Ist das Global Listening reliabel?

Studie zur Intrarater- und Interrater-Reliabilität des Global Listenings

Master Thesis zur Erlangung des Grades

Master of Science in Osteopathie

an der **Donauuniversität Krems** –

Zentrum für chin. Medizin & Komplementärmedizin

niedergelegt an der

Wiener Schule für Osteopathie

von Margit Rittler

Wien, Dezember 2010

Eidesstattliche Erklärung

Hiermit versichere ich, die vorgelegte Masterthese selbständig verfasst zu haben.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer übernommen wurden, wurden als solche gekennzeichnet. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit genützt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt weder im In- noch Ausland einer anderen Prüfungsbehörde vorgelegen.

Diese Arbeit stimmt mit der von dem/der Gutachter/in beurteilten Arbeit überein.

1.12.2010 _____

Danksagung

Ich danke Herrn Dr. Gebhard Woisetschläger für die Unterstützung bei der statistischen Auswertung der Ergebnisse, den Probanden, Testern und Assistenten für ihre Teilnahme an der Studie und meiner Familie für ihre vielfältige Unterstützung.

Margit Rittler

Abstract

Schlüsselwörter: Global Listening, Intrarater-Reliabilität, Interrater-Reliabilität, Erfahrung

Das Global Listening ist ein Faszientest, bei dem ein erster Eindruck über im Körper vorhandene Restriktionen gewonnen wird. In dieser Studie wird seine Intrarater- und Interrater-Reliabilität untersucht.

Nach Reflexion der – in der Literatur nicht einheitlich – beschriebenen Durchführung des Tests wurde ein standardisierter Ablauf für die Studie bestimmt und sieben Interpretationsmöglichkeiten für die Tester festgelegt. Weiters wurden anhand von bisherigen Studien Qualitätskriterien für die Durchführung der Datengewinnung erarbeitet. Sechs Tester testeten 18 symptomatische Probanden. Die ersten zehn Probanden wurden noch zwei weitere Male von jedem Tester getestet. Tester waren Osteopathen, die ihre Diplomprüfung im September 2009 abgelegt hatten. Ein Tester verwendet das Global Listening in der Praxis nicht, die anderen regelmäßig. Die Tester waren gegenüber den Symptomen der Probanden und den Ergebnissen der anderen Osteopathen blindiert und testeten mit verbundenen Augen. Ein Training wurde eine Woche vor der Datengewinnung durchgeführt. Die Probanden wussten über das Ziel des Tests Bescheid, nicht aber über die den Testern zur Verfügung stehenden Interpretationsmöglichkeiten. Die Auswertung erfolgte mittels Cohens Kappa. Die Ergebnisse wurden auch in Bezug auf eine Rechts-Links-Körperachse und eine anterior-posteriore Körperachse ausgewertet, was eine Reduktion auf drei Interpretationsmöglichkeiten bedeutet. Es konnte weder eine Interrater- noch eine Intrarater-Reliabilität nachgewiesen werden. Die Ergebnisse bewegen sich im Bereich zufälliger Übereinstimmung. Auffällig war, dass der einzige Tester, der das Global Listening in der Praxis nicht anwendet, das beste Ergebnis erreichte. Dieses befindet sich jedoch auch nur im Bereich schwacher Reliabilität und hebt sich somit nicht entscheidend von den Ergebnissen der anderen Tester ab.

Abstract

Keywords: intraobserver reliability, interobserver reliability, Global Listening, experience

The Global Listening focuses on the fascia to get a first impression of restrictions in the body. In this study intraobserver and interobserver reliability of the Global Listening are investigated.

After reviewing the literature a standardised procedure for performing and scoring the test was defined. The outcome had to be classified within seven categories. Six testers examined 18 symptomatic subjects. The first ten of them were examined two more times by each tester. Testers were osteopaths, who passed their final exam in September 2009. One tester never uses the Global Listening in his praxis, the others do regularly. Testers were blinded to the subject´s symptoms and the outcome of the other testers. A training was held one week before data collection. Subjects were informed about the aim of the test, but not about the possible interpretations. For calculation purposes Cohen´s Kappa was used. Additionally, results were evaluated in reference to an anterior-posterior axis and to a right-left axis through the body. This means a reduction of the possible answers to three. Neither an intraobserver nor an interobserver reliability has been found. Results were in a range of random agreement. Remarkably, the tester, who never uses the Global Listening, achieved the best intraobserver reliability.Still, his mean result shows poor agreement and thus, does not differ significantly from the other tester´sresults.

Inhalt

1 Einleitung

Das Global Listening erschien mir in meiner Osteopathieausbildung als einer der wesentlichsten Tests zum Finden der zu behandelnden Läsion. Gleichzeitig fragte ich mich oft, ob das Gespürte Tatsache oder Einbildung ist. Diese Frage stellte sich für mich bei diesem Test besonders, da er nicht auf der Palpation klar beschriebener anatomischer Strukturen beruht, sondern auf im klinischen Alltag Gespürtem und einem auf anatomischen Tatsachen aufgebauten Erklärungsmodell.

Eine wissenschaftliche Herangehensweise an diese Frage bietet die Durchführung einer Reliabilitätsstudie. Eine Reliabilitätsstudie misst den Grad der Genauigkeit, mit der ein Test das geprüfte Merkmal erfasst (Bortz und Döring, 2006).

Im Bereich struktureller Tests wurden bereits zahlreiche Reliabilitätsstudien durchgeführt. Ein Gewebeecoute wurde meines Wissens nur in einer Studie, nämlich der Masterthese von Podlesnic (2006) zur Reliabilität des Local Listenings, untersucht. Forschungsbedarf in diesem Bereich ist also gegeben.

In einigen Studien wurde untersucht, ob die Erfahrung der Tester sich auf das Ergebnis auswirkt, das heißt also, ob erfahrene Tester ein besseres Ergebnis als unerfahrene Tester erreichen. Zu diesem Thema fand ich vier Studien (Harvey und Byfield, 1991; Jensen et al. 1993; Kmita und Lucas, 2008, Podlesnic, 2006). Ich habe mich entschlossen auch dieser Frage nachzugehen.

Generell wird in Metaanalysen (Gemmel und Miller, 2005; Hartmann und Norton, 2002; Hestboek und Leboeuf-Yde, 2000; Van Trijffel et al., 2005) die Qualität bisheriger Studien kritisiert, weshalb die Durchführung weiterer, methodisch gut durchgeführter Studien wichtig ist.

Ziel einer Reliabilitätsstudie ist es, verlässliche Tests für den klinischen Alltag zu finden. Weiters kann es möglich sein, aus der Reflexion einer Studie Adaptationen zur Verbesserung eines Tests zu gewinnen.Bei der Durchführung einer Studie können aber auch Besonderheiten osteopathischer Tests bewusst werden. Es ist zu hinterfragen, ob die Studiendurchführung beziehungsweise das Design an sich diesen Besonderheiten gerecht wird. Eine Verbesserung der Forschungsmethode, vielleichtauch die Entwicklung neuer Forschungsansätze, kann eine Folge sein. Letztes Ziel der Forschung ist eine Verbesserung der Testung der Patienten, aber

auch ein besseres Verständnis der beeinflussenden und limitierenden Faktoren eines Tests.

Die in dieser Studie beantworteten Forschungsfragen sind:

1) Kommen verschiedene Tester, die am selben Probanden ein Global Listening durchführen, zum gleichen Ergebnis?
2) Kommt ein und derselbe Tester, der am selben Probanden mehrmals ein Global Listening durchführt, bei jeder Testung zum gleichen Ergebnis?
3) Hat die Erfahrung des Testers einen Einfluss auf die von ihm erreichte Übereinstimmung der Ergebnisse am selben Probanden?

In meiner Arbeit beschreibe ich zunächst welches Erklärungsmodell dem Global Listening zugrunde liegt. Anschließend stelle ich die in der Literatur beschriebenen Interpretations- und Durchführungsvarianten vor. Um Informationen über die Anwendung des Tests in der Praxis zu erhalten, habe ich ein Interview mit Beatrix Krall, die für ihre Masterthese Osteopathen zum Global Listening interviewte, durchgeführt. Nach Reflexion der Informationen aus der Literatur und jener aus dem Interview wurde die Durchführung des Tests festgelegt. Nach einer kurzen Erklärung der Reliabilität wende ich mich der Analyse bisheriger Reliabilitätsstudien zu. Dabei werden Details dieser Studien mit dem Ziel der Optimierung der eigenen Studiendurchführung besprochen.Zuletzt beschreibe ich die Durchführung meiner Studie, stelle die Ergebnisse vor und schließe mit einer Reflexion.

2 Das Global Listening

Das Global Listening ist ein Faszientest, bei dem ein erster Eindruck über im Körper vorhandene Restriktionen gewonnen wird. Viele Autoren schlagen es als Teil der osteopathischen Befundaufnahme vor (Barral, 2002; Croibier, 2006; Hinkelthein und Zalpour, 2005;Paoletti, 2001; Puylaert, 2005). Wie das Interview mit Krall (2010) und auch die Angaben der Tester zeigen, wird es in der osteopathischen Praxis häufig angewendet.

In der Literatur findet man sowohl den Begriff „Ecoute“ als auch „Listening“, was dasselbe, nur in verschiedenen Sprachen (Französisch, Englisch) bedeutet. In dieser Arbeit wird der Begriff „Listening“ verwendet.

Dieses Kapitel beschreibt, was unter einer Berührung, die als Listening bezeichnet wird, zu verstehen ist und stellt das dahinter stehende Erklärungsmodell vor.Weiters werden die in der Literatur beschriebenen Varianten der Durchführung und Interpretation und Informationen aus dem mit Krall (2010) geführten Interview (siehe Anhang) vorgestellt.

Aus Reflexion der in Kapitel 2.3 und 2.4 beschriebenen Durchführungs- und Interpretationsmöglichkeiten entstand die in dieser Studie verwendete Variante. Beschrieben wird diese im Methodikteil in Kapitel 6.2.3.

2.1 Der Begriff Listening

Ein Listening ist eine Form der Berührung, bei der man versucht, mit den Händen auf den Körper zu hören (Barral, 2002). „*Die Hand lauert gespannt auf jeden von den Geweben ausgeübten Zug“* (Croibier, 2006, S.61). Dabei soll sie passiv und aufnahmebereit zugleich sein (Barral, 2002; Paoletti, 2001). Wichtige Faktoren sind laut Paoletti (2001) ein guter manueller Kontakt, die Neutralität des Therapeuten und die Einstimmung auf den Patienten. Dazu gehört Respekt vor dem Menschen und seinem Körper. Man „bittet das Gewebe um Erlaubnis“, mit ihm in Kontakt zu treten.

Einer der ersten, die diesen Ansatz verwendet haben, ist Rollin E. Becker. Seiner Beschreibung, die vor allem ein „Gefühl“ für diesen Test vermitteln soll, wird aus diesem Grund hier besonders viel Aufmerksamkeit geschenkt und einige seiner Anleitungen wörtlich zitiert.

“Listen through the hands, not with the hands. The patient demonstrates; the physician listens, actively”(Becker in Brooks (Hrsg.), 1997, S. 148).

“In this process, stop thinking about it and surrender to the total anatomicophysiological output of the patient”(Beckerin Brooks (Hrsg), 1997, S. 149).

2.2 Erklärungsmodell

Das Global Listening wird zu den Faszientests gezählt. Paoletti beschreibt in seinem Buch Faszien wie folgt:

> *„Faszien bilden eine ununterbrochene Gewebseinheit, die sich von Kopf bis Fuß, aber auch von außen nach innen erstreckt. Es gibt keine Unterbrechung in der faszialen Kontinuität, denn auch ihre Ansätze an knöchernen Strukturen sind nur Relais- oder Übergangszonen, welche die Rolle der Faszien unterstützen. Faszien sind somit auf allen Ebenen des Körpers präsent, sie umhüllen alle anatomischen Strukturen […], dringen in das Innere der Strukturen ein, …“*(Paoletti, 2001, S. V).

Durch diese Kontinuität lässt sich über die Befundung des Fasziensystems ein Eindruck vom gesamten Körper gewinnen. Eine Dysfunktionstellt einen Fixpunkt im Fasziensystem dar, der ein mechanisches Ungleichgewicht erzeugt und einen Zug ausübt.Durch das „Hören“ auf diesen Zug mittels manuellem Kontakt kann der Fixpunkt und somit die Dysfunktion gefunden werden. (Barral, 2002; Croibier, 2006;Hinkelthein und Zalpour, 2005; Puylaert, 2005)

2.3 Durchführung

Ein Global Listening kann am stehenden, sitzenden oder am Rücken liegenden Patienten durchgeführt werden. Meiner Erfahrung nach wird der Test meist am stehenden Patienten durchgeführt, was auch mit den Angaben von Krall (2010) übereinstimmt. Daher habe ich mich für diese Variante entschieden und stelle sie hier vor.

In der Literatur zum Global Listening wird auf unterschiedliche Details Wert gelegt und der Handkontaktunterschiedlich beschrieben.

Paoletti (2001, S. 198) beschreibt eine genaue **Position des Patienten** wie folgt:

„... der Patient steht mit leicht gegrätschten Beinen, richtet den Blick horizontal nach vorne und schließt nach Aufforderung die Augen."

Croibier (2006) betont, dass der Patient aufrecht stehen und während des Tests die Augen schließen soll.Nach Hinkelthein und Zalpour (2005) steht der Patient mit einem Abstand von ca. 10 cm zwischen den parallel zueinander ausgerichteten Füßen und schließt während des Tests die Augen. Einig sind sich diese Autoren also darüber, dass der Patient während des Tests die Augen geschlossen hat. Dies ist auch das einzige Detail zum Thema Patientenposition, das in den von Krall (2010) geführten Interviews häufig beschrieben wird.

Was die **Position des Testers** betrifft, steht dieser nach Barral (2002), Croibier (2006),Puylaert (2005) und Paoletti (2001) hinter dem Patienten. Einzig Hinkelthein und Zalpour (2005) geben an, dass der Tester seitlich neben dem Patienten steht. In den Interviews von Krall (2010) hing die Position des Testers von der Positionierung seiner Hände ab. Angegeben wurde sowohl, dass der Tester hinten, seitlich hinten als auch seitlich steht.

Der **Handkontakt** kann mit einer Hand oder zwei Händen erfolgen. Paoletti (2001)beschreibt einen Kontakt mit einer Hand, Hinkelthein und Zalpour (2005) sowie Puylaert (2005) mit zwei Händen. Für Croibier (2006) und Barral (2002) ist beides möglich.

Paoletti (2001)beschreibt einen Kontakt mit einer Hand am Kopf, wobei ihm vor allem die Qualität der Berührung und weniger die exakte Position wichtig ist.Nach Puylaert (2005) wird eine Hand quer auf das Schädeldach und eine Handin Längsrichtung auf das Kreuzbein gelegt.Gemäß der Beschreibung von Hinkelthein undZalpour (2005) legt derTester eine Hand auf den Kopf und die zweite zwischen die Schulterblätter.Barral (2002) beschreibt nur die Handhaltung im Sitz. Dabei liegt eine Hand parallel oder quer zur Wirbelsäule auf der okzipitoparietalen Region (*Anmerkung: im „Lehrbuch der viszeralen Osteopathie" von Barral (2002) liegt die Hand auf der Abbildung auf S.6 am Scheitelpunkt)*. Die zweite Hand kann Kontakt unter dem Steißbein haben, muss aber nicht. Die Position der ersten Hand kann meiner Meinung nach für den Stand übernommen werden, die der zweiten Hand nicht.Nach Croibier (2006) legt der Tester eine Hand auf den Scheitelpunkt des Kopfes und eventuell eine zweite Hand auf den cervikothorakalen Übergang (*Anmerkung: im*

Buch „Diagnostik in der Osteopathie“ von Croibier (2006)hat der Tester die zweite Hand auf der Abbildung auf S. 214 im Bereich des Sakrums). Einigkeit herrscht somit darüber, dass eine Hand am Kopf im Bereich des Scheitelpunktes liegt. Auch Krall (2010) gibt an, dass fast alle Interviewpartner beim Global Listening eine Hand auf den Kopf des Patienten legen. Ungefähr die Hälfte verwendet eine zweite Hand, die meist am Sakrum oder auf der Brustwirbelsäule liegt.

Neben dem Ort der Kontaktaufnahme ist der **Druck**, mit dem der Kontakt aufgenommen wird, wichtig. Paoletti (2001) empfiehlt die Hand flach und ohne Druck auf den Kopf zu legen.Hinkelthein und Zalpour (2005) beschreiben eine Kontaktaufnahme mit einem sanften,durch 20-30 Gramm Gewicht verursachten Druck. Barral (2002) schreibt von einer flach aufgelegten Hand, die durch leichten Druck die Bewegung des Körpers zur Läsion spürt. Die Interviewpartner von Krall (2010) beschreiben, dass sie die Hand nur leicht auflegen, beziehungsweise vereinzelt, dass sie gar keinen Kontakt haben.

Was die **Dauer** des Tests betrifft, kann man aus der Angabe von Hinkelthein undZalpour (2005), gemäß der der Patient nach vier Sekunden im Normalfall wie ein Brett nach dorsal fällt, annehmen, dass der Test weniger als vier Sekunden dauern soll. Croibier (2006)gibt an, dass man das Ungleichgewicht der Gewebe meist in den ersten Sekunden, nachdem der Patient die Augen geschlossen hat, spürt. Auch in den Interviews von Krall (2010)wird meist ein kurzer Kontakt (bis fünf Sekunden) angegeben, vereinzelt aber auch ein Kontakt von zehn bis zwanzig Sekunden.

Allen Autoren (Barral, 2002; Croibier, 2006; Hinkelthein und Zalpour, 2005; Paoletti, 2001; Puylaert, 2005)ist die **Einstellung der Tester** wichtig. Die Hand soll neutral, passiv und aufnahmebereit sein. Paoletti (2001) spricht auch von einem „respektvollen Dialog“ mit dem Gewebe.Folgt man den Geweben, ist das bereits der Beginn einer Behandlung (Croibier, 2006). Will man den Test rein zur Diagnose anwenden und keine Veränderung des Körpers verursachen, darf man den Geweben also nicht folgen.

2.4 Zugrichtungen und ihre Interpretation

Von fast allen Autoren (Barral, 2002; Croibier, 2006; Hinkelthein und Zalpour, 2005; Puylaert, 2005)werden vorne, hinten und zur Seite als mögliche Zugrichtungen angegeben und besprochen. Ein Zug nach vorne wird auch als Flexion, ein Zug nach hinten als Extension und ein Zug zur Seite als Lateralflexion bezeichnet.

Ein Zug nach vorne weist auf ein viszerales Problem hin, ein Zug nach hinten auf ein parietales Problem. Die Lateralflexion zeigt an, auf welcher Körperseite das Problem liegt. Dies ist allerdings eine grobe Hilfestellung (Barral, 2002; Croibier, 2006; Hinkelthein und Zalpour, 2005; Puylaert, 2005).

Nach Barral (2002),Croibier (2006),Hinkelthein und Zalpour (2005) sowie Puylaert (2005) liegt das Problem umso caudaler, je größer das Ausmaß der Flexion/Extension ist. Das Problem liegt umso lateraler, je größer die Seitneigung ist.

Croibier (2006) beschreibt weiters die Möglichkeit, dass die Hand am Kopf nach innen gezogen wird, was auf eine Dysfunktion im kraniosakralen System schließen lässt.

Hinkelthein (2005) gibt zusätzlich die Möglichkeit einer Rotation, des Einsinkens und des Stehenbleibens an.

Paoletti (2001) gibt zur Interpretation des Tests nahezu keine Hinweise. Bei ihm stehen Informationen über die richtige Kontaktaufnahme und die Einstellung des Osteopathen im Vordergrund. Daraus lässt sich schließen, dass er die durch den Test gewonnenen Informationen für zu komplex hält, als dass man eindeutige Hilfestellungen zur Interpretation der Testergebnisse geben kann. Auch in der Ausbildung der Autorin wurde erwähnt, dass das Global Listening einen generellen, zum Teil schwer in Worte zu fassenden Eindruck vom Patientenvermittelt.

3 Die osteopathische Indikation

Den Erkenntnissen aus Kapitel 5.1.1 vorgreifend ist die Testung symptomatischer Probanden für die Qualität einer Reliabilitätsstudie wichtig. Das Global Listening ist ein Test, der bei allen Beschwerdebildern zu einer ersten Orientierung eingesetzt wird. Es erscheint also lediglich eine Einschränkung auf Beschwerden, bei denen Osteopathie indiziert ist, als sinnvoll. Dies ist auch im Sinne einer möglichst großen Streuung der Ergebnisse. Was aber ist eine osteopathische Indikation?

„Eine gute osteopathische Indikation kann immer dann gestellt werden, wenn der Patient eine Anzahl größerer mechanischer Dysfunktionen aufweist“(Croibier, 2006, S. 281).

Wie Croibier (2006) beschreibt, ist das Feststellen einer osteopathischen Indikation das Ergebnis eines Befundungsprozesses und manchmal auch erst nach einigen Behandlungen möglich.

In dieser Studie wurde eine Annäherung an die Feststellung einer osteopathischen Indikation durch eine von der Autorin durchgeführte Anamnese versucht. Es ist jedoch klar, dass damit lediglich ein Hinweis auf das Vorhandensein einer solchen gegeben ist.

4 Reliabilität

Fröhlich (1997, S. 21) beschreibt Reliabilität wie folgt:

„Ein Maß oder ein Test kann dann als reliabel bezeichnet werden, … wenn es/er bei einer Wiederholung der Messung/Testung unter gleichen Bedingungen und an denselben Gegenständen zu dem gleichen Ergebnis führt.“

Allein die Formulierung *„an denselben Gegenständen“* macht klar, dass bei Testungen an Menschen – also an einer lebendigen, ständigen Adaptationen unterworfenen Einheit von Körper, Geist und Seele – keine idealen Voraussetzungen zum Erreichen einer perfekten Reliabilität gegeben sein können. Ebenso können durch das Verstreichen von Zeit während der Testung und den währenddessen an Testern, Testperson und Umgebung stattfindenden Veränderungen nicht völlig gleiche Bedingungen garantiert werden.

In dieser Arbeit wird sowohl die Interrater- als auch die Intrarater-Reliabilität bestimmt. Als Interrater-Reliabilität wird die Übereinstimmung des Testergebnisses mehrerer Tester an derselben Testperson/demselben Gegenstand bezeichnet. Von Intrarater-Reliabilität spricht man, wenn ein und derselbe Tester bei mehrmaliger Testung am selben Probanden/Gegenstand zum gleichen Ergebnis kommt. In der Literatur wird davon ausgegangen, dass die Intrarater-Reliabilität höher ist als die Interrater-Reliabilität (Krause, 2007).

Die Prüfung der Validität - das heißt, ob der Test ausschließlich und genau das misst, was er messen soll (Bortz und Döring, 2006) – ist nicht Gegenstand dieser Arbeit. Es ist eine Annahme, dass durch das Global Listening fasziale Spannungen untersucht werden.

5 Bisherige Reliabilitätsstudien

Wissenschaftliche Texte suchte ich zunächst im Internet unter der Adresse www.medbioworld.com/med/journals/osteopathic.html. Über diese Seite hatte ich Zugang zu einigen relevanten Journalen.Weiters verwendete ich zur Suche das Osteopathic Research Web (http://www.osteopathicresearch.com) und die Bibliothek der WSO. Suchbegriff war „reliability“. Die Recherche wurde auf Artikel, die nach dem 1.1.2000 publiziert wurden, eingeschränkt. Rechtfertigung dafür war, dass meines Erachtens die Qualität der Artikel ständig zunimmt, die statistische Auswertung leichter vergleichbar ist (Verwendung von Kappa-Werten) und Artikel vor diesem Datum in Metaanalysen, die ich eingearbeitet habe, analysiert und zusammengefasst werden. Zusätzlich suchte ich nach in den Artikeln vorgestellten Arbeiten, die mir besonders interessant erschienen. Bei drei Texten nahm ich per E-Mail Kontakt zu den Autoren auf, was in einem Fall erfolgreich war.

In den folgenden Kapiteln wird – auf der Basis der Analyse gelesener Artikel und Studien – auf die für die Qualität einer Reliabilitätsstudie wichtigen Elemente eingegangen. Diese sind für die Aussagekraft einer Reliabilitätsstudie entscheidend. Ziel der Analyse ist eine Optimierung des eigenen Studiendesigns.

5.1 Elemente von Reliabilitätsstudien

5.1.1 Probanden

Ein wesentlicher Punkt ist die **Anzahl der Probanden**. Diese variiert in den von der Autorin gelesenen Studien (Brismee et al., 2006; Degenhardt et al., 2005; Fryer et al., 2005; Gemmel und Miller, 2005; Gibbons et al., 2002; Gonella et al., 1982;Halma et al., 2008;Hartman und Norton, 2002; Hestboek und Leboeuf-Yde, 2000;Humphreys et al., 2004; Johansson, 2005; Kmita und Lucas, 2007;Podlesnic, 2006; Potter et al., 2006; Tong et al., 2006; Van Trijffel et al., 2005) von drei (Humphreys et al., 2004) bis 119 (Degenhardt et al., 2005). Meist werden jedoch mindestens 20 Personen getestet. Eine adäquate Anzahl an Probanden ist für die statistische Auswertung wichtig.

Wichtig ist außerdem, ob die Probanden **Symptome** haben oder nicht.

Gemmel undMiller (2005) schlossen für ihre Metaanalyse Studien, die mit asymptomatischen Probanden arbeiteten, aus.

Van Trijffel et al. (2005) kritisierten in ihrer Metaanalyse, dass von den 19 analysierten Studien in nur vier Studien mit symptomatischen Probanden gearbeitet wurde.Laut den Autoren ist daran problematisch, dass die untersuchten Merkmale zu wenig variieren, was die statistische Auswertung verfälscht. Sie untermauern diese Behauptung mit anderen Studien.

Auch Hestboek undLeboeuf-Yde (2000) beschreiben die Testung asymptomatischer Personen als häufigstes Problem in den 37 analysierten Studien und begründen dies damit, dass die statistische Auswertung durch mangelnde Streuung der Ergebnisse erschwert wird. Weiters sind sie der Meinung, dass das Finden von Abnormalität bei asymptomatischen Personen nicht den klinischen Alltag reflektiert.

Dass dieses Kriterium noch nicht ausreichend Beachtung findet, zeigt sich für die Autorin auch darin, dass im Abstract einiger Artikel (Gibbons et al., 2002; Johansson, 2006;Degenhardt et al., 2005) nicht einmal erwähnt wird, ob mit symptomatischen oder asymptomatischen Probanden gearbeitet wird.

Es gibt jedoch auch Studien, die die (teilweise) Testung asymptomatischer Personen gut begründen:

Fryer et al. (2005) verwendeten bei ihrer Studie über die Auswirkung eines Trainings der Tester auf die Inter- und Intrarater-Reliabilität des Vorlauftests und der Palpation anatomischer Bezugspunkte am Becken asymptomatische Probanden. Sie erklärten dies damit, dass in der Literatur kein Zusammenhang zwischen Asymmetrien im Becken und Symptomen gefunden wurde.

Kmita undLucas (2008), die die Reliabilität der Palpation anatomischer Bezugspunkte am Becken untersuchten, versuchten die Streuung der Ergebnisse zu erhöhen, indem sie vier asymptomatische und fünf symptomatische Probanden untersuchten.

Für die Autorin hat die Verwendung symptomatischer Probanden in der Studie drei Vorteile:

- Es ist bei Probanden mit Symptomen wahrscheinlicher, dass eindeutige Züge vorhanden sind.

- Der klinische Alltag wird besser wiedergegeben.
- Die für die statistische Auswertung notwendige Streuung der Ergebnisse ist bei symptomatischen Testpersonen eher gegeben. Diese Streuung soll durch die Testung von Patienten mit unterschiedlichen Symptomen noch verstärkt werden.

5.1.2 Tester

Die **Anzahl** der Tester variiert in den von der Autorin gelesenen Studien von zwei (Halma et al., 2008; Tong et al., 2006) bis 20(Humphreys et al., 2004). Im Review von Van Trijffel et al (2005) testeten bei 19 vorgestellten Studien elfmal nur zwei Tester. Hestboek und Leboeuf-Yde (2000) geben in ihrer Metaanalyse an, dass in einer Studie mit 42 Testern gearbeitet wurde. Aus den beigelegten Tabellen war die Anzahl der Tester nicht für alle Studien klar herauszulesen. Es testeten anscheinend auch 45 (*Anmerkung: stimmt nicht mit der von den Autorinnen angegebenen Höchstzahl von 42 Testern überein)* Tester an einem mechanischen Modell. Ansonsten wird bei der Analyse der Literatur deutlich, dass auf die Anzahl der Tester meist wenig Wert gelegt wird und die Anzahl der getesteten Personen im Vordergrund steht (zum Beispiel Degenhardt et al., 2005: drei Tester testeten zwei Untergruppen mit 42 bzw. 77 Probanden).

Wenn man annimmt, dass ein Tester wegen schlechter Tagesverfassung oder sonstigen Gründen Probleme bei der Testung hat, so muss bei zwei Testern eine schlechte Interrater-Reliabilität erreicht werden. Für mich ist eine höhere Anzahl an Testern also wesentlich.

Einige Studien arbeiteten mit **erfahreneren und weniger erfahrenen** Testern (Harveyund Byfield, 1991; Jensen et al, 1993; Kmita und Lucas, 2008; Podlesnic, 2006). In der Studie von Kmita und Lucas (2008) palpierten zwei erfahrene Osteopathen und zwei Studenten des Abschlussjahrgangs anatomische Bezugspunkte am Becken. Es konnte kein Unterschied zwischen der von beiden Gruppen erreichten Reliabilität festgestellt werden. Auch Podlesnic (2006) fand in seiner Studie zur Reliabilität des Global Listenings keinen signifikanten Unterschied zwischen Osteopathen, die ihre Ausbildung vor 2002 abgeschlossen hatten, zu Osteopathen die ihre Ausbildung 2002 oder später abgeschlossen hatten. In den Studien von Harvey und

Byfield (1991) und Jensen et al (1993) testeten Studenten und erfahrene Osteopathen an einem mechanischen Modell. Die Studenten erzielten bessere Ergebnisse.

5.1.3 Verwendung von Testkombinationen

Wie auch die Interviewpartner von Krall (2010) angeben, verwenden Osteopathen meist mehrere Tests in verschiedenen Ausgangsstellungen um Dysfunktionen festzustellen. Erst wenn einige Tests auf das gleiche Problem hinweisen, wird auf eine für die Behandlung wesentliche Dysfunktion geschlossen. Aufgrund dieser Tatsache gibt es Studien, die Testkombinationen untersuchen. Problematisch daran ist, dass es mit zunehmender Zahl von Tests an einem Patienten immer schwieriger wird, für stabile Testbedingungen zu sorgen. Hestboek und Leboeuf-Yde (2000) geben außerdem zu bedenken, dass der Tester möglicherweise bei den weiteren Tests vom Ergebnis des ersten Tests bzw. der vorherigen Tests beeinflusst ist. Weiters kann man so nicht feststellen, welche Tests innerhalb der Kombination reliabel sind.

5.1.4 Einschulung und Training

Degenhardt et al. (2005) untersuchten, ob die Werte, die für die Reliabilität von Tests der Lendenwirbelsäule erreicht wurden, durch Training verbessert werden können. Dazu testeten drei Tester einmal 42 und einmal 77 Probanden, also nicht die gleiche Gruppe. Im ersten Durchgang wurde in zwei verschiedenen Positionen getestet. Durch die Anzahl an Tests, die zwei- bis dreimal an einem Probanden ausgeführt wurden, wurde ein Wirbel 48- bis 72-mal getestet. Nachdem vier Monate lang ein bis zwei Stunden pro Woche geübt wurde, fand die zweite Testung statt. Dabei wurden die Probanden in nur einer Position getestet, ein Wirbel nicht öfter als 18-mal. Die Werte für die Interrater-Reliabilität waren bei der zweiten Testung eindeutig höher.

Die stattgefundene Verbesserung der Werte lässt sich meiner Meinung nach nicht nur auf das Training zurückführen, da die Durchführung der Studie durch die Reduktion der Testanzahl und der Positionen deutlich verbessert wurde. Man kann auch interpretieren, dass ersichtlich ist, wie wichtig die gute Durchführung einer Studie ist, bzw. welche Auswirkung sie auf das Ergebnis hat. Es kann außerdem nicht notwendig sein, alle fertig ausgebildeten Therapeuten einem derart intensiven Training zu unterziehen, damit Tests reliabel sind.

Fryer et al. (2005) untersuchten den Effekt eines Trainings auf die Intra- und Interrater-Reliabilität des Vorlauftests und der Palpation anatomischer Bezugspunkte am Becken. Zehn Testpersonen wurden von zehn Testern untersucht, wobei fünf der Tester zweimal ein einstündiges Training absolvierten, bei dem mit einem erfahrenen Osteopathen sowohl Testablauf als auch die Interpretation des Tests geübt wurden. Die Intrarater-Reliabilität konnte durch das Training eindeutig verbessert werden. Die Verbesserung der Interrater-Reliabilität war zu gering, um einen eindeutigen Effekt des Trainings nachweisen zu können.

Van Trijffel et al. (2005) fanden in ihrer Analyse zahlreicher Artikel keinen Unterschied zwischen den Ergebnissen von vorab trainierten undvorab nicht trainierten Testern.

Zusammenfassend kann man sagen, dass eindeutige Verbesserungen durch Training und Einschulung bisher nicht nachgewiesen, aber auch nicht ausgeschlossen werden können. Zu bedenken ist auch, dass Osteopathen nach Absolvierung ihrer Ausbildung fähig sein sollten, eindeutige Abweichungen übereinstimmend zu finden – auch ohne vorheriges Training.

Ein anderer Aspekt eines Trainings ist, dass eine vorherige Einschulung in die geplante Durchführung des Tests und den Ablauf der Datengewinnung erst eine Standardisierung möglich macht.

5.1.5 Testablauf und Dokumentation

Zunächst ist die genaue **Durchführung des Tests** festzulegen, was Patientenposition, Position des Testers, Handhaltung, Dauer eines Tests und eventuelle Angaben an den Probandenbeinhaltet. Ziel ist eine möglichst einheitliche Durchführung des Tests. Zu bedenken ist, dass am Körper der Probanden möglichst wenige Veränderungen verursacht werden sollen, um stabile Bedingungen für alle Tester zu gewährleisten.

Um einer Ermüdung der Beteiligten vorzubeugen, sollten bei längerer Dauer der Testung **Pausen** eingeplant werden. In einigen Studien wurde die Datengewinnung zu diesem Zweck auf mehrere Tage verteilt (Degenhardt et al., 2005; Gibbons et al., 2002; Johansson et al., 2006).

Wichtig ist weiters **die Blindierung der Tester**gegenüber Informationen betreffend die Probanden. Dies beinhaltet auch visuelle Informationen. Bei Studien zur Intrarater-Reliabilität muss das Wiedererkennen eines Probanden verhindert werden. Aber auch bei Studien zur Interrater-Reliabilität sollen sich die Tester möglichst unbeeinflusst von verschiedenen Reizen für ein Testergebnis entscheiden. Blindierung gegenüber den Ergebnissen der anderen Tester muss ebenfalls gegeben sein.

Was in den von der Autorin gelesenen Studien nicht beachtet wurde, ist die **Blindierung der Probanden**. Diese müssen zwar über den Ablauf der Testung informiert sein, über das genaue Ziel und die Interpretationsmöglichkeiten sollten sie, um eine Einflußnahme zu verhindern, jedoch nicht informiert sein. Dies geben auch Gemmel und Miller (2005) in ihrer Metaanalyse an.

Bei Studien zur Intrarater-Reliabilität muss zumindest ein zweiter **Testungsdurchgang** stattfinden. In der von der Autorin gelesenen Literatur fanden in vier Fällen zwei (Gibbons et al., 2002; Kmita und Lucas, 2008; Podlesnic, 2006; Potter, 2006) und in zwei Fällen drei (Fryer et al., 2005; Halma et al., 2008) Untersuchungsdurchgänge statt. Für die statistische Auswertung ist eine möglichst große Anzahl an Durchgängen wichtig, mit zunehmender Wiederholungsanzahl an einemProbanden kann jedoch immer weniger für stabile Bedingungen gesorgt werden.

Die **Reihung** der Tester soll nach Gemmel und Miller (2005) **zufällig** erfolgen.

Es ist wesentlich, alle diese Punkte zu **dokumentieren**. Tabellen der Ergebnisse sollten in der Studie veröffentlicht werden, um allfällige Verzerrungen durch die statistische Auswertung nachvollziehbar zu machen (Van Trijffel et al., 2005).

5.1.6 Zusammenfassung

In diesem Kapitel werden die Erkenntnisse aus den Kapiteln 5.1.1 bis 5.1.5 zusammengefasst. DieUmsetzung der Erkenntnissein dieser Studiewird nach jeder Darstellung beschrieben.

1. Wesentlich ist eine ausreichende Anzahl an **Probanden**. Diese sollen Symptome aufweisen. Sie sind gegenüber den Interpretationsmöglichkeiten des Tests, dieden Testern vorgegeben sind, zu blindieren.

Umsetzung in dieser Studie:Es wurde angestrebt 30 Probanden zu testen. Am Tag der Datengewinnung standen jedoch nur 18 zur Verfügung. Getestet wurden Personen, die – wie in einer von der Autorin durchgeführten Anamnese festgestellt – eine osteopathische Indikation aufweisen. Die Probanden waren nicht über die den Testern zur Verfügung stehenden Interpretationsmöglichkeiten informiert.

2. Auch eine adäquate Anzahl an **Testern** ist anzustreben. Sie sind gegenüber den Symptomen der Probanden und den Ergebnissen der anderen Tester zu blindieren. Wird die Intrarater-Reliabilität bestimmt, ist ein Wiedererkennen einer Testperson durch den Tester zu verhindern. Eine Auswirkung der Erfahrung der Tester auf das Ergebnis konnte bisher nicht bzw. insofern festgestellt werden, als erfahrene Tester schlechtere Ergebnisse erzielten (Harvey und Byfield, 1991; Jensen et al, 1993; Kmita und Lucas, 2008; Podlesnic, 2006). Sie bleibt ein interessanter Forschungsgegenstand und wird deshalb auch in dieser Studie untersucht.

Umsetzung in dieser Studie: Sechs Tester führten das Global Listening mit verbundenen Augen an den Probanden durch. Das Ergebnis wurde aufgeschrieben und von Hilfspersonen in eine Tabelle übertragen. Das Wiedererkennen von Probanden wurde durch Vorgaben an die Probanden zu verhindern versucht. Es konnten keine Tester mit langjähriger Erfahrung gefunden werden. Daher wurde die Erfahrung über die Häufigkeit der Anwendung des Tests in der Praxis der einzelnen Tester erfasst.

3. Eine Verbesserung der Ergebnisse durch ein **Training** der Tester konnte nicht nachgewiesen werden (Fryer et al., 2005; Van Trijffel et al., 2005). Eine **Einschulung** der Tester wird jedoch im Sinne der Standardisierung des Ablaufs von der Autorin als sinnvoll erachtet.

Umsetzung in dieser Studie: Eine Woche vor der Datengewinnung fand eine Einschulung der Tester in Ablauf und Durchführung des Tests statt.

4. Die Verwendung von **Testkombinationen** mag den klinischen Alltag reflektieren. Es ist jedoch nicht möglich zu erkennen, welche Tests aus der Kombination reliabel sind. Außerdem führt sie zu einer höheren Anzahl an Tests an einem Probanden, was zu Veränderungen an diesem führen kann.

Umsetzung in dieser Studie: Es wurde ein einzelner Test untersucht.

5. Die **Testdurchführung** ist im Sinne der Standardisierung genau festzulegen. Um einer Ermüdung der Tester vorzubeugen, sind Pausen ratsam. Tester und Testpersonen sind zufällig zu reihen. Bei Bestimmung der Intrarater-Reliabilität ist eine ausreichende Anzahl an Durchgängen zu bedenken.

Umsetzung in dieser Studie: Patientenposition, Therapeutenposition, Handhaltung, Gewicht, mit dem die Hand aufgelegt wird, Position der Hände der Tester, Dauer des Handkontakts und mögliche Interpretationen wurden vorgegeben. Nach 18 Probanden fand eine Pause statt. Tester und Testpersonen wurden zufällig gereiht. Zehn der Probanden wurden zur Bestimmung der Intrarater-Reliabilität dreimal getestet.

6. Ein wesentlicher Punkt ist, für **stabile Bedingungen** während des Tests zu sorgen. Dies umfasst das Ausschaltenmöglicher Veränderungen am Tester und an der Testperson, die Einfluss auf das Ergebnis haben können.

Umsetzung in dieser Studie: Die Dauer des Tests war kurz. Die Tester wurden darauf hingewiesen den Zügen nicht zu folgen, um keinen Einfluss auf die Probanden zu nehmen. Es wurde auf eine ruhige Umgebung geachtet und eine Pause eingelegt.

5.2 Ergebnisse bisheriger Studien

In diesem Kapitel wird ein Überblick über die Ergebnisse bisheriger Studien gegeben. Dafür werden auch einige Metaanalysen zu Hilfe genommen. Da – dem Inhalt des Kapitels vorgreifend – Studien, die keine Reliabilität eines Tests nachweisen konnten, überwiegen, werden einige Studien, die einen Test als reliabel ausweisen konnten, genauer vorgestellt. Anhand ihrer Beschreibung kann überlegt werden, weshalb die akzeptablen Werte erreicht wurden. Vorgestellt wird auch die Arbeit von Podlesnic (2006), da sie wie diese Masterthese ein Listening untersucht.

Zum besseren Verständnis der besprochenen Ergebnisse wird die Interpretationsskala nach Landis und Koch (1977) vorgestellt. Sie beschreibt das Ausmaß der Reliabilität eines Tests.

Kappa-Wert	Interpretation (englisch)	Interpretation (deutsch)
$\kappa \leq 0{,}20$	poor	schwach
$0{,}20 \leq \kappa \leq 0{,}40$	fair	leidlich
$0{,}40 \leq \kappa \leq 0{,}60$	moderate	mittelmäßig
$0{,}60 \leq \kappa \leq 0{,}80$	substantial	beträchtlich
$0{,}80 \leq \kappa \leq 1{,}00$	almost perfect	fast ideal

Tabelle 1: Interpretationsskala nach Landis und Koch (1977)

Laut Fjellner et al. (1999) werden Werte von zumindest 0,4 als Indikator für eine akzeptable Interrater-Reliabilität angesehen.

Eine genaue Erklärung des Kappa-Werts folgt in Kapitel 6.3.1.

Degenhardt et al. (2005) untersuchten, ob die Werte, die für die Reliabilität von Tests der Lendenwirbelsäule erreicht wurden, durch Training verbessert werden können. Nach Durchführung eines intensiven Trainings konnten für Tests aus den Bereichen Gewebebeschaffenheit und Empfindlichkeit akzeptable Werte erreicht werden (Empfindlichkeit κ = 0,68, Gewebebeschaffenheit κ = 0,45). Für Tests zur Feststellung einer asymmetrischen Position oder Beweglichkeit konnte keine akzeptable Reliabilität erreicht werden (Position $\kappa = 0{,}34$, Beweglichkeit $\kappa = 0{,}20$). Die Autoren gaben zu bedenken, dass die Probanden keinen repräsentativen Anteil an der Gesamtbevölkerung darstellten. Es handelte sich um junge Probanden mit mäßigen Symptomen. Sowohl das Alter als auch die mäßige Ausprägung der Symptome machen Veränderungen im Körper während der Testung wahrscheinlicher.

Brismée et al. (2006) untersuchten die Interrater-Reliabilität eines „Passive Physiological Intervertebral Motion Tests“ (PPIM) der Beweglichkeit zwischen Brustwirbel (Th) sechs und sieben. Drei erfahrene Physiotherapeuten mit Zusatzausbildung in Manualtherapie untersuchten 41 asymptomatische Probanden. Es gibt keine Angaben über eine Einschulung der Tester. Ein Forscher, der kein Tester in der Studie war, markierte den sechsten Brustwirbel. Der Ablauf des Tests war klar, die Durchführung kurz. Jeder Proband wurde vom jeweiligen Tester einmal extendiert, zur einen Seite lateralflektiert und anschließend rotiert. Dann wurde die Extension mit Lateralflexion und Rotation zur anderen Seite kombiniert. Falls der erste Tester das

Segment Th6/7 zu immobil zur Durchführung des Tests fand, wurde das darüber liegende Segment getestet und Tester zwei und drei informiert. Diese testeten dann ebenfalls Th5/6. Um eine Ermüdung der Tester zu vermeiden, machte jeder Tester nach drei bis vier Probanden eine 20-30-minütige Pause.

Die Studie konnte zeigen, dass der PPIM Test reliabel ist. Die Kappa-Werte der Tester zueinander variierten von κ = 0,27 bis κ = 0,65. Insgesamt wurde ein Wert von κ = 0,41 erreicht. Allerdings erreichten zwei Paarungen von Testern ein Ergebnis unter 0,4 (0,27 und 0,30). Erst das gute Ergebnis des Vergleichs von Tester 1 mit Tester 2 (0,65) machte den insgesamt guten Wert möglich. Eine mögliche Ursache für das gute Ergebnis liegt im klar definierten Ablauf des Tests, durch den die Testpersonen nicht zu oft hintereinander getestet wurden. Auch das Ausweichen auf ein anderes Segment, falls im ursprünglichen Segmentkeine guten Bedingungen für die Testung gegeben waren, mag sich positiv ausgewirkt haben.

Halma et al. (2008) beschäftigten sich in ihrer Studie mit der Intrarater-Reliabilität der Feststellung dreier craniosakraler Parameter: des CRI (cranial rhythmic impuls), der Dysfunktion der Synchondrosis sphenobasilaris (SSB) und des betroffenen Quadranten. Dazu testeten zwei erfahrene Osteopathen jeweils 24 Patienten dreimal. Auf eine gute Blindierung der Tester wurde großer Wert gelegt. Getestet wurden jeweils acht Patienten aus folgenden Gruppen: asymptomatische Probanden, Patienten mit Kopfschmerz, Asthmapatienten. Da es für die Intrarater-Reliabilität keine allgemeine Grenze für den Kappa-Wert gibt, ab der man von akzeptabler Reliabilität spricht, setzten die Autoren sich selbst die Grenze eines Wertes κ> 0,60. Dieser Wert ist um 0,20 höher als die Grenze für eine akzeptable Interrater-Reliabilität.

Die Ergebnisse der Testung von Dysfunktionen der SSB konnten die gesetzte Grenze überschreiten und erreichten Kappa-Werte im Bereich „beträchtlich“. Bei der Festlegung auf einen betroffenen Quadranten konnte für drei Quadranten mittelmäßige Reliabilität und für einen Quadranten leidliche Reliabilität erreicht werden. Beides liegt unter der gesetzten Grenze. Am schlechtesten waren die Ergebnisse zur Palpation des CRI: Hier konnte ein Kappa-Wert von κ = 0,23 (leidlich) erreicht werden. Die Autoren vermuteten, dass der CRI ähnlich wie die Herzfrequenz ständigen Schwankungen unterworfen ist und damit das schlechte Ergebnis erklärbar ist.Bisher wurde im Bereich der craniosakralen Osteopathie die Frequenz des craniosakralen

Rhythmus untersucht. Mit ihrem neuen Ansatz und guter Methodik konnten die Autoren zum Teil akzeptable Werte erreichen.

Humphreys et al. (2004)untersuchten in ihrer Studie die Validität eines Tests zur Beurteilung der Beweglichkeit der Halswirbelsäule. Getestet wurden drei Probanden mit einem angeborenen Blockwirbel. Dadurch konnte ein „gold standard“ festgelegt und die Validität getestet werden. 24 Studenten für Chiropraktik, die nicht über das Vorhandensein der Blockwirbel informiert waren, testeten die Probanden. Die Höhe der einzelnen Wirbelsäulensegmente wurde zuvor auf der Haut markiert. Die Ergebnisse lagen mit κ = 0,65 im Bereich „beträchtlich“. Die Besonderheit dieser Studie liegt darin, dass es gelungen ist, an einem lebenden Menschen für stabile Bedingungen bezüglich des getesteten Merkmalszu sorgen. Das gute Ergebnis legt nahe, dass instabile Bedingungen ein wesentlicher Faktor für schlechte Ergebnisse sind.

Podlesnic (2006) untersuchte die Intra- und Interrater-Reliabilität des „Abdominal Local Listenings“. 14 Osteopathen untersuchten 15 Testpersonen. Drei Testpersonen wurden von den Osteopathen zweimal getestet. Es konnte weder eine Intra- noch eine Interrater-Reliabilität des Tests nachgewiesen werden.

Hestboek und Leboeuf-Yde (2000) kamen in ihrer Metaanalyse zu chiropraktischen Tests des Bereichs Lendenwirbelsäule-Sakrum zu dem Ergebnis, dass einzig Studien zur Schmerzpalpation akzeptable Ergebnisse erzielten. Die Analyse beinhaltete sowohl palpatorische als auch visuelle Tests. In zwei der besprochenen Studien (Harvey und Byfield, 1991; Jensen et al., 1993) wurde ein mechanisches Modell getestet.

Van Trijffel et al. (2005) kamen in ihrer Metaanalyse zur passiven Beurteilung der Beweglichkeit einzelner Wirbelsäulensegmente in der Hals- und Lendenwirbelsäule zu dem Ergebnis, dass die Interrater-Reliabilität in den analysierten Studien niedrig war. Sie kritisierten die Qualität der Studien. Zu ihrer Verbesserung erschien es ihnen besonders wichtig, dass symptomatische Probanden getestet werden, stabile Bedingungen im getesteten Bereich gegeben sind, die Tester ausreichend blindiert sind und die Intrarater-Reliabilität bestimmt wird.

Hartman und Norton (2002)analysierten Studien zur Reliabilität craniosakraler Tests. Abgesehen von der Studie von Upledger (1977) untersuchten alle Forscher die Frequenz des craniosakralen Rhythmus. Die Reliabilität dieser Testung konnte in

keiner Studie nachgewiesen werden. Allerdings konnte festgestellt werden, dass die einzelnen Tester einen Frequenzbereich bevorzugen, das heißt ein und derselbe Tester fand an verschiedenen Patienten ähnliche Frequenzen. Die Studie von Upledger(1977), die zu dem Schluss kam, dass die craniosakrale Befundaufnahme reliabel ist, wurde methodologisch stark kritisiert und ihr jegliche Aussagekraft abgesprochen. Zu bemerken ist, dass die Metaanalyse stark emotional geprägt ist und eine wissenschaftliche Neutralität gegenüber dem Forschungsgegenstand vermissen lässt.

Gemmel und Miller (2005)analysierten Studien, die eine Testkombination zum Finden einer manipulierbaren Läsion in der Wirbelsäule evaluierten. Von den wenigen Studien, die den Qualitätsanforderungen entsprachen, konnte keine eine reliable Testkombination nachweisen.

In den weiteren von der Autorin gelesenen Studien (Fryer et al., 2005; Gibbons et al., 2002; Gonella et a., 1982; Johannson, 2006; Kmita und Lucas, 2008; Potter et al., 2006; Tong et al., 2006) konnte keine akzeptable Reliabilität nachgewiesen werden.

6 Methodik

6.1 Fragestellung und Hypothese

Ziel der Studie ist die Beantwortung folgender **Fragestellungen**:

1)Kommen verschiedene Tester, die am selben Probanden ein Global Listening durchführen, zum gleichen Ergebnis? (Interrater-Reliabilität)

2) Kommt ein und derselbe Tester, der am selben Probanden mehrmals ein Global Listening durchführt, bei jeder Testung zum gleichen Ergebnis? (Intrarater-Reliabilität)

3) Hat die Erfahrung des Testers einen Einfluss auf die von ihm erreichte Übereinstimmung der Ergebnisse am selben Probanden?

Zu diesen Fragestellungen wurden folgende **Hypothesen** aufgestellt:

Nullhypothesen:

1) Führen mehrere Tester innerhalb eines kurzen Zeitraums am selben Probanden ein Global Listening durch, liegt die Übereinstimmung ihrer Ergebnisse im Bereich zufälliger Übereinstimmung.

2) Führt ein Tester innerhalb eines kurzen Zeitraums am selben Probanden mehrmals ein Global Listening durch, liegt die Übereinstimmung seiner Ergebnisse im Bereich zufälliger Übereinstimmung.

3)Führt ein Tester, der das Global Listening in der Praxis häufig anwendet, diesen Test mehrmals am selben Probanden durch, so erreicht er keine höhere Übereinstimmung seiner Ergebnisse als einTester, der das Global Listening in der Praxis nicht anwendet.

Alternativhypothesen:

1) Führen mehrere Tester innerhalb eines kurzen Zeitraums am selben Probanden ein Global Listening durch, liegt die Übereinstimmung ihrer Ergebnisse über den von

Fjellner et al. (1999) geforderten Kappa > 0,4 für eine akzeptable interindividuelle Übereinstimmung.

2) Führt ein Tester innerhalb eines kurzen Zeitraums am selben Probanden mehrmals ein Global Listening durch, liegt die Übereinstimmung seiner Ergebnisse über den von Halma et al.(2008) vorgeschlagenen Kappa > 0,6 für eine akzeptable intraindividuelle Übereinstimmung.

3) Führt einTester, der das Global Listening in seiner Praxis häufig anwendet, diesen Test mehrmals am selben Probanden durch, so erreicht er eine höhere Übereinstimmung seiner Ergebnisse als einTester, der das Global Listening in seiner Praxis nicht anwendet.

6.2 Durchführung der Datenaufnahme

6.2.1 Probanden

Um eine gute statistische Auswertung zu ermöglichen, wurde eine Probandenzahl von 30 angepeilt. Diese Zahl wurde unter der Annahme, dass es meist zu kurzfristigen Ausfällen kommt, festgelegt. Zur Gewinnung von Testpersonen wurde im Rahmen der Kurse an der Wiener Schule für Osteopathie Kontakt zu den Studenten aufgenommen. Von den 52 interessierten Studenten hatten am Tag der Testung jedoch nur 13 Zeit sowie Symptome, die eine osteopathische Indikation darstellen können. Deshalb nahmen zusätzlich fünf Patienten der Autorin teil.

Ob eine osteopathische Indikation gegeben ist, wurde durch eine von der Autorin durchgeführte Anamnese festgestellt. Die Symptome der Testpersonen wurden notiert (siehe Anhang) und bei Schmerz eine visuelle Analogskala (VAS - siehe Anhang) ausgefüllt.

Am Tag der Testung standen 18 Probanden zur Verfügung.

6.2.2 Tester

Bei der Rekrutierung von Testern war das ursprüngliche Ziel,fünferfahrene und fünfunerfahrene Tester zu gewinnen und die Ergebnisse der beiden Gruppen zu vergleichen. Da auf ein E-Mail an alle Osteopathen, die Mitglied der Österreichischen

Gesellschaft für Osteopathie sind, nur eine Antwort kam und diese Osteopathin am Tag der Testung keine Zeit hatte, muss auf diesen Vergleich verzichtet werden. Tester waren Osteopathen, die im Herbst 2009 ihre Diplomprüfung absolviert hatten. Die Häufigkeit der Anwendung des Global Listenings in der Praxis der Tester wurde erhoben (sieheTabelle 2) und in die Auswertung der Ergebnisse einbezogen.

Tester	**Häufigkeit der Anwendung**
T1	bei jedem Patienten vor und nach der Behandlung
T2	nie
T3	bei jedem Patienten
T4	bei jedem dritten Patienten
T5	bei jedem zweiten Patienten und bei jedem Erstbefund
T6	bei jedem zweiten Patienten

Tabelle 2: Häufigkeit der Anwendung des Global Listenings in der Praxis der Tester

Ein Tester verwendet das Global Listening in seiner Praxis nicht, zwei bei jedem Patienten, zwei bei jedem zweiten Patienten und ein Tester verwendet diesen Test bei ca. jedem dritten Patienten.

6.2.3 Durchführung des Tests

Die für die Studie gewählte Durchführung des Tests ist ein Ergebnis der in Kapitel 2.3vorgestellten Literatur. Um dies nachvollziehbar zu machen, wird auf die jeweilige Quelle verwiesen.

Die Position der Probanden wurde wie folgt festgelegt: stabiler, hüftbreiter Stand, Blick nach vorne gerichtet. Auf Aufforderung durch den Tester waren die Augen zu schließen. Dies entspricht der Beschreibung von Paoletti (2001), die am vollständigsten war und in einigen Punkten von Croibier (2006) und Hinkelthein und Zalpour (2005) unterstützt wird.

Die Tester standen hinter den Probanden (Barral, 2002; Croibier, 2006;Paoletti,2001; Puylaert, 2005) und legten eine Hand mit einem Gewicht von 20 bis 30 Gramm (Hinkelthein und Zalpour, 2005) auf den Scheitelpunkt.

Was die Position der ersten Hand betrifft (siehe oben), sind sich die vorgestellten Autoren einig. Die Verwendung einer zweiten Hand wurde bei der Einschulung (siehe Kapitel6.2.4) diskutiert, da in der Literatur verschiedene Varianten vorgestellt werden.

Ziel war eine Lösung, die allen Testern angenehm ist, sodass sie sich gut auf die Züge des Patienten konzentrieren können. So sollte eine Verzerrung der Ergebnisse durch eine ungewohnte und unangenehme Handhaltung vermieden werden. Es wurde die Positionierung der zweiten Hand zwischen den Schulterblättern festgelegt. Dies entspricht der Beschreibung von Hinkelthein und Zalpour (2005).

Die Dauer des Handkontakts zur Informationsaufnahme wurde auf vier Sekunden festgelegt (Hinkelthein und Zalpour, 2005).

Als Kategorien für das Testergebnis wurden ursprünglich fünf Möglichkeiten vorgegeben: Vorne, hinten, links, rechts, anders. Nach der Einschulung der Tester wurde auf deren Bitte und nach Rücksprache mit einem Statistiker auf sieben Kategorien erweitert: vorne, vorne links, vorne rechts, hinten, hinten links, hinten rechts, anders.

Abbildung 1: Durchführung des Tests

6.2.4 Einschulung

Eine Woche vor der Datengewinnung trafen sich die Tester zur Einschulung. Dabei wurde die Durchführung des Tests aneinander geübt und die Ergebnisse wurden besprochen. Die Positionierung der zweiten Hand wurde wie in Kapitel 6.2.3 beschrieben festgelegt. Da die Tester Probleme hatten, das Gespürte in die vorgege-

benen fünf Kategorien einzuordnen, wurde auf sieben Kategorien ausgeweitet (siehe Kapitel 6.2.3).

Mithilfe einer Küchenwaage wurde trainiert, eine Hand mit einem Gewicht von 20 bis 30 Gramm auf den Scheitelpunkt zu legen. Die Küchenwaage wurde dazu auf den Kopf einer Person gestellt, um der Handposition bei der Testung nahe zu kommen.

Die Tester wurden auf die Wichtigkeit der eigenen Einstellung hingewiesen und gebeten, bei der Testung in einen „respektvollen Dialog mit den Geweben“(Paoletti 2001) zu treten.

Außerdem wurde der Ablauf vorgestellt. Dabei äußerten die Tester, dass es ihnen angenehmer wäre, erst nach 20 Probanden und nicht wie ursprünglich geplant nach 10 Probanden eine Pause zu machen. Diesem Wunsch wurde entsprochen.

Zuletzt wurden die Tester gebeten, die vorgegebene Handhaltung in der Woche bis zur Testung bei den eigenen Patienten zu üben.

6.2.5 Vorinformationen an die Probanden

Die Probanden wurden einige Tage vor der Testung per E-Mail über den Ablauf und ihre Körperposition beim Test informiert. Außerdem wurden sie gebeten kein Parfum oder sonstiges auffällig Riechendes zu verwenden. Die Verwendung von Haarstylingprodukten und Haarspangen war nicht erlaubt. Bei der Oberbekleidung sollten keine auffälligen Materialien und keine Oberteile mit Kapuze getragen werden. Weiters wurde die Wichtigkeit einer neutralen Einstellung betont und insbesondere gebeten, nicht an eine Stelle im Körper oder die Beschwerden zu denken. Die Studenten unter den Probanden wussten, dass mit diesem Test ein Zug im Körper ermittelt werden soll. Über die im Rahmen der Testung vorgeschriebene Einordnung der Ergebnisse waren die Probanden nicht informiert.

6.2.6 Blindierung

Wie in Kapitel6.2.2 beschrieben, hatten die **Tester** keine Informationen über die Beschwerden der Probanden. Durch die sofortige räumliche Trennung der Probanden und Tester am Tag der Datengewinnung und das Verbinden der Augen während der Testung erhielten die Tester auch keine visuellen Informationen über die Probanden.

Den Testern wurde gesagt, dass sie 38 verschiedene Probanden testen werden. Weiters wurde durch die in Kapitel 6.2.5 beschriebenen Vorgaben an die Patienten das Risiko, dass ein Tester eine Testperson beim zweiten oder dritten Durchgang wiedererkennt, reduziert.

Wie in Kapitel 6.2.5 beschrieben, wussten die **Testpersonen** nicht, in welche Kategorien die Tester das Gespürte zu unterteilen hatten. Die Osteopathie-Studenten unter den Probanden wussten, dass mit dem Test Züge im Körper festgestellt werden.

6.2.7 Ablauf

Am Tag derDatengewinnung wurden Tester und Probanden beim Eintreffen im Therapiezentrum sofort räumlich voneinander getrennt. So erhielten die Tester auch keine visuellen Informationen über die Probanden.

Zusätzlich zur Autorin waren drei eingeschulte Hilfskräfte anwesend. Die Autorin nahm nicht an der Untersuchung teil. Der Ablauf wurde noch einmal mit Testern und Testpersonen durchgegangen. Die Tester wurden per Los zufälliggereiht, die Probanden wurden bereits vorher von der Autorin zufällig gereiht und über die Reihenfolge informiert. Fragen wegen Unklarheiten wurden auf Einladung dazu keine gestellt.

Die Tester wurden im Halbkreis aufgestellt und ihre Augen verbunden. Jeweils eine Hilfsperson war für zwei Tester zuständig. Der geplante Ablauf wurde an Proband 17 und Proband 18 noch einmal geübt. Dann begann die Datenaufnahme.

Die Probanden gingen von einem Tester zum nächsten. Der Tester nahm auf Aufforderung und zum Teil mit Hilfe der Hilfsperson den vorgeschriebenen **Handkontakt** auf. Sobald die Hände positioniert waren, forderten die Tester die Probanden auf, die Augen zu schließen. Vier Sekunden später wiesen die Hilfspersonen die Tester auf den Ablauf der Zeit zum Spüren eines Zugs hin. Das Ergebnis schrieben die Tester blind auf ein Blatt Papier. So wurde verhindert, dass ein Tester vom Testergebnis eines anderen Testers erfuhr.Anschließend wurde das Ergebnis von den Hilfspersonen in eine Tabelle (siehe Anhang) übertragen.

Nach 18 Probanden fand eine ca. zehnminütige Pause statt, in der Tester und Probanden getrennt waren.

Danach wurden die ersten zehn Probanden noch zweimal hintereinander getestet.

6.2.8 Rückmeldungen von Testern, Probanden und Hilfspersonen

Sowohl direkt nach der Testung als auch per E-Mail wurden Tester, Testpersonen und Hilfspersonen um Feedback zur Durchführung gebeten.

Einige **Testpersonen** gaben an, dass sie sich vom Tester in eine bestimmte Richtung bewegt gefühlt hatten. Auch, dass sich der Kontakt bei jedem Tester anders anfühlte, wurde von mehreren Personen beschrieben. Dies betraf den Druck, aber auch andere schwer zu beschreibende Qualitäten. So hatten zwei Probanden das Gefühl mit manchen Testern in eine Verbindung treten zu können, bei der ein Informationsaustausch möglich ist, bei anderen nicht.

Ein Proband spürte während der Testung in sich selbst Züge.

Ein Proband fühlte sich nach dem ersten Durchgang stabiler und mehr „in seiner Mitte“.

Einige **Tester** gaben an, dass es nicht leicht war, das Gespürte in die vorgegebenen Kategorien einzuteilen. Insbesondere die Möglichkeit, einen Zug nach ausschließlich links/rechts anzugeben, fehlte.

Eine Testerin fand es schwierig, blind zu testen, da sie selbst Probleme hatte, eine stabile Position beziehungsweise die eigene Mitte zu finden. Dieses Problem wurde mit zunehmender Dauer stärker.

Die **Hilfspersonen** gaben an, dass innerhalb des vorgegebenen Rahmens die Durchführung des Tests durch die von ihnen betreuten Tester variierte.

6.3 Ergebnisse

6.3.1 Auswertung der Untersuchungsergebnisse

Zur Auswertung der Ergebnisse wurde das Übereinstimmungsmaß **Cohen’s Kappa (κ)** für alle Zweierkombinationen der Tester berechnet. Der Kappa-Wert stellt das Ausmaß dar, um das die beobachtete Übereinstimmung über jene Übereinstimmung hinausgeht, die durch Zufall alleine zustande käme.

Kappa wird wie folgt berechnet:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

wobei P_o die Wahrscheinlichkeit der gemessenen Übereinstimmung und P_e die Wahrscheinlichkeit einer zufälligen Übereinstimmung darstellen (Sachs, 2004).

Die Werte von κ können konsensgemäß mit folgenden in der Literatur verwendeten Graden der Übereinstimmung interpretiert werden (Landis und Koch, 1977):

Kappa-Wert	Interpretation (englisch)	Interpretation (deutsch)
$\kappa \leq 0{,}20$	poor	schwach
$0{,}20 \leq \kappa \leq 0{,}40$	fair	leidlich
$0{,}40 \leq \kappa \leq 0{,}60$	moderate	mittelmäßig
$0{,}60 \leq \kappa \leq 0{,}80$	substantial	beträchtlich
$0{,}80 \leq \kappa \leq 1{,}00$	almost perfect	fast ideal

Tabelle 1: Interpretationsskala nach Landis und Koch (1977)

Laut Fjellner et al. (1999) werden Werte von zumindest 0,4 als Indikator für eine akzeptable Interrater-Reliabilität angesehen. Für die Intrarater-Reliabilität gibt es keine Interpretationsskala. Es wird davon ausgegangen, dass sie über der Interrater-Reliabilität liegt. Werte, die das Vorhandensein einer Intrarater-Reliabilität zeigen, sollten somit über den geforderten Werten für die Interrater-Reliabilität liegen.Halma et al. (2008) setzten in ihrer Studie einen Grenzwert von 0,6 für eine akzeptable Intrarater-Reliabilität fest.

Zusätzlich zu den einzelnen Kappa-Werten wurden für eine Gesamtbewertung des Tests deren **Mittelwerte**, sowie deren **Vertrauensbereiche (95 %-Konfidenzintervalle)** berechnet. Die 95 %-Konfidenzintervalle stellen jene Grenzen dar, die den wahren Mittelwert in 95 % aller Stichproben aus der gleichen Grundgesamtheit einschließen und hängen sowohl von der Anzahl der Werte als auch deren Streuung um den Mittelwert ab (Sachs, 2004). Weiters wurden die **Standardabweichungen** als Maß für die Streuung der Werte um den Mittelwert berechnet. Negative

Kappa-Werte wurden einerseits berücksichtigt, andererseits wurde die Berechnung nach deren Substitution durch null wiederholt. Dies geschah deshalb, da gemäß Definition für Cohens Kappa ein Wertebereich von null bis eins vorgesehen ist. Negative Werte entstehen aus einer Kombination von schlechter Reliabilität und niedriger Probandenzahl und verschieben die Mittelwerte ungerechtfertigt nach unten. Durch die Substitution negativer Werte durch null kann eine Verzerrung des Ergebnisses verhindert werden. Um die Ergebnisse zu veranschaulichen, wurde für die Interrater- und Intrarater-Reliabilität auch die **prozentuelle Übereinstimmung** berechnet. Sie wird am Anfang des jeweiligen Kapitels beschrieben.

6.3.1.1 Vorgehensweise bei derAuswertung

Zur Erklärung der Vorgehensweise bei der Auswertung sind die Testergebnisse in der folgenden Kontingenztabelle (Tabelle 3) schematisiert dargestellt. Die Anzahl der beobachteten Übereinstimmungen (O_{ii}) in der Diagonale ist jeweils dunkelblau, die Anzahl nicht übereinstimmender Ergebnisse hellblau gekennzeichnet. Die gelb gekennzeichneten Felder stellen berechnete Werte dar, wobei C_i die Spaltensummen und R_i die Zeilensummen bezeichnen.

Resultate		VL	V	VR	HR	H	HL	0	Reihensummen (R)
Osteopath 2	VL	O_{11}	O_{12}	O_{13}	O_{14}	O_{15}	O_{16}	O_{17}	R_1
	V	O_{21}	O_{22}	O_{23}	O_{24}	O_{25}	O_{26}	O_{27}	R_2
	VR	O_{31}	O_{32}	O_{33}	O_{34}	O_{35}	O_{36}	O_{37}	R_3
	HR	O_{41}	O_{42}	O_{43}	O_{44}	O_{45}	O_{46}	O_{47}	R_4
	H	O_{51}	O_{52}	O_{53}	O_{54}	O_{55}	O_{56}	O_{57}	R_5
	HL	O_{61}	O_{62}	O_{63}	O_{64}	O_{65}	O_{66}	O_{67}	R_6
	0	O_{71}	O_{72}	O_{73}	O_{74}	O_{75}	O_{76}	O_{77}	R_7
Spaltensummen (C)		C_1	C_2	C_3	C_4	C_5	C_6	C_7	n

Tabelle 3: 7x7-Kontingenztabelle der beobachteten Häufigkeiten von Beobachtungen

Die erwartete Anzahl aufgrund zufälliger Verteilung ergibt sich für jedes Datenfeld aus dem Produkt von Zeilen- und Spaltensumme dividiert durch die Gesamtzahl (n) der Beobachtungen:

$$E_{ij} = \frac{R_i \cdot C_j}{n}$$

Somit erhält man eine Kontingenztabelle mit den erwarteten Häufigkeiten, die schematisch in der folgenden Tabelle 4dargestellt ist:

Resultate		VL	V	VR	HR	H	HL	0	Reihensummen (R)
Osteopath 2	VL	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	R_1
	V	E_{21}	E_{22}	E_{23}	E_{24}	E_{25}	E_{26}	E_{27}	R_2
	VR	E_{31}	E_{32}	E_{33}	E_{34}	E_{35}	E_{36}	E_{37}	R_3
	HR	E_{41}	E_{42}	E_{43}	E_{44}	E_{45}	E_{46}	E_{47}	R_4
	H	E_{51}	E_{52}	E_{53}	E_{54}	E_{55}	E_{56}	E_{57}	R_5
	HL	E_{61}	E_{62}	E_{63}	E_{64}	E_{65}	E_{66}	E_{67}	R_6
	0	E_{71}	E_{72}	E_{73}	E_{74}	E_{75}	E_{76}	E_{77}	R_7
Spaltensummen (C)		C_1	C_2	C_3	C_4	C_5	C_6	C_7	n

Tabelle 4: 7x7-Kontingenztabelle der erwarteten Häufigkeiten von Beobachtungen

Somit ergibt sich die Wahrscheinlichkeit einer zufälligen Übereinstimmung P_e aus der Summe der erwarteten Übereinstimmungen (E_{ii}) in den Diagonalfeldern dividiert durch die Gesamtanzahl der Beobachtungen (n):

$$P_e = \frac{\left(\sum_{i=1}^{n} E_{ii}\right)}{n}$$

Die Wahrscheinlichkeit der beobachteten Übereinstimmung P_o wird analog aus der Summe der beobachteten Übereinstimmungen (aus den Diagonalfeldern der Originaltabelle) dividiert durch die Gesamtanzahl der Beobachtungen (n) berechnet:

$$P_o = \frac{\left(\sum_{i=1}^{n} O_{ii}\right)}{n}$$

Für die Berechnung von Cohen's Kappa benötigt man grundsätzlich zumindest zwei unterschiedliche Antwortmöglichkeiten und in SPSS (und auch anderer Statistiksoftware wie SAS) zusätzlich symmetrische Wertebereiche. Das heißt, jeder der beiden

Untersucher muss zumindest einmal ein Ergebnis des jeweils anderen gefunden haben, unabhängig davon ob bei derselben oder einer anderen Testperson. Hat beispielsweise einer der Tester auch nur einmal einen Zug nach links vorne angegeben und der zweite Tester diese Richtung bei keinem einzigen der 18 Patienten wahrgenommen, so ist Cohen's Kappa mit Hilfe der Statistiksoftware nicht berechenbar. Es handelt sich hierbei allerdings nicht um eine methodische Voraussetzung, sondern um einen Programmfehler.

Beim aktuellen Test treten bei annähernd allen Testerpaaren Probleme mit dieser Symmetrie der Wertebereiche auf. Dasselbe Problem tritt auch bei ein und demselben Tester bei jenen Wiederholungsuntersuchungen auf, bei denen einzelnen Untersuchungsergebnissen keine gleichwertigen Antworten in den Vergleichsuntersuchungen gegenüberstehen.

Daher wurde erwogen, diese Ergebnisse und die entsprechenden Resultate der zweiten Messung (d.h. beim selben Probanden) nicht in der Auswertung zu berücksichtigen und damit Maximalwerte für die Übereinstimmung zwischen den einzelnen Testern bzw. die Wiederholbarkeit der Untersuchung zu berechnen. Das hätte allerdings zur Folge, dass v.a. bei der Bestimmung der Intrarater-Reliabilität teilweise mehr als die Hälfte der Vergleiche nicht in die Auswertung miteinbezogen werden hätten können. Aus diesem Grund wurden die Kappa-Werte mit Hilfe einer eigens dafür angefertigten Datenbank (Microsoft® Access® 2000) berechnet.

Zur Validierung, ob die damit gewonnenen Ergebnisse mit denen der Statistiksoftware übereinstimmen, wurden diese anhand mehrerer Datensätze mit beiden Methodenausgewertet und die Ergebnisse verglichen, wobei für jene Fälle, in denen keine symmetrischen Wertebereiche vorlagen, Kappa nach paarweiser Elimination der bei einem der beiden Tester nie vorkommenden Werte berechnet wurde.

		Kappa	
Testerpaar	n	SPSS 14.0	Datenbank
T1/T6	18	-0,03	-0,03
T2/T5	16	0,31	0,31
T3/T5	13	0,05	0,05
T4/T5	11	0,15	0,15

Tabelle 5:Vergleich der mit SPSS 14.0 berechneten Werte gegenüber den mit der eigens für diese Studie angefertigten Datenbank berechneten Werten

Die mit Hilfe der Datenbankabfragen berechneten Werte stimmen mit den Ergebnissen aus SPSS 14.0 überein, es kann also davon ausgegangen werden, dass die Programmierung der Datenbank korrekt durchgeführt wurde (Tabelle 5).

Nachdem – den Ergebnissen vorgreifend – die Ergebnisse keine Größenordnung erreichten, die über eine zufällige Übereinstimmung der Untersuchungsergebnisse hinausgeht, wurde darauf verzichtet, für diesen Fall vorgesehene weitere Untersuchungen mit anderenProbanden, jedoch denselbenTestern, durchzuführen, die dem Zweck dienen sollten, einerseits symmetrische Wertebereiche zu gewährleisten und andererseits die Ergebnisse mit einer höheren Anzahl an Vergleichen abzusichern.

Um zu untersuchen, ob der Test in einer der beiden Körperachsen der Probanden (posterior/anterior und links/rechts) verlässlicher ist als in der anderen, wurden die Untersuchungsergebnisse auch in dieser Hinsicht ausgewertet. Es wird also evaluiert ob die Unterscheidung parietales/viszerales Problem verlässlicher ist als die Unterscheidung Dysfunktion links/rechts. Zusätzlich erreicht man mit diesem Schritt eine Verringerung der möglichen Ergebnisse von sieben (vorne/vorne rechts/ vorne links/hinten/hinten rechts/hinten links/anders) auf jeweils drei, indem jeweils nur die für die Achsen relevanten Informationen aus den Originalantworten verwendet wurden (vorne/hinten/weder vorne noch hinten, bzw. rechts/links/weder links noch rechts). Beispielsweise wurde dafür der Originalantwort „vorne“ bei der Betrachtung der Rechts-Links-Achse „weder links noch rechts“ zugeordnet, der Antwort „links vorne“ „links“.

6.3.2 Die Interrater-Reliabilität des Global Listenings

6.3.2.1 Die prozentuelle Übereinstimmung

In Tabelle 6ist die Anzahl der übereinstimmenden Untersuchungsergebnisse (n) aller 15 Testerpaare, sowie der prozentuelle Anteil (%) der übereinstimmenden an allen Untersuchungsergebnissen zusammengefasst. Weiters werden für jedes einzelne Testerpaar der Mittelwert und die Standardabweichung der prozentuellen Übereinstimmung der Ergebnisse aller drei Untersuchungsdurchgänge angegeben (MW(3), bzw. SD(3)), der Mittelwert und die Standardabweichung aller Prozentwerte jedes einzelnen Untersuchungsdurchgangs (MW(15) bzw. SD(15)) und zuletzt der Mittelwert und die Standardabweichung aller 45 Prozentwerte zusammen (MW(45) bzw. SD(45)).

Testerpaarung	**Durchgang 1**		**Durchgang 2**		**Durchgang 3**		**MW(3)**	**SD(3)**
Ex1/Ex2	**n**	**[%]**	**n**	**[%]**	**n**	**[%]**	**[%]**	**[%]**
T1/T2	1	5,6	0	0	2	20	8,5	10,3
T1/T3	3	16,7	1	10	1	10	12,2	3,8
T1/T4	3	16,7	4	40	2	20	25,6	12,6
T1/T5	3	16,7	2	20	2	20	18,9	1,9
T1/T6	2	11,1	0	0	0	0	3,7	6,4
T2/T3	2	11,1	4	40	1	10	20,4	17,0
T2/T4	6	33,3	2	20	4	40	31,1	10,2
T2/T5	7	38,9	4	40	2	20	33,0	11,2
T2/T6	1	5,6	5	50	1	10	21,9	24,5
T3/T4	1	5,6	4	40	2	20	21,9	17,3
T3/T5	3	16,7	1	10	2	20	15,6	5,1
T3/T6	2	11,1	3	30	1	10	17,0	11,2
T4/T5	4	22,2	3	30	2	20	24,1	5,3
T4/T6	2	11,1	4	40	1	10	20,4	17,0
T5/T6	6	33,3	2	20	1	10	21,1	11,7
MW(15)		17,0		26,0		16,0	**MW(45)**	19,7
SD(15)		10,6		15,9		9,1	**SD(45)**	12,8

Tabelle 6: Anzahl der übereinstimmenden Untersuchungsergebnisse (n) von jeweils zwei Testern und prozentueller Anteil (%) der Übereinstimmungen an der Gesamtanzahl der Untersuchungsergebnisse. MW(3)/SD(3) ist der Mittelwert bzw. die Standardabweichung der prozentuellen Übereinstimmung der drei Durchgänge jedes einzelnen Testerpaars, MW(15)/SD(15) der Mittelwert bzw. die Standardabweichung der 15 Vergleiche bei einem Durchgang und MW(45)/SD(45) bei allen 45 Vergleichen.

Im ersten Durchgang wurde eine mittlere Übereinstimmung von 17 % ermittelt, wobei die Ergebnisse minimal bei einem der 18 Patienten (5,6 %, Paare T1/T2, T2/T6, T3/T4) und maximal bei sieben (38,9 %, Paar T2/T5) übereinstimmen. Im zweiten Durchgang beträgt der Mittelwert 26 %. Hier wurde minimal keine einzige Übereinstimmung (Paare T1/T2, T1/T6) und wurden maximal fünf Übereinstimmungen (50 %, Paar T2/T6) bei den zehn Patienten gefunden. Im dritten Durchgangliegt der Mittelwert bei 16 %. Maximal stimmen vier Untersuchungsergebnisse bei den zehn Patienten überein (Paar T2/T4, 40 %) und minimal stimmte keines überein (Paar T1/T6).

Dem höheren Mittelwert im zweiten Durchgang, der auf eine bessere Übereinstimmung schließen lassen könnte, steht allerdings auch eine höhere Streuung der Einzelwerte um den Mittelwert gegenüber.

Der Gesamtmittelwert aller 45 Vergleiche (jeweils 15 Testerpaare in drei Durchgängen) liegt bei 19,7 %. Das bedeutet, dass im Durchschnitt nur bei etwa jedem fünften Patienten die Testergebnisse zweier Tester übereinstimmen.

6.3.2.2 Darstellung der Interrater-Reliabilität mittels Kappa-Wert

In Tabelle 7 sind die Kappa-Werte für sämtliche Testerpaare und alle drei Durchgänge angegeben (Auswertung mittels Datenbankabfragen).

Durchgang 1 (n=18) 95 %CI: -0,05 - 0,08 (0,01 - 0,10)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,08	0,01	0,00	0,06	-0,03
	T2		-0,07	0,21	0,27	-0,13
	T3			-0,17	0,03	-0,08
	T4				0,09	-0,07
	T5					0,20
Durchgang 2 (n=10) 95 %CI: 0,00 - 0,17 (0,06 - 0,18)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,15	-0,08	0,27	0,07	-0,16
	T2		0,14	-0,04	0,24	0,35
	T3			0,20	-0,11	0,11
	T4				0,16	0,26
	T5					0,01
Durchgang 3 (n=10) 95 %CI: -0,07 - 0,06 (0,00 - 0,08)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	0,12	-0,10	-0,01	0,04	-0,20
	T2		-0,05	0,30	0,07	-0,15
	T3			0,04	0,02	-0,05
	T4				0,05	-0,10
	T5					-0,06
Mittelwert Durchgang 1-3	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,04	-0,06	0,09	0,06	-0,13
	T2		0,01	0,16	0,19	0,02
	T3			0,02	-0,02	-0,01
	T4				0,10	0,03
	T5					0,05

Tabelle 7: Cohens Kappa für sämtliche Testerpaare und alle drei Durchgänge sowie Mittelwerte der drei Kappa-Werte aus den drei Durchgängen (grau: $\kappa \geq 0,2$), 95 %-Konfidenzintervalle vor (95 %-CI) und nach Substitution der negativen Werte durch null (in Klammer).

Die meisten Ergebnisse in allen drei Durchgängen deuten auf eine rein zufällige oder nur gering darüber hinausgehende Übereinstimmung der Untersuchungsergebnisse

hin. Werte über den von Fjellner et al. (1999) geforderten 0,40 konnten nicht erreicht werden. Die besten Werte liegen im Bereich leidlicher Reliabilität (κ >0,20).

Im ersten Durchgang beträgt der höchste ermittelte Wert für Cohen's Kappa κ= 0,27 (Testerpaar T2/T5). Lediglich zwei weitere Paare (T2/T4 und T5/T6) weisen κ-Werte über oder gleich 0,2 auf. Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt κ =0,02 (nach Substitution der negativen Werte durch null κ= 0,06), das heißt die aktuell beobachtete Übereinstimmung liegt nur vernachlässigbar über einer zufälligen Übereinstimmung derTestergebnisse.

Beim zweiten Durchgang beträgt der Maximalwert von Kappa κ= 0,35 (Testerpaar T2/T6), bei drei weiteren Paaren liegen die Werte über 0,2 (T1/T4, T2/T5 und T4/T6). Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt κ= 0,08 (nach Substitution der negativen Werte durch null κ= 0,12), das heißt die aktuell beobachtete Übereinstimmung ist zwar höher als im ersten Durchgang, liegt aber weiterhin nur gering über einer zufälligen Übereinstimmung.

Im dritten Durchgang beträgt der Maximalwert von Kappa κ= 0,30 (Testerpaar T2/T4), jedoch erreicht kein weiteres Paar Werte über 0,2. Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt κ= -0,01 (nach Substitution der negativen Werte durch null κ= 0,04), das heißt, die mittlere aktuell beobachtete Übereinstimmung liegt weiterhin nur vernachlässigbar über der Übereinstimmung, die die Tester zufällig erreichen würden.

Anhand dieser Ergebnisse kann abgeleitet werden, dass kein Lerneffekt auftritt: Die κ-Werte des dritten Durchgangs sind im Durchschnitt geringer als die des ersten.

Betrachtet man die Mittelwerte der drei κ-Werte der drei Durchgänge für die einzelnen Testerpaare, so erreicht kein Paar ein κ> 0,20, wobei der höchste Wert von T2/T5 κ= 0,19 beträgt. Jedoch erreicht auch dieses Paar bei einem Durchgang nur ein κ= 0,07.

Das 95 %-Konfidenzintervall des Mittelwerts aller 45 Kappa-Werte aus allen drei Durchgängen beträgt -0,01 - 0,07 (nach Substitution der negativen Werte durch null 0,00 - 0,10).

6.3.2.3 Darstellung der Ergebnisse in Bezug auf eine Rechts-Links-Körperachse

In Tabelle 8 sind die Kappa-Werte für die Körperachse rechts-links aller 15 Testerpaare angegeben (mögliche Richtungen des Zugs: links/rechts/weder links noch rechts; Auswertung mittels Datenbankabfragen).

Durchgang 1 (n=18)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,02	-0,27	-0,04	0,12	0,01
	T2		-0,21	0,06	0,30	-0,07
95 %CI:	T3			-0,16	-0,09	-0,14
-0,09 - 0,06	T4				0,18	0,00
(0,06 - 0,18)	T5					0,10
Durchgang 2 (n=10)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,09	-0,25	0,06	-0,13	-0,13
	T2		0,75	0,25	0,29	0,29
95 %CI:	T3			0,13	0,15	0,15
0,04 - 0,33	T4				0,57	0,15
(0,10 - 0,34)	T5					0,57
Durchgang 3 (n=10)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	0,31	-0,33	0,19	-0,31	-0,25
95 %CI:	T2		0,04	0,31	0,15	-0,25
-0,15 - 0,11	T3			-0,17	0,33	-0,15
(0,04 - 0,17)	T4				0,02	-0,43
	T5					0,25
Mittelwert Durchgang 1-3	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	0,07	-0,29	0,07	-0,11	-0,12
	T2		0,19	0,21	0,25	-0,01
	T3			-0,07	0,13	-0,05
	T4				0,26	-0,09
	T5					0,30

Tabelle 8:Cohens Kappa für sämtliche Testerpaare und alle drei Durchgänge berechnet aus den Untersuchungsergebnissen in der Rechts-Links-Achse, sowie Mittelwerte der drei Kappa-Werte (grau: $\kappa \geq 0{,}2$), 95 %-Konfidenzintervalle vor (95 %-CI) und nach Substitution der negativen Werte durch null (in Klammer).

In der Rechts-Links-Achse der Testpersonen sind im ersten Untersuchungsdurchgang kaum Übereinstimmungen innerhalb der Testerpaare erkennbar, die deutlich über eine zufällige Übereinstimmung hinausgehen. Lediglich ein Paar erreicht eine Übereinstimmung, die in $\kappa > 0{,}2$ resultiert (T2/T5: $\kappa = 0{,}30$). Werte im Bereich einer akzeptablen Reliabilität ($\kappa > 0{,}40$) konnten nicht gefunden werden. Der Mittelwert aller 15 Vergleiche dieses Durchgangs beträgt $\kappa = -0{,}02$; nach Substitution negativer Werte durch null: $\kappa = 0{,}05$.

Im zweiten Durchgang erreicht ein Testerpaar (T2/T3) mit $\kappa = 0{,}75$ den Maximalwert des Durchgangs und somit einen Wert, der im Bereich beträchtlicher Reliabilität liegt. Weitere fünf Paare (T1/T3, T2/T4, T2/T5, T2/T6, T4/T5) erreichen ein Kappa von $\kappa > 0{,}20$, die anderen Paare zeigen Übereinstimmungen, die durch $\kappa \leq 0{,}2$ charakterisiert werden. Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt $\kappa = 0{,}18$, nach Substitution der negativen Werte durch null $\kappa = 0{,}22$, das heißt die aktuell beobachtete Übereinstimmung ist deutlich höher als im ersten Untersuchungsdurchgang.

Im dritten Untersuchungsdurchgang beträgt der Maximalwert $\kappa = 0{,}33$ (T3/5), drei weitere Paare (T1/2, T2/4, T5/6) zeigen Übereinstimmungen mit $\kappa > 0{,}2$. Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt $\kappa = -0{,}02$, nach Substitution der negativen Werte durch null $\kappa = 0{,}11$.

Wie schon bei den Originaldaten ist die beste Übereinstimmung im zweiten Untersuchungsdurchgang zu beobachten. Die Werte des dritten unterscheiden sich kaum von denen des ersten Untersuchungsdurchgangs.

Berechnet man die 15 Mittelwerte der drei Kappawerte jedes der Testerpaare, so liegen lediglich drei über $\kappa = 0{,}20$, während sieben auf auch im Durchschnitt zufällige Übereinstimmung hinweisen. Selbst bei jenen drei Testerpaaren, die die höchste mittlere Übereinstimmung aufweisen, ist die Streuung der κ-Werte hoch, sodass auch bei diesen keine verlässliche Übereinstimmung abzuleiten ist.

Das 95 %-Konfidenzintervall des Mittelwerts aller 45 Kappa-Werte aus allen drei Untersuchungsdurchgängen beträgt -0,02 - 0,12 (nach Substitution der negativen Werte durch null 0,00 - 0,18).

6.3.2.4 Darstellung der Ergebnisse in Bezug auf eine anterior-posteriore Körperachse

In Tabelle 9 sind die Kappa-Werte für die anterior-posteriore Körperachse aller 15 Testerpaare angegeben (mögliche Richtungen des Zugs: vorne/hinten/weder vorne noch hinten; Auswertung mittels Datenbankabfragen).

Durchgang 1 (n=18) 95 %CI: -0,10 - 0,09 (0,02 - 0,14)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,23	0,00	0,16	-0,08	-0,11
	T2		-0,10	0,15	0,10	0,08
	T3			-0,35	0,24	-0,03
	T4				-0,19	-0,11
	T5					0,40
Durchgang 2 (n=10) 95 %CI: -0,07 - 0,11 (0,01 - 0,14)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,03	-0,17	0,06	0,10	-0,09
	T2		-0,07	-0,15	0,28	0,24
	T3			0,00	-0,29	0,06
	T4				-0,01	0,37
	T5					-0,04
Durchgang3 (n=10) 95 %CI: -0,11 - 0,11 (0,02 - 0,16)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	0,03	-0,16	-0,15	0,26	-0,17
	T2		0,10	0,49	-0,33	0,04
	T3			0,26	-0,06	-0,18
	T4				-0,21	0,17
	T5					-0,09
Mittelwert Durchgang 1-3	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0,08	-0,11	0,02	0,10	-0,12
	T2		-0,02	0,17	0,02	0,12
	T3			-0,03	-0,04	-0,05
	T4				-0,14	0,14
	T5					0,09

Tabelle 8: Cohens Kappa für sämtliche Testerpaare und alle Durchgänge, berechnet aus den Untersuchungsergebnissen in der anterior-posterioren Körperachse, sowie Mittelwerte der drei Kappa-Werte (grau: $\kappa \geq 0{,}2$), 95 %-Konfidenzintervalle vor (95 %-CI) und nach Substitution der negativen Werte durch null (in Klammer).

Im ersten Durchgang erreicht Testerpaar T5/6 mit κ =0,4, einem Wert der an der Grenze für akzeptable Reliabilität liegt, den Maximalwert des Durchgangs. Einen Wert κ> 0,2 erreicht ansonsten nur ein weiteres Paar (T3/4). Der Mittelwert aller 15 κ-Werte des Durchgangs beträgt κ= 0,00 (nach Substitution der negativen Werte durch null κ= 0,08)

Im zweiten Durchgang werden Kappa-Werte κ> 0,2 von drei Testerpaaren erreicht (Maximalwert T4/T6: κ= 0,37). Alle anderen Paare zeigen Übereinstimmungen, die durch κ< 0,2 charakterisiert werden. Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt κ= 0,02 (nach Substitution der negativen Werte durch null κ= 0,07), also ebenfalls ein sehr geringer Wert.

Im dritten Untersuchungsdurchgang beträgt der Maximalwertκ = 0,49 (T2/T4)und liegt somit im Bereich akzeptabler Reliabilität, zwei weitere Paare zeigen Übereinstimmungen mit κ> 0,2 (T1/5, T3/4). Der Mittelwert aller 15 κ-Werte dieses Durchgangs beträgt κ= 0,00 (nach Substitution der negativen Werte durch null κ= 0,09).

Im Vergleich zur Rechts-Links-Achse sind die Ergebnisse der drei Durchgängezwar einheitlicher, bewegen sich aber nur geringfügig über den Werten, die bei rein zufälliger Übereinstimmung zu erwarten wären. Keines der 15 Testerpaare weist einen Mittelwert $\kappa\geq$ 0,20 auf.

Das 95 %-Konfidenzintervall des Mittelwerts aller 45 Kappa-Werte aus allen drei Durchgängen beträgt -0,05 - 0,06 (nach Substitution der negativen Werte durch null 0,00 - 0,12).

6.3.3 Die Intrarater-Reliabilität des Global Listenings

6.3.3.1 Die prozentuelle Übereinstimmung

In Tabelle 10 ist die Anzahl der bei den Wiederholungsmessungen übereinstimmenden Untersuchungsergebnisse (n) der einzelnen Tester sowie der prozentuelle Anteil (%) der übereinstimmenden Untersuchungsergebnisse zusammengefasst. Weiters werden für jeden einzelnen Tester der Mittelwert und die Standardabweichung der prozentuellen Übereinstimmung bei allen drei Untersuchungsdurchgängen angegeben. In Spalte 1/2 werden die Ergebnisse des ersten Durchgangs mit jenen des zweiten verglichen, in Spalte 1/3 mit jenen des dritten. Die Ergebnisse des Vergleichs

des zweiten mit dem dritten Durchgang sind in Spalte 2/3 dargestellt. Aus diesen drei Werten wurde zusätzlich der Mittelwert und die Standardabweichung berechnet.

Tester	Test	1/2	1/3	2/3	Mittelwert (%)	SD (%)
T1	n	0	5	1		
	[%]	0	50	10	20,0	26,5
T2	n	2	2	4		
	[%]	20	20	40	26,7	11,5
T3	n	4	2	2		
	[%]	40	20	20	26,7	11,5
T4	n	2	1	4		
	[%]	20	10	40	23,3	15,3
T5	n	1	3	0		
	[%]	10	30	0	13,3	15,3
T6	n	2	4	2		
	[%]	20	40	20	26,7	11,5

Tabelle 10: Anteil der übereinstimmenden Untersuchungsergebnisse von jeweils zwei Untersuchungen ein und desselben Testers am selben Patienten an der Gesamtanzahl der Untersuchungsergebnisse (in % und absoluten Zahlen (n)).

Schlechtestenfalls wurde bei keinem der zehn Patienten vom selben Tester das Untersuchungsergebnis bestätigt (0 %, T1: Untersuchungen 1 und 2, T5: Untersuchungen 2 und 3), bestenfalls bei fünf (50 %, T1: Untersuchungen 1 und 3). Dass diese Extremwerte bei ein und demselben Tester auftreten, zeigt, wie sehr die Ergebnisse streuen.

Die im Durchschnitt geringste Wiederholbarkeit weisen die Untersuchungen bei Tester T5 auf (13,3 %), der bei nur vier der 30 Vergleiche Übereinstimmungen erzielte. Die im Durchschnitt höchste Wiederholbarkeit liefern die Tester T2, T3 und T6 mit jeweils 26,7 % aller Vergleiche.

6.3.3.2 Darstellung der Intrarater-Reliabilität mittels Kappa-Wert

In Tabelle 11 werden die Kappa-Werte für die Intrarater-Reliabilität wie folgt dargestellt: In Spalte 1/2 werden die Ergebnisse eines Testers aus dem ersten Durchgang mit seinen Ergebnissen aus dem zweiten verglichen, in Spalte 1/3 mit jenen des dritten. Die Ergebnisse des Vergleichs des zweiten mit dem dritten Durchgang sind in Spalte 2/3 dargestellt. Weiters sind der Mittelwert (MW(3)), die Standardabweichung (SD(3)) und der Vertrauensbereich (95 %-CI) der drei Kappa-Werte für jeden Tester angeführt. Das 95 %-Konfidenzintervall ist erst ab drei Werten berechenbar. In dieser Studie sind pro Tester drei Werte gegeben, die auch noch stark divergieren.Daher sind die Intervalle groß und bringen kaum Informationsgewinn. Die nach Substitution der negativen Werte durch null berechneten Bereiche (95 %-CI*) sind kleiner und bringen somit mehr Information. Der Mittelwert und die Standardabweichung für jeden Vergleich zweier Durchgänge sind mit MW(6) bzw. SD(6) angeführt, und zuletzt sind sowohl der Mittelwert als auch die Standardabweichung aller 18 Kappa-Werte angegeben.

Tester	1/2	1/3	2/3	**MW(3)**	**SD(3)**	**95 %-CI**	**95 %-CI***
T1	-0,14	0,38	-0,06	0,06	0,28	-0,63 – 0,75	-0,12 - 0,37
T2	0,04	0,01	0,30	0,12	0,16	-0,27 – 0,51	-0,06 - 0,30
T3	0,25	0,02	0,00	0,09	0,14	-0,25 – 0,43	-0,07 - 0,25
T4	-0,03	-0,11	0,27	0,04	0,20	-0,45 – 0,54	-0,09 - 0,27
T5	-0,08	0,11	-0,19	-0,05	0,15	-0,43 – 0,32	-0,04 - 0,11
T6	0,02	0,25	0,01	0,10	0,13	-0,24 – 0,43	-0,06 - 0,25
MW(6)	0,01	0,11	0,06	**MW(18)**	0,06	-0,02 – 0,14	0,03 – 0,15
SD(6)	0,13	0,18	0,19	**SD(18)**	0,17		

Tabelle 11: Kappa-Werte für die Übereinstimmung der Ergebnisse der einzelnen Tester in den drei Durchgängen (hellgrau: $\kappa \geq 0{,}2$), Mittelwerte (MW) und Standardabweichung (SD) wie im Text beschrieben, 95 %-Konfidenzintervall vor (95 %-CI) und nach (95 %-CI*) Substitution durch null

Wie schon die Interrater-Reliabilität ist auch die Intrarater-Reliabilität des Global-Listenings äußerst gering. Kappa-Werte > 0,2 treten nur bei vier der 18 Vergleiche auf(Maximalwert 0,38), wobei auch bei diesen vier Testern (T1, T2, T4 und T6) die Streuung sehr hoch ist (vgl. Standardabweichung der drei Einzelwerte SD(3)), und

bei den jeweils anderen beiden Vergleichen nur Werte zu beobachten sind, die kaum oder nicht über jene zufälliger Übereinstimmung hinausgehen. So weist der Tester mit der im Durchschnitt höchsten Wiederholbarkeit der Messergebnisse einen Mittelwert (MW(3)) von κ= 0,12 auf.

Der höchste Mittelwert aller Tester (MW(6)) kann beim Vergleich des ersten mit dem dritten Untersuchungsdurchgang beobachtet werden (κ= 0,11).

Der Mittelwert der Kappa-Werte aller 18 Vergleiche (MW(18)) beträgt 0,06 (nach Substitution der negativen Werte durch nullκ= 0,09). Der Vertrauensbereich des Mittelwerts (95 %-Konfidenzintervall) liegt zwischen κ= -0,02 und κ= 0,13 (nach Substitution der negativen Werte durch null zwischen κ= 0,03 und κ= 0,15).

6.3.3.3 Darstellung der Ergebnisse in Bezug auf eine Rechts-Links-Körperachse

In Tabelle 12 werden die Kappa-Werte für die Intrarater-Reliabilität in der Rechts-Links-Achse der Testpersonen wie folgt dargestellt (mögliche Richtungen des Zugs: links/rechts/weder links noch rechts; Auswertung mittels Datenbankabfragen): In Spalte 1/2 werden die Ergebnisse eines Testers aus dem ersten Durchgang mit seinen Ergebnissen aus dem zweiten verglichen, in Spalte 1/3 mit jenen des dritten. Die Ergebnisse des Vergleichs der zweiten mit der dritten Untersuchung sind in Spalte 2/3 dargestellt. Weiters sind der Mittelwert (MW(3)), die Standardabweichung (SD(3)), der Vertrauensbereich (95 %-CI) und der nach Substitution negativer Werte durch null berechnete Vertrauensbereich (95 %-CI*)der drei Kappa-Werte jedes Testers angeführt. Der Mittelwert und die Standardabweichung für jeden Vergleich zweier Durchgänge sind mit MW(6) bzw. SD(6) angeführt, und letztlich sind sowohl der Mittelwert als auch die Standardabweichung aller 18 Kappa-Werte angegeben (MW(18), SD(18)).

Tester	1/2	1/3	2/3	MW(3)	SD(3)	95 %-CI	95 %-CI*
T1	-0,29	0,25	-0,25	-0,09	0,30	-0,84 – 0,65	-0,08 - 0,25
T2	0,20	0,06	0,75	0,34	0,37	-0,57 – 1,24	-0,08 - 0,75
T3	0,33	-0,15	0,04	0,07	0,25	-0,52 – 0,67	-0,08 - 0,33
T4	-0,18	-0,31	0,48	0,00	0,42	-1,05 – 1,04	-0,15 - 0,47
T5	0,09	0,09	0,29	0,16	0,11	-0,13 – 0,44	0,03 - 0,29
T6	0,02	0,18	0,11	0,10	0,08	-0,09 – 0,30	0,01 - 0,19
MW(6)	0,03	0,02	0,24	MW(18)	0,10	-0,04 – 0,23	0,07 – 0,25
SD(6)	0,23	0,21	0,35	SD(18)	0,28		

Tabelle 12: Kappa-Werte für die Übereinstimmung der Ergebnissen der einzelnen Tester in den drei Durchgänge in der Rechts-Links-Achse der Testpersonen (hellgrau: $\kappa \geq 0{,}2$), Mittelwerte (MW) und Standardabweichung (SD) wie im Text beschrieben, 95 %-Konfidenzintervall vor (95 %-CI) und nach (95 %-CI*) Substitution durch null

Kappa-Werte > 0,2 treten nur bei sechs der 18 Vergleiche auf, wobei auch bei jenen fünf Testern (T1, T2, T3, T4, T5), die Werte über 0,2 erreichen (Maximalwert 0,75), bei den jeweils anderen beiden Vergleichen nur Werte zu beobachten sind, die nicht oder nicht nennenswert über jene zufälliger Übereinstimmung hinausgehen. Das ist auch bei Tester T2 der Fall, der den maximalen Mittelwert der drei Kappa-Werte (MW(3)) von $\kappa = 0{,}34$ erreicht.

Auffällig ist, dass die Vergleiche der Ergebnisse des zweiten mit jenen des dritten Durchgangs sich deutlicher von zufälliger Übereinstimmung abheben als die Vergleiche der Ergebnisse des ersten mit jenen des zweiten Durchgangs(MW(6): $\kappa = 0{,}24$ vs. $\kappa = 0{,}03$).

Der Mittelwert der Kappa-Werte aller 18 Vergleiche (MW(18)) beträgt 0,10 (nach Substitution der negativen Werte durch null$\kappa = 0{,}16$). Der Vertrauensbereich des Mittelwerts (95 %-Konfidenzintervall) liegt zwischen $\kappa = -0{,}03$ und $\kappa = 0{,}22$ (nach Substitution der negativen Werte durch null zwischen $\kappa = 0{,}07$ und $\kappa = 0{,}25$).

6.3.3.4 Darstellung der Ergebnisse in Bezug auf eine anterior-posteriore Körperachse

In Tabelle13 werden die Kappa-Werte für die Intrarater-Reliabilität in der anterior-posteriorenAchse der Testpersonen wie folgt dargestellt (mögliche Richtungen des Zugs: anterior/posterior/weder anterior noch posterior; Auswertung mittels Datenbankabfragen): In Spalte 1/2 werden die Ergebnisse eines Testers aus dem ersten Durchgang mit seinen Ergebnissen aus dem zweiten verglichen, in Spalte 1/3 mit jenen des dritten. Die Ergebnisse des Vergleichs der zweiten mit der dritten Untersuchung sind in Spalte 2/3 dargestellt. Weiters sind der Mittelwert (MW(3)), die Standardabweichung (SD(3)), der Vertrauensbereich (95 %-CI) und der nach Substitution negativer Werte durch null berechnete Vertrauensbereich (95 %-CI*) der drei Kappa-Werte für jeden Tester angeführt. Der Mittelwert und die Standardabweichung für jeden Vergleich zweier Durchgänge sind mit MW(6) bzw. SD(6) angeführt, und letztlich sind sowohl der Mittelwert als auch die Standardabweichung aller 18 Kappa-Werte angegeben (MW(18), SD(18)).

Tester	1/2	2/3	1/3	MW(3)	SD(3)	95 %-CI	95 %-CI*
T1	0,11	0,21	0,67	0,33	0,30	-0,41 – 1,07	-0,01 - 0,67
T2	0,00	0,13	0,17	0,10	0,09	-0,12 – 0,32	0,00 - 0,20
T3	0,43	-0,14	-0,06	0,08	0,31	-0,68 – 0,84	-0,14 - 0,42
T4	-0,23	0,12	-0,29	-0,13	0,22	-0,68 – 0,41	-0,04 - 0,12
T5	-0,09	-0,52	0,06	-0,18	0,30	-0,93 – 0,56	-0,02 - 0,06
T6	0,08	-0,09	0,33	0,11	0,21	-0,41 – 0,63	-0,06 - 0,33
MW(6)	0,05	-0,05	0,15	MW(18)	0,05	-0,08 – 0,18	0,04 - 0,21
SD(6)	0,22	0,27	0,33	SD(18)	0,27		

Tabelle 13:Kappa-Werte für die Übereinstimmung der Ergebnisse der einzelnen Tester in den drei Durchgängen in der anterior-posteriorenAchse der Testpersonen (hellgrau: $\kappa \geq 0,2$), Mittelwerte (MW) und Standardabweichung (SD) wie im Text beschrieben, 95 %-Konfidenzintervall vor (95 %-CI) und nach (95 %-CI*) Substitution durch null

Auch in der anterior-posterioren Achse der Testpersonen überwiegen Kappa-Werte, die kaum oder nicht über das Ausmaß einer zufälligen Übereinstimmung hinausgehen. Kappa-Werte > 0,2 treten nur bei vier der 18 Vergleiche auf (Maximalwert κ = 0,67). Betrachtet man jedoch die Mittelwerte der drei Einzelwerte(MW(3)), so weist nur einer der sechs Tester (T1) mitκ = 0,33 einen Mittelwert MW(3) κ> 0,2 auf.

Der Mittelwert der Kappa-Werte aller 18 Vergleiche (MW(18)) beträgt 0,05 (nach Substitution der negativen Werte durch nullκ= 0,13). Der Vertrauensbereich des Mittelwerts (95 %-Konfidenzintervall) liegt zwischen κ= -0,08 und κ= 0,18 (nach Substitution der negativen Werte durch null zwischen κ= 0,04 und κ= 0,21).

Im Gegensatz zur Körperachse rechts-links ist vom Mittelwert der jeweils sechs Kappa-Werte (MW(6)) keine deutliche Verbesserung der Wiederholbarkeit im Vergleich der Ergebnisse der Durchgänge 2 und 3 mit jenen der Durchgänge 1 und 2 abzuleiten (κ=0,05 bzw.κ=-0,05).

Erfahrung

Betrachtet man die Ergebnisse in Bezug auf die Erfahrung der Tester mit dem Global Listening,so fällt auf, dass derTester (T2), der diesen Test in der Praxis nicht anwendet, den besten Mittelwert (MW(3) κ = 0,12) aus allen Vergleichen erreicht. Er erzielt bei der Betrachtung in einer Rechts-Links-Körperachse wiederum das beste Ergebnis, das mit κ = 0,34 deutlich über dem Mittelwert aller 18 Vergleiche (MW(18)κ = 0,10) liegt. In der anterior-posterioren Körperachse liegt er mit κ = 0,10 über dem Mittelwert MW(18) von κ = 0,05, erreicht aber nicht das beste Ergebnis.

Das schlechteste Ergebnis (MW(3) κ = -0,05) erreicht jener Tester (T5), der das Global Listening bei jedem zweiten Patienten und bei jedem Erstbefund verwendet, was eine Erfahrung im Mittelfeld der Tester bedeutet.

Der Tester (T1), der den Test am häufigsten ausführt, nämlich bei jedem Patienten vor und nach der Behandlung, erreicht exakt den Mittelwert MW(18) vonκ = 0,06.

Eine bessere Intrarater-Reliabilität erfahrener Osteopathen kann somit nicht nachgewiesen werden.

6.4 Zusammenfassung

6.4.1 Interrater-Reliabilität

Die meisten der für die 15 Osteopathenpaare aus drei Durchgängen berechneten Werte deuten auf rein zufällige oder nur gering darüber hinausgehende Übereinstimmung und somit schlechte Interrater-Reliabilität des Tests hin. Dies zeigt sich in den Mittelwerten κ = 0,02 bzw. bei Berechnung nach Substitution der negativen Kappa-Werte durch null κ= 0,06 für den ersten, κ = -0,01 bzw. bei Berechnung nach Substitution der negativen Kappa-Werte durch null κ = 0,12 für den zweiten und κ = 0,08 bzw. bei Berechnung nach Substitution der negativen Kappa-Werte durch null κ= 0,04 für den dritten Durchgang. Der höchste erzielte Einzelwert liegt bei κ= 0,35 und wurde im zweiten Durchgang erzielt.

Werden die Interpretationsmöglichkeiten auf drei reduziert – sei es durch Betrachtung der Ergebnisse in Bezug auf eine Rechts-Links-Körperachse oder durch Betrachtung in Bezug auf eine anterior-posteriore Körperachse – werden weiterhin keine Werte erreicht, die den Test als reliabel ausweisen. Der Mittelwert bei Betrachtung in Bezug auf eine Rechts-Links-Achse liegt bei κ= -0,02 im ersten, κ = 0,18 im zweiten und κ = -0,02 im dritten Durchgang. Der Mittelwert bei Betrachtung in Bezug auf eine anterior-posteriore Körperachse liegt beiκ = 0,0 im ersten Durchgang, κ= 0,02 im zweiten Durchgang und κ =0,0 im dritten Durchgang. Somit konnte keine Verbesserung des Ergebnisses durch Reduktion auf drei Interpretationsmöglichkeiten erreicht werden.Auch ist kein Unterschied zwischen den in Bezug auf eine Rechts-Links-Achse zu den in Bezug auf eine anterior-posteriore Achse ausgewerteten Ergebnisse zu sehen.

Eine Steigerung vom ersten bis zum dritten Durchgang ist nicht feststellbar, weshalb ein Lerneffekt ausgeschlossen werden kann.

Für Fragestellung 1[1] hat sich somit die Nullhypothese[2] als richtig erwiesen.

---

[1] Fragestellung 1 lautet: Kommen verschiedene Tester, die am selben Probanden ein Global Listening durchführen, zum gleichen Ergebnis?

6.4.2 Intrarater-Reliabilität

Auch die Intrarater-Reliabilität des Global Listenings ist gering. So beträgt der Mittelwert aller Vergleiche aus drei Durchgängen κ = 0,06. Bemerkenswert ist die hohe Streuung, die sich auch dadurch zeigt, dass der Osteopath mit dem schlechtesten Ergebnis beim Vergleich von Durchgang eins mit Durchgang zwei (0 %) das beste Ergebnis beim Vergleich von Durchgang eins mit Durchgang drei (50 %) erzielte.

Bei Betrachtung des Ergebnisses in Bezug auf eineRechts-Links-Körperachse – was eine Reduktion der Interpretationsmöglichkeiten von sieben auf drei bedeutet - beträgt der Mittelwert κ = 0,10. Bei der Betrachtung in einer anterior-posterioren Körperachse (ebenfalls eine Reduktion auf drei Interpretationsmöglichkeiten) beträgt der Mittelwert κ = 0,05. Somit konnte auch durch Reduktion auf drei Interpretationsmöglichkeiten kein Ergebnis erreicht werden, das eine Intrarater-Reliabilität des Tests anzeigt.

Bemerkenswert ist, dass der einzige Osteopath, der den Test in der Praxis nicht verwendet, den besten Mittelwert aus allen Vergleichen erreicht (κ = 0,12). Auch dieser Wert ist jedoch im Bereich schwacher Reliabilität und die Streuung der Werte des Testers hoch (wie bei den anderen Testern auch). Nicht nachgewiesen werden kann, dass die Erfahrung der Tester die Intrarater-Reliabilität verbessert.

Für Fragestellung 2[3]hat sich somit die Nullhypothese[4] als richtig erwiesen.

Für Fragestellung 3[5] hat sich somit die Nullhypothese[6]als richtig erwiesen.

---

[2] Nullhypothese 1 lautet: Führen mehrere Tester innerhalb eines kurzen Zeitraums am selben Probanden ein Global Listening durch, liegt die Übereinstimmung ihrer Ergebnisse im Bereich zufälliger Übereinstimmung.

[3] Fragestellung 2 lautet: Kommt ein und derselbe Tester, der am selben Probanden mehrmals ein Global Listening durchführt, zum gleichen Ergebnis?

[4] Nullhypothese 2 lautet: Führt ein Tester innerhalb eines kurzen Zeitraumes am selben Probanden mehrmals ein Global Listening durch, liegt die Übereinstimmung seiner Ergebnisse im Bereich zufälliger Übereinstimmung.

[5] Fragestellung 3 lautet: Hat die Erfahrung des Testers einen Einfluss auf die von ihm erreichte Übereinstimmung der Ergebnisse am selben Probanden?

[6] Nullhypothese 3 lautet: Führt ein Tester, der das Global Listening in seiner Praxis häufig anwendet, diesen Test mehrmals am selben Probanden durch, so erreicht er keine höhere Übereinstimmung seiner Ergebnisse als ein Tester, der das Global Listening in seiner Praxis nicht anwendet.

7 Diskussion

Es konnte weder eine Intra- noch eine Interrater-Reliabilität des Global Listening nachgewiesen werden. Mögliche Ursachen werden nachfolgend diskutiert.

7.1 Tester

Zum Vergleich erfahrener Tester mit weniger erfahrenen Testern wurde angestrebt, dass zumindest zwei Tester über langjährige Erfahrung in der Anwendung des Global Listenings verfügen. Dies ist nicht gelungen. In der Literatur wird mehrfach darauf hingewiesen wird, dass das Global Listening ein Test ist, der Erfahrung erfordert (Becker in Brooks 1997;Croibier 2006; Paoletti 2001). In den bisherigen Reliabilitätsstudien findet sich freilich kein Hinweis darauf, dass bei unterschiedlichen Testverfahren erfahrene Tester ein besseres Ergebnis erreichen (siehe Kapitel 4.2).

Die Erfahrung der Tester wurde kategorisiert, indem die Tester angaben, bei wie vielen ihrer Patienten sie das Global Listening anwenden. Berücksichtigt wurde nicht, wie viele Patienten die Tester pro Woche behandeln.Es wurde nur ein unerfahrener Tester fünf erfahrenen Testern gegenübergestellt. Die Fähigkeiten des unerfahrenen Testers sind somit von entscheidender Bedeutung. Die Kraft der Aussage dieser Studie über den Einfluss der Erfahrung der Tester auf das Ergebnis ist also eingeschränkt.

7.2 Testpersonen

Die Anzahl der Probanden ist mit 18 für die Untersuchung der Interrater-Reliabilität bzw. zehn für die Intrarater-Reliabilität (Wiederholungsuntersuchungen) in Anbetracht der Anzahl der möglichen Testergebnisse (sieben) gering. Darauf deuten einerseits jene Untersuchungsergebnisse hin, die bei den paarweisen Vergleichen bei allen Testpersonen nur von einem einzigen Osteopathen, jedoch nicht von den anderen gefunden wurden, und andererseits die negativen Kappa-Werte, die auf eine Übereinstimmung hinweisen, die schlechter als die zufällig zu erwartende Übereinstimmung ist. Allerdings konnte auch bei Reduktion der möglichen Ergebnisse auf drei durch getrennte Berücksichtigung der beiden Körperachsen der Testpersonen keine Verbesserung der Ergebnisse erreicht werden.

Da ein Großteil der Probanden Osteopathiestudenten waren, wussten sie über die Verwendung des Tests Bescheid und waren somit nicht völlig gegenüber dem Ziel des Tests blindiert.

Es wurden angestrebt Probanden, die eine osteopathische Indikation aufweisen, zu testen. Dazu wurde von der Autorin eine Anamnese durchgeführt. Ob die Probanden eine osteopathische Indikation aufweisen ist somit eine subjektive Einschätzung der Autorin.Zu Bedenken ist weiters, dass aus der klinischen Erfahrung der Autorin aus der Stärke der Beschwerden nicht auf die Stärke eines beim Global Listening zu spürenden Zuges geschlossen werden kann. Nicht geachtet wurde darauf, dass die Probanden die Gruppe der Menschen, die einen Osteopathen aufsuchen, repräsentieren.

7.3 Ablauf

Zwar wurde ein Training der Tester zur Normierung der Durchführung eine Woche vor Studienbeginn durchgeführt, den Angaben der Testpersonen nach wurde die Berührung jedoch sehr unterschiedlich empfunden. Ein intensiveres Training hätte diesen Unterschied eventuell reduziert. Auf die Unmöglichkeit der Normierung von Berührung und deren fragliche Sinnhaftigkeit wird in Kapitel 7.4eingegangen.

Stabile Bedingungen in Bezug auf den vom Patienten verursachten Zug können nicht garantiert werden. Allerdings wurde versucht, durch die kurze Testdauer und die möglichst nicht invasive Berührung annähernd gleichbleibende Bedingungen zu gewährleisten. Adaptative Vorgänge im Körper finden freilich in kürzesten Zeiträumen statt. Wie wesentlich stabile Bedingungen während der Testung für das Ergebnis sind, lässt die Studie von Humphreys (2004) erahnen.

Stabile Bedingungen in Hinblick auf die Tagesverfassung bzw. augenblickliche Verfassung der Tester können nicht garantiert werden. Ermüdungserscheinungen können nicht ausgeschlossen werden. Die Anzahl an hintereinander durchgeführten Testungen wurde jedoch den Wünschen der Tester entsprechend gestaltet. Wert wurde auch auf ein ruhiges Umfeld ohne zusätzliche Reize gelegt.Durch die Anzahl an Testern, wirkt sich die Verfassung eines einzelnen Testers nicht so stark auf das Ergebnis aus. Weiters sollte ein in der klinischen Praxis brauchbarer Test auch dann korrekte Ergebnisse liefern, wenn sich der Tester nicht am Höhepunkt seiner Leistungsfähigkeit befindet.

Eine Stärke dieser Studie besteht darin, dass zur Bestimmung der Intrarater-Reliabilität drei Durchgänge durchgeführt wurden, was über den in Studien häufig durchgeführten zwei Durchgängen liegt (siehe Kapitel4.5) Eine größere Anzahl an Durchgängen und damit eine höhere Anzahl an Tests macht jedoch auch Veränderungen an den Probanden und eine Ermüdung der Tester wahrscheinlicher.

7.4 Standardisierung des Tests

Die Durchführung des Tests wurde nach Reflexion der Literatur klar festgelegt und mit den Testern geübt. Trotzdem zeigen die Rückmeldungen aller Beteiligten, dass es nicht gelang, eine einheitliche Durchführung des Global Listenings zu erreichen.

Dabei ist auch zu bedenken, dass jeder Tester mit seinen körperlichen Gegebenheiten an den körperlichen Gegebenheiten des Probanden testet. Testet z.B. ein großer Tester einen kleinen Probanden, bedeckt seine Hand die ganze Brustwirbelsäule. Testet ein Tester mit kleinen Händen einen großen Probanden, bedeckt seine Hand nur einen Teil der Brustwirbelsäule. Auch die Haltung der beiden Tester wird sich anpassen. Nach Beobachtung der Autorin lernt jeder Student im Laufe seiner Ausbildung mit seinen körperlichen Gegebenheiten umzugehen und Griffe und Haltung so anzupassen, dass er sich am besten auf das Gespürte konzentrieren kann. Eine zu strenge Normierung kann das Ergebnis somit negativ beeinflussen.

Wie schwer bis unmöglich eine einheitliche (geschweige denn idente) Durchführung des Tests ist, zeigt sich aber auch durch Rückmeldungen,in denen Begriffe vorkommen, die sich jeder Standardisierung entziehen („in Verbindung treten“, „die eigene Mitte finden“).Diese Begriffe wurdenunter anderem auch von Probanden verwendet, die Patienten der Autorin und keine Osteopathiestudenten waren. Das Problem ist die Normierung der zwischen zwei Individuen durchgeführten Berührung. Es ist nicht das Ziel dieser Arbeit, das Thema Berührung in seiner ganzen Dimension zu besprechen. Es gänzlich auszulassen ginge aber an der Problematik von Reliabilitätsstudien vorbei. So werden hier einige Fragen, die Peter Sommerfeld in seinem Artikel „Berührung – Wahrnehmung des Fernen im Nahen?“ stellt, zitiert:

> *„Geht Berührung allein im Nehmen, im Vernehmen von etwas auf? Ist sie nicht gleichzeitig auch immer ein Wirken und Bewirken? Wohin laufen diese Momente? Wer ist „in“ der Berührung die/der Berührende und wer ist die/der Berührte? Gibt es überhaupt eine Ausrichtung, der die Berührung gehorcht? Was*

ist am Ende dieses Weges, dem die Berührung – in welche Richtung auch immer – folgt? An welcher Grenze läuft Berührung entlang, welche Grenze überwindet, durchstößt, durchbricht oder infiltriert bzw. unterwandert sie?"(Sommerfeld 2006, S. 26)

Problematisch wurde von den Testern die Vorgabe von sieben Interpretationsmöglichkeiten empfunden. Wie bereits in Kapitel 7.2. dargestellt, war für eine statistische Auswertung die Anzahl der Interpretationsmöglichkeiten hoch und eine Ausweitung aus diesem Grund nicht möglich.

Fraglich ist, ob jeder Proband nur eine Läsion, die einen Zug ausübt, aufweist. Es besteht die Möglichkeit, dass verschiedene Läsionen im Körper der Probanden Züge ausgeübt haben, und verschiedene Tester verschiedene Läsionen gespürt haben.

7.5 Externe Validität

Unter externer Validität versteht man nach Bortz und Döring (2006) die *„Generalisierbarkeit der Ergebnisse einer Untersuchung auf andere Personen, Objekte, Situationen und/oder Zeitpunkte" (Bortz und Döring, 2006, S. 33)*.

Wie die Autoren schreiben, besteht bei einer Studie

„die Gefahr, dass der zu erforschende Realitätsausschnitt durch zu strenge Definitions-,Operationalisierungs- oder Messvorschriften nur verkürzt, unvollständig bzw. verzerrt erfasst wird, sodass die Gültigkeit der so gewonnenen Erkenntnisse anzuzweifeln ist". (Bortz und Döring, 2006, S.32)

Einige Punkte, die in den vorangegangenen Kapiteln besprochen wurden, betreffen die externen Validität dieser Studie:

1) Die Probanden repräsentieren nicht die Gruppe der Menschen, die einen Osteopathen aufsuchen. Zu dieser wurde keine Information eingeholt.
2) Die Normierung der Durchführung, das Verbinden der Augen während des Tests und das Durchführen mehrerer Tests hintereinander, sowie die Reduktion auf sieben Interpretationsmöglichkeiten spiegeln nicht die Anwendung des Tests im klinischen Alltag wieder. Nach Paoletti (2001) ist eine Kategorisierung der komplexen Information, die man durch ein Global Listening erhält, nicht möglich.

3) Weiters wird nach den von Krall (2010) geführten Interviews das Global Listening immer nur als Teil eines Befundungsprozesses gesehen, der ein Gesamtbild entstehen lässt. Passt das Global Listening zu den anderen Ergebnissen hat es Einfluss auf das Gesamtbild. Ansonsten wird sein Ergebnis verworfen.

8 Zusammenfassung

Das Global Listening ist ein Faszientest, bei dem versucht wird, mit den Händen auf Züge im Körper zu „hören“ und so einen ersten Eindruck vom Körper zu gewinnen. Verursacht werden fasziale Züge durch Dysfunktionen. Sie erzeugen einen Fixpunkt im Fasziensystem, zu dem der Zug leitet.Die Durchführung und Interpretation des Tests werden in der Literatur zum Teil unterschiedlich beschrieben. Für diese Studie wurden von der Autorin nach Reflexion der Literatur eine standardisierter Ablauf und sieben Kategorien zur Einordnung des Gespürten erarbeitet.

Anhand bisheriger Reliabilitätsstudien und anhand von Metaanalysen wurden Qualitätskriterien für diese Studienform bestimmt und Probleme von Reliabilitätsstudien aufgezeigt.

Sechs Tester testeten 18 symptomatische Probanden, wobei die ersten zehn Probanden dreimal getestet wurden. Ein Tester verwendet das Global Listening in seiner Praxis nicht, die anderen wenden es regelmäßig an. Eine Woche vor Testablauf fand eine Einschulung zur Normierung der Durchführung statt. Die Tester waren gegenüber den Symptomen der Probanden und den Ergebnissen der anderen Tester blindiert und hatten während der Testung die Augen verbunden. Die Probanden wussten nicht über die den Testern zur Verfügung stehenden Interpretationsmöglichkeiten Bescheid.

Es konnte weder eine Intrarater- noch eine Interrater-Reliabilität nachgewiesen werden. Die Ergebnisse bewegen sich im Bereich zufälliger Übereinstimmung. Der Tester, der den Test in seiner Praxis nicht anwendet, erreichte das beste Ergebnis. Dieses liegt jedoch nur im Bereich schwacher Reliabilität und die Streuung der von diesem Tester erreichten Werte ist groß.

Fraglich ist, ob stabile Bedingungen sowohl im Hinblick auf die faszialen Züge im Patienten als auch im Hinblick auf die Leistung der Tester erreicht werden konnten. Eine Standardisierung der Durchführung des Tests wurde angestrebt, konnte allerdings laut den Rückmeldungen von Probanden und Hilfspersonen nicht erreicht werden. Die Sinnhaftigkeit einer strengen Standardisierung ist zweifelhaft. Auch die externe Validität – d. h. ob es gelang, einen Realitätsausschnitt zu erzeugen, der ein generalisierbares Ergebnis bringt– ist anzuzweifeln. Es sollte Ziel zukünftiger For-

schung sein, die klinische Realität beziehungsweise einen repräsentativen Ausschnitt davonzu beurteilen. Die Verwendung des Global Listenings in der Praxis sollte mit Vorsicht erfolgen. Eine Anwendung in Kombination mit anderen Tests erscheint empfehlenswert.

Literaturverzeichnis

Barral J-P. 2002. Lehrbuch der Viszeralen Osteopathie. Band 2. München Jena. Urban&Fischer, 5-6.

Brismée J-M, Gipson D, Ivie D,Lopez A, Moore M, Matthijs O, Phelps V, Sawyer S, Sizer P. 2006.Interrater Reliability of a Passive Physiological Intervertebral Motion Test in the Mid-Thoracic Spine. Journal of Manipulative and Physiological Therapeutics 29, 369-373.

Becker R E. Learning to Listen. In: Brooks R E. (Hrsg.) 1997. Life in Motion. The Osteopathic Vision of Rollin E. Becker. Portland. Oregon. Rudra Press. 148-149.

Bortz J, Döring N. 2006. Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4. Auflage. Heidelberg.Springer Medizin Verlag. 32-33, 196, 200.

Croibier A. 2006. Diagnostik in der Osteopathie. München Jena. Urban&Fischer. 60-62, 212-215, 280-281.

Degenhardt B F, Snider K T, Snider E J, Johnson J C. 2005.Interobserver Reliability of Osteopathic Palpatory Diagnostic Tests of the Lumbar Spine: Improvement from Consensus Training. Journal of the American Osteopathic Association 105, 465-473

Fjellner A, Bexander C, Faleij R, Strender L-E. 1999.Interexaminer reliability in physical examination of the cervical spine. Journal of Manipulative and Physiological Therapeutics 22, 511-516.

Fröhlich W D. 1997. Wörterbuch Psychologie. 21., überarbeitete Auflage. München. Deutscher Taschenbuchverlag.350.

Fryer G, Mc Pherson H C, O´Keefe P. 2005.The effect of training on the inter-examiner and intra-examiner reliability of the seated flexion test and assessment of pelvic anatomical landmarks with palpation. International Journal of Osteopathic Medicine 8, 131-138.

Gemmel H. und Miller P. 2005.Interexaminer reliability of multidimensional examination regimes used for detecting spinal manipulable lesions: A systematic review. Clinical Chiropractic 8, 199-204.

Gibbons P, Dumper C, Gosling C. 2002. Interexaminer and intra-examiner agreement of assessing simulated leg length inequitality using palpation and observation during a standing assessment. Journal of Osteopathic Medicine 5, 53-58.

Gonella C, Paris S V, Kutner M. 1982. Reliability in Evaluating Passive Intervertebral Motion. Physical Therapy 62. 436-444.

Halma K D, Degenhardt B F, Snider K T, Johnson J C, Schaun Flaim M, Bradshaw D. 2008.Intraobserver Reliability of Cranial Strain Patterns as Evaluated by Osteopathic Physicians: A Pilot Study. Journal of the American Osteopathic Association 108, 493-502.

Hartman S Eund Norton J M. 2002.Interexaminer Reliability and Cranial Osteopathy.The scientific Review of Alternative Medicine 6, 23-34.

Harvey D und Byfield D. 1991. Preliminary studies with a mechanical model for the evaluation of spinal motion palpation. Clinical Biomechanics 6, 79-82

Hawk C, Phongphua C, Bleecker J, Swank L, Lopez D, Dubley T. 1999. Preliminary study of the reliability of assessment procedures for indications for chiropractic adjustment of the lumbar spine. Journal of Manipulative and Physiological Therapeutics 22, 382-389.

Hestboek L undLeboeuf-Yde C. 2000.Are Chiropractic Tests for Lumbo-Pelvic Spine Reliable and Valid? A Systematic Critical Literature Review. Journal of Manipulative and Physiological Therapeutics 23, 258-275.

Hinkelthein E und Zalpour C. 2005. Diagnose und Therapiekonzepte in der Osteopathie. Berlin. Springer, 13.

Humphreys B K, Delahaye M, Peterson C K. 2004. An investigation into the validity of cervical spine motion palpation using subjects with congenital block vertebrae as a “gold standard”. BMC Musculosceletal Disorders 5, s.p..

Kmita A und Lucas NP. 2008.Reliability of physical examination to assess asymmetry of anatomical landmarks indicative of pelvic somatic dysfunction in subjects with and without low back pain. International Journal of Osteopathic Medicine 11, 16-25.

Krause R. 2007. Validität und Reliabilität. Deutsche Zeitschrift für Osteopathie 3, 29.

LandisJ R and Koch G G. 1977.The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

Paoletti S. 2001. Faszien. Anatomie Strukturen Techniken Spezielle Osteopathie. München Jena. Urban&Fischer. 194-198.

PodlesnicW. 2006. Local Listening – a General Diagnostic Tool?An Experimental Examination of its Reliability. Masterthese, Universität Krems

Potter L,Mc Carthy C, Oldham J. 2006. Intraexaminer Reliability of Identifying aDysfunctional Segment in the Thoracic and Lumbar Spine. Journal of Manipulative and Physiological Therapeutics 29, 203-207.

Puylaert M. 2005. Osteopathische Diagnostik. In: Liem T,Dobler T K, Puylaert M.(Hrsg.): Leitfaden viszerale Osteopathie. München Jena. Urban&Fischer, 54-56.

Sachs L. 2004. Angewandte Statistik. 11. überarbeitete und ergänzte Auflage. Springer Berlin, Heidelberg, New York,90, 473.

Sommerfeld P. 2006. Berührung – Wahrnehmung des Fernen im Nahen. Osteopathische Medizin 4, 25-32.

Tong H C, Heyman O G, Lado D A, Isser M M. 2006. Interexaminer Reliability of Three Methods of Combining Test Results to Determine Side of Sacral Restriction, Sacral Base Position, and Innominate Bone Position. Journal of the American Osteopathic Association 106, 464-468.

Van Trijffel E, Anderegg Q, Bossuyt P M M, Lucas C.2005.Inter-examiner reliability of passive assessment of intervertebral motion in cervical and lumbar spine: A systematic review. Manual Therapy 10,256-269.

9 Anhang

9.1 Interviewleitfaden für das Interview mit Beatrix Krall

Einstiegsfrage: Für Ihre Masterthese haben Sie leitfadengestützte Experteninterviews geführt um Informationen zum Global Listening oder Globalen Ecoute Test zu erhalten. Können Sie die Themenschwerpunkte ihrer Arbeit kurz beschreiben?

Weitere Punkte:

- Wie viele Interviews?
- Wie weit ist die Ausarbeitung?
- Bedeutung des Tests in der Praxis (Regelmäßig eingesetzt?Zu welchem Zweck? Stellenwert beim Finden der osteopathischen Diagnose)
- Durchführung des Tests:
  Patientenposition
  Position des Testers
  Handkontakt (Eine Hand oder zwei Hände? Wo werden die Händeaufgelegt?)
  Mit wieviel Druck wird der Kontakt aufgenommen?
  Interpretation, was wird zu spüren versucht
  Dauer

9.2 Interview mit Beatrix Krall

A: Für ihre Masterthese haben Sie leitfadengestützte Experteninterviews geführt, um Informationen zum Global Listening oder Globalen Ecoute Test zu bekommen. Können Sie kurz die Themenschwerpunkte Ihrer Arbeit beschreiben?

B: Die Themenschwerpunkte in meiner Masterthese… Ich kann Ihnen nur sagen, was die Themenschwerpunkte in meinem Leitfaden sind, also was ich eigentlich im Interview herausfinden wollte: Was versteht man unter dem Begriff Global Listening, weil es da ja sehr viel in der Literatur gibt: von Ecoute über Listening über Global Ecoute über General Listening. Dann wollte ich wissen, ob diese Begriffe auch verwendet werden, ob die Personen, Probanden, die Interviewkollegen, dass dieser Begriff auch das trifft, was die testen wollen damit. Also da waren die meisten damit einverstanden. Dann gab's indirekt Fragen zum Stellenwert, insofern, dass man fragt, wird es angewendet, regelmäßig/ja/nein. Dann war ein Schwerpunkt in meiner Arbeit, herauszufinden, welche Zielsetzung wird damit verfolgt, warum macht man das Global Listening überhaupt, wird's zur Feststellung von Dysfunktionen verwendet oder zur Evaluierung verwendet und wenn es nicht regelmäßig verwendet wird, warum nicht regelmäßig. Dann war ein riesengroßer Bereich darin zum Thema Durchführung, sowohl was die Ausgangsstellung der Patienten betrifft, weil man da sehr viel Verschiedenes in der Literatur findet, von Liegen über Sitzen über Stehen. Dann war die Frage: Wie steht der Therapeut, weil es unterschiedliche Handhaltung einfach gibt, gibt's irgendwelche die sich durchgesetzt haben, hat der Therapeut eigene entwickelt und sind es noch immer die, die er in der Ausbildung kennen gelernt hat. Das ist ein Schwerpunkt. Und Abschlusspunkt ist dann zum Thema einfach gewesen: wie war die Ausbildung. Hätte man sich Veränderungen der Ausbildung gewünscht zum Thema Global Listening oder war es gut so wie es unterrichtet worden ist. War der Zeitpunkt gut, war der Ablauf gut, ist es ausreichend oft geübt worden, ist es zufällig gekommen oder war das ein eigener Punkt, dass man gesagt hat „Ok, heute geht es ums Thema Global Listening", als Teil der Befundaufnahme oder ist das vorausgesetzt worden, dass man es dann einbaut als Befundaufnahme. Das sind so die Schwerpunkte drin. Im Endeffekt geht es darum nur herauszufinden, wie wird das Global Listening in der Praxis angewendet.

A: Wie viele Interviews haben Sie geführt?

B: Ich habe zwölf Interviews geführt.

A: Und bei der Ausarbeitung, wie weit sind Sie da bisher gekommen?

B: Bis jetzt ist es so, dass alle Interviews transkribiert sind. Die Interviews insofern ausgearbeitet sind, dass alles in eigene Kategorien zugeordnet worden ist. Kategorien sind eigentlich die Hauptüberschriften der Hauptkapiteln in meinem Leitfaden. Und inzwischen haben sich neue Kategorien gebildet, wenn man sagt, die Antworten passen eigentlich nicht zu dem und dem Punkt, was ich gefragt habe, sondern es gehört eigentlich ein neuer Punkt hin oder es gehört unter einen neuen Überbegriff. Die sind jetzt alle zugeordnet und jetzt sind Teile schon so ausgearbeitet, man schaut innerhalb der Zuordnung in der Kategorie und Unterkategorien, gibt es jetzt etwas, dass großer gemeinsamer Nenner ist. Zum Beispiel gibt es immer wieder, dass so Sachen kommen, ich spür hinein, wo zieht's mich hin. Ich schau wie groß und wo das Spannungszentrum ist, dass man sagen kann, aha, da gibt es irgendeinen gemeinsamen Nenner, der beschäftigt sich mit Spannung. Und da darunter fallen sehr viele Antworten und was gibt es in der Literatur dazu. Das ist jetzt momentan der Stand, einzelne Kapitel sind schon weiter ausgereift, die man interpretiert hat und dann nur alles zugeordnet, in Kategorien zugeordnet, aber noch nicht weiter geschaut, ob es gemeinsame große Überbegriffe gibt.

A: Und wird das Global Listening regelmäßig angewendet von den Therapeuten?

B: Also die zwölf, die ich interviewt habe, würde ich sagen 90 % wenden es beim ersten Mal regelmäßig an – so ganz grob. Ob es am Ende der Behandlung auch gemacht wird, ist sehr unterschiedlich. Hängt davon ab, ob es ein auffälliger Test war – am Anfang überhaupt. Wenn der auffällig war bei der Befundaufnahme am Anfang, dann wird er am Ende auch nicht gemacht. Manche verwenden es zu Evaluierung der Behandlung im Verlauf von drei oder vier Therapien. Manche machen es wirklich nur beim 1. Mal, dann ist das ein Teilaspekt, passt der hinein in die Befundaufnahme, in die restlichen Ergebnisse, passt es im Gesamtbild, ist es ok, passt es nicht ins Gesamtbild, dann bleibt er dort stehen, wo er ist und das war's auch.

A. Das heißt der Stellenwert beim Finden der osteopathischen Diagnose?

B: Ist schon sehr wichtig bis hin hat keine Bedeutung.

A: Und Zweck ist also einerseits das Finden der Läsion, wieder Evaluierung, kann man das…

B: Zweck ist so herauszufinden, wo gibt es eigentlich hauptsächlich Spannungszentren im Körper und passt es irgendwie mit der Anamnese und den anderen Befunden zusammen. Insofern hilft es zur Formulierung der osteopathischen Diagnose, wobei man das nur ableiten kann, weil direkt wird es nicht so beantwortet.

A: Aber immer im Gesamtbild?

B: Aber immer im Gesamtbild! Das heißt man muss zu allen anderen Ergebnissen aus der Anamnese und aus sonstigen Befundungstechniken, die angewendet werden, muss dazu passen. Dann findet es Einfluss. Wenn nicht, dann findet es keinen Einfluss.

A: Und jetzt zur Durchführung des Global Listenings. Vielleicht können Sie die Varianten vorstellen, die es da gegeben hat. Das erste ist einmal die Position des Patienten. Gibt es überhaupt Instruktionen dazu? Wenn ja, wie schauen die aus?

B: Wenig, wenn dann ist es Ausgangsstellung Stand, da ist es schon so, dass manche sagen, man soll hüftbreit stehen, andere sagen, es soll nur eine Hand zwischen den Füßen sein, andere sagen, es ist ganz egal. Also ich kann da noch keine Tendenz jetzt – ich habe es noch nicht zusammengezählt – herausfinden. Häufig, dass der Patient die Augen schließt, aber nicht alle. Und er sollte mal nach vorne schauen.

A: Und immer im Stand und nicht im Sitz?

B: Sitz ganz selten, nur vereinzelt. Liegen wieder deutlich mehr. Rückenlage durchaus auch Ausgangsstellung. Sitz wie gesagt sehr selten. Wenn, 90 % im Stehen.

A: Die Position des Osteopathen, gibt es da allgemeine Angaben?

B: Hängt von der Handhaltung ab und die variiert zwischen einer Hand und mit zwei Händen. Variiert von einer Hand am Kopf und wenn es zwei Hände sind, dann ist es eine am Kopf, eine in der BWS, eine am Kopf, eine am Sakrum, oder eine am Kopf und die zweite nur zur Sicherheit vom Patienten, die fühlt er aber nicht. Von gerade dahinter stehen über seitlich dahinter stehen oder seitlich stehen oder beide Hände auf den Schultern.

A: Haupttendenz ist eine Hand am Kopf?

B: Eine Hand am Kopf haben alle, also fast alle. Manche haben noch eine zweite Hand irgendwo.

A: Zweite Hand ist selten, oder…

B: Sakrum ist nicht so selten, eher selten nur eine Hand am Kopf, würde ich jetzt sogar aus dem Gedächtnis sagen. Oder 50/50, so irgendwie.

A: Und wie sich der Therapeut dann wohl fühlt, stellt er sich dann auf?

B: Je nachdem wie nah. Mit einer Hand - dann stellen wir uns meistens dahinter, bei zwei Händen ist dahinter stehen nicht so günstig, aber wenn man Spannungen hat, dann ist es eher seitlich und ob die linke oder rechte Hand am Kopf ist, hängt von der Vorliebe des Therapeuten ab.

A: Und der Druck, der aufgenommen wird, dieser Ecoute, ist das mit Druck, ist das ganz leicht aufgelegt?

B: Nein, also das ist nur leicht aufgelegte Berührung.

A: Kontakt ist vorhanden, oder?

B: Kontakt ist zu 90 % vorhanden.

A: Und wie lang dauert dieser Test?

B: Im Schnitt kann man sagen die meisten waren zwischen 0 und 5 Sekunden, manche 10 Sekunden bis zu 20 Sekunden im Schnitt. Einige Sekunden, sagen wir es mal so.

A: Zur Interpretation des Tests: Gibt es da einheitliche Tendenzen oder ist es sehr, sehr unterschiedlich?

B: Es gibt schon ein paar Grobtendenzen, das sind so die Tendenzen, die in der Literatur beschrieben sind mit nach vorne ist eher viszerales Problem, nach hinten eher parietales Problem. So ganz grob. Es gibt aber dann Differenzierungen, dass einzelne Organe damit gefunden werden, manche Kollegen führen dann noch genaue Differenzierungen durch, z.B. Zug nach rechts vorne unten, dann ist es eher die Leber oder nach rechts hinten unten, dann könnte es die Niere auf dieser Seite sein,

oder nach links vorne unten – dann eher der Magen, oder ganz gerade nach vorne unten könnte auf ein Problem der Blase hinweisen. Allerdings sind diese genauen Differenzierungen nicht in der Literatur angeführt und die Mehrheit der interviewten Kollegen hat eher global interpretiert, das heißt Zug nach vorne eher viszerales Problem und Zug nach hinten eher parietales Problem.

9.3 Beschwerden der Probanden:

Proband 1:1) rechter Ellbogen: nach deutlicher Belastung Schmerz (VAS 5cm), kann bis zu einem Tag dauern,

2)linke Hüfte: Gefühl von Festigkeit, nach Sport, dauert ein paar Stunden bis zu einem Tag

3) Schultern: endgradige Bewegungseinschränkung bei Elevation, nicht schmerzhaft

Proband 2: Cervikalsyndrom, VAS: 4cm innerhalb der letzten 24 Stunden

Proband 3: Schmerz paravertebral rechts in der BWS, täglich, im Lauf eines anstrengendes Arbeitstages ansteigend, VAS: 3-4cm

Proband 4:1) starke Spannungszustände vom Becken bis zur BWS und nach ventral, konstant vorhanden

2)ziehende, nicht schmerzhafte Spannungszustände im Nierenbereich, die sich in der Nacht aufbauen und morgens am stärksten sind, täglich

Proband 5: Schmerzen in der linken Hüfte, VAS 6,5cm innerhalb der letzten Stunden

Proband 6: 1) rechtes Handgelenk und rechtes Daumengrundgelenk: Schmerzen bei Belastung, dann einige Tage dauernd

2)Verdauungsbeschwerden: Wechsel Verstopfung/Durchfall, v.a. nach unregelmäßigem Essen

Proband 7: 1) rechte Fußsohle zwischen 2 und 3 Zehengrundgelenk: Schmerz beim Abrollen – VAS 4cm

2) Sprunggelenk links: Schmerz auch in Ruhe, Dauer: einige Stunden, täglich, VAS: 3 cm

Proband 8: schmerzhafte Verspannung Schulter-Nackenbereich links, z.T. ausstrahlend über den Thorax links bis zum Magen, kurz auftretend bis permanent mäßig vorhanden, VAS 3 cm

Proband 9: 1) Obstipation, nach dem Essen geblähter, harter Bauch
2) Schmerzen in der Lendenwirbelsäule nach längerem Stehen, geht nach dem Hinsetzen weg, VAS 5 cm

Proband 10:Schmerzen linkes ISG und LWS, meist nachts und morgens, Dauer ca. 30 min, tritt ca. zwei mal im Monat auf, VAS 5 cm

Proband 11: Status post Thalamusblutung mit Hemiparese und Parästhesien rechts

Proband 12: Schmerzen im rechten ISG und in der LWS, Dauerschmerz – VAS 2cm, bei bestimmten Bewegungen kurzzeitig Schmerz VAS 5cm (innerhalb der letzten 24 Stunden)

Proband 13: Schmerzen in der rechten Schulter, in unregelmäßigen Abständen nachts – VAS 6cm, ansonsten bei bestimmten Alltagsbewegungen, VAS 4cm

Proband 14: 1)Cervikalsyndrom mit deutlicher Bewegungseinschränkung der HWS, bei Rotation nach rechts Schmerzpunkt, VAS 2,5cm
2)Schmerz im Bereich der linken Hüfte – wetterabhängig?, dauert, wenn er da ist, ca. 2 Tage, VAS 4 cm

Proband 15: Supinationstrauma links 01/2010, beim Laufen und bei maximaler Dorsalextension stechende Schmerzen, VAS 3cm (innerhalb der letzten 24 Stunden)

Proband 16: 1) Verdauungsprobleme je nach Nahrung – Blähungen, auffälliger Stuhl
2) Schmerz beim Kauen, nur anfangs, 2- bis 3-mal pro Woche, VAS 7cm

Proband 17: Schmerzen im Hüft/Beckenbereich seit einem Sturz, täglich spürbar, Stärke variiert je nachdem, wieviel sich die Probandin bewegt, zwischen VAS 2-5 cm

Proband 18: rechte Hand, wie „eingeschlafen“, z.T. Schwächegefühl, manchmal Schmerzen, in der Nacht, bei längerem Arbeiten über Kopf, längerem Beibehalten einer Position, Besserung durch Bewegung, Schmerzen z.T. auch in Unter-und Oberarm weitergeleitet,
VAS innerhalb der letzten 24 Stunden: 5cm

9.4 Visuelle Analogskala

Visuelle Analogskala – Schmerz

Name Patient:

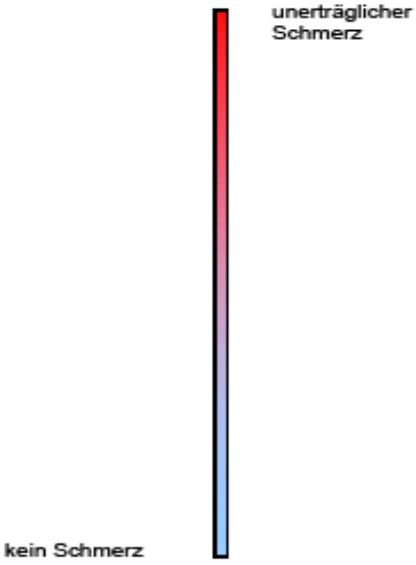
Datum:

Bitte machen Sie auf der senkrechten Linie eine Markierung, die Ihren stärksten Schmerz

- innerhalb der letzten Stunde
- innerhalb der vergangenen 24 Stunden
- innerhalb der letzten 7 Tage

angibt.

(nur eine Möglichkeit auswählen – bitte vereinbaren Sie mit Ihrer Physiotherapeutin/Ihrem Therapeuten, welche der drei Möglichkeiten für Sie am aussagefähigsten ist; die anderen beiden Möglichkeiten streichen Sie bitte durch).



Schmerzangabe des Patienten in cm
(Abstand vom unteren Ende der VAS):

Therapeutin/Therapeut:

9.5 Tabelle für die Ergebnisse

Therapeut:

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1			
Proband 2			
Proband 3			
Proband 4			
Proband 5			
Proband 6			
Proband 7			
Proband 8			
Proband 9			
Proband 10			
Proband 11		----	----
Proband 12		----	----
Proband 13		----	----
Proband 14		----	----
Proband 15		----	----
Proband 16		----	----
Proband 17		----	----
Proband 18		----	----
Proband 19		----	----
Proband 20		----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

9.6 Ergebnisse

Tester 1

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1	HL	VL	0
Proband 2	H	HL	HR
Proband 3	0	HR	0
Proband 4	V	VR	VR
Proband 5	HL	0	HL
Proband 6	HL	HR	H
Proband 7	HR	H	HR
Proband 8	HL	H	HL
Proband 9	0	VR	0
Proband 10	VL	H	0
Proband 11	0	----	----
Proband 12	V	----	----
Proband 13	VL	----	----
Proband 14	0	----	----
Proband 15	VR	----	----
Proband 16	HL	----	----
Proband 17	0	----	----
Proband 18	0	----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

Tester 2

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1	H	V	V
Proband 2	0	0	H
Proband 3	VL	H	H
Proband 4	H	0	HR
Proband 5	HR	HL	HL
Proband 6	VR	H	H
Proband 7	V	R	V
Proband 8	0	0	H
Proband 9	VR	HL	VL
Proband 10	H	0	H
Proband 11	V	----	----
Proband 12	V	----	----
Proband 13	0	----	----
Proband 14	V	----	----
Proband 15	HR	----	----
Proband 16	V	----	----
Proband 17	H	----	----

Proband 18	VL	----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

Tester 3

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1	V	H	V
Proband 2	H	H	0
Proband 3	0	0	0
Proband 4	VL	0	H
Proband 5	H	HL	VR
Proband 6	L	0	VL
Proband 7	V	H	H
Proband 8	HL	HR	0
Proband 9	VL	HL	VR
Proband 10	0	0	HL
Proband 11	HR	----	----
Proband 12	H	----	----
Proband 13	0	----	----
Proband 14	VR	----	----
Proband 15	0	----	----
Proband 16	H	----	----
Proband 17	VL	----	----
Proband 18	H	----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

Tester 4

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1	H	H	VR
Proband 2	HL	0	0
Proband 3	VL	0	H
Proband 4	0	HL	HR
Proband 5	0	HL	HL
Proband 6	HL	HR	HR
Proband 7	H	H	VL
Proband 8	0	H	0
Proband 9	H	V	0
Proband 10	VR	H	H
Proband 11	VR	----	----
Proband 12	V	----	----
Proband 13	0	----	----
Proband 14	H	----	----
Proband 15	VL	----	----
Proband 16	VR	----	----
Proband 17	H	----	----
Proband 18	0	----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

Tester 5

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1	VR	V	0
Proband 2	VR	0	H
Proband 3	0	H	VR
Proband 4	V	0	V
Proband 5	HR	VL	LH
Proband 6	VR	HR	VR
Proband 7	V	VR	H
Proband 8	0	V	0
Proband 9	VR	0	HR
Proband 10	H	H	VR
Proband 11	VR	----	----
Proband 12	H	----	----
Proband 13	HR	----	----
Proband 14	H	----	----
Proband 15	HR	----	----
Proband 16	HR	----	----
Proband 17	V	----	----
Proband 18	0	----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

Tester 6

	1 Durchgang	2 Durchgang	3 Durchgang
Proband 1	VR	H	H
Proband 2	0	0	0
Proband 3	HR	VR	HR
Proband 4	V	0	V
Proband 5	H	HL	H
Proband 6	HR	H	0
Proband 7	0	HR	V
Proband 8	H	0	HR
Proband 9	V	V	H
Proband 10	HL	V	HR
Proband 11	VL	----	----
Proband 12	H	----	----
Proband 13	HR	----	----
Proband 14	H	----	----
Proband 15	H	----	----
Proband 16	0	----	----
Proband 17	HR	----	----
Proband 18	0	----	----

vorne: V, vorne links: VL, vorne rechts: VR, hinten: H, hinten links: HL, hinten rechts: HR, anders: 0

Intraobserver and Interobserver Reliability of the Global Listening

Margit Rittler

Vienna School of Osteopathy, Danube University Krems – Center for Traditional Chinese Medicine and Complementary Medicine

Abstract

Objective: To investigate the intra- and interobserver reliability of the Global Listening and to observe whether the experience of the tester has an influence on his intraobserver reliability.

Methods: 18 subjects were tested by six testers. The first ten of the subjects were tested two more times. Subjects had to have symptoms, which were judged as an indication for osteopathy. The testers´ experience was assessed on the basis of the frequency of the use of the Global Listening in their practice. One week before data collection a training was held. A standardised procedure of the test was elaborated and practised. During data collection testers had their eyes blindfolded and were assisted by additional persons. Testers were blinded to each others´ results and to the patients´ symptoms. Patients were blinded to the possible outcomes of the test. Cohen´s Kappa was used for calculation purposes.

Results: Neither an intraobserver reliability nor an interobserver reliability could be found. Results were in a range of random agreement. The only tester, who never uses the test in his practice, achieved the best intraobserver reliability.

Conclusion: As no reliability of the Global Listening could be found, the test should be used in practice with caution. However, it is doubtful, whether stable conditions

during the test were reached and whether external validity of the study is given. These problems are shared by most reliability studies.

Introduction

The Global Listening is an osteopathic test, which is proposed in literature to get a first impression of the patient´s restrictions. By feeling the traction of the fascia the tester gets a hint, whether the main problem is a viszeral or structural one, whether it is on the left or right side, more cranial or caudal (Barral, 2002; Croibier, 2006; Hinkelthein und Zalpour, 2005; Puylaert, 2005).

The question, whether the outcome of a test can be trusted,is often posed. An attempt to answer it, is to determine the test´s reliability. Reliability can be defined as the extent to which a repeated test will produce the same result when evaluating an unchanged characteristic. Is the test repeated by one person on the same subject, the intraobserver reliability can be determined. If two or more examiners test the same subject, the interobserver reliability can be established (Krause, 2007).

Up to date reliability studies have hardly found reliable tests so far. Hestboek and Leboeuf-Yde (2000) concluded in their review on the reliability of chiropractic tests for the lumbo-pelvic spine, that only tests focusing on palpation for pain had consistently acceptable reliability values. The authors criticized a lack of quality in many studies. Van Trijffel et al. (2005) reviewed studies on the interobserver reliability of passive assessment of intervertebral motion in the cervical or lumbar spine. The authors defined criteria for assessing the methodological quality of studies. Amongst others they demanded an adequate number of subjects and testers, and the blinding of testers to patients´ symptoms and each other´s results. None of the studies included in the review could find a reliable test. The authors criticized the quality of the studies. In both reviews the authors criticize the testing of asymptomatic subjects, which on the one hand does not reflect the daily practice, on the hand other leads to problems in statistical analysis because of the lack of varying outcomes. Another problem described by both teams is to guarantee stable conditions concerning the subjects´ characteristics. Especially, when repeating a test several times or when using a combination of tests, changes in the system might be induced. Humphreys et al. (2004) managed to achieve stable conditions by testing subjects with a congenital

block vertebrae. Testers achieved substantial agreement for identifying the segment of greatest hypomobility in the cervical spine.

Gemmel and Miller (2005) reviewed studies dealing with multidimensional examination procedures, which aimed at detecting spinal manipulable lesions. The authors advocate the opinion, that the evaluation of a combination of tests is closer to clinical practice than evaluating only one test. Again the quality of the reviewed studies was criticized. An additional criteria for the methodological quality was implemented in this review: the testing of naïve subjects, who are unable to influence the result. From the few papers done on this topic none could find a reliable combination of tests. The argument that a combination of tests better represents the clinical practice is striking. Yet a combination of tests is more likely to have an influence on the tested subject, which means, that it is even more difficult to care for stable conditions during data collection.

Hartman and Norton (2002) summarized studies on the reliability of tests used in craniosacral osteopathy. They found no sudy showing a craniosakral test to be reliable. The authors strongly criticized the concept of craniosakral osteopathy, which they claim to have "*little science in anyaspect within it*" (Hartman and Norton, 2002, S 32). Halma et al. (2008) observed the intraobserver reliability of diagnostic tests for the cranial rhythmic impuls rate, cranial strain patterns and quadrants of restriction. Two experienced examiners tested 24 symptomatic subjects. The scientists took a great effort in blinding the testers to the subjects. As intraobserver reliability is supposed to be higher than interobserver reliability, the authors raised the standard level for acceptable reliability in this study from $\kappa > 0{,}40$ (as proposed for interobserver reliability) to $\kappa > 0{,}60$. With a value of $\kappa = 0{,}67$ the testing of cranial strain patterns showed the highest reliability and could exceed the level for acceptable reliability. The testing of the cranial rhythmic impuls and the quadrants of restrictions achieved values in the fair to moderate range (as proposed in the interpretation scale of Landis and Koch 1977 for interobserver reliability, view table 2). It is remarkable that a test out of this concept achieved that good intraobserver reliability.

Podlesnic (2006) investigated the intra- and interobserver reliability of the "abdominal local listening". No other study on the reliability of a "Listening" could be found by the author. Neither an intra- nor an interobserver reliability was found in this study.

An interesting detail in reliability studies is the influence of the testers´ experience. One can assume that experienced testers achieve better reliability values than juniors, but so far the opposite has been shown. Harvey and Byfield (1991) and Jensen et al. (1993) compared the results of students testing a mechanical model to the results of experts testing the same model. In both studies the students did better than the experts. Kmita and Lucas (2008) compared in their study on the physical examination to assess asymmetry of anatomical landmarks indicative of pelvic somatic dysfunction the results of experienced osteopaths to those of students. They could not find any significant difference between both groups. Similar result were achieved by Podlesnic (2006) for a group of more and a group of less experienced osteopaths. The testers´ experience was defined by the testers´ final year of osteopathic education at the “Wiener Schule für Osteopathie”.

Worth consideration is furthermore the effect of a training of the testers. As testers often develop their personal variation of a test, a training helps to standardise the procedure. Fryer et al. (2005) investigated the effect of training on the intraobserver and interobserver reliability of the seated flexion test and assessment of pelvic anatomical landmarks with palpation. Testers had to perform a training of one hour for two times. Intraobserver reliability raised significantly, interobserver reliability did not. Degenhardt et al. (2005) investigated the effect of a consensus training for osteopathic palpatory diagnostic tests of the lumbar spine on the interobserver reliability. They split their study into three phases: pretraining interobserver reliability assessment, consensus training and posttraining interobserver reliability assessment. The achieved values indicate an increase of reliability in phase three. As methods were changed remarkably from phase one to phase three (for example one vertebrae was tested 48 to 72 times by each tester in phase one and maximum 18 times in phase three) an improvement from consensus training cannot be concluded. Van Trijffel et al. (2005) could not find any difference between the results of trained and untrained testers.

In this study the intra- and interobserver reliability of the Global Listening is investigated. In addition, the influence of the testers experience with the test is evaluated.

Methods

Subjects:

Students of the “Wiener Schule für Osteopathie” were contacted during their courses and asked to give their name and e-mail address if interested to take part in the study. The Global Listening is a test used for all patients coming to an osteopath, no matter which symptoms they present. Therefore the only inclusion criteria for the subjects was, that they present an osteopathic indication. 51 interested students were contacted two weeks before the study took place. At the time of data collection 13 Students were available and had symptoms. To augment the number of subjects five patients of the author also took part in the study. A medical history of each subject was taken by the author to assure that subjects probably present an osteopathic indication.

Some days before data collection subjects were informed about the procedure. They were asked to stay as “neutral as possible” during the test in order not to influence the outcome. Furthermore they were asked not to use any perfume, hair-care product, hairslide or wear a top with a hood. Subjects knew that the test serves to find tractions in the body. They were not informed about the possible interpretations.

Testers

The original aim was to compare the findings of experienced osteopaths with osteopaths who had just passed their final exam (September 2009). Unfortunately no experienced osteopath was available. For this reason the testers´ experience was judged by the frequency of the use of the Global Listening in their practice. One of the six osteopaths taking part in the study never uses the test in his practice, the others regularly use it (view table 1).

tester	**frequency of use**
T1	each patient before and after treatment
T2	never
T3	every second patient
T4	every third patient
T5	every second patient and patients examined for the first time
T6	every second patient

Table 1: frequency of the use of the Global Listening

Training

One week before data collection a training in order to standardise the proceduretook place. Testers performed the test on each other and discussed the outcome. It was practised with kitchen scales to lay the hand with a weight of 20 – 30 grams on top of the head. The importance of the attitude during the test was emphasised and testers were asked to be in a “dialogue full of respect” during the test. Furthermore, testers were informed about the mode of data collection. They were told, that they will have to test 38 subjects (so they were blinded about testing eight subjects three times).

Test procedure

After reviewing the literature a test procedure was fixed.

Subjects had to stand feet under their hips, looking straight forward. When instructed by the tester, they had to close their eyes.

Testers stood behind the subjects and placed one hand on the apex with a weight of 20-30 grams. Whether a second hand is used and - if yes - where it is placed varies in literature. Therefore the use of a second hand was discussed with the testers and a possibility most convenient for all was fixed. This was to lay the hand between the shoulder blades.

Handcontact had to last no longer than four seconds.

Originally five possible outcomes were proposed. As testers found it difficult to sort the outcome of the Global Listening in five categories seven possible outcomes were fixed: front, left front, right front, back, left back, right back, other.

Data collection:

On the day of data collection testers and subjects were guided into two separate rooms when entering the building. By doing so it was guaranteed, that they had no information about each other. Three assistants were present. The course of data collection was talked through with each group and participants were asked to pose questions if something was unclear to them.

Testers were arranged in a semicircle and their eyes were blindfolded. One assistant was responsible for two testers. Subjects went from one tester to the next. Partly with

the help of the assistants the testers placed their hands on the subject and asked the subject to close the eyes. Four seconds later the assistants asked the testers to write down one of the possible outcomes, which then was written in a table by the assistants. After the testing of 18 subjects a break of ten minutes was held. Then the first ten subjects were tested two more times by each tester.

Participants were asked to give feedback.

Statistical analysis

Cohen´s Kappa was used to quantify the intra- and interobserver reliability. Kappa is the proportion of observed agreement above chance divided by the maximum possible agreement above chance. The guidelines proposed by Landis and Koch were used to interpret the kappa values (see table 2).

Value of kappa	**Agreement**
$\kappa<0,20$	Slight
$0,20<\kappa<0,40$	Fair
$0,40<\kappa<0,60$	Moderate
$0,60<\kappa<0,80$	Substantial
$0,80<\kappa<1,00$	Almost perfect

Table 2: interpretation of kappa coefficient (after Landis and Koch, 1977)

According to Fjellner et al. (1999) values higher than at least 0.4 are an indicator for acceptable interobserver reliability. According to Halma et al. (2008) values higher than at least 0.6 are an indicator for acceptable intraobserver reliability.

In order to investigate whether the Global Listening is more reliable in one axis of the body (left/right, anterior/posterior) results were also evaluated regarding this aspect. For example the answer “front” was assigned to “other”, the answer “left front” to “left” regarding a left/right axis. A reduction of the possible answers down to three could also be achieved by this step.

Results

Intraobserver reliability

The mean intraobserver reliability calculated out of all comparisons from the three rounds (view table 3) indicates slight agreement (κ = 0.06). The calculation shows a high statistical spread.

Tester	1/2	1/3	2/3	**M(3)**	**SS(3)**
T1	-0.14	0.38	-0.06	0.06	0.28
T2	0.04	0.01	0.30	0.12	0.16
T3	0.25	0.02	0.00	0.09	0.14
T4	-0.03	-0.11	0.27	0.04	0.20
T5	-0.08	0.11	-0.19	-0.05	0.15
T6	0.02	0.25	0.01	0.10	0.13
M(6)	0.01	0.11	0.06	**M(18)**	0.06
SS(6)	0.13	0.18	0.19	**SS(18)**	0.17

Table 3: Kappa-values for the comparisons of two rounds for each single tester, mean (M) and statistical spread (SS) for the three values of each tester (M(3), SS(3)), M and SS out of the values achieved by the six testers in the comparison of two rounds (M(6), SS(6)), M and SS for all comparisons (M(18), SS(18))

Evaluating the outcomes regarding a left/right axis the mean is κ = 0.10, regarding an anterior/posterior axis the mean is κ = 0.05. Conclusively, reducing the possible outcomes to three, again, only slight agreement is achieved.

Remarkably, the only osteopath, who never uses the test in his clinical practice, achieved the best mean (κ = 0.12). Still, his value is within the range of slight agreement.

Interobserver reliability

The achieved values indicate no reliability of the Global Listening (view table 4). The mean for the first round is κ = 0.02, for the second κ = -0.01 and for the third κ = 0.08.

Round 1 (n=18)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0.08	0.01	0.00	0.06	-0.03
	T2		-0.07	0.21	0.27	-0,13
	T3			-0.17	0.03	-0.08
	T4				0.09	-0.07
	T5					0,20
Round 2 (n=10)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0.15	-0.08	0.27	0.07	-0.16
	T2		0.14	-0.04	0.24	0.35
	T3			0.20	-0.11	0.11
	T4				0.16	0.26
	T5					0.01
Round 3 (n=10)	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	0.12	-0.10	-0.01	0.04	-0.20
	T2		-0.05	0.30	0.07	-0.15
	T3			0.04	0.02	-0.05
	T4				0.05	-0.10
	T5					-0.06
Mean round 1-3	Ex1/Ex2	T2	T3	T4	T5	T6
	T1	-0.04	-0.06	0.09	0.06	-0.13
	T2		0.01	0.16	0.19	0.02
	T3			0.02	-0.02	-0.01
	T4				0.10	0.03
	T5					0.05

Table 4: Cohen´s Kappa for all pairs of testers from all three rounds, means of the three values out of the three rounds (grey: $\kappa \geq 0.2$)

Analysing the results regarding a left/right axis and an anterior/posterior axis, again, the results are in the range of random agreement. Therefore, a reduction to three

possible interpretations has not brought any amelioration of the result. Regarding a left/right axis the mean for the first round is $\kappa =$ -0,02, for the second κ = 0,18 and for the third κ = -0,02. Regarding an anterior/posterior axis the mean for the first round is $\kappa = 0{,}00$, for the second κ = 0,02 and for the third κ = 0,00. No difference when evaluating the results regarding a left/right axis to evaluating the results regarding an anterior/posterior axis could be found.

Discussion

The results of this study demonstrate that examiners were unreliable in their interpretation of the Global Listening. When evaluating the influence of the testers` experience, the examiner having no experience with the test achieved the best result. A similar outcome has been published in former studies (Harvey and Byfield, 1991; Jensen et al., 1993). However, the difference between the achieved results of both groups was slight in this study. One has to be aware of the fact, that when comparing one tester with a group of other testers, the abilities of the one tester have a great influence on the outcome. Therefore statements about the effect of the tester´s experience should be made with caution on the basis of this study.

The number of subjects was low with regard to the possible outcomes (seven). A reduction to three - by analysing the data regarding a left/right and alternatively an anterior/posterior axis - did not change the results significantly. Therefore, one can assume that the low agreement is not caused by too many possible outcomes.

A great problem of reliability studies is their external validity, i.e. the question whether one can generalize results to other persons, objects, situations and/or times (Bortz and Döring, 2006). This concerns different aspects of this study. The subjects included in the study had symptoms as demanded by Hestboeuk and Leboeuf-Yde (2000), Van Trijffel et al. (2005) and Gemmel and Miller (2005), but they did not represent the group of people coming to see an osteopath. A standardised procedure of the test was elaborated and practised. In their daily practise clinicians develop a way of performing of a test, that suits their personal conditions best. When using a different procedure testers may not be as sensitive. Feedback of the subjects indicated that the testers´ touch felt different albeit standardisation and training. There-

fore, the standardisation of touch may be doubted. Another aspect is, that for statistical analyses possible outcomes of the test have to be provided. Although guidelines for the interpretation of the test have been given in literature – these were used in this study -, it is always emphasised that the Global Listening gives a complex impression of the patient (Barral, 2002; Croibier, 2006; Hinkelthein und Zalpour, 2005; Paoletti, 2001; Puylaert, 2005). This cannot be evaluated by statistics. Gemmel and Miller (2005) argued that the evaluation of a combination of tests is better representing the clinical practice. In an interview on the use of the Global Listening with Krall (2010), who writes a masterthesis on this topic, she emphasises, that the Global Listening is used as a part of a diagnostic process, which in the end conveys a picture of the patient. It is part of an integrae whole and should not stand alone.

One of the problems in reliability studies is the difficulty of guaranteeing stable conditions for all testers. An effort to stabilize conditions was made by asking the testers not to be invasive and by fixing a duration of the test of four seconds (i.e. a short time). A calm surrounding and a break should avoid tiring of the testers. Nevertheless the body has its one dynamic and cannot be forced not to change. Though, it can be assumed that a strong dysfunction in the body is not changed by such a slight contact as applied in this study. Unclear is, whether each subject had one lesion with a distinct traction into one direction. Perhaps different lesions have been felt by different testers.

Conclusion

As neither an intraexaminer nor an interexaminer reliability of the Global Listening could be found, the test should be used with caution. A proposition is, to use the information of the test in cases, when it is clear and fits into a picture of the patient rising during the diagnostic process. It will be a challenge for future studies to analyse the clinical practice and not an artificial situation.

Acknowledgement

I would like to thank Dr. Gebhard Woisetschläger for his help with the statistical analysis of the study. Thanks also to the subjects, testers and assistants for taking part in the study.

References

Barral J.-P. 2002. Lehrbuch der Viszeralen Osteopathie. Band 2. München Jena. Urban&Fischer, 5-6.

Bortz J, Döring N. 2006. Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4. Auflage. Heidelberg. Springer Medizin Verlag. 32-33, 196, 200.

Croibier A. 2006. Diagnostik in der Osteopathie. München Jena. Urban&Fischer. 60-62, 212-215, 280-281.

Degenhardt B F, Snider K T, Snider E J, Johnson J C. 2005. Interobserver Reliability of Osteopathic Palpatory Diagnostic Tests of the Lumbar Spine: Improvement from Consensus Training. Journal of the American Osteopathic Association 105, 465-473

Fjellner A, Bexander C, Faleij R, Strender L-E. 1999. Interexaminer reliability in physical examination of the cervical spine. Journal of Manipulative and Physiological Therapeutics 22, 511-516.

Fryer G, Mc Pherson H C, O´Keefe P. 2005. The effect of training on the inter-examiner and intra-examiner reliability of the seated flexion test and assessment of pelvic anatomical landmarks with palpation. International Journal of Osteopathic Medicine 8, 131-138.

Gemmel H. und Miller P. 2005. Interexaminer reliability of multidimensional examination regimes used for detecting spinal manipulable lesions: A systematic review. Clinical Chiropractic 8, 199-204.

Halma K D, Degenhardt B F, Snider K T, Johnson J C, Schaun Flaim M, Bradshaw D. 2008. Intraobserver Reliability of Cranial Strain Patterns as Evaluated by Osteopathic Physicians: A Pilot Study. Journal of the American Osteopathic Association 108, 493-502.

Hartmann S E und Norton J M. 2002. Interexaminer Reliability and Cranial Osteopathy. The scientific Review of Alternative Medicine 6, 23-34.

Harvey D und Byfield D. 1991. Preliminary studies with a mechanical model for the evaluation of spinal motion palpation. Clinical Biomechanics 6, 79-82

Hestboek L und Leboeuf-Yde C. 2000. Are Chiropractic Tests for Lumbo-Pelvic Spine Reliable and Valid? A Systematic Critical Literature Review. Journal of Manipulative and Physiological Therapeutics 23, 258-275.

Hinkelthein E und Zalpour C. 2005. Diagnose und Therapiekonzepte in der Osteopathie. Berlin. Springer. 13.

Humphreys B K, Delahaye M, Peterson C K. 2004. An investigation into the validity of cervical spine motion palpation using subjects with congenital block vertebrae as a “gold standard”. BMC Musculosceletal Disorders 5, s.p..

Kmita A und Lucas N P. 2008. Reliability of physical examination to assess asymmetry of anatomical landmarks indicative of pelvic somatic dysfunction in subjects with and without low back pain. International Journal of Osteopathic Medicine 11, 16-25.

Krause R. 2007. Validität und Reliabilität. Deutsche Zeitschrift für Osteopathie 3. 29.

Landis J R and Koch G G. 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159-174.

Paoletti S. 2001. Faszien. Anatomie Strukturen Techniken Spezielle Osteopathie. München Jena. Urban&Fischer. 194-198.

Podlesnic W. 2006. Local Listening – a General Diagnostic Tool? An Experimental Examination of its Reliability. Masterthese, Universität Krems

Puylaert M. 2005. Osteopathische Diagnostik. In: Liem T, Dobler T K, Puylaert M. (Hrsg.): Leitfaden viszerale Osteopathie. München Jena. Urban&Fischer, 54-56.

Van Trijffel E, Anderegg Q, Bossuyt P M M, Lucas C. 2005. Inter-examiner reliability of passive assessment of intervertebral motion in cervical and lumbar spine: A systematic review. Manual Therapy 10, 256-269.

Appendices

Tester 1

	1stRound	2ndRound	3rdRound
Subject 1	bl	fl	0
Subject 2	b	bl	br
Subject 3	0	br	0
Subject 4	f	fr	fr
Subject 5	bl	0	bl
Subject 6	bl	br	b
Subject 7	br	b	br
Subject 8	bl	b	bl
Subject 9	0	fr	0
Subject 10	fl	b	0
Subject 11	0	----	----
Subject 12	f	----	----
Subject 13	fl	----	----
Subject 14	0	----	----
Subject 15	fr	----	----
Subject 16	bl	----	----
Subject 17	0	----	----
Subject 18	0	----	----

front: f, front left: fl, front right: fr, back: b, back left: bl, back right: br, other: 0

Tester 2

	1stRound	2ndRound	3rdRound
Subject 1	b	f	f
Subject 2	0	0	b
Subject 3	fl	b	b
Subject 4	b	0	br
Subject 5	br	bl	bl
Subject 6	fr	b	b
Subject 7	f	R	f
Subject 8	0	0	b
Subject 9	fr	bl	fl
Subject 10	b	0	b
Subject 11	f	----	----
Subject 12	f	----	----
Subject 13	0	----	----
Subject 14	f	----	----
Subject 15	br	----	----
Subject 16	f	----	----
Subject 17	b	----	----
Subject 18	fl	----	----

front: f, front left: fl, front right: fr, back: b, back left: bl, back right: br, other: 0

Tester 3

	1stRound	2ndRound	3rdRound
Subject 1	f	b	f
Subject 2	b	b	0
Subject 3	0	0	0
Subject 4	fl	0	b
Subject 5	b	bl	fr
Subject 6	L	0	fl
Subject 7	f	b	b
Subject 8	bl	br	0
Subject 9	fl	bl	fr
Subject 10	0	0	bl
Subject 11	br	----	----
Subject 12	b	----	----
Subject 13	0	----	----
Subject 14	fr	----	----
Subject 15	0	----	----
Subject 16	b	----	----
Subject 17	fl	----	----
Subject 18	b	----	----

front: f, front left: fl, front right: fr, back: b, back left: bl, back right: br, other: 0

Tester 4

	1stRound	2ndRound	3rdRound
Subject 1	b	b	fr
Subject 2	bl	0	0
Subject 3	fl	0	b
Subject 4	0	bl	br
Subject 5	0	bl	bl
Subject 6	bl	br	br
Subject 7	b	b	fl
Subject 8	0	b	0
Subject 9	b	f	0
Subject 10	fr	b	b
Subject 11	fr	----	----
Subject 12	f	----	----
Subject 13	0	----	----
Subject 14	b	----	----
Subject 15	fl	----	----
Subject 16	fr	----	----
Subject 17	b	----	----
Subject 18	0	----	----

front: f, front left: fl, front right: fr, back: b, back left: bl, back right: br, other: 0

Tester 5

	1stRound	2ndRound	3rdRound
Subject 1	fr	f	0
Subject 2	fr	0	H
Subject 3	0	H	fr
Subject 4	f	0	f
Subject 5	br	fl	bl
Subject 6	fr	br	fr
Subject 7	f	fr	H
Subject 8	0	f	0
Subject 9	fr	0	br
Subject 10	H	H	fr
Subject 11	fr	----	----
Subject 12	H	----	----
Subject 13	br	----	----
Subject 14	H	----	----
Subject 15	br	----	----
Subject 16	br	----	----
Subject 17	f	----	----
Subject 18	0	----	----

front: f, front left: fl, front right: fr, back: b, back left: bl, back right: br, other: 0

Tester 6

	1stRound	2ndRound	3rdRound
Subject 1	fr	H	H
Subject 2	0	0	0
Subject 3	br	fr	br
Subject 4	f	0	f
Subject 5	b	bl	b
Subject 6	br	b	0
Subject 7	0	br	f
Subject 8	b	0	br
Subject 9	f	f	b
Subject 10	bl	f	br
Subject 11	fl	----	----
Subject 12	b	----	----
Subject 13	br	----	----
Subject 14	b	----	----
Subject 15	b	----	----
Subject 16	0	----	----
Subject 17	br	----	----
Subject 18	0	----	----

front: f, front left: fl, front right: fr, back: b, back left: bl, back right: br, other: 0