenztes Konfidenzintervall (KI) bestimmt werden.

$$KI \;=\; \mu \pm z_{krit} \cdot \sigma_e,\; mit\; \sigma_e \;=\; \sqrt{\frac{\sigma^2}{n}}\; und\; z_{krit} = z_{\frac{\alpha}{2}} \tag{3.3}$$

Bei einem einseitigen Test hingegen kann der Annahmebereich nur in eine Richtung begrenzt werden.

z_{krit} kann aus jeder beliebigen Tabelle der Standardnormalverteilung abgelesen werden. Bei einem gewählten α-Niveau von $0,05$ ergeben sich für den ungerichteten zweiseitigen Test $z_{1-\frac{\alpha}{2}} = +1,96, z_{\frac{\alpha}{2}} = -1,96$ und für den rechts gerichteten einseitigen Test $z_{1-\alpha} =$

$1,64$ als kritische z-Werte. Für eine links gerichtete Hypothese müsste $z_\alpha = -1,64$ bestimmt werden.

Mit

$$z_P = \frac{\mu_x - \mu_y}{\frac{\sigma_{x,y}}{\sqrt{n}}},\ mit\ n = n_x = n_y \tag{3.4}$$

lässt sich eine Prüfgröße z_P definieren, wobei μ_x und μ_y hier beispielhaft für die beiden Mittelwerte der Kennwerte AV_x und AV_y und $\sigma_{x,y}$ für deren durchschnittliche Standardabweichung stehen. Der Parameter n bezeichnet die Anzahl der Werte, aus denen sich die Kennwerte zusammensetzen. Anhand der Prüfgröße z_P ist anhand folgender Kriterien die Nullhypothese H_0 abzulehnen oder beizubehalten:

- Wenn $z_{\frac{\alpha}{2}} < z_P < z_{1-\frac{\alpha}{2}}$, dann behalte H_0 bei und lehne H_1 ab. *(2-seitig)*
- Wenn $z_P < z_{\frac{\alpha}{2}}$ oder $z_P > z_{1-\frac{\alpha}{2}}$, dann verwerfe H_0 und nimm H_1 an. *(2-seitig)*
- Wenn $z_P > z_\alpha$ *(links-gerichtet)* oder $z_P < z_{1-\alpha}$ *(rechts-gerichtet)*, dann behalte H_0 bei und lehne H_1 ab.
- Wenn $z_P < z_\alpha$ *(links-gerichtet)* oder $z_P > z_{1-\alpha}$ *(rechts-gerichtet)*, dann verwerfe H_0 und nimm H_1 an.

Weiterhin ist mit z_P eine zusätzliche Irrtumswahrscheinlichkeit p zu berechnen. Diese gibt diejenige Wahrscheinlichkeit an, mit der die Nullhypothese fälschlicherweise abgelehnt würde. Darüber hinaus gibt sie Auskunft darüber, wie signifikant sich die Ergebnisse der Kennwerte voneinander unterscheiden. Man spricht mit $p \leq 0,05$ von einem signifikanten, mit $p \leq 0,01$ von einem sehr signifikanten und bei $p \leq 0,001$ von einem hoch signifikanten Unterschied [Bortz und Döring, 2006, S.494]. Das Signifikanz-Niveau α kann demnach als maximal tolerierbare Irrtumswahrscheinlichkeit angesehen werden.

Ein bestandener Signifikanztest mit $p \leq \alpha$ sagt nichts über die relative Größe des signifikanten Unterschiedes aus. Diese Größe, welche zugleich das von der H_0 abweichende Maß beschreibt, wird als Effektgröße bezeichnet. Effektgrößen werden in Abhängigkeit vom Signifikanztest unterschiedlich berechnet und werden nach ihrer Größe klassifiziert. Die wichtigsten Effektgrößen nach Cohen [1992] sind in Tabelle 3.2 zu sehen. Eine nahe Null liegende Effektgröße deutet auf eine anzunehmende H_0 hin.

Weitere Prüfverteilungen, an denen Signifikanztests durchgeführt werden, sind beispielsweise die t-Verteilung, F-Verteilung und die χ^2-Verteilung. Die diesen Verteilungen entsprechenden kritischen t-, F- und χ^2-Werte sind analog dem z-Wert der Standardnormalverteilung aus Tabellen abzulesen (siehe Anhang A). Für diese Werte gilt dasselbe wie bei der Standardnormalverteilung. Entstammen die aus den Kennwerten berechneten Prüfgrößen normalverteilten Stichprobengruppen und überschreiten

Signifikanztest	**Effektgröße**	**Klassifikation**		
		Klein	Mittel	Groß
Test auf Mittelwertdifferenzen	$\delta = \frac{\mu_x - \mu_y}{\sigma}$	0,2	0,5	0,8
Korrelationstest	r	0,1	0,3	0,5

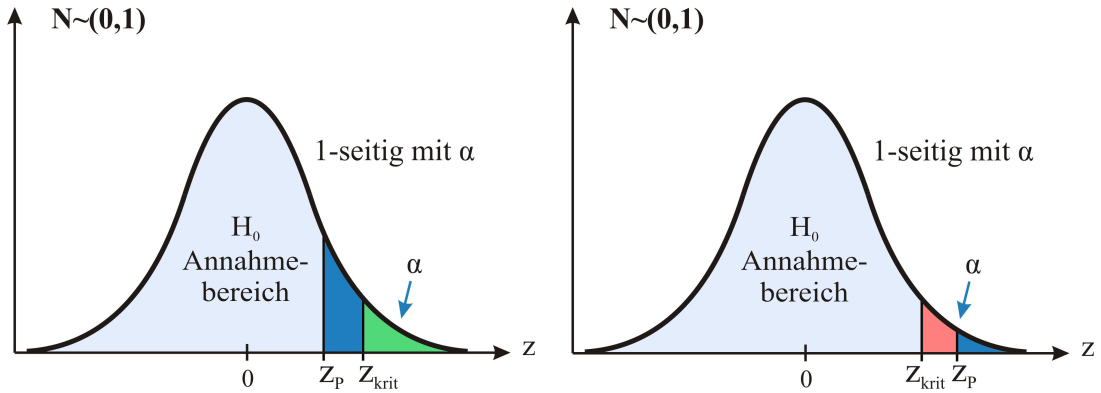
Tabelle 3.2: Effektgrößen für Mittelwertdifferenzen und Korrelationstests in Anlehnung an COHEN [1992].

sie die entsprechenden kritischen Werte, ist die Nullhypothese abzulehnen. Bei nicht-normalverteilten Stichproben-Messwerten hingegen werden verteilungsfreie Verfahren zur Hypothesenprüfung angewendet.

Insgesamt gibt es also zwei Anzeichen für ein signifikantes Ergebnis, so dass die H_0 abgelehnt und die H_1 angenommen werden kann:

- Wenn $p \leq \alpha$, mit $\alpha \in \{0,05, 0,01, 0,001\}$
- Wenn z_P außerhalb des durch z_{krit} begrenzten Annahmebereichs bzw. Konfidenzintervalls der H_0 liegt

Abbildung 3.2 zeigt die Lage zweier möglicher z-Prüfwerte und deren Auswirkung auf das Ergebnis einer ein-seitigen bzw. gerichteten Hypothesenprüfung.



(a) $z_P < z_{krit}$ (b) $z_P > z_{krit}$

Abbildung 3.2: In Bild (a) liegt z_P innerhalb des durch z_{krit} begrenzten Annahmebereiches von H_0, so dass die H_0 beibehalten werden muss. In Bild (b) hingegen liegt z_P außerhalb dieses Bereiches, so dass die H_0 verworfen und somit die H_1 angenommen werden kann.

3.2.3 Fehler 1. und 2. Art

Jede Entscheidung bezüglich der Annahme oder Ablehnung der H_0 oder H_1 geht mit einer bestimmten Fehlerwahrscheinlichkeit einher. Wird die H_0 aufgrund des Ergebnisses eines statistischen Tests abgelehnt, obwohl diese in Wirklichkeit gilt, spricht man von einem Fehler 1. Art. Dieser Fehler entspricht der im vorherigen Abschnitt betrachteten Irrtumswahrscheinlichkeit p und wird auch als α-Fehler bezeichnet. Lehnt man die H_1 ab, obwohl diese gilt, wird ein Fehler 2. Art begangen. Der Fehler 2. Art ist in

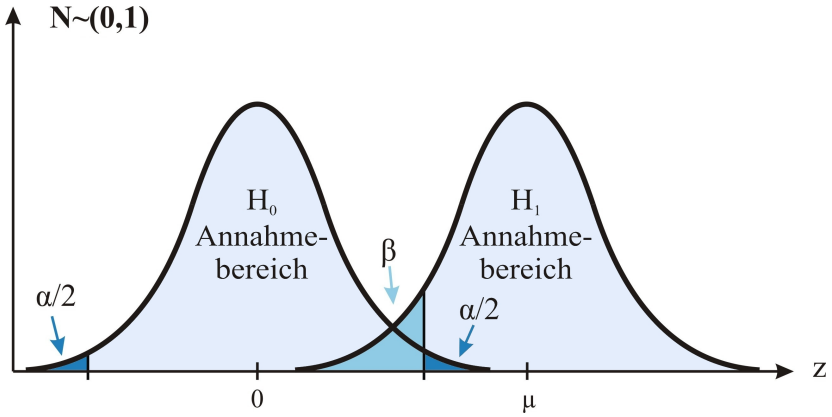


Abbildung 3.3: α -und β- Fehler.

der Literatur auch unter dem Namen β-Fehler zu finden und $1 - \beta$ wird als Teststärke beziehungsweise Power des angewandten Signifikanztests bezeichnet. Die Teststärke gibt somit diejenige Wahrscheinlichkeit an, mit der die Entscheidung für eine gültige Forschungshypothese H_1 richtig ist. Je höher die Teststärke, desto wahrscheinlicher ist es, dass sich aufgrund des Signifikanztests für die H_1 entschieden wird, wenn diese richtig ist. Diese Wahrscheinlichkeit nimmt zudem mit der Erhöhung des Signifikanz-Niveaus α und des Stichprobenumfangs zu. Nach Cohen [1988] ist jedoch ein α-Fehler gravierender als ein β-Fehler, weshalb eine Erhöhung von α nicht sinnvoll ist. Auch der Stichprobenumfang lässt sich in der Praxis nicht immer beliebig vergrößern. *Cohen* schlägt daher eine standardisierte Teststärke von $1 - \beta = 0,8$ vor, welche mittlerweile in der experimentellen Psychologie sowie in den Sozialwissenschaften gebräuchlich ist. Aus dieser Größe, dem α-Niveau und der Effektgröße δ lässt sich eine optimale Stichprobengröße n_{opt} berechnen. Mit ihr ist eine vorläufige Abschätzung des Arbeitsaufwandes bezüglich des durchzuführenden Experiments möglich.

3.3 Experimentelles Design

Die in den Abschnitten 3.1 und 3.2 dargelegten Zusammenhänge sind bei der Planung und Vorbereitung eines Experimentes zu berücksichtigen, damit eine zuverlässige Aus-

sage über die Beziehungen zwischen den UV und AV getroffen werden kann. Während Störvariablen mittels Eliminierung oder Randomisierung kontrolliert werden, wird dies für die UV und AV durch die Erstellung fester Versuchspläne erreicht. Ein solcher Versuchsplan, auch experimentelles Design genannt, dient der einheitlichen Vorgehensweise für jede einzelne Testreihe und ermöglicht die Reproduzierbarkeit des gesamten Experimentes. Darüber hinaus ermöglicht ein experimentelles Design die Auswahl geeigneter Analyseverfahren zur Hypothesenprüfung.

3.3.1 Operationalisierung

Die aus der Hypothese operationalisierten UV und AV bestimmen den Aufbau sowie die Durchführung eines Experimentes. Hieraus ergeben sich folgende Informationen für das experimentelle Design:

- Untersuchungsgegenstand bzw. -merkmale: UV
- Kennwerte: AV
- Methode: Messinstrument
- Stichprobe: Vpn

Ein zuverlässiges und abgesichertes Ergebnis kann nicht durch eine einmalige, sondern nur durch wiederholte Messungen gewonnen werden. Aus diesen Messungen ergibt sich die Zusammensetzung der AV. Abhängige Variablen repräsentieren somit den Durchschnitt aller Messungen bezüglich einer UV. Hierfür ist eine bestimmte Mindestgröße des Stichprobenumfangs erforderlich, auf die zum Ende dieses Abschnitts eingegangen wird.

Inwiefern die Anzahl der UV sowie die Anzahl ihrer Unterstufen den Aufbau des Untersuchungsplans beeinflussen, soll im Folgenden dargestellt werden.

3.3.2 Faktorielle Versuchspläne

Von einem ein-faktoriellen Versuchsplan wird dann gesprochen, wenn durch die Hypothese nur eine UV gegeben ist [Fisher, 1935]. Die UV entspricht somit dem Faktor und seinen k Variationen, den Faktorstufen UV_i, mit $i \in \{1, 2, .., k\}$. Daraus ergibt sich der in Tabelle 3.3 dargestellte Versuchplan nach Fisher [1935].

Das im Rahmen dieser Arbeit durchgeführte Experiment entspricht einem ein-faktoriellen Design mit den Faktorstufen der drei verschiedenen Hervorhebungstechniken und keiner Hervorhebung.

Vpn	**Faktorstufen der UV**					
	$\mathbf{UV_1}$	$\mathbf{UV_2}$	...	$\mathbf{UV_i}$	...	$\mathbf{UV_k}$
1	x_{11}	x_{12}	...	x_{1i}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2i}	...	x_{2k}
...	...	...	...	...	...	...
j	x_{j1}	x_{j2}	...	x_{ji}	...	x_{jk}
...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{ni}	...	x_{nk}
Mittelwert	μ_1	μ_2	...	μ_i	...	μ_k

Tabelle 3.3: Ein-faktorieller Versuchsplan mit k Faktorstufen und einem Stichprobenumfang von n. Die Gruppenmittelwerte sind durch μ_i und die Mittelwerte pro Vpn durch τ_j gegeben (Tabelle in Anlehnung an Fisher [1935]).

Im Fall voneinander unabhängiger Faktorstufen ist in jeder Zeile jeweils nur ein Messwert vorzufinden, da eine Vpn nur einer UV-Stufe ausgesetzt wird. Bei abhängigen Stufen hingegen ist je Proband entweder kein oder ein Messwert für jede Stufe vorzufinden.

		Faktorstufen der $\mathbf{UV_A}$			
		$\mathbf{UV_{A1}}$	$\mathbf{UV_{A2}}$	...	$\mathbf{UV_{Ak}}$
Faktorstufen der $\mathbf{UV_B}$	$\mathbf{UV_{B1}}$	x_{111}	x_{121}	...	x_{1k1}
		x_{112}	x_{122}	...	x_{1k2}
		...	...	...	...
		x_{11n}	x_{12n}	...	x_{1kn}
	$\mathbf{UV_{B2}}$	x_{211}	x_{221}	...	x_{2k1}
		x_{212}	x_{222}	...	x_{2k2}
		...	...	...	...
		x_{21n}	x_{22n}	...	x_{2kn}
	...	...	...	...	...
	$\mathbf{UV_{Bl}}$	x_{l11}	x_{l21}	...	x_{lk1}
		x_{l12}	x_{l22}	...	x_{lk2}
		...	...	...	...
		x_{l1n}	x_{l2n}	...	x_{lkn}

Tabelle 3.4: Zwei-faktorieller Versuchsplan mit $k \cdot l$ Faktorstufen und einem Stichprobenumfang von n (Tabelle in Anlehnung an [Fisher, 1935]).

Häufig ist auch das Zusammenspiel mehrerer Faktoren auf eine AV von Bedeutung. Fisher [1935] schlägt daher einen mehr-faktoriellen Versuchsplan vor, in dem jede Faktorstufe des einen Faktors mit allen Faktorstufen eines anderen Faktors kombiniert wird. Hieraus ergeben sich beispielsweise in einem zwei-faktoriellen Versuchsplan $k \cdot l$ Faktorstufenkombinationen, wobei jede dieser Kombinationen n Versuchspersonen

zugewiesen werden. Tabelle 3.4 zeigt einen solchen zwei-faktoriellen Plan mit $k \cdot l \cdot n$ Untersuchungsobjekten. Je größer die Dimensionen eines faktoriellen Versuchsplanes, desto größer und komplexer ist die zu untersuchende Datenmenge. Mehr-faktorielle Pläne eignen sich außerdem für die explorative Datenanalyse, auf deren Grundlage neue Hypothesen formuliert werden können.

Neben der Anzahl der UV ist auch die Anzahl der AV für den Aufbau eines Experimentes von Bedeutung. Wird eine UV durch nur eine AV beschrieben, handelt es sich um ein univariates Design. Wird sie durch zwei oder mehrere Kennwerte beschrieben, handelt es sich um bivariate bzw. multivariate Versuchspläne. In diesem Fall sind die Kennwerte bzw. abhängigen Variablen auf Korrelationen hin zu überprüfen. Abbildung 3.4 zeigt vier solcher möglichen experimentellen Versuchspläne. Die Beispiele für univariate Versuchspläne (a) und (c) entsprechen den in den Tabellen 3.3 und 3.4 dargestellten Versuchsplänen.

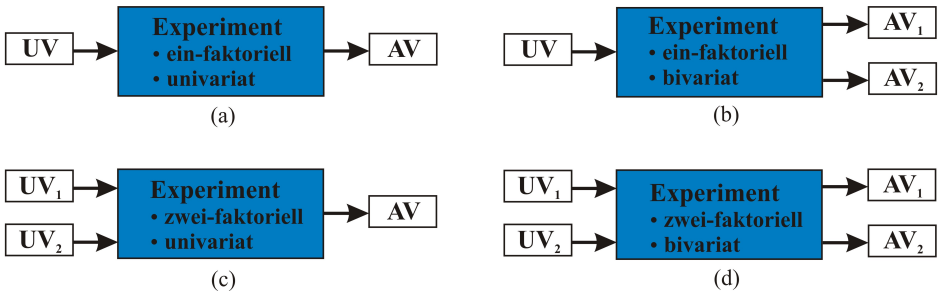


Abbildung 3.4: Mögliche aus der H_1 ableitbare experimentelle Designs: Ein- faktorielles uni- und bivariates Design (a und b), zwei-faktorielles uni- und bivariates Design (c und d).

Des Weiteren muss für die AV ein geeignetes Skalenniveau festgelegt werden, damit entsprechende Relationen zwischen den Variablen bestimmt werden können. Personenmerkmale wie Name, Geschlecht und Beruf werden zumeist nominal skaliert festgehalten. Dagegen sind Rangordnungen ordinal- und messbare Merkmale wie Gewicht, Temperaturen oder Reaktionszeiten intervall-skaliert.

3.3.3 Signalentdeckungstheorie

Wird die Reaktion von Versuchspersonen auf einen gegebenen Reiz im Rahmen einer wahrnehmungsspezifischen Studie untersucht, geschieht dies mit Hilfe psychophysischer Methoden. In sogenannten Entscheidungsexperimenten werden den Vpn verschiedene Stimuli präsentiert, auf die mit einer von zumeist zwei Antwortmöglichkeiten reagiert werden soll. Generell wird zwischen *yes/no*-, *rating*-, und *forced-choice* Experimenten unterschieden. In *yes/no*-Experimente wird einer Vpn entweder ein Rausch-Stimulus N (Noise) oder ein Ziel-Stimulus $N + S$ (Noise+Signal) präsentiert. Das Rauschen ist dabei immer präsent. Hebt sich das Signal deutlich vom Rauschen ab und wird dies

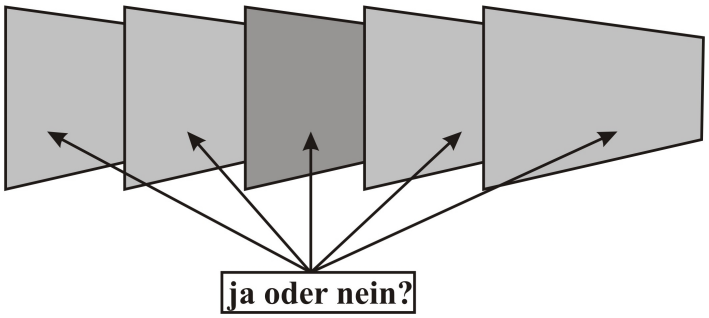


Abbildung 3.5: Beispielaufgabe eines psychophysischen Experimentes zur Bestimmung einer absoluten Wahrnehmungsschwelle. Empfindet die Vpn einen Reiz dunkler als einen vorgegebenen Standardreiz, hat diese mit *ja*, andernfalls mit *nein* zu antworten.

von der Vpn wahrgenommen, muss die Vpn entsprechend mit *ja* antworten. Nimmt sie das Signal nicht wahr, muss sie sich für *no* entscheiden. Bei *rating*- Experimenten hingegen erfolgt eine Beurteilung über verschiedene Rangordnungen. In *forced-choice*-Experimenten werden Rausch- und Signal- Stimuli immer zusammen präsentiert, wobei die Vpn das Signal S spezifizieren können muss [Stanislaw und Todorov, 1999]. Abbildung 3.5 zeigt ein einfaches Beispiel eines *yes/no*- Experimentes, wie es auch in dieser Arbeit durchgeführt wird.

Der Begründer der Psychophysik [Fechner, 1860], versuchte zunächst anhand verschiedener Methoden eine absolute Reiz-Schwelle zu bestimmen. Damit ist diejenige Reizintensität gemeint, mit der ein Reiz von der Vpn physiologisch gerade noch wahrgenommen oder von einem anderen Reiz unterschieden werden kann.

Die meisten psychophysischen Methoden gehen von einer normal verteilten Empfindungs- bzw. Entscheidungsstärke der Vpn aus. Diese Annahme beruht auf Thurstones Theorie *Law of Comparative Judgement* [Thurstone, 1927]. Es wird davon ausgegangen, dass bei wiederholter Darbietung eines Reizes, mit um den Mittelwert oszillierenden Werten zu rechnen ist. Mit Hilfe wiederholter Messungen bezüglich eines Reizes kann ein solcher Mittelwert gefunden werden. Abbildung 3.6 zeigt, je näher sich die Mittelwerte zweier Reize N und S_i sind, desto weniger lässt sich S_i von N unterscheiden, da deren normalverteilte Empfindungsstärken sich überlappen. Ist ein großer Intensitätsunterschied zwischen zwei Reizen gegeben und wird dieser auch deutlich wahrgenommen, liegen die Mittelwerte der jeweiligen Verteilungen weit auseinander und überschneiden sich nur wenig.

Aufbauend auf der Schwellen-Theorie von *Fechner* wurde von Green und Swets [1966] die Signalentdeckungstheorie entwickelt. Diese sieht von der absoluten Schwelle als einzige Einflussgröße für eine Entscheidung ab und geht vielmehr davon aus, dass die Urteilsfähigkeit der Vpn nicht allein von ihrer physiologisch bedingten Unterscheidungsfähigkeit, sondern auch von ihrer psychologisch bedingten Entscheidungsbereitschaft abhängt. Die Diskriminationsfähigkeit beschreibt dabei die Fähigkeit der Vpn,

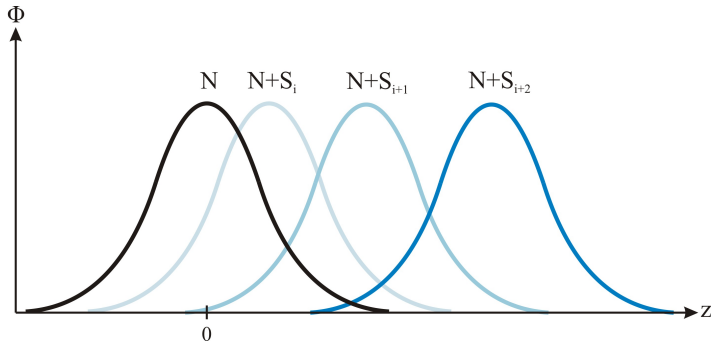


Abbildung 3.6: Normalverteilung der Rauschreize N und der Rausch-Plus-Signal-Reize N+S. Je stärker sich das Signal S vom Rauschen N abhebt, desto geringer ist die Wahrscheinlichkeit einer Verwechslung zwischen S und N.

einen Reiz entweder wahrnehmen (absolute Sensitivität) oder ihn von einem anderen Reiz unterscheiden (differentielle Sensitivität) zu können. Die Reaktionsneigung wird entweder durch die Motivation der Vpn beeinflusst oder durch unzureichende Instruktionen. Sie wird durch das Antwortkriterium L_x repräsentiert.

Grundsätzlich handelt es sich bei experimentellen Studien, basierend auf der Signalentdeckungstheorie, um Entscheidungsexperimente mit nur zwei Antwortmöglichkeiten. Abhängig von der gegebenen Bedingung ob ein Reiz präsentiert wurde oder nicht, können ja- Antworten in *Treffer* (Hit) und *Falscher Alarm* (False Alarm) und nein-Antworten in *Verpasser* (Miss) und *korrekte Zurückweisung* (Correct Rejection) unterschieden werden. In der Statistik wird unter einem *Treffer* diejenige bedingte Wahrscheinlichkeit verstanden, mit der ein vorhandener Zielreiz wahrgenommen wurde, und wird dort als *Sensitivität* bezeichnet. Hingegen entspricht die *korrekte Zurückweisung* dem statistischen Begriff der *Spezifität*.

Tabelle 3.5 stellt die in einem Entscheidungsexperiment möglichen Antworten in einem Reiz-Antwort-Schema nach Green und Swets [1966] gegenüber. Die Diskriminations-

		Antwort	
		ja „S vorhanden“	**nein „S nicht vorhanden“**
Stimuli	**N + S**	Treffer $P(ja\|(N+S))$	Verpasser $P(nein\|(N+S))$
	N	Falscher Alarm $P(ja\|N)$	Korrekte Zurückweisung $P(nein\|N)$

Tabelle 3.5: Antwortklassifikationen nach Green und Swets [1966].

leistung d' sowie das Antwortkriterium L_x können somit anhand der z-transformierten Treffer- und Falsch-Alarm-Werte berechnet werden [Bortz und Döring, 2006]. Die

z-transformierten Werte können der Tabelle der Standardnormalverteilung in Anhang A entnommen werden.

$$d' = z_{Treffer} - z_{Falscher\ Alarm} \tag{3.5}$$

Das Antwortkriterium L_x, siehe Gleichung 3.6, wird anhand der Wahrscheinlichkeitsdichten berechnet. Im Idealfall beträgt $L_x = 1$, was eine neutrale Reaktionsneigung bedeutet. Je mehr sich L_x jedoch von 1 unterscheidet, desto größer die Verzerrung der Ergebnisse, beeinflusst entweder durch eine liberale Reaktionsneigung mit $L_x < 1$ oder einer konservativen Reaktionsneigung mit $L_x > 1$.

$$L_x = \frac{\phi_{z_{Treffer}}}{\phi_{z_{Falscher\ Alarm}}},\ mit\ \phi_{z_i} = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot z_i^2}\ und\ i \in \{Treffer,\ Falscher\ Alarm\} \tag{3.6}$$

Abbildung 3.7 zeigt am Beispiel zweier Stimuli (Rauschen und Signal), inwiefern die Werte der Treffer und Falschen Alarme Aufschluss über die Reaktionsneigung einer Vpn geben. Die Messungen sind dabei für jeden Probanden einzeln durchzuführen.

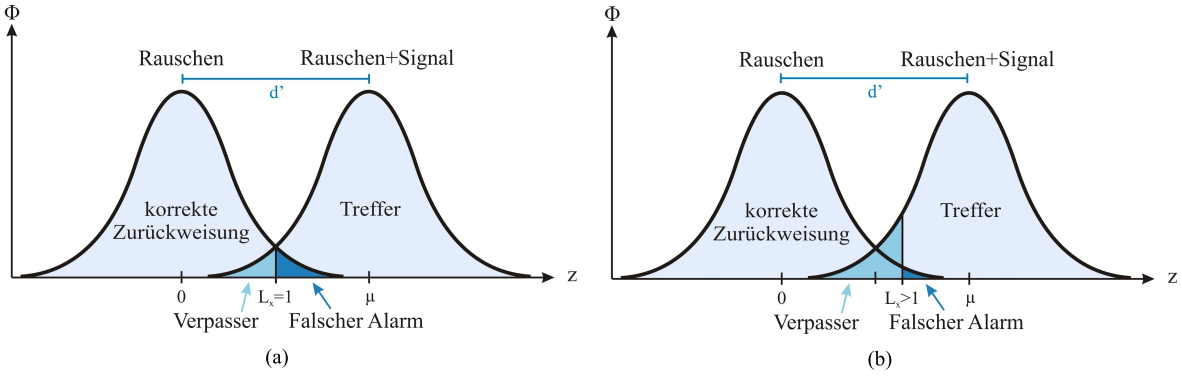


Abbildung 3.7: Antwortkriterium am Beispiel normalverteilter Entscheidungen bezüglich zweier Reize. In Bild (a) gibt es keine Reaktionsverzerrung, da $L_x = 1$ und die Vpn somit nur aufgrund seiner bzw. ihrer physiologisch bedingten Sensitivität d' entscheidet. Mit $L_x > 1$ in Bild (b) hat die Vpn eine konservative Reaktionsneigung und tendiert somit häufiger zum *nein*-Sagen um Falsch-Alarm-Aussagen zu vermeiden. Gleichzeitig riskiert die Vpn eine geringere Trefferquote.

Stanislaw und Todorov geben in [Stanislaw und Todorov, 1999] einen guten Überblick über die Anwendung der Signalentdeckungstheorie und der mit ihr verbundenen Rechnungen.

Für die Durchführung sogenannter *ja-/nein-* Experimente werden häufig Reaktionszeittests angewendet, in denen nicht nur die Entscheidungen der Vpn, sondern auch die für diese Entscheidungen benötigten Zeiten aufgezeichnet werden.

3.3.4 Bestimmung des optimalen Stichprobenumfangs

Zu Beginn dieses Abschnitts wurde kurz auf die Notwendigkeit von Messwiederholungen eingegangen, damit zuverlässige und repräsentative Ergebnisse gewonnen werden können. Hierbei ist unter Messwiederholung die Durchführung des Experimentes an mehreren Vpn gemeint. Bei der Auswahl der Stichprobengröße stellen sich folgende Probleme:

- Ist der Stichprobenumfang zu groß gewählt, ist fast immer ein signifikantes Ergebnis zu erwarten.
- Je kleiner der Stichprobenumfang, desto geringer ist die Wahrscheinlichkeit ein signifikantes Ergebnis zu erhalten.

Wird das Experiment also an zu vielen Vpn durchgeführt, wird unabhängig davon wie klein der tatsächliche Unterschied zwischen zwei Mittelwerten ist, immer ein signifikantes Ergebnis festgestellt. Es stellt sich dann jedoch die Frage, ob dieser signifikante Unterschied für die jeweilige Studie von Relevanz ist. Von daher ist es vorteilhaft Hypothesen spezifisch zu formulieren, so dass die darin definierte relevante Stärke δ im Rahmen der Hypothesenüberprüfung zusätzlich berücksichtigt wird.

Allerdings lässt sich diese Stärke bereits an einer kleineren Stichprobengröße feststellen. Eine solche Mindestgröße n_{opt} einer Stichprobe, ab der ein gewünschter relevanter Unterschied eintritt, kann mit Hilfe der Effektstärke δ, dem Signifikanzniveau α und der Teststärke $1 - \beta$ berechnet werden. Die Berechnung einer solchen optimalen Stichprobengröße ist unter anderem mit der in 3.7 gegebenen Gleichung nach [Kähler, 2002] möglich.

$$n_{opt} \geq (\frac{z_\alpha - z_{1-\beta}}{\delta})^2 \tag{3.7}$$

Cohen führt in seinem Buch [Cohen, 1988] Tabellen auf, in denen abhängig vom durchzuführenden Signifikanztest optimale Größen für eine benötigte Stichprobe abgelesen werden können.

Da die tatsächliche Effektgröße häufig während der Versuchsplanung unbekannt ist, kann diese geschätzt werden. Eine solche Schätzung hängt dabei von der jeweiligen Studie und dem von ihr geforderten Mindestunterschied ab. Sind in einer Studie beispielsweise bereits kleine Unterschiede zwischen zwei zu vergleichenden Merkmalen von Interesse, kann nach *Cohen* eine Effektgröße von $\delta = 0.2$ angenommen werden. Sind hingegen nur eindeutige und große Unterschiede von Bedeutung, wird $\delta = 0.8$ angenommen.

Die häufigsten Stichprobengrößen zu den gängigsten Effektgrößen und Niveaus, sind in Tabelle 3.6 zu sehen. Hieraus wird ersichtlich, dass die Stichprobengröße sowohl mit der Erhöhung des Signifikanzniveaus α als auch mit kleiner werdender Effektgröße zunimmt. Ein maximal tolerierbarer β-Fehler bzw. Fehler 2. Art von 20%, wie in Tabelle

α-Niveau	Signifikanzgrenzen	$\delta = \mathbf{0,2}$	$\delta = \mathbf{0,5}$	$\delta = \mathbf{0,8}$
0,05	1-seitig	155	25	10
	2-seitig	196	31	12
0,01	1-seitig	251	40	16
	2-seitig	292	47	18

Tabelle 3.6: Optimale Stichprobenumfänge für Signifikanztest mit einer Teststärke $1 - \beta$ von 0.8.

3.6 aufgeführt, gewährleistet bei optimalen Stichprobenumfang, dass ein Signifikanztest mit einer Wahrscheinlichkeit von 80% zu einem signifikanten Ergebnis kommt. Auch wenn eine solche Teststärke gering erscheint, hat sie sich in der Praxis als zuverlässig bewährt. Wird eine höhere Teststärke angestrebt, muss mit einem entsprechend größerem Stichprobenumfang gearbeitet werden.

Die bisherige Berechnung geht von einem Vergleich zwischen gleichgroßen Stichprobengruppen aus. Die Verteilung der ermittelten Stichprobengröße hängt dabei vom Versuchsdesign ab. Während in einem within-subject Experiment n_{opt} als die tatsächlich benötigte Stichprobengröße angesehen werden kann, ist bei einem *between-subject*-Design n_{opt} mit der Anzahl der zu vergleichenden Gruppen zu multiplizieren. Dabei sollte $k \cdot n_{opt}$ auf die k Stichprobengruppen gleichverteilt werden. In der Praxis ist dies jedoch nicht immer möglich. Stehen in einem Zwei-Gruppenvergleich für eine Gruppe A beispielsweise nur n_A Vpn zur Verfügung, kann mit der Gleichung 3.8 die benötigte Anzahl n_B an Vpn für die Gruppe B berechnet werden.

$$n_B = \frac{n_A \cdot n_{opt}}{2 \cdot n_A - n_{opt}} \tag{3.8}$$

Für die Bestimmung geeigneter Effektgrößen sowie opimaler Stichprobenumfänge für unabhängige Gruppenvergleiche und ungleich große Stichproben empfiehlt sich weiterführende Literatur wie [Cohen, 1988],[Cohen, 1992] und [Bortz und Döring, 2006]. Hier werden Sonderfälle aufgezeigt sowie weitere Definitionen von Effektgrößen für beispielsweise Korrelationen oder multivariate Varianzanalysen gegeben.

3.4 Analyseverfahren

Der Wert einer empirischen Untersuchung wird vor allem an ihrer statistischen Validität gemessen. Deskriptive Analysen bezüglich der gemessenen Häufigkeiten, Mittelwerte und Standardabweichungen sowie deren grafische und tabellarische Darstellungen geben nur unzureichend Auskunft über die Beziehungen zwischen den Messwerten. Aus diesem Grund sind sie für die Überprüfung von Hypothesen nicht geeignet und dienen

lediglich der Veranschaulichung und Zusammenfassung der Ergebnisse. Um der H_1 entsprechende signifikante Ergebnisse ermitteln zu können, sind inferentielle statistische Verfahren, sogenannte Hypothesen- bzw. Signifikanztests, anzuwenden.

Die Auswahl solcher Verfahren ist an bestimmte Voraussetzungen gebunden, deren Gültigkeit zunächst anhand der Messdaten überprüft werden muss. Wird ein Verfahren zur Interpretation der Untersuchungsergebnisse gewählt, welches für diese Art von Untersuchung nicht vorgesehen oder ungeeignet ist, da die gegebenen Voraussetzungen verletzt sind, sind sämtliche darauf aufbauende Schlüsse bezüglich der Hypothesen ungültig.

Im Folgenden werden Prüfverfahren hinsichtlich der von den Signifikanztests geforderten Bedingungen vorgestellt. Jedes dieser Prüfverfahren selbst ist ein Signifikanztest bezüglich der zu prüfenden Größe bzw. Eigenschaft, dessen Ergebnis entweder für oder gegen die untersuchte Bedingung spricht. Aufbauend auf diesen Eigenschaften sind entsprechende Signifikanztests für die eigentliche Forschungshypothese zu wählen und im Falle eines signifikanten Ergebnisses deskriptiv darzustellen.

Die im Folgenden vorgestellten Prüfverfahren setzen unterschiedliche Prüfverteilungen wie die Standardnormalverteilung nach *Gauß*, die t-Verteilung nach Student [Gosset, 1908], die F-Verteilung nach Fisher [1924] sowie die χ^2-Verteilung nach Pearson [1900] voraus. Die berechneten Prüfgrößen sind mit den kritischen Werten der jeweiligen Verteilung zu vergleichen. Mit Hilfe der gegebenen Freiheitsgrade *df* können diese den entsprechenden Tabellen der Prüfverteilungen entnommen werden. Anhang A beinhaltet die für diese Arbeit relevanten Tabellen.

3.4.1 Überprüfung der Daten

Für die Auswahl eines geeigneten Signifikanztests sind die erhobenen Daten auf folgende Eigenschaften zu überprüfen:

- Ausreißer
- Anzahl der Faktoren und deren Gruppen bzw. Faktorstufen
- Verteilung
- Varianz
- Abhängigkeit

Vor Anwendung der Prüfverfahren sind die vorliegenden Messdaten auf Ausreißer zu überprüfen, die das Ergebnis verzerren können. Eine Möglichkeit hierfür wurde bereits in Abschnitt 3.3.3 angesprochen. Mit Hilfe der Berechnungen der Signalentdeckungstheorie können Vpn mit einem zu hohen oder niedrigen Antwortkriterium L_x entdeckt

und von der Analyse ausgeschlossen werden. Darüber hinaus kann mit Hilfe des Mittelwertkriteriums festgestellt werden, inwiefern sich die Messdaten einer Vpn von den gemittelten Messwerten über alle Vpn unterscheiden. Messwerte außerhalb des Intervalls von $\mu_i \pm m \cdot \sigma_i$ mit $m \in [2,0; \infty)$ sind ergebnisverzerrende Werte bzw. Ausreißer und sollten nicht weiter betrachtet werden.

Die Anzahl der Faktorstufen einer UV legt fest, ob ein Signifikanztest für zwei (Abschnitt 3.4.2) oder für mehr als zwei Faktorstufen bzw. Gruppen (Abschnitt 3.4.3) durchgeführt werden muss. Anschließend werden in Abhängigkeit von der Verteilung der Messdaten einer Gruppe entweder parametrische oder nicht-parametrische Verfahren zur Hypothesenprüfung angewandt. Parametrische Tests setzen dabei immer eine Normalverteilung der Daten voraus. Nicht-parametrische bzw. verteilungsfreie Methoden hingegen gehen von einer unbestimmten Verteilung aus und benutzen daher häufig verschiedene Rangordnungsverfahren. Die in Frage kommenden Signifikanztests werden in diesem Fall auf den Rängen der Messdaten durchgeführt.

Für eine erste Eingrenzung der möglichen hypothesenprüfenden Verfahren, müssen die Daten zunächst auf Normalverteilung überprüft werden. Hierfür gibt es wiederum viele verschiedene Verfahren. Die am häufigsten verwendeten sind

- Chi-Quadrat-Test für mehr als 100 Messdaten,
- Kolmogorov-Smirnov-Test für mehr als 50 Messdaten,
- Shapiro-Wilk-Test für weniger als 50 Messdaten.

Können parametrische Verfahren aufgrund gegebener Normalverteilung angewandt werden, muss weiterhin auf Varianzhomogenität der Daten geprüft werden. Dies kann im Fall mehrerer Gruppen mit Hilfe des *Levene*-Tests überprüft werden.

Sind lediglich zwei Gruppenmittelwerte miteinander zu vergleichen, kann ein einfacher F-Test angewandt werden. Beim F-Test berechnet sich die Prüfgröße aus dem Quotienten der beiden Gruppenvarianzen.

$$F \;=\; \frac{\sigma^2_{max}}{\sigma^2_{min}},\ mit\ df_1 = n_{min} - 1\ und\ df_2 = n_{max} - 1 \tag{3.9}$$

Die Nullhypothese für den F-Test H_0(Varianzhomogenität) geht von gleichen Varianzen mit $\sigma_1^2 = \sigma_2^2$ aus und entspricht im Falle eines nicht-signifikanten Ergebnisses der Varianzhomogenität. Mit $F < F_{\alpha,df_1,df_2}$ ergibt sich somit die Varianzhomogenität mit $p > \alpha$ und andernfalls die Varianzheterogenität.

Sind mehrere Gruppen auf Varianzhomogenität zu überprüfen, findet der *Levene*-Test Anwendung. Der sich hieraus ergebende Prüfwert F_{Levene} ist mit den Freiheitsgraden $df_1 = k - 1$ und $df_2 = n - k$ F-verteilt und ist dementsprechend anhand des kritischen F-Werts F_{α,df_1,df_2} auf Signifikanz zu prüfen.

Die Abhängigkeit der Faktorgruppen wird dadurch bestimmt, ob die Vpn einer oder allen Faktorstufen ausgesetzt sind. Im letzteren Fall sind die Faktorgruppen voneinander abhängig.

Sind die Beziehungs-, Verteilungs- und Varianzeigenschaften bekannt, können entsprechende Hypothesentests ausgewählt und durchgeführt werden.

Im nächsten Abschnitt werden mögliche Signifikanztests und deren Auswahlkriterien zur Überprüfung der Forschungshypothese vorgestellt. Zunächst werden Tests vorgestellt, die nur für die Überprüfung von zwei Gruppen geeignet sind. Anschließend wird auf Verfahren für mehr als zwei Gruppen eingegangen. Der Einfachheit halber sind die hier aufgezeigten Gleichungen zur Berechnung der Prüfgrößen mit dem Versuchsplan 3.3 in Abschnitt 3.3.2 abgestimmt.

3.4.2 Signifikanztests für Mittelwertdifferenzen zwischen zwei Gruppen

Im einfachsten Fall eines Signifikanztests sind die jeweiligen Mittelwerte zweier Gruppen miteinander zu vergleichen. Die hier zu überprüfende Nullhypothese $H_0(Test)$, welche von gleichen Mittelwerten beider Gruppen ausgeht, soll anhand eines geeigneten Signifikanztests verworfen oder angenommen werden. In Abhängigkeit von der Forschungshypothese kann hier entweder einseitig bei gerichteten oder zweiseitig bei ungerichteten Hypothesen geprüft werden. Die im Folgenden aufgezeigten Beispiele werden hier jedoch nur einseitig dargestellt. Sind die zu vergleichenden Gruppen jeweils normalverteilt, kann ein einfacher parametrischer Paarungstest, wie der *t-Test*, durchgeführt werden [Gosset, 1908]. Bei nicht-normalverteilten Gruppen finden nichtparametrische Paarungstests, wie der *Wilcoxon* [Wilcoxon, 1945]- oder *Mann-Whitney-U*-Test [Mann und Whitney, 1947], Anwendung. Abbildung 3.8 zeigt eine mögliche Auswahl von Signifikanztests in Abhängigkeit der Verteilung und Gruppenbeziehungen.

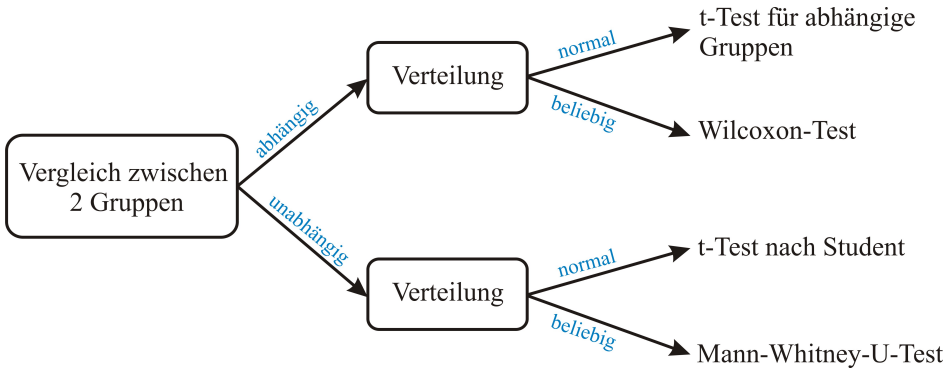


Abbildung 3.8: Mittelwert-Vergleich zwischen zwei Stichproben-Gruppen.

Parametrische Verfahren

Sind die beiden zu vergleichenden Gruppen normalverteilt und unabhängig voneinander wird der *t-Test nach Student* angewendet. Die Prüfverteilung ist in diesem Fall t-verteilt und die Prüfgröße t ist bei Varianzhomogenität mit der Gleichung 3.10 und bei Varianzheterogenität mit der Gleichung 3.11 zu berechnen. Ein signifikanter Unterschied mit $p < 0,05$ ist dann gegeben, wenn $t > t_{krit}$, mit $t_{krit} = t_{\alpha,df}$, zutrifft.

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{(n_1-1)\cdot\sigma_1^2+(n_2-1)\cdot\sigma_2^2}{n_1+n_2-2}}} \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1+n_2}},\ mit\ df = n_1 + n_2 - 2 \tag{3.10}$$

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},\ mit\ df = \frac{n_1 + n_2 - 2}{2} \tag{3.11}$$

Der *t-Test für abhängige Stichproben* unterscheidet sich lediglich bei der Berechnung des Prüfwerts t und dessen Freiheitsgrades df. Der Prüfwert ergibt sich aus der Differenz $\bar{d}$ beider Mittelwerte mit $\bar{d} = \mu_1 - \mu_2$ und deren Standardabweichung σ, entsprechend Gleichung 3.12 und 3.13.

$$t = \frac{|\bar{d}| \cdot \sqrt{n}}{\sigma},\ mit\ df = n - 1 \tag{3.12}$$

und

$$\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (d_i - \frac{|\bar{d}|}{n})^2} \tag{3.13}$$

, wobei $d_i = x_{1i} - x_{2i}$ die Differenz zweier Messwerte der Stichprobengruppen UV_1 und UV_2 ist.

Nicht-parametrische Verfahren

Sind die zu vergleichenden Gruppen nicht normalverteilt, finden nicht-parametrische Verfahren Anwendung. Dabei werden die Messdaten häufig durch Rangplätze ersetzt. Die Berechnung der entsprechenden Prüfgröße erfolgt dann aus den jeweiligen Rängen der Messdaten.

Bei unabhängigen Stichproben eignet sich der *Mann-Whitney-U*-Test. Dieser Test von Mann und Whitney [1947] fasst beide Stichproben in eine gemeinsame Rangreihe zusammen, in der der kleinste Messwert aus beiden Stichproben, sofern er nur einmal vorkommt, den Rangplatz 1 bekommt. Kommt ein Messwert mehrfach vor, werden die in Frage kommenden Rangplätze gemittelt und diesen Messwerten zugewiesen. Sind

alle Rängplätze vergeben, werden beide Stichproben wieder getrennt betrachtet. Die Prüfgrößen U_1 und U_2 der Stichproben berechnen sich dann aus der Summe ihrer Rangplätze R_1 und R_2 sowie aus ihren Stichprobengrößen n_1 und n_2.

$$U_1 = R_1 - \frac{n_1 \cdot (n_1 + 1)}{2} \tag{3.14}$$

$$U_2 = R_2 - \frac{n_2 \cdot (n_2 + 1)}{2} \tag{3.15}$$

Die kleinere der beiden Prüfgrößen $U_{min} = min(U_1, U_2)$ wird dann mit dem kritischen Wert U_{krit} verglichen. Im Gegensatz zu anderen Signifikanztests ist ein signifikantes Ergebnis dann gegeben, wenn $U_{min} < U_{krit}$ vorliegt.

Ein weiterer parameterfreier Test ist der *Wilcoxon*-Test von Wilcoxon [1945]. Dieser wird vor allem auf abhängige Stichprobenpaare angewandt, indem Rangplätze bezüglich der Differenzen der abhängigen Messwerte vergeben werden. Die Ränge werden nach positiven und negativen Differenzen gruppiert und anschließend aufsummiert. Die sich daraus ergebene kleinste Rangsumme stellt die Prüfgröße W dar und lie-

UV_1	UV_2	**d**	**|d|**	**Rang**	**Rang mit d > 0**	**Rang mit d < 0**
20	16	4	4	8,5	8,5	-
17	12	5	5	10	10	-
17	19	-2	2	3	-	3
15	14	1	1	1	1	-
18	15	3	3	6	6	-
19	21	-2	2	3	-	3
20	17	3	3	6	6	-
16	16	0	0	-	-	-
21	17	4	4	8,5	8,5	-
21	18	3	3	6	6	-
20	18	2	2	3	3	-
$n_r = 11 - 1$					$W_+ = \sum = 49$	$W_- = \sum = 6$

Tabelle 3.7: *Wilcoxon*- Test am Beispiel zweier Faktorstufen: Die Anzahl der relevanten Paare n_r mit $d \neq 0$ sowie das Testniveau α bestimmen den kritischen Wert W_{krit}. Die kleinere Rangsumme $W_- = 6$ ist kleiner als der kritische Wert W_{krit}, da $W_{krit} = W_{\alpha,n_r} = W_{0,05,10} = 8$ (siehe Anhang A). Da die Differenz 2 insgesamt dreimal vorkommt, werden die möglichen Ränge von $2, 3$ und 4 gemittelt. Somit wird jeder Vpn dessen Messwerte eine Differenz von 2 aufweisen, der Rang 3 vergeben.

fert mit $W < W_{krit}$ ein signifikantes Ergebnis. In diesem Fall kann die Nullhypothese $H_0(Wilcoxon) := \mu_1 = \mu_2$ verworfen werden. Liegt keine Tabelle für W_{krit} vor, kann

die Prüfgröße W in die Prüfgröße z der Standardnormalverteilung mit der Gleichung 3.16 umgewandelt werden mit $n_r = n - Anzahl\ der\ Differenzpaare$ mit $d \neq 0$.

$$z = \frac{\frac{n_r \cdot (n_r+1)}{4} - W}{\sqrt{\frac{n_r \cdot (n_r+1) \cdot (2 \cdot n_r+1)}{24}}} \tag{3.16}$$

Für das Beispiel aus Tabelle 3.7 ergibt sich somit die normalverteilte Prüfgröße mit

$$z = \frac{\frac{10 \cdot 11}{4} - 6}{\sqrt{\frac{10 \cdot 11 \cdot 21}{24}}} = 2,19 \tag{3.17}$$

, welche den kritischen z-Wert $z_{1-0,05} = 1,64$ für den einseitigen Test übersteigt und somit einen signifikanten Unterschied zwischen den Gruppen feststellt.

Kommen geteilte Rangplätze mehrfach vor, ist sowohl die Anzahl der geteilten Ränge m als auch die Anzahl der Teiler t_i mit zu berücksichtigen. Gleichung 3.16 wird dementsprechend wie folgt modifiziert:

$$z = \frac{\frac{n_r \cdot (n_r+1)}{4} - W}{\sqrt{\frac{n_r \cdot (n_r+1) \cdot (2 \cdot n_r+1)}{24} - \frac{\sum_{i=1}^{m} (t_i^3 - t_i)}{48}}} \tag{3.18}$$

Der Tabelle 3.7 entsprechend ergibt sich dann

$$z = \frac{\frac{10 \cdot 11}{4} - 6}{\sqrt{\frac{10 \cdot 11 \cdot 21}{24} - \frac{\sum_{i=1}^{3} (t_i^3 - t_i)}{48}}} = 2,28 \tag{3.19}$$

, wobei der Rangplatz 3 und 6 jeweils dreimal und der Rangplatz $8,5$ zweimal auftraten.

Ein signifikantes Ergebnis führt nur dann zur Annahme der Forschungshypothese, wenn es sich um eine unspezifische Hypothese handelt. Bei spezifischen Hypothesen muss darüber hinaus auch der erforderliche Mindestunterschied δ_{min} zwischen den Gruppen eingehalten sein. Ist der tatsächliche Gruppenunterschied δ größer gleich δ_{min}, kann die Forschungshypothese angenommen werden.

3.4.3 Signifikanztests für Mittelwertdifferenzen zwischen mehr als zwei Gruppen

Für den Vergleich der Mittelwerte von mehr als zwei Gruppen wurden vielfältige Signifikanztests entwickelt. Die verbreitetsten und anerkanntesten sind die *ANOVA* nach Fisher [1925], der *Friedman*-Test von Friedman [1937] oder der *H*-Test von Kruskal und Wallis [1952]. In Abbildung 3.9 ist dargestellt, unter welchen Voraussetzungen welcher dieser Tests anzuwenden sind. Sowohl die *ANOVA* als auch der *Friedman*-Test und der *H*-Test gehen von der Annahme aus, dass alle Gruppenmittelwerte mit $\mu_1 = \mu_2 = ... = \mu_i = ... = \mu_k$ gleich sind. Diese Annahme entspricht dabei den Nullhypothesen der Tests und wird dann abgelehnt, sobald sich nur ein μ_i von den anderen Mittelwerten unterscheidet. Die Richtung (negativ oder positiv) der Unterschiede ist in dieser Untersuchung nicht von Bedeutung, weshalb zweiseitig getestet wird. Anschließend werden Paarungstests, wie in Abschnitt 3.4.2 aufgeführt, zwischen allen Gruppen durchgeführt, damit der entsprechende Unterschied lokalisiert und dessen Stärke bestimmt werden kann. Dabei ist zu beachten, dass mit jedem Paarungstest ein α-Fehler zustande kommt, der sich bei Durchführung mehrerer Paarungstests aufsummiert. Diese sogenannte α-Kumulierung kann mit Hilfe einer α-Adjustierung nach *Bonferroni* minimiert werden, indem jeder Paarungstest mit dem Testniveau $\alpha_{adj} = \frac{\alpha}{s}$ durchgeführt wird. Wobei $s = k \cdot \frac{(k-1)}{2}$ der Anzahl der benötigten Paarvergleiche und k der Anzahl der zu vergleichenden Faktorstufen entspricht.

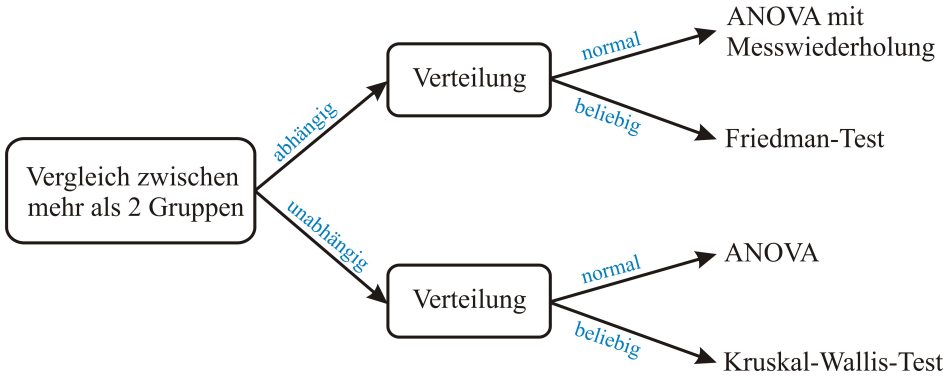


Abbildung 3.9: Mittelwerts-Vergleich zwischen mehr als zwei Stichproben-Gruppen.

Im Folgenden werden nur Verfahren erläutert, die bei ein-faktoriellen Versuchsplänen Anwendung finden. Häufig lassen sich diese auf mehr-faktorielle Versuchspläne erweitern. Für die Darstellung des experimentellen Designs im Rahmen dieser Arbeit ist dies jedoch nicht relevant. Eine ausführliche Beschreibung mehr-faktorieller Versuchspläne ist in [Bortz, 2005] und [Bortz und Döring, 2006] zu finden.

Wie bei Signifikanztests für Mittelwertdifferenzen zwischen zwei Stichprobengruppen ist auch bei mehr als zwei Gruppen zwischen parametrischen Verfahren für normal-

verteilte Stichproben und nicht-parametrischen Verfahren bei nicht-normalverteilten Stichproben zu unterscheiden.

Parametrische Verfahren

Bei gegebener Normalverteilung aller Faktorstufen bzw. UV-Variationen wird auf den unabhängigen Faktorstufen die normale *ANOVA* (Analysis of Variance) und auf abhängige Faktorstufen eine *ANOVA mit Messwiederholung* durchgeführt [Fisher, 1925]. Beide Verfahren unterscheiden sich in der Aufteilung ihrer Varianzen. Die Gesamtvariabilität SAQ_{Ges} aller Faktorstufen berechnet sich für beide Verfahren gleich und ergibt sich aus der Summe der quadrierten Abweichungen vom Gesamtmittelwert $\bar{\mu}$.

$$\bar{\mu} \;=\; \frac{1}{k \cdot n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} \;=\; \sum_{i=1}^{k} \mu_i \tag{3.20}$$

Die für die *ANOVA* benötigte Prüfgröße F unterliegt der F-Verteilung und berechnet sich aus der Varianz zwischen den einzelnen Faktorstufen MQ_{Zw} sowie aus der Varianz innerhalb der k Faktorstufen, Gruppen oder Bedingungen MQ_{In}.

$$F \;=\; \frac{MQ_{Zw}}{MQ_{In}} \tag{3.21}$$

Die Varianz bzw. mittlere Quadratsumme MQ ist die Summe der Abweichungsquadrate SAQ dividiert durch ihre Freiheitsgrade df, siehe Tabelle 3.8 und 3.9. Die Summen der Abweichungsquadrate SAQ ergeben sich anhand der aufsummierten Abweichungsquadrate der Gruppenmittelwerte μ_i vom Gesamtmittelwert $\bar{\mu}$ mit SAQ_{Zw} und anhand der aufsummierten Abweichungsquadrate der Messwerte innerhalb einer Gruppe vom Gruppenmittelwert μ_i mit SAQ_{In}. Letztere wird auch als Fehlervarianz bezeichnet und wird für die *ANOVA mit Messwiederholung* nochmals in zwei Teile zerlegt. Tabelle 3.8 und 3.9 gibt eine Übersicht über alle benötigten Gleichungen zur Berechnung der ein-faktoriellen *ANOVA*s. Bei der *ANOVA mit Messwiederholung* wird die

SAQ	**df**	**MQ**	**F**
$SAQ_{Zw} = \sum_{i=1}^{k} n_i \cdot (\mu_i - \bar{\mu})^2$	$df_{Ges} = k - 1$	$MQ_{Zw} = \frac{SAQ_{Zw}}{df_{Zw}}$	$F = \frac{MQ_{Zw}}{MQ_{In}}$
$SAQ_{In} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mu_i})^2$	$df_{In} = n - k$	$MQ_{In} = \frac{SAQ_N}{df_{In}}$	
$SAQ_{Ges} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mu})^2$	$df_{Ges} = n - 1$	$MQ_{Ges} = \frac{SAQ_{Ges}}{df_{Ges}}$	

Tabelle 3.8: Berechnungen einer ein-faktoriellen *ANOVA ohne Messwiederholung.*

Variabilität zwischen den Vpn SAQ_e nicht weiter betrachtet und dementsprechend aus

SAQ_{In} herausgenommen. Hierfür werden die Abweichungen der Vpn-Mittelwerte vom Gesamtmittelwert quadriert, aufsummiert und mit der Anzahl k an Faktorstufen bzw. Gruppen multipliziert.

$$SAQ_e = k \cdot \sum_{j=1}^{n} (\tau_j - \bar{\mu})^2, \; mit \; \tau_j \; = \; \frac{1}{k} \sum_{i=1}^{k} x_{ij} \tag{3.22}$$

Die Prüfgröße F ergibt sich dann aus der mittleren Quadratsumme der Gruppenvariabilität MQ_{Zw} sowie aus der mittleren Quadratsumme der restlichen Fehlervarianz $MQ_{Res} = \frac{SAQ_{Res}}{df_{Res}}$ mit $SAQ_{Res} = SAQ_{Ges} - SAQ_{Zw} - SAQ_e$ [Kähler, 2002].

$$F \; = \; \frac{MQ_{Zw}}{MQ_{Res}} \tag{3.23}$$

Die genauen Rechenschritte sowie die den SAQ entsprechenden Freihheitsgrade df sind in Tabelle 3.9 zusammengefasst, wobei k der Anzahl an Faktorstufen entspricht.

SAQ	**df**	**MQ**	**F**
$SAQ_{Zw} = \sum_{i=1}^{k} n \cdot (\mu_i - \bar{\mu})^2$	$df_{Zw} = k - 1$	$MQ_{Zw} = \frac{SAQ_{Zw}}{df_{Zw}}$	$F = \frac{MQ_{Zw}}{MQ_{Res}}$
$SAQ_e = k \cdot \sum_{j=1}^{n} (\tau_j - \bar{\mu})^2$	$df_e = n - 1$	$MQ_e = \frac{SAQ_e}{df_e}$	
$SAQ_{Res} =$ $SAQ_{Ges} - SAQ_{Zw} - SAQ_e$	$df_{Res} =$ $(k-1) \cdot (n-1)$	$MQ_{Res} = \frac{SAQ_{Res}}{df_{Res}}$	
$SAQ_{Ges} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{\mu})^2$	$df_{Ges} = n - 1$	$MQ_{Ges} = \frac{SAQ_{Ges}}{df_{Ges}}$	

Tabelle 3.9: Berechnungen einer ein-faktoriellen *ANOVA mit Messwiederholung.*

Wurde eine entsprechende Prüfgröße F mit Hilfe der *ANOVA* oder *ANOVA mit Messwiederholung* ermittelt, ist diese mit dem kritischen Wert der F-Prüfverteilung zu vergleichen. Unterschreitet F diesen Wert F_{krit}, besteht kein signifikanter Unterschied zwischen den Mittelwerten der Faktorstufen und die Nullhypothese H_0 muss mit $p > 0,05$ beibehalten werden. Mit $F > F_{krit}$ ist die H_0 abzulehnen und die globale Hypothese, dass es irgendeinen Unterschied gibt, anzunehmen. Welche Faktorstufe für die Signifikanz verantwortlich ist, kann mit Hilfe des t-Tests nachgewiesen werden.

Nicht-parametrische Verfahren

Der H-Test von Kruskal und Wallis [1952] findet vor allem bei unabhängigen oder ordinalskalierten Messwerten der Faktorstufen Anwendung. Bei diesem Test handelt es

sich um ein Rangordnungsverfahren. In ihm werden den Messwerten aller k Faktorstufen Rangplätze entsprechend ihrer Größe vergeben. Den mehrfach vorkommenden Werten werden, wie beim *Wilcoxon*-Test, gemittelte Rangplätze zugeteilt. Im Anschluss werden die in einer Faktorstufe UV_i enthaltenen Ränge zu einer Rangsumme mit T_i zusammengefasst. Die χ^2-verteilte Prüfgröße H berechnet sich dann nach Gleichung 3.24, wobei n_i der Anzahl der in der Faktorstufe UV_i enthaltenen Messwerte und k der Anzahl an Faktorstufen entspricht.

$$H = \frac{12}{n \cdot (n+1)} \cdot \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3 \cdot (n+1), \; mit \; df = k-1 \; und \; n = \sum_{i=1}^{k} n_i \qquad (3.24)$$

Die Nullhypothese kann dann abgelehnt werden, wenn H größer oder gleich dem kritischen χ^2_{krit} ist. Andernfalls muss sie beibehalten werden.

Sind die Messwerte der Faktorstufen voneinander abhängig, kann der *Friedman*-Test nach Friedman [1937] durchgeführt werden. In diesem Test wird für jede Vpn eine eigene Rangreihe gebildet, indem der höchste Rang maximal der Anzahl der Faktorstufen entspricht, für die die Vpn einen Messwert enthält. Anschließend werden alle Ränge pro Faktorstufe bzw. Gruppe zu T_i aufsummiert. Wie beim H-Test werden bei der Berechnung der Prüfgröße χ^2 alle Rangsummen berücksichtigt. Die Prüfverteilung ist in diesem Fall χ^2-verteilt [Hussy und Jain, 2002].

$$\chi^2 = \frac{12}{n \cdot k \cdot (k+1)} \cdot \sum_{i=1}^{k} T_i^2 - 3 \cdot n \cdot (k+1), \; mit \; df = k-1 \qquad (3.25)$$

Mit $\chi^2 > \chi^2_{\alpha;df}$ ergibt sich ein signifikantes Ergebnis, so dass die $H_0(Friedman - Test)$ abgelehnt und die $H_1(Friedman - Test)$ angenommen werden kann. Ein weiterführender *Wilcoxon*-Test kann dann die Signifikanz anhand von Paarungsvergleichen lokalisieren. Tabelle 3.10 zeigt die Rangvergabe beim *Friedman*-Test und die daraus resultierenden Rangsummen T_i, die für die Berechnung der Prüfgröße χ^2 notwendig sind. Mit der Gleichung 3.25 berechnet sich die Prüfgröße wie folgt:

$$\chi^2 = \frac{12}{10 \cdot 3 \cdot 4} \cdot (12^2 + 20^2 + 28^2) - 3 \cdot 10 \cdot 4 = 13,84 \qquad (3.26)$$

Mit $\chi^2_{krit} = \chi^2_{0,05;2} = 5,99$ überschreitet χ^2 den kritischen Wert und weist somit auf einen signifikanten Unterschied zwischen den Faktorstufen hin.

3.4.4 Auswertung nominal- und ordinalskalierter Messreihen

Oftmals sind die Messdaten nicht intervallskaliert bzw. auf metrischem Messniveau, sondern nominal- oder ordinalskaliert. Diese Skalenniveaus erhält man häufig bei der

UV_1 mit Messwert → **Rang**	UV_2 mit Messwert → **Rang**	UV_3 mit Messwert → **Rang**
14 → **1,5**	14 → **1,5**	16→ **3**
10 → **1**	11 → **2,5**	11 → **2.5**
12 → **2,5**	11 → **1**	12 → **2.5**
13 → **1**	14 → **2**	15 → **3**
11 → **1**	13 → **2**	14 → **3**
9 → **1**	11 → **3**	10 → **2**
12 → **1**	13 → **2**	14 → **3**
14 → **1**	15 → **2**	16 → **3**
11 → **1**	13 → **2**	14 → **3**
9 → **1**	11 → **2**	12 → **3**
$T_1 = 12$	$T_2 = 20$	$T_3 = 28$

Tabelle 3.10: *Friedman*-Test am Beispiel von drei Faktorstufen. Für jede Vpn bzw. Zeile beginnt eine neue Rangreihe mit eventuell gemittelten Rangplätzen.

Auswertung von Fragebögen, in denen den Vpn k Antwortmöglichkeiten vorgegeben werden. In diesem Fall sind die Häufigkeiten der gegebenen Antworten, auf Signifikanz zu überprüfen. In der Regel besteht für jede der k Antwortmöglichkeiten dieselbe Wahrscheinlichkeit, gewählt zu werden. Die erwartete Häufigkeit f_e ist demnach für alle Antwortmöglichkeiten gleich, mit $f_e = \frac{n}{k}$. Diese Gleichverteilung kann mit Hilfe des χ^2-Test nach Pearson [1900] überprüft werden. Die Prüfverteilung ist in diesem Fall also χ^2-verteilt und die Prüfgröße berechnet sich anhand der Summe der quadrierten und standardisierten Abweichungen der beobachteten Häufigkeiten f_{o_i} von den erwarteten Häufigkeiten f_{e_i}, siehe 3.27.

$$\chi^2 \;=\; \sum_{i=1}^{k} \frac{(f_{o_i} - f_{e_i})^2}{f_{e_i}},\; mit\; df = k - 1 \tag{3.27}$$

Wird der kritische Wert $\chi^2_{krit} = \chi^2_{\alpha,df}$ vom Prüfwert χ^2 überschritten, kann von einem signifikanten Unterschied zwischen den Häufigkeiten mit $p < 0,05$ ausgegangen werden.

Besteht nicht für jede Antwortmöglichkeit k, die gleiche theoretische Wahrscheinlichkeit gewählt zu werden, berechnet sich die erwartete Häufigkeit wie folgt:

$$f_{e_i} = \frac{n \cdot v_i}{s},\; mit\; s = \sum_{i=1}^{k} v_i \tag{3.28}$$

v_i entspricht dabei der Verhältniszahl der jeweiligen Antwortmöglichkeit.

Die Anwendung des χ^2-Test macht häufig erst dann Sinn, wenn mit einer erwarteten Mindesthäufigkeit von 5 gerechnet werden kann und der Freiheitsgrad mindestens

1 beträgt. Ist diese Voraussetzung nicht erfüllt, kann ein *Exakter Test nach Fisher* durchgeführt werden [Fisher, 1925]. Mit diesem Test können auch bei einer geringen Stichprobengröße zuverlässige Ergebnisse erzielt werden.

3.5 Präsentation der Ergebnisse

Die Präsentation der Ergebnisse kann sowohl deskriptiv, grafisch als auch durch Darstellung statistischer Kennwerte erfolgen. Je weniger Informationen über eine Studie einsehbar sind, desto kritischer kann diese betrachtet werden und die Zuverlässigkeit der Ergebnisse kann in Frage gestellt werden.

Die wichtigsten und für den Betrachter interessantesten Informationen können durch die deskriptive Beschreibung der Messdaten vermittelt werden. Darunter fallen vor allem Häufigkeiten, Durchschnittswerte, Summen, Varianzen, Maximal- und Minimalwerte. Mit Hilfe von Diagrammen können diese Informationen grafisch zusammengefasst und dargestellt werden. Oftmals werden hierfür Balkendiagramme, Tortendiagramme, Histogramme oder Boxplots verwendet. Während mit Torten- und Balkendiagrammen lediglich Häufigkeiten und Mittelwerte dargestellt werden können, ist mit Hilfe von Boxplots eine Aussage über die Verteilung der Daten möglich, siehe Abbildung 3.10. Die rechteckige Box der Boxplots umfasst dabei die mittleren 50% der Daten, die wiederum durch den Median in zwei Bereiche aufgeteilt werden. Der obere Bereich der Box, begrenzt durch das 3. Quartil enthält 25% der Daten die größer als der Median sind. Entsprechend enthält der untere Bereich die restlichen 25% der Daten innerhalb der Box. Liegt der Median genau in der Mitte der Box und die Box selbst in der Mitte der Whisker, kann auf eine Normalverteilung der Daten geschlossen werden. Die Definition der Whisker sowie der Fehlerbalken in den Balkendiagrammen erfolgt häufig unterschiedlich. Sie können entweder die Extremwerte der Daten, aber auch die Abweichungen repräsentieren. Häufig sind die Whisker aber auch durch das $1,5$-fache der Quartilabstände begrenzt. Neben der grafischen Darstellung ist auch die Zuverlässigkeit der Ergebnisse von besonderem Interesse. Aus diesem Grund sollten neben Angaben zum Versuchsdesign auch Angaben zu den durchgeführten Signifikanztests gemacht werden. Zudem können das verwendete Testniveau α, die Teststärke $1 - \beta$, die Effektgröße δ sowie die Größe der Stichprobe präsentiert werden.

3.6 Zusammenfassung

Zum besseren Verständnis der weiteren Ausführungen wurden in diesem Kapitel zunächst die im Zusammenhang mit Evaluierungen und Hypothesenprüfung verwendeten

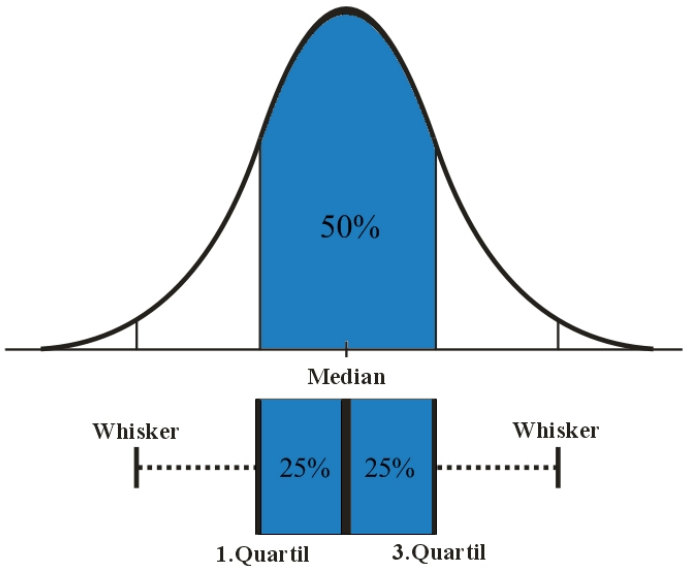


Abbildung 3.10: Mit Hilfe von Boxplots können Angaben über Mittelwert, Extremwerte,Streuung und Verteilung der Daten gemacht werden.

Fachbegriffe definiert und erklärt. Darauf aufbauend wurde beschrieben, wie die verschiedenen Hypothesenformen ausgehend von der Problemstellung und unter Berücksichtigung möglicher Fehlerquellen zu formulieren sind. Damit wurde die Grundlage für die Erstellung eines experimentellen Designs auf dem Wege der Operationalisierung dargestellt.

Ein Versuchsplan und die Bestimmung der optimalen Stichprobengröße ermöglichen eine Aufwandsabschätzung für die Durchführung des Experimentes. Darüber hinaus wurde in Abschnitt 3.4 ausgeführt, wie die durch das Experiment gewonnenen Daten bei der Hypothesenüberprüfung analysiert werden können.

Unter Berücksichtigung der grundlegenden Struktur einer experimentellen Evaluierung wird im folgenden Kapitel ein spezielles Versuchsdesign zur Bewertung der Hervorhebung von Fokus-Strukturen in der medizinischen Visualisierung konzipiert.

4 Entwurf des Versuchsdesigns

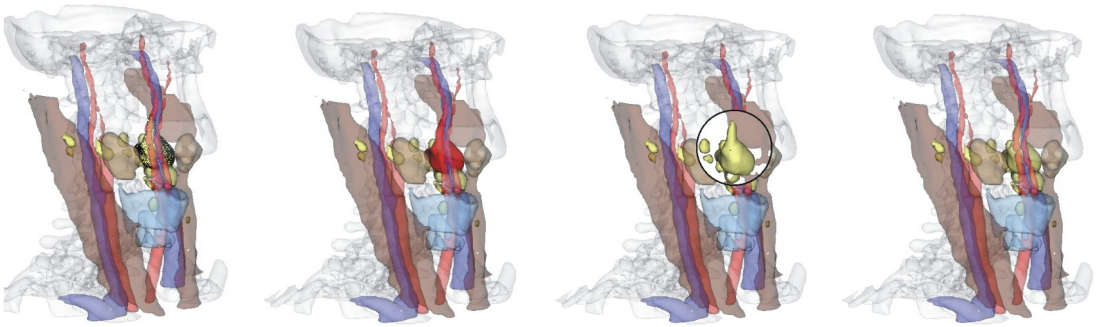
Die im Rahmen dieser Arbeit durchgeführte Studie untersucht am Beispiel ausgewählter Hervorhebungstechniken, ob diese die Detektion von Fokusstrukturen innerhalb komplexer medizinischer Visualisierungen signifikant unterstützen. Für die Durchführung der Studie wird ein entsprechendes experimentelles Design entworfen, so dass weitere auf diesem Design aufbauende Studien reproduzierbar und vergleichbar sind. Voraussetzung dafür ist, dass die Gestaltung des Designs entsprechend den in Kapitel 3 dargelegten wissenschaftlichen Erkenntnissen hinsichtlich der Planung und Durchführung von Experimenten sowie der validen Hypothesenprüfung erfolgt. Zudem sind die in Abschnitt 2.3 besprochenen Theorien zur visuellen Suche in Bildern zu berücksichtigen.

4.1 Aufgabenstellung

Für die Hervorhebung vergrößerter Lymphknoten in 3D-Renderings von Halsdatensätzen wurden verschiedene Techniken implementiert, die die Detektion pathologischer Strukturen in 3D-Umgebungen erleichtern sollen. Bisher konnte jedoch nicht bewiesen werden, ob dies tatsächlich der Fall ist.

Nach dem Vorbild psychophysischer Methoden ist ein Design für einen Reaktionszeittest zu konzipieren, mit dem die visuelle Wahrnehmung bezüglich der Techniken der lokalen Hervorhebung mit *roter Einfärbung*, dem illustrativen Pen&Ink-Vertreter *Stippling* sowie der regionalen Hervorhebungstechnik *CutAway* gemessen werden kann. Anhand der mit diesem Test erfassten Messdaten sind die jeweiligen Hervorhebungstechniken miteinander zu vergleichen und daraufhin zu untersuchen, ob mit ihrer Hilfe ein hervorgehobener vergrößerter Lymphknoten leichter wahrgenommen werden kann als ein nicht hervorgehobener vergrößerter Lymphknoten.

Abbildung 4.1 zeigt einen vergrößerten Lymphknoten hervorgehoben mit jeweils einer dieser drei Techniken. Im Vergleich dazu ist in Bild (d) derselbe Lymphknoten ohne Hervorhebung dargestellt.



(a) Stippling (b) Rote Einfärbung (c) CutAway (d) Keine Hervorhebung

Abbildung 4.1: Hervorhebung eines vergrößerten Lymphknoten mit *Stippling, CutAway* und *roter Einfärbung* innerhalb einer komplexen medizinischen Halsvisualisierung. In Bild (d) ist derselbe Lymphknoten nicht hervorgehoben.

4.2 Hypothesen

In Abschnitt 3.2.1 wurden die verschiedenen Formen, wie eine Forschungshypothese zu formulieren ist, vorgestellt. Für diese Studie werden Forschungshypothesen sowohl bezüglich der Häufigkeit der visuellen Erfassung hervorgehobener vergrößerter Lymphknoten als auch bezüglich der Reaktionszeit, mit der sie erfasst wurden, postuliert. Anschließend sind die für die Untersuchung notwendigen Variablen UV und AV aus den Hypothesen heraus zu operationalisieren, so dass ein entsprechender Versuchsplan entworfen werden kann.

4.2.1 Hypothesenpostulierung

Bisher gibt es keine vergleichbaren Untersuchungen bezüglich der Detektionsgenauigkeit von hervorgehobenen vergrößerten Lymphknoten in medizinischen Visualisierungen. Dennoch wird davon ausgegangen, dass mit Hilfe von Hervorhebungstechniken die visuelle Erfassung der Lymphknoten unterstützt wird, so dass im Folgenden gerichtete unspezische Hypothesen formuliert werden können. Aus den Forschungshypothesen abgeleitet ergeben sich entsprechende Nullhypothesen, die anhand geeigneter Signifikanztests zu überprüfen sind.

- ***Forschungshypothese* $\mathbf{H_1}$(Detektion)*:***

  Mit *Stippling, CutAway* oder rot hervorgehobene vergrößerte Lymphknoten in farbigen 3D-Renderings von Halsdatensätzen werden häufiger vom Betrachter erfasst als vergrößerte Lymphknoten ohne Hervorhebung.

- ***Nullhypothese* $\mathbf{H_0}$(Detektion)*:***

> Es besteht kein Unterschied zwischen mit *Stippling, CutAway* oder rot hervorgehobenen vergrößerten Lymphknoten und vergrößerten Lymphknoten ohne Hervorhebung bezüglich ihrer Erfassbarkeit in farbigen 3D-Renderings von Halsdatensätzen.

Des Weiteren stellt sich die Frage, ob hervorgehobene vergrößerte Lymphknoten schneller wahrgenommen werden als Lymphknoten ohne Hervorhebung. Auch hierzu gibt es keine vergleichbaren Studien in diesem Bereich, so dass die Forschungshypothesen bezüglich der erwarteten Reaktionszeit ebenfalls unspezifisch, dennoch gerichtet zu formulieren sind.

- ***Forschungshypothese* $\mathbf{H_1}$(Reaktionszeit)*:***

> Mit *Stippling, CutAway* oder rot hervorgehobene vergrößerte Lymphknoten in farbigen 3D-Renderings von Halsdatensätzen werden schneller erfasst als vergrößerte Lymphknoten ohne Hervorhebung.

- ***Nullhypothese* $\mathbf{H_0}$(Reaktionszeit)*:***

> Es besteht kein Unterschied in der Reaktionszeit zur Erfassung mit *Stippling, CutAway* oder rot hervorgehobener vergrößerter Lymphknoten sowie vergrößerter Lymphknoten ohne Hervorhebung in farbigen 3D-Renderings von Halsdatensätzen.

In Folge eines explorativen Pilotexperimentes können weitere Forschungshypothesen in Bezug auf die besondere Eignung einer Technik formuliert werden.

4.2.2 Operationalisierung

Entsprechend den Forschungshypothesen sind die Hervorhebungstechniken, mit denen vergrößerte Lymphknoten dargestellt werden können, die drei Faktorstufen UV_i mit $i \in \{1, 2, 3\}$ des Faktors vergrößerter Lymphknoten (vLK). Diese sind einer vierten Faktorstufe, der Kontrollbedingung vLK ohne Hervorhebung, gegenüber zu stellen. Die Detektionsleistung und die benötigte Zeit zum Erfassen eines vergrößerten Lymphknotens bilden die abhängigen Variablen AV_{TQ} und AV_{RT}. Diese sollen mit Hilfe eines Reaktionszeitexperimentes, welches in Abschnitt 4.5 vorgestellt wird, gemessen werden. Zusammengefasst ergeben sich folgende Untersuchungsvariablen:

- UV: vergrößerte Lymphknoten
  - UV_1: Hervorhebung mit Stippling
  - UV_2: Hervorhebung mit roter Einfärbung
  - UV_3: Hervorhebung mit CutAway
  - UV_4: Keine Hervorhebung
- AV_{TQ}: Detektionsgenauigkeit in %
- AV_{RT}: Detektionszeit in ms

Bisherige Studien wie [Kosara et al., 2003], [Kobsa, 2004], [Kobsa, 2001] und [Isenberg et al., 2006], die sich mit der Wahrnehmung von Visualisierungstechniken befasst haben, führten ihren Signifikanztest mit einem α-Niveau von $0,05$ durch. Die in dieser Arbeit angewendeten Tests werden ebenfalls auf diesem Signifikanzniveau durchgeführt. Weiterhin wird ein β-Fehler bzw. ein Fehler 2.Art von $0,2$ riskiert, so dass sich eine Teststärke von $1 - \beta = 0,8$ ergibt. Diese Teststärke findet vor allem in sozialwissenschaftlichen Studien Anwendung [Bortz und Döring, 2006, S.604]. Die für diese Studie entsprechenden Prüfparameter lassen sich demnach wie folgt zusammenfassen:

- Testniveau: $\alpha = 0,05$
- Teststärke: $1 - \beta = 0,8$

4.3 Effektgrößen und Stichprobenumfang

Kann nach Cohen [1992] eine Effektstärke bzw. -größe δ hinsichtlich der Detektionsgenauigkeit einer Hervorhebungstechnik geschätzt werden, ist eine erste Aufwandsabschätzung bezüglich der benötigten Mindeststichprobengröße n_{opt} für das durchzuführende Experiment möglich.

4.3.1 Berechnung einer optimalen Stichprobengröße

Sind bereits kleinere Unterschiede von praktischer Bedeutung, sollte eine kleine Effektgröße und somit eine größere Stichprobe gewählt werden. Die Anwendung von Hervorhebungstechniken ist nicht gerechtfertigt, wenn die Wahrnehmung vergrößerter Lymphknoten nur minimal verbessert wird. Neben statistisch nachgewiesenen signifikanten Unterschieden wird eine eindeutige, praktisch relevante Verbesserung der Erfassung erwartet, welche die Sicherheit der Erkennung der vergrößerten Lymphknoten erhöhen soll. Daher wird zunächst ein mittlerer Effekt mit $\delta_{Acc} = 0,5$ als praktisch

relevante Effektgröße für die Detektionsgenauigkeit gefordert. Daraus lässt sich die benötigte Stichprobengröße mit Hilfe der Gleichung 3.7 wie folgt ermitteln:

$$n_{opt} \geq (\frac{z_{0,05} - z_{0,8}}{0,5})^2 = (\frac{-1,64 - 0,84}{0,5})^2 = 24,6 \tag{4.1}$$

Ein zu erwartender Effekt hinsichtlich der benötigten Zeit zur Erfassung eines vergrößerten Lymphknotens kann ebenfalls nur geschätzt werden, da keine vergleichbaren Studien vorliegen, aus denen eine solche Größe abgeleitet werden könnte. Erwartet werden Reaktionszeiten im Millisekunden-Bereich. Für den Anwender wird jedoch höchstwahrscheinlich erst ein Unterschied in der aufgewendeten Zeit im Sekunden-Bereich von Bedeutung sein. Aus diesem Grund wird hier nur ein großer Effekt für δ_{RT} eine praktische Relevanz haben. Tabelle 4.1 zeigt den benötigten Stichprobenumfang, um einen entsprechenden Effekt für δ_{Acc} und δ_{RT} feststellen zu können. Hierbei ergeben sich mit $\alpha = 0,05$ und $\beta = 0,2$ $n_{opt} = 25$ für δ_{Acc} und $n_{opt} = 10$ für δ_{RT}.

α-**Niveau**	$\delta_{\mathbf{Acc}} = \mathbf{0,5}$	$\delta_{\mathbf{RT}} = \mathbf{0,8}$
0,05	**25**	10
0,01	40	16

Tabelle 4.1: Optimale Stichprobenumfänge für einen 1-seitigen Signifikanztest mit einer Teststärke $1 - \beta$ von 0.8 und für zwei unterschiedliche Signifikanz-Niveaus.

Da n_{opt} den benötigten Mindeststichprobenumfang angibt, sollte die Entscheidung immer für die kleinere Effektgröße auf dem entsprechenden Signifikanz-Niveau und somit für die größere Stichprobengröße getroffen werden. Das in dieser Arbeit durchgeführte Experiment ist mit $\alpha = 0,05$ somit an mindestens 25 Vpn durchzuführen. Jedoch kann es vorkommen, dass einige Versuchsdurchläufe oder die ganze Messreihe einzelner Vpn aufgrund technischer Probleme oder ergebnisverzerrender Leistungen einiger Teilnehmer als Ausreißer von der Datenanalyse ausgenommen werden. Daher sollten über die berechnete Mindeststichprobe hinaus ein paar mehr Probanden an der Studie teilnehmen.

Nach Abschluss der Experimentdurchführung an allen Vpn erfolgt die statistische Datenanalyse. Werden hier signifikante Ergebnisse festgestellt, können mit Gleichung 4.2 die tatsächlichen Effektgrößen δ'_{Acc} und δ'_{RT}, die innerhalb der Stichprobe vorliegen, berechnet werden.

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{4.2}$$

Für σ kann dabei entweder eine oder der Durchschnitt der beiden Standardabweichungen σ_1 und σ_2 der Gruppen UV_1 und UV_2 gewählt werden. In dieser Arbeit entspricht σ immer dem Durchschnitt beider Standardabweichungen.

Ist $\delta'_{Acc} \geq \delta_{Acc}$, liegt ein praktisch relevanter Unterschied bezüglich der Detektionsleistung vor. Mit $\delta'_{RT} \geq \delta_{RT}$ gibt es einen relevanten Unterschied in der benötigten Zeit, die für die Detektion eines vLK aufgebracht wurde.

Liegen bereits die ersten Messergebnisse vor, kann n_{opt} anhand der vorläufigen Mittelwerte und Standardabweichungen angepasst werden. Die hierfür benötigte Effektgröße lässt sich ebenfalls anhand der Gleichung 4.2 ermitteln.

4.3.2 Akquirierung von Versuchspersonen

Teilnehmer an experimentellen Studien sollten nach Möglichkeit der Zielpopulation angehören. In diesem Fall wären dies idealerweise Ärzte, die bereits über Erfahrungen mit medizinischen Visualisierungen verfügen. Verwandte Arbeiten zeigten jedoch, dass Personen aus diesem Arbeitsfeld, zumeist durch zeitliche aber auch örtliche Faktoren bedingt, sich nur sehr schwierig für eine solche Studie gewinnen lassen ([Mirschel, 2004], [Oeltze, 2004]). Daher wird diese Studie weitestgehend mit Studenten und Mitarbeitern der Fakultät für Informatik sowie dem Institut für Psychologie durchgeführt. Dies entspricht dem Prinzip der Zufallsstichprobe. Vermutlich würden Ärzte aufgrund ihrer Erfahrungen durch die tägliche Praxis bessere Ergebnisse erzielen, als ungeübte Personen. Dennoch sollten die aus den mit dieser Stichprobe gewonnenen Messdaten in der Tendenz hinsichtlich der Wahrnehmungsunterstützung sich auf die Zielpopulation übertragen lassen. Anhand einer Aufschlüsselung der Ergebnisse nach Tätigkeit und Erfahrung der Vpn kann dies überprüft werden. Eine ausreichend große Stichprobengruppe mit erfahrenen Vpn kann jedoch aus oben genannten Gründen nicht erreicht werden, so dass das Ergebnis der Aufschlüsselung nicht repräsentativ ist. Das Design kann auch im Rahmen des Medizinstudiums für Ausbildungszwecke verwendet werden, in diesem Fall sind Zielgruppe und Zufallsstichprobe nahezu kongruent.

Soweit Untersuchungen mit Vpn durchgeführt werden, sind bestimmte Faktoren zu berücksichtigen, wie sie auch in der Psychodiagnostik angewendet werden. Besonders wichtig sind dabei Transparenz, Unverfälschbarkeit, Zumutbarkeit und Fairness. Weiterhin sollten den Probanden nur soviel Informationen zur Verfügung stehen, so dass ihr Handeln keinem bestimmten Motiv zugrunde liegt.

4.4 Versuchsplan

Die in 4.2.2 operationalisierten Variablen lassen sich in einen ein-faktoriellen Versuchsplan mit vier Faktorstufen übertragen. Tabelle 4.2 zeigt den aus Abschnitt 3.3.2 vorgestellten Versuchsplan nach Fisher [1935], angepasst auf die vorliegende Studie. Die in Abschnitt 4.3.1 berechnete Mindeststichprobengröße ist in der ersten Spalte der Tabelle 4.2 dargestellt. Um den experimentellen Aufwand gering zu halten, wird ein

Vpn	Faktorstufen eines vergrößerten Lymphknoten			
	Stippling	**Rote Einfärbung**	**CutAway**	**Keine Hervorhebung**
1	x_{11}	x_{12}	x_{13}	x_{14}
2	x_{21}	x_{22}	x_{23}	x_{24}
...	...	...	...	...
j	x_{j1}	x_{j2}	x_{j3}	x_{j4}
...	...	...	...	...
n	x_{n1}	x_{n2}	x_{n3}	x_{n4}
	μ_1	μ_2	μ_3	μ_4

Tabelle 4.2: Ein-faktorieller Versuchsplan mit 4 Faktorstufen des Faktors *vergrößerter Lymphknoten* vLK und einem Stichprobenumfang von n. Die Gruppenmittelwerte sind durch μ_i und die Mittelwerte pro Vpn durch τ_j gegeben.

within-subject-Design verwendet. Dabei wird jede Vpn auf jede Faktorstufe hin untersucht. In Tabelle 4.2 wird demnach jede Zelle x_{ij} mit einem Messwert belegt sein und n ist mindestens so groß wie n_{opt}. Vorteil eines solchen Versuchdesigns ist zum einen die automatische Parallelisierung der Vpn über alle Faktorstufen. Zum anderen erhält man dabei gleich große Stichprobengruppen. Ein weiterer Vorteil besteht in der automatischen Kontrolle zeitlicher Störvariablen [Sedlmeier und Renkewitz, 2008].

In einem *between-subject*-Design würde jede Vpn nur auf eine der vier Hervorhebungstechniken untersucht werden, so dass pro Vpn nur ein Zelle x_{ij} belegt wäre. In diesem Fall muss n_{opt} für jede der vier Faktorstufen gelten, so dass $n \geq 4 \cdot n_{opt}$ gilt. Für dieses Design würden also erheblich mehr Vpn benötigt als bei einem *within-subject*-Design.

4.5 Versuchsanordnung

Die Forschungshypothesen gehen davon aus, dass Hervorhebungstechniken die visuelle Erfassung vergrößerter Lymphknoten (vLK) in medizinischen Halsvisualisierungen erleichtern. Anhand eines Reaktionszeittests, wie er in der experimentellen Psychologie angewendet wird, soll diese Annahme überprüft werden.

In den folgenden Abschnitten wird dargestellt, wie die Stimuli den Vpn zu präsentieren sind und wie deren Aufmerksamkeit während des Experimentes kontrolliert werden kann. Da die objektiv erfassten Messdaten anschließend mit den subjektiven Eindrücken der Vpn verglichen werden sollen, ist hierfür ein Fragebogen zu entwerfen.

4.5.1 Aufgabenstellung für die Versuchspersonen

In einem einfachen *ja-/nein*-Experiment werden den Vpn verschiedene Halsvisualisierungen gezeigt, die sich im Wesentlichen darin unterscheiden, ob sie einen vergrößerten Lymphknoten (vLK) enthalten oder nicht. Die Aufgabe der Vpn besteht darin, den vLK innerhalb einer festgelegten Zeit zu finden. Hat die Vpn den Zielreiz vLK im Bild gefunden, muss sie mit *ja* antworten, andernfalls mit *nein*. Abbildung 4.2 zeigt links eine Halsvisualisierung ohne den Zielreiz vLK, welche dem Rauschbild N entspricht. Im rechten Bild hingegen ist dasselbe Rauschbild mit dem Zielreiz $N+vLK$ dargestellt.

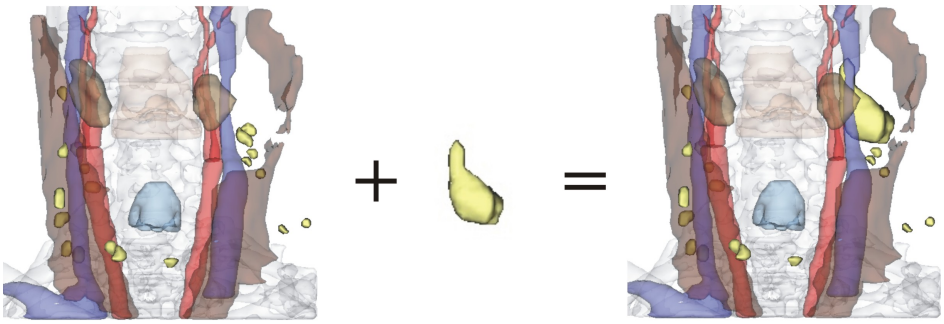


Abbildung 4.2: Den Vpn werden zwei verschiedene Arten von Stimuli präsentiert. Entweder ein Rauschbild N ohne Darstellung eines vergrößerten Lymphknotens (links) oder ein Rauschbild, welches einen Zielreiz, den vergrößerten Lymphknoten vLK, enthält $N + vLK$ (rechts). Aufgabe der Vpn ist es, diesen Zielreiz im Rauschbild zu detektieren.

Alle Anweisungen sind grundsätzlich schriftlich in kurzen, präzisen Sätzen zu erteilen. Dies gewährleistet, dass alle Vpn dieselben Informationen über den Versuchsablauf erhalten. Zudem dürfen gegenüber den Vpn keine Aussagen über das Ziel der Studie getroffen werden, da ansonsten Strategien zur Bewältigung der Aufgabe entwickelt werden könnten. Aus demselben Grund wird allen Vpn die Instruktion erst direkt vor Beginn des Experimentes gegeben.

In der Instruktion werden den Vpn zunächst diejenigen Strukturen und Hervorhebungstechniken vorgestellt, die ihnen in den Halsvisualisierungen gezeigt werden sollen. Anschließend werden die Vpn angewiesen, in jedem Bild einen vergrößerten Lymphknoten, sofern vorhanden, zu detektieren, unabhängig davon, ob dieser hervorgehoben ist oder nicht. Ob die Aufgabenstellung verstanden wurde, wird anhand eines kurzen Testdurchlaufs überprüft.

4.5.2 Anforderungen an die Stimuli

Um die Messergebnisse sowohl einer Vpn als auch mehrerer Vpn und Gruppen miteinander vergleichen zu können, muss die Erfassung der Daten unter den gleichen Be-

dingungen stattfinden. Zum einen muss jedes Stimulusbild gleich groß sein, da nach Treisman und Gelade [1980] die Dauer einer Konjunktionssuche mit der Größe des Bildes zunimmt. Zum anderen lassen die Ergebnisse der Studie von Tory [2003] vermuten, dass frontale Ansichten die günstigste Variante für die Orientierung innerhalb dreidimensionaler Szenen sind. Daher sind für einen konstanten Suchaufwand in jedem Stimulusbild immer dieselben Strukturen aus coronaler Sicht und orthogonaler Projektion darzustellen. Lediglich die Form der einzelnen Strukturen darf sich unterscheiden, wie dies auch der medizinischen Praxis entspricht. Auf diese Weise soll auch ein Wiedererkennungseffekt der Bilder und der zu suchenden Lymphknoten vermieden werden. Insgesamt werden sieben verschiedene Strukturen innerhalb der Stimuli zu sehen sein, die mögliche Risikostrukturen darstellen können (siehe Abbildung 4.3). Dazu gehören die paarigen Strukturen der Muskeln, Gefäße und Speicheldrüsen sowie der Kehlkopf, die Luftröhre, Knochen als auch kleine bis große Lymphknoten. Der Transparenzgrad ist dabei für jede Struktur unterschiedlich.

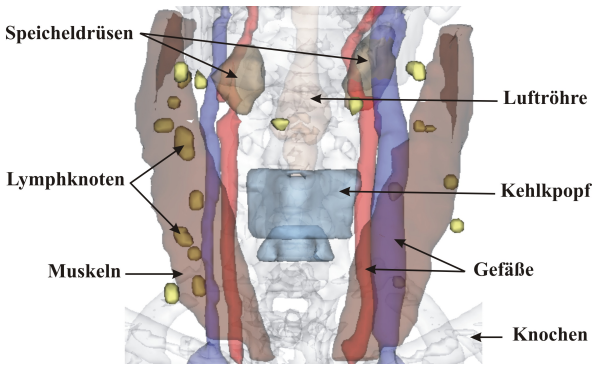
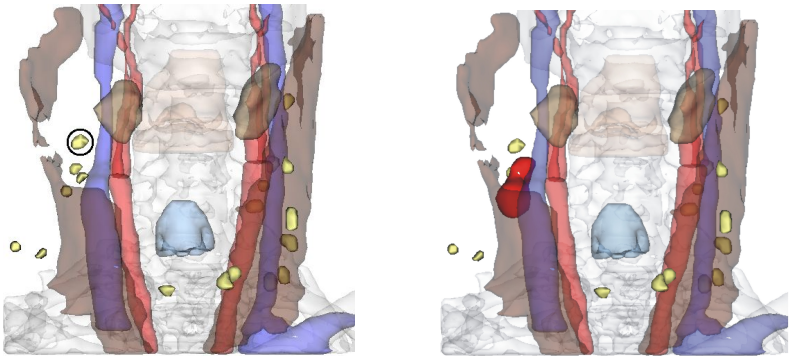


Abbildung 4.3: Strukturen der Halsvisualisierungen, wie sie in den Stimuli visualisiert sind.

Die Zielreize vLK werden mit den verschiedenen Hervorhebungstechniken innerhalb der Rauschbilder N dargestellt. Aufgrund des *within-subject*-Designs werden allen Vpn die selben Zielreize präsentiert. Dabei muss jede Reaktion einer Vpn darauf zurückzuführen sein, ob ein Zielreiz in einem Rauschbild vorhanden war oder nicht. Es darf maximal ein Zielreiz in einem Rauschbild enthalten sein, da andernfalls nicht zwischen mehreren Zielreizen differenziert werden kann.

Um zu vermeiden, dass die Vpn nur auf die ihnen präsentierten Hervorhebungstechniken reagieren, sollten Distraktoren innerhalb der Rauschbilder dargestellt werden. Damit soll die Aufmerksamkeit der Vpn während des Experimentes kontrolliert werden. Es handelt sich hierbei um nicht vergrößerte Lymphknoten, die ebenfalls verschieden hervorgehoben sind. Abbildung 4.4 zeigt jeweils eine mögliche Darstellung für einen Stimulus ohne (a) und mit einem Zielreiz (b) aus coronaler Sicht.



(a) Rauschbild mit einem Distraktor (hervorgehobener nicht vergrößerter Lymphknoten)

(b) Rauschbild mit hervorgehobenen Zielreiz

Abbildung 4.4: In Bild (a) ist ein Rauschbild mit einem falsch-positiven Lymphknoten dargestellt. In Bild (b) ist dasselbe Rauschbild mit einem vergrößerten Lymphknoten in rot zu sehen.

4.5.3 Präsentation der Stimuli

Die Anordnung der Stimuli-Reihenfolge sowie deren Anzeigedauer haben einen erheblichen Einfluss darauf, wie die Vpn auf die ihnen präsentierten Stimuli reagieren. Für die Vergleichbarkeit der Messergebnisse ist es daher wichtig, dass allen Vpn die Stimuli in gleicher Anzahl und in gleich vielen Durchgängen präsentiert werden.

Die Anzeigedauer der Stimuli ist immer gleich und unabhängig von der Reaktion der Vpn. Bei der Suchaufgabe handelt es sich um eine Konjunktionssuche. Das heißt, dass das Bild seriell nach dem vergrößerten Lymphknoten abgesucht wird [Treisman und Gelade, 1980]. Aufgrund eigener Erfahrung als Teilnehmer an mehreren Experimenten sowie unter Rücksprache mit einem Diplom-Psychologen wird die Anzeigedauer eines Stimulus zunächst auf eine Sekunde festgelegt. Eine kürzere Anzeigedauer wäre aufgrund der Komplexität der Bilder nicht sinnvoll. Würden die Stimuli deutlich länger als eine Sekunde gezeigt, bestünde die Gefahr zu hoher Trefferquoten bei allen Faktorstufen. Eine Differenzierung zwischen diesen wäre dann anhand der Resultate kaum mehr möglich und man könnte die Treffer nicht mehr auf die Hervorherbungstechniken zurückführen. Die genaue Anzeigedauer der Bilder soll im Rahmen eines Pilotexperimentes ermittelt werden. Ein Experiment sollte die Dauer von einer Stunde nicht übersteigen, da ansonsten Ermüdungserscheinungen auftreten können.

Die Rauschbilder werden mit und ohne Zielreiz in zufälliger Reihenfolge gezeigt, wobei sich zwischen jedem Stimulus ein Übergangsbild mit einem Fokussierungskreuz befindet, siehe Abbildung 4.5. Insbesondere bei solchen komplexen Szenarien, wie sie in dieser Studie verwendet werden, wäre eine Differenzierung zwischen den Stimuli bei direktem Übergang zu schwierig und könnte die Reaktion und somit den Ausgang der Untersuchung beeinflussen. Daher ist es in der experimentellen Psychologie üblich, ein

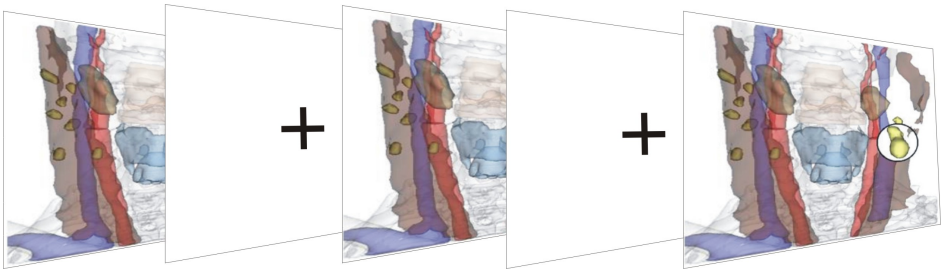


Abbildung 4.5: Anordnung der Stimuli in zufälliger Reihenfolge mit jeweils einem Interstimulus-Intervall zwischen den Stimuli. Die Vpn haben zu jedem Stimulus eine Entscheidung dahingehend zu fällen, ob sie einen *vLK* sehen oder nicht.

sogenanntes Interstimulus-Intervall einzuführen, in dem häufig ein Bild mit weißem oder schwarzen Hintergrund, im Fall visueller Stimuli, angezeigt wird. Das Anzeigeintervall dieses Bildes sollte unterschiedlich lang sein, so dass sich die Vpn nicht auf den Zeitpunkt des nächsten Stimulus vorbereiten können und ihre Aufmerksamkeit gehalten werden kann. Ist der Übergang zwischen zwei Stimuli zu lang, könnten die Vpn dazu verleitet werden, den Blick vom Bildschirm abzuwenden. Bei zu kurzer Anzeigedauer des Übergangbildes könnten die Vpn wiederum überfordert sein.

Um den *Positions-* und *Carry-Over*-Effekten entgegen zu wirken, werden die verschieden hervorgehobenen *vLK* und Distraktoren ebenfalls in zufälliger Reihenfolge in den Rauschbildern präsentiert.

4.6 Konzeption der Instruktionsanleitung und des Fragebogens

In der schriftlich formulierten Aufgabenstellung werden die Versuchsteilnehmer zunächst darauf aufmerksam gemacht, was für Bilder sie zu sehen bekommen und worin ihre Aufgabe bestehen wird. Da die diesem Experiment zugrunde liegenden Stimuli komplexe dreidimensionale Darstellungen von Halsdatensätzen darstellen und die Vpn überwiegend wenig Erfahrungen mit dieser Art von Bildern haben, werden ihnen zu Beginn die in den Bildern enthaltenen Halsstrukturen vorgestellt. Dabei werden die Vpn angewiesen sich vor allem das Erscheinungsbild der Lymphknoten einzuprägen, da unter diesen die vergrößerten Lymphknoten detektiert werden sollen. Weiterhin werden die Hervorhebungstechniken vorgestellt, in denen vergrößerte Lymphknoten und Distraktoren dargestellt sein können. Anschließend wird konkret die Aufgabenstellung formuliert. Folgende Fragen der Vpn sollen mit Hilfe der Instruktionsanleitung beantwortet werden:

- Was für Bilder werde ich zu sehen bekommen?

- Worauf soll ich in den Bildern achten?
- Welche Strukturen sind in den Bildern enthalten?
- Wie können vergrößerte Lymphknoten und Distraktoren dargestellt sein?
- Was muss ich tun, wenn ich einen bzw. keinen vergrößerten Lymphknoten sehe?

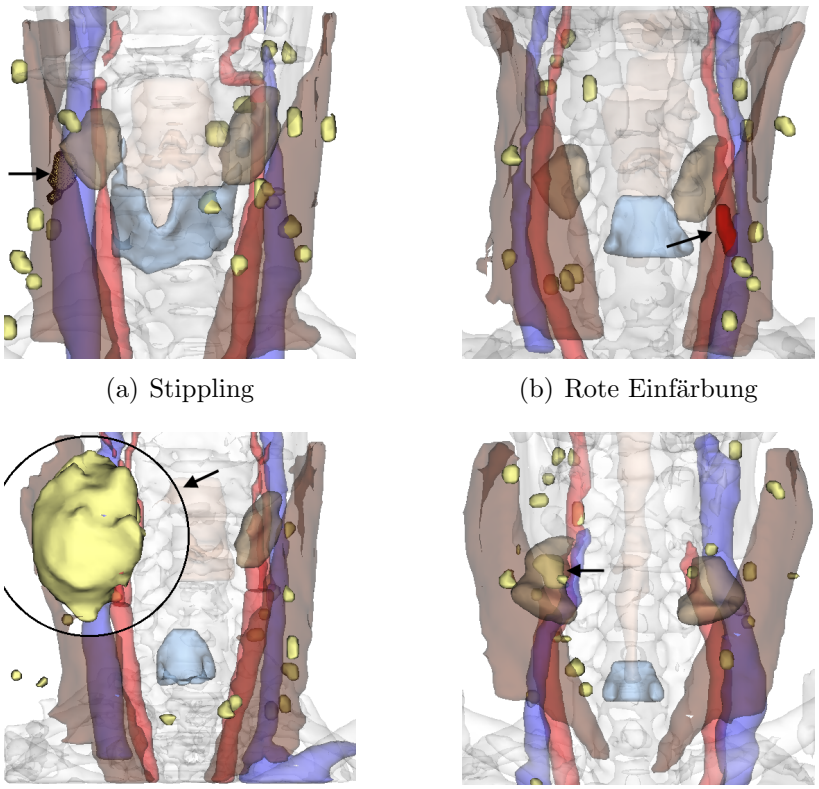
Zuletzt werden die Vpn darauf hingewiesen, dass jedes Bild nur sehr kurz angezeigt wird und sie sich daher nach Möglichkeit schnell entscheiden müssen. Die für dieses Experiment formulierte Instruktionsanleitung ist in Anhang C aufgeführt.

Im Anschluss an das Experiment werden die Ergebnisse mit den subjektiven Einschätzungen der Vpn verglichen. Hierfür wird ein Fragebogen erstellt, in welchem die Vpn Angaben zu ihrer Person und zu den von ihnen bevorzugten Hervorhebungstechniken machen sollen. Die Bewertung der Hervorhebungstechniken soll ohne Verwendung von Beispielbildern erfolgen, da bereits durch die Auswahl dieser die Entscheidung der Vpn beeinflusst werden können, siehe Abbildung 4.6. Die im Wege einer explorativen Datenanalyse gewonnenen Erkenntnisse aus den Fragebögen, können für weiterführende Studien von Bedeutung sein. Die Erhebung der demografischen Daten ermöglicht eine differenzierte Aufschlüsselung der Testergebnisse nach Alter, Geschlecht, Tätigkeit, mögliche Sehschwächen und Erfahrungen der Vpn mit medizinischen Visualisierungen. Weiterhin kann so überprüft werden, inwieweit sich die Ergebnisse die mit Hilfe der Zufallsstichprobe gewonnen wurden, auf die Zielpopulation übertragen lassen.

Darüber hinaus werden die Vpn dahingehend befragt, welche Strukturen sie während der Suche nach dem vergrößerten Lymphknoten als störend empfunden haben. Diese Informationen könnten für weitere Studien hinsichtlich der Kontextstrukturen von Interesse sein. Die Antworten der Vpn sind dabei ordinalskaliert aufzuzeichnen und entsprechend mit einem Analyseverfahren für ordinalskalierte Messdaten zu überprüfen.

4.7 Zusammenfassung

Der Entwurf eines Versuchsdesigns ist zum einen notwendig für die Reproduzierbarkeit des Experimentes, zum anderen können Ergebnisse verschiedener Studien nur dann miteinander verglichen werden, wenn diese nach einheitlichen und wissenschaftlichen Standards durchgeführt wurden. Ausgangspunkt für das in diesem Kapitel entworfene Versuchsdesign war die Aufgabenstellung. Im Rahmen einer Evaluierung soll ermittelt werden, welche der Hervorhebungstechniken *rote Einfärbung*, *CutAway* und *Stippling* die visuelle Wahrnehmung pathologischer Lymphknoten am besten unterstützt. Von



(a) Stippling (b) Rote Einfärbung

(c) CutAway (d) Keine Hervorhebung

Abbildung 4.6: Ungünstig ausgewählte Bilder für den direkten Vergleich der Hervorhebungstechniken *Stippling, CutAway* und *roter Einfärbung* und keiner Hervorhebung innerhalb einer komplexen medizinischen Halsvisualisierung. Die in Bild (a), (b) und (d) enthaltenen vergrößerten Lymphknoten (mit Pfeil markiert) heben sich nur wenig von den hinter liegenden oder von den davor liegenden Strukturen ab. In Bild (c) hingegen wäre der vergrößerte Lymphknoten auch ohne Hervorhebung eindeutig zu erkennen.

dieser Aufgabenstellung ausgehend wurden die Forschungs- und Nullhypothesen formuliert. Hieraus ließen sich die vier Faktorstufen des *vLK* als UV und die Detektionsgenauigkeit und Reaktionszeit als AV operationalisieren. Die optimale Stichprobengröße wurde mit 25 Vpn bestimmt. Ein weiterer wesentlicher Punkt in Vorbereitung auf die Durchführung des Experimentes ist der Aufbau der Versuchsanordnung selbst. Hier fließen vor allem die Erfahrungen aus der experimentellen Psychologie ein. Demnach müssen alle teilnehmenden Probanden das Experiment unter denselben Bedingungen durchlaufen. Das bedeutet, dass allen dieselbe Aufgabe zu stellen ist und die Stimuli in gleicher Art und Weise präsentiert werden müssen. Hierbei musste ein Kompromiss aus Bildern des klinischen Alltags und den zur Verfügung stehenden Testbildern geschlossen werden, so dass immer dieselben Bedingungen zur Bewältigung der Aufgabe gegeben sind. Eine einheitliche Instruktion wird durch die Formulierung einer schriftlichen Aufgabenstellung garantiert. Darüber hinaus wurde beschrieben, welche weiteren

Anforderungen an die Stimuli gestellt werden und wie die Konzentration der Teilnehmer kontrolliert werden kann. Unabhängig davon, und über die Aufgabenstellung hinausgehend, wurde ein Fragebogen entworfen, um die objektiv gemessenen Daten den Bewertungen der Probanden gegenüber zustellen. Anhand dieses Versuchsdesigns wird im Folgenden das Experiment implementiert und durchgeführt.

5 Umsetzung und Durchführung des Experimentes

In diesem Kapitel soll zunächst die für die Erstellung, Durchführung und Auswertung des in dieser Studie durchgeführten Experimentes benötigte Software vorgestellt werden. Weiterhin wird auf die Generierung der diesem Experiment zugrunde liegenden Stimuli, die Durchführung des Experimentes und die Aufbereitung der gewonnenen Messdaten eingegangen.

5.1 Verwendete Software

Die für das Experiment benötigten Stimuli wurden mit Hilfe des StimulusGenerator-Moduls in der Entwicklungsumgebung MeVisLab 1.6 erzeugt. Der Aufbau und die Durchführung des Versuchs wird mit der *Presentation 12.2* Software von *Neurobehavioral Systems* realisiert [Neurobs, 2009]. Die Datenerfassung erfolgte ebenfalls mit diesem Programm. Für die Weiterverarbeitung und Auswertung der erfassten Daten wurde die Statistik- und Analyse-Software *SPSS 15.0* [SPSS, 2009] verwendet. *SPSS* ist eine weltweit etablierte Statistik- und Analyse-Software. Aufgrund ihrer Zuverlässigkeit, dem großem Funktionsumfang sowie ihrem umfassenden Angebot an Analyseverfahren findet sie häufig Anwendung in der experimentellen Psychologie.

5.1.1 MeVisLab

MeVisLab wird von einer Forschungsgruppe von *Fraunhofer MEVIS* in Bremen entwickelt, welches ursprünglich als gemeinnütziges Forschungszentrum für medizinische Diagnostiksysteme und Visualisierung an der Universität Bremen gegründet wurde [MEVIS, 2009]. Hierbei handelt es sich um ein Entwicklungswerkzeug zur Bearbeitung und Visualisierung medizinischer Bilddaten. Dabei können bereits vorhandene Algorithmen und Datenstrukturen aus verschiedenen Bibliotheken genutzt sowie eigene Verfahren umgesetzt werden. Zur Beschreibung von 3D-Szenen wird auf die objektorientierte Grafikbibliothek *OpenInventor* zurückgegriffen. Die Beschreibung dieser Szenen erfolgt mittels Szenengraphen, wobei deren Darstellung über *OpenGl* realisiert

wird. Um auf bereits vorhandene Methoden zurückgreifen zu können, dienen Module als deren Schnittstellen. Für eigen-realisierte Verfahren muss dementsprechend eine Schnittstelle in Form eines Moduls zur Verfügung gestellt werden. Die Entwicklung solcher Module erfolgt in *C++*.

In dieser Entwicklungsumgebung werden die für das Experiment benötigten Stimuli aus Halsdatensätzen erzeugt.

5.1.2 Presentation

Presentation ist ein von Neurowissenschaftlern entwickeltes Programm zur Durchführung von Experimenten, in denen Reaktionen von Vpn auf verschiedene Reizdarbietungen gemessen werden. Mit Hilfe dieses Programmes können zum einen Stimuli erzeugt, vorhandene Stimuli benutzt und deren Präsentationart bestimmt werden. Die dieser Arbeit zugrunde liegenden Stimuli stellen komplexe, anatomische Halsvisualisierungen dar und können nicht mit dem in Presentation integrierten Programm zur Stimuligenerierung erzeugt werden. Hierfür wird die Entwicklungsumgebung MeVisLab benutzt. Zum anderen können die Reaktionen der Vpn auf die gezeigten Stimuli mit Hilfe von Eingabe- oder externen Geräten (fMRT, MEG, Eye-Tracker) aufgenommen und weiterverarbeitet werden.

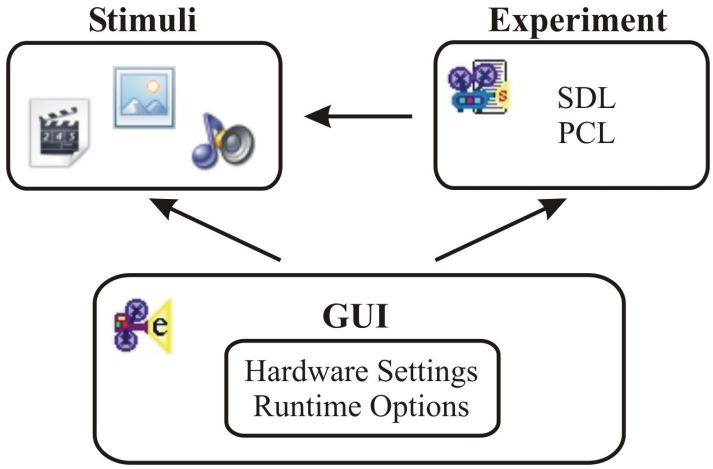


Abbildung 5.1: Aufgabenbereiche in *Presentation*. Zum einen können Stimuli erzeugt und importiert werden. Zum anderen können die Szenen- und Kontroll-Dateien direkt in dem in *Presentation* integrierten Editor implementiert werden. Einstellungen und Durchführung des Experimentes werden mit Hilfe der Benutzeroberfläche getätigt.

Mit Hilfe von *Presentation* können Vpn visuelle Stimuli in *JPG*-, *BMP*-, *AVI*-Formaten und auditive Stimuli im *WAVE*-Format einzeln, aber auch in kombinierter Form dargeboten werden. Zu jedem Stimulus können eine oder mehrere Reaktionen gemessen werden. Ein in *Presentation* durchgeführtes Experiment basiert auf Text-Dateien, in

denen die Abläufe des Experimentes in der *Scenario Description Language* (SDL) oder *Presentation Control Language* (PCL) beschrieben sind und vom Programm interpretiert werden. In *SDL* werden lediglich die zu verwendenden Stimuli als Objekte, deren Präsentationsform und -dauer sowie die mit ihnen assoziierten Eingaben definiert. Mit *PCL* hingegen können komplexere Darstellungsformen wie randomisierte, kombinierte oder auf bestimmte Reaktionen angepasste Stimuli bestimmt werden. Zudem können die in den *SDL*-Dateien definierten Objekte aufgerufen und manipuliert werden. *SDL*- und *PCL*-Dateien werden in **.sce* bzw. in **.pcl*-Formaten abgespeichert.

Mit Hilfe der *Presentation* Benutzeroberfläche werden die für ein Experiment benötigten Hardware-Einstellungen festgelegt. Dazu gehört die Zuweisung von Monitoren und der zu überwachenden Eingabe-Schnittstellen für Tastatur, Maus oder Eye-tracker und fMRT-Geräte. Abbildung 5.1 zeigt die wichtigsten Komponenten der *Presentation* Software.

Darüber hinaus erstellt *Presentation* nach jeder Durchführung eines Experimentes automatisch *Log*-Dateien, in denen sämtliche Ereignisse, die während des Experimentes aufgetreten sind, aufgelistet sind. Mit Hilfe von Analyse-Dateien (*.sdf) können benutzerdefinierte Ausgabe-Dateien aus diesen *Log*-Dateien erstellt werden, die anschließend für weitere Analysen in *Microsoft Excel* oder *SPSS* importiert werden können.

5.2 Implementierung

Die Umsetzung des in Kapitel 4 vorgestellten Konzepts beinhaltet zunächst die Erzeugung von visuellen Stimuli, die den Vpn gezeigt werden sollen. Danach werden die einzelnen Implementierungsschritte des Versuchdesigns erläutert. Abschließend wird darauf eingegangen, wie nach dem Experiment die entsprechenden *Log*-Dateien für die Analyse in *Microsoft Excel* und *SPSS* aufbereitet werden.

5.2.1 Aufbereitung und Erzeugung der Stimuli

Die Bewertung der in Abschnitt 4.1 vorgestellten Hervorhebungstechniken soll am Beispiel von pathologischen Lymphknoten im Halsbereich des Menschen erfolgen. Hierfür werden Bilder erstellt, in denen die Vpn vergrößerte Lymphknoten, falls vorhanden, in repräsentativen Halsvisualisierungen detektieren sollen. Insgesamt stehen 16 verschiedene Halsdatensätze in MeVisLab zur Verfügung, aus denen die für das Experiment benötigten Stimuli generiert werden. Dabei wurden die in Abschnitt 4.5.2 genannten Anforderungen an die Stimuli berücksichtigt. Da nicht alle Halsdatensätze die gleiche Anzahl an Lymphknoten und dieselben Strukturen beinhalten, wurden einigen Visualisierungen Lymphknoten sowie fehlende Strukturen aus anderen Datensätzen hinzugefügt, um den Suchaufwand für jede Vpn in jedem Bild gleich zu halten.

Die Stimuli werden mit Hilfe von Screenshots der 3D-Halsrenderings aus coronaler Sicht mit orthogonaler Projektion erstellt. Hierfür wird das *Makro*-Modul StimulusGenerator benutzt, welches die Bilder in einer 512×512-Auflösung im *TIF*-Format abspeichert. Das Sichtfeld der Halsvisualisierungen in den Screenshots ist dabei von den beiden äußeren Rändern der Muskeln begrenzt. Als vergrößerte Lymphknoten werden diejenigen Lymphknoten angesehen, deren Durchmesser größer als 3 *cm* ist. Aufgrund der in den Bildern vorliegenden 2D-Ansicht der Halsvisualisierungen reicht dieses Größenkriterium jedoch nicht mehr aus, da für den Betrachter keine Tiefeninformationen vorliegen. Aus diesem Grund werden nur diejenigen Lymphknoten als vergrößert angesehen, deren sichtbare Pixelanzahl größer als 30 Pixel ist.

Mit dem Modul StimulusGenerator ist die Auswahl der darzustellenden Strukturen möglich. Zudem generiert es für jeden vergrößerten Lymphknoten ein Bild, indem dieser Lymphknoten jeweils in einer der vier Faktorstufen dargestellt ist. Die Namenskodierung der Bilder setzt sich dabei aus der Art des Stimulus, der Seite des vergrößerten Lymphknotens, der Hervorhebungsart, dem Namen des Datensatzes aus dem dieses Bild generiert wurde, der Lymphknoten-ID sowie der Pixelgröße des vergrößerten Lymphknotens zusammen. Somit ist sichergestellt, dass jede Aktion einer Vpn auf einen bestimmten Lymphknoten zurückgeführt werden kann. Zur Anonymisierung werden die Datensätze von 1 bis 16 nummeriert. Die Hervorhebungsarten sind mit 0 für *CutAway*, 1 für *rote Einfärbung*, 2 für *keine Hervorhebung* und 3 für *Stippling* kodiert. Im Folgenden ist jeweils eine Beispielkodierung eines *Target*- und *Noise*-Bildes dargestellt:

TL_0_Hals01_LK15_85 Namenskodierung für ein *Target*-Bild (T), indem ein vergrößerter Lymphknoten (85 Pixel) mit der ID 15 aus dem Halsdatensatz 1 (Hals01) mit der Technik 0 (*CutAway*) auf der linken Seite des Bildes (L) hervorgehoben ist.

NR_3_Hals04_LK11_12 Namenskodierung für ein *Noise*-Bild (N), indem ein kleiner Lymphknoten (12 Pixel) mit der ID 11 aus dem Halsdatensatz 4 (Hals04) mit *Stippling* (3) auf der rechten Seite des Bildes (R) hervorgehoben ist.

Insgesamt konnten 701 Bilder aus den Datensätzen gewonnen werden, wovon 352 Rauschbilder mit vergrößerten und 349 Rauschbilder ohne vergrößerten Lymphknoten sind. Viele der Bilder liegen dabei in gespiegelter Form vor, um die Anzahl der Stimuli zu erhöhen. Dadurch soll der Wiedererkennungseffekt der Bilder nach Möglichkeit gering gehalten werden. Vergrößerte Lymphknoten sind in gleicher Anzahl links und rechts innerhalb der Halsvisualisierung angeordnet, so dass sich die Vpn stets auf beide Seiten des Halses konzentrieren müssen. Die Rauschbilder mit Zielreiz werden im Folgenden als *Targets* und die Rauschbilder ohne Zielreiz als *Noise* bezeichnet, da dies den konventionellen Bezeichnungen der Stimuli entspricht.

5.2.2 Umsetzung des Versuchsaufbaus

Die *Target*- und *Noise*-Bilder werden den Vpn in zufälliger Reihenfolge präsentiert. Jedem Stimulus folgt ein Interstimulus-Intervall in Form eines Fokusbildes, welches wiederum von einem Stimulus gefolgt wird. Dieses Fokusbild besteht aus einem zentrierten schwarzen Fokussierungskreuz auf weißem Hintergrund. Das Experiment wird in acht gleich lange Blöcke unterteilt, wobei jeder Block mit einem Instruktionsbildschirm beginnt.

Sowohl das Fokusbild als auch der Instruktionstext wurden für jeden dieser acht Blöcke als Objekte in einer *Presentation* Szenen-Datei angelegt. Aufgrund der Vielzahl an Stimuli wurde ein Array angelegt, in dem eine bestimmte Anzahl an Bildern gespeichert werden kann. Darüber hinaus wurden in dieser Datei die in den Hardware-Optionen deklarierten Eingabetasten, linke und rechte Maustaste, mit Event-Codes belegt. Die linke Maustaste wurde mit der Ziffer 1 und die rechte Maustaste mit der Ziffer 2 kodiert.

```
active_buttons = 2;
button_codes=1,2;
target_button_codes=11,22;

stimulus_properties = sStimulusArtSeite, string,
                      nHT, number,
                      sDataset, string,
                      sLK, string,
                      nPixel, number;
event_code_delimiter = "_";
```

Listing 5.1: Kodierung der Maustasten sowie der Stimuli-Eigenschaften in der Szenen-Datei *scene.sce*.

Wurden entsprechend den Stimuli die richtigen Maustasten betätigt, werden diese mit 11 bzw. mit 22 gekennzeichnet und als *target_button_codes* bezeichnet. Anhand der Namenskodierung der Stimuli können den dargebotenen Bilder ebenfalls Event-Codes vergeben werden, so dass diese nach ihren Eigenschaften gruppiert und analysiert werden können. Listing 5.1 zeigt einen Ausschnitt aus der Szenen-Datei, in der den Maustasten und den Stimuli-Eigenschaften *stimulus_properties* Event-Codes vergeben werden.

Das Laden der Stimuli, die Bestimmung der zufälligen Präsentations-Reihenfolge und die Anzeigedauer der Stimuli und der Fokusbilder wurden in der *PCL*-Datei *scene.pcl* festgelegt und sind prinzipiell in Pseudocode 5.1 in den Zeilen 14 bis 30 dargestellt. Wird nach Anzeige des Instruktionsbildschirmes T eine Maustaste m betätigt, soll das Experiment mit der Anzeige des Fokusbildes F beginnend starten. Jedes Fokusbild wird zufällig zwischen 750 ms und 1250 ms gezeigt. Die Anzeigedauer der Stimuli beträgt jeweils eine Sekunde. Eine Aktion A kann zum einen das Drücken der linken Maustaste im Falle eines gesichteten Zielreizes sein, oder das Drücken der rechten Maustaste, wenn die Vpn keinen Zielreiz gefunden hat. Wird keine Maustaste in der genannten

Zeitspanne betätigt, wird dies mit 0 kodiert, das Betätigen der linken und rechten Maustaste entsprechend mit 1 und 2.

In jedem der acht Blöcke werden den Vpn 145 Stimuli und Fokusbilder präsentiert. Daraus ergibt sich ein Block von ca. 5 *min*, so dass das gesamte Experiment im Durchschnitt 40 *min* exklusive der Pausen dauern wird. Die Länge der Pausen kann von jeder Vpn selbst bestimmt werden. Alle Ereignisse während der acht Teilexperimente sowie die Zeitpunkte ihres Auftretens werden anschließend in jeweils einer *Log*-Datei gespeichert.

```
EXPERIMENT(Bilder B, Fokusbild F, Instruktionstext T)
 1  Stimuli S ← Liste der dargebotenen Stimuli
 2  Aktionen A ← Liste der gedrückten Maustasten
 3  Reaktionszeiten RT ← Liste der gemessenen Reaktionszeiten
 4  Stimulus st
 5  Maustaste m ← 0 \\ 0 ← keine, 1 ← linke und 2 ← rechte Maustaste
 6  count ← 1
 7
 8  \\ Startbildschirm mit Instruktionstext
 9  while m = 0 do
10      Zeige T
11  end
12
13  \\ Präsentation der Stimuli und Fokusbilder
14  if m ≠ 0 then
15      for i ← 0 to 2 · n
16          if i mod 2 = 0 then
17              Zeige F zwischen 750 und 1250ms
18              if i ≠ 0 then
19                  A[count] ← m
20                  RT[count] ← gemessene Zeit zwischen Stimuluspräsentation und Mausdruck
21                  increase(count)
22                  m ← 0
23              end
24          else
25              st ← random(B) \\ wählt ein zufälliges Bild aus B
26              Zeige st für 1s
27              S[count] ← st
28          end
29      end
30  end
```

Pseudocode 5.1: Pseudocode für die Durchführung des Experimentes mit n Stimuli aus einer Bildmenge B, welche in gleicher Anzahl *Target* und *Noise* Bilder enthält.

Der Pseudocode 5.1 zeigt den prinzipiellen Ablauf des Experimentes. Die Zeilen $1-11$ des Pseudocodes entsprechen dem in den Szenen-Dateien definierten Teil in *SDL*. Der ab Zeile 14 dargestellte Teil zur Randomisierung und Präsentationsweise der Stimuli wurde in *PCL* umgesetzt.

5.2.3 Aufbereitung der Messdaten

Für die Aufbereitung der gewonnenen Daten wurde ein *Set Definition File* (SDF) geschrieben, in dem Gruppierungs- und Berechnungsvorschriften definiert und auf die *Log*-Dateien angewendet werden. Die in diesen *Log*-Dateien aufgezeichneten Ereignisse des Experimentes werden so anhand ihrer Kodierungen in sogenannten *Event Sets* unterteilt. Die wichtigsten *Event Sets* sind die Stimuli- und Reaktionsereignisse. Diese werden wiederum in *Event Sub Sets* untergliedert. Für die Stimuli ergeben sich die *Noise-* und *Target-Sub Sets* und für die Reaktionen die *Sub Sets* der betätigten Maustasten (0, 1, 2). Abbildung 5.2 zeigt die einzelnen Verarbeitungsschritte, wie sie von dem in *Presentation* integrierten Analyse-Tool durchgeführt werden. Anhand der

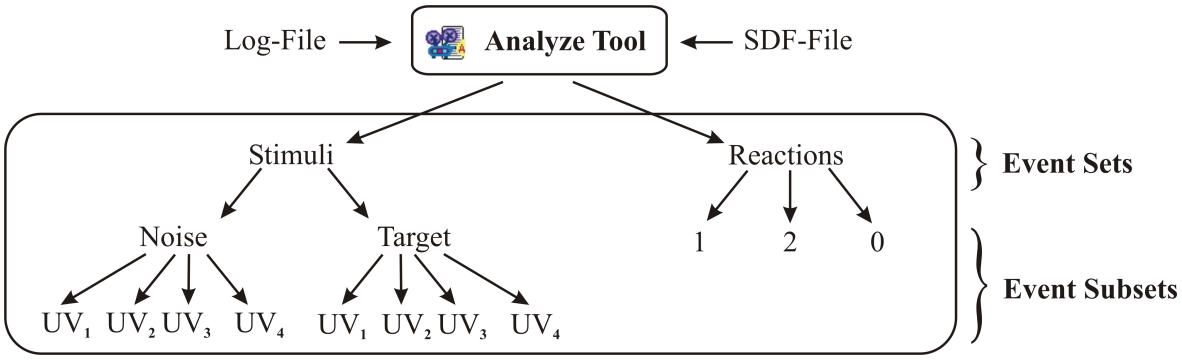


Abbildung 5.2: Schematische Darstellung der Datenaufbereitung in *Presentation* mit Hilfe des integrierten Analyse-Tools. In dieses werden die *Log*-Dateien und die Definitions-Datei (SDF) eingelesen. Anschließend werden die Daten in *Event Sets* und *Event Subsets* gruppiert.

Ereigniszeitpunkte können die *Sub Sets* der Stimuli und der Reaktionen miteinander in Verbindung gebracht und gepaart werden, so dass sich sogenannte *Event Pair Sets* ergeben. Zudem werden anhand der Ereigniszeitpunkte der Paarungen die Reaktionszeiten berechnet, welche sich aus deren Differenz ergeben. Darüber hinaus werden diese Paarungen nach den zu untersuchenden Faktorstufen gruppiert, so dass für diese die entsprechenden Antwortklassifikationen erstellt werden können. Tabelle 5.2 zeigt die entsprechenden Klassifikationen für die sechs Paarungsmöglichkeiten (Target-0, Target-1, Target-2, Noise-0, Noise-1, Noise-2) in einer Reiz-Antwort-Matrix nach Green und Swets [1966]. Trägt man die Anzahl der vorkommenden Klassifikationen in diese Matrix ein, kann eine Aussage darüber gemacht werden, wie häufig eine Vpn einen vergrößerten Lymphknoten detektiert oder übersehen sowie einen Distraktor fälschlicherweise als *vLK* angenommen oder einen Noise-Stimulus korrekt zurückgewiesen hat.

	Antwort	
	linke Maustaste (1) „vLK vorhanden“	**rechte/keine Maustaste (2/0) „vLK nicht detektiert“**
Target	Treffer	Verpasser
Noise	Falscher Alarm	Korrekte Zurückweisung

Tabelle 5.2: Antwortklassifikationen für die gegebenen Stimuli, wobei ein vergrößerter Lymphknoten *vLK* derjenige Zielreiz ist, der von der Vpn detektiert werden muss. Hat die Vpn keine Maustaste betätigt, wird davon ausgegangen, dass sie keinen Zielreiz wahrgenommen hat.

- Trefferrate $= \frac{Anzahl\ Treffer\ \cdot 100}{Anzahl\ Targetstimuli}$
- Falsche Alarmrate $= \frac{Anzahl\ Falscher\ Alarme\ \cdot 100}{Anzahl\ Noisestimuli}$

Diese Angaben können für jede der Faktorstufen-Paarungen gemacht werden. Die Ergebnisse der SDF für jede *Log*-Datei werden anschließend in eine *Text*-Datei geschrieben, so dass sich folglich acht **.txt*- Dateien ergeben.

Für jede Vpn wurde anschließend eine *Microsoft Excel* Datei erstellt. In dieser wurden die entsprechenden *Text*-Dateien eingelesen und tabellarisch aufgeführt. Hier werden die Ergebnisse der Trefferraten und Reaktionszeiten für alle Faktorstufen über alle acht Teilexperimente gemittelt und anhand des Mittelwertkriteriums auf Ausreißer hin überprüft. *Microsoft Excel* verfügt nicht über alle benötigten Verfahren zur statistischen Hypothesenprüfung. Die gemittelten Werte der Vpn werden daher in das Statistik-Programm *SPSS* exportiert.

Für die statistische Auswertung der gemittelten Messdaten braucht nur eine Datei in *SPSS* erstellt zu werden, in die alle Daten der Vpn eingetragen werden. Aufgrund des *within-subject*-Designs wird für jede AV der Faktorstufen eine eigene Variable angelegt. In Abbildung 5.3 sind die Variablen der Faktorstufen bezüglich der Trefferraten und Reaktionszeiten in den Zeilen 2 bis 9 sowie deren Eigenschaften und Skalenniveau zu sehen. In Abbildung 5.4 sind die entsprechenden gemittelten Werte der Vpn dargestellt. Weitere Variablen enthalten Angaben der Vpn, welche sie in den Fragebögen gegeben haben.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ID	Numeric	2	0	ID der Vp	None	None	3	Center	Nominal
2	hit_red	Numeric	8	2	Rote Einf	{,00, Kein	,00	8	Center	Scale
3	hit_cutaway	Numeric	8	2	CutAway	{,00, Kein	,00	8	Center	Scale
4	hit_stippling	Numeric	8	2	Stippling	{,00, Kein	,00	8	Center	Scale
5	hit_noHT	Numeric	8	2	Keine Her	{,00, Kein	,00	8	Center	Scale
6	rt_red	Numeric	8	2	Rote Einf	None	None	8	Center	Scale
7	rt_cutaway	Numeric	8	2	CutAway	None	None	8	Center	Scale
8	rt_stippling	Numeric	7	2	Stippling	None	None	8	Center	Scale
9	rt_noHT	Numeric	8	2	Keine Her	None	None	8	Center	Scale

Abbildung 5.3: Variablendeklaration in *SPSS*.

	ID	hit_red	hit_cutaway	hit_stippling	hit_noHT	rt_red	rt_cutaway	rt_stippling	rt_noHT
1	1	83,67	92,35	86,09	60,06	645,23	581,44	625,92	715,26
2	2	47,68	59,50	49,25	45,73	652,99	654,32	660,02	728,22
3	3	69,48	74,62	79,70	42,29	675,02	649,93	644,53	701,61
4	4	82,89	94,58	85,32	81,33	769,56	756,93	747,31	849,08
5	5	68,12	84,96	77,65	53,65	737,87	672,40	729,08	777,44
6	6	72,80	87,70	76,96	55,72	685,90	680,54	704,06	762,15
7	7	29,29	52,73	35,60	36,92	828,68	758,48	829,53	808,25
8	8	40,47	68,26	57,38	38,75	808,33	739,73	732,72	753,81

Abbildung 5.4: Eintragung der gemittelten Werte in *SPSS*.

Abbildung 5.5 zeigt schematisch den Ablauf der Datenaufbereitung. Zunächst wird das Experiment in *Presentation* aus den acht Szenen-Dateien geladen. Anschließend werden sämtliche Stimuli, Aktionen und Messdaten in *Log*-Dateien gespeichert. Mit Hilfe des Analyse-Tools wird eine SDF-Datei geladen und ausgeführt, welche diese Daten gruppiert und Berechnungen durchführt, die wiederum in Text-Dateien ausgegeben werden. Die Zusammenfassung der Daten aus allen acht Teilexperimenten erfolgt in *Microsoft Excel* und ihre Auswertung in *SPSS*.

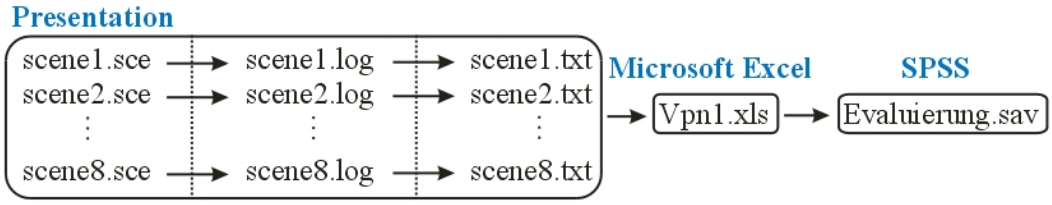


Abbildung 5.5: Pipeline der Datenerakquirierung- und aufbereitung.

5.3 Durchführung des Experimentes

Die visuelle Stimulation erfolgte auf einem PC (Dual-Core AMD Opteron mit $2,6\,GHz$, $2,75\,GB$ RAM und dem Betriebssystem XP Professional) unter der Verwendung von *Presentation* in einem Computer-Labor. Hierbei wurden den Vpn die Stimuli über einen 24“ Monitor in 512×512-Auflösung in acht gleich langen Blöcken und einem Interstimulus-Intervall von $750 - 1250\,ms$ (Fokusbild) dargeboten. Da noch keine vergleichbare Studie vorliegt, sollen die bisherigen Überlegungen im Rahmen eines kurzen Pilot-Experimentes auf ihre Praktikabilität hin überprüft werden.

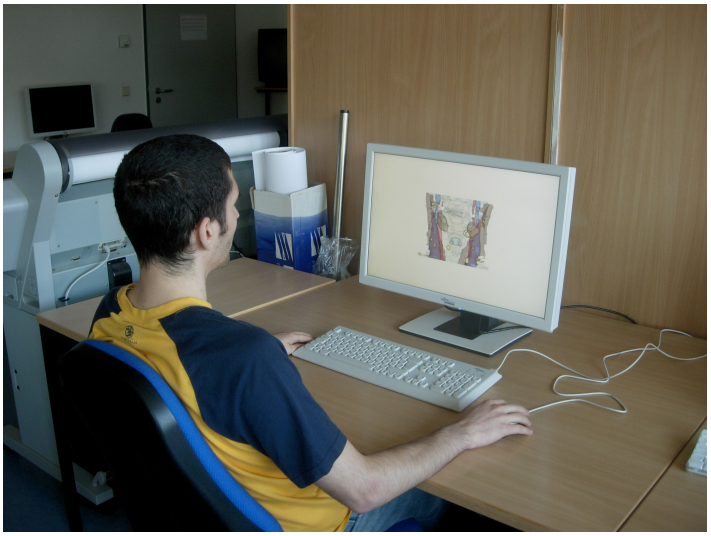


Abbildung 5.6: Versuchsperson während des Experimtes.

5.3.1 Pilot-Experiment

Damit die Reliabilität des Versuchplans überprüft und für das Experiment die richtigen Einstellungen bezüglich der Anzeigedauer der Stimuli sowie der Gesamtlänge des Experimentes gefunden werden können, wurde ein Vorexperiment mit sieben Versuchspersonen durchgeführt. Diesen wurde zunächst eine schriftliche Instruktion (Anhang C) zur Erläuterung der Aufgabenstellung gegeben. Das Vorexperiment selbst dauerte für jede dieser Vpn ca. 20 *min*. Während dieser Zeit wurden die Vpn allein in einem Raum gelassen, damit diese nach Möglichkeit nicht gestört werden. Im Anschluss wurde nach der subjektiven Meinung der Vpn bezüglich der ihnen präsentierten Stimuli, der Experimentalumgebung und -bedingungen gefragt. Die Anzeigedauer der Stimuli von 1 *s* wurde überwiegend als zu kurz empfunden. Da dennoch gute Detektionsergebnisse erzielt wurden, wird die Anzeigedauer nur geringfügig um 100 *ms* verlängert.

Anhand der subjektiven Meinungen sowie der explorativen Analyse der Messdaten konnte festgestellt werden, dass *CutAway* von den meisten Vpn bevorzugt wurde und dass es diejenige Technik war, mit der die meisten vergrößerten Lymphknoten wahrgenommen wurden. Darüber hinaus wurden mit *CutAway* die schnellsten Reaktionszeiten gemessen. Aus den in diesem Vorexperiment gewonnenen Erkenntnissen wurden daher zwei weitere Forschungshypothesen postuliert:

- ***Forschungshypothese* $\mathbf{H_2}$(Detektion)*:***

  Mit *CutAway* hervorgehobene vergrößerte Lymphknoten in farbigen 3D-Renderings von Halsdatensätzen werden häufiger vom Betrachter erfasst als vergrößerte Lymphknoten mit *roter Einfärbung* oder *Stippling*.

- ***Forschungshypothese* $\mathbf{H_2}$(Reaktionszeit)*:***

  Mit *CutAway* hervorgehobene vergrößerte Lymphknoten in farbigen 3D-Renderings von Halsdatensätzen werden schneller vom Betrachter erfasst als vergrößerte Lymphknoten, die mit *roter Einfärbung* oder *Stippling* hervorgehoben werden.

Bei einer Vpn wurde festgestellt, dass sie die Aufgabenstellung nur teilweise verstanden hatte. So betätigte sie nur die linke Maustaste, wenn sie einen Zielreiz wahrgenommen hatte. Nahm sie keinen vergrößerten Lymphknoten wahr, drückte sie keine Taste. Infolgedessen, wird im Hauptexperiment für jede Vpn zunächst ein kurzer Übungsdurchgang durchgeführt, so dass diese mit der Funktionsweise der Maus vertraut und Unsicherheiten hinsichtlich der Aufgabenstellung ausgeschlossen werden können.

5.3.2 Haupt-Experiment

Bei der Durchführung des eigentlichen Experimentes ist darauf zu achten, dass die Vpn des Vorexperimentes nicht an diesem Experiment teilnehmen, da diese mit den Stimuli bereits vertraut sind. Wie beim Pilotexperiment wurden die Vpn mit Hilfe einer schriftlichen Instruktion (C) in die Aufgabenstellung eingewiesen. Anschließend führten die Vpn einen kurzen zwei minütigen Übungsdurchgang durch, nach dem die Treffer der Vpn sofort eingesehen wurden. Im Fall einer sehr geringen Trefferrate konnte somit ein erneuter Übungsdurchlauf durchgeführt werden. Hat die Vpn die Aufgabenstellung sowie die Bedienung der Maustasten verstanden, wurde das Hauptexperiment durchgeführt.

Im Anschluss an das Experiment wurden den Vpn die Fragebögen ausgehändigt. Insgesamt dauerte ein Experiment ca. 67 *min* (10 *min* Einführung in die Aufgabenstellung, 40 *min* Experiment, ca. $7 \cdot 1$ *min* Pause und 10 *min* Beantwortung des Fragebogens). Jede Vpn musste zusätzlich eine Teilnehmerliste unterschreiben, um nachweisen zu können, dass 33 verschiedene Personen an der Studie teilgenommen haben.

6 Auswertung der Messdaten

An dem im Rahmen dieser Arbeit durchgeführten Experiment nahmen insgesamt 33 Personen teil. Aus dem *within-subject*-Design ergibt sich somit eine Stichprobengruppe von 33 Vpn für jede Faktorstufe UV_i, mit $i \in \{1, 2, 3, 4\}$. Die Abhängigkeit der Faktorstufen untereinander ist ebenfalls durch das Design gegeben. Vor der Hypothesenüberprüfung werden die abhängigen Variablen bezüglich der Trefferraten AV_{Acc} auf Ausreißer hin überprüft. Anschließend werden die AV_{Acc} sowie die abhängigen Variablen (AV) bezüglich der Reaktionszeiten AV_{RT} auf Normalverteilung und Varianzhomogenität hin untersucht. Darauf aufbauend werden geeignete Signifikanztests zur Hypothesenüberprüfung, welche in Abschnitt 3.4 vorgestellt wurden, angewendet. Die dazu angewendeten Analyseverfahren werden mit dem Statistikprogramm *SPSS* durchgeführt. Die Auswertung der Fragebögen (Anhang D) erfolgt ebenfalls mit diesem Programm. Da *SPSS* lediglich die Ergebnisse einer Datenanalyse ausgibt, wurden die Rechenwege der einzelnen durchgeführten Verfahren in *Microsoft Excel* nachvollzogen und in Tabellenform in Anhang B dargestellt.

6.1 Überprüfung der Daten

Im Folgenden sollen anhand des Mittelwertkriteriums diejenigen Messwerte bestimmt werden, die eine zu große Abweichung vom Gesamtmittelwert aufweisen. Des Weiteren werden signalentdeckungstheoretische Berechnungen bezüglich der Diskriminationsfähigkeit und Reaktionsneigung der Vpn durchgeführt.

6.1.1 Ausreißerbestimmung

Die Untersuchung der Messdaten auf ergebnisverzerrende Werte erfolgt anhand der Trefferraten und falschen Alarme. Dabei erfolgte zunächst ein Ausschluss von Ausreißern der Stichprobe, deren Trefferrate in einen Experimentdurchgang kleiner oder größer als die zweifache Standardabweichung vom Mittelwert war. Diese Methode wird häufig dann angewandt, wenn von Messfehlern innerhalb der Daten auszugehen ist. Dahingehend wurden die gemittelten Trefferraten der jeweils acht Versuchsdurchläufe einer Vpn auf Abweichungen untersucht. Dabei wurden bei vier Vpn (Vpn: 1, 3, 21 und 33) Trefferraten gemessen, die um mehr als die zweifache Standardabweichung

vom Gesamtmittelwert abwichen. Gründe für diese Abweichungen können zum einen Ermüdungserscheinungen und anfängliche Probleme mit der Aufgabenstellung sein. Bei der Vpn 1 kam es aus technischen Gründen zu fehlenden Messwerten, welche bei der Mittelwertbildung der Trefferrate der Vpn durch Ausschluss dieser Versuchsdurchläufe berücksichtigt wurden. Auch bei den anderen drei Versuchsteilnehmern wurden diejenigen Versuchsdurchläufe von der Mittelwertbildung über alle Versuchsdurchläufe ausgeschlossen, bei denen die Trefferrate kleiner als die zweifache Standardabweichung vom Gesamtmittelwert war. Dies entsprach jeweils einem ungültigen Versuchsdurchgang.

Bei der Überprüfung, ob alle Vpn eine hinreichende Diskriminationsfähigkeit d' aufweisen, wurde bei keinem Probanden ein negatives oder nahe bei Null liegendes Diskriminationsmaß festgestellt. Daraus lässt sich schließen, dass alle Versuchsteilnehmer die Aufgabenstellung verstanden haben und dass sie zwischen den Stimuli differenzieren konnten.

Des Weiteren sind die Reaktionsneigungen L_x der Vpn zu betrachten. Aus den Berechnungen von L_x ergeben sich ausschließlich Werte > 1. Hierbei wurde beobachtet, dass alle Vpn ein konservatives Antwortverhalten zeigten, was sich bei einigen Vpn negativ auf deren Trefferrate auswirkte. Dies könnte auf eventuelle Unsicherheit der Vpn infolge der kurzen Anzeigedauer der Stimuli oder die große Zahl der Kontextstrukturen, die die Suche erschweren, zurückzuführen sein.

Fünf Probanden zeigten eine sehr geringe Trefferrate ($< 50\%$) bei einer sehr geringen Falsch-Alarm-Rate ($< 3\%$). Ob dies allein auf die Unsicherheiten der Probanden zurückzuführen ist oder auch Desinteresse oder mangelnde Aufmerksamkeit eine Rolle spielen, kann nur vermutet werden. Nach Rücksprache mit einem in der experimentellen Psychologie erfahrenen Diplom-Psychologen wurden die Daten jener Vpn von der weiteren Analyse ausgeschlossen, deren Detektionsgenauigkeit unter 50% lag. Dies betraf die Vpn 7, 10, 28, 30 und 31. Im Wesentlichen haben diese Vpn nur sehr große und eindeutige vLK ($> \ 100$ *Pixel*) detektieren können. Trotz des Ausschlusses der fünf genannten Teilnehmer wird mit einer Stichprobengröße von 28 Personen die geforderte Mindeststichprobengröße von 25 Vpn nicht unterschritten.

6.1.2 Deskriptive Statistiken

Nach Ausschluss der Ausreißer können die wichtigsten Eigenschaften der abhängigen Variablen tabellarisch dargestellt werden. Sowohl die Gruppenmittelwerte μ_i und Mittelwerte der einzelnen Vpn innerhalb dieser Gruppen als auch die Standardabweichungen σ_i der Gruppen sind für die Signifikanztests ausschlaggebend. Die Maximalwerte der abhängigen Variablen bezüglich der Detektionsgenauigkeit zeigen, dass die Bewältigung der Aufgabenstellung möglich war. Die Minimalwerte hingegen lassen darauf schließen, dass einigen Teilnehmern dies jedoch Probleme bereitete. *CutAway* zeigt im

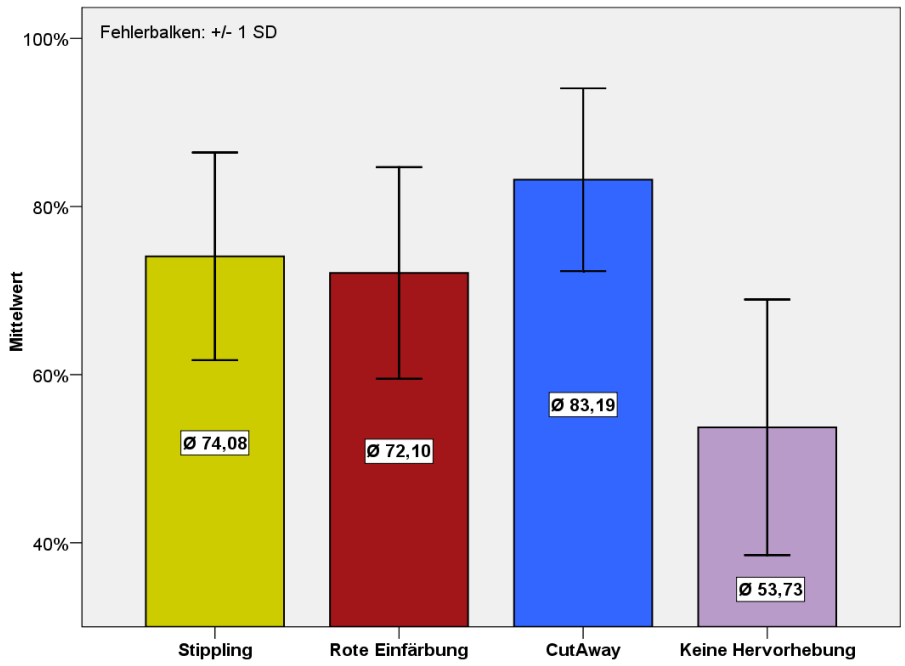
	Faktorstufe	n	μ	σ	Minimum	Maximum
$\mathbf{AV_{Acc}}$	Stippling	28	74,08 %	12,36 %	48,34 %	91,86 %
	Rote Einfärbung	28	72,10 %	12,58 %	40,47 %	84,76 %
	CutAway	28	83,19 %	10,88 %	59,50 %	96,06 %
	Keine Hervorhebung	28	53,73 %	15,21 %	14,51 %	81,33 %
$\mathbf{AV_{RT}}$	Stippling	28	726,94 ms	75,48 ms	583,71 ms	890,83 ms
	Rote Einfärbung	28	725,43 ms	72,84 ms	614,20 ms	905,90 ms
	CutAway	28	705,12 ms	77,94 ms	573,68 ms	907,89 ms
	Keine Hervorhebung	28	800,1 ms	91,33 ms	677,31 ms	1011,91 ms

Tabelle 6.1: Deskriptive Statistiken der Faktorstufen bezüglich der abhängigen Variablen der Trefferraten AV_{Acc} und Reaktionszeiten AV_{RT}. Mit μ und σ sind die jeweiligen Mittelwerte und Standardabweichungen der Faktorstufen dargestellt. Die Stichprobengröße für jede UV_i ist mit n gegeben. Weiterhin werden die minimalen und maximalen Extremwerte aufgezeigt.

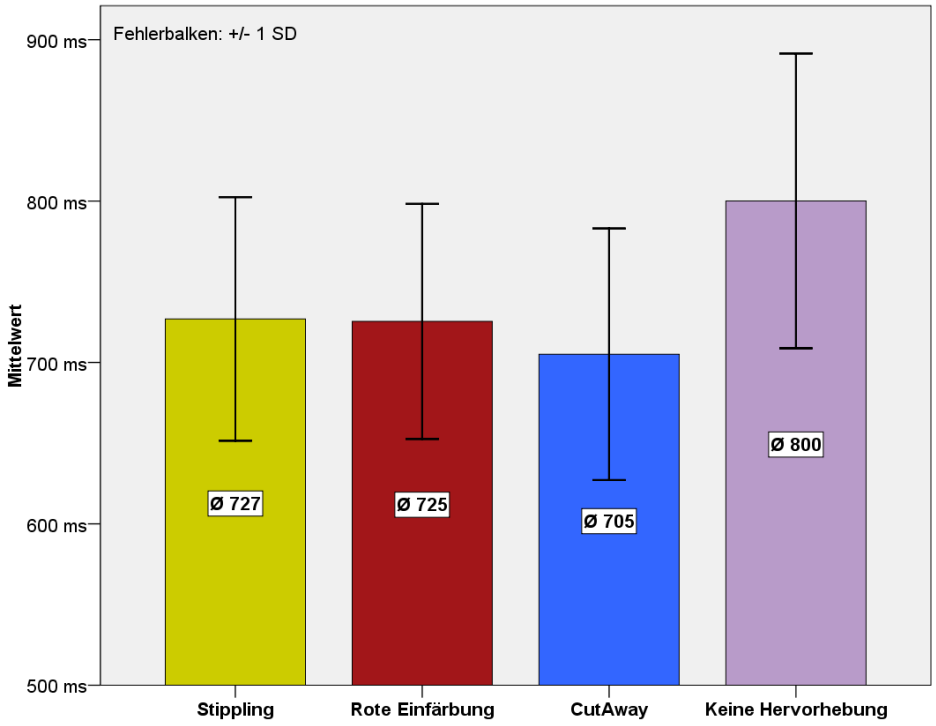
Vergleich zu den anderen Hervorhebunsgtechniken einen sehr hohen Minimalwert von 59,5%, wogegen eine Vpn vergrößerte Lymphknoten ohne Hervorhebung mit 14,51% kaum wahrgenommen hat. Lediglich eine Vpn erzielte mit *Stippling* die höchste Treffergenauigkeit, alle anderen Vpn hingegen mit *CutAway*.

Abbildung 6.1 zeigt die grafische Darstellung der in Tabelle 6.1 aufgezeigten deskriptiven Statistiken mit den zugehörigen Standardabweichungen. Wie bereits aus dieser Tabelle ersichtlich ist, wurden mit *CutAway* die meisten vergrößerten Lymphknoten in den Stimuli gefunden. Außerdem wurden diese mit dieser Technik auch am schnellsten detektiert, siehe Bild (b). *Rote Einfärbung* und *Stippling* unterscheiden sich nur sehr wenig sowohl in Bezug auf die Trefferquoten als auch hinsichtlich der Reaktionszeiten. Die Faktorstufe *keine Hervorhebung* schneidet in beiden Fällen am schlechtesten ab.

Weiterhin wurden die falschen Alarme (FA) der Vpn betrachtet. Aus Tabelle 6.2 wird ersichtlich, dass die Falsch-Alarm-Raten der Faktorstufen insgesamt sehr klein ($< 10\,\%$) sind und sich nur geringfügig unterscheiden. Dennoch wurden mit *CutAway* oder nicht hervorgehobene Distraktoren die höchsten Extremwerte von $17,64$ und $26,78$ erzielt. Mit hervorgehobenen Lymphknoten waren den Vpn immer potentielle vergrößerte Lymphknoten gegeben die mit den umliegenden Lymphknoten hinsichtlich ihrer Größe verglichen werden konnten. Diese Möglichkeit bestand nicht, wenn kein Lymphknoten hervorgehoben war, was ein Grund für deren hohen Extremwert sein könnte. Mit einem Kreis hervorgehobene Lymphknoten könnten nicht vergrößerte Lymphknoten größer erscheinen lassen, da sie über die Strukturgrenzen der Lymphknoten hinausgehen und somit vereinzelt hohe Extremwerte hervorrufen.



(a) Durchschnittliche Trefferquoten die mit Hilfe der Faktorstufen erzielt wurden.



(b) Durchschnittliche Reaktionszeiten die mit Hilfe der Faktorstufen erreicht wurden.

Abbildung 6.1: Deskriptive grafische Darstellung der Trefferquoten und Reaktionszeiten.

	Faktorstufe	n	μ	σ	Minimum	Maximum
FA	Stippling	28	3,28 %	2,51 %	0,00 %	8,29 %
	Rote Einfärbung	28	2,64 %	2,08 %	0,00 %	8,53 %
	CutAway	28	4,17 %	3,43 %	1,25 %	17,64 %
	Keine Hervorhebung	28	3,89 %	6,24 %	0,00 %	26,78 %

Tabelle 6.2: Falsche Alarme hervorgerufen durch die verschieden hervorgehobenen kleinen und normal großen Lymphknoten (Distraktoren).

6.1.3 Überprüfung auf Normalverteilung

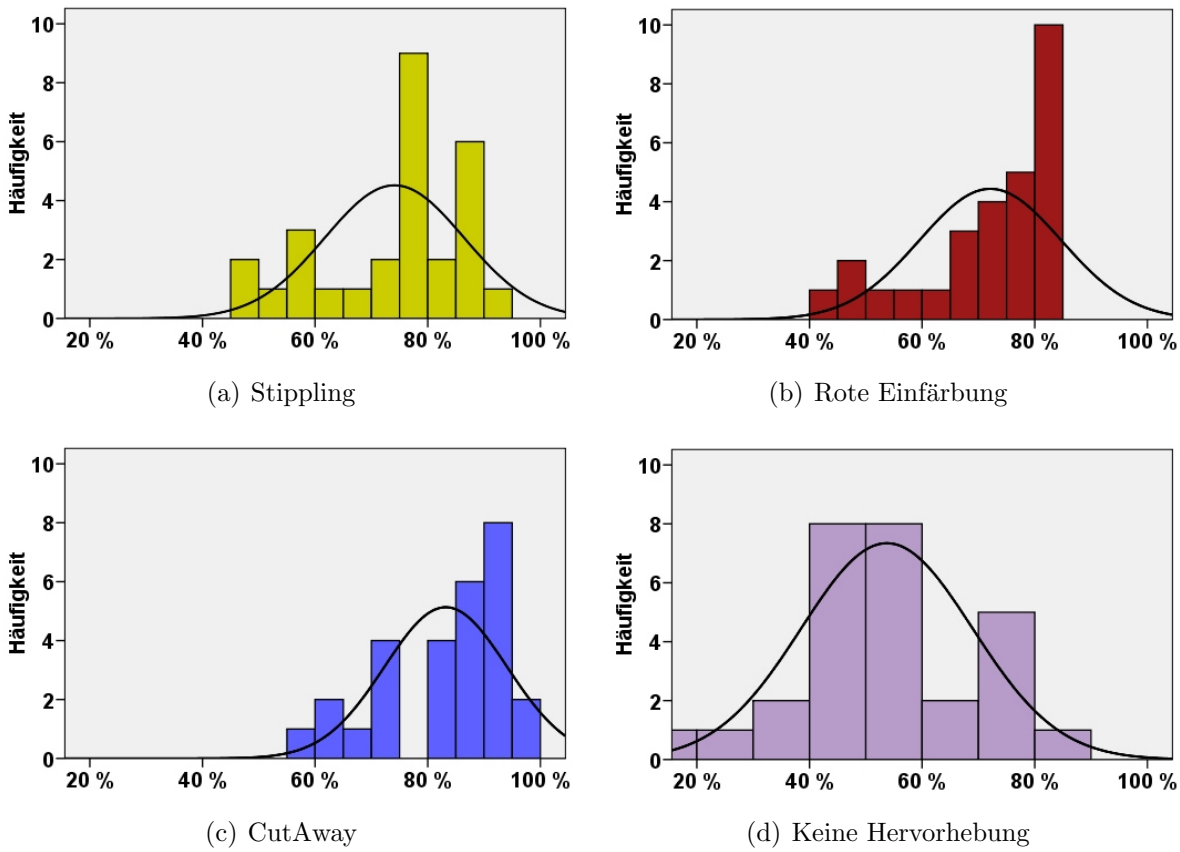
Die Überprüfung auf Normalverteilung erfolgt anhand der Häufigkeiten der Messdaten innerhalb der Faktorstufen in *SPSS* mit dem *Shapiro-Wilk*-Test auf dem Signifikanz-Niveau $\alpha = 0,05$. Ist die Wahrscheinlichkeit, dass die Annahme auf Normalverteilung zutrifft, größer als die Signifikanzschwelle $\alpha = 0,05$, kann eine Normalverteilung angenommen werden. Tabelle 6.3 zeigt die in *SPSS* berechneten Wahrscheinlichkeiten, ob eine Normalverteilung zutrifft oder nicht.

Die Abbildungen 6.2 und 6.3 zeigen grafische Histogrammdarstellungen der Häufigkeitsverteilungen der AV_{Acc} und AV_{RT} einer jeden Faktorstufe. Die in den Abbildungen dargestellten Normalverteilungskurven entsprechen einer theoretischen Normalverteilung mit den entsprechenden Mittelwerten und Standardabweichungen der Faktorstufen. Weichen die tatsächlichen Verteilungen deutlich von den Normalverteilungskurven ab, kann keine Normalverteilung der Messdaten angenommen werden. Dies wird in den Histogrammen (a), (b) und (c) der Abbildung 6.2 deutlich. Die Messdaten der Faktor-

	Faktorstufe	df	p	Normalverteilung
$\mathbf{AV_{Acc}}$	Stippling	28	0,006	nein
	Rote Einfärbung	28	0,001	nein
	CutAway	28	0,006	nein
	Keine Hervorhebung	28	0,655	ja
$\mathbf{AV_{RT}}$	Stippling	28	0,911	ja
	Rote Einfärbung	28	0,323	ja
	CutAway	28	0,621	ja
	Keine Hervorhebung	28	0,046	ja

Tabelle 6.3: Überprüfung auf Normalverteilung mit dem *Shapiro-Wilk*-Test in *SPSS* mit den Freiheitsgraden $df = n$. Die Wahrscheinlichkeit, ob den Messdaten eine Normalverteilung zu Grunde liegt, ist durch p gegeben.

stufen *Stippling, rote Einfärbung* und *CutAway* unterscheiden sich signifikant von einer Normalverteilung bezüglich der AV_{Acc}. Aus diesem Grund sind die diesbezüglichen Hypothesen $H_1(Detektion)$ und $H_2(Detektion)$ mit parameterfreien Verfahren zu überprüfen. Hingegen sind dieselben Faktorstufen bezüglich der AV_{RT} mit $p > 0,05$ normal-



(a) Stippling

(b) Rote Einfärbung

(c) CutAway

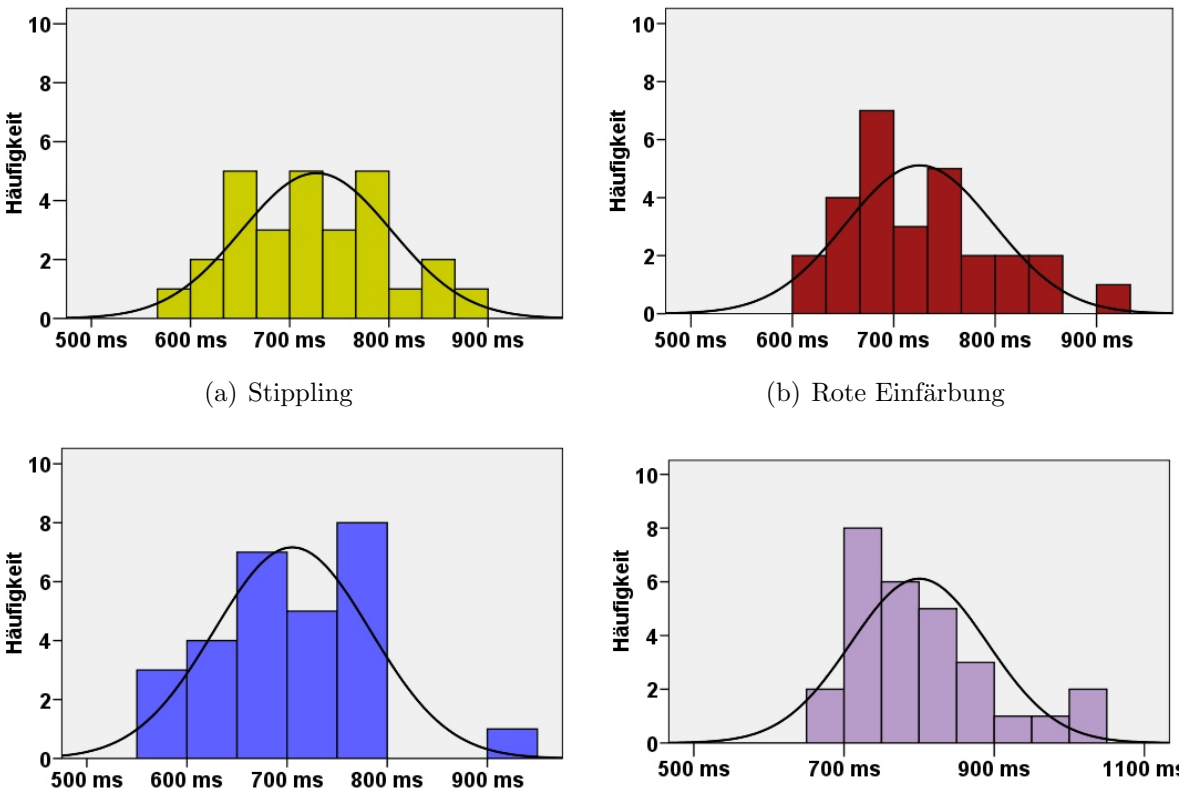
(d) Keine Hervorhebung

Abbildung 6.2: Verteilungen der AV_{Acc} mit Normalverteilungskurve. In den Bildern (a), (b) und (c) ist deutlich zu erkennen, dass sie nicht normalverteilt sind. Nur den vergrößerten Lymphknoten ohne Hervorhebung (d) liegt eine Normalverteilung zu Grunde.

verteilt. Die Faktorstufe *keine Hervorhebung* unterliegt mit $p = 0,046$ keiner Normalverteilung hinsichtlich der gemessenen Reaktionszeiten. Da deren gemessene Irrtumswahrscheinlichkeit p dem Grenzwert α jedoch sehr nah ist und die restlichen Faktorstufen einer Normalverteilung unterliegen, wurden die Hypothesen $H_1(Reaktionszeit)$ und $H_2(Reaktionszeit)$ dennoch mit Hilfe parametrischer Signifikanztests geprüft. Diese sind gegenüber der Verletzung der Normalverteilung relativ robust [Bortz, 2005, S.145, S.328].

6.1.4 Überprüfung auf Varianzhomogenität

Die Varianzhomogenität zwischen den Faktorstufen wird anhand des *Levene*-Tests ebenfalls in *SPSS* überprüft. Die dem Test zu Grunde liegende Nullhypothese geht davon aus, dass die Standardabweichungen der Faktorstufen sich nicht unterscheiden. Mit $p > 0,05$ für AV_{Acc} und AV_{RT} kann diese Nullhypothese nicht abgelehnt wer-



(a) Stippling

(b) Rote Einfärbung

(c) CutAway

(d) Keine Hervorhebung

Abbildung 6.3: Verteilungen der AV_{RT} der Reaktionszeiten mit Normalverteilungskurve.

den. Das bedeutet, dass sich die Standardabweichungen σ_i der Faktorstufen UV_i, mit $i \in \{1, 2, 3, 4\}$ nicht unterscheiden und somit Varianzhomogenität vorliegt. In Tabelle 6.4 sind die Ergebnisse für die abhängigen Variablen dargestellt.

	$\mathbf{df_1}$	$\mathbf{df_2}$	**p**	**Varianzhomogenität**
$\mathbf{AV_{Acc}}$	3	108	0,588	ja
$\mathbf{AV_{RT}}$	3	108	0,699	ja

Tabelle 6.4: Überprüfung auf Varainzhomogenität mit dem *Levene*-Test in SPSS mit den Freiheitsgraden $df_1 = k - 1$ und $df_2 = k \cdot n - k$, wobei k für die Anzahl der Faktorstufen steht. Die Wahrscheinlichkeit, ob Varianzhomogenität zwischen den Faktorstufen vorliegt, ist durch p gegeben.

6.2 Hypothesenprüfung

In diesem Abschnitt werden die in den Abschnitten 4.2.1 und 5.3.1 postulierten Forschungshypothesen mit ausgewählten statistischen Verfahren überprüft. Die Annahmen über Detektionsgenauigkeit und Reaktionszeiten wurden in verschieden Hypothesen postuliert und sind daher jeweils univariat bzw. getrennt zu untersuchen.

6.2.1 Signifikanzüberprüfung zwischen allen Faktorstufen

Anhand der Mittelwerte μ_i der abhängigen Variablen AV_{Acc} und AV_{RT}, mit $i \in \{1, 2, 3, 4\}$ soll überprüft werden, ob mit Hilfe der Hervorhebungstechniken vergrößerte Lymphknoten tatsächlich häufiger und schneller wahrgenommen werden als nicht hervorgehobene vergrößerte Lymphknoten. Die Auswahl des Signifikanztests erfolgt dabei unter Berücksichtigung der in Abschnitt 6.1.3 und 6.1.4 dargestellten Eigenschaften der AV_{Acc} und AV_{RT}.

Detektionsgenauigkeit

Zunächst werden die Techniken anhand der mit ihnen erzielten Trefferquote daraufhin untersucht, ob sich eine der Techniken hinsichtlich ihres Mittelwertes signifikant von den anderen unterscheidet. Da bei drei von vier Faktorstufen den Messwerten für die AV_{Acc} keine Normalverteilung zugrunde liegt, wird hierfür ein nicht-parametrischer Signifikanztest angewendet. Aufgrund der Abhängigkeit der Faktorstufen untereinander ist der *Friedman*- Test geeignet.

In *SPSS* kann dieser Test entweder über die Benutzeroberfläche oder über den *Syntax Editor* mit den in Listing 6.1 dargestellten Befehlen aufgerufen werden.

```
NPAR TESTS
  /FRIEDMAN = hit_red hit_cutaway hit_stippling hit_noHT.
```

Listing 6.1: *SPSS*-Syntax für den *Friedman*-Test auf den Trefferraten der Hervorhebungstechniken.

Aus Tabelle 6.5 ergeben sich die für die Berechnung der Prüfgröße benötigten Rangsummen der Faktorstufen bezüglich der Trefferraten AV_{Acc}. Anhand dieser Rangsummen wird die Prüfgröße mit Hilfe der Gleichung 3.25 wie folgt berechnet:

$$\chi^2 \;=\; \frac{12}{28 \cdot 4 \cdot 5} \cdot (75^2 + 66^2 + 111^2 + 28^2) - 3 \cdot 28 \cdot 5 \;=\; 74,7 \tag{6.1}$$

Die in 6.1 ermittelte Prüfgröße χ^2 ist größer als der kritische Wert $\chi^2_{0,05;3} = 7,81$ der

	Faktorstufe	Rangsumme T	Mittlerer Rang
$\mathbf{AV_{Acc}}$	Stippling	75	2,68
	Rote Einfärbung	66	2,36
	CutAway	111	3,96
	Keine Hervorhebung	28	1,0

Tabelle 6.5: *Friedman*-Test auf die AV_{Acc}. Die ermittelten Rangsummen T gehen in Berechnung der Prüfgröße χ^2 ein. Anhand der mittleren Ränge kann zudem eine Rangfolge der Faktorstufen hinsichtlich der erzielten Trefferquoten abgelesen werden.

Prüfverteilung, womit sich ein hoch signifikanter Unterschied mit $p < 0,001$ zwischen den Mittelwerten der Hervorhebungstechniken ergibt. Daraus kann geschlossen werden, dass sich mindestens eine der vier Faktorstufen signifikant von den anderen unterscheidet. Daher ist die Nullhypothese $H_0(Detektion)$ abzulehnen. Daraus kann jedoch nicht die Annahme der Forschungshypothese gefolgert werden, da dieser mehrere Bedingungen zugrunde liegen. Dennoch lässt sich aus den mittleren Rängen der Faktorstufen ablesen, mit welcher Technik die höchste Trefferquote erzielt wurde. Demnach wurden mit *CutAway* mehr vergrößerte Lymphknoten detektiert als mit den anderen Faktorstufen, da sie den höchsten Rang aufweist.

Mit Hilfe von Paarungstests wird überprüft, welche und wie viele der Hervorhebungstechniken sich signifikant von der Faktorstufe *ohne Hervorhebung* unterscheiden, und ob *CutAway* sich als die geeignetste Hervorhebungstechnik erweist.

Reaktionszeit

In der Forschungshypothese $H_1(Reaktionszeit)$ wird die Behauptung aufgestellt, dass mit Hilfe der Hervorhebungstechniken die vergrößerten Lymphknoten schneller wahrgenommen werden als ohne Hervorhebung. Diese Annahme soll anhand der Reaktionszeiten AV_{RT}, die für die einzelnen Techniken gemessen wurden, überprüft werden. Aus Abschnitt 6.1 geht hervor, dass für alle Faktorstufen eine Normalverteilung der Messergebnisse vorliegt. Aufgrund der Abhängigkeit der Faktorstufen ist ein *ANOVA-Test mit Messwiederholung* geeignet, welcher in *SPSS* durchgeführt werden soll. Dieser lässt sich ebenfalls über die Benutzeroberfläche oder anhand der in Listing 6.2 aufgezeigten Syntax ausführen.

```
GLM
  rt_red rt_cutaway rt_stippling rt_noHT
  /WSFACTOR = vLK 4
  /CRITERIA = ALPHA(.05).
```

Listing 6.2: *SPSS*-Syntax für den *ANOVA-Test mit Messwiederholung* auf den gemessenen Reaktionszeiten der Hervorhebungstechniken.

Der *ANOVA-Test mit Messwiederholung* wird dabei von dem Befehl */WSFACTOR* (within-subject-factor) mit dem Signifikanz-Niveau $\alpha = 0,05$, dem Faktor UV=vLK und der Anzahl an Faktorstufen = 4 aufgerufen. Die Prüfgröße F berechnet sich nach Gleichung 3.23 wie folgt:

$$F = \frac{48624,64}{850,82} = 57,15 \tag{6.2}$$

In Tabelle 6.6 sind die Ergebnisse des *SPSS* Programms der Tabelle 3.9 entsprechend zusammengefasst dargestellt. Die Prüfgröße F ist mit $57,15$ um ein vielfaches größer als

	SAQ	**df**	**MQ**	**F**	$\mathbf{F_{krit} = F_{0,05;3;81}}$	**p**
$\mathbf{SAQ_{Zw}}$	145873,93	3	48624,64	57,15	~2,72	<0,001
$\mathbf{SAQ_{e}}$	617422,57	27	22867,50			
$\mathbf{SAQ_{Res}}$	68916,55	81	850,82			
$\mathbf{SAQ_{Ges}}$	832213,06	108				

Tabelle 6.6: *ANOVA-Test mit Messwiederholung* auf die AV_{RT}.

der kritische Wert $F_{0,05;3;81} \cong 2,72$ der Prüfverteilung, siehe Anhang A. Es wurde somit ein hoch signifikanter Unterschied mit $p < 0,001$ zwischen den Mittelwerten der Reaktionszeiten der AV_{RT} nachgewiesen. Folglich ist die Nullhypothese $H_0(Reaktionszeit)$ ungültig. Ob die Forschungshypothese $H_1(Reaktionszeit)$ zutrifft, muss auch hier erst mit Hilfe von Paarungstests untersucht werden.

Im Unterschied zum *Friedman*-Test kann aus dem *ANOVA-Test mit Messwiederholung* keine Rangfolge der Hervorhebungstechniken hinsichtlich der Reaktionszeiten abgeleitet werden.

6.2.2 Signifikanzüberprüfung zwischen jeweils zwei Faktorstufen

Die im Folgenden verwendeten Faktorstufen UV_i mit $i \in \{1, 2, 3, 4\}$ stehen repräsentativ für die vier Faktorstufen der vergrößerten Lymphknoten.

- UV_1:= Stippling
- UV_2:= Rote Einfärbung
- UV_3:= CutAway
- UV_4:= Keine Hervorhebung

SPSS führt sämtliche Paarungstest zweiseitig durch. Die aufgestellten Hypothesen sind jedoch gerichtet, so dass die von *SPSS* ausgegebene Irrtumswahrscheinlichkeit p für einen einseitigen Test halbiert und entsprechend mit α verglichen werden muss [Duller, 2007]. Weiterhin ist der kritische Wert der Prüfverteilung des zweiseitigen Tests mit $\frac{\alpha}{2}$ durch α zu ersetzen.

Detektionsgenauigkeit

In Abschnitt 6.2.1 wurde bereits dargestellt, dass ein signifikanter Unterschied zwischen den Mittelwerten der Faktorstufen besteht. Im Folgenden soll zum einen herausgefunden werden, ob ein negativer oder positiver Unterschied vorliegt und zum anderen, welche der Faktorstufen die höchste Detektionsleistung aufweist. Aufgrund der Abhängigkeit und Nicht-Normalverteilung der Faktorstufen wird der in Abschnitt 3.4.2 beschriebende *Wilcoxon*-Test angewendet. Dieser wird mit Hilfe der in Listing 6.3 aufgezeigten *SPSS*-Syntax-Befehle aufgerufen. Möglich ist auch ein Aufrufen dieses Tests über die Benutzeroberfläche. Hierbei ist zu beachten, dass in der weitergehenden Menü-Führung der *Wilcoxon*-Test auszuwählen ist.

```
NPAR TEST
  /WILCOXON=hit_stippling hit_cutaway hit_red hit_cutaway
      hit_red hit_red
   WITH hit_noHT hit_noHT hit_noHT hit_stippling
       hit_stippling hit_cutaway
 (PAIRED).
```

Listing 6.3: *SPSS*-Syntax für den *Wilcoxon*-Test auf den Trefferraten aller Faktorstufen-Paarungen.

Aus den vier Faktorstufen ergeben sich insgesamt sechs Paarungstests, von denen die drei Paarungen $UV_1 - UV_4, UV_2 - UV_4$ und $UV_3 - UV_4$ relevant für die Forschungshypothesen $H_1(Detektion)$ und $H_1(Reaktionszeit)$ sind. Die beiden Paarungen $UV_1 - UV_3$ und $UV_2 - UV_3$ sind für die Forschungshypothesen $H_2(Detektion)$ und $H_2(Reaktionszeit)$ von Bedeutung. In Tabelle 6.7 sind zu diesen Paarungen die entsprechenden Prüfparamter sowie die von *SPSS* berechneten Prüfgrößen z dargestellt. Da für keine der sechs Paarungen geteilte Ränge ($Rang_T = 0$) vorkommen, lassen sich

Paarung	$\mathbf{Rang_-}$	$\mathbf{Rang_+}$	$\mathbf{Rang_T}$	$\mathbf{W_-}$	$\mathbf{W_+}$	z	p
$\mathbf{UV_1 - UV_2}$	10	18	0	**133**	273	1,594	0,055
$\mathbf{UV_1 - UV_3}$	27	1	0	401	**5**	4,509	0,000
$\mathbf{UV_1 - UV_4}$	0	28	0	**0**	406	4,623	0,000
$\mathbf{UV_2 - UV_3}$	28	0	0	406	**0**	4,623	0,000
$\mathbf{UV_2 - UV_4}$	0	28	0	**0**	406	4,623	0,000
$\mathbf{UV_3 - UV_4}$	0	28	0	**0**	406	4,623	0,000

Tabelle 6.7: *Wilcoxon*-Tests auf die AV_{Acc}. Den Werten der positiven und negativen Ränge $Rang_{-,+}$ entspricht der Anzahl der positiven und negativen Differenzen zwischen den Wertepaarungen. $Rang_T$ gibt die Anzahl der geteilten Rangplätze und p die Irrtumswahrscheinlichkeit für einen einseitigen Test an. W_- und W_+ stellen die Rangsummen dar.

die Prüfgrößen entsprechend der Gleichung 3.16 berechnen. Dabei geht jeweils die kleinere Rangsumme W_- oder W_+ der Paarungen in die Formel ein. Da die Rangsummen

für vier Paarungen gleich sind, ergeben sich nur drei verschiedene Prüfgrößen z_A, z_B und z_C.

$$z_A = \frac{\frac{28 \cdot 29}{4} - 133}{\sqrt{\frac{28 \cdot 29 \cdot 57}{24}}} = 1,594 \tag{6.3}$$

$$z_B = \frac{\frac{28 \cdot 29}{4} - 5}{\sqrt{\frac{28 \cdot 29 \cdot 57}{24}}} = 4,509 \tag{6.4}$$

$$z_C = \frac{\frac{28 \cdot 29}{4} - 0}{\sqrt{\frac{28 \cdot 29 \cdot 57}{24}}} = 4,623 \tag{6.5}$$

Aus $z_{krit} = 1,64$ für den einseitigen Test ergibt sich, dass sich die Detektionsgenauigkeit vergrößerter Lymphknoten bei allen drei Hervorhebungstechniken UV_1, UV_2 und UV_3 von der Detektionsgenauigkeit vergrößerter Lymphknoten ohne Hervorhebung UV_4 unterscheidet. Anhand der p-Werte lässt sich die Signifikanzstärke ablesen. Mit $p < 0,001$ sind hier hoch signifikante Unterschiede zu verzeichnen. Die $H_1(Detektion)$ kann demnach angenommen werden. Zudem unterscheidet sich die Hervorhebungstechnik *Cut-Away* (UV_3) ebenfalls hoch signifikant von den anderen beiden Hervorhebungstechniken UV_1 und UV_2. Daher kann die Forschungshypothese $H_2(Detektion)$ angenommen werden. Zwischen den Hervorhebungstechniken *rote Einfärbung* UV_2 und *Stippling* UV_1 hingegen besteht mit $p > 0,05$ kein signifikanter Unterschied, da $z_3 = 1,594 < 1,64$.

Reaktionszeit

Für die Überprüfung der Forschungshypothesen auf signifikante Unterschiede in Bezug auf ihre Reaktionszeit ($H_1(Reaktionszeit)$ und $H_2(Reaktionszeit)$) sind dieselben Paarungen wie bei der Überprüfung hinsichtlich der Detektionsgenauigkeit zu untersuchen. Aufgrund der Normalverteilung der AV_{RT} und der Abhängigkeit der Faktorstufen ist ein *t-Test für abhängige* Stichprobengruppen anzuwenden. In Listing 6.4 ist der entsprechende *SPSS*-Aufruf dargestellt. Es besteht auch die Möglichkeit, den Test über die Benutzeroberfläche in *SPSS* aufzurufen.

```
T-TEST
  PAIRS = rt_stippling rt_cutaway rt_red rt_cutaway rt_red
      rt_red
  WITH rt_noHT rt_noHT rt_noHT rt_stippling rt_stippling
      rt_cutaway (PAIRED)
  /CRITERIA = CI(.9916666).
```

Listing 6.4: *SPSS*-Syntax für den *t-Test* für abhängige Stichprobengruppen auf den gemessenen Reaktionszeiten mit *Bonferroni*-Korrektur. Das korrigierte Signifikanz-Niveau ist in diesem Fall durch das Konfidenzintervall /CRITERIA angegeben.

Hierbei ist zu beachten, dass eine manuelle *Bonferroni*-Korrektur angewendet werden muss, indem das Signifikanz-Niveau durch die Anzahl an Paarungen dividiert wird, $\alpha_{adj} = \frac{0{,}05}{6} = 0,0083333$. Daraus folgt, dass die Prüfgrößen t mit dem kritischen Wert $t_{0{,}008;27} = 2,86$, siehe Anhang A, zu vergleichen sind.

	df = n − 1	**|d|**	σ	**t**	**p**
$\mathbf{UV_1 - UV_2}$	27	42,25 ms	32,551 ms	0,245	0,404
$\mathbf{UV_1 - UV_3}$	27	611,07 ms	23,804 ms	4,851	0,000
$\mathbf{UV_1 - UV_4}$	27	2048,4 ms	47,879 ms	8,085	0,000
$\mathbf{UV_2 - UV_3}$	27	568,82 ms	32,459 ms	3,311	0,001
$\mathbf{UV_2 - UV_4}$	27	2090,6 ms	54,724 ms	7,219	0,000
$\mathbf{UV_3 - UV_4}$	27	2659,4 ms	47,36 ms	10,612	0,000

Tabelle 6.8: t-Tests auf die AV_{RT} mit den kumulierten Differenzen $|d|$ zwischen den abhängigen Messwerten der Faktorstufen und der Standardabweichung σ für alle Differenzen.

Die in der Tabelle 6.8 angegebenen Prüfgrößen berechnen sich mit der Gleichung 3.12 wie folgt:

$$t_A = \frac{2048,4 \cdot \sqrt{28}}{47,879} = 8,085 \qquad t_B = \frac{2090,6 \cdot \sqrt{28}}{54,724} = 7,219 \tag{6.6}$$

$$t_C = \frac{2659,4 \cdot \sqrt{28}}{47,36} = 10,612 \qquad t_D = \frac{611,07 \cdot \sqrt{28}}{23,804} = 4,851 \tag{6.7}$$

$$t_E = \frac{42,25 \cdot \sqrt{28}}{32,551} = 0,245 \qquad t_F = \frac{568,82 \cdot \sqrt{28}}{32,349} = 3,311 \tag{6.8}$$

Aus den ermittelten Werten für t_A, t_B und t_C ist ersichtlich, dass zwischen den Faktorstufen der Hervorhebungstechniken UV_1, UV_2, UV_3 und der Faktorstufe *keine Hervorhebung* (UV_4) ein hoch signifikanter Unterschied hinsichtlich der gemessenen Reaktionszeiten besteht. Aus den Werten der Reaktionszeiten, siehe Tabelle 6.1, wird auch die Richtung des Unterschiedes ersichtlich. Alle drei Hervorhebungstechniken bewirkten eine schnellere Reaktion als bei der Darstellung ohne Hervorhebung. Somit kann die Forschungshypothese $H_1(Reaktionszeit)$ als gültig angenommen werden. Die Hervorhebungstechnik mit der größten kumulierten Differenz $|d| = 2659,4\ ms$ ist *Cut-Away*, so dass diese als die Technik mit der schnellsten Detektionsunterstützung identifiziert und folglich die Gültigkeit der $H_2(Reaktionszeit)$ nachgewiesen wurde. Die nicht-hypothesenrelevante Überprüfung der Paarung UV_1 und UV_2 führte zu einem nicht signifikanten Ergebnis mit $t_E < t_{krit}$ und $p > 0,05$ bezüglich der Mittelwertunterschiede der beiden Techniken.

Die kritischen Werte sowie die Prüfgrößen können gemäß Abbildung 3.1 grafisch dargestellt werden. Anhand der kritischen Werte der z- und t- Prüfverteilungen kann ein Konfidenzintervall für den Annahmebereich der Nullhypothesen festgelegt werden. Die

hierfür benötigten Grenzen können mit der Gleichung 3.3 berechnet werden. Im Fall der t-Verteilung ist der z-Wert der Gleichung mit dem t-Wert zu ersetzen. Die Lage der Prüfgrößen ist nach derselben Gleichung mit den Werten der Prüfgrößen zu berechnen. Für den einseitigen Test muss die Berechnung entsprechend der Richtung einer Hypothese (+ oder -) durchgeführt werden.

6.3 Überprüfung auf praktische Bedeutsamkeit

Mit den durchgeführten Auswertungen in Abschnitt 6.2 wurde nachgewiesen, dass zwischen allen hypothesenrelevanten Paarungen $UV_1 - UV_4, UV_2 - UV_4$, $UV_3 - UV_4$, $UV_1 - UV_3$ und $UV_2 - UV_3$ ein signifikanter Unterschied bezüglich der Trefferraten und der Detektionsschnelligkeit existiert. UV_4 entspricht der Faktorstufe der nicht hervorgehobenen vergrößerten Lymphknoten. Die Faktorstufen UV_1, UV_2 und UV_3 entsprechen den mit den Techniken *Stippling*, *rote Einfärbung* und *CutAway* hervorgehobenen vergrößerten Lymphknoten.

Nachfolgend soll geklärt werden, ob diese Unterschiede für die praktische Anwendung von Bedeutung sind. In Abschnitt 4.3 wurden die hierfür benötigten Effektgrößen im Zuge der Berechnung der optimalen Stichprobengröße bereits bestimmt. Von einer praktischen Relevanz wäre dann auszugehen, wenn die geforderten Mindesteffekte von $0,5$ für AV_{Acc} und $0,8$ für AV_{RT} erreicht oder überschritten werden. In Tabelle 6.9 sind die auf Grundlage der Gleichung 4.2 ermittelten Effektstärken für die Paarungen der Faktorstufen dargestellt. Für die AV_{Acc} ergeben sich für die hypothesenrelevanten Paarun-

	$\mathbf{AV_{Acc}}$			$\mathbf{AV_{RT}}$		
Paarung	**Signifikanz**	δ	**Relevanz**	**Signifikanz**	$\delta_{\mathbf{RT}}$	**Relevanz**
$\mathbf{UV_1 - UV_2}$	nein	0,158	nein	nein	0,02	nein
$\mathbf{UV_1 - UV_3}$	ja	2,258	ja	ja	0,284	nein
$\mathbf{UV_1 - UV_4}$	ja	1,47	ja	ja	0,877	ja
$\mathbf{UV_2 - UV_3}$	ja	0,945	ja	ja	0,269	nein
$\mathbf{UV_2 - UV_4}$	ja	1,322	ja	ja	0,909	ja
$\mathbf{UV_3 - UV_4}$	ja	2,258	ja	ja	1,122	ja

Tabelle 6.9: Effektstärken der Paarungen.

gen mit $\delta > 0,5$ praktisch bedeutsame Unterschiede. Die Paarung $UV_1 - UV_2$ hingegen weist weder für die Trefferraten noch für die Reaktionszeiten eine ausreichende Effektstärke auf. Auch für $UV_1 - UV_3$ und $UV_2 - UV_3$ sind die Unterschiede der Reaktionszeiten nicht von Bedeutung. Für die hypothesenrelevanten Paarungen der $H_1(Reaktionszeit)$ hingegen wurden praktisch bedeutsame Unterschiede nachgewiesen.

6.4 Zusammenhangsanalyse

Bisher wurden die abhängigen Variablen der Faktorstufen univariat betrachtet. Durch die Berechnung der Korrelationskoeffizienten soll nun geklärt werden, ob zwischen den Trefferraten und den Reaktionszeiten der vier Faktorstufen ein Zusammenhang besteht. Aufgrund der unterschiedlichen Verteilung der Messdaten innerhalb der Faktorstufen wird der Rangkorrelationskoeffizient mit der Gleichung 3.2 nach Spearman [1907] berechnet. Die ermittelten Koeffizienten r der jeweiligen Korrelationspaarungen sind in Tabelle 6.10 wiedergegeben. Der Tabelle kann man entnehmen, dass zwischen den Tref-

$\mathbf{AV_{Acc}}$: $\mathbf{AV_{RT}}$	r
Stippling	-0,371
Rote Einfärbung	-0,262
CutAway	-0,209
Keine Hervorhebung	0,011

Tabelle 6.10: Korrelationen zwischen den abhängigen Variablen der Faktorstufen nach Spearman [1907].

ferraten und den Reaktionszeiten bei vergrößerten Lymphknoten ohne Hervorhebung mit r=0,011 kein Zusammenhang besteht. Bei den anderen Paarungen ist ein geringer, negativer Zusammenhang zu beobachten. Das bedeutet, dass die Vpn mit hoher Detektionsgenauigkeit auch tendenziell weniger Zeit zum detektieren der vergrößerten Lymphknoten benötigten. Nach Cohen [1992] sind die beobachteten Stärken der Zusammenhänge $|r| = 0,371$, $|r| = 0,262$ und $|r| = 0,209$ dennoch von eher kleiner bis mittlerer relevanter Bedeutung.

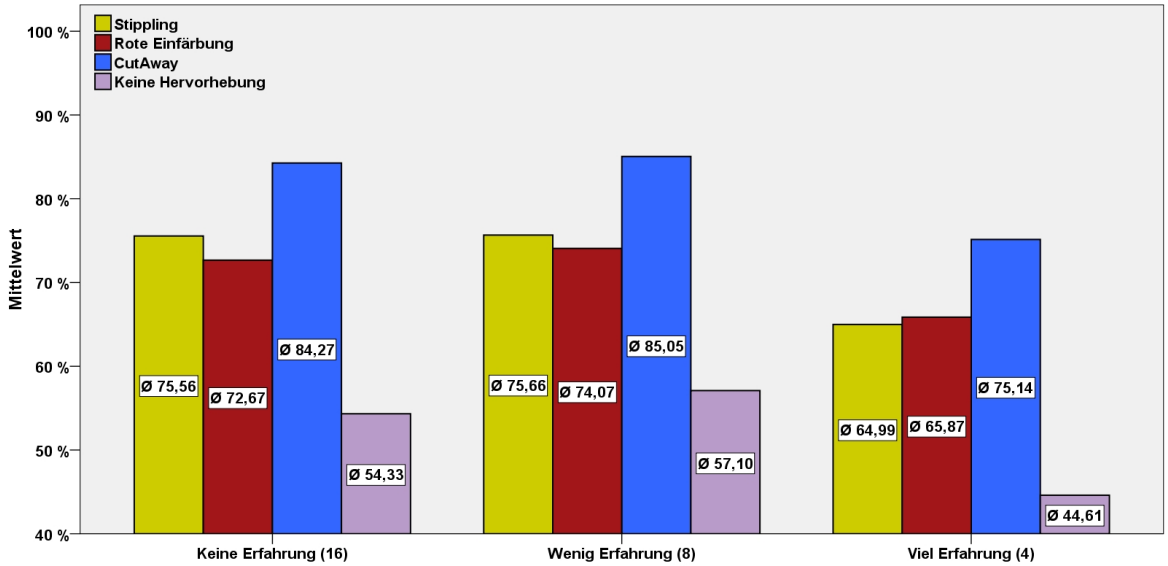
6.5 Explorative Datenanalyse

Im Anschluss an das Experiment am PC wurden die Teilnehmer gebeten einen Fragebogen auszufüllen. Die hieraus gewonnenen Angaben werden im Folgenden zum einen für sich betrachtet zum anderen mit den Messdaten verglichen.

Die von Tietjen [2004] und Mirschel [2004] entworfenen Fragebögen dienten dabei als Vorlage des in dieser Arbeit verwendeten Fragebogens (siehe Anhang D) und wurden unter Berücksichtigung von Hinweisen eines Diplom-Psychologen entsprechend angepasst. Die in den Fragebögen erhobenen demographischen Daten sind für die Hypothesenprüfung im Speziellen und die Studie insgesamt von untergeordneter Bedeutung. Zum einen kann hierdurch die Zusammensetzung der Zufallsstichprobe kontrolliert werden, zum anderen soll anhand der der Daten überprüft werden, inwieweit die subjektiven Meinungen mit den objektiven Messdaten übereinstimmen.

Das Alter der an dieser Studie teilnehmenden 33 Personen lag zwischen 19 und 51 Jahren. Davon waren 18 Vpn weiblichen und 15 Vpn männlichen Geschlechts. Die Stichprobengruppe enthielt zwei Ärzte und zwei Diplom-Designer, zehn Vpn aus dem Ingenieur-Bereich, 15 Psychologiestudenten und vier Vpn aus anderen Betätigungsfeldern. 20 Probanden verfügten über keine und neun nur über geringfügige Erfahrungen mit medizinischen Visualisierungen. Lediglich vier Vpn gaben an, viel Erfahrung darin zu haben. 25 Vpn waren rechtshändig, sechs Vpn linkshändig und zwei Vpn beidhändig. Zudem gaben 15 Vpn an kurzsichtig und zwei Vpn weitsichtig zu sein. Von einer Rot-Grün-Sehschwäche war keiner der Teilnehmer betroffen.

Die verschiedenen Alters-, Geschlecht- und Tätigkeits- Gruppen unterscheiden sich hinsichtlich der Trefferraten und Reaktionszeiten kaum voneinander. Aufgrund der Minderheit der Links- und Beidhänder wird anhand dieses Unterscheidungskriteriums kein Vergleich zwischen den Vpn gezogen. Bei der Gegenüberstellung von Vpn die entweder



(a) Erfahrungen der Vpn

Abbildung 6.4: Vergleich der Trefferraten zwischen Vpn mit unterschiedlichen Erfahrungen mit medizinischen Visualisierungen. Die in Klammern gefassten Zahlen entsprechen der Anzahl der in den Gruppen enthaltenen Vpn.

keine, wenig oder viel Erfahrung mit medizinischen Visualisierungen hatten, unterscheiden sich die Resultate der Vpn mit wenig und keiner Erfahrung nur geringfügig voneinander (Abbildung 6.4). Da diese Stichprobengruppe lediglich aus vier Vpn bestand, ist eine Verallgemeinerung dieses Ergebnisses nicht möglich. Dennoch zeigt der Vergleich mit den Personen, welche über keine oder nur geringe Erfahrungen in diesem Bereich haben, dass in der Tendenz die Ergebnisse übereinstimmen. Mit *CutAway* wurden jeweils die besten, *ohne Hervorhebungstechnik* jeweils die schlechtesten Ergebnisse erzielt. Auch sind die Trefferraten mit *CutAway* deutlich höher als mit den beiden

anderen Hervorhebungstechniken. Eine Signifikanzprüfung ist hierbei nicht sinnvoll, da zu wenige Vpn innerhalb dieser Gruppen enthalten sind und zudem die Gruppen ungleich groß sind. Aus diesem Grund wurde sich auf eine rein deskriptive Darstellung der Ergebnisse beschränkt.

In den Fragebögen konnten sich die Vpn dazu äußern, ob und welche der in den Stimuli abgebildeten Halsstrukturen sie als störend für die Aufgabenbewältigung empfanden. Dabei waren den Vpn jeweils fünf Antwortmöglichkeiten (2=*sehr störend*, 1=*störend*, 0=*egal*, -1= *nicht störend*, -2=*gar nicht störend*) vorgegeben (ordinales Messniveau). Abbildung 6.5 zeigt, dass Speicheldrüsen als die einzigen störenden Strukturen bewertet wurden, was wahrscheinlich an der Ähnlichkeit mit den Lymphknoten liegt. Dieser Meinung waren auch die beiden an der Studie teilnehmenden Ärzte. Dennoch kann auf die Darstellung dieser nicht verzichtet werden, da sie unter Umständen von pathologischen Lymphknoten infiltriert sein können. Die Muskeln, Knochen sowie der Kehlkopf hingegen wurden nicht als störend empfunden. Die mit dieser Umfrage gewonnenen Erkenntnisse können insbesondere für weitergehende Arbeiten von Bedeutung sein. Durch eine gezielte Manipulation der, als am störendsten empfundenen Strukturen, kann möglicherweise die Detektionsunterstützung durch Hervorhebungstechniken weiter verbessert werden.

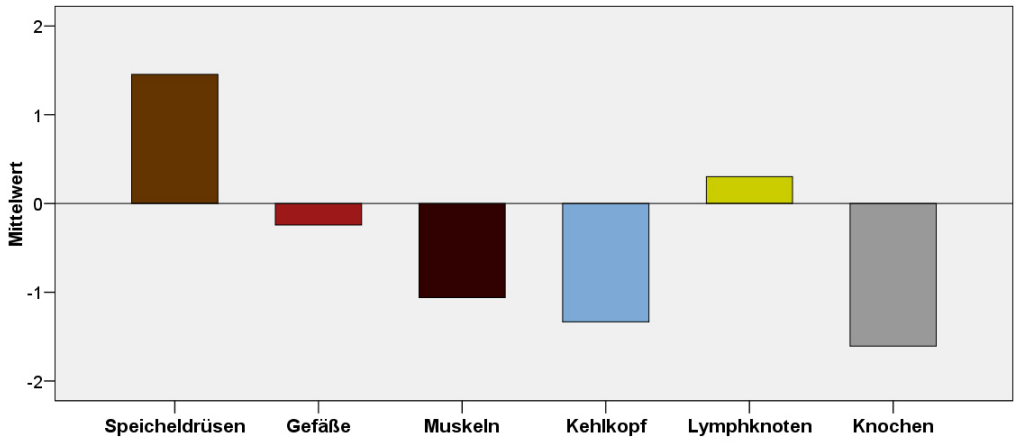


Abbildung 6.5: Bewertung der Halsstrukturen in den Stimuli hinsichtlich ihres Störeinflusses. Die negativen Bewertungen sagen aus, dass die Strukturen entweder *gar nicht*, mit -2, oder als *wenig störend*, mit -1, empfunden wurden. Die positiven Bewertungen entsprechen den komplementären Angaben von *sehr störend* und *störend*.

Weiter wurden die Vpn befragt, inwiefern sie die im Experiment gezeigten Faktorstufen für die praktische Anwendung als geeignet erachten. Den Vpn wurden wieder jeweils fünf Antwortmöglichkeiten von -2 *gar nicht geeignet* bis 2 *sehr geeignet* vorgegeben (ordinales Messniveau). Anhand der grafischen Darstellung des Umfrageergebnisses in Abbildung 6.6 ist zu erkennen, dass die subjektive Einschätzung der Hervorhebungstechniken im Wesentlichen mit den objektiv gemessenen Ergebnissen übereinstimmt.

Es ist eine eindeutige Präferenz der Technik *CutAway* mit einem Mittelwert von $1,67$ festzustellen. Die Techniken *Stippling* und *rote Einfärbung* wurden ebenfalls positiv bewertet, wobei die *rote Einfärbung* mit einem Mittelwert von $0,7$ dem *Stippling* mit $0,48$ vorgezogen wurde. Die meisten Vpn empfanden die Darstellung von Lymphknoten *ohne Hervorhebung* mit $-0,79$ als ungeeignet.

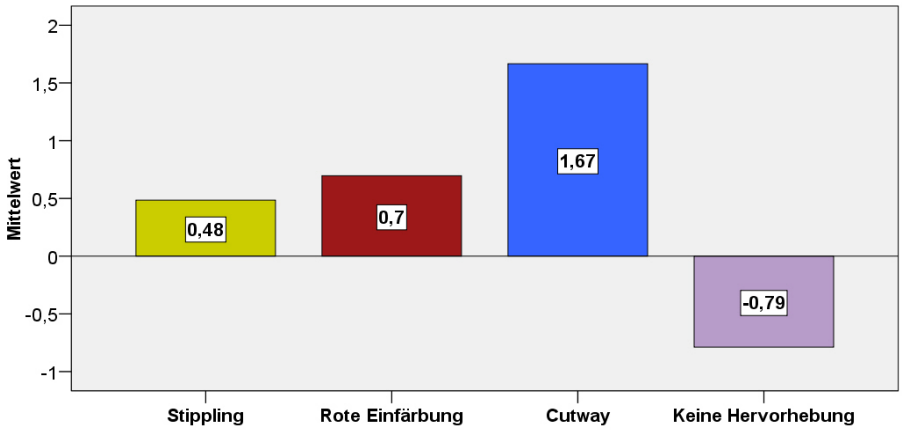


Abbildung 6.6: Einschätzung der Faktorstufen hinsichtlich ihrer Eignung in der praktischen Anwendung. Die negativen Bewertungen bedeuten, dass die Strukturen entweder als *gar nicht geeignet*, mit -2, oder als *nicht geeignet* mit -1, empfunden wurden. Die positiven Bewertungen entsprechen den komplementären Angaben von *sehr geeignet* mit 2 und *geeignet* mit 1.

Darüber hinaus wurden die Vpn aufgefordert, eine Faktorstufe zu nennen, die ihnen für die Detektion von vergrößerten Lymphknoten am geeignetsten erscheint (nominales Messniveau). Das Ergebnis des Tests ist in Abbildung 6.7 in einem Tortendiagramm dargestellt. Bei dieser Umfrage wurden die Techniken *Stippling* und *rote Einfärbung* mit je $15,15\%$ gleich bewertet. Dies entspricht nicht ganz dem vorherigen Vergleich der Faktorstufen, bei dem die *rote Einfärbung* dem *Stippling* vorgezogen wurde. Auch hier geht *CutAway* mit $69,7\%$ wieder als die meist bevorzugte Technik hervor. Insgesamt votierten 23 von 33 Vpn für die Hervorhebungstechnik *CutAway* als die Technik mit der vergrößerte Lymphknoten am besten detektiert werden können. Die Faktorstufe *Keine Hervorhebung* hingegen wurde von keinem Probanden bevorzugt. Für die Techniken *Stippling* und *rote Einfärbung* haben sich jeweils fünf Vpn entschieden. Die Antworten wurden zusätzlich mit dem *Chi-Quadrat*-Test auf Signifikanz untersucht, dessen Ergebnis in Tabelle 6.11 zu finden ist. Mit einem Prüfwert $\chi^2 = 37,18$ ist der kritische Wert der Prüfverteilung mit $\chi^2_{krit} = 7,81$ auf dem Signifikanzniveau $\alpha = 0,05$ weit überschritten, so dass die Antworten der Vpn nicht gleichverteilt sind. Die beobachteten Bewertungen von *CutAway* heben sich demnach signifikant von den beobachteten Bewertungen der anderen Faktorstufen ab.

Für die Gegenüberstellung der subjektiven Angaben der Vpn mit deren objektiv erfassten Messdaten werden die in Abschnitt 6.1.1 als Ausreißer deklarierten Vpn nicht

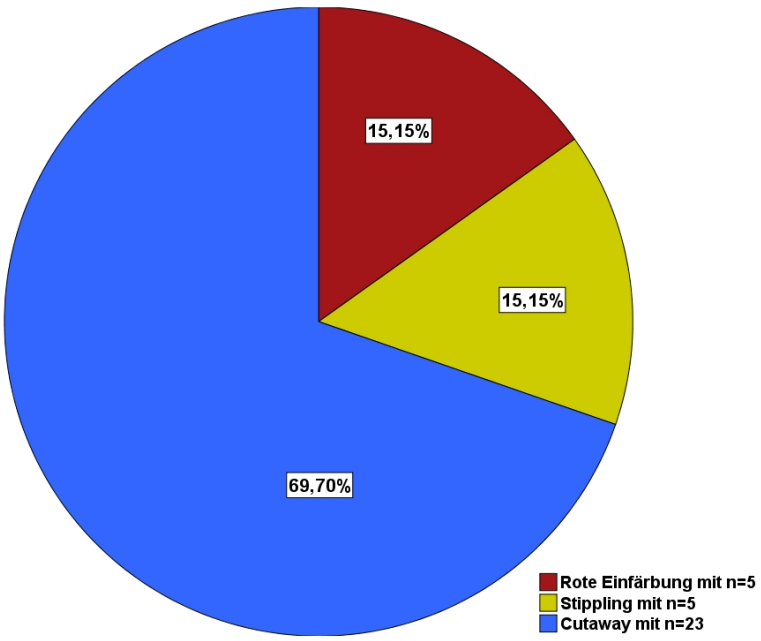


Abbildung 6.7: Subjektive Einschätzung der 33 Vpn hinsichtlich der Hervorhebungstechniken.

Hervorhebungstechniken	$\mathbf{f_o}$	$\mathbf{f_e}$	$\frac{(\mathbf{f_o}-\mathbf{f_e})^2}{\mathbf{f_e}}$	χ^2	**p**
Stippling	5	8,25	1,2803	37,18	0,0000
Rote Einfärbung	5	8,25	1,2803		
CutAway	23	8,25	26,3712		
Keine Hervorhebung	0	8,25	8,25		

Tabelle 6.11: *Chi-Quadrat*-Test auf die von den Vpn favorisierten Faktorstufen mit 3 Freiheitsgraden. Der kritische Wert der Prüfverteilung ist $\chi^2_{0,05;3} = 7,81$.

betrachtet. Insgesamt erzielten 27 von 28 Vpn mit *CutAway* die besten Trefferraten im Experiment. Jedoch nur 19 Vpn bewerteten diese Technik als die für die Praxis geeignetste von den drei vorgestellten Hervorhebungstechniken. Vier Vpn hingegen bevorzugten *Stippling* wobei nur eine dieser drei Vpn mit *Stippling* auch die beste Detektionsgenauigkeit erreichte. Fünf Versuchsteilnehmer votierten für die *rote Einfärbung* von vergrößerten Lymphknoten, erreichten jedoch mit dieser nicht die höchste Trefferqenauigkeit. Somit ergibt sich, dass nur $71,43\%$ der Befragten Personen mit den von ihnen bevorzugten Hervorhebungstechniken auch die besten Ergebnisse hinsichtlich der Detektionsgenauigkeit erreichten.

6.6 Zusammenfassung

Die Hervorhebungstechniken verbessern sowohl die Treffergenauigkeit als auch die Reaktionszeit bei der Detektion von pathologischen Lymphknoten. Dies konnte bereits im Rahmen der deskriptiven Analyse festgestellt werden. Zuvor wurden einige Ausreißer anhand ihrer Trefferquote und zu großer Mittelwertabweichung von der Analyse ausgeschlossen. Die Signifikanzprüfung ergab nicht nur, dass jede Hervorhebungstechnik die Detektion signifikant verbessert, sondern auch, das mit *CutAway* signifikant bessere Ergebnisse erzielt wurden als mit *roter Einfärbung* und *Stippling*. Alle vier Forschungshypothesen konnten durch diesen Nachweis angenommen werden.

Darüber hinaus konnte im Rahmen der Relevanzprüfung nachgewiesen werden, dass die Verbesserungen auch praktisch bedeutend sind. Der indirekte Zusammenhang zwischen den Trefferraten und den Reaktionszeiten bei den Hervorhebungstechniken ist hingegen nur von untergeordneter Bedeutung.

Abschließend wurden die Angaben der Probanden in den Fragebögen ausgewertet und mit den Messdaten verglichen. Auch in den Fragebögen wurde *CutAway* von den Teilnehmern präferiert. Allerdings bevorzugten lediglich knapp zweidrittel der Vpn *CutAway*, den Messdaten zufolge erzielten jedoch alle Teilnehmer, bis auf eine Ausnahme, mit der regionalen Hervohebungstechnik ihr bestes Ergebnis.

7 Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war es, ein Design zu entwickeln und umzusetzen, mit dessen Hilfe wahrnehmungsunterstützende Visualisierungen evaluiert werden können. Die Grundlage hierfür bildeten die in der experimentellen Psychologie verwendeten Reaktionszeittests. Am Beispiel der Evaluation der Hervorhebungstechniken konnte nachgewiesen werden, dass eine Adaption auf komplexe dreidimensionale anatomische Szenarien möglich ist. Das entwickelte Design ist geeignet, die Grundlage eines Standards für Evaluierungen in der Computervisualistik zu bilden. Ein solcher Standard ist unerlässlich damit der Rückstand zu anderen Wissenschaften überwunden werden kann. Hierfür erforderlich ist auch die transparente Darstellung des Experimentes und der Ergebnisse, d.h.

- die Angabe der zugrunde liegenden Fragestellung oder Hypothese und der daraus operationalisierten Variablen sowie
- die Beschreibung der verwendeten Stimuli, Stichprobengröße und Stichprobenzusammensetzung, als auch
- die Erläuterung des Versuchsdesigns,
- die Angabe der angewendeten Analyseverfahren mit eventueller Begründung und
- der deskriptiven Darstellung und Präsentation der Ergebnisse unter Angabe der verwendeten Signifikanzniveaus.

Die Ergebnisse einer solchen experimentellen Evaluierung sind objektiv, valide und reliabel. Sie können mit den Ergebnissen anderer Studien verglichen werden, sofern diese nach demselben Muster durchgeführt wurden. Für die Experimentdurchführung und die Auswertung der Messdaten wurde die Software *Presentation* verwendet. Die Aufbereitung der erhobenen Daten erfolgte in dieser Arbeit manuell und war sehr zeitaufwendig. Eine Automatisierung der Datenaufbereitung könnte diesen Prozess für zukünftige Arbeiten vereinfachen und beschleunigen. Die für die Auswertung dieser Messdaten elementaren statistischen Auswertungsverfahren wurden eingehend erläutert, in *SPSS* durchgeführt und in *Microsoft Excel* nachvollzogen.

Das Design ist nicht auf die experimentelle Evaluierung von Visualisierungsstechniken beschränkt. Eine individuelle Anpassung ermöglicht es, Laufzeiten neu entwickelter Algorithmen sowie neu entwickelter Interaktionssysteme auf Signifikanz und Relevanz

zu überprüfen. Darüber hinaus können neben Unterschieden auch Zusammenhänge und Veränderungen untersucht werden.

Über die grundlegende Entwicklung eines Versuchsdesigns hinaus sollte im Wege dieser Arbeit eruiert werden, inwieweit Hervorhebungstechniken die Detektion pathologischer Lymphknoten verbessern. Die Ergebnisse der stellvertretenden Hervorhebungstechniken *CutAway*, *rote Einfärbung* und *Stippling* wurden mit Resultaten der Versuchspersonen für nicht hervorgehobene vergrößerte Lymphknoten verglichen. Dabei zeigte sich, dass mit allen drei Techniken gegenüber keiner Hervorhebung der vergrößerten Lymphknoten signifikant bessere Ergebnisse erzielt wurden. *CutAway* hat sich als regionale Hervorhebungstechnik deutlich von den beiden lokalen Techniken absetzen können. Sowohl die Reaktionszeiten als auch die Trefferraten waren signifikant und relevant besser als die der anderen Hervorhebungstechniken. Ob generell mit regionalen Hervorhebungstechniken bessere Ergebnisse erzielt werden als mit lokalen, kann nur vermutet werden. Dies kann im Wege weiterer Studien herausgefunden werden. Vorstellbar ist auch, dass eine Kombination verschiedener Hervorhebungstechniken geeigneter ist als eine einzelne Technik. Auch blieb der Zusammenhang von Fokusstruktur, Hervorhebungstechnik und Kontextstrukturen in dieser Arbeit unberücksichtigt. Daher lassen sich die Ergebnisse der vorliegenden Studie nicht einfach für die Detektion anderer Fokusstrukturen in anderen Umgebungen übertragen. Wahrnehmungsunterstützende Kombinationen von Fokus- und Kontextstrukturen in mehr-faktoriellen Versuchsplänen können Gegenstand weiterführender Studien sein. Weiterhin könnte eine Evaluierung verschiedener Ansichten anatomischer Szenen in Abhängigkeit bestimmter Fokusbereiche von Bedeutung sein. Und auch die Verwendung eines Eye-Trackers kann möglicherweise interessante Ergebnisse hervorbringen, z.B. könnte so geklärt werden, ob die Suche nach pathologischen Strukturen in medizinischen Visualisierungen unter Verwendung von Hervorhebungstechniken weiterhin seriell abläuft.

Da mit der Durchführung und Auswertung eines Experimentes regelmäßig ein hoher Zeitaufwand einhergeht, sollte zuvor eine Aufwand-Nutzen-Abschätzung durchgeführt werden. Um Wiederholungen hinsichtlich des Evaluierungsgegenstandes zu vermeiden, ist es vorstellbar, dass die Ergebnisse und erhobenen Daten in einer Datenbank gespeichert werden. Aufgrund es Zeitaufwandes lassen sich Ärzte für derartige Studien nur schwer gewinnnen. Eine selbstständige Durchführung am Computer der Ärzte ist keine geeignete Alternative, denn die notwendige Kontrolle der Parameter und Störvariablen ist in diesem Fall nur sehr eingeschränkt möglich.

Mit einer experimentellen Evaluierung können Hypothesen lediglich überprüft werden, so dass ein Fragebogen immer dann zusätzlich notwendig ist, wenn darüber hinausgehende Ansatzpunkte für Weiterentwicklungen herausgefunden werden sollen. Wie aus dieser Arbeit hervorgeht, wurden Speicheldrüsen als die störenste Kontextstruktur empfunden. Eine gezielte Manipulation der Darstellung dieser kann möglicherweise die Detektionsleistung weiter verbessern.

Der Einsatz von Fragebögen ist oftmals weniger zeitaufwendig. Zudem können mehr Teilnehmer mit dieser Evaluierungsmethode erreicht werden, z.B. durch Verteilen in Lehrveranstaltungen, durch Online-Umfragen oder das Versenden an den Anwender. Fragebögen können eine experimentelle Evaluierung jedoch nur in Ausnahmefällen ersetzen und haben in aller Regel nur eine Ergänzungsfunktion. Zwar können sowohl die erhobenen Daten einer experimentellen Studie als auch die einer Umfrage auf Signifikanz und Relevanz hin überprüft werden, jedoch sind diese Angaben nur subjektive Meinungswiedergaben, wohingegen die Messdaten der experimentellen Erhebung selbst bereits objektiv sind. Zumal der subjektive Eindruck der Probanden, wie in dieser Arbeit gezeigt wurde, deutlich von den objektiven Messungen abweichen kann.

Durch die hier beschriebene experimentelle Evaluierung werden fundierte und wissenschaftlich belastbare Ergebnisse erzielt. Mit der hieraus gewonnenen neuen Qualität der Ergebnisse und der damit verbundenen gesteigerten Glaubwürdigkeit kann der Skepsis und den Berührungsängsten, welche die Praxis gegenüber neu entwickelten Verfahren hat, entgegengetreten werden. Die hierdurch erhöhte Akzeptanz durch die Ärzte führt zu einer schnelleren und verbreiterten Übernahme der neuen Visualisierungstechniken. Das kann in einem qualitativ und quantitativ gesteigerten Feedback resultieren, was wiederum eine Beschleunigung und Verbesserung der Entwicklungen in der Visualisierung zur Folge haben kann. Die Evaluierung nach dem Vorbild dieser Arbeit kann eine wichtige Schnittstelle zwischen Entwickler und Praktiker sein.

Literaturverzeichnis

[MEVIS 2009] *Fraunhofer MEVIS*. 2009. – URL `http://www.mevis.de/mre/`. – Letzte Aktualisierung: 27.07.2009

[Neurobs 2009] *Neurobehavioral Systems Presentation*. 2009. – URL `http://www.neurobs.com`. – Stand: 12.08.2009, Version: 12.2

[Bortz 2005] Bortz, J.: *Statistik für Human- und Sozialwissenschaftler*. Bd. 6. Springer Verlag, 2005

[Bortz und Döring 2006] Bortz, J. ; Döring, N.: *Forschungsmethoden und Evaluation*. 4. Springer Medizin Verlag, 2006

[Bulmer 2003] Bulmer, M.: *Francis Galton: Pioneer of Heredity and Biometry*. Johns Hopkins University Press, 2003

[Burgert et al. 2007] Burgert, O. ; Örn, V. ; Gessat, M. ; Joos, M. ; Strauss, G. ; Preim, B. ; Tietjen, C. ; Hertel, I.: Evaluation of perception performance in neck dissection planning using eye tracking and attention landscapes. In: *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment* (2007)

[Card et al. 1983] Card, S. ; Moran, T. ; Newell, A.: *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates, 1983

[Cohen 1988] Cohen, J.: *Statistic Power Analysis for the Behavioural Science*. Erlbaum, 1988

[Cohen 1992] Cohen, J.: A Power Primer. In: *Psychological Bullitin* 112 (1992), S. 155–159

[DeGEval 2009] DeGEval: *Gesellschaft für Evaluation e.V.* 2009. – URL `http://www.degeval.de/`. – Stand: 13.08.2009

[Denning 1980] Denning, P. J.: What is experimental computer science? In: *Communication of the ACM* 23(10) (1980), S. 543–544

[Duller 2007] Duller, C.: *Einführung in die Statistik mit Excel und SPSS*. Physica-Verlag, 2007

[Fechner 1860] Fechner, G. T.: *Elemente der Psychophysik*. Breitkopf und Härtel, 1860

[Fisher 1924] Fisher, R. A.: On A Distribution Yielding The Error Functions Of Several Well Known Statistics. In: *Proceedings of the International Congress of Mathemathics* 2 (1924), S. 805–813

[Fisher 1925] Fisher, R. A.: *Statistical Methods for Research Workers*. Oliver & Boyed, 1925

[Fisher 1935] Fisher, R. A.: *The Design of Experiments*. Reprint 1990, Oxford University Press, 1935

[Friedman 1937] Friedman, M.: The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. In: *Journal of the American Statistical Association* 32 (1937)

[Gosset 1908] Gosset, W.S.: The Probable Error Of A Mean. In: *Biometrika* 6 (1908), S. 1–25

[Gray 1918] Gray, H.: *Anatomy of the Human Body*. 20. Lea & Febiger, 1918

[Green und Swets 1966] Green, D.M. ; Swets, J. A.: *Signal Detection and Psychophysics*. Reprint from 1966. New York: Wiley, 1966

[Hansen 2006] Hansen, C.: *Verwendung von Textur in der Gefäßvisualisierung*, Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2006

[Huber 2005] Huber, O.: *Das psychologische Experiment: Eine Einführung*. Bd. 4. 2005

[Hussy und Jain 2002] Hussy, W. ; Jain, A.: *Experimentelle Hypothesenprüfung in der Psychologie*. Hogrefe, 2002

[Isenberg et al. 2006] Isenberg, T. ; Neumann, P. ; Carpendale, S. ; Sousa, M. C. ; Jorge, J. A.: Non-Photorealistic Rendering in Context: An Observational Study. (2006), S. 115–126. – URL `http://cpsc.ucalgary.ca/~isenberg/paperpages/Isenberg_2006_NPR.html`

[Kellermann 2009] Kellermann, K.: *Automatische Ableitung und Verarbeitung semantischer Informationen zur Generierung adaptiver Interventions-Planungs-Visualisierungen*, Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2009

[Kähler 2002] Kähler, M.-W.: *Statistische Datenanalyse*. Vieweg, 2002

[Kobsa 2001] Kobsa, A.: An Empirical Comparison of Three Commercial Information Visualization Systems. In: *Information Visualization, IEEE Symposium on* 0 (2001), S. 123. – ISSN 1522-404X

[Kobsa 2004] Kobsa, A.: User Experiments with Tree Visualization Systems. In: *IEEE Symposium on Information Visualization* (2004), S. 9–16

[Kosara et al. 2003] Kosara, R. ; Hauser, H. ; Miksch, S. ; Schrammel, J. ; Giller, V. ; Tscheligi, M.: Experimental Evaluation of Semantic Depth of Field, a Preattentive Method for Focus+Context Visualization. In: *Human-Computer Interaction* (2003), S. 888–891

[Kosara et al. 2002] Kosara, R. ; Miksch, S. ; Hauser, H.: Focus+Context Taken Literally. 22 (2002), Nr. 1, S. 22–29

[Kruskal und Wallis 1952] Kruskal, W. H. ; Wallis, W. A.: Use of Ranks in One-Criterion Variance Analysis. In: *Journal of the American Statistical Association* 47 (1952), S. 583–621

[Lukowicz et al. 1994] Lukowicz, P. ; Heinz, E.A. ; Prechelt, L. ; Tichy, W.F.: Experimental Evaluation in Computer Science: A Quantitative Study. In: *Journal of Systems and Software* (1994)

[Mann und Whitney 1947] Mann, H. B. ; Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. In: *The Annals of Mathematical Statistics* 18 (1947), S. 50–60

[Mirschel 2004] Mirschel, S.: *Erstellung eines Prototypen für ein fallbasiertes Lernsystem in der Leberchirurgie*, Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2004

[Neugebauer 2006] Neugebauer, M.: *Entwicklung eines Verfahrens zur parametrisierbaren Kamerapositionierung in der medizinischen Visualisierung*, Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2006

[Oeltze 2004] Oeltze, S.: *Visualisierung baumartiger anatomischer Strukturen mit Convolution Surfaces*, Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2004

[Pearson 1900] Pearson, K.: On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. In: *Philosophical Magazine* 50 (1900), S. 157–175

[Preim 1999] Preim, B.: *Entwicklung interaktiver Systeme: Grundlagen, Fallbeispiele und innovative Anwendungsfelder*. Springer-Verlag, 1999

[Preim und Bartz 2007] Preim, B. ; Bartz, D.: *Visualization in Medicine*. 2007

[Preim und Ritter 2002] Preim, B. ; Ritter, F.: Techniken zur Hervorhebung von Objekten in medizinischen 3d-Visualisierungen. In: *SimVis2002*, 2002, S. 187–200

[proEval 2009] proEval: *Gesellschaft zur Förderung von professioneller Evaluation*. 2009. – URL `http://www.proeval.com/`. – Stand: 13.08.2009

[Ritter et al. 2006] Ritter, F. ; Hansen, C. ; Dicken, V. ; Konrad, O. ; Preim, B. ; Peitgen, H. O.: Real-time illustration of vascular structures. In: *IEEE Transactions on Visualization and Computer Graphics* 12(5) (2006), S. 877–884

[Sedlmeier und Renkewitz 2008] Sedlmeier, P. ; Renkewitz, F.: *Forschungsmethoden und Statistik in der Psychologie*. PEARSON Studium, 2008

[Spearman 1907] Spearman, C.: Demonstration Of Formulae For True Measurement Of Correlation. In: *The American Journal of Psychology* (1907)

[SPSS 2009] SPSS: *SPSS*. 2009. – URL `http://www.spss.com/de/`. – Stand: 12.08.2009, Version: 15.0

[Stanislaw und Todorov 1999] Stanislaw, H. ; Todorov, N.: Calculation of signal detection theory measures. In: *Behaviour Research Methods, Instruments & Computers* 31(1) (1999), S. 137–149

[Strothotte und Schlechtweg 2002] Strothotte, T. ; Schlechtweg, S.: *Non-Photorealistic Computer Graphics - Modelling, Rendering and Animation*. Morgan Kaufmann, 2002

[Thorpe et al. 1996] Thorpe, S.J. ; Fize, D. ; Marlot, C.: Speed of processing in the human visual system. In: *Nature* 381 (1996), S. 520–522

[Thurstone 1927] Thurstone, L.L.: A Law of Comparative Judgment. In: *Psychology Review* 34 (1927), S. 273–286

[Tietjen 2004] Tietjen, C.: *Evaluierung und Modifkation von Methoden zur Generierung von Liniengrafken in der medizinischen Visualisierung*, Otto-von-Guericke-Universität Magdeburg, Diplomarbeit, 2004

[Tietjen 2009] Tietjen, C.: *Illustrative Visualisierungstechniken zur Unterstützung der präoperativen Planung von chirurgischen Eingriffen*, Otto-von-Guericke-Universität Magdeburg, Dissertation, 2009

[Tory 2003] Tory, M.: Mental Registration of 2D and 3D Visualizations (An Empirical Study). (2003)

[Tory und Möller 2004] Tory, M. ; Möller, T.: Human Factors In Visualization Research. In: *IEEE Transactions on Visualization and Computer Graphics* 10 (2004)

[Tory und Möller 2005] Tory, M. ; Möller, T.: Evaluating Visualizations: Do Expert Reviews Work? In: *IEEE Transactions on Visualization and Computer GraphicsTory.Moeller2005* (2005)

[Treisman und Gelade 1980] Treisman, A. ; Gelade, G.: A Feature-Integration Theory of Attention. In: *Cognitive Psychology* 10 (1980), S. 97–136

[Wilcoxon 1945] Wilcoxon, F.: Individual Comparisons By Ranking Methods. In: *Biometrics Bulletin* 1 (1945), S. 80–83

[Wittekind et al. 2005] Wittekind, C. ; Klimpfinger, M. ; Sobin, L. H.: *TNM-Atlas: Illustrierter Leitfaden zur TNM/pTNM-Klassifikation maligner Tumoren.* Springer, Berlin, 2005

[Wolfe et al. 1989] Wolfe, J. M. ; Cave, K. R. ; Franzel, S. L.: Guided search: An alternative to the feature integration model for visual search. In: *Journal of Experimental Psychology: Human Perception and Performance* 15 (1989), S. 419–433

Stichwortregister

A Tabellen

Im Folgenden sind die Tabellen der Prüfverteilungen dargestellt, aus denen die entsprechenden kritischen Werte entnommen werden können.

- Tabelle der Standardnormalverteilung
- F-Tabelle
- t-Tabelle
- χ^2-Tabelle
- Wilcoxon-Tabelle

Tabelle der Standardnormalverteilung

	Zahlen der 2. Nachkommastelle von z									
z*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tabelle wurde erzeugt in Microsoft Excel mit der Funktion STANDNORMVERT(z).

Aufgrund der Symmetrie der Verteilung werden hier nur die positiven Werte dargestellt.

Kritische Werte bei F-Verteilungen

für $F_{\alpha,df1,df2}$ mit $\alpha=0{,}05$

	df_1									
df_2	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
32	4,15	3,29	2,90	2,67	2,51	2,40	2,31	2,24	2,19	2,14
34	4,13	3,28	2,88	2,65	2,49	2,38	2,29	2,23	2,17	2,12
36	4,11	3,26	2,87	2,63	2,48	2,36	2,28	2,21	2,15	2,11
38	4,10	3,24	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93
200	3,89	3,04	2,65	2,42	2,26	2,14	2,06	1,98	1,93	1,88

Tabelle erstellt in Microsoft Excel mit der Funktion FINV(0,05; df1; df2)

Kritische Werte bei t-Verteilungen

für $t_{\alpha,df}$

	kritische Werte bei einem 1-seitigen T-Test					
Testniveau:	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**	**0,004**
	kritische Werte bei einem 2-seitigen T-Test					
Testniveau: / **df**	**0,2**	**0,1**	**0,05**	**0,02**	**0,01**	**0,008**
1	3,08	6,31	12,71	31,82	63,66	79,57
2	1,89	2,92	4,30	6,96	9,92	11,11
3	1,64	2,35	3,18	4,54	5,84	6,32
4	1,53	2,13	2,78	3,75	4,60	4,91
5	1,48	2,02	2,57	3,36	4,03	4,26
6	1,44	1,94	2,45	3,14	3,71	3,90
7	1,41	1,89	2,36	3,00	3,50	3,67
8	1,40	1,86	2,31	2,90	3,36	3,51
9	1,38	1,83	2,26	2,82	3,25	3,39
10	1,37	1,81	2,23	2,76	3,17	3,30
11	1,36	1,80	2,20	2,72	3,11	3,23
12	1,36	1,78	2,18	2,68	3,05	3,17
13	1,35	1,77	2,16	2,65	3,01	3,13
14	1,35	1,76	2,14	2,62	2,98	3,09
15	1,34	1,75	2,13	2,60	2,95	3,06
16	1,34	1,75	2,12	2,58	2,92	3,03
17	1,33	1,74	2,11	2,57	2,90	3,00
18	1,33	1,73	2,10	2,55	2,88	2,98
19	1,33	1,73	2,09	2,54	2,86	2,96
20	1,33	1,72	2,09	2,53	2,85	2,95
21	1,32	1,72	2,08	2,52	2,83	2,93
22	1,32	1,72	2,07	2,51	2,82	2,92
23	1,32	1,71	2,07	2,50	2,81	2,90
24	1,32	1,71	2,06	2,49	2,80	2,89
25	1,32	1,71	2,06	2,49	2,79	2,88
26	1,31	1,71	2,06	2,48	2,78	2,87
27	1,31	1,70	2,05	2,47	2,77	**2,86**
28	1,31	1,70	2,05	2,47	2,76	2,86
29	1,31	1,70	2,05	2,46	2,76	2,85
30	1,31	1,70	2,04	2,46	2,75	2,84
40	1,30	1,68	2,02	2,42	2,70	2,79
60	1,30	1,67	2,00	2,39	2,66	2,74

Tabelle erstellt in Microsoft Excel mit der Funktion TINV(α; df) für 2-seitigen Test
und TINV(2α; df) für den 1-seitigen Test

Kritische Werte bei χ^2-Verteilungen

für $\chi^2_{\alpha,df}$

Testniveau: / df	0,1	0,05	0,01
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21
11	17,28	19,68	24,72
12	18,55	21,03	26,22
13	19,81	22,36	27,69
14	21,06	23,68	29,14
15	22,31	25,00	30,58
16	23,54	26,30	32,00
17	24,77	27,59	33,41
18	25,99	28,87	34,81
19	27,20	30,14	36,19
20	28,41	31,41	37,57
21	29,62	32,67	38,93
22	30,81	33,92	40,29
23	32,01	35,17	41,64
24	33,20	36,42	42,98
25	34,38	37,65	44,31
26	35,56	38,89	45,64
27	36,74	40,11	46,96
28	37,92	41,34	48,28
29	39,09	42,56	49,59
30	40,26	43,77	50,89
40	51,81	55,76	63,69
60	74,40	79,08	88,38

Tabelle erstellt in Microsoft Excel mit der Funktion CHIINV(0,05; df)

Kritische Werte für den Wilcoxon- Test

für $F_{\alpha,df}$

kritische Werte bei einem 1-seitigen T-Test			
Testniveau:	**0,05**	**0,025**	**0,01**
kritische Werte bei einem 2-seitigen T-Test			
Testniveau:	**0,1**	**0,5**	**0,02**
df			
5	0		
6	2	0	
7	3	2	0
8	5	3	1
9	8	5	3
10	10	8	5
11	13	10	7
12	17	13	9
13	21	17	12
14	25	21	15
15	30	25	19
16	35	29	23
17	41	34	27
18	47	40	32
19	53	46	37
20	60	52	43
21	67	58	49
22	75	65	55
23	83	73	62
24	91	81	69
25	100	89	76
26	110	98	84
27	119	107	92
28	130	116	101
29	140	126	110
30	151	137	120
31	163	147	130
32	175	159	140
33	187	170	151
34	200	182	162
35	213	195	173
36	227	208	185
37	241	221	198
38	256	235	211
39	271	249	224
40	286	264	238

Tabelle in Anlehnung an [Kaehler2002]

B Signifikanz-Tests

Die in dieser Studie angewendeten Hypothesentests wurden sowohl in *SPSS* als auch in *Microsoft Excel* durchgeführt. Damit die einzelnen Rechenschritte nachvollzogen werden können, sind diese im Folgenden in Tabellenform dargestellt.

- Friedman-Test
- ANOVA mit Messwiederholung
- Wilcoxon-Test
- t-Test für abhängige Stichproben

Friedman-Test auf die AV_{Acc}

Faktorstufen				Rangplätze der Faktorstufen			
Rot	CutAway	Stippling	noHT	Rot'	CutAway'	Stippling'	noHT'
83,67	92,35	86,09	60,06	2	4	3	1
47,68	59,5	49,25	45,73	2	4	3	1
69,48	74,62	79,7	42,29	2	3	4	1
82,89	94,58	85,32	81,33	2	4	3	1
68,12	84,96	77,65	53,65	2	4	3	1
72,8	87,7	76,96	55,72	2	4	3	1
29,29	52,73	35,6	36,92	1	4	2	3
40,47	68,26	57,38	38,75	2	4	3	1
71,86	86,15	76,85	48,55	2	4	3	1
42,2	33,86	36,2	37,74	4	1	2	3
84,76	89,13	84,83	76,94	2	4	3	1
71,1	71,34	64,23	55,05	3	4	2	1
55,61	74,49	57,42	49,4	2	4	3	1
48,61	60,36	52,84	46,53	2	4	3	1
79,84	85,26	79,77	56,3	3	4	2	1
77,03	90,42	73,97	14,51	3	4	2	1
61,91	62,7	58,86	42,33	3	4	2	1
74,82	89,38	75,9	54,14	2	4	3	1
65,43	82,68	65,49	40,64	2	4	3	1
78,73	90,67	76,53	72,13	3	4	2	1
76,95	84,03	74,82	61,92	3	4	2	1
84,53	96,06	85,29	57,3	2	4	3	1
80,2	91,02	85,84	59,14	2	4	3	1
80,53	80,69	78,19	46,07	3	4	2	1
82,71	95,95	85,33	71,63	2	4	3	1
53,43	73,99	48,34	29,26	3	4	2	1
80,18	85,7	78,32	38,01	3	4	2	1
50,61	62,91	47,12	29,51	3	4	2	1
79,99	92	86,9	59,19	2	4	3	1
49,58	59,8	48,58	35,99	3	4	2	1
32,37	55,14	37,32	35,45	1	4	3	2
84,46	91,47	80,3	71,23	3	4	2	1
81,07	93,87	91,86	76,74	2	4	3	1

n:	**28**	**Rangsumme T:**	**66**	**111**	**75**	**28**
p:	**4**	**Mittlerer Rang:**	**2,36**	**3,96**	**2,68**	**1,00**
df:	**3**					

$\chi^2_{krit} = \chi^2_{0,05;3}$	**7,81**
χ^2	**74,70**

p	0,0000

Die rot markierten Messwerte, sind die von der Datenanalyse ausgeschlossenen Ausreißer.

ANOVA mit Messwiederholung (1/2)

	Faktorstufen							SAQ$_{Ges}$			
Vpn	**UV$_2$**	**UV$_3$**	**UV$_1$**	**UV$_4$**	**τ**	**τ^2**		**(UV$_2$-MW)^2**	**(UV$_3$-MW)^2**	**(UV$_1$-MW)^2**	**(UV$_4$-MW)^2**
1	645,23	581,44	625,92	715,26	2567,85	6593853,62		8867,74	24950,94	12877,41	582,67
2	652,99	654,32	660,02	728,22	2695,55	7265989,80		7466,46	7238,38	6300,97	124,96
3	675,02	649,93	644,53	701,61	2671,09	7134721,79		4144,61	8004,64	9000,06	1427,98
4	769,56	756,93	747,31	849,08	3122,88	9752379,49		909,71	307,35	62,59	12030,00
5	737,87	672,40	729,08	777,44	2916,79	8507663,90		2,34	4488,82	106,47	1447,14
6	685,90	680,54	704,06	762,15	2832,65	8023906,02		2862,11	3464,34	1248,82	517,62
8	808,33	739,73	732,72	753,81	3034,59	9208736,47		4751,53	0,11	44,60	207,69
9	621,04	574,74	629,45	703,79	2529,02	6395942,16		14008,77	27112,47	12088,71	1267,98
11	671,53	624,05	662,32	741,37	2699,27	7286058,53		4606,16	13305,31	5941,12	3,89
12	905,90	907,89	890,83	1011,91	3716,53	13812595,24		27722,70	28389,33	22931,45	74262,43
13	709,21	722,29	718,10	772,14	2921,74	8536564,63		911,36	292,71	453,63	1072,00
14	839,42	780,68	805,10	857,99	3283,19	10779336,58		10004,27	1704,15	4316,67	14063,91
15	655,11	620,61	652,64	688,74	2617,10	6849212,41		7104,58	14110,75	7527,07	2566,30
16	764,85	784,54	798,74	1009,46	3357,59	11273410,61		647,77	2037,74	3521,39	72933,13
17	699,77	690,29	681,09	729,98	2801,13	7846329,28		1570,43	2411,66	3399,90	88,71
18	843,90	751,89	788,33	901,37	3285,49	10794444,54		10920,53	156,03	2394,28	26234,71
19	797,91	775,47	843,28	881,64	3298,30	10878782,89		3423,58	1301,14	10791,33	20232,60
20	745,41	724,03	769,83	793,94	3033,21	9200362,90		36,14	236,20	926,07	2974,76
21	716,43	733,74	753,64	782,19	2986,00	8916196,00		527,56	32,02	202,82	1831,10
22	691,81	710,98	728,68	829,01	2960,48	8764441,83		2264,68	807,62	114,89	8030,19
23	614,20	573,68	583,71	677,31	2448,90	5997111,21		15674,70	27462,67	24238,96	3855,00
24	764,31	797,62	794,93	872,13	3228,99	10426376,42		620,57	3389,72	3083,73	17617,61
25	645,50	641,97	656,29	808,07	2751,83	7572568,35		8816,96	9492,34	6907,05	4715,75
26	698,52	697,69	757,53	716,10	2869,84	8235981,63		1671,06	1739,61	328,75	542,83
27	736,36	766,64	782,33	837,76	3123,09	9753691,15		9,23	742,09	1843,10	9674,95
29	701,18	663,35	676,14	806,22	2846,89	8104782,67		1460,67	5783,40	4001,66	4465,09
32	698,04	670,14	692,13	741,12	2801,43	7848010,04		1710,54	4796,76	2234,33	2,96
33	816,85	795,75	845,67	952,96	3411,23	11636490,11		5998,71	3175,47	11293,60	45608,45
MW:	**725,43**	**705,12**	**726,94**	**800,10**	**82812,65**	**247395940,28**	**$\sum$**	**148715,45**	**196933,79**	**158181,41**	**328382,41**
	739,40							**832213,06**			

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

ANOVA mit Messwiederholung (2/2)

S:	61231562,5	#Vpn:	28	Bedingungen:	4

	$(UV_2'-MW)^2$	$(UV_3'-MW)^2$	$(UV_1'-MW)^2$	$(UV_4'-MW)^2$
	195,01	1175,10	155,15	3684,52
SAQ_{Zw}	145873,93			
df	3			

	QS	df	MQ	F
SAQ_{Zw}	**145873,93**	**3,00**	**48624,64**	**57,15**
SAQ_e	**617422,57**	**27,00**	**22867,50**	
SAQ_{Res}	**68916,55**	**81,00**	**850,82**	
SAQ_{Ges}	**832213,06**	**108,00**		

Wilcoxon-Test auf die Faktorstufen UV_4 und UV_1

UV_4:= Keine Hervorhebung

UV_1:= Stippling

Vpn	UV_1	UV_4	d	|d|	Rang	Rang > 0	Rang < 0
1	86,09	60,06	26,03	26,03	20	20	0
2	49,25	45,73	3,52	3,52	1	1	0
3	79,7	42,29	37,41	37,41	26	26	0
4	85,32	81,33	3,99	3,99	2	2	0
5	77,65	53,65	24	24	18	18	0
6	76,96	55,72	21,24	21,24	15	15	0
8	57,38	38,75	18,63	18,63	13	13	0
9	76,85	48,55	28,3	28,3	24	24	0
11	84,83	76,94	7,89	7,89	5	5	0
12	64,23	55,05	9,18	9,18	8	8	0
13	57,42	49,4	8,02	8,02	6	6	0
14	52,84	46,53	6,31	6,31	4	4	0
15	79,77	56,3	23,47	23,47	17	17	0
16	73,97	14,51	59,46	59,46	28	28	0
17	58,86	42,33	16,53	16,53	12	12	0
18	75,9	54,14	21,76	21,76	16	16	0
19	65,49	40,64	24,85	24,85	19	19	0
20	76,53	72,13	4,4	4,4	3	3	0
21	74,82	61,92	12,9	12,9	9	9	0
22	85,29	57,3	27,99	27,99	23	23	0
23	85,84	59,14	26,7	26,7	21	21	0
24	78,19	46,07	32,12	32,12	25	25	0
25	85,33	71,63	13,7	13,7	10	10	0
26	48,34	29,26	19,08	19,08	14	14	0
27	78,32	38,01	40,31	40,31	27	27	0
29	86,9	59,19	27,71	27,71	22	22	0
32	80,3	71,23	9,07	9,07	7	7	0
33	91,86	76,74	15,12	15,12	11	11	0
					∑	**406**	**0**
					$Rang_T$	**0**	
					z	4,62259895	
				1-seitig	p	0,0000	
				2-seitig	p	0,0000	

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

Wilcoxon-Test auf die Faktorstufen UV_4 und UV_2

UV_4:= Keine Hervorhebung

UV_2:= Rote Einfärbung

Vpn	UV_2	UV_4	d	|d|	Rang	Rang > 0	Rang < 0
1	83,67	60,06	23,61	23,61	21	21	0
2	47,68	45,73	1,95	1,95	3	3	0
3	69,48	42,29	27,19	27,19	24	24	0
4	82,89	81,33	1,56	1,56	1	1	0
5	68,12	53,65	14,47	14,47	11	11	0
6	72,8	55,72	17,08	17,08	14	14	0
8	40,47	38,75	1,72	1,72	2	2	0
9	71,86	48,55	23,31	23,31	19	19	0
11	84,76	76,94	7,82	7,82	8	8	0
12	71,1	55,05	16,05	16,05	13	13	0
13	55,61	49,4	6,21	6,21	6	6	0
14	48,61	46,53	2,08	2,08	4	4	0
15	79,84	56,3	23,54	23,54	20	20	0
16	77,03	14,51	62,52	62,52	28	28	0
17	61,91	42,33	19,58	19,58	15	15	0
18	74,82	54,14	20,68	20,68	16	16	0
19	65,43	40,64	24,79	24,79	23	23	0
20	78,73	72,13	6,6	6,6	7	7	0
21	76,95	61,92	15,03	15,03	12	12	0
22	84,53	57,3	27,23	27,23	25	25	0
23	80,2	59,14	21,06	21,06	18	18	0
24	80,53	46,07	34,46	34,46	26	26	0
25	82,71	71,63	11,08	11,08	9	9	0
26	53,43	29,26	24,17	24,17	22	22	0
27	80,18	38,01	42,17	42,17	27	27	0
29	79,99	59,19	20,8	20,8	17	17	0
32	84,46	71,23	13,23	13,23	10	10	0
33	81,07	76,74	4,33	4,33	5	5	0
					∑	**406**	**0**

	$Rang_T$	**0**
	z	4,62259895
1-seitig	p	0,0000
2-seitig	p	0,0000

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

Wilcoxon-Test auf die Faktorstufen UV_4 und UV_3

UV_4:= Keine Hervorhebung

UV_3:= CutAway

Vpn	UV_3	UV_4	d	|d|	Rang	Rang > 0	Rang < 0
1	92,35	60,06	32,29	32,29	18	18	0
2	59,5	45,73	13,77	13,77	3	3	0
3	74,62	42,29	32,33	32,33	19	19	0
4	94,58	81,33	13,25	13,25	2	2	0
5	84,96	53,65	31,31	31,31	15	15	0
6	87,7	55,72	31,98	31,98	17	17	0
8	68,26	38,75	29,51	29,51	14	14	0
9	86,15	48,55	37,6	37,6	23	23	0
11	89,13	76,94	12,19	12,19	1	1	0
12	71,34	55,05	16,29	16,29	5	5	0
13	74,49	49,4	25,09	25,09	12	12	0
14	60,36	46,53	13,83	13,83	4	4	0
15	85,26	56,3	28,96	28,96	13	13	0
16	90,42	14,51	75,91	75,91	28	28	0
17	62,7	42,33	20,37	20,37	9	9	0
18	89,38	54,14	35,24	35,24	22	22	0
19	82,68	40,64	42,04	42,04	25	25	0
20	90,67	72,13	18,54	18,54	7	7	0
21	84,03	61,92	22,11	22,11	10	10	0
22	96,06	57,3	38,76	38,76	24	24	0
23	91,02	59,14	31,88	31,88	16	16	0
24	80,69	46,07	34,62	34,62	21	21	0
25	95,95	71,63	24,32	24,32	11	11	0
26	73,99	29,26	44,73	44,73	26	26	0
27	85,7	38,01	47,69	47,69	27	27	0
29	92	59,19	32,81	32,81	20	20	0
32	91,47	71,23	20,24	20,24	8	8	0
33	93,87	76,74	17,13	17,13	6	6	0
					∑	**406**	**0**
					$Rang_T$	**0**	
					z	4,622599	
				1-seitig	p	0,0000	
				2-seitig	p	0,0000	

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

Wilcoxon-Test auf die Faktorstufen UV_1 und UV_3

UV_1:= Stippling

UV_3:= CutAway

Vpn	UV_1	UV_3	d	|d|	Rang	Rang > 0	Rang < 0
1	86,09	92,35	-6,26	6,26	9	0	9
2	49,25	59,5	-10,25	10,25	17	0	17
3	79,7	74,62	5,08	5,08	5	5	0
4	85,32	94,58	-9,26	9,26	15	0	15
5	77,65	84,96	-7,31	7,31	11	0	11
6	76,96	87,7	-10,74	10,74	19	0	19
8	57,38	68,26	-10,88	10,88	21	0	21
9	76,85	86,15	-9,3	9,3	16	0	16
11	84,83	89,13	-4,3	4,3	4	0	4
12	64,23	71,34	-7,11	7,11	10	0	10
13	57,42	74,49	-17,07	17,07	26	0	26
14	52,84	60,36	-7,52	7,52	13	0	13
15	79,77	85,26	-5,49	5,49	8	0	8
16	73,97	90,42	-16,45	16,45	25	0	25
17	58,86	62,7	-3,84	3,84	3	0	3
18	75,9	89,38	-13,48	13,48	23	0	23
19	65,49	82,68	-17,19	17,19	27	0	27
20	76,53	90,67	-14,14	14,14	24	0	24
21	74,82	84,03	-9,21	9,21	14	0	14
22	85,29	96,06	-10,77	10,77	20	0	20
23	85,84	91,02	-5,18	5,18	7	0	7
24	78,19	80,69	-2,5	2,5	2	0	2
25	85,33	95,95	-10,62	10,62	18	0	18
26	48,34	73,99	-25,65	25,65	28	0	28
27	78,32	85,7	-7,38	7,38	12	0	12
29	86,9	92	-5,1	5,1	6	0	6
32	80,3	91,47	-11,17	11,17	22	0	22
33	91,86	93,87	-2,01	2,01	1	0	1
					∑	**5**	**401**
					$Rang_T$	**0**	
					z	4,5087418	
				1-seitig	p	0,0000	
				2-seitig	p	0,0000	

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

Wilcoxon-Test auf die Faktorstufen UV_1 und UV_2

UV_1:= Stippling

UV_2:= Rote Einfärbung

Vpn	UV_1	UV_2	d	|d|	Rang	Rang > 0	Rang < 0
1	86,09	83,67	2,42	2,42	12	12	0
2	49,25	47,68	1,57	1,57	6	6	0
3	79,7	69,48	10,22	10,22	26	26	0
4	85,32	82,89	2,43	2,43	13	13	0
5	77,65	68,12	9,53	9,53	25	25	0
6	76,96	72,8	4,16	4,16	17	17	0
8	57,38	40,47	16,91	16,91	28	28	0
9	76,85	71,86	4,99	4,99	20	20	0
11	84,83	84,76	0,07	0,07	2	2	0
12	64,23	71,1	-6,87	6,87	23	0	23
13	57,42	55,61	1,81	1,81	7	7	0
14	52,84	48,61	4,23	4,23	19	19	0
15	79,77	79,84	-0,07	0,07	3	0	3
16	73,97	77,03	-3,06	3,06	16	0	16
17	58,86	61,91	-3,05	3,05	15	0	15
18	75,9	74,82	1,08	1,08	5	5	0
19	65,49	65,43	0,06	0,06	1	1	0
20	76,53	78,73	-2,2	2,2	10	0	10
21	74,82	76,95	-2,13	2,13	9	0	9
22	85,29	84,53	0,76	0,76	4	4	0
23	85,84	80,2	5,64	5,64	22	22	0
24	78,19	80,53	-2,34	2,34	11	0	11
25	85,33	82,71	2,62	2,62	14	14	0
26	48,34	53,43	-5,09	5,09	21	0	21
27	78,32	80,18	-1,86	1,86	8	0	8
29	86,9	79,99	6,91	6,91	24	24	0
32	80,3	84,46	-4,16	4,16	17	0	17
33	91,86	81,07	10,79	10,79	27	27	0
					∑	**272**	**133**
					$Rang_T$	**0**	
					z	1,59399964	
				1-seitig	p	0,0555	
				2-seitig	p	0,1109	

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

Wilcoxon-Test auf die Faktorstufen UV_3 und UV_2

UV_3:= CutAway

UV_2:= Rote Einfärbung

Vpn	UV_2	UV_3	d	|d|	Rang	Rang > 0	Rang < 0
1	83,67	92,35	-8,68	8,68	10	0	10
2	47,68	59,5	-11,82	11,82	15	0	15
3	69,48	74,62	-5,14	5,14	5	0	5
4	82,89	94,58	-11,69	11,69	13	0	13
5	68,12	84,96	-16,84	16,84	24	0	24
6	72,8	87,7	-14,9	14,9	23	0	23
8	40,47	68,26	-27,79	27,79	28	0	28
9	71,86	86,15	-14,29	14,29	21	0	21
11	84,76	89,13	-4,37	4,37	4	0	4
12	71,1	71,34	-0,24	0,24	2	0	2
13	55,61	74,49	-18,88	18,88	26	0	26
14	48,61	60,36	-11,75	11,75	14	0	14
15	79,84	85,26	-5,42	5,42	6	0	6
16	77,03	90,42	-13,39	13,39	20	0	20
17	61,91	62,7	-0,79	0,79	3	0	3
18	74,82	89,38	-14,56	14,56	22	0	22
19	65,43	82,68	-17,25	17,25	25	0	25
20	78,73	90,67	-11,94	11,94	16	0	16
21	76,95	84,03	-7,08	7,08	9	0	9
22	84,53	96,06	-11,53	11,53	12	0	12
23	80,2	91,02	-10,82	10,82	11	0	11
24	80,53	80,69	-0,16	0,16	1	0	1
25	82,71	95,95	-13,24	13,24	19	0	19
26	53,43	73,99	-20,56	20,56	27	0	27
27	80,18	85,7	-5,52	5,52	7	0	7
29	79,99	92	-12,01	12,01	17	0	17
32	84,46	91,47	-7,01	7,01	8	0	8
33	81,07	93,87	-12,8	12,8	18	0	18
					$\sum$	**0**	**406**

	$Rang_T$	**0**
	z	4,62259895
1-seitig	p	0,0000
2-seitig	p	0,0000

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

T-Test auf die Faktorstufen UV_1,UV_2,UV_3,UV_4

UV_1:= Stippling UV2:= Rote Einfärbung
UV_3:= CutAway UV4:= Keine Hervorhebung

Vpn	UV_4	UV_1	d
1	715,26	625,92	89,34
2	728,22	660,02	68,2
3	701,61	644,53	57,08
4	849,08	747,31	101,77
5	777,44	729,08	48,36
6	762,15	704,06	58,09
8	753,81	732,72	21,09
9	703,79	629,45	74,34
11	741,37	662,32	79,05
12	1011,9	890,83	121,08
13	772,14	718,1	54,04
14	857,99	805,1	52,89
15	688,74	652,64	36,1
16	1009,5	798,74	210,72
17	729,98	681,09	48,89
18	901,37	788,33	113,04
19	881,64	843,28	38,36
20	793,94	769,83	24,11
21	782,19	753,64	28,55
22	829,01	728,68	100,33
23	677,31	583,71	93,6
24	872,13	794,93	77,2
25	808,07	656,29	151,78
26	716,1	757,53	-41,43
27	837,76	782,33	55,43
29	806,22	676,14	130,08
32	741,12	692,13	48,99
33	952,96	845,67	107,29
#Vpn	28	∑	2048,4
		σ	47,879
		t	8,085
	1-seitig	p	0,0000
	2-seitig	p	0,0000

UV_4	UV_2	d
715,26	645,23	70,03
728,22	652,99	75,23
701,61	675,02	26,59
849,08	769,56	79,52
777,44	737,87	39,57
762,15	685,9	76,25
753,81	808,33	-54,52
703,79	621,04	82,75
741,37	671,53	69,84
1011,9	905,9	106,01
772,14	709,21	62,93
857,99	839,42	18,57
688,74	655,11	33,63
1009,5	764,85	244,61
729,98	699,77	30,21
901,37	843,9	57,47
881,64	797,91	83,73
793,94	745,41	48,53
782,19	716,43	65,76
829,01	691,81	137,2
677,31	614,2	63,11
872,13	764,31	107,82
808,07	645,5	162,57
716,1	698,52	17,58
837,76	736,36	101,4
806,22	701,18	105,04
741,12	698,04	43,08
952,96	816,85	136,11
	∑	2090,6
	σ	54,724
	t	7,2197
1-seitig	p	0,0000
2-seitig	p	0,0000

UV_4	UV_3	d
715,26	581,44	133,82
728,22	654,32	73,9
701,61	649,93	51,68
849,08	756,93	92,15
777,44	672,4	105,04
762,15	680,54	81,61
753,81	739,73	14,08
703,79	574,74	129,05
741,37	624,05	117,32
1011,9	907,89	104,02
772,14	722,29	49,85
857,99	780,68	77,31
688,74	620,61	68,13
1009,5	784,54	224,92
729,98	690,29	39,69
901,37	751,89	149,48
881,64	775,47	106,17
793,94	724,03	69,91
782,19	733,74	48,45
829,01	710,98	118,03
677,31	573,68	103,63
872,13	797,62	74,51
808,07	641,97	166,1
716,1	697,69	18,41
837,76	766,64	71,12
806,22	663,35	142,87
741,12	670,14	70,98
952,96	795,75	157,21
	∑	2659,4
	σ	47,36
	t	10,612
1-seitig	p	0,0000
2-seitig	p	0,0000

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

p kann mit Hilfe der Microsoft Excel Funktion TVERT(t;n-1) ermittelt werden.

T-Test auf die Faktorstufen UV_1,UV_2,UV_3

UV_1:= Stippling

UV_2:= Rote Einfärbung

UV_3:= CutAway

Vpn	UV_1	UV_3	d
1	625,92	581,44	44,48
2	660,02	654,32	5,7
3	644,53	649,93	-5,4
4	747,31	756,93	-9,62
5	729,08	672,4	56,68
6	704,06	680,54	23,52
8	732,72	739,73	-7,01
9	629,45	574,74	54,71
11	662,32	624,05	38,27
12	890,83	907,89	-17,06
13	718,1	722,29	-4,19
14	805,1	780,68	24,42
15	652,64	620,61	32,03
16	798,74	784,54	14,2
17	681,09	690,29	-9,2
18	788,33	751,89	36,44
19	843,28	775,47	67,81
20	769,83	724,03	45,8
21	753,64	733,74	19,9
22	728,68	710,98	17,7
23	583,71	573,68	10,03
24	794,93	797,62	-2,69
25	656,29	641,97	14,32
26	757,53	697,69	59,84
27	782,33	766,64	15,69
29	676,14	663,35	12,79
32	692,13	670,14	21,99
33	845,67	795,75	49,92
#Vpn	28	∑	611,07
		σ	23,804
		t	4,8514
	1-seitig	p	0,0000
	2-seitig	p	0,0000

UV_1	UV_2	d
625,92	645,23	-19,31
660,02	652,99	7,03
644,53	675,02	-30,49
747,31	769,56	-22,25
729,08	737,87	-8,79
704,06	685,9	18,16
732,72	808,33	-75,61
629,45	621,04	8,41
662,32	671,53	-9,21
890,83	905,9	-15,07
718,1	709,21	8,89
805,1	839,42	-34,32
652,64	655,11	-2,47
798,74	764,85	33,89
681,09	699,77	-18,68
788,33	843,9	-55,57
843,28	797,91	45,37
769,83	745,41	24,42
753,64	716,43	37,21
728,68	691,81	36,87
583,71	614,2	-30,49
794,93	764,31	30,62
656,29	645,5	10,79
757,53	698,52	59,01
782,33	736,36	45,97
676,14	701,18	-25,04
692,13	698,04	-5,91
845,67	816,85	28,82
	∑	42,25
	σ	32,551
	t	0,2453
1-seitig	p	0,4040
2-seitig	p	0,8081

UV_2	UV_3	d
645,23	581,44	63,79
652,99	654,32	-1,33
675,02	649,93	25,09
769,56	756,93	12,63
737,87	672,4	65,47
685,9	680,54	5,36
808,33	739,73	68,6
621,04	574,74	46,3
671,53	624,05	47,48
905,9	907,89	-1,99
709,21	722,29	-13,08
839,42	780,68	58,74
655,11	620,61	34,5
764,85	784,54	-19,69
699,77	690,29	9,48
843,9	751,89	92,01
797,91	775,47	22,44
745,41	724,03	21,38
716,43	733,74	-17,31
691,81	710,98	-19,17
614,2	573,68	40,52
764,31	797,62	-33,31
645,5	641,97	3,53
698,52	697,69	0,83
736,36	766,64	-30,28
701,18	663,35	37,83
698,04	670,14	27,9
816,85	795,75	21,1
	∑	568,82
	σ	32,459
	t	3,3118
1-seitig	p	0,0013
2-seitig	p	0,0026

Die Vpn 7, 10, 28, 30 und 31 sind von der Analyse ausgeschlossen.

p kann mit Hilfe der Microsoft Excel Funktion TVERT(t;n-1) ermittelt werden.

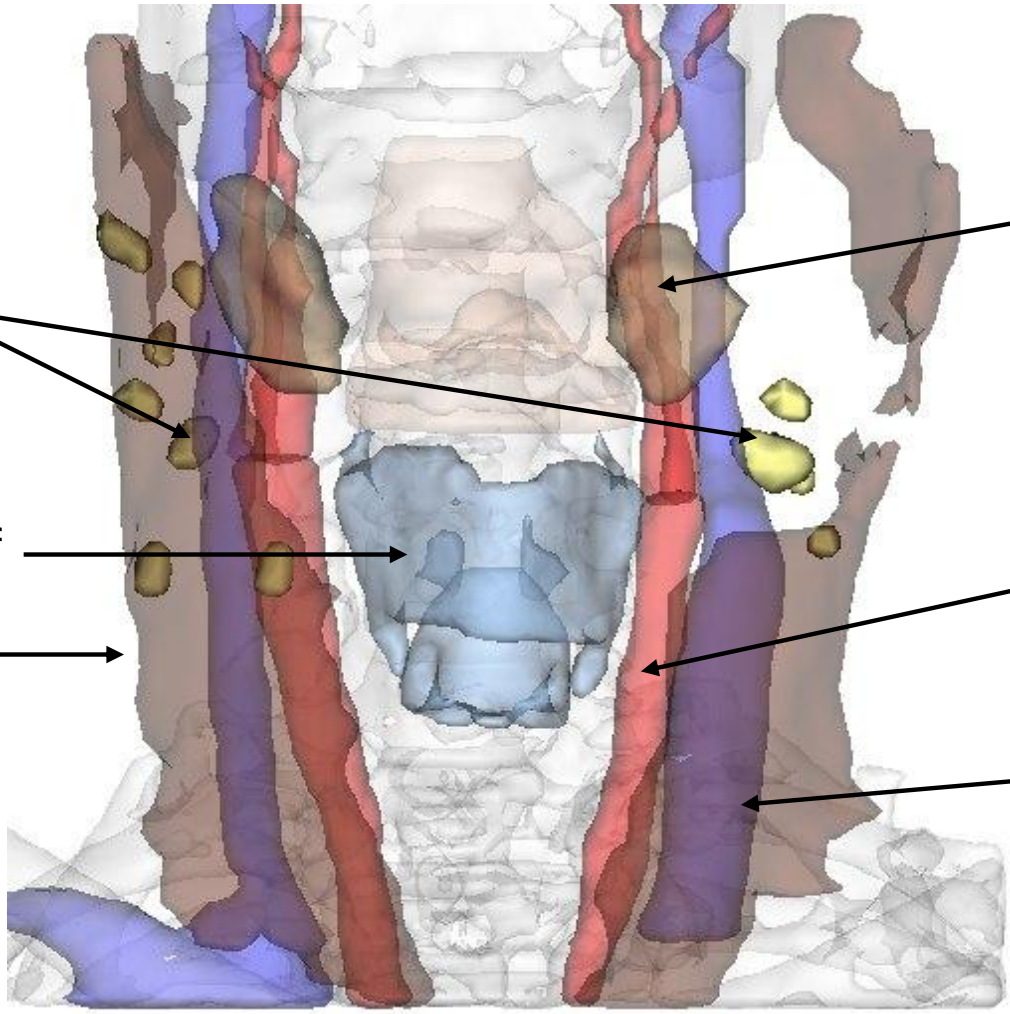
C Instruktion

Experiment

zur Bewertung von Hervorhebungstechniken in der medizinischen Visualisierung.

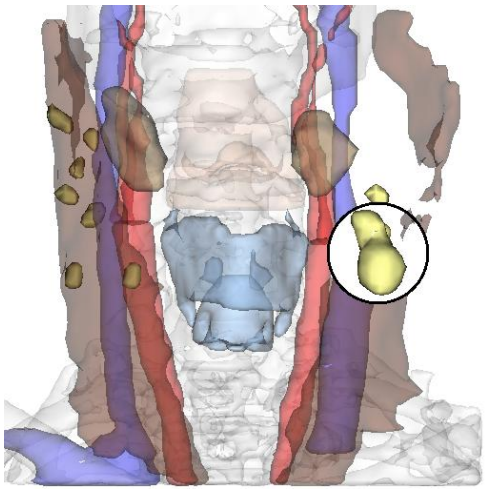
- Im folgenden Experiment werden Ihnen 2D-Bilder von Hals-Datensätzen präsentiert
- Ihre Aufgabe besteht darin zu entscheiden, ob im Bild ein vergrößerter Lymphknoten zu sehen ist oder nicht

Anatomie des Halses

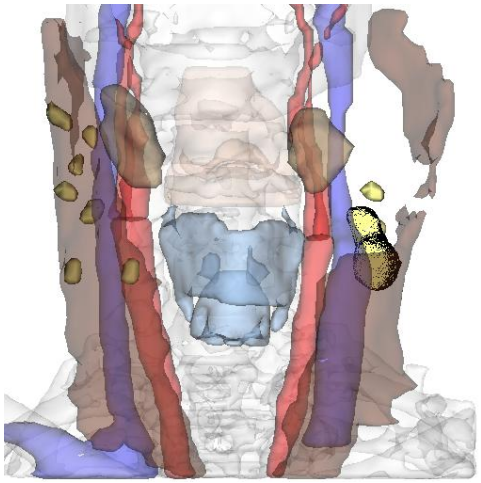


Bitte prägen Sie sich nur das Erscheinungsbild der Lymphknoten ein.

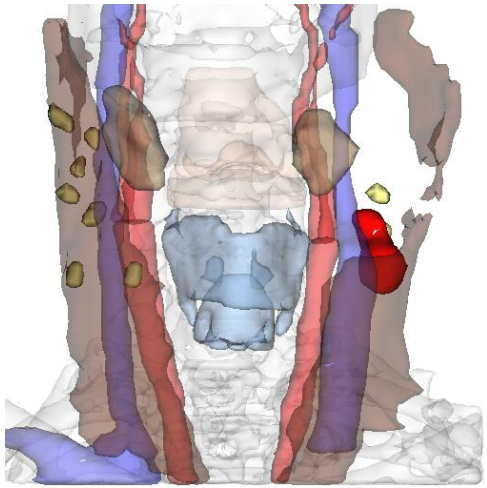
Folgende Hervorhebungstechniken für Lymphknoten können vorkommen:



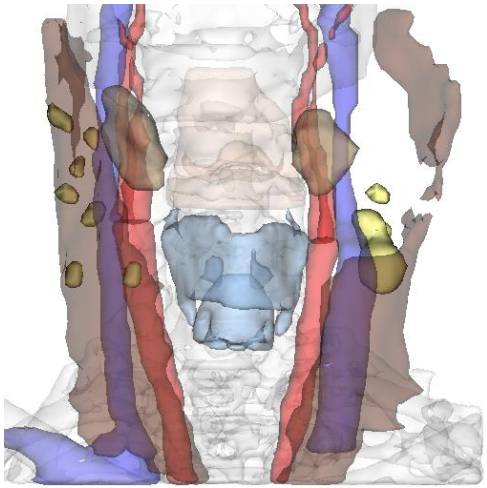
CutAway



Stippling



Rote Hervorhebung

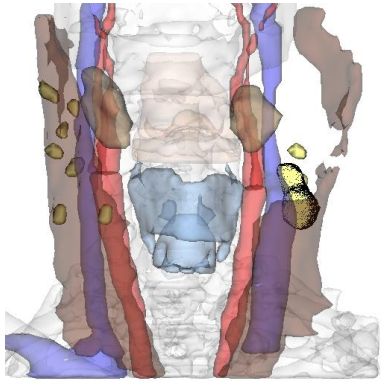


Keine Hervorhebung

Es gibt zwei verschiedene Arten von Bildern. Alle Bilder enthalten Lymphknoten.

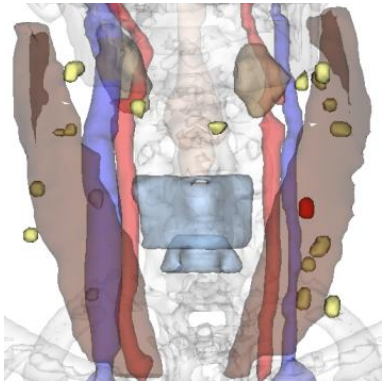
Variante 1: Das Bild enthält einen vergrößerten Lymphknoten.

Dieser ist entweder hervorgehoben oder auch nicht.



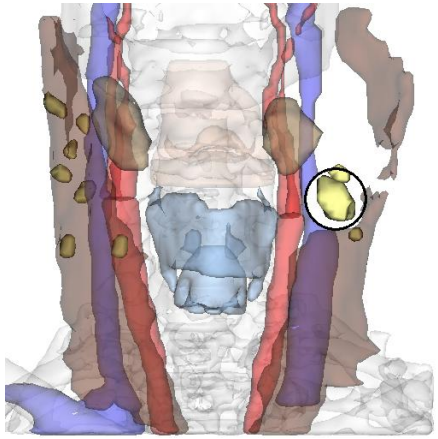
Variante 2: Das Bild enthält <u>keinen</u> vergrößerten Lymphknoten.

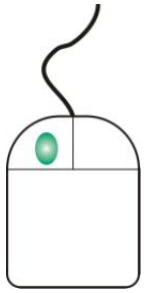
Eines der kleinen normalen Lymphknoten ist entweder hervorgehoben oder auch nicht.



Drücken Sie die **linke Maustaste**, wenn ein **vergrößerter** Lymphknoten zu sehen ist!

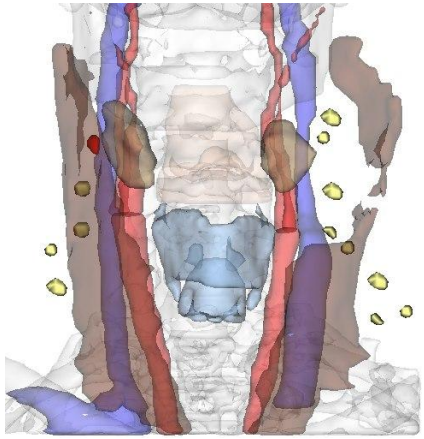
Variante 1:

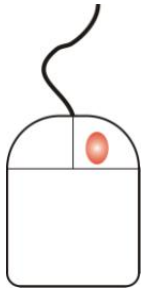




Drücken Sie die **rechte Maustaste**, wenn **<u>kein</u> vergrößerter** Lymphknoten zu sehen ist!

Variante 2:





Bitte beachten Sie, dass hervorgehobene Lymphknoten nicht zwingend den vergrößerten Lymphknoten kennzeichnen!

Jedes Bild wird nur sehr kurz angezeigt, entscheiden Sie sich deswegen schnell!

Zunächst kommt ein kleiner Testdurchlauf.

Viel Spaß!

D Fragebogen

Sehr geehrte Damen und Herren,

im Namen der Arbeitsgruppe Visualisierung der Otto-von-Guericke-Universität Magdeburg möchte ich mich bei Ihnen für ihre Teilnahme an diesem Experiment bedanken.

Die Bilder die sie bisher gesehen haben, sind 3D-Renderings von Hals-Datensätzen.

Ihre Aufgabe bestand darin vergrößerte Lymphknoten in den Ihnen präsentierten Bildern zu detektieren. Die verwendeten Hervorhebungstechniken für die Lymphknoten, sollen es dem Anwender erleichtern pathologische Strukturen schnellst möglich zu erkennen. Welche Hervorhebungstechnik für Lernsysteme oder die Diagnostik am besten geeignet ist soll aus diesem Experiment hervorgehen.

Abschließend bitte ich Sie darum, die auf den nächsten Seiten gestellten Fragen für die statistische Auswertung zu beantworten.

Vielen Dank für Ihre Teilnahme

Friederike Adler

Zunächst sollen Sie jedoch ein paar Fragen beantworten, die für die Auswertung der Experiment-Ergebnisse von Interesse sind.

Ihre Experiment ID?

Alter?

Geschlecht?

Haben Sie eine Sehschwäche?

Wenn ja, welche?

Sind Sie Links- oder Rechtshänder?

Tätigkeit?

Haben Sie eine medizinische Fachausbildung?

Wenn ja, welche?

Wie oft haben Sie mit medizinischen 3D-Visualisierungen zu tun?

Wie schätzen Sie den generellen Einsatz von Hervorhebungstechniken in medizinischen Visualisierungen ein?

Welche der vorgestellten Hervorhebungstechniken gefällt Ihnen am besten?

Welche der vorgestellten Hervorhebungstechniken finden sie am besten, zur Hervorhebung von pathologischen Lymphknoten, geeignet?

Wie schätzen sie den Gebrauch der folgenden Techniken für die Hervorhebung von vergrößerten Lymphknoten ein?
(gar nicht zu gebrauchen(--) bis sehr gut geeignet(++))

	--	-	0	+	++
Keine Hervorhebungstechnik?	☐	☐	☐	☐	☐
Stippling (gepunktet)?	☐	☐	☐	☐	☐
Rote Hervorhebung?	☐	☐	☐	☐	☐
CutAway (Kreis)?	☐	☐	☐	☐	☐

Bitte begründen Sie ihre Entscheidung:

Welche Strukturen zur Detektion des vergrößerten Lymphknotens empfanden Sie als störend?(nicht störend(--) bis sehr störend(++))

	--	-	0	+	++
Speicheldrüsen(braun)	☐	☐	☐	☐	☐
Gefäße(blau,rot)	☐	☐	☐	☐	☐
Muskeln(braun)	☐	☐	☐	☐	☐
Kehlkopf(hellblau)	☐	☐	☐	☐	☐
Knochen(grau)	☐	☐	☐	☐	☐
Andere Lymphknoten(gelb)	☐	☐	☐	☐	☐