Psychometrische Beurteilung verhaltensgestützter Schmerzassessments für Menschen mit Demenz

Potenziale von Item-Response-Theorie und Latent Variable Modellen am Beispiel der Verhaltensinventare CNPI und BESD

Inauguraldissertation
zur Erlangung des Grades eines Dr. phil. an der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Roman Kaspar

Heidelberg, im Juni 2009
Disputation am 1. Oktober 2009

Gutachter

Prof. Dr. Andreas Kruse
Institut für Gerontologie an der Universität Heidelberg

Prof. Dr. Joachim Funke
Psychologisches Institut der Universität Heidelberg

Danksagung
Mein besonderer Dank gilt Herrn Prof. Andreas Kruse für seine offene und kritisch-wertschätzende Begleitung bei der Erstellung dieser Arbeit. Herrn Prof. Joachim Funke danke ich für seine Bereitschaft zur Zweitbegutachtung der vorliegenden Arbeit. Weiterhin möchte ich den Kollegen vom Arbeitskreis Schmerz im Alter der Deutschen Gesellschaft zum Studium des Schmerzes für die fruchtbare Projektkooperation und ihre hilfreichen Anmerkungen zu dieser weiterführenden Arbeit danken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die vorliegende Arbeit untersucht die Möglichkeiten, auf der Grundlage beobachteten Verhaltens zu einer Einschätzung der Schmerzbelastung von Pflegeheimbewohnern zu gelangen, die sich aufgrund ihrer demenzbedingten Kommunikationsbeeinträchtigungen nicht mehr eindeutig verbal zu ihrem Schmerzerleben äußern können. Damit soll eine gegenwärtig bestehende Lücke bei der Sicherung einer optimalen Versorgung dieser besonders vulnerablen Klientel geschlossen werden, die in einem aktuellen Praxishandbuch zum Schmerzmanagement für Pflegeberufe wie folgt beschrieben ist:

> Das vorliegende Wissen lässt vermuten, dass das Schmerz-Assessment Pflegender beschränkt und oft auch ungenau ist [. . .]. Pflegende neigen noch immer dazu, sich an ihr eigenes Urteil zu halten, und verlassen sich lieber auf körperliche Zeichen und Verhaltensweisen, die irreführend und ungenau sein können. Formelle Schmerz-Assessment-Instrumente erleichtern eine effiziente Kommunikation und das Assessment, indem sie die Möglichkeit von Fehlern oder Verzerrungen (Bias) verringern. (Carr & Mann, 2002 , S. 52)

Diese Arbeit prüft, inwiefern strukturierte Verfahren des beobachtungsgestützten Schmerzassessments diesem Anspruch tatsächlich gerecht werden.

Das hohe Lebensalter scheint mit einem beträchtlichen Ausmaß erlebter Schmerzen verbunden zu sein (AGS, 2002). Dies scheint im besonderen Maße auch für ältere Menschen zuzutreffen, die aufgrund körperlicher oder kognitiver Beeinträchtigungen die eine selbstständige Lebensführung unmöglich machen, in Einrichtungen der stationären Altenpflege versorgt werden (Marzinski, 1991; Proctor & Hirdes, 2001). Im Vergleich zu kognitiv gesunden altersgleichen Personen klagen an einer Demenz erkrankte Menschen signifikant seltener über Schmerzen (Cook et al., 1999; Feldt, 2000; Huffman et al., 2000; Mäntyselkä et al., 2004; Marzinski, 1991; Proctor & Hirdes, 2001; Shega et al., 2004). Die Prävalenz selbstberichteter Schmerzen scheint dabei mit stärkerer kognitiver Beeinträchtigung abzunehmen (Parmelee et al., 1997; Shega et al., 2004). Eine Reihe von klinischen Studien wies darüber hinaus darauf hin, dass demenzkranke Menschen im Vergleich zu kognitiv gesunden Personen gleichen Alters deutlich weniger Schmerzmedikation erhalten, obwohl sie vergleichbar häufig mit schmerzrelevanten Erkrankungen belastet sind (Closs, Barr & Briggs, 2004; Horgas & Tsai, 1998; Kaasalainen et al., 1998; Mäntyselkä et al., 2004; Scherder, 2000; Scherder & Bouma, 1997; Wolf-Klein et al., 1988). Die Befundlage zu demenzbedingten Veränderungen im Schmerzerleben selbst sind gegenwärtig noch uneinheitlich. Allerdings kann angenommen werden, dass die sensorisch-diskriminativen Anteile der Schmerzempfindung auch in fortgeschrittenen Stadien der Demenz noch weitgehend erhalten sind (Kunz, 2006; Scherder, Sergeant & Swaab, 2003). Die nicht selten bis in das späte Krankheitsstadium hinein erhaltene Fähigkeit demenzkranker Menschen, Auskunft über ihr Schmerzerleben zu geben wird gegenwärtig sicherlich noch nicht hinreichend systematisch genutzt. Die überwiegende Mehrheit von Menschen in späten Stadien der Demenzerkrankung kann jedoch keine verbale Auskunft mehr zu erlebten Schmer-

zen geben, weswegen für diese Population die schmerzbezogene Verhaltensbeobachtung mitunter die einzige Möglichkeit der Schmerzerfassung darstellt. Das Schmerzmanagement bei Menschen mit Demenz stellt damit aber insbesondere die stationäre Versorgung vor große Herausforderungen, da hier der Anteil mittelschwer bis schwer beeinträchtigter Bewohner ungefähr zwei Drittel beträgt (Weyerer, Schäufele & Hendlmeier, 2005).

In den letzten Jahrzehnten wurden annähernd zwei Dutzend Instrumente vorgeschlagen, die sich auf die Beobachtung schmerzbezogenen Verhaltens stützen (Hadjistavropoulos et al., 2007; Herr et al., 2006; Schofield et al., 2005; Smith, 2005; Stolee et al., 2005; van Herk et al., 2007; Zwakhalen, Hamers, Abu-Saad & Berger, 2006). Eine angemessene Beurteilung der psychometrischen Güte dieser Verfahren scheitert gegenwärtig zum einen an der unzureichenden empirischen Datenlage, zum anderen aber auch am gewählten testtheoretischen Zugang. Die ausschließliche Orientierung an den sehr restriktiven Konzepten und Methoden der klassischen Testtheorie verstellt den Blick auf die in Frage stehenden potenziell schmerzbezogenen Verhaltensweisen und den Kontext der Messung (z.B. Aktivierung). Die Empfehlungen aktueller Übersichtsarbeiten müssen damit nicht nur vorläufig, sondern auch vergleichsweise undifferenziert bleiben.

Die vorliegende Arbeit hat sich darum zum Ziel gesetzt, diese Lücke durch die vergleichende Analyse von zwei im anglo-amerikanischen Raum weitverbreiteten Verhaltensinventaren, die in einer deutschen Version bei einer substanziellen Stichprobe demenzkranker Menschen eingesetzt wurden, zu schließen. Dabei werden die Potenziale neuerer Verfahren aus den methodischen Traditionen der Item-Response-Theorie und Latent Variable Modellen genutzt, um einen maximalen Auflösungsgrad der psychometrischen Überprüfung zu leisten und verschiedene Grade der Aktiviertheit der Probanden als einen für die schmerzbezogene Verhaltensbeobachtung unverzichtbaren Kontextfaktor berücksichtigen zu können.

Die Erkenntnisse dieser Arbeit sollen dazu beitragen, zukünftig Schmerzen anhand von Verhaltensweisen einschätzen zu können, deren Indikationscharakteristiken mit Blick auf die Kernmerkmale der spezifischen Zielgruppe demenzkranker Menschen und die situative Rahmung der Schmerzbeobachtung detailliert beschrieben sind.

Aufbau der Arbeit

Im einleitenden theoretischen Teil dieser Arbeit sollen zunächst die Kernmerkmale der Demenz als gewissermaßen kleinster gemeinsamer Nenner der insgesamt außerordentlich heterogenen Population älterer demenziell erkrankter Menschen herausgearbeitet werden. Für jedes Instrument, das sich zum Ziel setzt, die Schmerzbelastung dieser hoch vulnerablen Gruppe abzubilden, muss zumindest gefordert werden, dass diese Kernmerkmale angemessen berücksichtigt sind. Für einen Teil der großen Unterschiedlichkeit im klinischen Erscheinungsbild bzw. der Symptomatik der Demenz können verschiedene Grunderkrankungen oder Ätiologien verantwortlich gemacht werden. Wo aber eine systematische spezifische Beeinträchtigung durch eine bestimmte Demenzätiologie beschrieben werden kann, bieten sich auch für die Schmerzmessung Chancen – über eine

globale Beeinträchtigung der Gedächtnis- und Denkleistung hinaus – die Effekte dieser enger umrissenen krankhaften Prozesse auf das Schmerzerleben selbst, nicht zuletzt aber auch deren Folgen für die Möglichkeit zur Schmerzmessung zu adressieren. Die Darstellung der heterogenen Krankheitsbilder, die gegenwärtig unter dem Sammelbegriff der Demenzen gefasst werden, orientiert sich dabei weniger stark an einer detailreichen fallbezogenen klinischen Beschreibung möglicher Symptome, sondern versucht dem Leser über den engen Bezug zu den Forschungskriterien der Internationalen Klassifikation psychischer Störungen (ICD-10; Dilling et al., 1997) eine schnelle Orientierung mit Blick auf die wesentlichsten Ätiologien und Leitsymptome zu ermöglichen.

Dem schließt sich ein für die Zwecke der vorliegenden Arbeit ebenso kursorisch gehaltener Überblick zu grundlegenden neurologischen Strukturen und Mechanismen der Schmerzwahrnehmung und -verarbeitung an. Die am Schmerzgeschehen beteiligten zentralen und peripheren Anteile des nozizeptiven Systems, aber auch (psychologische) Prozesse der Schmerzverarbeitung und -modulation sind bei demenzkranken Menschen nicht allein durch die demenzielle Erkrankung bestimmt, sondern unterliegen darüber hinaus selbstverständlich auch einer Veränderung durch normale Altersprozesse. Um die Spezifika des Schmerzerlebens demenzkranker Menschen deutlicher herausstellen zu können, werden darum zunächst die verfügbaren klinischen und neurologischen Befunde zu einer im Alter veränderten Schmerzbelastung kurz dargestellt, bevor der gegenwärtige Stand der lebhaften Diskussion um eine gesteigerte oder verminderte Schmerzbelastung bei (verschiedenen Formen der) Demenz umrissen wird. Es versteht sich von selbst, dass für einen stärker medizinisch orientierten Schmerzforscher beim hier geleisteten Überblick an einigen Stellen Wünsche nach einer tiefergehenden Darstellung unbeantwortet bleiben müssen.

Im zweiten Teil werden zentrale Herausforderungen beim Schmerzassessment durch die Beobachtung potenziell schmerzbezogener Verhaltensweisen demenzkranker Menschen identifiziert und strukturiert. Dabei wird auch das non-verbale Ausdrucksverhalten als ein Bestandteil einer unterschiedlich bewusst ablaufenden, und gegebenenfalls unterschiedlich stark reziprok charakterisierten Kommunikation von Schmerzen begriffen, und somit stärker als bisher auch an die – über weite Strecken der Demenzerkrankung hinweg erhaltene – Selbstauskunft angeschlossen. Zunächst wird diskutiert, wie der überaus umfangreiche und heterogene Pool von Verhaltensweisen, die mittlerweile als potenzieller Schmerzausdruck vorgeschlagen wurden, sinnvoll geordnet werden kann. Anschliessend wird ein komprimierter Überblick über die wesentlichsten Befunde zu Aspekten der Reliabilität, Praktikabilität und Validität von insgesamt mehr als zwei Dutzend für die Schmerzmessung bei Demenz vorgeschlagenen Verhaltensinventaren aus dem größtenteils anglo-amerikanischen Sprachraum gegeben. Dabei werden sowohl die konzeptuellen als auch die methodischen die Grenzen herausgearbeitet, die eine systematische Weiterentwicklung der Assessmentinstrumente selbst, aber auch deren Überprüfung und Beurteilung gegenwärtig zu behindern scheinen. Dabei wird deutlich, dass insbesondere die Kontextmerkmale der Schmerzbeobachtung zur Zeit keine angemessene Berücksichtigung finden. Der Umstand, dass die vorgeschlagenen Verfahren kaum über ihre Initi-

altestung hinaus Bestand haben und kontinuierlich modifiziert oder untereinander neu kombiniert werden, kann als Ausdruck dieses konzeptionellen Versäumnisses gelten. Eine angemessenere Berücksichtigung der mit Blick auf die beobachteten Situationen oder demenzkranken Menschen tatsächlich vorzufindenden Heterogenität wird wohl auch dadurch erschwert, dass die Methoden zur Entwicklung und Überprüfung von Verfahren zum Schmerzassessment nahezu ausschließlich der Logik der klassischen Testtheorie folgen, die für eine Berücksichtigung verschiedener Facetten einer Messung konzeptionell nur wenige Möglichkeiten bereithält. In der Folge blieben beispielsweise Fragen zur Aussagekraft einzelner Verhaltensweisen als Schmerzindikator, zum Einfluss von Bewegung und Aktivität auf die Schmerzbeobachtung, oder zur Angemessenheit der Instrumente bei Menschen mit verschiedenen Schweregraden oder Ätiologien ihrer demenziellen Beeinträchtigungen weitestgehend unbeantwortet. Die zentralen Anforderungen, die sich für eine zukünftige Schmerzforschung daraus ableiten, bestehen zum einen in der Überwindung der Testorientierung zugunsten einer stärkeren Orientierung an konkreten Verhaltensweisen, und zum zweiten in einer systematischeren Berücksichtigung des praktischen Anwendungskontextes der vorgeschlagenen Verfahren.

Um jedoch nicht allein den Forschungsbedarf aufzudecken, sondern darüber hinaus auch Möglichkeiten für eine bessere Bearbeitung der aufscheinenden Probleme einer verhaltensgestützten Schmerzmessung aufzuzeigen, wird dem Leser mit dem sich anschliessenden Kapitel 4 eine Übersicht der Annahmen, Voraussetzungen und Potenziale verschiedener klassischer und probabilistischer testtheoretischer Konzeptionen und der auf dieser Grundlage entwickelten statistischen Verfahren angeboten. Dabei werden insbesondere die Merkmale von Latent Variable Modellen und Item-Response-Modellen beschrieben und deren spezifische Eignung zur Entwicklung und Beurteilung von Instrumenten zur Schmerzbeobachtung herausgearbeitet. Besonderes Augenmerk liegt dabei auf den Möglichkeiten, die Funktionsweise einzelner konkreter Verhaltensweisen im Kontext der Schmerzmessung zu bestimmen, mehrere konkurrierende Verfahren hinsichtlich ihres Aussagebereiches und ihrer Präzision miteinander zu kontrastieren sowie den Einfluss situativer Umstände auf die Güte der Schmerzmessung abzuschätzen und die Äquivalenz einer Schmerzmessung über verschiedene Beobachtungssituationen, Personengruppen, oder Zeitpunkte hinweg sicherzustellen. Selbstverständlich kann diese Zusammenschau aktuell verfügbarer Methoden aus zwei testtheoretischen Traditionen, die sich über weite Strecken unabhängig voneinander – und vom Anwendungsfeld der verhaltensgestützten Schmerzmessung weitestgehend unbemerkt – entwickelt haben, keinerlei Anspruch auf Vollständigkeit erheben. Bei der Vorstellung der wesentlichen Konzepte der verschiedenen Strömungen wird darum die notwendige formale Darstellung so wenig technisch wie möglich gehalten und wann immer möglich auch grafisch veranschaulicht. Ich hoffe, damit insgesamt ein Abstraktionsniveau zu erreichen, das insbesondere den Bedürfnissen derjenigen Personen entgegenkommt, die vor einem (pflege-)wissenschaftlichen Hintergrund Verfahren zur Schmerzmessung bei demenzkranken Menschen entwickeln und prüfen.

Im empirischen Teil dieser Arbeit werden die wichtigsten der zuvor identifizierten

Problemlagen der Schmerzerfassung bei Demenz aufgegriffen, und im Rückgriff auf die beschriebenen aktuellen Analyseansätze bearbeitet. Dazu werden Daten aus einem größeren Forschungsprojekt zur Lebensqualität Demenzkranker Menschen herangezogen. Der Hauptaugenmerk liegt dabei auf der Analyse des schmerzbezogenen Verhaltensausdruckes von annähernd zweihundert demenzkranken Pflegeheimbewohnern, die auf der Grundlage zweier international gut rezipierter Verhaltensinventare gewonnen werden konnten. Um für die Überprüfung der psychometrischen Eigenschaften einen maximalen Auflösungsgrad zu erreichen, wurden die in beiden Verfahren beschriebenen schmerzbezogenen Verhaltensweisen dabei als dichotome Verhaltensindikatoren in einer Ruhe- und einer Aktivitätssituation zur Beobachtung vorgegeben.

Der doppelten Zielsetzung der vorliegenden Arbeit entsprechend wird auch bei der Darstellung der empirischen Ergebnisse besonderer Wert darauf gelegt, neben den inhaltlichen Erkenntnissen zur verhaltensgestützten Schmerzbeobachtung bei der untersuchten Stichprobe demenzkranker Menschen auch den jeweiligen Mehrwert der verwendeten Methoden deutlich zu machen. Die Analysestrategie wurde darum so gewählt, dass in aufeinander aufbauenden Bearbeitungsschritten sowohl die Überprüfung messtheoretischer Annahmen durch die Spezifizierung eines Latent-Variable-Messmodells veranschaulicht, als auch die Vorteile einer weiterführenden Überprüfung inhaltlich konzeptioneller Annahmen im Rahmen eines Latent-Variable-Strukturmodells verdeutlicht werden konnten. Neben den damit aufgezeigten prinzipiellen Vorzügen von Strukturgleichungsmodellen kann anhand der erweiterten Generalisierbarkeit der test- und stichprobenunabhängig geschätzten Itemparameter jedoch auch der Mehrwert methodischer Anleihen aus dem Bereich der Item-Response-Theorie konkret nachvollzogen werden.

Abschließend werden die erarbeiteten Erkenntnisse zu den gegenwärtigen konzeptionellen Herausforderung der Entwicklung von Verhaltensinventaren zur Schmerzmessung bei Demenz, den erwarteten und eingelösten Potenzialen aktuellerer methodischer Analysestrategien und zur psychometrischen Güte der hier detailliert überprüften Verhaltensinventare CNPI und BESD hinsichtlich ihres Geltungsanspruches und ihrer Implikationen für die Praxis kritisch diskutiert.

2 Schmerzbelastung bei demenziellen Erkrankungen

Einige der Merkmale, die eine Schmerzerfassung in der Population demenzkranker Menschen vor besondere Herausforderungen stellen lassen sich stärker der Demenzsymptomatik selbst, andere dagegen eher den im höheren Lebensalter häufig kumulierten gesundheitlichen Problemlagen, oder auch nur normalen Alterungsprozessen zuschreiben. Um diese Faktoren soweit als möglich zu entflechten, werden in den folgenden Kapiteln wesentliche Ätiologien und Symptome demenzieller Prozesse und die theoretischen Vorstellungen sowie empirischen Befunde zu Veränderungen im Schmerzerleben bei Hochaltrigen und schließlich bei demenzkranken Menschen getrennt voneinander dargestellt. Aufgrund der methodischen Schwerpunktsetzung der vorliegenden Arbeit können die hierbei

referierten Forschungsergebnisse selbstverständlich keine Vollständigkeit beanspruchen, sondern sollen dem Leser vielmehr einen Eindruck von der klinischen Heterogenität des Demenzsyndroms und der die aktuelle demenzbezogene Schmerzforschung bestimmenden Modellbildung vermitteln.

2.1 Demenzielle Erkrankungen

Unter der Bezeichnung Demenz werden im alltäglichen Sprachgebrauch eine Reihe verschiedener Kompetenzverluste vor allem in den kognitiven Leistungsbereichen Gedächtnis und Denkvermögen, daneben aber auch hinsichtlich der Affektkontrolle, des Antriebs- oder des Sozialverhaltens verstanden. In den nachfolgenden Abschnitten werden die wesentlichen Definitionskriterien für ein demenzielles Syndrom vorgestellt und Möglichkeiten der Abgrenzung von einer leichten kognitiven Beeinträchtigung (Mild Cognitive Impairment MCI) und Verwirrtheitszuständen (Delir) diskutiert. Eine Vielzahl körperlicher und psychischer Veränderungsprozesse kann für die Ausbildung eines Demenzsyndroms verantwortlich gemacht werden. Im Rahmen dieser Arbeit sollen die wichtigsten neurodegenerativen, entzündlichen, ischämischen und psychogenen Ätiologien der Demenzerkrankung kurz charakterisiert werden.

2.1.1 Definition des Demenzsyndroms

Gemäß der Internationalen Klassifikation psychischer Störungen (ICD-10-GM Version; WHO, 2009) ist Demenz ein Syndrom, als Folge einer meist chronischen oder fortschreitenden Krankheit des Gehirns mit Störung vieler höherer kortikaler Funktionen, einschließlich Gedächtnis, Denken, Orientierung, Auffassung, Rechnen, Lernfähigkeit, Sprache und Urteilsvermögen. Das Bewusstsein ist nicht getrübt. Die kognitiven Beeinträchtigungen werden gewöhnlich von Veränderungen der emotionalen Kontrolle, des Sozialverhaltens oder der Motivation begleitet, gelegentlich treten diese auch eher auf. Dieses Syndrom kommt bei Alzheimer-Krankheit, zerebrovaskulären Störungen und bei anderen Zustandsbildern vor, die primär oder sekundär das Gehirn betreffen (WHO, 2009; Onlineressource: http//www.dimdi.de).

Für die Vergabe der Diagnose Demenz (F00-F03) sind die folgenden Kriterien als notwendig beschrieben (vgl. Dilling et al., 1997; Förstl, 2009):

G1 Nachweis aller folgenden Bedingungen:

1. Abnahme des Gedächtnisses, am deutlichsten beim Lernen neuer Information (verbales und non-verbales Material).

2. Abnahme anderer kognitiver Fähigkeiten (z.B. Verminderung der Urteilsfähigkeit, des Denkvermögens und der Informationsverarbeitung).

Sowohl für die Gedächtnisleistungen, wie auch für die anderen kognitiven Fähigkeiten werden jeweils drei Schweregrade beschrieben, die sich neben der Schwere der Symptomatik auch an den verbliebenen Möglichkeiten einer selbständigen Lebensführung orientieren. Die Kategorie der leichten Beeinträchtigung mit erhaltener

Selbständigkeit in den Alltagsvollzügen wird dabei jeweils als Schwellenwert definiert, d.h. für eine Diagnose muss eine deutliche Beeinträchtigung der Alltagsbewältigung in mindestens einem dieser Funktionsbereiche gegeben sein. Der Gesamtschweregrad der Demenz ergibt sich aus der jeweils höheren Beeinträchtigungsstufe beider Leistungsbereiche.

G2 Um G1 eindeutig nachweisen zu können, muss die Wahrnehmung der Umgebung ausreichend lange erhalten geblieben sein (Ausschluss eines akuten Verwirrtheitszustandes bzw. Delirs).

G3 Die Verminderung der Affektkontrolle, des Antriebs oder des Sozialverhaltens manifestiert sich in mindestens einem der Merkmale emotionaler Labilität, Reizbarkeit, Apathie oder Vergröberung des Sozialverhaltens.

G4 Dauer der unter G1 genannten Leistungs- und Funktionsbeeinträchtigungen für mindestens sechs Monate.

2.1.1.1 Abgrenzung zum Delir

Der Ausschluss einer vorübergehenden Leistungsbeeinträchtigung im Sinne eines akuten Verwirrtheitszustandes (Delir) für die Diagnose einer Demenz wird sowohl explizit, als auch indirekt über das 6-monatige Zeitkriterium gefordert.

Im Gegensatz zur Demenz entwickelt sich ein Verwirrtheitszustand (ICD-10: F05) gewöhnlich innerhalb eines vergleichsweise kurzen Zeitraums (wenige Stunden oder Tage). Kernsymptome des Delirs sind Bewusstseinseintrübung, kognitive Beeinträchtigung, ein rascher Beginn und fluktuierender Verlauf, sowie die Verursachung durch einen spezifischen medizinischen Krankheitsfaktor. Die Dauer des Verwirrtheitszustandes kann bis zu mehreren Monaten betragen.

2.1.1.2 Abgrenzung zur leichten kognitiven Beeinträchtigung

Vom Vollbild einer Demenz wurden in den letzten Jahren verstärkt mildere Formen von insbesondere Kurzzeitgedächtnis-, Aufmerksamkeits- und Auffassungsstörungen ohne auffällige Beeinträchtigungen der psychosozialen Kompetenz abgegrenzt.

Im Gegensatz zur leichten kognitiven *Störung* (F06.7), für die laut ICD-10-Klassifikation eine eindeutig bestimmbare organische Ursache und die prinzipielle Reversibilität gefordert wird, sind für die Diagnosestellung einer leichten kognitiven *Beeinträchtigung* (F06.8) spezifische organische Ursachen und psychische Störungen (z.B. Depression) auszuschließen. Eine Reversibilität wird für die leichte kognitive Beeinträchtigung (LKB) nicht gefordert.

Die LKB gewinnt als mögliches Vorläuferstadium einer sich später entwickelnden Demenz besondere diagnostische Bedeutung. Tatsächlich ist das Risiko, nach fünf Jahren eine manifeste Demenz zu entwickeln mit bis zu 50 Prozent sehr hoch (Bickel & Cooper, 1994; Gauthier, 2006).

Andererseits wird die LKB auch als Ausdruck einer nichtprogredienten gutartigen Altersvergesslichkeit (benign senescent forgetfulness, BSF, Kral, 1962) oder altersassoziierter Gedächtnis- bzw. allgemein kognitiver Beeinrächtigungen diskutiert. Insbesondere mit Blick auf die in den bisherigen Verlaufsstudien einheitlich gefundenen hohen Konversionsraten wird dennoch zu einer regelmäßigen Testung alle 6-12 Monate geraten, um Veränderungen in Richtung auf eine manifeste Demenz möglichst frühzeitig zu erkennen (Zaudig, 2009).

2.1.1.3 Epidemiologie

Gegenwärtig leiden zwischen 6 und 8 Prozent der über 65-Jährigen in den westlichen Ländern unter einer mittelgradig oder schwer ausgeprägten Demenz. Die Prävalenz fraglicher oder leichter Demenzstadien wird in dieser Altenpopulation auf nochmals denselben Prozentsatz geschätzt. In Deutschland leiden gegenwärtig über eine Million Menschen an einer Demenz (Bickel, 2001; Förstl, 2009). In Anbetracht der absehbaren demographischen Entwicklungen kann bis zum Jahr 2050 in den westlichen Industrienationen von einem Bevölkerungsanteil über 80-jähriger Personen von nahezu 10 Prozent ausgegangen werden (UN, 2006). Die Prävalenzraten demenzieller Erkrankungen steigen mit dem Lebensalter expotenziell an (Förstl, 2009). So waren in einer kanadischen Studie nach Graham und Kollegen (1997) in der Altersgruppe der 65 bis 74-Jährigen ungefähr 3 Prozent, bei den 75 bis 84-Jährigen ca. 11 Prozent und ungefähr 35 Prozent der über 84-Jährigen an einer Demenz erkrankt. Die Gesamtzahl der bis 2050 an einer manifesten Demenz Erkrankten wird in den Entwicklungsnationen auf 30 Millionen Menschen geschätzt (Wimo et al., 2003). Weltweit muss mit bis zu 114 Millionen Demenzkranken im Jahr 2050 gerechnet werden.

Diejenigen Krankheiten, die am häufigsten als alleinige Ursache eines Demenzsyndroms berichtet werden, sind die Alzheimer-Krankheit (ca. zwei Drittel bzw. gegenwärtig ca. 65 000 Erkrankte in Deutschland; Bickel, 2000) und zerebrovaskuläre Erkrankungen (vaskuläre Demenzen; 10-30%; Haberl & Schreiber, 2009). Wahrscheinlicher als eine unikausale Verursachung sind jedoch Kombinationen verschiedener Grunderkrankungen, vor allem neurodegenerativer und zerebrovaskulärer Prozesse (mixed dementia). Es wird angenommen, dass zwischen 70 und 90 Prozent aller an einer manifesten Demenz Erkrankten (ggfs. zusätzlich zu weiteren Demenzursachen) Alzheimerveränderungen aufweisen (Förstl, Kurz & Hartmann, 2009).

Sowohl der Zeitpunkt der Ersterkrankung als auch der weitere Verlauf und die Lebenserwartung sind wesentlich durch die verschiedenartigen psychischen und/oder somatischen Grunderkrankungen bestimmt. Demenzerkrankungen in mittleren Lebensaltern sind vergleichsweise selten und häufig an familiäre Erbkrankheiten oder Gendefekte (z.B. Pick-Komplex, Chorea Huntington, Trisomie 21) gekoppelt. Das Erkrankungsrisiko für eine (präsenile) AD vor dem 65. Lebensjahr ist vergleichsweise gering. Sowohl die altersbezogene Prävalenz eines Demenzsyndromes im Allgemeinen, als auch der AD im Besonderen steigt mit dem Lebensalter jedoch expotenziell an (Bickel et al., 2006).

Mit wenigen Ausnahmen (z.B. Demenzsyndrom der Depression oder bei Substanzmissbrauch) handelt es sich bei der Demenz um eine progredient fortschreitende Erkrankung, die zur vollkommenen Disintegration geistiger und körperlicher Steuerungsmechanismen und schließlich zum Tode führt. Wiederum in Abhängigkeit von der spezifischen Ätiologie können sowohl kontinuierlich schleichende (z.B. Alzheimer Demenz), als auch diskontinuierlich stufenförmige (vaskuläre Demenzen) oder sehr rapide Verläufe (z.B. Demenz bei Prionerkrankung) beobachtet werden.

Die durchschnittliche Lebenserwartung beträgt bei der Alzheimer Demenz nach der klinischen Diagnosestellung weitere 5-8 Jahre (Kurz & Greschniok, 1994), bei der Creutzfeld-Jakob-Demenz dagegen lediglich 3-12 Monate (Kretzschmar & Förstl, 2009). Je nach Ausmaß und Lokalisation der zerebrovaskulären Schädigung sind die Prognosen für die Überlebenszeit bei vaskulären Demenzen unterschiedlich, liegen durchschnittlich aber unter derjenigen für AD (Haberl & Schreiber, 2009).

2.1.2 Ätiologien und Differentialdiagnostik

Im ICD-10-GM-System werden inhaltlich die drei übergeordneten Krankheitskategorien Demenz bei Alzheimer-Krankheit (F00), Vaskuläre Demenz (F01) und Demenz bei anderenorts klassifizierten Krankheiten (F02) unterschieden, wobei in letztere Kategorie eine ganze Reihe vergleichsweise prominenter Demenztypen mit sehr unterschiedlichen Ätiologien fallen. Aus diesem Grunde werden spezifische Gruppen von Demenzen, die durch Veränderungen der Basalganglien (Parkinson plus, Lewy-Körper-Demenz, Chorea-Huntington), spezifische bakterielle oder virale Infektionen (Prionkrankheit, z.B. Creutzfeldt-Jakob-Krankheit) oder einen Niedergang umgrenzter Hirnareale i.S. atrophischer Veränderungen (Pick-Komplex) bedingt sind, exemplarisch herausgegriffen.

2.1.2.1 Demenz bei Alzheimer-Krankheit

Die Demenz vom Alzheimer-Typ (auch präsenile/senile Demenz, primär degenerative Demenz, Alzheimer Demenz AD) kann nur neuropathologisch durch das Vorliegen von Alzheimer-Plaques, Neurofibrillen und Neuronenverlust diagnostiziert werden. Entsprechende Veränderungen zerebraler Strukturen lassen sich in geringerem Maße auch bei anderen Demenzformen und sogar bei alten Menschen ohne manifestes Demenzsyndrom finden. Es muss davon ausgegangen werden, dass Alzheimer-Veränderungen an einem Großteil demenzieller Erkrankungen verschiedenster Ätiologie zumindest mitbeteiligt sind.

Neurobiologische Korrelate

Eine wichtige Rolle im degenerativen Prozess der Alzheimer-Krankheit spielt das vor allem intraneuronal toxische β-Amyloid, das durch die ungünstige Spaltung des Membran- und Amyloidvorläuferproteins (APP) entsteht. Eine vermehrte Freisetzung des aus 42

Aminosäuren bestehenden Grundbausteins der Alzheimer-Plaques bei bestimmten Gendefekten (Apolipoprotein-E4-Allel ApoE4 auf Chromosom 19, autosomal-dominante Mutationen der Präsenilin-Gene auf den Chromosomen 14 und 1) erscheint gesichert. Neben der Agglomeration zu extrazellulären Plaques lagert sich β-Amyloid auch perivaskulär ab (kongophile Angiopathie) und begünstigt damit Marklagerveränderungen (Leukoaraiose).

Alzheimer-Plaques bestehen zum großen Teil aus extrazellulär aggregiertem β-Amyloid, Apolipoprotein E und Präsenilin. Bereits Jahrzehnte vor der eigentlichen Manifestation einer Demenz können Plaques als vergleichsweise diffuse Ablagerungen nachgewiesen werden. Über den Krankheitsverlauf hinweg nimmt die Dichte und das Volumen der Plaques zu und in den Plaques finden sich vermehrt Ausläufer von degenerativ veränderten Neuronen (dystrophe Neuriten).

Den Grundbaustein der stärker lokalisierten *Neurofibrillen* stellt hyperphosphoryliertes Tau-Protein, ein pathologisch verändertes mikrotubulläres Transporteiweiß, dar. Neurofibrillen (auch paired helical filaments PHF) treten nicht nur bei AD, sondern auch bei weiteren (z.B. zerebrovaskulären) Erkrankungen auf. Das regelhafte Muster der Ausbreitung von Neurofibrillen über den Krankheitsverlauf hinweg kann in sechs Stadien unterteilt werden, wobei messbare klinische Defizite erst in den beiden letzten Ausbreitungsstadien nachweisbar sind, wenn die Ablagerungen über den (trans)entorhinalen Kortex und das limbische System hinaus auch die Assoziationsareale der Großhirnrinde betreffen (Braak & Braak, 1994).

Die beschriebenen strukturellen Hirnveränderungen äußern sich funktionell durch eine De-Afferenzierung bzw. Efferenzierung des limbischen Systems, einer nachhaltigen Schädigung der neokortikalen Feed-forward- und Feed-back-Systeme zwischen niedrigen und höheren Assoziationsarealen und einer cholinergen Denervation des Neokortex durch den Niedergang cholinerger Zellverbände des basalen Vorderhirns (v.a. Nucleus basalis Meynert). Da Acetylcholin wesentlich an der Steuerung der Aufmerksamkeit, einer geordneten neokortikalen Verarbeitung und dem Abspeichern und Abruf von Gedächtnisinhalten beteiligt ist, sind diese kognitiven Funktionen durch die degenerativen Veränderungen des basalen Vorderhirns und das daraus resultierende cholinerge Defizit besonders betroffen.

Die Möglichkeiten, den Diagnoseverdacht AD mit bildgebenden Verfahren (kraniale Computertomographie cCT, Magnetresonanztomographie MRT) zu sichern erscheinen begrenzt, weshalb diese im Wesentlichen differenzialdiagnostisch zum Ausschluss möglicher alternativer Ätiologien der Demenz (z.B. zerebrovaskuläre oder raumfordernde Prozesse) beitragen kann. Als AD-typischer struktureller Befund gilt eine Hippokampusatrophie bzw. eine Temporalhornaufweitung, wie bereits angemerkt sind aber auch ausgedehnte Marklagerveränderungen mit einer AD vereinbar.

Da es sich bei der Diagnose Demenz bei Alzheimer Krankheit (F00) um eine vorläufige Ausschlussdiagnose handelt, fordert das ICD-10 neben dem Vorliegen aller notwendigen Kriterien für die Diagnose eines Demenzsyndroms weiterhin, dass in der Anamnese, bei der körperlichen Untersuchung oder aufgrund spezieller Untersuchungen keine Hinweise auf eine andere Ursache der Demenz (z.B. zerebrovaskuläre Erkrankung,

HIV, Normaldruck-Hydrozephalus, Parkinson oder Huntington), eine Systemerkrankung (z.B. Hypothyreose, Vitamin-B12- oder Folsäuremangel, Hyperkalzämie) oder auf einen Alkohol- oder Substanzmissbrauch zu finden sind.

Für die Diagnosestellung können nach ICD-10 Alzheimer-Demenzen mit frühem Beginn (d.h. vor dem 65. Lebensjahr; präsenile AD) und solche mit Beginn ab dem 65. Lebensjahr (senile AD) unterschieden werden. Inhaltlich werden mit dem Verlauf des kognitiven Abbaus und der Gewichtung von Gedächtnis- zu sonstigen kognitiven Fähigkeitseinbußen zwei weitere Symptomaspekte genannt, anhand derer die beiden AD-Typen differenziert werden können (s. nachfolgende Übersicht).

F00.1 Demenz bei Alzheimer-Krankheit mit frühem Beginn

1. Die Kriterien für die Demenz bei Alzheimer-Krankheit (F00) müssen erfüllt sein und der Krankheitsbeginn liegt vor dem 65. Lebensjahr.
2. Außerdem muss mindestens eine der folgenden Bedingungen erfüllt sein:
   (a) Nachweis eines relativ plötzlichen Beginns und einer raschen Progredienz
   (b) Zusätzlich zur Gedächtnisstörung eine amnestische oder sensorische Aphasie, Agraphie, Alexie, Akalkulie oder Apraxie (als Hinweis auf das Vorliegen einer temporalen, parietalen und/oder frontalen Beteiligung).

F00.2 Demenz bei Alzheimer-Krankheit mit spätem Beginn

1. Die Kriterien für die Demenz bei Alzheimer-Krankheit (F00) müssen erfüllt sein und der Krankheitsbeginn liegt bei 65 Jahren oder darüber.
2. Außerdem muss mindestens eine der folgenden Bedingungen erfüllt sein:
   (a) Nachweis eines sehr langsamen Beginns und einer allmählichen Progredienz (die Geschwindigkeit der letzteren wird nur retrospektiv nach einem Verlauf von drei oder mehr Jahren deutlich)
   (b) Vorherrschen der Gedächtnisstörung (G1.1) gegenüber der intellektuellen Beeinträchtigung (G1.2) (siehe allgemeine Kriterien für Demenz).

Es muss angenommen werden, dass die beschriebene Altersgrenze von 65 Jahren für die meisten Betroffenen eine passende Zuordnungshilfe darstellt, jedoch auch in jüngeren Jahren AD vom senilen Typus und auch im Senium entsprechend eine AD mit präseniler Symptomatik auftreten kann. Um diese Passung oder Nichtpassung zu adressieren, besteht zudem die Möglichkeit zur Kodierung einer gemischten oder atypischen AD (F00.2).

Klinische Einschätzung

In der *klinischen Einschätzung des Schweregrades* der chronisch progredienten AD werden gewöhnlich ein leichtes, mittelschweres und schweres Demenzstadium unterschieden. Mitunter kann bei entsprechend detaillierter neurokognitiver Testung bereits Jahre vor der Manifestation eines Demenzsyndroms mit den zu fordernden deutlichen Beeinträchtigungen in Alltagsgestaltung und selbständiger Lebensführung ein *subklinisches Vorstadium* festgestellt werden. Durch die Nutzung von Gedächtnishilfen oder anderer Strategien

oder auch nur die Vermeidung kognitiv anspruchsvoller Aufgaben bis hin zum sozialen Rückzug kann die Früherkennung demenzieller Symptome der AD jedoch regelmäßig erschwert sein.

Leichtes Stadium. Die beginnende AD ist vor allem durch Schwierigkeiten mit dem Lernen und Erinnern gekennzeichnet. Während das Ultrakurzzeit-, das Kurzzeit- und das Altgedächtnis vergleichsweise gut erhalten sind, erscheint das Neugedächtnis besonders stark betroffen. Die Sprache erscheint stockend, weniger reichhaltig (abnehmendes Vokabular) und wenig präzise. In kognitiven Tests sind Wortfindungsstörungen und eine Reduktion der freien Wortwiedergabe zu identifizieren. Sowohl die Orientierung (v.a. zum Ort hin) als auch die konstruktive Praxie (Zeichenaufgaben) sind beeinträchtigt. Zu diesen kognitiven Demenzsymptomen treten häufiger auch nicht-kognitive Störungen, wie vor allem depressive Episoden (wohl zum Teil auch als Reaktion auf die Kompetenzverluste und deren Auswirkungen). Vielen demenzkranken Menschen gelingt trotz der merkbaren Beeinträchtigungen im Alltagsleben die weitgehende Aufrechterhaltung einer selbständigen Lebensführung. Insbesondere komplexe Alltagsverrichtungen wie Behördengänge oder Geldgeschäfte, die hohe Anforderungen an das planende Handeln und die Organisations- und Urteilsfähigkeiten stellen, müssen jedoch häufig durch Angehörige oder nahestehende Personen unterstützt werden.

Mittelschweres Stadium. Durchschnittlich 3 Jahre nach der Diagnosestellung zeigen sich deutlich reduzierte Gedächtnis- und Denkleistungen, die eine selbständige Lebensführung nahezu unmöglich machen. Das Neugedächtnis ist in dieser mittleren Erkrankungsphase schwerwiegend beeinträchtigt, und das logische Denken, Planen und Handeln massiv gestört. Die Betroffenen sind leicht ablenkbar und verkennen häufig Umgebungsreize, wobei bis zu 20 Prozent der mittelschwer Erkranken vorwiegend optische Halluzinationen ausbilden. Der sprachliche Ausdruck ist zunehmend lückenhaft mit einer deutlichen Häufung von Wortfindungsstörungen und Paraphrasien. Zu den kognitiven Einbußen gesellen sich nun häufiger auch Störungen der emotionalen Kontrolle und Bewegungsdrang (Unruhe, Umherwandern). Die Krankheitseinsicht der Betroffenen geht zunehmend verloren und insgesamt erscheint eine umfassende Supervision und Betreuung notwendig.

Schweres Stadium. Das finale Erkrankungsstadium wird im Mittel ungefähr 6 Jahre nach Diagnosestellung erreicht. In allen Bereichen kognitiver Funktion sind die Beinträchtigungen stark ausgeprägt. So sind beispielsweise auch gut überlernte frühe Erinnerungen kaum mehr verfügbar und die Sprache im Allgemeinen auf einfache Phrasen oder einzelne Wörter beschränkt. Vermehrt treten nun auch körperliche Symptome in den Vordergrund, wie beispielsweise massive Störungen der zirkadianen Rhythmik, Inkontinenz, Bewegungsstereotypien oder neurologische Störungen (Myoklonie, epileptische Anfälle, parkinsonoider Rigor). Ein großer Teil der schwer demenzkranken Menschen wird schließlich bettlägrig. Die häufigsten Todesursachen sind Lungenentzündung, Herzinfarkt und Sepsis. Insgesamt kann nach der Stellung der Diagnose AD mit einer voraussichtlichen Lebenserwartung von weiteren fünf bis acht Jahren gerechnet werden (Kurz & Greschniok, 1994).

2.1.2.2 Vaskuläre Demenzen

Unter dem Überbegriff vaskuläre Demenz kann eine mit Blick auf das klinische Zustandsbild recht heterogene Gruppe von Demenzen verstanden werden, deren gemeinsame Verursachung in zerebralen Durchblutungsstörungen liegt.

Für die Diagnose einer vaskulär bedingten Demenz muss, über die zur Feststellung eines demenziellen Syndroms notwendigen Symptome hinaus, durch Anamnese, klinische Urteilsbildung oder bildgebende Verfahren eine zerebrovaskuläre Erkrankung nachgewiesen werden, die mit der Demenzsyndromatik in einem hinreichend engen zeitlichen Zusammenhang (3 Monate) steht (Haberl & Schreiber, 2009).

Die Definition einer vaskulären Demenz nach dem ICD-10 hebt demgegenüber stärker auf die ungleichgewichtige Beeinträchtigung verschiedener kognitiver Funktionsbereiche ab und fordert die klinische Feststellung spezifischer funktionaler Ausfälle aufgrund einer fokalen Hirnschädigung.

F01 vaskuläre Demenz

1. Die allgemeinen Kriterien für eine Demenz (G1.-G4.) müssen erfüllt sein.
2. Ungleiche Verteilung der Defizite höherer kognitiver Funktionen, von denen einige betroffen, andere relativ verschont sind. So kann das Gedächtnis eindeutig eingeschränkt sein, während das Denken, Urteilen und die Informationsverarbeitung nur mäßig beeinträchtigt sind.
3. Nachweis einer fokalen Hirnschädigung, die durch ein oder mehrere der folgenden Merkmale angezeigt wird:
   (a) einseitige spastische Hemiparese der Gliedmaßen
   (b) einseitig gesteigerte Muskeleigenreflexe
   (c) positiver Babinskireflex
   (d) Pseudobulbärparalyse
4. Eindeutiger Nachweis einer zerebrovaskulären Krankheit aus der Anamnese, aufgrund von Untersuchungen oder besonderen Tests, die für die Demenz verantwortlich gemacht werden kann (z.B. Insultanamnese, Nachweis einer zerebralen Infarzierung).

Klinisch äußert sich eine vaskuläre Demenz durch ein vergleichsweise plötzliches Auftreten von kognitiven Störungen in zeitlichem Zusammenhang mit einer zerebrovaskulären Erkrankung, sowie einen fluktuierenden, schubweisen Verlauf. Bereits zu Beginn der Erkrankung lassen sich ein auffällig kleinschrittiges, engbasiges, mitunter schlurfendes oder auch spastisches Gangbild und eine erhöhte Sturzgefährdung feststellen. Daneben kommt es häufig zu einer Miktionsstörung im Sinne häufigeren Kontinenzdranges. Da die Trias Demenz, Gangstörung, Blaseninkontinenz daneben auch für einen Normaldruck-Hydrozephalus charakteristisch ist, bleibt die klinische Unterscheidung hier häufig unscharf. Auf fokalneurologische Zeichen, die sich je nach Lokalität der Ischämie unterschiedlich darstellen können, wurde im Rahmen der ICD-10-Klassifikation bereits hin-

gewiesen. Typische pyramidale Zeichen sind beispielsweise Hemiparese und/oder zentrale Faszialisparese mit positivem Babinski-Zeichen. Extrapyramidale Symptome umfassen Tonussteigerung und Akinese. Darüber hinaus kommt es häufig auch zum pseudobulbären Syndrom mit Sprech- und Schluckstörungen und affektiver Labilität mit pathologischem Lachen oder Weinen. Einbußen in der Affektivität (v.a. depressiv gefärbte Grundstimmung und Gleichgültigkeit) und dem Antriebsverhalten (Rückzug und Teilnahmslosigkeit) kennzeichnen das ebenfalls häufige Frontalhirnsyndrom.

Eine gängige, wenn auch nicht einheitlich vertretene Subdifferenzierung vaskulärer Demenzen unterscheidet Infarkte im Versorgungsgebiet großer Hirnarterien (territoriale Infarkte) von Mikroangiopathien kleinerer (Arteriolen) und kleinster (Kapillaren) Hirngefäße (lakunäre Infarkte). Beide Formen der vaskulären Schädigung können daneben isoliert auftreten und an entsprechend ungünstigen Stellen ein demenzielles Syndrom bedingen, oder aber als aggregierte multiple Infarkte auch stärker multilokal wirken. Für das klinische Zustandsbild erscheint darüber hinaus eine Unterscheidung der betroffenen kortikalen oder subkortikalen Strukturen bedeutsam. Kortikale vaskuläre Demenzen, die sich häufig aufgrund von atherothrombotischen oder kardiogen-embolischen Infarkten ergeben sind durch plötzlich auftretende Lähmungen, Sensibilitätsstörungen und aphasische Symptome bestimmt. Subkortikale Infarkte gehen häufig einher mit isolierten Pyramidenbahnzeichen wie Haltungs- und Tonusanomalien, einem Pseudobulbärhirnsyndrom oder einem Frontalhirnsyndrom.

2.1.2.3 Demenz bei weiteren Krankheitsbildern

Auch wenn der Großteil der von einem demenziellen Syndrom Betroffenen Veränderungen vom Alzheimer-Typ, eine zerebrovaskuläre Erkrankung oder eine Kombination dieser Erkrankungskomplexe aufweist, sollen im folgenden einige der wichtigsten neurodegenerativen, entzündlichen, ischämischen und psychogenen Ätiologien der Demenzerkrankung zumindest angerissen werden.

Demenz bei Basalganglienerkrankung

Zu den prominentesten neurologischen Krankheitsbildern, die u.a. die Basalganglien betreffen, gehören Morbus Parkinson und Chorea Huntington, wenngleich eine ganze Reihe weiterer Krankheiten primär oder sekundär einen Niedergang frontosubkortikaler Strukturen mitbedingen. Erkrankungen der Basalganglien beeinträchtigen die Funktionsfähigkeit mehrerer motorisch und nicht-motorisch relevanter frontosubkortikaler Schaltkreise. Zu den kognitiven und anderen psychischen Symptomen gehören die Störung exekutiver Funktionen, des Arbeitsgedächtnisses und komplexer motorischer Programme (dorsolateraler präfrontaler Schaltkreis), Störungen in der Affektsteuerung, Antwortunterdrückung und Selbstkontrolle mit den Folgen Persönlichkeitsveränderung und unangemessenem Verhalten (lateraler orbitofrontaler Schaltkreis) und Apathie, Antriebsminderung, akinetischer Mutismus sowie Zwangsverhalten bei einer Reihe verschiedener Bewe-

gungsstörungen (anteriorer zingulärer mediofrontal-limbischer Schaltkreis). Gewöhnlich fehlen Symptome der Aphasie, Apraxie, Agnosie und Amnesie, sodass einzelne Schritte einer komplexen Aufgabe gelöst werden können, eine Integration der Einzelkomponenten jedoch nicht möglich ist.

In einem aktuellen Übersichtskapitel unterscheidet Weindl (2009) symptomatisch zwischen überwiegend hypokinetischen (z.B. Morbus Parkinson) und hyperkinetischen Bewegungsstörungen (z.B. Chorea Huntington). Im folgenden sind die ICD-10-Diagnosekriterien für diese beiden prominenten Vertreter subkortikal gelagerter Demenzen aufgeführt.

Demenz bei Prionkrankheit

Die Creutzfeld-Jakob-Demenz (CJD) ist eine durch infektiöse Eiweißpartikel (proteinaceous infectious agents, Prionen) verursachte spongiforme (vakuoläre) Hirnveränderung mit Neuronenverlusten. Neben der in den meisten Fällen (90%) spontan entstehenden (sporadischen bzw. idiopathischen) CJD kommen jedoch auch hereditäre (beispielsweise eine familiäre Form der CJD (10-15%), das seltene Gerstmann-Sträussler-Scheinker-Syndrom oder die noch seltenere tödliche familiäre Insomnie) oder erworbene Prionenerkrankungen (v.a. die vermutlich durch BSE verursachte *neue Variante* der CJD) vor.

Die klinische Diagnose einer CJD wird insbesondere durch eine rasche Progredienz der Beeinträchtigungen (weniger als zwei Jahre Dauer), pyramidale und extrapyramidale Zeichen, Myoklonus, Sehstörungen oder andere Kleinhirnstörungen und in späteren Stadien durch akinetischen Mutismus und apallisches Syndrom nahegelegt. Durch apparative (typisches Muster der EEG-Ableitung i.S. periodischer Sharp-Wave-Komplexe) und laborchemische (positiver 14-3-3-Liquorbefund) Untersuchungen kann der Verdacht auf eine CJD bestätigt bzw. die Diagnose gesichert werden.

Demenz bei fokalen Hirnathrophien

Eine vergleichsweise häufig (bis zu 20% aller Demenzen) zu beobachtende Gruppe demenzieller Erkrankungen wird als Pick-Komplex bezeichnet. Bei dieser häufig auch als frontotemporale Lobärdegeneration (FTLD) bezeichneten Demenzform werden keinerlei Hinweise auf eine Alzheimer-Krankheit, vaskuläre Demenz oder Verursachung durch Prionen gefunden. In Abhängigkeit von der Lokalisation des degenerativen Prozesses werden die drei prototypischen Syndrome frontotemporale Demenz (FTD), progrediente unflüssige Aphasie (PA) und semantische Demenz (SD) unterschieden (vgl. Danek, Wekerle & Neumann, 2009).

Bei der *frontotemporalen Demenz* (auch Frontallappendemenz) stehen Veränderungen der Persönlichkeit und Störungen der Sozialbeziehungen im Vordergrund. Die gestörte Fähigkeit, Affektivität und (soziales) Verhalten angemessen zu steuern äußert sich in beruflicher und sozialer Unzuverlässigkeit, Missachtung von Normen bis hin zu kriminellem Verhalten, Taktlosigkeit und verändertem Sexualverhalten. Daneben sind auch Apathie und sozialer Rückzug, oder aber Unruhe, Hyperaktivität und Bewegungsdrang charak-

teristisch. Dagegen bleiben die Sinnesfunktionen, die räumlich-konstruktive Praxie und das Gedächtnis weitgehend unbeeinträchtigt. Diese recht heterogenen Symptome können häufig einem pseudoneurasthenischen (Apathie, Rückzug) oder einem pseudopsychopathischen (enthemmten) Typus zugeordnet werden.

Die *progrediente unflüssige Aphasie* ist in ihrem gesamten Verlauf durch eine Störung der Sprachproduktion mit Agrammatismus (Telegrammstil), phonematischen Paraphrasien (Lautfehlern) oder Störungen der Wortfindung und Benennung gekennzeichnet. Andere kognitive Funktionsbereiche sind dagegen kaum betroffen.

Bei der *semantischen Demenz* ist der Abruf des Bedeutungsgehaltes eines Wortes oder Gegenstandes beeinträchtigt. Der Sinn von Worten beispielsweise kann nicht erklärt werden, oder Bedeutungszuschreibungen bleiben auf einer vergleichsweise oberflächlichen Ebene verhaftet (Oberflächendyslexie). Diese Störung zeigt sich auch in einer der Spontansprache entsprechenden Schreibweise (Oberflächendysgraphie). Schließlich kann ein gestörtes Wissen um Sinn und Zweck von Objekten (assoziative Agnosie) auch zu fehlerhafter, gegebenenfalls selbst- und fremdgefährdender Benutzung von Gegenständen führen. Auch bei der semantischen Demenz bleiben weitere kognitive Funktionsbereiche (v.a. parietale und okzipitale Funktionen wie räumlich-konstruktive Praxie) weitestgehend erhalten.

Die vorgenannte inhaltliche Differenzierung verschiedener Syndromtypen fokaler Hirnathrophie wird durch die ICD-10-Kriterien zur Klassifizierung eines Demenzsyndroms bei Pick-Krankheit nicht explizit angesprochen.

Nicht zuletzt wegen der vergleichsweise eng umrissenen Defizitbereiche und der bei allen drei Leitsyndromen nur nachgeordneten Rolle des Gedächtnisverlustes soll hier darauf hingewiesen werden, dass für die Kodierung einer Demenz bei Pick-Krankheit nach ICD-10 (F02.0) selbstverständlich alle notwendigen Kriterien eines Demenzsyndroms gegeben sein müssen.

Demenz bei Depression

Kognitive Leistungseinbußen i.S. eines Demenzsyndroms können auch im Kontext einer schweren Depression auftreten. Dabei sind vor allem ein Nachlassen des Interesses, eine kognitive Verlangsamung oder auffällige Umständlichkeit zentrale Symptome. Wesentliche Unterschiede zwischen dieser prinzipiell reversiblen demenziellen Symptomatik und derjenigen bei organisch begründeten Demenzen sind das gewöhnlich akute Auftreten der Beeinträchtigungen, die spezifisch auf Leistungen in neuropsychologischen Tests beschränkt bleiben und durch die Betroffenen explizit bzw. expressiv beklagt werden. Gewöhnlich kann die zugrunde liegende Störung des Affektes nachgewiesen werden. Die selbständige Lebensführung (z.B. Hygiene oder Orientierung) erscheint dagegen vergleichsweise gut erhalten. Apparative bzw. bildgebende Verfahren zeigen keine Hinweise auf eine organische Ursache der Beeinträchtigungen.

Auch wenn für die kognitiven Folgeerscheinungen schwerer Depressionen häufig der Begriff *depressive Pseudodemenz* verwendet wird, sind die demenziellen Symptome nicht

lediglich eingebildet oder vorgetäuscht. Mit der Besserung der affektiven Lage der Betroffenen geht gewöhnlich auch eine Verbesserung der kognitiven Leistungsfähigkeit einher.

Das gemeinsame Auftreten demenzieller und affektiver Krankheitssymptome kann darüberhinaus auch eine ursächlich nicht zusammenhängende Komorbidität darstellen, da Depressionen zu den häufigsten psychischen Erkrankungen im höheren Lebensalter zählen. Insbesondere in frühen Stadien der Demenzerkrankung kann eine depressive Symptomatik schließlich auch reaktiv als Folge der erlebten demenzbedingten Kompetenzverluste entstehen.

2.1.3 Behandlung und Versorgung demenzkranker Menschen

Therapie. Eine ursächliche Behandlung der Demenz ist gegenwärtig nicht verfügbar. Symptomatische medikamentöse, psychosoziale und psychotherapeutische Therapieansätze adressieren sowohl die kognitiven Leistungseinbußen als auch die affektiv-behavioralen Begleitsymptome der Demenz (Hirsch, 2001). Psychopharmakologisch erscheint im Kontext der Alzheimer-Demenz durch die rechtzeitige Gabe von Antidementiva (z.B. Acetylcholinesterasehemmern) eine Verzögerung des kognitiven Leistungsabbaus um bis zu 24 Monate möglich (Förstl et al., 2009). Auch bei Patienten mit vaskulärer Demenz erwiesen sich Acetylcholinesteraseinhibitoren als wirksam (Roman, 2005). Psychopathologische Symptome können mit entsprechenden Psychopharmaka behandelt werden (Giacobine, 2000; Greig et al., 2005). Die Befunde zur Effektivität und Nebenwirkungen von Neuroleptika, (trizyklischen) Antidepressiva und Benzodiazepinen deuten jedoch darauf hin, das nicht-medikamentösen Therapieversuchen auf jeden Fall Vorrang vor psychopharmakologischer Intervention eingeräumt, und falls nötig auf nebenwirkungsärmere Substanzgruppen zurückgegriffen werden sollte. Allerdings erscheinen demenzkranke Heimbewohner gegenwärtig mit antidepressiven Medikamenten eher unterversorgt zu sein, obwohl für die Substanzgruppe der neueren selektiven Serotonin-Wiederaufnahme-Hemmer (SSRIs) sowohl eine gute Wirksamkeit als auch Verträglichkeit beschrieben werden konnte (s. zusammenfassend Schröder, 2006).

Versorgung. Gegenwärtig werden immernoch ungefähr 75 Prozent aller Demenzkranken familiär versorgt (Bruder, 2009). Im späteren Verlauf der Erkrankung jedoch wird der Großteil der Demenzkranken stationär weiterversorgt (Bickel, 1996). Über einen Zeitraum von zwölf Monaten hinweg wurden in Abhängigkeit der untersuchten Population Einweisungsraten zwischen 20 und 33 Prozent berichtet. Für längere Zeiträume (bis 5 Jahre) stieg die Institutionalisierungsrate deutlich auf bis zu 50 Prozent an (Luppa et al., 2008). Die Dauer zwischen Diagnosestellung und Heimunterbringung wurde in einer bundesweit repräsentativen Längsschnittstudie kürzlich auf ca. 35 Monate geschätzt (Luck et al., 2008).

Die Entscheidung für den Wechsel in die stationäre Versorgung wird im Kontext des von Andersen vorgestellten Modells der Inanspruchnahme von Pflegeleistungen durch soziokulturelle Hintergrundvariablen (Soziodemographie, Beziehungsqualität zwischen Pflegenden und Betroffenen), konkrete Pflegebedarfe und Bedürfnislagen der Demenzkranken und Pflegenden (z.B. Schweregrad der Erkrankung), sowie persönliche (z.B. Bewälti-

gungsstrategien) und soziale bzw. gesellschaftliche Ressourcen und Unterstützungsangebote mitbestimmt angenommen (Andersen, 1995). Einen aktuellen Überblick der empirischen Studien zu Prädiktoren der Heimaufnahme geben Luppa und Kollegen (Luck et al., 2008; Luppa et al., 2008). Insbesondere nicht-kognitive Demenzsymptome wie Aggressivität, Wahnvorstellungen oder Depressionen werden von den pflegenden Angehörigen als besonders belastend empfunden und erhöhen das Risiko einer Heimeinweisung (Bruder, 2009; Coehlo, Hooker & Bowman, 2007; Luppa et al., 2008).

Der Anteil von mittelschwer bis schwer demenziell beeinträchtigten Personen in Altenpflegeheimen beträgt gegenwärtig ungefähr zwei Drittel (Weyerer, Schäufele & Hendlmeier, 2005). In Anbetracht der vorauszusehenden demographischen Entwicklung ist davon auszugehen, dass die Gruppe der demenzkranken Heimbewohner das Aufgabenfeld der stationären Pflege zukünftig in noch stärkerem Maße bestimmen wird.

2.2 Schmerzbelastung demenzkranker Menschen

In diesem Unterkapitel wird der Wissenstand zur Schmerzbelastung demenzkranker Menschen dargestellt. Zunächst werden die wichtigsten neurobiologischen Grundlagen des Phänomens Schmerz kurz erläutert und verschiedene Möglichkeiten der Definition des Schmerzes vorgestellt. Da Demenzen alterskorrelierte Erkrankungen darstellen, werden dann zunächst Veränderungen im Schmerzerleben beschrieben, die im Kontext des normalen, nicht-pathologischen Alterns, erwartet werden können. Von diesen schmerzbezogenen Mechanismen normalen Alterns werden danach solche Veränderungen im Schmerzerleben und Schmerzausdruck abgegrenzt, die spezifisch mit neurodegenerativen Prozessen im Rahmen der demenziellen Erkrankung verschiedener Ätiologie assoziiert sind.

2.2.1 Definition Schmerz

Bei der Definition von Schmerzen steht das subjektive Erleben zumeist im Vordergrund. So definieren beispielsweise McCaffery und Beebe (1994; S.15) Schmerz als „stets so, wie die empfindende Person sagt, dass er ist, und vorhanden, wann immer sie sagt, dass er vorhanden ist“.

Die Notwendigkeit einer differenzierten Introspektions- und Kommunikationsfähigkeit der betroffenen Person für das Assessment von Schmerzen ist dabei offensichtlich. Ebenso offensichtlich ist die notwendige grundsätzliche Bereitschaft, die erhaltene Selbstauskunft als verlässlich und valide anzuerkennen. Erscheinen Kommunikations- und Einsichtsfähigkeit der Betroffenen, wie beispielsweise im Verlauf der Demenz, nun jedoch zunehmend eingeschränkt, oder muss angenommen werden, dass die Selbstauskunft zugunsten eigener Vorteilsnahme unaufrichtig ist, richtet sich der Blick nicht selten auf potenzielle krankheits- oder situativ bedingte Schmerzursachen zur vermeintlichen Bestätigung oder Entkräftung unklarer Schmerzäußerungen (vgl. Hadjistavropoulos, Craig & Fuchs-Lacelle, 2004, p. 93).

Die durch den Verweis auf den individuellen Charakter des Schmerzerlebens implizierte Dichotomie *subjektiv vs. objektiv* erscheint problematisch, da Subjektivität häufig mit Unzuverlässigkeit und Unwissenschaftlichkeit assoziiert wird. Solange nicht anerkannt wird, dass die objektive Erfassung subjektiver Phänomene in erster Linie die Unabhängigkeit des Messwertes von Merkmalen des Versuchsleiters bzw. Fragenden verlangt, stellt diese Dichotomie einen Hemmschuh für die Entwicklung und Akzeptanz von Schmerzassessments dar. Die Herausforderungen, die sich mit der sozialen Eingebundenheit einer Schmerzmessung ergeben können, werden ausführlich in Kapitel 3.4.6 diskutiert.

Manche Schmerzdefinitionen verweisen warnend auf die nur schwer bestimmbare Assoziation zwischen der Schmerzempfindung und vom Betroffenen unabhängig feststellbaren (objektivierbaren) Noxen:

> Schmerz ist eine unangenehme sensorische und emotionale Erfahrung in Verbindung mit einer tatsächlichen oder möglichen Gewebsschädigung oder beschrieben in Begriffen einer solchen Schädigung. (IASP, 1994)

Personengruppen, die sich nicht (mehr) verbal zu ihrem Schmerzerleben äußern können laufen jedoch auch nach dieser Definition Gefahr, zu selten als schmerzbelastet diagnostiziert zu werden (Anand & Craig, 1996; Craig & Hadjistavropoulos, 2004). Die Unfähigkeit, Schmerzen verbal zu kommunizieren bedeutet keinesfalls, dass ein Individuum keine Schmerzen verspürt und einer angemessenen Schmerzbehandlung bedarf. Die implizierte Dichotomisierung des zugrundeliegenden Kontinuums der Kommunikationsfähigkeit in auskunftsfähig oder -unfähig und die Fokussierung auf den Ausdrucksbereich der verbalen Kommunikation müssen als Schwächen dieser Definition berücksichtigt werden (Hadjistavroupolous, von Baeyer & Craig, 2001).

Auch somatisch orientierte Arbeiten zu Schmerz und Schmerzerleben schlagen mitunter recht verschiedene Definitionen des Schmerzes vor. Die naive Vorstellung eines fest verkabelten größtenteils exterozeptiven Schmerzsystems, welches den Körper in Notsituationen in überlebensnotwendige Bereitschaft versetzt, wurde mittlerweile durch eine Perspektive überwunden, die dem Schmerz als interozeptiver Empfindung komplexe homöostatische Funktionen zuschreibt:

> [...] pain is one aspect of the physiological condition of the body, which homeostatic (autonomic, neuroendocrine, and behavioral) mechanisms serve to maintain in an optimally balanced state. (Craig, 2003, p. 22)

Schmerz wird aus dieser integrierenden Sicht ähnlich wie Hunger oder Durst als homöostatische Emotion mit sowohl sensorischen als auch motivationalen Qualitäten begriffen.

Die Berücksichtigung des Einflusses psychologischer Mechanismen auf den Schmerz, wie sie beispielsweise durch die Gate-Control-Theorie von Melzack und Wall (1965) eingefordert wurde, kann als wichtiger Schritt zur Überwindung der im cartesischen Denken verhafteten Spezifitätsperspektive auf Schmerz gelten. Weiterentwicklungen der Gate-Control-Theorie verschoben den Fokus noch deutlicher in Richtung einer komplexen zentralen Körperwahrnehmung und -steuerung:

> These views are in sharp contrast to the classical specificity theory in which the qualites of experience are presumed to be inherent in peripheral nerve fibers. Pain is not injury; the *quality of pain experience* must not be confused with the physical event of breaking skin or bone. Warmth and cold are not "out there"; temperature changes occur "out there", but the *qualities of experience* must be generated by structures in the brain. (Melzack & Katz, 2004, p. 23; Hervorhebungen im Original)

Der Dualismus zwischen für Schmerz und einzelne Schmerzqualitäten spezifischen neuronalen Strukturen und einer weitgehenden multilokal-synthetischen Generierung von Schmerzqualitäten wird auch in der nachfolgenden Darstellung der neuroanatomischen und -psychologischen Grundlagen menschlichen Schmerzerlebens deutlich.

2.2.2 Neurophysiologische und -psychologische Grundlagen

Aus den zuvor aufgeführten Definitionsversuchen des Phänomens Schmerz wird deutlich, dass die Beziehung zwischen objektiv feststellbaren physikalischen Gegebenheiten (z.B. der Stärke eines elektrischen Reizes) und den erlebten introspektiv zugänglichen emotionalen, motivationalen oder kognitiv-evaluativen Qualitäten der Schmerzempfindung nicht als eine analoge Funktion gedacht werden kann.

In Anbetracht der hohen Variabilität der individuellen Schmerzverarbeitung verwundert es nicht, dass die Suche nach dem neurophysiologischen Substrat regelhafter Schmerzverarbeitung auch gegenwärtig noch nicht abgeschlossen ist. Eine ausführliche Darstellung angenommener Schmerzmechanismen oder Strukturen des Schmerzsystems kann an dieser Stelle selbstverständlich nicht geleistet werden. Dennoch sollen kurz einige neurophysiologische und neuropsychologische Grundlagen erläutert werden, die für das Verständnis des multidimensionalen Schmerzerlebens und insbesondere der im Rahmen einer demenziellen Erkrankung zu erwartenden Veränderungen des Schmerzerlebens notwendig erscheinen.

2.2.2.1 Zentrale und periphere Anteile des nozizeptiven Systems

Den peripheren Teil des Schmerzsystems bilden Schmerzrezeptoren (sog. Nozizeptoren) der Haut, Muskeln, Gelenke oder der Viszera, die durch die bei einer Gewebeschädigung freigesetzten chemischen Substanzen erregt werden. Dabei sind verschiedene Nozizeptoren unterschiedlich stark auf einzelne Reizmodalitäten (Temperatur, Druck) spezialisiert. Andere Nozizeptoren reagieren erst auf eine tatsächliche Gewebeschädigung oder entzündliche Prozesse (Craig, 2003). An der Übermittlung der noxischen Impulse sind verschieden schnell leitende Nervenfasern beteiligt. Die markarmen Aδ-Fasern leiten die Schmerzimpulse deutlich schneller (bis zu 40 m/s) als die marklosen C-Fasern (unter 2 m/s). Für das Verständnis von Schmerzwahrnehmung und -kontrolle sind daneben die ebenfalls sehr schnell leitenden Aβ-Fasern von Bedeutung. Diese im wesentlichen

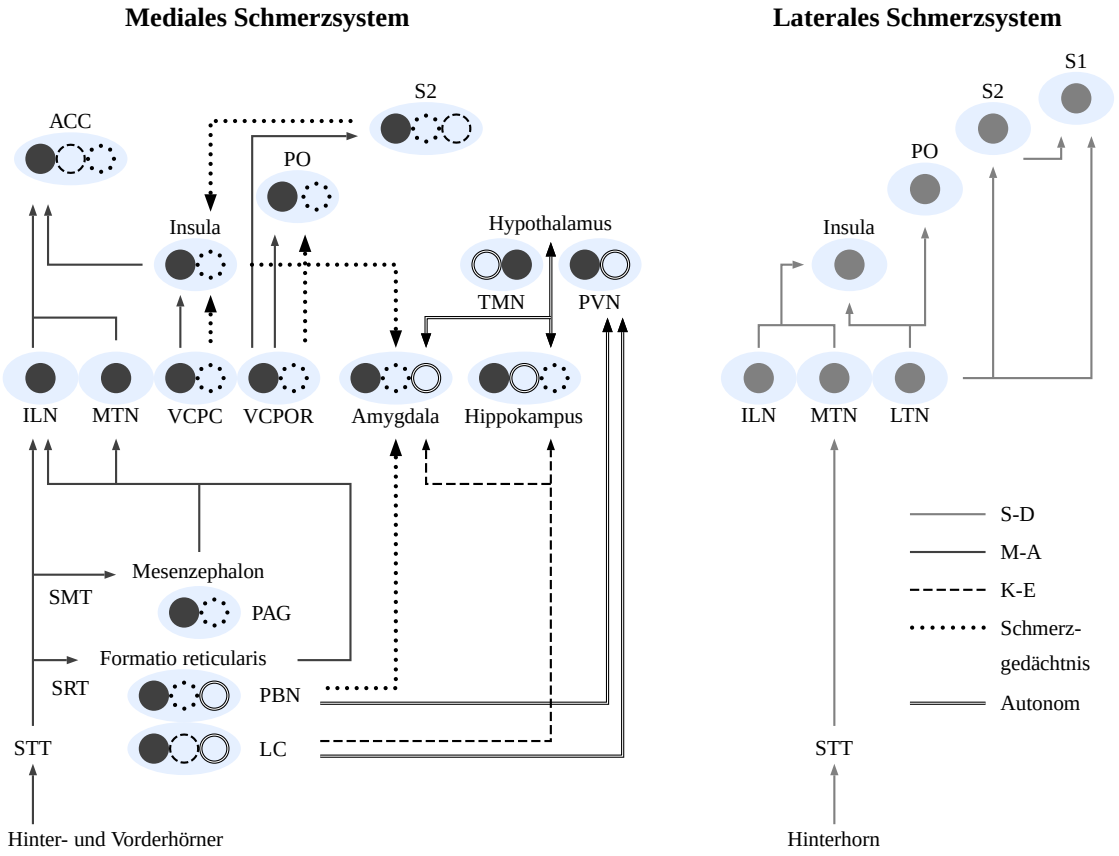
in der Haut konzentrierten Fasern sind für sich genommen allerdings nicht schmerzspezifisch, sondern übermitteln Reize und Empfindungen, die in der Regel nicht als schmerzhaft wahrgenommen werden. Da sie jedoch vergleichsweise schnell leiten und zum Teil denselben Weg zum ZNS zurücklegen, können sie die eigentliche Schmerzübermittlung beeinflussen. Die lokal evozierten Signale ziehen über Aδ-Fasern, Aβ-Fasern und C-Fasern zunächst zum Hinterhorn des Rückenmarks, wo sie in verschiedenen Zellschichten (Laminae) und insbesondere in der Substantia gelatinosa (Laminae I und II) auf sekundäre Neurone umgeschaltet werden. Diese sekundären Neurone sind ihrerseits unterschiedlich schmerzspezifisch. Wide Dynamic Range (WDR)-Neurone können auch durch Reize von geringerer Intensität (leichte Berührung, Reiben) erregt werden, während Nociceptor-Specific (NS)-Neurone überwiegend durch starke, noxische Reize erregt werden. Beide Neuronenarten übermitteln interozeptive Signale zum Stammhirn und zu somatosensorischen Bereichen der Hirnrinde, also den zentralen Anteilen des Schmerzsystems, in denen die Schmerzwahrnehmung und -verarbeitung stattfindet. Die aktuelle Sichtweise auf Schmerz begreift das Schmerzsystem als Teil eines komplexen homöostatischen Steuerungsmechanismus, bei dem die Impulse von nach Lokalisation und Reizqualität differenzierbaren Rezeptoren und solchen mit unspezifischerem Erregungsmuster in zentralen Strukturen integriert werden (Craig, 2003). Die über Aδ-Fasern vermittelten Schmerzimpulse erlauben eine recht genaue Lokalisation des noxischen Reizes und führen zu einer augenblicklichen Reflexantwort. Da diese Fasern an ihrer Oberfläche keine Opioidrezeptoren tragen, bleibt die unmittelbare Reaktion auf den bereits 1942 durch Sir Thomas Lewis beschriebenen sog. *ersten Schmerz* (z.B. Zusammenzucken bei Nadelstich) in der Regel auch bei analgetisch versorgten Patienten erhalten. Die durch die C-Fasern etwas langsamer geleiteten Schmerzimpulse sind dagegen gewöhnlich mit einer etwas weniger stark lokal begrenzten, dumpferen Schmerzempfindung, dem sog. *zweiten Schmerz* assoziiert, der sich gut durch Opioid-Analgetika kontrollieren lässt.

2.2.2.2 Mediales und laterales Schmerzsystem

Das komplexe nozizeptive Verarbeitungssystem im Gehirn kann in zwei untereinander nur wenig verbundene Subsysteme unterteilt werden, deren subkortikale und kortikale Strukturen für die spezifische Ausgestaltung der sensorischen, affektiven und kognitiven Qualitäten der Schmerzempfindung verantwortlich gemacht werden. Eine Übersichtsarbeit zur Struktur der Schmerzverarbeitung im zentralen Nervensystem bei Gesunden und Demenzkranken mit verschiedenen Demenzätiologien wurde von Scherder und Kollegen (2003) veröffentlicht.

Abbildung 1 fasst die gegenwärtigen Erkenntnisse zu an der Schmerzverarbeitung beteiligten subkortikalen und kortikalen Gebiete und ihrer Bedeutung für verschiedene Qualitäten des Schmerzerlebens zusammen. Das laterale Schmerzsystem erscheint dabei vorrangig für die sensorisch-diskriminativen Anteile des Schmerzerlebens verantwortlich. Diese Subkomponente der Schmerzverarbeitung ist insbesondere für die Identifikation der Lokalisation, Intensität und Perseveranz noxischer Reize zuständig. Vom Hinterhorn wer-

Abbildung 1: Strukturen des medialen und lateralen Schmerzsystems (adaptiert nach Scherder et al., 2003).



Subkortikale und kortikale Strukturen und Projektionsbahnen des medialen und lateralen Schmerzsystems. TMN=Tuberomamillarkern; PO=parietales Operculum; PVN=Paraventrikularkern; ILN=intralaminare Thalamuskerne; MTN=mediale Thalamuskerne; PBN=Parabrachialkern; PAG=periaquäduktales Grau; SMT=spinomesenzephalitischer Trakt; SRT=spinoreticularer Trakt; STT=spinothalamischer Trakt; LC=Locus Coeruleus; S-D=sensorisch-diskriminativ; M-A=motivational-affektiv; K-E=kognitiv-evaluativ.

den die sensuo-diskriminatorisch relevanten Informationen über den spinothalamischen Trakt zu den ventrolateralen Kerngruppen des Thalamus (Intralaminare, mediale und laterale Thalamuskerne), und weiter zu den somatosensorischen Kortexarealen I und II, der Insula und dem parietalen Operculum projiziert.

Motivational-affektive Schmerzqualitäten werden dagegen vorrangig durch das mediale Schmerzsystem projiziert und und in den Strukturen des limbischen Systems verarbeitet. Das mediale Schmerzsystem leitet nozizeptive Reize entweder direkt über spinothalamische Bahnen zu den intralaminaren und medialen Thalamuskernen oder zunächst über den spinoreticularen Trakt in den Parabrachialkern und Locus coeruleus der Formatio reticularis bzw. über den spinomesenzephalitischen Trakt in das periaquäduktale Grau des Mittelhirns, bevor die nozizeptiven Impulse an thalamische Strukturen, die Amygdala oder den Hippocampus weitergeleitet werden. Weitere am medialen Schmerzsystem beteilig-

te Strukturen sind der Hypothalamus, sowie der anteriore cingulate und parasylvianische Kortex. Letzterer umfasst die Insula, das parietale Operculum und den somatosensorischen Kortex II.

Auch an der Verarbeitung kognitiv-evaluativer Schmerzqualitäten (Aufmerksamkeitssteuerung, Antizipation und Schmerzgedächtnis) sind überwiegend Strukturen des medialen Schmerzsystems beteiligt.

2.2.2.3 Schmerzmodulation

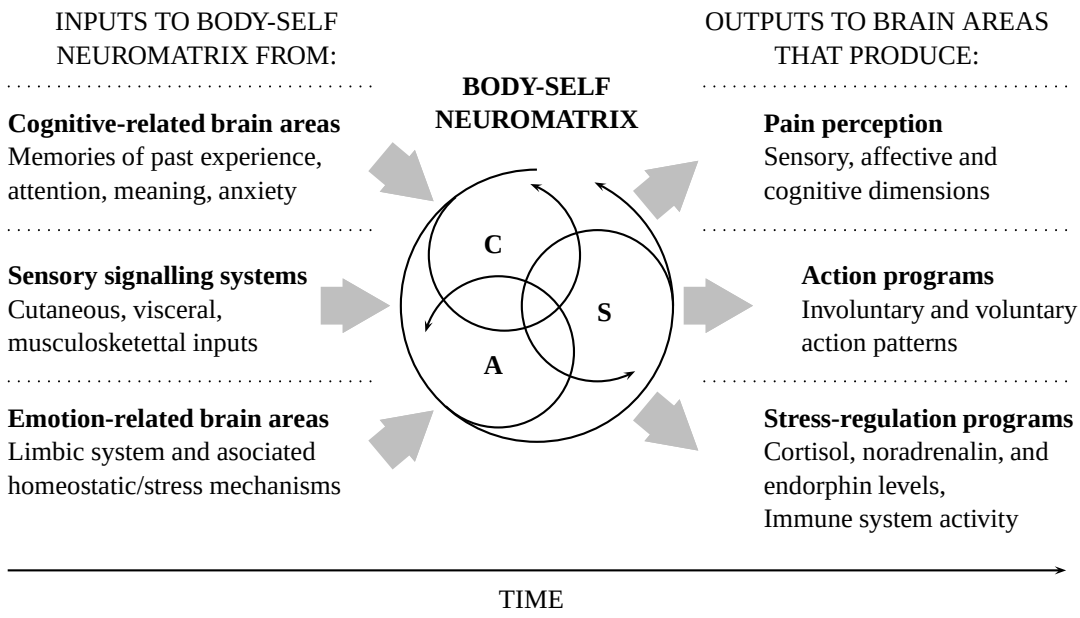
Neben den beschriebenen afferenten Impulsen aus der Peripherie, die auf zentraler Ebene entsprechende exzitatorische Vorgänge auslösen findet stets auch eine efferente inhibitorische Antwort auf nozizeptive Reize statt. Die für die absteigende Schmerzhemmung wichtigsten Hirnareale sind dabei das periaquäduktale Grau im Mesencephalon, der Locus coeruleus in der Formatio reticularis und der Nucleus paraventricularis im Hypothalamus. Die Hirnstammareale weisen eine besonders hohe Dichte an körpereigenen (endogenen) Opioiden und Opioidrezeptoren auf und bilden den Ausgangspunkt absteigender inhibitorischer Schmerzsysteme mit serotoninerger und noradrenerger Neurotransmission (Kunz, 2006; Scherder, Sergeant & Swaab, 2003).

Zentrale Aufmerksamkeits- oder Gedächtnisprozesse können das individuelle Erleben von Schmerzen verstärken. Entspannung, Ablenkung oder sozial-kulturelle Vorstellungen können andererseits die Empfindung auch starker Reizintensitäten reduzieren. Die insgesamt wohl einflussreichste Theorie dieser Steuerung des Schmerzerlebens ist die bereits 1965 von Melzack und Wall vorgestellte Gate-Control-Theory. Dabei wird stark vereinfacht angenommen, dass die Weiterleitung der ins Hinterhorn eintretenden nozizeptiven Impulse an höhere kortikale Strukturen durch absteigende inhibitorische Prozesse moduliert werden kann. Auch die gleichzeitige Aktivierung nicht-schmerzbezogener Empfindungen über Aβ-Nervenleitungen durch Reiben oder (Reizstrom-)Massage kann danach „das Tor zur Schmerzwahrnehmung schließen“ indem die WDR-Neuronen des Hinterhorns aktiviert und gewissermaßen mit nicht-schmerzbezogenen Reizen belegt werden.

Wenngleich die Gate-Control-Theorie wesentlich dazu beigetragen hat zu verstehen, wie in der Peripherie des Nervensystems wahrgenommene potenziell schmerzhafte Reize an zentrale Hirnstrukturen weitergeleitet oder gehemmt werden können, können einige der markantesten Befunde der klinischen Schmerzforschung, beispielsweise die erhaltene (Schmerz-)Empfindlichkeit in amputierten Gliedmaßen oder Körperbereichen, die durch eine Unterbrechung des Rückenmarks (z.B. Querschnittslähmung) keine afferenten Impulse mehr an zentrale Hirnstrukturen mehr leiten können, im Rahmen dieser Theorie nicht erklärt werden.

In aktuelleren Arbeiten entwickelten Melzack und Kollegen aus diesem Grunde eine umfassendere Theorie der zentralen Repräsentation und Steuerung des eigenen Körpers (Melzack, 2001; Melzack & Katz, 2004). Dabei wird angenommen, dass die innere Repräsentation des eigenen Körpers auf neuronaler Ebene durch kreisförmige Verschaltungen thalamisch-limbisch-kortikaler Strukturen, der sogenannten *Neuromatrix*, angelegt sei (Abb. 2).

Abbildung 2: Das Konzept der Neuromatrix von Melzack (2001)



In dieser zu einem großen Teil genetisch vorbestimmten Verarbeitungsstruktur sind alle notwendigen Mechansismen zur sensorischen, affektiven und kognitiven Steuerung der Homöostase des menschlichen Körpers bereits angelegt. Danach ist es nicht unbedingt notwendig, dass der Körper sich durch externe Reize im Laufe seiner Entwicklung erst „erfährt“ und ein Körperbild formt. Auch ohne aktive Impulse aus der Peripherie scheinen auf zentraler Ebene kontinuierlich aufeinander bezogene Input- und Outputprozesse stattzufinden. Die Integration verschiedener sensorischer Inputs und entsprechender Regulationsmechanismen ist damit durch das individuelle Raster der Neuromatrix geprägt. Die in der Neuromatrix kontinuierlich ablaufenden sowohl parallelen als auch zyklisch aufeinander bezogenen synthetischen Prozesse bilden eine individuelle *Neurosignatur*, die ihrerseits die Grundlage des kontinuierlichen Stromes eigenen Körperbewußtseins und den Input für Aktionsprogramme darstellt.

Die Neuromatrix wird als vergleichsweise statische Repräsentation menschlicher Selbstwahrnehmung gedacht, kann durch langandauernde oder extreme Ereignisse bzw. Erfahrungen jedoch nachhaltig moduliert werden.

2.2.3 Schmerzerleben und -ausdruck im höheren Lebensalter

Ältere Menschen leiden häufiger an Schmerzen als jüngere Personen (Gibson & Chambers, 2004). Insbesondere die Häufigkeit chronischer Schmerzzustände nimmt mit dem Alter zu, was sicherlich zum Teil auf eine im höheren Lebensalter gesteigerte Multimorbidität zurückgeführt werden kann (Gagliese & Melzack, 1997; Galicia-Castillo & McElha-

ney, 2003; Helme & Gibson, 2001). In einer kanadischen Studie von Proctor und Hirdes (2001) litten insgesamt annähernd 50 Prozent der untersuchten 3195 Pflegeheimbewohner an Schmerzen, die Hälfte davon sogar täglich.

Zur Klärung der Frage, ob die im Alter gesteigerte Schmerzprävalenz neben der Multimorbidität auch durch spezifische alterskorrelierte Veränderungen in der Wahrnehmung und dem Ausdruck von Schmerz bedingt sein könnten, wurden mittlerweile eine Vielzahl experimenteller Studien durchgeführt (für einen Überblick s. Kunz, 2006; Kunz & Lautenbacher, 2005).

Eine Reihe von Untersuchungen zur alterskorrelierten Veränderung in der Schmerzhemmung deuten darauf hin, dass die endogene Schmerzinhibition mit dem Alter nachlässt (Edwards & Fillingim, 2001; Lautenbacher et al., 2005; Larivière, Goffaux, Marchand & Julien, 2007). Dabei konnten Edwards und Kollegen (2001) nachweisen, dass der protektive Effekt einer Blutdrucksteigerung im Alter signifikant reduziert war. Alterskorrelierte Veränderungen konnten auch für die endogene wechselseitige Hemmung mehrerer Schmerzempfindungen (z.B. die gleichzeitige Applikation von elektrischen Reizen und Kälteschmerz an unterschiedlichen Körperteilen, auch als Konterirritation oder Diffuse noxious inhibitory controls (DINCs) bezeichnet, nachgewiesen werden (Edwards, Ness & Fillingim, 2004; Larivière et al., 2007).

Die Schmerzschwelle, also diejenige Intensität oder Dauer eines Reizes, ab der dieser als schmerzhaft erlebt bzw. beschrieben wird, erscheint bei älteren Menschen in Abhängigkeit von der gewählten Schmerzinduktionsmethode herabgesetzt (mechanische Reize), unverändert (elektrische Reize) oder erhöht (thermische Reize). Dabei werden durch elektrische und thermische Reize vor allem Nozizeptoren der oberen Gewebeschichten und der Haut erregt, während mechanische Reize stärker auch Nozizeptoren der tieferliegenden Gewebestrukturen (Muskeln und Sehnen) reizen (Lautenbacher et al., 2005). Diese tiefliegenden Schmerzrezeptoren unterliegen ihrerseits einem deutlich stärkeren Einfluss endogener Schmerzinhibition als weniger tiefliegende Nozizeptoren (Mense, 1993a,b). Die herabgesetzte mechanische Schmerzschwelle könnte somit durch die altersassoziierte Abnahme endogener Schmerzhemmung erklärt werden (Larivière et al., 2007).

Die Toleranzschwelle, also diejenige Intensität oder Dauer eines Reizes, ab der dieser im experimentellen Setting nicht mehr toleriert wird, wurde in der überwiegenden Zahl der Studien als im Alter deutlich herabgesetzt berichtet (Collins & Stone, 1966a,b; Edwards & Fillingim, 2001; Pickering et al., 2002; Walsh et al., 1989; Woodrow et al., 1972). Auch für die über verschiedene Methoden der Schmerzinduktion hinweg vergleichsweise einheitlich gefundene altersassoziierte Absenkung der Toleranzschwelle werden alterskorrelierte Defizite in der endogenen Schmerzhemmung verantwortlich gemacht.

2.2.4 Veränderungen des Schmerzerlebens und -ausdrucks bei Demenz

Im folgenden sollen zunächst sowohl die verfügbaren Befunde zur klinischen Schmerzeinschätzung demenzkranker Menschen dargestellt werden, als auch ein Überblick die Erkenntnisse der experimentellen Schmerzforschung in dieser Population gegeben werden.

Im Anschluss werden die wesentlichen Befunde der neuroexperimentellen und neuropathologischen Schmerzforschung beschrieben.

2.2.4.1 Klinische und experimentelle Befunde

Im Vergleich zu kognitiv gesunden altersgleichen Personen klagen an einer Demenz erkrankte Menschen signifikant seltener über Schmerzen (Cook et al., 1999; Feldt, 2000; Huffman et al., 2000; Mäntyselkä et al., 2004; Marzinski, 1991; Proctor & Hirdes, 2001; Shega et al., 2004). Die Prävalenz selbstberichteter Schmerzen scheint dabei mit stärkerer kognitiver Beeinträchtigung abzunehmen (Parmelee et al., 1997; Shega et al., 2004). Eine Reihe von klinischen Studien wies darüberhinaus darauf hin, dass demenzkranke Menschen im Vergleich zu kognitiv gesunden Personen gleichen Alters deutlich weniger Schmerzmedikation erhalten (Closs, Barr & Briggs, 2004; Horgas & Tsai, 1998; Kaasalainen et al., 1998; Mäntyselkä et al., 2004, Scherder, 2000; Scherder & Bouma, 1997; Wolf-Klein et al., 1988).

Eine naheliegende Erklärung dieser klinischen Befunde wäre ein reduzierter Schmerzbericht aufgrund der häufig mit einer Demenz einhergehenden Beeinträchtigung kommunikativer Fähigkeiten bei eventuell unveränderter Schmerzempfindung. Eine alternative – und in ihren Implikationen für die Versorgung nahezu entgegengesetzte – Interpretation geht von einem durch die neurodegenerativen Demenzprozesse tatsächlich substanziell reduzierten Schmerzerleben aus. Der jeweilige empirische Gültigkeitsbereich beider Annahmen kann freilich nur durch experimentelle Studien abgeschätzt werden, in denen eine Kontrolle und systematische Variation des Schmerzreizes und eine multidimensionale Erfassung verbaler und nicht-verbaler Schmerzreaktionen möglich ist. Eine nach dem verwendeten Outcomeparameter differenzierte Übersicht der bislang veröffentlichten Arbeiten zur Schmerzwahrnehmung von Menschen mit einer Demenz bei Alzheimer-Erkrankung leistet Kunz (2006).

Die Schmerzschwelle scheint bei Menschen mit einer Alzheimer-Demenz gegenüber altersgleichen Kontrollpersonen nicht systematisch verändert zu sein (Benedetti et al., 1999; Cornu, 1975; Gibson et al., 2001; Rainero et al., 2000), wohingegen eine deutlich erhöhte Toleranzschwelle der Demenzpatienten belegt werden konnte (Benedetti et al., 1999). Damit darf angenommen werden, dass Demenzkranke bei gleichen Fähigkeiten zur Schmerzdiskrimination ein breiteres Intervall empfundener Schmerzen tolerieren. Geht man weiterhin davon aus, dass die Toleranzschwelle die Einforderung bzw. Nutzung von Analgetika wesentlich stärker bestimmt als die Schmerzschwelle, ergibt sich ein auch hinsichtlich der Unterschiede in der Schmerzmedikation zwischen demenzkranken und nicht kognitiv beeinträchtigten Menschen kongruentes Bild.

Die Hinweise auf ein durch die Demenzerkrankung differenziell verändertes Erleben von Schmerzen in Bereichen geringer und hoher Intensitäten werden auch durch die Schmerzbeurteilung mit Ratingskalen gestützt. Schmerzreize geringer bis mittlerer Intensität werden dabei von Alzheimerpatienten und kognitiv unbeeinträchtigten Personen als vergleichbar intensiv erlebt eingeschätzt (Gibson et al., 2001; Porter et al., 1996; Rainero

et al., 2000). Bei Verwendung stärkerer thermischer Reize schätzten die auskunftsfähigen Alzheimerpatienten diese jedoch als deutlich weniger schmerzhaft ein als Kontrollpersonen (Rainero et al., 2000).

Einschränkend muss angemerkt werden, dass alle zuvor genannten subjektiven bzw. verbal gestützten Outcomeparameter in verschiedenem Maße die Introspektions- und Kommunikationsfähigkeit der demenzkranken Menschen voraussetzen. Dabei sind die kognitiven Anforderungen bei diskriminativen Aufgaben wie der Beurteilung als nicht-schmerzhaft oder schmerzhaft (Schmerzschwelle) bzw. erträglich oder nicht mehr zu ertragen (Toleranzschwelle) wahrscheinlich etwas geringer als bei verbalen Ratingskalen mit mehreren Intensitätskategorien oder abstrakteren Verfahren wie beispielsweise der visuellen Analogskala (VAS) oder der Faces Pain Skala (FPS) anzunehmen. In Kapitel 3 werden verschiedene Verfahren der Schmerzbeurteilung und ihre Eignung für Menschen mit Demenz eingehender diskutiert. Besonderes Augenmerk wird dabei auf Inventare für die Fremdbeobachtung schmerzbezogenen Ausdrucksverhaltens gelegt werden.

In mehreren Studien, bei denen die mimische Schmerzreaktion als verhaltensbezogenes Maß erlebter Schmerzen beobachtet wurde, konnte eine bei Demenzkranken gesteigerte mimische Schmerzreaktion nachgewiesen werden (Hadjistavropolous et al., 1997; 2000; Kunz, 2006; Porter et al., 1996). Obgleich Porter und Kollegen davon ausgehen, dass die Steigerung im schmerzbezogenen Ausdrucksverhalten demenzkranker Personen die Folge einer global exzessiveren Mimikreaktion darstellt und darum nicht als Hinweis auf ein bei Demenz gesteigertes Schmerzerleben gewertet werden kann, weisen die Ergebnisse der Studie von Kunz (2006) darauf hin, dass während der Schmerzinduktion vermehrt mit schmerzspezifischem Ausdrucksverhalten reagiert wurde, weitere nichtschmerzbezogene Mimikreaktionen dagegen nicht häufiger auftraten.

Diejenigen Studien, die eine Steigerung des systolischen Blutdrucks und der Herzrate als vegetative Schmerzreaktionen untersuchten, konnten eine zum Teil deutlich reduzierte vegetative Schmerzreaktion demenzkranker Menschen nachweisen (Benedetti et al., 2004; Kunz, 2006; Porter et al., 1996; Rainero et al., 2000).

2.2.4.2 Degeneration schmerzrelevanter Hirnstrukturen

Eine naheliegende Erklärung der klinisch und experimentell herausgearbeiteten demenzbezogenen Veränderungen im Schmerzerleben ist die fortschreitende Degenerierung schmerzrelevanter zerebraler Strukturen. Die bisherigen neuroanatomischen Befunde legen nahe, dass Demenzen verschiedener Ätiologien das schmerzverarbeitende System auf unterschiedliche Weise betreffen können (Scherder et al., 2003). In Abbildung 1 wurden die bei der Schmerzverarbeitung beteiligten medialen und lateralen Strukturen des Gehirns und ihre Beziehung zueinander bereits dargestellt. Einen Überblick über nachgewiesene Schädigungen dieser schmerzrelevanten zerebralen Arealen bei verschiedenen Grunderkrankungen bzw. Demenzätiologien leistet Tabelle 1.

Im Kontext der Alzheimer-Erkrankung erscheinen nahezu alle (sub)kortikalen Strukturen des medialen Schmerzsystems deutlich beeinträchtigt, während die dem lateralen

Tabelle 1: Beeinträchtigungen des medialen und lateralen Schmerzsystems bei verschiedenen Demenzätiologien (nach Scherder, Sergeant & Swaab, 2003).

		Demenzätiologie		
		Alzheimer	Vaskulär	Frontotemporal
Mediales Schmerzsystem				
Locus coeruleus (LC)		+		–
Parabrachialkern (PBN)		+		
Periaquäduktales Grau (PAG)		+		
Thalamus	Medialer Thalamuskern (MTN)	+		
	Intralaminare Kerne (ILN)	+		
Insula		+		+
Anteriorer zingularer Kortex (ACC)		+		++
Hippocampus		++		+
Amygdala		++		+
Hypothalamus	Paraventrikularkern (PVN)		+Dis	+
	Tubermamillarkern (TMN)	+	+Dis	+
Präfrontaler Kortex		+	+Dis	++
Laterales Schmerzsystem				
Thalamus	Laterale Thalamuskerne (LTN)	+		
Somatosensorischer Kortex I (S1)		–		
Somatosensorischer Kortex II (S2)		+		
Marklagerveränderungen		+	++	+

– = nicht betroffen, + = betroffen, ++ = im Vergleich zu anderen Demenzformen stark betroffen; +Dis = Entkopplung (durch Marklagerveränderungen).

Schmerzsystem zugerechneten Hirnareale bis in späte Erkrankungsstadien hinein noch vergleichsweise unbeeinträchtigt erscheinen. Aufgrund dieses spezifischen Beeinträchtigungsmuster wären bei Alzheimerpatienten im Wesentlichen Veränderungen der motivational-affektiven, kogntiv-evaluativen und autonomen Schmerzkomponente zu erwarten, während die sensorisch-diskriminativen Anteile des Schmerzerlebens weitgehend erhalten bleiben sollten. Die zuvor beschriebenen klinischen und experimentellen Befunde scheinen sich schlüssig aus den genannten differenziellen Hirnschädigungen erklären zu lassen. Die erhöhte Toleranzschwelle der Alzheimerpatienten und die reduzierten vegetativen Schmerzreaktionen können als veränderte affektiv-motivationale Qualität des Schmerzerlebens begriffen werden, wohingegen die unveränderte Schmerzschwelle als Ausdruck erhaltener sensorisch-diskriminativer Funktionen angesehen werden kann.

Im Vergleich zur Alzheimer-Erkrankung ist die empirische Grundlage zur Bestimmung spezifischer Veränderungen im Schmerzerleben von Menschen mit andersweitig bedingten Demenzen gegenwärtig noch recht dünn. Der Großteil der klinischen und experimentellen Studien zum Schmerzerleben bei Demenz bezog ausschließlich (wahrscheinliche) Alzheimerpatienten ein, und auch die neuropathologischen Veränderungen der durch die Erkrankung betroffenen schmerzrelevanten Hirnareale wurden bei Personen mit Alzheimer am vollständigsten untersucht.

Das wesentlichste neuropathologische Korrelat vaskulärer Demenzen sind infarktiös bedingte Marklagerveränderungen (auch white matter lesions WML). Die Deafferenzierung bzw. Unterbrechung der Verbindung verschiedener schmerzverarbeitender Areale können zu schlaganfallbedingten zentralen Schmerzen oder Kopfschmerzen als Spätfolge von Marklagerläsionen in tiefen Hirnschichten (ohne manifesten strategischen Schlaganfall) führen. Insgesamt lässt die Neuropathologie der vaskulären Demenz gänzlich andere demenzbedingte Veränderungen im Schmerzerleben, nämlich eine Steigerung des Schmerzerlebens, erwarten (Scherder et al., 2003; Scherder et al., 2005). Die klinischen Befunde und experimentellen Ergebnisse zum Schmerzerleben von Menschen mit vaskulärer Demenz sind jedoch uneinheitlich. Während Scherder und Kollegen (2003) ein im Vergleich zu nicht-dementen Älteren und Personen mit schmerzhaften chronischen Krankheiten höheres berichtetes Schmerzniveau bei Patienten mit einer (möglichen) vaskulären Demenz beschreiben, konnten in den experimentellen Studien von Kunz und Kollegen (2005; 2007) keine eindeutigen Hinweise auf eine im Vergleich zu Alzheimerpatienten gesteigerte Schmerzempfindlichkeit gefunden werden.

Auch mit Blick auf die frontotemporale Demenz besteht sowohl hinsichtlich der klinischen und experimentellen Untersuchung von Schmerzbelastung und -erleben als auch mit Blick auf die neuropathologischen Veränderungen in schmerzrelevanten Hirnstrukturen gegenwärtig noch enormer Forschungsbedarf. Aufgrund der teilweisen Übereinstimmung im Muster betroffener schmerzverarbeitender Areale, vor allem des medialen Schmerzsystems, wird angenommen, dass Personen mit frontotemporaler Demenz ähnliche Veränderungen in motivational-affektiven, kognitiv-evaluativen, autonom-neuroendokrinologischen und auf das Schmerzgedächtnis bezogenen Qualitäten des Schmerzerlebens aufweisen wie Personen mit einer Demenz vom Alzheimer-Typ (Scherder et al., 2003). In einer Untersuchung von Bathgate und Kollegen (2001) schienen für Personen mit frontotemporaler Demenz insbesondere Defizite in der Antizipation gefährdender Situationen (kognitiv-evaluativ), jedoch auch in der unmittelbaren Reaktion auf Schmerzreize (sensorisch-diskriminativ) auf.

3 Herausforderungen der Schmerzmessung bei Demenz

Die zuvor berichteten empirischen Arbeiten zur Schmerzverarbeitung und Schmerzbelastung hochaltriger und demenziell erkrankter Menschen verwendeten eine Reihe recht unterschiedlicher Schmerzmaße. Neben der mitunter auch als *subjektivem* Schmerzmaß bezeichneten Selbstauskunft zur Schmerz- und Toleranzschwelle oder Einschätzungen der Schmerzintensität auf Ratingskalen wurden dabei auch weniger stark verbal gestützte Marker erlebten Schmerzes erfasst. Zu Letzteren zählen insbesondere von Außen beobachtbare oder apparativ ableitbare schmerzevozierte mimische und physiologische (vegetative, nozifensiv-reflexhafte und zentralnervöse) Reaktionen auf schmerzhafte Reize.

Im folgenden Kapitel werden verschiedene Schmerzmarker und die Verfahren für ihre Erfassung eingehend dargestellt. Besonderes Augenmerk wird dabei auf die Anwend-

barkeit im Kontext der stationären Versorgung von Menschen mit einer Demenz gelegt. Neben der Selbstauskunft bei hinreichend gut erhaltenen kommunikativen Fähigkeiten werden darum vor allem Verfahren der Fremdauskunft und Verhaltensbeobachtung eingehend diskutiert. Auf eine detaillierte Darstellung neurophysiologischer Schmerzmarker, die aufgrund ihrer technisch-apparativen und fachlichen Voraussetzungen im Pflegealltag gewöhnlich nicht erhoben werden können, soll im Rahmen dieser Arbeit dagegen verzichtet werden.

Verschiedene Schmerzmaße geben dabei Hinweise auf z.T. unterschiedliche Qualitäten des Schmerzerlebens. So erscheinen beispielsweise Angaben zur Schmerzschwelle stärker die sensorisch-diskriminative Qualität des Schmerzes auszudrücken als beispielsweise die Toleranzschwelle, die deutlicher auch durch motivational-affektive Komponenten des Schmerzerlebens bestimmt wird. Auch vegetative Schmerzreaktionen wie Herzfrequenz, Blutdruck und Hautleitfähigkeit werden als Ausdruck der affektiven Schmerzkomponente interpretiert. Zeitbezogene Vergleiche von Schmerzintensitäten (z.B. im Rahmen von Schmerztagebüchern) oder Inter-Modalitäts-Vergleiche (wie bei der Visuellen Analogskala) adressieren hingegen stärker auch kognitiv-evaluative oder gedächtnisbezogene Qualitäten des Schmerzes.

Wenngleich für das Verständnis der zugrunde liegenden Veränderungsprozesse im Schmerzerleben bei Demenz experimentelle Forschungsansätze unabdingbar sind, kommt diesen im Kontext eines alltagspraktischen Schmerzassessments in der stationären Pflege zunächst eine nur nachgeordnete Bedeutung zu. Dennoch wird die Idee der systematischen (i.S. einer mehr oder weniger stark standardisierten) Variation von Schmerzreizen (z.B. über Bewegung oder Aktivierung) von mehreren für die klinische und pflegerische Schmerzerfassung vorgeschlagenen Assessmentverfahren aufgegriffen (Feldt, 2000; Husebo et al., 2007; Morello et al., 2007; Warden, Hurley & Volicer, 2003).

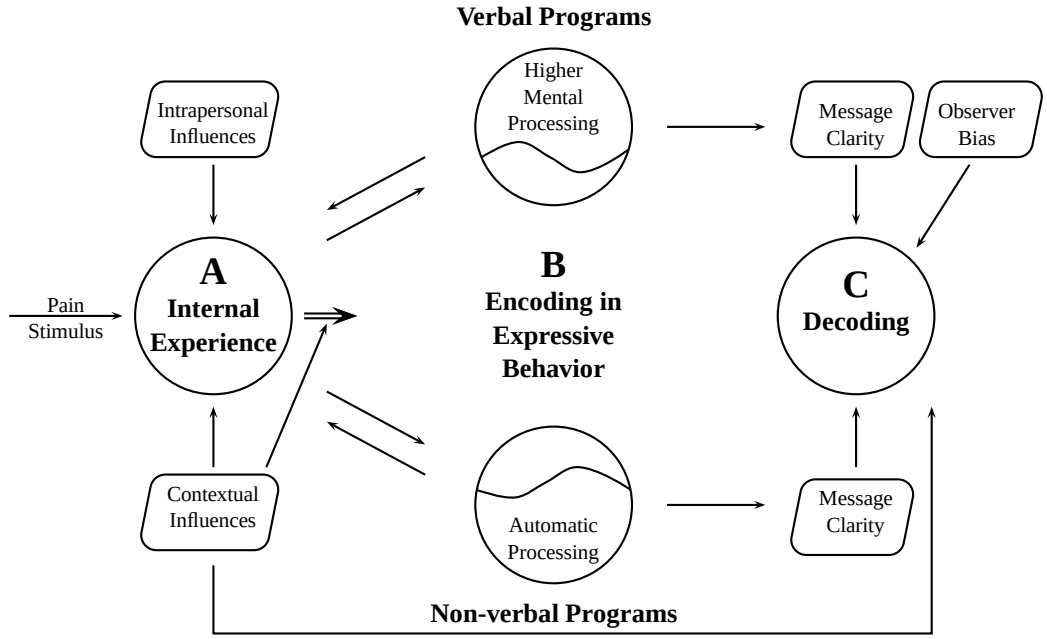
3.1 Schmerzkommunikation

Die in Fokus dieser Arbeit stehenden Verhaltensinventare zur Schmerzbeobachtung wurden mitunter explizit für Populationen entwickelt, die in ihren kommunikativen Möglichkeiten substanziell beeinträchtigt sind. Dabei besteht die Gefahr, Kommunikationsfähigkeit ausschließlich mit verbaler Auskunftsfähigkeit gleichzusetzen und die grundsätzliche soziale Rahmung des Schmerzerlebens sowie den kommunikativen Gehalt non-verbalen Verhaltensausdrucks zu übersehen.

In einer Reihe konzeptioneller Arbeiten sprachen sich Hadjistavropoulos, Craig und Kollegen dafür aus, die Erfassung von Schmerzen stärker als bisher als einen kommunikativen Akt zu verstehen, der durch Eigenschaften der Schmerzen erlebenden und kommunizierenden Person selbst, aber auch durch Kontextfaktoren und nicht zuletzt durch Eigenschaften des Fragenden oder Beobachtenden bestimmt wird (Hadjistavropoulos & Craig, 2002; Hadjistavropoulos, Craig & Fuchs-Lacelle, 2004; Hadjistavropoulos, von Baeyer & Craig, 2001). In Abbildung 3 sind die Prozesse des Schmerzerlebens (A), der Enkodierung in Verhalten (B) und der Entschlüsselung durch einen Kommunikationspartner (C)

schematisch dargestellt.

Abbildung 3: Das (Sozial-)kommunikative Schmerzmodell (Hadjistavropoulos & Craig, 2002, p.555)



Die Autoren erweitern das ursprünglich von Rosenthal (1982) vorgeschlagene und im Kontext der Schmerzmessung durch Prkachin und Craig (1995) auf den mimischen Schmerzausdruck bezogene A→B→C Kommunikationsmodell um verbale (Selbstauskunft) und nicht-mimische behaviorale Ausdrucksweisen (Hadjistavropoulos & Craig, 2002).

Das Enkodieren schmerzbezogenen Erlebens in verbale und nicht-verbale Signale erscheint dabei in unterschiedlichem Maße durch automatisierte (reflexhafte) und höhere geistige Verarbeitungsprozesse (Antizipation von Folgen, soziale Normen etc.) bestimmt. Grundsätzlich wird dabei angenommen, dass der verbale Selbstbericht erlebter Schmerzen als bewusste mentale Leistung höhere Gefahren einer Verfälschung bzw. Täuschung durch Einstellungen und Motive des Senders birgt als die weniger stark bewusst kontrollierbare motorische Verhaltensreaktion auf (unmittelbar erlebte) Schmerzen. Das schließt im Umkehrschluss jedoch nicht aus, dass auch nicht-verbale Verhaltensweisen bewusst zur (irreführenden) Kommunikation erlebter Schmerzen genutzt werden können.

Wird der Schmerzbericht absichtsvoll beispielsweise auch durch Gestik, Mimik oder Körperhaltung gestaltet, muss aber die bisher eng an den verbalen Ausdruck gekoppelte Definition von Selbstauskunft um willentlichen non-verbalen Schmerzausdruck erweitert werden.

> [...] under certain circumstances, nonverbal behaviours (e.g., grimaces, body postures) are purposefully used to communicate pain and distress. When the

> term self-report is used to refer to nonverbal behaviour, the deliberate and conscious nature of the communicative act, as implied by the word 'report', needs to be appreciated. (Hadjistavropoulos & Craig, 2002, p. 553, Hervorhebungen im Original)

Tatsächlich kann auch der nonverbale Schmerzausdruck durch höhere mentale Prozesse zumindest teilweise vorgetäuscht, abgewandelt, unterdrückt oder übertrieben werden (Craig, Hill & McMurty, 1999), so dass man also auch im nonverbalen Bereich nicht von einer strikten Kongruenz zwischen Schmerzerleben und -ausdruck ausgehen kann.

> There are no manifestations of pain – verbal, nonverbal or physiological – that are exclusive markers of the experience and not subject to observer doubts about credibility. (Hadjistavropoulos & Craig, 2002, p. 552)

Aus der Perspektive des dargestellten Kommunikationsmodells können auch schmerzbezogene Verhaltensweisen, die primär keinem kommunikativen Zweck dienen, sondern beispielsweise auf die Beseitigung oder Modulierung schmerzhafter Reize (nozizeptiv bedingte Flexorreflexe, Massage etc.) ausgerichtet sind, einem Beobachter wichtige Hinweise auf unmittelbar erlebte Schmerzen geben.

In ähnlicher Weise können aufmerksame Beobachter mit einer guten Kenntnis der schmerzleidenden Person Veränderungen in deren Gewohnheiten oder einen Rückzug aus sozialen Beziehungen als subtile Schmerzmarker erfassen, obwohl diese langfristigen Verhaltensänderungen gewöhnlich nicht (primär) auf die Kommunikation von Schmerz oder die Beseitigung unmittelbarer Schmerzreize hin ausgerichtet sind.

Hadjistavropoulos, Craig und Fuchs-Lacelle (2004) weisen in einer Folgearbeit nochmals stärker auf die Bedeutung sozialer Determinanten für die Schmerzkommunikation insbesondere im Kontext der pflegerischen Versorgung schmerzbelasteter Menschen hin. Mit Blick auf den in Abbildung 3 beschriebenen Beobachterbias werden dabei Prozesse der Aufmerksamkeitssteuerung, geistige Dispositionen und individuelle Werthaltungen oder die Beziehung des Pflegenden zur schmerzbelasteten Person als soziale Rahmenbedingungen der Schmerzeinschätzung diskutiert. Da sowohl der selbstinitiierte Schmerzbericht als auch das an den Bewohner herangetragene pflegerische Schmerzassessment dem Zweck dienen, schmerzhafte Zustände nicht nur zu offenbaren sondern auch zu vermindern, wird das Modell sozialer Schmerzkommunikation durch die Autoren um eine Komponente pflegerischer Folgehandlungen ergänzt. Medizinische oder psychosoziale Interventionen können dabei ebenso mögliche Resultate des Schmerzmanagements sein wie Änderungen in der Sensibilität der Pflegenden für die Wahrnehmung von Schmerzen bei Anderen. Andererseits könnten Verfahren der Schmerzmessung, welche die Aufmerksamkeit des Befragten auf sein inneres Schmerzerleben richten, die Schmerzwahrnehmung unter Umständen intensivieren oder durch die Unterbrechung von Coping- und Selbstmanagementmechanismen zu einem Anstieg der Schmerzbelastung führen (Carr & Mann, 2002; Leventhal, Leventhal, Shacham & Easterling, 1989; Webb, Campbell, Schwartz & Sechrest, 1966).

3.2 Verbal gestützte Verfahren

Verfahren, bei denen die Betroffenen direkt Auskunft über ihr subjektives Schmerzerleben geben, sind in den meisten Fällen sprachgebunden. Ein intaktes Sprachverständnis erscheint dabei sowohl für das Verständnis der gestellten Fragen oder Testanweisungen, als auch für die Differenzierung und Auswahl angemessener Antwortoptionen nötig. Werden im Rahmen der Schmerzmessung schriftliche Anweisungen oder Antwortformate vorgegeben, müssen daneben die Lesbarkeit des Arbeitsmaterials und die Lesefähigkeit der Probanden sichergestellt sein. Prinzipiell gelten diese Voraussetzungen selbstverständlich auch für Schmerzassessments per Fremdbeurteilung durch Proxys, die im pflegerischen Setting nicht immer über hinreichende Sprachkenntnisse verfügen. Im Folgenden werden zunächst Verfahren der sprachgestützten Schmerzmessung beschrieben, und diese anschließend mit Blick auf ihre Anwendbarkeit in der Population demenzkranker Menschen beurteilt.

3.2.1 Stimulusabhängige Messung

Im Rahmen experimenteller Verfahren mit kontrollierter Induktion von Schmerzreizen mit verschiedener Intensität, Ausbreitung oder Dauer werden häufig die Schmerz- und Toleranzschwelle als individuelle Schmerzmarker per Selbstauskunft erfasst. Dabei soll die betroffene Person angeben, ab wann sie einen applizierten Reiz als schmerzhaft (Schmerzschwelle) empfindet, und ab wann der applizierte Reiz nicht mehr toleriert werden kann (Toleranzschwelle). Beide Marken können dabei in den verwendeten Reizstärke- (z.B. °C, kg/cm^2, mA) oder Zeiteinheiten gemessen werden, weshalb hier von einer stimulusabhängigen Messung gesprochen wird (Lautenbacher, 2004). Werden die bei Schmerzreizen verschiedener Intensität empfundenen Schmerzen auf feiner gegliederten metrischen Ratingskalen eingeschätzt, kann der psychophysische Zusammenhang zwischen Reizstärke und -empfindung als gewöhnlich nicht-lineare Funktion abgeschätzt werden (sog. power function; s. Price et al., 1983; Price, Riley & Wade, 2001).

Die häufigsten Verfahren der experimentellen Schmerzinduktion sind dabei die Applizierung elektrischer, thermischer oder mechanischer Reize. Seltener werden Schmerzen durch eine künstlich herbeigeführte Ischämie eines Körperteiles induziert (für einen Überblick siehe Kunz, 2006).

Selbst wenn vom Probanden lediglich eine dichotome Entscheidung getroffen werden soll, kann das Verständnis abstrakter Begriffe wie ‚Toleranz' für demenzkranke Menschen problematisch sein (Kunz, 2006). Daneben ist fraglich, welche Bedeutung das Konzept der Schmerz- und Toleranzschwelle im Bereich des klinischen Schmerzmanagements überhaupt besitzt. So erfordert beispielsweise die Bestimmung beider qualitativer Übergänge eine besondere Aufmerksamkeit hinsichtlich des potenziellen Schmerzreizes, womit Prozessen der Sensibilisierung und Erwartungshaltung eine besondere Bedeutung zukommen könnte. Die Schmerzschwelle mag für das alltägliche klinische Schmerzmanagement eine geringere Rolle spielen als die Toleranzschwelle, die als ein im Wesentlichen affektiver

Marker des Schmerzerlebens den klinischen Schmerzbericht und die Medikamentennachfrage wesentlich mitbestimmen sollte. Einschränkend sollte jedoch berücksichtigt werden, dass das experimentelle Setting als solches eine außergewöhnliche soziale Situation darstellt und unerwünschter Weise bestimmte Erwartungshaltungen und Leistungsmotivationen fördern könnte.

3.2.2 Ratingskalen

Bei Ratingverfahren wird das eigene oder bei Anderen vermutete Schmerzerleben mit verbal beschriebenen Kategorienlabels oder assoziativ mit einem – aus anderen Erfahrungszusammenhängen bereits bekanntem – Merkmalskontinuum (z.B. Temperatur) in Beziehung gesetzt. Einen Überblick über die wichtigsten Grundformen von Skalen zur (Selbst-)Beurteilung erlebter Schmerzen gibt Abbildung 4.

3.2.2.1 Kategorialskalen

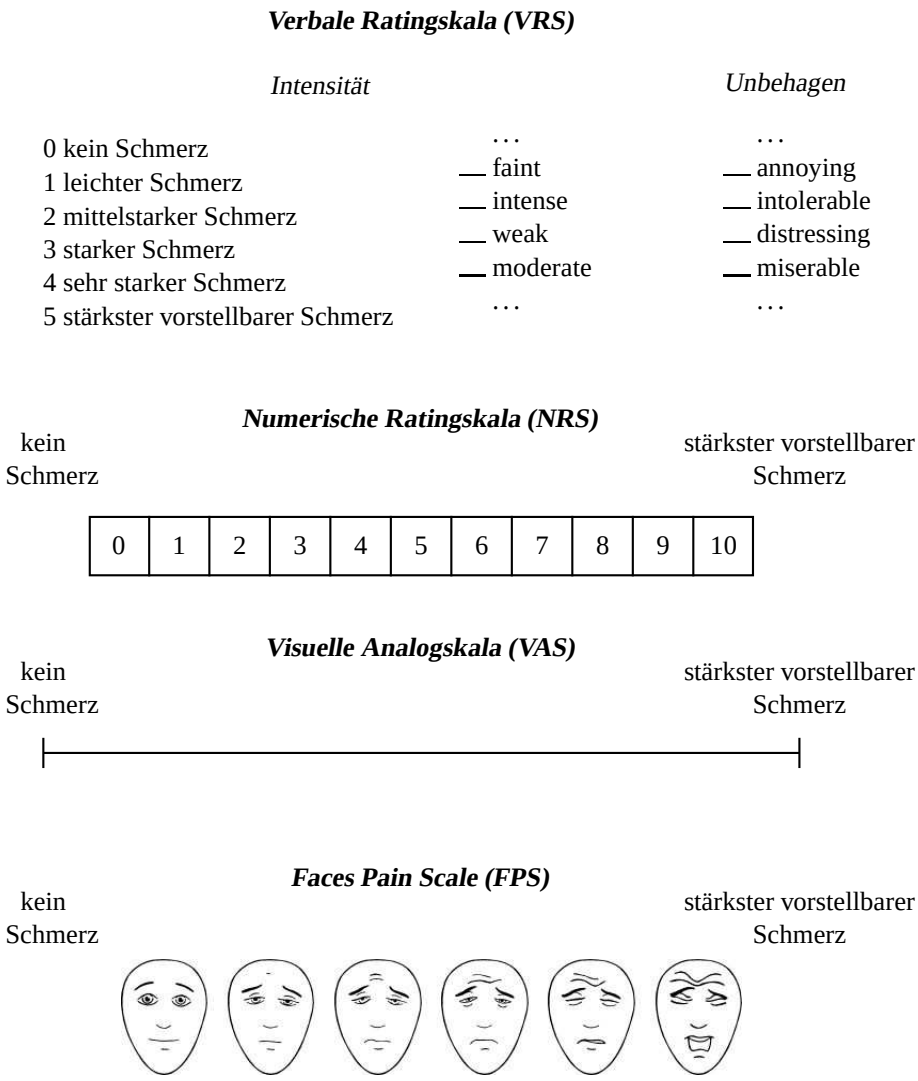
Kategorialskalen unterteilen das gedachte Merkmalskontinuum empfundenen Schmerzes in verschiedene Abschnitte, die zumeist sprachlich (Verbale Ratingskala, VRS bzw. Verbal Descriptor Scale, VDS) oder durch Zahlen (Numerische Ratingskala, NRS) näher bezeichnet sind. Die sprachlichen Kategorienlabels selbst stellen gewöhnlich Steigerungsformen dar (z.B. leicht-mäßig-stark), die empfundene qualitative Übergänge markieren. Die Pole der numerischen Ratingskalen sind häufig ebenfalls mit sprachlichen Ankern (z.B. 0=„kein Schmerz“, 10=„stärkster vorstellbarer Schmerz“) versehen (s. Abb. 4).

Kategorialskalen zur Schmerzmessung nehmen zumindest ein ordinales Skalenniveau der erfassten Messwerte an, und werden gewöhnlich interpretiert als vergröberte Abbildung der eigentlich kontinuierlichen latenten Schmerzintensität. Insofern können auch Dichotomisierungen des Schmerzkontinuums in „nicht vorhanden“ und „vorhanden“ oder von Reizstärken in „(nicht) schmerzhaft“ oder „(nicht mehr) tolerierbar“ als kategoriale Antwortformate begriffen werden.

In der Regel werden zwischen drei und 10 Kategorien differenziert und diesen aufsteigende Zahlenwerte für die statistische Verrechnung zugeordnet. Je stärker das gedachte Merkmalskontinuum gegliedert wird, d.h. je mehr Antwortalternativen bzw. Kategorienstufen erfasst werden, desto eher weisen die Messwerte wünschenswerte metrische Eigenschaften auf. Für verbale Ratingskalen kann gewöhnlich keine Äquidistanz zwischen den (wenigen) beschriebenen Intensitätsstufen angenommen werden, weshalb für diese Maße Verfahren für Ordinaldaten angemessen erscheinen. In der gegenwärtigen Praxis des Schmerzassessments werden mitunter doch bereits Schmerzmaße mit vergleichsweise wenigen Ausprägungen als intervallskaliert angenommen oder zumindest so behandelt. Andererseits weisen Untersuchungen zur Verwendung der NRS und VRS darauf hin, dass mehr vorgegebene Antwortkategorien nicht automatisch auch zu einer differenzierteren Einschätzung führen. Jensen und Kollegen (1994) wiesen nach, dass ungefähr drei von vier Befragten eine Skala mit 101 Punktwerten (Range 0-100) lediglich in Zehnerschrit-

ten, und damit analog zu einer 11-stufigen Skala bearbeiten (Jensen, Turner & Romano, 1994).

Abbildung 4: Verschiedene sprachgestützte Ratingverfahren zur Schmerzmessung.



Rosier und Kollegen gaben bei wiederholten Schmerzassessments jeweils eine Liste mit 15 verbalen Deskriptoren für die Qualitäten Intensität und Unbehagen (unpleasantness) vor. Die durch die Probanden erstellten Rangreihen der Deskriptoren erwiesen sich über mehrere Sitzungen hinweg als nicht vollständig konstant. Außerdem wurden insbesondere mit Blick auf die Einschätzung der Unbehaglichkeit der induzierten Schmerzreize nur wenige der verfügbaren Deskriptoren genutzt (Rosier, Iadarola & Coghill, 2002). Insbesondere bei Verwendung vieler verbaler Kategorienlabels besteht die Gefahr, dass den Probanden nicht alle Begriffe geläufig sind und einzelne Bereiche des Schmerzkontinuums darum de facto weniger detailliert unterschieden werden können als andere. Als ein

möglicher Nachteil der VRS wird daneben ihre geringe Änderungssensitivität bei wenigen Kategorienstufen angeführt (DNQP, 2004).

Soll dagegen angenommen werden, dass das zu messende Schmerzmerkmal nicht kontinuierlich sondern diskret verteilt ist, ist auch die Abbildung in die beschriebenen Kategorien (z.B. „Schmerz vorhanden“ oder „Schmerz nicht vorhanden“) nicht als ordinal oder metrisch höherwertig, sondern als nominal anzusehen. Das Augenmerk der Schmerzforschung sollte sich dann stärker auf Verfahren zur Differenzierung der postulierten qualitativ unterschiedlichen Gruppen schmerz(un)belasteter Menschen richten (Latent Class Analyse). Gegenwärtig dominiert jedoch die Vorstellung eines kontinuierlichen latenten Schmerzkontinuums (Williamson & Hoggart, 2005). Es wird der zukünftigen Schmerzforschung vorbehalten bleiben, zu entscheiden, ob und wie diese doch noch stark mechanistisch geprägte Konzentration auf den Aspekt der Schmerzintensität zugunsten einer qualitative Aspekte des Schmerzerlebens stärker betonenden Perspektive überwunden werden sollte.

3.2.2.2 Direkte Skalierung

Bei der direkten Skalierung sollen proportional zur empfundenen Merkmalsstärke Zahlenwerte vergeben werden (Größenschätzung), oder Analogieschlüsse auf die Stärke anderer Sinnesmodalitäten, z.B. die Länge einer Linie, die (Rot-)Sättigung eines Farbverlaufes, die Temperatur auf einem Thermometer o.ä. geleistet werden (Inter-Modalitäts-Vergleich).

Im Kontext der Schmerzmessung sind die Visuelle Analogskala (VAS) und die Faces Pain Scale (FPS) die am häufigsten verwendeten Verfahren der direkten Skalierung, und sollen darum hier knapp umrissen werden.

Bei der VAS wird der Proband gebeten, die Intensität seiner Schmerzempfindung auf einer gewöhnlich 100 Millimeter langen horizontalen Linie mit den verbalen Ankern „kein Schmerz“ und „schlimmster vorstellbarer Schmerz“ einzuzeichnen (Abbildung 4). Der Messwert liegt damit in einem Bereich zwischen 0 und 100 Punkten.

Price und Kollegen fanden Hinweise darauf, dass die auf einer VAS eingeschätzten Schmerzintensitäten Verhältnisskalenniveau besitzen. Dazu ermittelten sie zunächst den (nicht-linearen) Zusammenhang zwischen experimentell induzierten thermischen Schmerzreizen und der relativen Länge der VAS-Strecken (sog. power function). Diejenige Temperatur, die durch diese Funktion als verglichen mit der Ausgangstemperatur doppelt so schmerzhaft erlebt vorhergesagt wurde, stimmte recht gut mit demjenigen Punkt in einer induzierten Sequenz ansteigender Temperaturen überein, den die Probanden explizit als doppelt so schmerzhaft wie das Basisniveau identifizierten (Price et al., 1983).

Die Gesichterskala (Faces Pain Scale, FPS, Brieri et al., 1990) misst die Schmerzintensität mit Hilfe von mehreren gezeichneten Gesichtern mit unterschiedlichem Schmerzausdruck (s. Abb. 4). Der Proband wird gebeten, dasjenige Gesicht zu wählen, dass seine erlebten Schmerzen am besten repräsentiert. Bei der von Brieri und Kollegen entwickelten naturalistisch orientierten Skala sind einzelne Aspekte typischer schmerzbezogener

Gesichtsmuskelkontraktionen (wie z.B. der Augenbrauen) berücksichtigt. Daneben wurden mehrere Skalen mit abstrakteren Gesichtern (sog. Smileys) vorgeschlagen, die stärker die affektive Komponente des Schmerzerlebens abbilden sollen (Facial Affective Scale, FAS; McGrath, de Veber & Hearn, 1985; McGrath et al., 1996). Gewöhnlich wird dabei ein lachendes Gesicht als Gegenpol zum schmerzverzerrten, weinenden Gesichtsausdruck maximalen Schmerzes vorgegeben. Dabei bleibt unklar, wie der abgebildete positive Affektbereich theoretisch mit erlebten Schmerzen zusammenhängt, da bereits ein neutraler Gesichtsausdruck Schmerzfreiheit anzeigen könnte. Daneben erscheinen – aufgrund der tatsächlich eher kontinuierlich intensitätssteigernd denn qualitativ diskret angelegten Skala – Differenzierungen verschiedener schmerzbezogener Emotionsqualitäten im negativen Affektbereich (Angst, Trauer/Resignation, Aggression/Wut) nur unzureichend möglich.

3.2.3 Anwendbarkeit sprachgestützter Verfahren

Da es im Verlauf der demenziellen Erkrankung zu einem systematischen Verlust sprach- und gedächtnisbezogener Kompetenzen kommt, muss abgeschätzt werden, wie lange der demenzkranke Mensch selbst mithilfe von Schmerzassessments, die entsprechende Funktionen voraussetzen, verlässliche Informationen zu seinem Schmerzerleben geben kann.

Diese Abschätzung wird mitunter durch die vielen verschiedenartigen Vorgabe- und Antwortformate erschwert. Analogskalen arbeiten über herkömmliche paper-and-pencil-Versionen hinaus häufig mit mechanischen Umsetzungen (Schieber, Regler etc.). Auch die in der gegenwärtigen Praxis eingesetzten *Kategorialskalen*, und hier insbesondere die VRS, unterscheiden sich z.T. deutlich mit Blick auf die Anzahl unterschiedener Schmerzausprägungen oder der erfragten Schmerzqualitäten und damit in ihren sprachlichen und kognitiven Anforderungen an die demenzkranken Informanten. Verbale Ratingskalen mit einer überschaubaren Kategorienanzahl (zwischen 3 und 6) und Deskriptoren aus dem normalen Sprachgebrauch können auch noch von einem substanziellen Anteil demenziell beeinträchtigter Menschen genutzt werden. Krulewitch und Kollegen (2000) untersuchten die Möglichkeit zur Selbstauskunft von 156 kognitiv beeinträchtigten Altenheimbewohnern (mittlerer MMSE=15,7) anhand einer VAS, NRS und VRS. Dabei konnten immerhin zwei Drittel der Bewohner Auskünfte mittels zumindest einer dieser Skalen machen. In der Gruppe schwerbeeinträchtigter Bewohner (mittlerer MMSE-Score $<$11 Punkte) konnten immerhin noch 59 Prozent eine Selbstauskunft geben.

Zu ähnlichen Ergebnissen kommen auch Closs und Kollegen (2004), die fünf Skalen zur Selbstauskunft (VRS, NRS, FPS, CAS und VAS) bei 113 Bewohnern mit unterschiedlichem Demenzschweregrad untersuchten. Dabei konnten über 80 Prozent der leicht bis mittelschwer erkrankten und immerhin noch über 36 Prozent der schwerwiegend beeinträchtigten Menschen Selbstauskünfte mittels VRS geben (Closs, Barr, Briggs, Cash & Seers, 2004).

Wie zuvor bereits allgemein für das höhere Lebensalter berichtet, erscheinen damit Schmerzskalen mit verbalen Deskriptoren für die Schmerzintensität auch bei kognitiver Beeinträchtigung einfacher zu beantworten als abstraktere Antwortformate (NRS und vor

allem VAS; vgl. DNQP, 2004). In einer kürzlich veröffentlichten Konsensvereinbarung zur Schmerzmessung bei älteren Menschen sprach sich auch ein interdisziplinäres Expertengremium dafür aus, selbst bei demenzbedingt eingeschränkten verbalen Fähigkeiten wenn irgend möglich sowohl Selbstauskünfte als auch Verhaltensbeobachtungen zu berücksichtigen. Für ältere Menschen mit intakten bis mittelgradig beeinträchtigten kognitiven Fähigkeiten werden dabei neben verbalen Skalen allerdings auch die Coloured Analog Scale (CAS) und die Numerische Ratingskala explizit vorgeschlagen (Hadjistavropoulos et al., 2007).

Fisher und Kollegen (2006) verweisen jedoch darauf, dass eine Abschätzung der Möglichkeiten kognitiv beeinträchtigter Personen zu einer validen Selbstauskunft in den allermeisten Fällen dadurch erschwert wird, dass lediglich Bearbeitungsraten berichtet werden, der Prozess der Datengenerierung selbst (besonders mit Blick auf die geleistete Unterstützung und die Kriterien für den Abbruch einer Erfassung) jedoch weitestgehend unberücksichtigt bleibt (Fisher, Burgio, Thorn & Hardin, 2006). Nicht zuletzt darum wird ein über den einmaligen Einsatz von Standardinstrumenten deutlich hinausgehendes umfassendes Schmerzassessment gefordert (Davies et al., 2004a,b; Hadjistavropoulos et al., 2007). Eine Möglichkeit der Systematisierung des Assessments und -managements negativer Befindenszustände in der Pflege wurde mit der Serial Trial Intervention (STI) von Kovach und Kollegen vorgestellt (Kovach, Cashin & Sauer, 2006) und vor kurzem im deutschsprachigen Raum auf die Schmerzerfassung bezogen (Fischer, Spahn & Kovach, 2007).

Die für die Demenz vorgeschlagenen *visuellen Analogskalen* wurden als sprachfreie Tests zunächst für die Schmerzerfassung bei Kleinkindern entwickelt. Von der Übertragung auf die Klientel kognitiv beeinträchtigter Menschen erhoffte man sich aufgrund der geringen Sprachgebundenheit zunächst reliablere Schmerzauskünfte.

> These [. . .] scales minimize cognitive demands, and are, therefore, particularly suitable for adult populations with cognitive deterioration like AD. (Scherder & Bouma, 2000, p. 49)

Auch wenn Verfahren der direkten Skalierung häufig mit apparativen oder grafischen Antwortformaten (Gleitschieber etc.) arbeiten, ist für die vergleichsweise komplexe Instruktion der geforderten Abstraktionsleistung dennoch ein hinreichendes Sprachverständnis nötig. Da es sich um eine Übertragung einer Empfindung in einen anderen (Mess)-Zusammenhang handelt, müssen die Pole der grafischen Antwortformate in der Regel mit entsprechenden schmerzbezogenen verbalen Deskriptoren beschriftet werden. Im Vergleich zur VRS und NRS stellt die VAS höhere Anforderungen an die kognitive Leistungsfähigkeit der Probanden, und wird darum für den Einsatz bei älteren und insbesondere demenzkranken Personen nicht empfohlen (Basler et al., 2001; Herr et al., 2004; Kremer et al., 1981). Auch die betroffenen Demenzkranken selbst ziehen zur Schmerzbeurteilung verbale Ratingskalen den abstrakteren Analogskalen vor (Radbruch et al., 2000). Scherder und Bouma (2000) untersuchten die Anwendbarkeit mehrerer Analogskalen in Gruppen verschieden stark kognitiv beeinträchtigter älterer Menschen. Dabei konnten

mehr als die Hälfte der leicht dementen und immerhin noch 30 Prozent der mittelschwer beeinträchtigten Probanden die FPS korrekt bearbeiten, während für die etwas komplexere FAS bereits bei den nicht-dementen Vergleichspersonen substanzielle (25%) Verständnisschwierigkeiten beobachtet wurden.

3.3 Physiologische Marker

Erlebte Schmerzen sind gewöhnlich mit einer Reihe unmittelbarer vegetativer Zustandsänderungen, sowie mit expressiven und defensiven Verhaltensweisen verbunden. Anhaltende oder wiederkehrende Schmerzen führen darüber hinaus häufig auch zu längerfristigeren Veränderungen beispielsweise des Sozialverhaltens oder der Befindlichkeit. Nicht-verbale Verfahren der Schmerzmessung greifen diese prinzipiell beobachtbaren Schmerzmarker auf.

Bei Personen, die sich aufgrund eines veränderten Bewusstheitsgrades (z.B. Komapatienten), lebenserhaltender apparativer Maßnahmen (z.B. beatmete Patienten), oder (noch) nicht entwickelter oder verlorener geistiger Fähigkeiten (z.B. frühes Kindesalter, Demenz) nicht bzw. nur eingeschränkt verbal zu ihrem Schmerzerleben äußern können, stellen nicht-verbale Verfahren gegebenenfalls den einzig möglichen Zugang zur Schmerzerfassung dar.

Trotzdem muss betont werden, dass akute physiologische und reflexhafte Reaktionen auf Schmerzreize oder komplexere Verhaltensweisen zur Vermeidung und Bewältigung von Schmerzen auch bei voll erhaltener Auskunftsfähigkeit wichtige Informationen beispielsweise zur Schmerzlokalisation oder -verursachung liefern können. An der Vermittlung zumindest von akut erlebten Schmerzen sind neben der sprachlichen Auskunft in der Regel immer auch weitere Kommunikationskanäle beteiligt. Retrospektive Berichte erlebter Schmerzen werden nicht selten durch eine entsprechende begleitende Mimik, Körperhaltung oder Gestik anschaulicher gestaltet.

Auf der Suche nach Indikatoren erlebter Schmerzen, die in einem möglichst geringen Maß der bewussten Beeinflussung unterliegen, und in diesem Sinne als objektiv gelten können, wurden Möglichkeiten zur apparativen Ableitung und (funktionalen) Bildgebung verschiedener peripherer und zentraler Körpersignale entwickelt. Einen Überblick über gängige in der (experimentell-klinischen) Schmerzforschung verwendete Verfahren zur Ableitung und graphischen Veranschaulichung schmerzkorrelierter psychophysiologischer Signale gibt Flor (2001).

3.3.1 Periphere Schmerzmarker

Als *periphere Schmerzmarker* werden dabei beispielsweise Muskelspannungen (z.B. des nozizeptiven Flexor-Reflexes) mit Hilfe der Elektromyographie (EMG) aufgezeichnet und ausgewertet. Weitere apparativ relativ einfach ableitbare Schmerzreaktionen äussern sich in sympathischen Veränderungen des vegetativen Nervensystems, beispielsweise des Blutflusses und -druckes, der Herzrate oder der Hauttemperatur und -leitfähigkeit. Als sym-

pathische Reaktionen, die in Beziehung zur Bereitstellung und Entladung von Energie dienen, bilden diese Biomarker eher den emotionalen Aspekt des Schmerzerlebens ab (Möltner et al., 1990). Aufgrund ihrer Unabhängigkeit von kognitiven Fertigkeiten wird der Erfassung der physiologischen Schmerzreaktion bei Demenzpatienten von manchen Autoren einen besondere Bedeutung zugesprochen (Kunz, 2006). Allerdings sind physiologische Schmerzmarker im Allgemeinen nur unzureichend schmerzspezifisch, beispielsweise finden sich ähnliche vegetative Reaktionen auch bei anderen Defensivreaktionen wie dem Schreckrefex (Janig, 1995). Daneben können vegetative Schmerzmarker verschiedene Schmerzintensitäten nur unzureichend differenzieren (Möltner et al., 1990) und findet mit der Zeit eine Anpassung der physiologischen Reaktion auf wiederkehrende bzw. anhaltende Schmerzzustände statt. Insgesamt eignen sich diese Marker darum wohl eher im Kontext der experimentellen Schmerzforschung als für das alltägliche Schmerzmanagement in der Pflege.

3.3.2 Zentralnervöse Schmerzmarker

Zu den aktuellen Möglichkeiten der Veranschaulichung *zentralnervöser Schmerzmarker* zählen zum einen die Gruppe von Verfahren, die metabolistische Veränderungen in bestimmten Hirnregionen abzubilden erlauben, von denen dann auf neuronale Signale der Schmerzverarbeitung geschlossen wird. Das blutflussbasierte Neuroimaging, denen die Verfahren der Positronen-Emissions-Tomographie (PET) und der Single-Photonen-Emissions-Computertomographie (SPECT) angehören, arbeitet dazu mit radioaktiven Kontrastmitteln. Die funktionelle Magnet-Resonanz-Tomographie (fMRT) bildet dagegen die lokale Sauerstoffsättigung bzw. den stimuluskorrelierten Sauerstoffverbrauch in bestimmten Hirnstrukturen ab. Ein besonderer Vorteil dieser Abbildungen ist ihre sehr gute räumliche Auflösung, wohingegen die neuronale Aktivität selbst in ihrer zeitlichen Auflösung weniger gut dargestellt werden kann.

Eine zweite Gruppe von Verfahren des Neuroimaging misst die ereigniskorrelierte neurologische Aktivität direkt und nicht nur deren metabolische Korrelate. Hierzu gehören im einzelnen neuroelektrische und neuromagnetische Verfahren wie Elektro- und Magnetoenzephalographie (EEG, MEG). Durch eine angemessene Verstärkung der abgeleiteten elektrischen Potenziale wird eine sehr gute zeitliche Auflösung schmerzkorrelierter Hirnströme erreicht, während die räumliche Lokalisierung der Hirnaktivität bedeutend schwerer fällt.

Einige der zuvor bereits referierten Befunde zu alters- und demenzbezogenen Veränderungen im Schmerzerleben (vgl. Kapitel 2.2.4) gründen ganz wesentlich auf diesen modernen Verfahren des Neuroimaging. Es ist aber offensichtlich, dass der apparativen Schmerzmessung für das alltagspraktische Schmerzmanagement in der stationären Versorgung demenzkranker Menschen eine bestenfalls nachgeordnete Bedeutung zukommt, weswegen dieser Zugang zur Schmerzmessung hier nicht vertieft werden soll.

3.4 Verhaltensbeobachtung

Zur Abschätzung der akuten Schmerzbelastung solcher Personengruppen, die nicht zu einer verbalen Selbstauskunft in der Lage sind, wurden in den letzten Jahren mehr als zwei Dutzend schmerzbezogene Verhaltensinventare vorgeschlagen. Tabelle 2 gibt einen nach Erscheinungsjahr sortierten Überblick über die gegenwärtig diskutierten Instrumente aus dem anglo-amerikanischen und französischem Sprachraum.

Einige der neueren Verfahren stellen dabei Weiterentwicklungen früherer Skalen dar oder stellten ihre Erfassungsinhalte aus mehreren zuvor verfügbaren Verhaltensinventaren zusammen. Die PAINAD beispielsweise wurde auf der Grundlage der DS-DAT und der FLACC entwickelt. Die CNPI basiert auf der *University of Alabama Birmingham Pain Behaviour Scale* (UAB-PBS; Richards et al., 1982), die als solche gegenwärtig jedoch kaum mehr Aufmerksamkeit erfährt. ECPA gründet auf der für Kleinkinder entwickelten *Douleur Enfant Gustave Roussy Skala* (DEGR; Gauvain-Piquard & Pichard-Leandri, 1989). DOLOPLUS2 ist eine Weiterentwicklung der von Wary (1999) vorgestellten DOLOPLUS, die ihrerseits auf der DEGR beruht. Das *Observational Pain Behaviour Tool* (Simons & Malabar, 1995) ist eine Adaptation der PBM (Keefe & Block, 1982).

Die *Abbey Pain Scale* basiert auf Simon und Malabars Beobachtungsinstrument und der DS-DAT. Andererseits stellen beispielsweise Morello und Kollegen (2007) bei der Beschreibung der Konstruktion der EPCA-2 überraschenderweise keinerlei direkten Bezug zu ihren vorangegangenen Arbeiten mit der ECPA her (Morello et al., 1998; Morello et al., 2007).

Aufgrund der vielfältigen Verflechtungen bei der Skalenentwicklung darf angenommen werden, dass die gegenwärtig verfügbaren Verfahren zur Schmerzbeobachtung um einen als besonders relevant erachteten Pool einzelner Verhaltensindikatoren oder Indikatorbereiche herum gruppiert sind. Die Rationale der Beibehaltung, Modifikation oder Verwerfung einzelner Verhaltensweisen, sowie die empirische Grundlage für entsprechende Entscheidungen bleiben dabei jedoch häufig unklar. Da sich die Darstellung und Diskussion der psychometrischen Befunde der Originalarbeiten in den allermeisten Fällen auf den Gesamtskalenscore beschränkt, ist anzunehmen, dass die Itemselektion und -überarbeitung häufig auf der Ebene der Augenscheinvalidität verhaftet bleibt.

3.4.1 Behaviorale Schmerzindikatoren

Wenngleich Schmerzen häufig und systematisch mit bestimmten physiologischen Körperreaktionen, einem charakteristischen mimischen Ausdrucksverhalten oder komplexen Verhaltensweisen beispielsweise zur Beseitigung oder Vermeidung eines Schmerzreizes verbunden sind, konnten bislang keine über Populationen, Schmerzzustände und schmerzrelevante Situationen hinweg konsistent beobachtbaren und damit universell gültigen behavioralen Schmerzindikatoren identifiziert werden.

Die vielschichtigen Veränderungen, die mit einer demenziellen Erkrankung in den Bereichen kognitiver, motorischer, aber auch sozial- und alltagsbezogenenen Kompeten-

zen einhergehen, stellen auch für die Decodierung potenziell schmerzbezogenen Verhaltensausdrucks eine besondere Herausforderung dar. So erscheinen neben den kognitiven Abbauprozessen, die schließlich zu einer massiven Beeinträchtigung der Möglichkeiten für eine valide Selbstauskunft führen, insbesondere die häufig zu beobachtenden *nicht-kognitiven Demenzsymptome* wie Unruhe, Aggressivität, Enthemmung oder Apathie eine schlüssige schmerzbezogene Interpretation auffälligen Verhaltens zu erschweren.

Tabelle 2: Verfahren zur Schmerzerfassung durch Verhaltensbeobachtung.

Instrument	**Abkürzung**	**Autoren**
Mahoney Pain Scale	MPS	Mahoney & Peters, 2008
Elderly Pain Caring Assessment 2	EPCA-2	Morello et al., 2007
Mobilization-Observation-Behavior-Intensity-Dementia Pain Scale	MOBID	Husebo et al., 2007
Abbey Pain Scale	–	Abbey, Piller & De Bellis, 2004
Noncommunicative Patient's Pain Assessment Instrument	NOPPAIN	Snow et al., 2004
Pain Assessment Checklist for Seniors with Limited Ability to Communicate	PACSLAC	Fuchs-Lacelle & Hadjistavropoulos, 2004
Pain Assessment Tool for Use with Cognitive Impaired Adults	–	Davies et al., 2004a,b
Pain Assessment for the Dementing Elderly	PADE	Villanueva et al., 2003
Pain Assessment In Advanced Dementia	PAINAD	Warden, Hurley & Volicer, 2003
Pain Assessment in the Communicatively Impaired	PACI	Kaasalainen & Crook, 2003
Pain Assessment Tool in Confused Older Adults	PATCOA	Decker & Perry, 2003
Rating Pain in Dementia	RaPID	Sign & Orrell, 2003
Proxy Pain Questionnaire	PPQ	Fisher et al., 2002
DOLOPLUS-2	DOLOPLUS-2	Lefebvre-Chapiro et al., 2001
Checklist of Nonverbal Pain Indicators	CNPI	Feldt, 2000
Amy's Guide	Amy's Guide	Galloway & Turner, 1999
Assessment of Discomfort in Dementia Protocol	ADD	Kovach et al., 1999
Face, Legs, Activity, Cry, and Consolability Pain Assessment Tool	FLACC	Merkel et al., 1997
Behavior Checklist	–	Baker et al., 1996
Echelle Comportementa le simplifiée	ECS	le Quintrec et al., 1995
Observational Behavior Tool (keine offiz. Bezeichn.)	–	Simons & Malabar, 1995
Minimum Data Set	MDS	Morris et al., 1989
Echelle Comportementa le pour personne Agées	ECPA	Morello et al., 1998
Discomfort in Dementia of the Alzheimer's Type	DS-DAT	Hurley et al., 1992
Comfort Checklist	–	Volicer et al., 1988
Observational Pain Behavior Assessment Instrument	OPBAI	Teske et al., 1983
Pain Behavior Measure/Method	PBM	Keefe & Block, 1982
Facial Action Coding System	FACS	Ekman & Friesen, 1978

In ihren grundlegenden Empfehlungen zum Schmerzmanagement ordnet das Panel on Persistent Pain in Older Adults der American Geriatrics Society (AGS) das von kognitiv

beeinträchtigten Menschen häufig gezeigte Schmerzverhalten sechs übergeordneten Verhaltenskategorien zu (s. Tabelle 3).

Tabelle 3: Verhaltensbezogene Schmerzindikatoren nach der AGS-Leitlinie (AGS, 2002)

Category	**Examples**
Facial expressions	slight frown; sad, frightened face; grimacing, wrinkled forehead, closed or tightened eyes; any distorted expression; rapid blinking
Verbalizations, vocalizations	sighing, moaning, groaning; grunting, chanting, calling out; noisy breathing; asking for help; verbally abusive
Body movements	rigid, tense body posture, guarding; fidgeting; increased pacing, rocking; restricted movement; gait or mobility changes
Changes in interpersonal interactions	aggressive, combative, resisting care; decreased social interaction; socially inappropriate, disruptive; withdrawn
Changes in activity patterns or routines	refusing food, appetite change; increase in rest periods; sleep, rest pattern changes; sudden cessation of common routines; increased wandering
Mental status changes	crying or tears; increased confusion; irritability or distress

Das Schmerzverhalten kann nach dieser Auflistung als sowohl durch Annäherungs- (z.B. um Hilfe bitten, ängstlicher Gesichtsausdruck) als auch durch Vermeidungstendenzen bestimmt angenommen werden, wobei der letztere Aspekt in allen Ausdrucksbereichen deutlich überwiegt.

Das Schmerzverhalten muss nach dieser Auflistung immer als bidirektional angenommen werden, in dem Sinne, dass erlebter Schmerz sowohl zu einem Mehr an spezifischem Verhalten führt, als auch bestimmtes (normales) Verhalten verhindern kann (vgl. Hadjistavropoulos et al., 2007). Diese Perspektive hat auch Auswirkungen auf die gewählten Verfahrensweisen zur Schmerzmessung, beispielsweise wenn im Rahmen des *Pain Behavior Measurements* (PBM; Keefe & Block, 1982; Keefe, Williams & Smith, 2001) oder der MOBID2-Skala (Husebo et al., 2007) bewusst eine Reihe potenziell schmerzinduzierender Aktivitäten initiiert werden, oder in den Leitlinien zum pflegerischen Schmerzmanagement der *Australian Society for Pain Management Nursing (ASPMN)* analgetische Trials empfohlen werden.

Insbesondere für die Feststellung schmerzbedingt reduzierten oder fehlenden Verhaltens ist selbstverständlich eine Kenntnis der üblichen Routinen, Gewohnheiten und Gemütslagen des Betroffenen nötig. Die durch die drei letzten Indikatorbereiche abgebildeten veränderungsbezogenen Schmerzindikatoren machen darum ggfs. eine Einbeziehung von Proxys (z.B. Angehörige) mit entsprechenden gemeinsamen Erfahrungen oder zumindest wiederholte Beobachtungen und die Definition einer schmerzbezogenen Baseline nötig.

> It is quite common for caregivers to become familiar with residents' typical behavior and functioning. This information acts as an individual baseline.

> Abrupt, unexplained shifts from this baseline may indicate pain. (Mahoney & Peters, 2008, p. 252)
>
> Unusual behavior in a patient with severe dementia should trigger assessment for pain as a potential cause. (AGS, 2002, p. S210)

Schmerzverhalten erlangt seine Indikationskraft also zu einem nicht unerheblichen Teil durch den Abgleich mit situativ erwartetem Verhalten.

Auch ohne Kenntnis des Betroffenen und bei einmaliger Beobachtung kann ein Verhalten dann auf potenziellen Schmerz hinweisen, wenn es *nicht situationsadäquat* erscheint. So könnte eine starre Körperhaltung und eingeschränkte Bewegung in einer Aktivitätssituation eher auf Schmerzen hinweisen als in einer Ruhesitation. Umgekehrt benötigt beispielsweise die Feststellung sozial unangemessenen Verhaltens genaugenommen keine Referenz auf früheres sozialbezogenes Verhalten. Die im Verlauf der Demenz häufig auftretenden nicht-kognitiven Verhaltensauffälligkeiten stellen ebensolche unerwarteten Verhaltensweisen dar, und erschweren eine eindeutige Ursachenzuschreibung.

Die in den AGS-Kriterien genannten schmerzbezogenen Verhaltensweisen sind durch eine recht unterschiedliche zeitliche Erstreckung und Nähe zum Schmerzerleben gekennzeichnet. Insbesondere die Indikatoren aus den Bereichen Mimik, Lautäußerung und Körperhaltung weisen auf unmittelbar erfahrenen Schmerz hin, während die verbleibenden Kategorien vorrangig auf andauernde oder wiederkehrende Schmerzen hinweisen. Eine Veränderung von Gewohnheiten und Vorlieben vollzieht sich im Vergleich dazu (z.B. zu Mobilitätsveränderungen) entsprechend langsam, oder benötigt zumindest eine vergleichsweise lange Zeit für deren Feststellung.

Der Komplexitätsgrad der beschriebenen Verhaltensweisen reicht von isolierten Muskelbewegungen (z.B. Falten über die Stirn) bis hin zur Verletzung soziokulturell bestimmter Verhaltensnormen (sozial unangemessenes Verhalten). Beispiele für eine intermediäre Stufe schmerzbezogenen Verhaltensausdruckes in den AGS-Kriterien wären beispielsweise ein trauriger oder ängstlicher Gesichtsausdruck. Im Bereich körperbezogenen Schmerzausdruckes könnten Aspekte der Körperhaltung als komplexere Verhaltensform herausgearbeitet, mit Blick auf sozialbezogenes Schmerzverhalten dagegen Aspekte der Gestik (Abwehr, Rückzug, Verweigerung) näher beschrieben werden, um Beobachtungseinheiten zu definieren, die möglichst auch den natürlichen Wahrnehmungsgewohnheiten der Pflegenden entsprechen.

Aufgrund der Erkenntnis, dass sich die physiologische Schmerzreaktion bei chronischen Schmerzzuständen adaptiert, ist es nachvollziehbar, dass der AGS Panel for Persistent Pain in Older Adults keine physiologischen Marker in sein Kategoriensystem aufgenommen hat. Andererseits werden auch hier zumindest ein paar Beispielindikatoren aufgeführt, die in hohem Maße vegetativ bestimmt sind (lautstarke Atmung, angespannte Körperhaltung, Schlaf-Wach-Rhythmus).

In ihrem Reviewartikel ordnen Herr und Kollegen (2006) die in insgesamt zehn Verhaltensinventaren (Abbey, ADD, CNPI, DS-DAT, DOLOPLUS-2, FLACC, NOPPAIN, PACSLAC, PADE, PAINAD) aufgeführten Indikatorbereiche, konkreten Verhaltenswei-

sen und Beispielitems diesen sechs AGS-Indikatorbereichen zu. Für den überwiegenden Teil der Instrumente wurden dabei einzelne Verhaltensbereiche bzw. -weisen beschrieben, die sich nicht einfach in das vorgegebene AGS-Raster implementieren ließen. So werden bei PACSLAC beispielsweise zusätzlich physiologische Marker erfasst, und die Inventare FLACC und PAINAD berücksichtigen jeweils den Indikatorbereich Trost bzw. Tröstbarkeit (die m.E. allerdings der Kategorie Mental Status Change zugeordnet werden könnten). Die für einzelne Items oder Itembereiche gewählte Zuordnung erscheint nicht durchgängig leicht nachvollziehbar, beispielsweise wenn Indikatoren beschleunigter Atmung (PAINAD, PADE) dem Bereich der Lautäußerung zugeschrieben werden.

Einen Überblick über die in neun Verhaltensinventaren (Behaviour Checklist, CNPI, DS-DAT, DOLOPLUS, EPCA, FACS, PADE, PAINAD) enthaltenen Ausdrucksbereiche gaben bereits im Vorjahr auch Stolee und Kollegen (Stolee et al., 2005). Dabei differenzieren sie schmerzbezogenes Verhalten in die Ausdruckskategorien Vokalisation (Stöhnen, Weinen), Gesichtsausdrücke (Grimassieren, Gesicht verziehen), Verhaltensänderung (soziale Aktivitäten, Essen, Schlafen, Kommunikation), Körpersprache (Schutzverhalten, Körperhaltung), Stimmungsänderung (Ruhelosigkeit), Reaktionen während Alltagsaktivitäten (Waschen, Ankleiden, Gehen), Physiologische Veränderungen (Temperatur, Blutdruck, Rötung, Blässe), Verbale Klagen über Schmerzen, Lautstarkes Atmen und Tröstbarkeit.

In ihrer Beschreibung der strukturellen Eigenschaften von dreizehn verschiedenen Verfahren zur schmerzbezogenen Verhaltensbeobachtung (FACS, PBM, DS-DAT, DOLOPLUS2, Behavior Checklist, CNPI, ADD, PAINAD, PATCOA, PADE, PACSLAC, NOPPAIN und Abbey) unterschieden van Herk und Kollegen kürzlich (2007) die Verhaltenskategorien Gesichtsausdruck, Bewegung (motor behavior), Sozialverhalten oder Stimmung, Lautäußerung, Ernährungs- oder Schlafgewohnheiten (patterns) und physiologische Indikatoren.

Eine Kategorisierung des Schmerzverhaltens von Menschen mit schweren kognitiven Einschränkungen wurde für den *deutschsprachigen Raum* durch den vom *Deutschen Netzwerk für Qualitätsentwicklung in der Pflege (DNQP)* veröffentlichten *Expertenstandard Schmerzmanagement in der Pflege* geleistet (DNQP, 2004). Dabei wurden die in insgesamt sechs englischsprachigen qualitativen und literaturgestützten Arbeiten beschriebenen Indikatoren übersetzt und den übergeordneten Kategorien lautsprachlicher (verbal und vokal), mimischer, verhaltensbedingter und physischer, sowie veränderungsbezogener Schmerzindikatoren (Verhalten und Stimmung) zugeordnet (AGS, 2002; Carr & Mann, 2002; Cohen-Mansfield & Creedon, 2002; Epps, 2001; Herr & Garrand, 2001; Kovach et al., 2000). Leider blieb die tabellarische Übersicht weitestgehend unkommentiert, und naheliegende Fragen zur Äquivalenz der Begriffe in beiden Sprachräumen (z.B. crying, wincing) und möglichen weiteren Unsicherheiten bei der Übersetzung oder Kategorienzuordnung damit unbeantwortet.

Selbst wenn im Einzelfall die Zuordnung bestimmter schmerzbezogener Verhaltensweisen zu übergeordneten Kategorien schwierig ist, leisten die AGS-Verhaltensbereiche eine in aller Regel sinnvolle allgemeine Strukturierung potenziellen Schmerzverhaltens. Dennoch könnte die zusätzliche Berücksichtigung vegetativer bzw. physiologischer Mar-

ker zu einer vollständigeren Repräsentation theoretisch schmerzbezogener Verhaltensreaktionen beitragen. Da Verhalten eigentlich immer Veränderung bedeutet, die sinnvoller nach Komplexität, Unmittelbarkeit, zeitlicher Entwicklung oder Perseveranz bzw. Nachhaltigkeit beschrieben werden sollte, erscheinen insbesondere mit Blick auf die Kategorien subtileren Schmerzausdruckes (Veränderungen im Sozialverhalten, der Befindlichkeit usw.) präzisere Beobachtungs- und Entscheidungkriterien nötig.

3.4.2 Psychometrische Beurteilung der verfügbaren Verfahren

Mittlerweile liegen einige umfangreiche Übersichtsarbeiten vor, die den Aufbau und die psychometrischen Eigenschaften mehrerer Beobachtungsskalen auf der Grundlage der bislang publizierten empirischen Befunde miteinander vergleichen und bewerten (Hadjistavropoulos et al., 2007; Herr, Bjoro & Decker, 2006; Schofield et al., 2005[1]; Smith, 2005; Stolee et al., 2005; van Herk et al., 2007; Zwakhalen et al., 2006). Für einen großen Teil der insgesamt 26 verschiedenen in diesen Reviews berücksichtigten Skalen liegen damit allerdings paradoxerweise mehr Reviews vor als empirische Originalarbeiten (vgl. Tabelle 4).

Die Arbeitsgruppe um Keela Herr (2006) legt ihrer Beurteilung verfügbarer Verhaltensinventare zur Schmerzmessung die fünf Kriterienbereiche *Konzeptualisierung*, *Bezugsgruppe*, *Praktikabilität*, *Reliabilität* und *Validität* zugrunde, für die sie jeweils zwischen 0 und 3 Punktwerte vergeben. Zwakhalen und Kollegen (2006) differenzieren für ihr Review insgesamt zehn Beurteilungskriterien (Herkunft der Items, Stichprobengröße, Inhalts-, Kriteriums- und Konstruktvalidität, Skalenhomogenität, Interrater-, Intrarater- bzw. Retest-Reliabilität und Praktikabilität) und vergeben jeweils zwischen 0 und 2 Punkten. Beide Arbeiten weisen entsprechende Gesamtscores für die psychometrische Güte der berücksichtigten Instrumente aus. Stolee und Kollegen (2005) ziehen für ihre Beurteilung der Skalengüte das erreichte Skalenniveau der Testscores, die Itemanzahl, die Anzahl empirischer Studien, in denen das Instrument eingesetzt wurde, sowie die Differenziertheit und die Ergebnisse der Analysen zur Reliabilität und Validität heran. Für einen Teil dieser Kriterien vergibt auch diese Autorengruppe zwischen 0 und 3 Punkte, auch wenn von einer einfachen Summierung zu einem Gesamturteil hier abgesehen wird. Hadjistavropoulos und Kollegen (2007) gründen ihr Review auf einer Zusammenstellung der Itemanzahl, des Itemformates, der Durchführungsdauer, internen Konsistenz und Interraterreliabilität, sowie der jeweils berichteten Hinweise auf die Validität der Verfahren. Ihre Empfehlungen leitet die Gruppe aus einer narrativen Zusammenschau der Limitationen in allen zuvor genannten Kriterienbereichen ab. In gleicher Weise leisten auch Smith (2005), Schofield und Kollegen (2005), sowie die Autorengruppe um van Herk (2007) eine stärker narrative Zusammenschau der in einzelnen Studien berichteten psychometrischen Befunde.

Im Folgenden sollen Problembereiche der psychometrischen Güte der gegenwärtig verfügbaren "*tool box*" zur schmerzbezogenen Verhaltensbeobachtung bei nicht-auskunfts-

[1] Die zitierte Arbeit ist unter http://auraserv.abdn.ac.uk als Manuskript verfügbar, wurde jedoch, entgegen der dortigen Angaben und der Zitation in Schofield & Reid, 2006 bislang nicht veröffentlicht.

Tabelle 4: Aktuelle Überblicksartikel zur Schmerzerfassung durch Verhaltensbeobachtung.

Instrument	van Herk et al., 2007	Hadjistavropoulos et al., 2007	Herr et al., 2006	Zwakhalen et al., 2006	Smith, 2005	Stolee et al., 2005	Schofield et al., 2005
Abbey Pain Scale	■	■	■	■			■
CNPI	■	■	■	■	■	■	■
DS-DAT	■	■	■		■	■	■
DOLOPLUS-2	■	■	■	■		■	■
NOPPAIN	■	■	■	■			■
PACI		■					
PAINAD	■	■	■	■	■	■	■
PATCOA	■	■					
PACSLAC	■	■	■	■			■
PADE	■	■	■	■	■	■	■
Amy's Guide		■					
Observational Behavior Tool		■		■	■		
ADD	■		■				■
FLACC			■				
ECPA				■			
l'ECS				■			
RaPID				■			
Pain Assessment Tool for Use with Cognitive Impaired Adults				■			
Comfort Checklist					■		
PPQ					■	■	
Behavior Checklist	■					■	
EPCA[1]						■	
FACS	■					■	
MDS						■	
PBM	■					■	

[1] Hierbei handelt es sich um eine frühe Version der französischen EPCA-2-Skala, die auf dem Weltkongress für Gerontologie 1992 vorgestellt wurde.

fähigen demenzkranken Menschen mit Blick auf die übergeordneten, wenngleich auch nicht voneinander unabhängigen Kriterienbereiche *Reliabilität, Praktikabilität bzw. klinische Nützlichkeit* und *Validität* zusammenfassend skizziert werden, ohne auf spezifische Details einzelner Instrumente einzugehen. Eine detaillierte Diskussion der beiden im Rahmen dieser Arbeit berücksichtigten Verfahren (dt. Übersetzungen der PAINAD und CNPI) wird im Kapitel zur Datenbasis (Kap. 5.3.4 und 5.3.5) geleistet.

3.4.2.1 Reliabilität

Die Reliabilität eines Verfahrens ist abhängig von dem Ausmaß, in dem ein Instrument durch den wahren Merkmalswert bestimmt ist. Gemäß den Annahmen der klassischen Testtheorie gilt ein Instrument dann als reliabel, wenn es bei mehrmaligem Einsatz und gleichbleibendem Merkmalswert zu konstanten Ergebnissen führt. Hinweise auf die Reliabilität der vorgeschlagenen Verhaltensinventare zur Schmerzbeobachtung gibt dementsprechend die Übereinstimmung zweier Messwertreihen, die zu unterschiedlichen Zeitpunkten (Retest-Reliabilität, Intra-Rater-Reliabilität), durch unterschiedliche Beurteiler (Inter-Rater-Reliabilität) oder anhand von parallelen Testformen (Split-Half-, Paralleltest-Reliabilität, Interne Konsistenz) erhoben worden sind. Eine ausführliche kritische Würdigung des Reliabilitätskonzeptes im Kontext der klassischen Testtheorie (KTT) und dessen Erweiterungen durch die Generalisierungstheorie und probabilistische Testtheorie erfolgt im Methodenteil dieser Arbeit.

Bei der Diskussion der psychometrischen Eigenschaften der vorgeschlagenen Verhaltensinventare wurden die konzeptionellen Beschränkungen, die mit dem Reliabilitätsbegriff der KTT verbunden sind, bislang nahezu vollständig ausgeklammert. Beispielsweise werden die verschiedenen Typen von Reliabilitätskoeffizienten (Interrater, Retest, Konsistenz) in der Arbeit von Zwakhalen und Kollegen (2006) getrennt voneinander dargestellt und diskutiert, obgleich alle Maße dieselbe *eine* Skalenreliabilität anzeigen. Bei der Gesamtbeurteilung der Verfahren kommt damit dem Aspekt der Reliabilität folglich ein vergleichsweise großes Gewicht zu.

Aufgrund der häufig kurzfristigen und wechselhaften Natur akuten Schmerzes erscheint bei Inventaren zur Beobachtung akut schmerzbezogenen Verhaltens die Angemessenheit von *Test-Retest-Designs* zur Abschätzung der Verlässlichkeit eines Verfahrens fraglich, da sowohl Retest- als auch Intrarater-Reliabilitäten prinzipiell unveränderte Merkmalswerte voraussetzen. Für Verfahren, die sowohl unmittelbaren Schmerzausdruck, als auch subtilere längerfristige Verhaltensänderungen berücksichtigen (z.B. PACSLAC), ist die Bestimmung eines sinnvollen Intervalles für Retests zur Reliabilitätsbestimmung besonders schwierig.

Auch der unzweifelhafte Umstand, dass praktisch kein reales Instrument ein Merkmal über den gesamten Merkmalsraum hinweg, also sowohl im Bereich geringen, als auch exzessiven Schmerzes, gleich reliabel zu messen in der Lage ist, sondern sich je nach der Verteilung leichterer und schwierigerer Items unterschiedlich präzise Abbildungsbereiche ergeben, wird bei der Diskussion der Schmerzskalen vollständig außen vor gelassen. Dabei

scheinen zumindest einzelne Skalen, beispielsweise DOLOPLUS2, FLACC und PAINAD ihre Verhaltensindikatoren auch nach der Intensität des anzuzeigenden Schmerzerlebens zu sortieren, und gehen damit von unterschiedlichen Itemschwierigkeiten aus.

Der Zusammenhang zwischen der internen Konsistenz einer Itembatterie und der Itemanzahl wird dagegen erkannt und angesprochen. Je weniger Items eine Skala umfasst, desto schwieriger können mit dieser Statistik hohe Reliabilitätswerte erreicht werden. Wenngleich dahingehend Einigkeit besteht, dass im Klinikalltag praktikable Instrumente relativ kurz sein müssen, bleibt offen, wie die Skalenkonsistenz im Verhältnis zur Itemanzahl beurteilt werden soll. Während Hadjistavropoulos und Kollegen (2007) die geringen Konsistenzwerte der CNPI und BESD Skala als Hinweis darauf werten, dass manche der Items möglicherweise ein anderes Konstrukt als Schmerz abbilden, sehen beispielsweise Zwakhalen und Kollegen (2006) die geringe Itemanzahl (N=6) der CNPI als Grund für die geringe interne Konsistenz an, bewerten aber die ähnlich moderate Konsistenz der PAINAD „angesichts der begrenzten Itemanzahl (N=5) als bemerkenswert gering“ (Übersetzung durch den Autor).

Insgesamt werden die verfügbaren Reliabilitätsnachweise vorwiegend als unvollständig oder ungenügend bewertet. Die zehn von Herr und Kollegen berücksichtigten Instrumente erreichen im Mittel gerademal 1,5 von 3 möglichen Punkten, und auch die von Stolee und Kollegen eingeschätzten Skalen müssen mit im Mittel 1,3 von maximal drei Punkten als insgesamt wenig zufriedenstellend gelten. Bei vier der zwölf von Zwakhalen und Kollegen eingeschätzten Skalen war überhaupt keine (angemessene) Information bezüglich der Reliabilität verfügbar, und nur zwei Skalen berichten Reliabilitätsindizes in allen drei differenzierten Kategorien (Konsistenz, Interrater- und Retest-Reliabilität). Bei der Einschätzung der Reliabilität einzelner Skalen zeigen sich bei gleicher Befundlage zum Teil große Differenzen zwischen den Autorengruppen, die mitunter auch durch eine unterschiedlich strenge Bewertung der Angemessenheit der berichteten Reliabilitätskennwerte bedingt scheinen. Daneben berücksichtigen beispielsweise Herr und Kollegen auch unveröfffentlichte Ergebnisse zu laufenden Forschungsprojekten (NOPPAIN), und kommen auf dieser Grundlage zu deutlich anderen Reliabilitätsbeurteilungen als beispielsweise Zwakhalen und Kollegen. Die verbleibenden Arbeiten bewerten und diskutieren die verfügbaren Reliabilitätsnachweise im klassisch narrativen Stil, kommen aber ebenfalls zu dem Schluss, dass die für einige Skalen durchaus zufriedenstellend hohen Reliabilitätskennwerte einer Replizierung in umfangreicheren, und stärker auf nicht-kommunikative demenzkranke Menschen ausgerichteten Studien bedürfen. Nicht zuletzt aufgrund der sehr wenigen empirischen Studien und der für manche Instrumente doch sehr geringen Stichprobengröße folgern alle Arbeiten einheitlich, dass die Möglichkeiten, verlässliche Aussagen zur psychometrischen Qualität einzelner Skalen zu machen, gegenwärtig – beim durch die referierten Arbeiten gut bekundeten besten Willen – sehr beschränkt bleiben.

3.4.2.2 Praktikabilität

Ein Verfahren zur Schmerzmessung durch Verhaltensbeobachtung wird im Praxisalltag nur dann Akzeptanz und Verbreitung finden, wenn es nicht nur eine hinreichend valide und reliable Abbildung von Schmerzen erlaubt, sondern darüber hinaus auch einfach anzuwenden ist. Praktikabilität für den Versorgungsalltag demenzkranker Menschen bedeutet im engeren Sinne, dass ein Verfahren hinreichend kurz und überschaubar ist, um auch bei knappen zeitlichen Ressourcen durchgeführt zu werden. Daneben müssen die Inhalte und Instruktionen zur Durchführung und Interpretation auch für wenig vorgebildete Personen hinreichend verständlich gestaltet sein.

Der Aufwand für Schulung und Einarbeitung, und die Dauer des Instrumenteneinsatzes bzw. der Auswertung werden entsprechend auch in den meisten der genannten Reviewverfahren als Kriterien für die Alltagstauglichkeit berücksichtigt.

Werden Praktikabilität und Effizienz nicht nur als Nebengütekriterien verstanden, dann bestimmen diese wesentlich auch den möglichen Auflösungsgrad der Schmerzmessung mit. Die Frage, welche Information zum Schmerz (beispielsweise Intensität vs. Vorliegen) in der Praxis vorrangig benötigt wird, scheint bei der Beurteilung der vorgeschlagenen Verhaltensinventare jedoch kaum eine Rolle zu spielen. Würde man sich beispielsweise darauf einigen, dass das Ziel der Schmerzmessung eher in der Unterscheidung vorhandener und nicht-vorhandener Schmerzen liegt als in der Bestimmung der Schmerzintensität, könnte auch für die Abschätzung der Validität eine engere Auswahl angemessener Vergleichskriterien erfolgen.

Besteht das Ziel der Instrumentenentwicklung darin, möglichst keine vorhandenen Schmerzzustände unerkannt zu lassen (hohe Sensitivität, geringe Falsch-Negativ-Rate), erscheint eine umfangreiche Batterie breitgefächerter potenziell schmerzbezogener Verhaltensweisen angezeigt. Besteht das Ziel dagegen eher in der Entwicklung eines effizienten Screeningverfahrens mit hoher Schmerzspezifität (geringe Falsch-Positiv-Rate), erscheint eine Zusammenstellung weniger, besonders eindeutiger Schmerzindikatoren sinnvoll. Obschon sicherlich ein gerechtfertigter Bedarf für beide Instrumentengewichtungen besteht, weisen die mittlerweile auch empirisch gut belegten Befunde zur Unterversorgung demenzkranker Menschen mit Analgetika darauf hin, dass insbesondere sensitive Verfahren zur Entdeckung nicht verbal geäußerter Schmerzzustände dringend benötigt werden. Diese Forderung ist umso gerechtfertigter, als die zur Behandlung fraglicher Schmerzen eingesetzten Analgetika - entgegen sich hartnäckig haltender Vorurteile und Befürchtungen - nur selten Nebenwirkungen und ein geringes Abhängigkeitspotenzial besitzen, und der potenzielle Schaden einer falsch-positiven Zuordnung damit sehr überschaubar bliebe. Allerdings muss an dieser Stelle bereits einschränkend angemerkt werden, dass weder die Sensitivität noch die Spezifität eines Verfahrens ausschließlich über die Itemanzahl hinreichend gesteuert werden können, sondern weitergehende Informationen zur Indikationsgüte einzelner Verhaltensindikatoren nötig sind. Die vorliegende Arbeit überwindet die gegenwärtig dominierende Orientierung an der Gesamtskala zugunsten einer solchen, stärker auf die Eigenschaften der enthaltenen Einzelitems abgestellten Perspektive.

In ihrem Überblick berücksichtigen beispielsweise Hadjistavropoulos und Kollegen als Merkmal für die Praktikabilität eines Verfahrens vorrangig die Itemanzahl, die Dauer der Bearbeitung und das Itemformat. Im Gegensatz zu herkömmlichen paper-and-pencil-Verfahren (Selbstauskunft) bestimmt sich die Bearbeitungszeit bei Verhaltensbeobachtungen aber weniger an der Itemanzahl, sondern wird häufig ein definiertes Beobachtungsintervall vorgegeben. Verfahren mit vielen Indikatoren erfordern natürlich dennoch ein höheres Maß an Vorbereitung und Schulung, sowie eine gesteigerte Aufmerksamkeit und Sensitivität während der Beobachtung. Auch gemischte Antwortformate oder Items mit vielen Antwortkategorien (z.B. Intensitätseinschätzung vs. beobachtet/nicht beobachtet) steigern die Komplexität eines Verfahrens und können die Praktikabilität beeinträchtigen.

Die Anzahl der in den vorgeschlagenen Instrumenten berücksichtigten Items variiert beträchtlich (PAINAD: 5, PACSLAC: 60). In ihrem Überblick unterscheiden Hadjistavropoulos und Kollegen (2007) grob zwischen Instrumenten mit bis zu 10 Items und umfangreicheren Inventaren. Diese Einteilung kann helfen, die klinische Praktikabilität der Skalen einzuschätzen, da zumindest im Bereich der Altenpflege die zeitlichen Ressourcen eher knapp sind. Andererseits unterscheiden sich die in verschiedenen Instrumenten enthaltenen Items deutlich auch in ihrem Aufbau und damit dem eigentlichen Bearbeitungsaufwand.

Nicht zuletzt aus Gründen der Übersichtlichkeit eines Instrumentes ist es sinnvoll, nicht nur eine ungeordnete Anzahl von – vielleicht sogar in ihrer Reihenfolge randomisierten – konkreten Verhaltensweisen vorzugeben, sondern diese Verhaltensweisen vorab, z.B. nach Körperregionen oder Ausdruckskanälen und -formen zu gliedern. Damit wird zum einen die Aufmerksamkeit des Beobachters gesteuert und daneben die in-situ-Dokumentation beobachteten Verhaltens erleichtert. Einige Verfahren beschränken sich allerdings darauf, lediglich übergeordnete Verhaltenskategorien (z.B. Mimik, Gestik, Lautäußerung vorzugeben) als Items im eigentlichen Sinne vorzugeben. Entsprechend werden diese Verfahren in vergleichenden Reviews als kurze und damit für den Klinikalltag angemessene Verfahren rezipiert.

Um dem Beobachter aber dennoch eine Hilfestellung zu geben, welche konkrete Verhaltensweisen in diesen Bereichen zu scoren sind (und welche nicht), werden dabei in der Regel Beispiele entsprechenden schmerzbezogenen Ausdrucksverhaltens aufgeführt. Tatsächlich besteht die Aufgabe des Beobachters damit darin, eine theoretisch unbegrenzte Anzahl möglicher (vergleichbarer) konkreter Verhaltensweisen, praktisch zumindest aber eine deutlich über die Kategorienanzahl hinausgehende Zahl beispielhafter (vergleichbarer) Verhaltensweisen zu beobachten oder eben nicht zu beobachten.

Wird aber nicht dokumentiert, welche konkreten schmerzbezogenen Verhaltensweisen beobachtet wurden, bleiben die Möglichkeiten einer Abschätzung sowohl der Praktikabilität als auch der Güte des Assessmentverfahrens sehr begrenzt. Es ist unklar, inwiefern die Beobachter von der Freiheit Gebrauch machen, neben dem aufgeführten Beispielverhalten noch weitere vergleichbar erscheinende Verhaltensweisen berücksichtigen zu können. Aufgrund der in der Pflege herrschenden knappen zeitlichen Ressourcen und der immernoch weitverbreiteten Unsicherheit im Hinblick auf Schmerzmessung und -management

darf angenommen werden, dass Pflegende Instrumente mit möglichst konkreten Vorgaben gegenüber solchen mit großer individueller Deutungsnotwendigkeit bevorzugen. Der praktische Nutzen eines Instrumentes aber ist wesentlich von der Akzeptanz durch die in der Praxis Tätigen abhängig. Stolee und Kollegen (2005) bemängeln, dass für kaum eines der vorgeschlagenen Beobachtungsinstrumente Bearbeitungsraten (completion rates) oder ähnliche Hinweise darauf gegeben werden "whether they were able to respond to all of the items or the extent to which they were confident in their responses or were providing best "guesstimates" " (Stolee et al., 2005, p. 325; Hervorhebung im Original).

Neben dem Skalenumfang kommt auch Aspekten des Settings eine wichtige Bedeutung für die Abschätzung der Praktikabilität eines Instrumentes zu. Während der Großteil der Instrumente im und für den stationären Kontext entwickelt wurden, steht die Überprüfung der Einsetzbarkeit einzelner Skalen, wie beispielsweise der CNPI in diesem Setting noch aus. Die vorliegende Arbeit trägt dazu bei, diese Forschungslücke zu schließen.

Ein Teil der Instrumente (z.B. PAINAD, CNPI, NOPPAIN) bestimmt die Beobachtungssituation inhaltlich präziser, indem der Betroffene beispielsweise in Ruhe, während Alltagsaktivitäten oder in Pflegesituationen hinsichtlich ihres schmerzbezogenen Verhaltensausdruckes eingeschätzt werden sollen. Die Notwendigkeit, bestimmte Rahmenbedingungen für die Schmerzbeobachtung herzustellen oder abzuwarten schränkt die klinische Nützlichkeit dieser Verfahren unter Umständen ein. Ausdifferenzierte Protokolle für das pflegerische Schmerzassessment, die einen systematischen Algorithmus von Messung, Intervention und Wirksamkeitsüberprüfung vorsehen, bergen sowohl das Potenzial das Schmerzmanagement transparenter und effektiver zu machen, als auch die Gefahr, flexibles pflegerisches Handeln in ein büro- oder technokratisches Korsett zu zwängen. Die Herausforderungen bei der Implementierung eines detaillierten Schmerzprotokolls in die Pflegepraxis wurden beispielsweise von Davies und Kollegen beschrieben (Davies et al., 2004a,b). Die Praktikabilität eines durchstrukturierten Assessments und Managements besonderer Pflegeanforderungen (z.B. herausfordernden Verhaltens und Schmerz) wird für den deutschen Versorgungsbereich gegenwärtig evaluiert (Fischer, Spahn & Kovach, 2007).

Die Autorengruppe um van Herk (2007) ergänzt die Einschätzung der Alltagstauglichkeit bzw. Anwendbarkeit explizit um den Aspekt der klinischen Nützlichkeit. Dabei betrachten sie Skalen als klinisch nützlich, wenn der ermittelte Skalenwert für klinische Entscheidungen, z.B. medikamentöse Intervention o.ä. herangezogen werden kann. Für die DOLOPLUS2 beispielsweise wird ein Cut-off-Wert für das Vorliegen von Schmerzen vorgegeben. Bei der ADD ist die Schmerzerfassung direkt mit der Interventionsplanung (z.B. Verabreichung analgetischer Bedarfsmedikation) kombiniert (Kovach et al., 2002).

3.4.2.3 Validität

Ein Instrument wird dann als valide bezeichnet, wenn es die Abbildung des in Frage stehenden Zielkonzeptes erlaubt, einfach ausgedrückt also das misst, was es messen soll. Damit aber ist die Validität das wichtigste Kriterium zur Bestimmung der psychometri-

schen Güte einer Schmerzmessung.

Die Schmerzmessung ist kein Selbstzweck. Ziel der Messung könnte beispielsweise sein, ein Verdachtsmoment vorliegender Schmerzbelastung zu bestätigen, die weitere klinische Versorgung zu informieren, oder durchgeführte Interventionen zu evaluieren. Da die Messung damit verschiedene Ziele verfolgen kann, muss auch deren Validität mit Blick auf die jeweils intendierte Verwendung bzw. den Zweck des Einsatzes beurteilt werden.

Entspricht die Auswahl von Probanden einer empirischen Studie nicht der intendierten Zielgruppe, oder sind deren Kernmerkmale aufgrund einer zu geringen Stichprobengröße nicht hinreichend gut repräsentiert, kann die Validität eines Verfahrens auf dieser Datengrundlage prinzipiell nicht nachgewiesen werden.

Das prinzipiell unlösbare Grundproblem der Validität besteht demnach darin, dass man nicht etwas Bekanntes zur Abbildung bringen, sondern durch ein bestimmtes alternatives Verfahren seinen Gegenstand detaillierter kennenlernen will. Ob dieses Vorhaben gelingt, wird jedoch daran fest gemacht, ob die apriorischen (d.h. dem bisherigen Forschungsstand entsprechenden) Annahmen zum in Frage stehenden Merkmal bestätigt werden oder nicht.

Insgesamt erstreckt sich die Aufgabe der Validitätssicherung von der inhaltlichen Bestimmung des Zielkonzeptes, über die Definition der Zielpopulation bis hin zur Festlegung des eigentlichen praktischen Verwendungszweckes und der implizierten Folgen des Instrumenteneinsatzes. Die häufig separat diskutierten Kriterienbereiche Konzeptualisierung und Untersuchungs- bzw. Zielpopulation lassen sich damit als Subaspekte des Validitätskriteriums begreifen.

Wie zuvor bereits angemerkt, basiert ein Teil der vorgeschlagenen Beobachtungsverfahren auf Verfahren, die ursprünglich für andere nicht-auskunftsfähige Populationen entwickelt wurden. Sowohl Zwakhalen und Kollegen (2006), als auch Herr und Kollegen (2005) verzichten beispielsweise bewusst auf eine Berücksichtigung der PATCOA (Decker & Perry, 2003), da diese für die Zielpopulation verwirrter, jedoch kognitiv unbeeinträchtigter Personen entwickelt wurde. Zwakhalen und Kollegen sehen darüber hinaus in ihrem Review von einer Berücksichtigung der DS-DAT (Hurley et al., 1992) ab, da das Konzept des Unwohlseins theoretisch nur unzureichend mit dem Konzept Schmerz übereinstimmt.

Die genannten Übersichtsarbeiten orientieren sich im Gegensatz zu der soeben skizzierten wesentlich auch erkenntnistheoretisch orientierten Auffassung von Validität an einem vergleichsweise eng gefassten Validitätsbegriff, auch wenn sie dabei eine ganze Reihe verschiedener Validitätskriterien berücksichtigen (vgl. van Herk et al., 2007).

Augenscheinvalidität

Insbesondere bei der Sammlung bzw. Zusammenstellung möglicher Merkmalsindikatoren in frühen Phasen der Instrumentenentwicklung werden potenzielle Items auf der Grundlage ihrer augenscheinlichen Verbundenheit mit dem Zielkonstrukt beurteilt. Dieser erste Eindruck von einer „trefflichen Formulierung“ kann dazu beitragen, die nomologische Struktur des Merkmalsraumes zu verdeutlichen und das zeitgeistliche oder kultu-

rell bedingte Verständnis eines Merkmales aus dem Sprachgebrauch heraus einzuschätzen bzw. die wissenschaftliche Sprache mit der Alltagssprache abzugleichen. Folgerichtig wird diesem Aspekt in besonders sorgfältig durchgeführten Skalenentwicklungen auch entsprechende Aufmerksamkeit gewidmet (vgl. Morello et al., 2007). Andererseits ist es nicht notwendig – und gelegentlich auch gar nicht erwünscht –, dass der Gegenstand der Messung allzu offensichtlich ist, beispielsweise wenn die Gefahr einer bewussten Täuschung bestünde. Sprachgebundene Verfahren der Schmerzmessung adressieren das Schmerzerleben in aller Regel direkt und unverschleiert. Trotzdem ist bekannt, dass insbesondere ältere Menschen häufig auch weniger offensichtlich mit Schmerz verbundene Erlebensweisen (z.B. Schlafstörungen, Unwohlsein, Appetitlosigkeit) schildern, anstatt unmittelbar von Schmerzen zu sprechen (vgl. DNQP, 2004). Auch im Bereich der verhaltensgestützten Schmerzmessung können beispielsweise die durch die AGS vorgeschlagenen veränderungsbezogenen subtilen Ausdruckskategorien (Veränderungen in Gewohnheiten und Alltagsroutinen, in sozialen Beziehungen und im Affektleben) nur bedingt Augenscheinvalidität beanspruchen. Limitationen für ein augenscheinvalides Verhaltensinventar ergeben sich darüber hinaus auch durch die, durch verschiedene Demenzätiologien und Multimorbiditäten bedingten, hochgradig spezifischen Muster individuellen Schmerzausdruckes.

Bei ihrer Beurteilung der psychometrischen Güte der bislang vorgeschlagenen Verhaltensinventare zur Schmerzbeobachtung messen die Autoren der hier beschriebenen Reviewarbeiten der Augenscheinvalidität darum – wenn überhaupt – nur eine nachrangige Bedeutung zu (vgl. van Herk et al., 2007).

Inhaltsvalidität

Ein Instrument gilt gemeinhin dann als inhaltsvalide, wenn es durch die spezifische Zusammenstellung von Items oder Subskalen alle relevanten Facetten des abzubildenden Merkmales repräsentiert. Da es sich beim Schmerzerleben um ein vergleichsweise komplexes Geschehen mit sensorisch-diskriminativen, affektiv-motivationalen und kognitiv-evaluativen Komponenten handelt, erscheint auch ein entsprechend multidimensionales Messinstrument nötig, um der Forderung nach umfassender inhaltlicher Validität nachzukommen. Tatsächlich beschränken sich die meisten der vorgeschlagenen Verfahren darauf, einzelne Dimensionen wie die Intensität des sensorisch-diskriminativen Erlebens oder der affektiven Schmerzkomponente abzubilden. Insbesondere in Verhaltensinventaren zur Schmerzmessung, bei denen die Indikatoren häufig nach Körper- bzw. Ausdrucksbereich gegliedert werden, bleibt die Zuordnung zu einzelnen Erlebensqualitäten unklar. Das Reiben schmerzender Körperteile, abwehrende Gesten und fugale Körperbewegungen beispielsweise könnten stärker die sensorisch-diskriminativen Anteile (z.B. Nozifensivreflex) des Schmerzerlebens anzeigen, während das Anklammern an Gegenstände den kognitiv-evaluativen Aspekt der Antizipation von Schmerzen, und ein trauriger Gesichtsausdruck die (langfristige) affektive Folge andauernder Schmerzexposition darstellen könnte.

In gleicher Weise könnte man auch die Frage stellen, ob durch die getroffene Zu-

sammenstellung von schmerzbezogenen Verhaltensweisen der gesamte theoretisch oder auch praktisch relevante Ausschnitt verschiedener Schmerzintensitäten abgebildet werden kann, oder aber bestimmte – z.B. für die Zuordnung von Patienten zu verschiedenen Interventionsstrategien besonders relevante – Abschnitte auf dem angenommenen latenten Schmerzkontinuum weniger gut repräsentiert sind als andere. Eine detaillierte Analyse der Beobachtungsraten einzelner Verhaltensweisen (Basisraten) und eine entsprechende Diskussion der Itemschwierigkeiten wird zwar vereinzelt gefordert (Herr et al., 2006), wurde bislang jedoch weder bei der Instrumententwicklung noch bei der Beurteilung der psychometrischen Güte der vorgeschlagenen Verfahren systematisch verfolgt.

In den angeführten Reviews wird die Inhaltsvalidität der vorgeschlagenen Verfahren unter anderem daran gemessen, inwiefern alle sechs durch die AGS beschriebenen potenziellen Kategorien schmerzbezogenen Verhaltensausdruckes berücksichtigt sind. Verfahren zur Schmerzmessung, die auf einzelne oder wenige Ausdrucksbereiche fokussieren (wie beispielsweise FACS auf die Mimik) wird entsprechend eine geringere Inhaltsvalidität zugesprochen.

Konstruktvalidität

Von *Konstruktvalidität* kann dann gesprochen werden, wenn ein Instrument das in Frage stehende theoretische Konstrukt bzw. Merkmal erfasst. In einem engen Sinn können darunter auch die Verhältnisse zwischen einzelnen Items oder Subskalen eines Instrumentes verstanden werden. Eine etwas breitere Definition versteht darunter ganz allgemein die Übereinstimmung eines Verfahrens mit den theoretischen Annahmen zur Struktur des Merkmalsraumes. Inhaltsvalidität bezieht sich auf die Angemessenheit der Zusammenstellung von Iteminhalten bzw. abgebildeten Merkmalsfacetten, und kann somit (wie ferner auch die Augenscheinvalidität) als Komponente der Konstruktvalidität begriffen werden.

Um abzuschätzen, ob ein zu prüfendes Verfahren das gewünschte Merkmal abbildet, wird gewöhnlich ein Vergleich mit Instrumenten durchgeführt, die entweder dasselbe Konstrukt (konvergente Validität) oder ein anderes Merkmal (divergente Validität) messen. Selbstverständlich ist ein solcher Abgleich nur dann informativ, wenn die Konstruktvalidität der dazu herangezogenen Verfahren bereits gut belegt ist. Verschiedene konkurrierende Instrumente zur verhaltensgestützten Schmerzmessung bei nicht-auskunftsfähigen Personen können darum zumindest beim gegenwärtigen Stand der empirischen Überprüfung nur bedingt wechselseitig aneinander validiert werden. Dennoch stellen einige empirische Studien verschiedene alternative Inventare zur Verhaltensbeobachtung auch zu diesem Zweck einander direkt gegenüber (Cohen-Mansfield, 2008; Cohen-Mansfield & Lipson, 2008; Warden, Hurley & Volicer, 2003; Zwakhalen, Hamers & Berger, 2006). Wird ein zu überprüfendes Instrument mit einer Skala verglichen, die ein anderes Konstrukt als Schmerz misst, werden prinzipiell geringere empirische Zusammenhänge erwartet. Stellt man jedoch einen Bezug zu einer Skala her, die zwar ein anderes Konstrukt, beispielsweise Depressivität, erfasst, das jedoch theoretisch mit dem Schmerzerleben verbunden ist, können auch moderate Korrelationen als Hinweis auf die Konstruktvalidität der Schmerz-

erfassung gelten. Allgemein wird für eine Abschätzung der Konstruktvalidität eines Instrumentes also überprüft, wie stimmig sich dieses in das Netzwerk theoretischer Bezüge und bisheriger empirischer Erkenntnisse einbinden lässt. Skrondal und Rabe-Hesketh (2004) sprechen darum in diesem Zusammenhang auch von der *nomologischen* Validität eines Verfahrens.

Kriteriumsvalidität

Die *Kriteriumsvalidität* bezieht sich auf die relative Übereinstimmung der gemessenen Merkmalswerte mit einem externen Kriterium, dessen Validität bereits als bestätigt gilt (auch als konkurrente Validität bezeichnet). Verfahren der Schmerzmessung für auskunftsfähige Populationen ziehen als solches Kriterium gewöhnlich den „Goldstandard" der Selbstauskunft durch den Betroffenen heran. In nicht-auskunftsfähigen Populationen, für die kein solcher Goldstandard verfügbar ist, werden häufig stellvertretende Schmerzeinschätzungen durch nahestehende Angehörige oder Pflegende als Vergleichskriterium herangezogen. Die Angemessenheit dieses *'silver-standards'* (van Herk et al., 2007) muss jedoch bezweifelt werden (Bergh & Sjöström, 1999). Sjöström (1995) konnte Hinweise darauf finden, dass Pflegekräfte ihren Referenzrahmen zur Einschätzung von Schmerzen ihren professionellen Erfahrungen entsprechend anpassen, und so häufig zu geringeren Schmerzeinschätzungen gelangen als die Patienten selbst (vgl. auch Bergh, Jakobsson & Sjöström, 2008). Daneben gibt es Hinweise darauf, dass Proxys zwar in der Lage sind, das Vorhandensein, nicht jedoch die Intensität von Schmerzen verlässlich einzuschätzen (Cohen-Mansfield & Lipson, 2002). Zwakhalen und Kollegen kommentieren diesen Lösungsvorschlag für das Problem des fehlenden Aussenkriteriums nicht ohne Humor wie folgt:

> If nurses' pain impression was a valid and reliable measure, a more complex behavioural scale to assess pain would become redundant. (Zwakhalen et al., 2006)

Ein alternatives Kriterium für die Validitätsbestimmung stellt die Vorhersage eines zukünftigen Outcomes wie beispielsweise ein reduzierter Merkmalswert nach einer analgetischen Intervention dar (prädiktive Validität). Dabei muss jedoch nicht immer direkt interveniert werden, da Schmerzzustände (z.B. nach akuten Verletzungen oder Operationen) gewöhnlich nicht persistent sind, sondern häufig auch ohne Schmerzmedikation nach einer gewissen Zeit abklingen und so abnehmende Messwerte bedingen sollten (vgl. Jensen, 2003). Auch wenn für ältere Probanden in ruhigen Situationen geringere Schmerzwerte ermittelt werden als bei Aktivierung, wird das mitunter als Hinweis auf die prädiktive Validität des Verfahrens verstanden, da der Übergang von Ruhe zu Bewegung in dieser Population häufig schmerzhaft ist. Sowohl die CNPI als auch die PAINAD geben jeweils eine Ruhe und eine Aktivitätssituation zur Beobachtung vor. Werden in einer realisierten Stichprobe älterer Menschen die erwarteten Unterschiede nicht gefunden, bleibt unklar, ob dieser Befund gegen die Validität des Verfahrens spricht, oder nur die Vorannahmen zur

bewegungsinduzierten Schmerzsteigerung in dieser Stichprobe nicht zutreffen. Auf die Problematik der Abhängigkeit von Merkmalen eines Tests und Merkmalen der Stichprobe wird bei der Diskussion der Limitationen der klassischen Testtheorie im Methodenteil dieser Arbeit (Kapitel 4.1.4) ausführlich eingegangen.

Grundvoraussetzung für die prädiktive Validität ist die Fähigkeit eines Instrumentes, überhaupt Merkmalsveränderungen abbilden zu können. Insbesondere wenn Schmerzmessungen für die Verlaufskontrolle schmerzbezogener Interventionen eingesetzt werden sollen, muss die Änderungssensitivität des Instrumentes nachgewiesen werden. Diese Forderung stellt gerade für die Schmerzmessung auf der Grundlage beobachtbaren Schmerzverhaltens eine außerordentliche Herausforderung dar, da in den vorgeschlagenen Inventaren in der Regel unmittelbare, mittelbare, und langfristige behaviorale Schmerzäußerungen gemischt vorkommen. Diejenigen Instrumente, die beispielsweise neben mimischen, lautsprachlichen oder posturalen Indikatoren unmittelbar erlebten Schmerzes auch längerfristige schmerzbezogene Verhaltensänderungen (z.B. sozialer Rückzug, Veränderung von Gewohnheiten etc.) berücksichtigen, besitzen unter Umständen eine deutlich verzögerte Responsivität und machen damit verbunden zumindest implizit auch Annahmen zur (akuten, persistierenden oder wiederkehrenden) zeitlichen Natur des abbildbaren Schmerzerlebens.

Neben bekannten Verläufen der Schmerzbelastung bei Erkrankung und Genesung kann auch ein querschnittlicher Vergleich der Schmerzbelastung in bestimmten Gruppen (z.B. chronische Schmerzpatienten, postoperative Patienten) als Validitätskriterium herangezogen werden. Erlaubt ein Verfahren beispielsweise die Differenzierung von Personengruppen mit a priori als unterschiedlich angenommenen Merkmalswerten anhand der Skalenwerte, so spricht man von der diskriminanten Validität des Verfahrens. Für die PADE wurden beispielsweise keine unterschiedlichen Scores bei Personen mit und ohne schmerzkorrelierte Erkrankung erzielt. Personen, die als klinisch schmerzbelastet galten erreichten dagegen höhere Werte (Villanueva et al., 2003). Mit der CNPI wurden nach einem chirurgischen Eingriff höhere Schmerzwerte ermittelt als davor (Feldt, 2000). Auch die Möglichkeit einer Skala, schmerzhafte, ruhige, und anstrengende Tätigkeiten von Probanden voneinander zu unterscheiden, gibt Hinweise auf dessen diskriminative Validität. Entsprechende Nachweise wurden beispielsweise für CNPI, FACS, PBM, DS-DAT, PAINAD und PACSLAC angeführt (für einen Überblick siehe van Herk et al., 2007).

Schwierig erscheint die in den meisten Reviewarbeiten gewählte Mischung aus Beurteilungskriterien, die sich zum einen auf die Anlage der Instrumentenentwicklung und empirischen Testung (z.B. die Stichprobengröße im Verhältnis zur Itemanzahl), zum zweiten auf Merkmale des Instrumentes selbst (z.B. Anzahl der Items, Inhaltsbereiche oder Durchführung) beziehen. Daneben bleiben selbst bei den in allen Arbeiten dezidiert dargelegten Beurteilungskriterien und einem systematischen Verfahren der Konsensbildung zwischen mehreren Beurteilern natürlich immer auch individuelle Ermessensspielräume bestehen.

Eine weitere fundamentale Einschränkung der bisher geleisteten Skalenvalidierung liegt darin, dass in den gegenwärtigen Übersichtsarbeiten Reliabilität und Validität weit-

gehend nebeneinander betrachtet, jedoch nicht systematisch aufeinander bezogen werden. Validität kann aber überhaupt nur für hinreichend reliable Maße nachgewiesen werden, da eine Verschätzung wahrer Merkmalswerte auch entsprechende Unterschieds- und Zusammenhangsanalysen zur Validitätsbestimmung kompromittiert.

Da die Abschätzung von Reliabilität und Validität zu großen Teilen auf Zusammenhangsanalysen beruht, kann ferner auch das Scoring der Einzelindikatoren bzw. die Skalierung des Gesamtmaßes die abgebildete Merkmalsvariabilität und damit die Möglichkeit zur Kovariation bzw. Korrelation mit anderen Merkmalen beeinflussen.

Zusammenfassend kann die Validität eines Verfahrens also nicht durch eine einzelne Studie belegt, sondern lediglich aus der Stimmigkeit der in verschiedenen Studien mit unterschiedlichem Zweck und Design gewonnenen Informationen abgeschätzt werden. In Anbetracht der vielen verschiedenen Instrumente, aber jeweils nur wenigen Originalarbeiten, in denen diese eingesetzt wurden (siehe Stolee et al., 2005; van Herk et al., 2007), erscheint eine Beurteilung der Validität einzelner Verfahren gegenwärtig noch überhaupt nicht möglich.

Insbesondere beklagen beispielsweise Zwakhalen und Kollegen, dass in den letzten Jahren zu viele Verhaltensinventare neu entwickelt wurden, anstatt die bestehenden Verfahren im Praxisalltag hinreichend zu prüfen.

> It is the researchers' as well as the funding agencies' and journals' responsibility to prevent excessive growth of newly developed tools. Thus, further psychometric evaluation of existing scales should be given priority over developing new scales for future use. (Zwakhalen et al., 2006)

Dem soll im Rahmen dieser Arbeit entgegengehalten werden, dass bei einer Fortführung der gegenwärtig dominierenden en-bloc-Beurteilung auch die reliablen und validen Indikatoren und Kategorien eines Verfahrens wieder verloren gehen könnten. Es wäre dagegen von Vorteil, die Stärken verschiedener Instrumente zu kombinieren. Mit der vorliegenden Arbeit wird die hier aufscheinende sehr simple dichotome Logik eines entweder guten oder schlechten Verhaltensinventares um notwendige Zwischentöne erweitert, indem den eingeschränkten Möglichkeiten der psychometrischen Beurteilung der klassischen Testtheorie eine deutlicher an den behavioralen Inhalten der vorgeschlagenen Inventare ausgerichtete und den Erfassungskontext angemessener berücksichtigende probabilistische Perspektive gegenübergestellt wird.

In ihrer Gesamtbeurteilung der bisher verfügbaren Verhaltensinventare zur Schmerzbeobachtung bei nicht-auskunftsfähigen Personen gelangen alle Übersichtsarbeiten zu dem Schluss, dass "[. . .] most observational scales are still under development and show moderate pychometric properties. Most scales lack validity, reliability and clinical usefulness." (Zwakhalen, Hamers & Berger, 2006, p. 211)

Dennoch halten die meisten Autorengruppen mindestens eines der beurteilten Verfahren für vielversprechend, und fordern deren weitere psychometrische Überprüfung. Den insgesamt größten, wenn auch zurückhaltenden Zuspruch erhalten dabei die Skalen

DOLOPLUS-2, PACSLAC und PAINAD (Hadjistavropoulos et al., 2007; Schofield et al., 2005; van Herk et al., 2007; Zwakhalen, Hamers, Abu-Saad & Berger, 2006).

3.4.3 Neuentwicklungen

Zwischenzeitlich wurden weitere Verfahren neu konstruiert oder weiterentwickelt, die in den angeführten Übersichtsartikeln noch nicht berücksichtigt werden konnten. Im einzelnen sind dies die zuvor bereits in Tabelle 2 aufgeführten Skalen MOBID, EPCA-2 und MPS. Wie die bisherigen Skalen orientiert sich die Zusammenstellung und Überprüfung der Neuentwicklungen dabei ebenfalls weitgehend am Konzept und der Methodik der klassischen Testtheorie, so dass manche der angeführten prinzipiellen Probleme einer integrativen Bestimmung der psychometrischen Güte von Verfahren zur Schmerzbeobachtung auch zukünftig fortbestehen werden.

Dennoch lässt sich eine Tendenz erkennen, die wechselseitigen Beziehungen zwischen den Items und spezifische Erwartungen für bestimmte Verhaltensweisen in bestimmten Patientengruppen oder Situationen stärker in den Blick zu nehmen.

3.4.3.1 Standardisierte Mobilisation – MOBID Schmerzskala

Ein konzeptionell neuartiges und für die vorliegende Arbeit besonders relevantes Verfahren verhaltensgestützten Schmerzassessments, die *Mobilization-Observation-Behavior-Intensity-Dementia Pain Scale* (MOBID), wurde in jüngster Zeit von Husebo und Kollegen (2007, 2009) vorgestellt. Ausgehend von der Annahme, dass ältere Menschen häufig an bewegungsinduzierten Schmerzen leiden, die beispielsweise aufgrund von Vermeidungsstrategien durch herkömmliche Verhaltensinventare in Alltagssituationen unerkannt bleiben, schlagen die Autoren die Beobachtung schmerzbezogenen Ausdrucksverhaltens bei einer Sequenz von fünf standardisierten Mobilisationen verschiedener Körperteile vor. Dabei gruppieren sie das zu beobachtende schmerzbezogene Ausdrucksverhalten in die Kategorien *Schmerzlaute*, *Gesichtsausdruck* und *Abwehr*, wobei für jeden Ausdrucksbereich vier beispielhafte Verhaltens- oder Ausdrucksweisen angeführt werden. Beobachtet wird der Schmerzausdruck während der Patient die Hand öffnet (beide Hände nacheinander), die Arme in Richtung Kopf ausstreckt (beide Arme nacheinander), die Beine streckt und anwinkelt (beide Beine nacheinander), sich im Bett auf beide Seiten dreht, und sich an der Bettkante aufsetzt. Die ursprüngliche Version der Skala berücksichtigte daneben zusätzlich die beiden Beobachtungssituationen Ruhe und Mund- bzw. Zahnpflege, die aufgrund der geringen empirischen Übereinstimmung mit den restlichen Beobachtungssituationen jedoch aus dem endgültigen Instrument herausgenommen wurden. Die einschätzende Pflegekraft führt und unterstützt alle Bewegungen bis zur Vollständigen Extension bzw. Deklination, sofern keine Schmerzäußerungen erkennbar sind. Neben der Beobachtung von Schmerzlauten, Schmerzmimik und schmerzbezogenen Defensivreaktionen schätzt der Beobachter die vermutliche Intensität des bei einer Mobilisation jeweils vom Patienten erlebten Schmerzes auf einer elfstufigen NRS (0-10 Punkte) ein. Zum Ab-

schluss der Erfassung wird mit dem gleichen Skalenformat eine Gesamtbeurteilung der Intensität der erlebten Schmerzen erfragt.

Entwickelt wurde das Verfahren in einer norwegischen Stichprobe von 26 kognitiv beeinträchtigten Altenheimbewohnern (MMST: 4,3±4,3) mit chronischer Schmerzbelastung. Die strukturierte MOBID-Einschätzung erfolgte direkt im Anschluss an die morgendliche Grundpflege durch insgesamt 11 Pflegende und wurde für anschließende Nachratings durch die Pflegenden selbst und drei externe Personen per Video aufgezeichnet. Als Indikator für die Reliabilität des Verfahrens wurden die Intensitätsscores aller sieben Beobachtungssituationen auf ihre interne Konsistenz überprüft. Da die Ruhesituation und die Mundhygiene geringe Korrelationen mit dem Rest der Skala aufwiesen, wurden diese Items entfernt. Die Interne Konsistenz der verbleibenden fünf Items betrug für die Einschätzung durch drei externe Rater zwischen α=.90 und .91. Bemerkenswert ist dabei, dass das durch die MOBID erfasste Konstrukt nunmehr die über verschiedene Mobilisationssituationen hinweg abgebildete *bewegungsinduzierte* Schmerzbelastung der Bewohner ist, und nicht – wie in allen zuvor berichteten Verfahren – die über verschiedene schmerzbezogene Verhaltensweisen in einer bestimmten Beobachtungssituation abgebildete Schmerzbelastung. Insofern erscheint es aber auch nicht verwunderlich, dass der in der Ruhesituation beobachtete Schmerz (=Ruheschmerz) nur gering mit dem Rest der Skala korreliert. Problematisch erscheint dagegen der logische Sprung von *Verhaltens*indikatoren zu *Situations*indikatoren. Selbst wenn die Kategorien des zu beobachtenden Schmerzverhaltens wie hier recht breit definiert sind, bleibt dennoch fraglich, ob alle drei jeweils zugrundegelegten Indikatorkategorien den bewegungsinduzierten Schmerz in allen sieben Bewegungssituationen ähnlich (gut) abzubilden in der Lage sind. Wird der Bewohner beispielsweise in seinem Bett auf die Seite gedreht, könnten mimische Schmerzreaktionen unter Umständen weniger gut beobachtet werden, da auch das Gesicht vom Pflegenden abgewandt wird. Auch abwehrende Reaktionen könnten bei verschiedenen Mobilisationsversuchen unterschiedlich gut möglich oder offensichtlich sein. In einer Ruhesituation ohne soziale Interaktion oder pflegerische Intervention hätte die Verhaltenskategorie Abwehr von vorneherein als wenig sinnvoll erachtet werden können. Diese Unsicherheiten erschweren eine eindeutige Interpretation der durch die Autoren berichteten Befunde, dass bei der Bewegung der Arme durchschnittlich mehr schmerzbezogene Verhaltensweisen beobachtet werden konnten als bei der Bewegung der Beine, oder in liegender Position. Auf das damit implizierte Problem der Invarianz einer Messstruktur über verschiedene Beobachtungszeitpunkte oder -situationen hinweg wird im methodischen Teil dieser Arbeit (Kapitel 4.3.3.5) detailliert eingegangen.

Eine weitere Herausforderung des durch Husebo und Kollegen vorgeschlagenen Ansatzes stellt die Erklärung bzw. der Nachweis des Zusammenhanges zwischen beobachteten Verhaltensindikatoren und der Einschätzung der Schmerzintensität dar. Auch die Arbeitsgruppe um Husebo geht von der Annahme aus, dass ein intensiveres Schmerzerleben mit stärker ausgeprägtem, länger andauerndem, und/oder vielfältigerem Verhaltensausdruck einhergeht. Tatsächlich können für alle fünf in der Endfassung des Instrumentes berücksichtigten Beobachutngssituationen bzw. Mobilisationen entsprechende Trends

höherer eingeschätzter Schmerzintensität bei Beobachtung von Verhaltensindikatoren aus mehreren Ausdrucksbereichen nachgewiesen werden. Andererseits räumen die Autoren ein, dass "Transforming pain behavior indicators into pain is an individual process, reflecting the observers' experience and attention. High reliability of pain intensity seems, however, to be based on the overall concept of the behavior indicators, rather than the presence of it." (Husebo et al., 2007, p. 76).

Eine neuere Arbeit der MOBID-Autorengruppe zur Reliabilität der Verhaltensbeobachtung und Intensitätseinschätzung auf der Grundlage des in der zuvor beschriebenen Studie gewonnenen Videomaterials erscheint für die vorliegende Arbeit weniger wegen der berichteten Befunde, sondern hauptsächlich wegen der testtheoretisch anspruchsvolleren Auseinandersetzung mit dem Konzept der (Retest-)Reliabilität der MOBID-Einschätzungen interessant (Husebo et al., 2009).

Wie zuvor bereits angesprochen, ist für ein Instrument zur Schmerzmessung zu fordern, dass es zu konsistenten Aussagen über beispielsweise die Intensität erlebter Schmerzen dann gelangt, wenn sich das Schmerzerleben über zwei mehrere Erfassungzeitpunkte nicht verändert (=Reproduzierbarkeit), andererseits aber tatsächliche Schmerzveränderungen durch unterschiedliche Schmerzscores abzubilden erlaubt (=Responsivität). Dabei ist die Reproduzierbarkeit eines Skalenwertes eine Voraussetzung dafür, Merkmalsveränderungen abbilden zu können (Beckerman et al., 2001; van Baalen et al., 2006).

Husebo und Kollegen (2009) berichten für die durch externe Rater zu insgesamt drei Zeitpunkten (Re-test nach 4 und 8 Tagen) eingeschätzten MOBID-Schmerzintensitäten jeweils drei verschiedene Reproduzierbarkeits-Indizes.

Der *Intra-Klassen-Korrelationskoeffizient (ICC)* wird dabei als Marker für die relative Reliabilität der MOBID-Einschätzungen herangezogen. Unterschiede in den zu zwei verschiedenen Zeitpunkten (nachträglich) durch einen Rater eingeschätzen Schmerzwerten der 26 beobachteten Bewohner (within-subject-Varianz) werden dabei als Fehlerkomponente begriffen und an der Gesamtunterschiedlichkeit der Messwerte relativiert.

Daneben zeigen die Autoren, dass die individuellen *Standardmessfehler (Standard Error of Measurement, SEM)* der einzelnen Rater als Maß der absoluten Abweichung der beiden Messwertreihen voneinander (within-subject standard deviation, s_w) mit steigender Erfahrung in der MOBID-Einschätzung kleiner werden.

Mit einer gesteigerten Reproduzierbarkeit der MOBID-Werte aber können Schmerzveränderungen z.B. im Verlauf einer medikamentösen Intervention besser abgebildet werden. Als ein Maß für die *Responsivität* der MOBID-Skala berechnen die Autoren auf der Grundlage der für jeden Rater individuell geschätzten SEMs die kleinste Schmerzdifferenz, die mit MOBID nachgewiesen werden kann. Husebo bezeichnen diesen Index als *Smallest Detectable Difference (SDD),* der aber auch als *Smallest Real Difference (SRD)* bekannt ist (Beckerman et al., 2001). Die Ergebnisse belegen, dass die tatsächlichen Veränderungen in der Intensität bewegungsinduzierter Schmerzen, die mit MOBID identifiziert werden können, für verschiedene externe Rater, in verschiedenen Beobachtungssituationen und in Abhängigkeit von der Expertise bzw. Erfahrung des Beobachters unterschiedlich groß sind. Damit aber wird auf ein erweitertes Verständnis des Re-

liabilitätskonzeptes verwiesen, bei dem Merkmalen des Erfassungskontextes im Gegensatz zum Reliabilitätsbegriff der klassischen Testtheorie ein systematisches Gewicht eingeräumt wird. Die Grundzüge einer solchen Weiterentwicklung testtheoretischer Grundannahmen im Kontext der Generalisierungstheorie werden detailliert im Methodenteil dieser Arbeit (Kapitel 4.1.3) erläutert.

3.4.3.2 Schmerz versus Agitation – *Mahoney Pain Scale*

Eine weitere interessante Neuentwicklung stellt auch die von Mahoney und Peters (2008) vorgestellte *Mahoney Pain Scale* (MPS) dar. Die konzeptionelle Neuerung ist bei dieser Arbeit darin zu sehen, dass der schmerzbezogene Verhaltensausdruck in einer konkreten Beobachtungssituation mit dem Verhalten verglichen wird, das der Bewohner üblicherweise zeigt. Ist das normale Verhalten des Bewohners durch ein hohes Maß an Agitiation bestimmt, könnten auch die verwendeten Schmerzindikatoren zu einem großen Teil durch diese nicht-kognitive Demenzsymptomatik bestimmt sein. Eine detaillierte Diskussion des Potenzials dieses innovativen Ansatz der Schmerzmessung bei Demenz wird in Kapitel 3.4.7.3 ab Seite 78 geführt.

3.4.4 Verhaltensinventare für den deutschen Sprachraum

Im deutschsprachigen Raum sind gegenwärtig nur wenige der zuvor besprochenen Verfahren verfügbar (PAINAD, ECPA, DOLOPLUS-2). Offizielle Übersetzungen ins Deutsche wurden bislang lediglich für die ursprünglich französische ECPA und die in dieser Arbeit diskutierte PAINAD erarbeitet und empirisch überprüft. In einigen Einrichtungen werden jedoch bereits selbst übersetzte, inoffizielle deutsche Versionen beispielsweise von ECPA oder DOLOPLUS-2 zur Ergänzung des Schmerzmanagements verwendet.

3.4.4.1 Die Skala BISAD

Eine von den ursprünglichen Autoren offiziell authorisierte deutsche Version der ECPA wird gegenwärtig im Rahmen einer Dissertationsarbeit an der Berliner Charité unter der Bezeichnung *Beobachtungsinstrument für das Schmerzassessment bei alten Menschen mit Demenz (BISAD)* erstellt und evaluiert (Fischer, 2005, 2007).

Inhaltlich erfasst die BISAD schmerzbezogenen Verhaltensausdruck vor und während einer pflegerischen Mobilisation. Die vor der Mobilisation einzuschätzenden Verhaltenskategorien umfassen den Gesichtsausdruck (Blick und Mimik), die spontane Ruhehaltung bzw. die Suche nach einer schmerzfreien Schonhaltung, Bewegungen bzw. Mobilisation der Person (innerhalb und/oder außerhalb des Bettes), sowie Merkmale der Beziehung zu Anderen (mittels Blicken, Gesten oder verbalem Ausdruck). Während der Mobilisation sollen Verhaltensweisen aus den Kategorien Ängstliche Erwartung bei der Pflege, Reaktion während der Mobilisation, Reaktion während der Pflege der schmerzenden Bereiche, sowie während der Pflege vorgebrachte Klagen eingeschätzt werden. Alle acht Verhaltenskategorien werden mit 0 bis 4 Punkten gescort, wobei höhere Werte auf eine

stärkere Intensität vorliegender Schmerzen hinweisen. Innerhalb jeder Verhaltenskategorien sind konkrete Verhaltensindikatoren genannt, die unterschiedliche Schmerzintensitäten repräsentieren, also als theoretisch unterschiedlich *schwierig* angenommen werden. Die für die Indikatorbereiche Bewegung und Beziehung zu Anderen beschriebenen Verhaltensweisen (z.B. Person bewegt sich wie gewohnt oder Kontakt ist schwerer herzustellen als gewohnt) sollen im Abgleich mit dem (gewöhnlichen) Verhalten des Bewohners am/an den vorangegangenen Tagen eingeschätzt werden. Als Kennwert für die Schmerzbelastung wird der Gesamtscore aller acht Items aus der Ruhe- und Mobilisations- bzw. Pflegesituation gebildet.

Die Übersichtlichkeit (8 Items), die Berücksichtigung von sowohl Ruhe- und Pflege- bzw. Mobilisationssitationen, und der Bezug zum üblichen Verhalten des Bewohners sind sicherlich als konzeptionelle Stärken dieses Verfahrens zu werten. Dennoch wirft die konkrete Umsetzung aller drei Strukturmerkmale auch Fragen auf.

Da für jeden der acht Indikatorbereiche jeweils fünf konkrete Verhaltensweisen beschrieben sind, müssen für die Bearbeitung der Skala genaugenommen 40 Verhaltensindikatoren hinsichtlich ihres Vorliegens eingeschätzt werden. Die Angemessenheit des Itemscorings selbst erscheint nicht durchgängig auch augenscheinvalide, beispielsweise wenn für *Reaktion während der Pflege, nicht darüber hinausgehend* ein Punkt, für *Reaktion auf Anfassen der schmerzenden Bereiche* dagegen 2 Punkte vergeben werden. Daneben is die Zuordnung einzelner Verhaltensweisen (z.B. *ängstlicher Blick, ängstlicher Eindruck* bzw. *Person blickt angespannt und scheint Mobilisation und Pflege zu fürchten* zu bestimmten Kategorien (Ängstliche Erwartung bei Pflege und/oder Reaktion während Pflege) unklar.

Eine Schwierigkeit bei der Interpretation des BISAD-Scores ergibt sich aus der Verrechnung der ruhe- und mobilisationsbezogenen Beobachtungen, selbst wenn unter beiden Bedingungen jeweils unterschiedliche Items eingeschätzt wurden und so die Frage der Äquivalenz des Instrumentes gar nicht erst aufkommt. Der erfasste Schmerzwert beschreibt demnach die Schmerzbelastung der Bewohner nicht mehr situationsspezifisch, womit ein wichtiger Teil der erfassten Information zu potenziellen Auslösern des Schmerzes wieder verloren zu gehen droht.

Während Veränderungen in den Sozialbeziehungen auch von der AGS gefordert werden, sind längerfristige Veränderungen der Mobilität bislang seltener berücksichtigt worden. Der angegebene Vergleichszeitraum ist mit dem vorangegangenen Tag bzw. den vorangegangenen Tagen unklar definiert. Die gewählte Beschränkung auf einen so kurzen Zeitraum kann dazu beitragen, langfristige Folgen von Schmerzen zu vermeiden; dennoch wird es schwerer fallen, Verhaltensänderungen über so kurze Vergleichsintervalle als hinreichend stabile Schmerzfolgen auszuweisen bzw. gegen nicht-schmerzbezogene (Tages-)Schwankungen abzugrenzen.

Eine Veröffentlichung der Arbeitsergebnisse, die eine auch empirisch stärker informierte Einschätzung der Potenziale der BISAD für den deutschen Versorgungskontext erlauben werden, ist für die kommenden Monate im Verlag Hans Huber vorangekündigt (Fischer, 2009).

3.4.4.2 Die Skala DOLOPLUS-2

Wie die ECPA wird auch das im französischsprachigen Raum (Frankreich, Kanada) ebenfalls weit verbreitete Instrument DOLOPLUS-2 gegenwärtig als informelle Übersetzung in einzelnen deutschen Pflegeheimen eingesetzt. Allerdings gibt es zur Zeit keine offizielle deutsche Fassung und es sind auch keine publizierten Befunde zur Testung einer deutschsprachigen Version verfügbar (Stand 10.11.2008 Datenbanken CINAHL, EMBASE, MedLine und CareLit; Fischer, 2008).

3.4.4.3 Die Skala ZOPA©

Am Zentrum für Entwicklung und Forschung in der Pflege (ZEPF) des UniversitätsSpitals Zürich wurde in Kooperation mit zwei Masterstudentinnen des Institutes für Pflegewissenschaft der Universität Witten/Herdecke zwischen 2002 und 2007 ein neues Beobachtungsinstrument zur Schmerzeinschätzung bei nicht verbal auskunftsfähigen Patienten erarbeitet (s. Gnass & Sirsch, 2007).

Das Instrument ZOPA© (*Zurich Observation Pain Assessment*) enthält in seiner gegenwärtigen Fassung 13 diskrete Verhaltensmerkmale, die den vier Ausdruckskategorien Lautäußerung (Stöhnen/Klagen, Brummen), Gesichtsausdruck (verzerrter Gesichtsaudruck, Zähne zusammenpressen/auf Tubus beißen, Augen zusammenkneifen, starrer Blick, Tränenfluss), Körpersprache (Ruhelosigkeit, Massieren oder Berühren eines Körperteiles, angespannte Muskeln) und physiologische Schmerzmerkmale (Veränderung in den Vitalzeichen: Blutdruck/Puls, Atmung; Veränderungen in der Gesichtsfarbe: Schwitzen/Röte) zugeordnet sind (Sirsch, 2009).

Bei der Entwicklung und Überprüfung des Instrumentes lag der Fokus allerdings nicht ausschließlich auf demenzkranken Menschen, was auch an der Zusammenstellung der Items, insbesondere dem Hinweis auf mögliche Intubation und der großen Zahl physiologischer Indikatoren, deutlich wird. Zumindest für den Bereich der klinischen Versorgung im Akutkrankenhaus beansprucht das Instrument Gültigkeit und Nützlichkeit jedoch auch für die Gruppe kognitiv beeinträchtiger älterer Menschen.

Das Instrument ZOPA© wird gegenwärtig auf den Bettenstationen der Neurochirurgie, Neurologie, sowie den Intensivstationen des UniversitätsSpitals Zürich routinemäßig eingesetzt. Für den September 2009 wurde eine Buchpublikation zum Instrument (Bern: Verlag Hans Huber & Hogrefe) in Aussicht gestellt.

3.4.4.4 Die Skala BESD

Die PAINAD-Skala wurde vom Arbeitskreis Schmerz und Alter der Deutschen Gesellschaft zum Studium des Schmerzes (DGSS) als *Skala zur Beurteilung von Schmerzen bei Demenz* (BESD) ins Deutsche übersetzt (Basler et al., 2006). Die Ergebnisse eines ersten Praxiseinsatzes im Rahmen einer multizentrischen Studie mit insgesamt 150 demenzkranken Heimbewohnern, an der der Autor der vorliegenden Arbeit wesentlich beteiligt war,

sollen als Vorarbeiten verstanden werden und sind im Kapitel zur Durchführung der empirischen Studie (Kap. 5.3.4.2) im Detail dargestellt (vgl. auch Basler et al., 2006; Schuler, 2008; Schuler et al., 2007).

3.4.5 Vergleich mehrerer Verhaltensinventare

Eine Reihe neuerer empirischer Arbeiten setzte mehrere Skalen zur Verhaltensbeobachtung parallel ein, so dass die Eigenschaften der Inventare unmittelbarer miteinander verglichen werden konnten (Cohen-Mansfield, 2008; Cohen-Mansfield & Lipson, 2008; Zwakhalen, Hamers & Berger, 2006).

In einer niederländischen Studie (N=12 gerontopsychiatrische Stationen, N=144 Pflegeheimbewohner; n=128 kognitiv Beeinträchtigte) überprüften Zwakhalen und Kollegen (2006) die psychometrischen Eigenschaften von PAINAD, PACSLAC und DOLOPLUS-2. Schmerzbeobachtungen wurden dabei in einer Ruhesituation (T1, n=128) und zwei potenziellen Schmerzsituationen parallel von einer Pflegekraft (n=12, davon 5 mit Pflegeexamen) und einem Rater des Autorenteams (n=1) durchgeführt. Die erste potenzielle Schmerzsituation (T2, Grippeimpfung subkutan, n=127) wurde am gleichen Tag wie T1 durchgeführt. Das Beobachtungsintervall war für die Situationen T1 und T2 jeweils 2 Minuten, die drei Beobachtungsinstrumente wurden direkt im Anschluss an das Beobachtungsintervall ausgefüllt (Zufallsreihenfolge). Eine weitere bewohnerspezifische potenzielle Schmerzsituation (Waschen, Transfer, Mobilisation, Wundversorgung) wurde für n=35 Bewohner zu einem späteren Zeitpunkt (größtenteils innerhalb von 3 Wochen nach T2) durchgeführt. Bewohner, für die nach der Einschätzung der Pflegekräfte keine Schmerzsituationen erwartet werden konnten, wurden zu T3 nicht mehr berücksichtigt. Da DOLOPLUS-2 nicht zur in-situ Schmerzbeobachtung, sondern zur retrospektiven Beurteilung der Schmerzveränderung eingesetzt wurde, soll auf deren Beschreibung hier verzichtet werden. Zum Abschluss der Studie schätzten die beteiligten Pflegenden die Praktikabilität bzw. klinische Nützlichkeit der verschiedenen Verhaltensinventare ein.

Die Darstellung der Beobachtungsraten einzelner Items der PAINAD und PACSLAC beschränkt sich auf die beiden Schmerzsituationen (T2 und T3) und berücksichtigt nur die Gruppe der als jeweils schmerzbelastet eingeschätzten Bewohner (Einschätzung der Autoren, Cut-off-Wert VAS $\geq$30, n=53 Beobachtungen). In dieser Schmerzgruppe wurden 28 der 60 PACSLAC-Items von beiden Ratern in weniger als 10 Prozent der potenziellen Schmerzsituationen beobachtet. Die Autoren verbinden mit einer geringen Auftretensrate zumindest implizit eine mangelnde Indikationsgüte, auch wenn sie zurückhaltend interpretieren, dass "[...] these items could be 'less likely' candidates for inclusion in a general pain instrument for the target group of elderly people with dementia." (Zwakhalen, Hamers & Berger, 2006, p. 214, Hervorhebungen im Original).

Zu den sowohl während der Impfung als auch der Pflegesituation häufig (d.h. bei jeweils mindestens zwei Dritteln der Bewohner) beobachteten Einzelindikatoren zählen beispielsweise *angespanntes Gesicht, Veränderungen in den Augen/Blick, düsterer Blick, Stirnrunzeln, Ausdruck von Schmerz,* oder *schmerzspezifische Lautäußerung.* Für den

weitaus überwiegenden Teil der durch Zwakhalen und Kollegen näher beschriebenen 20 PACSLAC-Items sind die Beobachtungsraten in den beiden Situationen allerdings sehr verschieden. Dabei können beispielsweise höhere Beobachtungsraten der Items *möchte nicht berührt werden* (43,5% vs. 28,6%), *ängstlich* (39,1% vs. 21,4%) oder *starre Körperhaltung* (30,4% vs. 14,3%) in der bewohnerspezifischen Pflegesituation gegenüber der Impfsituation als Hinweis auf systematische Unterschiede der Situationen hinsichtlich Aktiviertheit/Bewegung und Vorerfahrungen/Antizipation gewertet werden. Eine Diskussion der in diesen Beobachtungssituationen unterschiedlichen Beobachtungsraten und der a priori erwartbaren Verhaltensweisen (z.B. Stillhalten bei Impfung, (Mit)Bewegen bei Pflegehandlungen) bleiben die Autoren allerdings schuldig. Auf die möglicherweise unterschiedliche Angemessenheit verschiedener Indikatoren in den unterschiedenen Situationen wird auch bei den Empfehlungen für die Optimierung des Verhaltensinventars kein Bezug genommen. Stattdessen wird prinzipiell die Eliminierung wenig trennscharfer und selten beobachteter Items vorgeschlagen, um die interne Konsistenz zu erhöhen. Unberücksichtigt bleiben dabei auch die Folgen einer Streichung selten beobachteter und damit schwieriger Items für den Bereich von Schmerzintensitäten, die durch die PACSLAC schließlich abgebildet werden können.

Für die PAINAD werden Beobachtungsraten für die fünf Einzelitems lediglich dichotomisiert (0=kein Schmerz, 1+2=Schmerz) angegeben, d.h es wurde nicht differenziert, wie häufig innerhalb der Indikatorbereiche Mimik, Körpersprache, Negative Lautäußerung, Atmung und Trost Verhaltensweisen beobachtet wurden, die mit 0, 1, oder 2 Punkten zu werten waren. Schmerzbezogene Verhaltensweisen, die mit den in den PAINAD-Kategorien genannten Beispielitems identisch oder vergleichbar waren, wurden im Bereich der Mimik in 92,9 bzw. 95,7 Prozent, im Bereich der Körpersprache in 89,2 bzw. 95,7 Prozent und im Bereich negativer Lautäußerung in 75,0 bzw. 92,3 Prozent der Impf- bzw. Pflegesituationen beobachtet. Die Schmerzindikatoren aus dem Bereich Atmung wurden dagegen in jeweils weniger als 20 Prozent der Schmerzsituationen beobachtet. Keine differenzierten Angaben machen die Autoren dagegen zum Indikatorbereich Tröstbareit, was bei genauerer Betrachtung daran liegen muss, dass dieses Item de facto eine Minimalintervention immer dann vorsieht, wenn der Beobachter *den Wunsch zu trösten verspürt.* Das Ergebnis dieser Intervention wird entsprechend im Itemscore mit einem (Bewohner kann getröstet/abgelenkt werden) oder zwei Punkten (Bewohner kann nicht getröstet werden) festgehalten. Damit kann PAINAD aber nur in situ, nicht aber retrospektiv im Anschluss an eine Beobachtung eingeschätzt werden. Beim in der beschriebenen Studie gewählten Prozedere muss also kritisch hinterfragt werden, wie die (für weniger augenfällige Folgeanalysen dann doch berücksichtigten) Beurteilungen dieses Items überhaupt zustande kamen.

Insgesamt erscheinen die auf einzelne Verhaltensindikatoren abgestellten Analysen für beide Skalen auf den zweiten Blick doch weit weniger konsequent verfolgt worden zu sein, als möglich gewesen wäre. Insbesondere den Übereinstimmungen und Differenzen in den inhaltsgleichen Verhaltenskategorien beider Verhaltensinventare hätten die Autoren – trotz des unterschiedlichen Skalenaufbaus – genauer nachgehen können.

Alle weiteren berichteten Analysen setzen PACSLAC und PAINAD dagegen auf der Ebene des Gesamtskalenscores zueinander in Beziehung. Aufgrund der deutlichen Unterschiede in Testlänge und Scoring bleiben die Möglichkeiten eines direkten Vergleichs beider Skalen jedoch auf die Rangplätze der erhobenen Schmerzwerte beschränkt. Die Korrelation zwischen PAINAD und PACSLAC wird mit Pearsons r=.85 berichtet, wobei unklar bleibt, auf welche Situationen und Beobachter sich dieser Wert bezieht.

Die klinische Nützlichkeit von PACSLAC wurde durch die beteiligten Pflegenden als durchschnittlich höher eingeschätzt als die der PAINAD (7,0$\pm$0,5 vs. 5,9$\pm$1,7 Punkte auf einer 11-stufigen NRS).

Zusammenfassend leistet die durch Zwakhalen und Kollegen vorgelegte Arbeit hinsichtlich des Vergleiches konkurrierender Verhaltensinventare zur Schmerzmessung – abgesehen von der Korrelation beider Skalenscores und der Einschätzung der klinischen Nützlichkeit der Skalen – kaum einen Beitrag, der deutlich über denjenigen separater Studien hinausgeht. Diesem vergleichsweise kleinen Mehrgewinn stehen jedoch nicht unerhebliche konzeptionelle und praktische Herausforderungen, beispielsweise mit Blick auf die Trostintervention bei PAINAD gegenüber, die eine Vergleichbarkeit mit bisherigen Forschungsarbeiten erschweren.

Cohen-Mansfield und Lipson verglichen in ihrer kürzlich erschienenen Arbeit neben verschiedenen Instrumenten zur Selbst- und Fremdauskunft auch die Beobachtungsskalen PAINAD, CNPI und das bereits etwas ältere OPBAI hinsichtlich ihrer Möglichkeiten miteinander, Schmerzen zu identifizieren und die Wirkung von Schmerzmedikation abzubilden (Cohen-Mansfield & Lipson, 2008). Weder für die Schmerzeinschätzung zur Baselineerhebung, noch für die Identifikation unterschiedlicher Schmerzentwicklungen in medikamentös behandelten vs. unbehandelten Schmerzpatienten über einen zweiwöchigen Zeitraum hinweg erschienen die eingesetzten Verhaltensinventare besonders geeignet. Leider berichten die Autoren keinerlei psychometrische Kriterien zu den Verhaltensinventaren und stellen keinen Vergleich der verschiedenen Beobachtungsinstrumente an. In einer neueren Arbeit mit dem vielversprechenden Titel “The Relationship Between Different Pain Assessments in Dementia” beschränkt sich Cohen-Mansfield (2008) für den Vergleich der drei berücksichtigten Verhaltensinventare leider auf die Korrelation der Gesamtskalenscores (PAINAD-CNPI: r=.85, PAINAD-OPBAI: r=.88, CNPI-OPBAI: r=.84).

Der Differenzierungsgrad vergleichender Analysen zu den psychometrischen Eigenschaften mehrerer Verhaltensinventare zur Schmerzmessung, die an derselben Stichprobe eingesetzt wurden, muss zusammenfassend gegenwärtig als bemerkenswert gering beurteilt werden.

3.4.6 Kontextbedingungen der Schmerzerfassung

Selbstverständlich ist jede Messung zu einem gewissen Teil auch durch die Merkmale des Erfassungskontextes mitbestimmt. Experimentelle Zugänge zur Schmerzmessung versuchen situative Bedingungen soweit wie möglich zu kontrollieren, um einen möglichst großen Teil der erfassten Variabilität in den Schmerzwerten auf die im jeweiligen Stu-

diendesign berücksichtigten inhaltlichen Faktoren zurückführen zu können. Besteht das Ziel des Schmerzassessments darin, eine möglichst adäquate Abbildung der gewöhnlichen bzw. alltäglichen Schmerzbelastung demenzkranker Menschen zu leisten, muss nicht selten jedoch zwischen der Sicherung der internen und der externen Validität der Erfassung abgewogen werden. Für die Verhaltensbeobachtung von Schmerzen bei demenzkranken Menschen erscheint es dabei besonders wichtig, für die Demenzkranken alltägliche Situationen zu berücksichtigen und die in die direkte Pflege dieser Menschen involvierten Personen zu beteiligen.

Das Erleben von Schmerzen ist für alle Menschen eine kaum zu vermeidende Erfahrung. Für die Untersuchung schmerzbezogener Reaktionen werden häufig jedoch Schmerzzustände auch bewusst hergestellt. In den vorangegangenen Kapiteln wurden bereits einige Verfahrensweisen beschrieben, bei denen das Schmerzerleben selbst durch die kontrollierte Verabreichung mechanischer, thermischer, oder elektrischer nozizeptiver Reize experimentell kontrolliert wurde. Neben diesen vergleichsweise standardisierten Verfahren, wurden jedoch auch alltagsnähere Verfahren der bewussten Schmerzinduktion angesprochen, beispielsweise durch einen Nadelstich im Zuge einer Impfung, das Berühren potenziell schmerzhafter Körperteile, oder die Mobilisation von Körperteilen zum Zwecke der Schmerzbeurteilung oder während einer gewöhnlichen Pflegehandlung. In einer Studie zur Schmerzerfassung bei beatmeten Patienten der Klinik für Intensivmedizin des Inselspitals in Bern wurde auch das intratracheale Absaugen als kontrollierte Schmerzinduktion untersucht (Jeitziner, 2008).

Während man darüber streiten kann, ob das bewusste Auslösen von Schmerzen zum Zwecke einer Schmerzmessung für eine angemessene pflegerische Versorgung nötig und ethisch vertretbar ist, lassen sich einige Situationen, die bei älteren und demenzkranken Menschen häufig mit Schmerzen verbunden sind, im Pflegealltag nicht vollständig vermeiden. Zu diesen Situationen gehören insbesondere Pflegehandlungen, bei denen die Bewohner mobilisiert (z.B. bei Transfer oder Lagerung) oder an schmerzenden Körperstellen berührt werden müssen (z.B. bei der Wundversorgung).

3.4.6.1 Aktivität

Aufgrund der im höheren Lebensalter häufigen degenerativen Veränderungen des Bewegungsapparates und der damit verbundenen nozizeptiven Schmerzen bei Bewegung und Aktivität sehen einige der vorgeschlagenen Schmerzassessments die Beobachtung des Bewohners sowohl in Ruhe, als auch in Aktivität oder während der Pflege vor (BESD, CNPI, ECPA, EPCA-2 und BISAD). Der angenommene Zusammenhang zwischen Aktivierung (i.S. von Mobilität oder Herz-Kreislauf-Status) und Schmerzerleben jedoch wird in keinem Verfahren genauer beschrieben. Im Gesamtpool der für die Schmerzbeobachtung mittlerweile eingesetzten Verhaltensindikatoren finden sich Einzelindikatoren, die auf die Transaktion von Verhalten und Erleben hinweisen. Sowohl Aktivität (z.B. *Massage eines bestimmten Körperteiles/-bereiches*) als auch das Fehlen von Aktivität (z.B. *starre Körpersprache* und Vermeidungsverhalten) werden als potenzielle Schmerzindika-

toren berücksichtigt. Zwakhalen, Hamers und Berger (2006) weisen allerdings darauf hin, dass die prinzipielle Annahme einer geringeren Schmerzbelastung älterer Menschen in einer Ruhe- gegenüber einer Aktivitätssituation auch in Zweifel gezogen werden könnte, da ältere Pflegeheimbewohner häufig unter chronischen Schmerzzuständen leiden.

Bei der BESD und der CNPI sollen in beiden Beobachtungssituationen dieselben schmerzbezogenen Verhaltensweisen beobachtet werden. Bei der zuvor bereits näher beschriebenen MOBID werden fünf diskrete Bewegungsabläufe beobachtet und anhand derselben Verhaltenskategorien eingeschätzt. Bei der Skala NOPPAIN, die speziell für den Einsatz in Pflegesituationen entwickelt wurde, können neun verschieden stark mit Bewegung bzw. Mobilisation verbundene Pflegetätigkeiten differenziert werden, und auch hier sind die beschriebenen Indikatoren jeweils die gleichen.

Für Ruhe- und Pflegesituation unterschiedliche Verhaltensweisen sind dagegen bei der ECPA und den darauf basierenden Verfahren EPCA und BISAD aufgeführt. Dabei sind jedoch nur für die Verhaltenskategorie *Reaktionen während der Mobilisation* auch besonders bewegungsspezifische Verhaltensweisen wie Festhalten, Schonhaltung oder Abwehr beschrieben, während die Verhaltensweisen in den verbleibenden drei während der Mobilisation zu beobachtenden Verhaltenskategorien (*Ängstliche Erwartung*, *Reaktion während Pflege schmerzender Bereiche* und *Klagen*) einen weniger deutlichen Bezug zu Bewegung aufweisen.

Alle Verfahren sehen die Summation der in den unterschiedenen Situationen beobachteten schmerzbezogenen Verhaltensweisen zu einem Gesamtskalenscore vor. Eine inhaltliche Interpretation eventueller Differenzen zwischen verschiedenen Erfassungssituationen (z.B. als Hinweis auf bewegungsinduzierten Schmerz) ist in keinem Instrument angedacht.

Beim wiederholten Einsatz eines Verhaltensinventares in a priori als systematisch unterschiedlich angenommenen Situationen stellt sich die Frage, ob die einzelnen Verhaltensweisen den Schmerz tatsächlich gleich gut anzeigen können. Zum ersten könnte beispielsweise eine Ruhesituation nicht nur durch ein geringeres *Niveau* derselben Schmerzqualität (z.B. bewegungsinduzierten Schmerz) charakterisiert sein, sondern gegebenenfalls auch durch einen *qualitativ* anderen Schmerz (Ruheschmerz). Einzelne Indikatoren (z.B. Schaukeln vs. starre Körperhaltung) könnten besser zur Abbildung einer bestimmten Schmerzqualität beitragen als zur Abbildung anderer Qualitäten. Zum zweiten ist anzunehmen, dass bestimmte Schmerzindikatoren nur in spezifischen Kontexten als Hinweis auf vorliegende Schmerzen gewertet werden können, in anderen Situationen dagegen nur eine geringe Indikationskraft besitzen. Obgleich diesem Umstand in keiner der dargestellten Reviewarbeiten ein besonderes Gewicht beigemessen wird, ist die Problematik die gleiche, die auch hinter der weit verbreiteten Forderung steht, situativ unerwartetes bzw. ungewöhnliches Verhalten stärker als bisher als einen Hinweis auf mögliche Schmerzen anzusehen.

3.4.6.2 Zeit

Eine Reihe weiterer Unterschiede zwischen Beobachtungssituationen lassen sich auf einer zeitlichen Dimension beschreiben. Dazu gehört beispielsweise die bislang selten thematisierte Dauer des Beobachtungsintervalles oder der Abstand zwischen der Beobachtung und dem Ausfüllen des Untersuchungsmaterials. Daneben können aber auch Prozesse (Eigendynamik des Schmerzes, Expertise bzw. Lernprozesse beim Beurteiler), die zwischen zwei inhaltlich verschiedenen Beobachtungssituationen stattfinden, die Interpretation gefundener Schmerzunterschiede als Folge der beispielsweise unterschiedlichen Aktiviertheit des Betroffenen erschweren. Auf potenzielle Schmerzveränderungen in Abhängigkeit von der zirkadianen Rhythmik der Bewohner oder der Tageszeit (morgens vs. abends) soll im Rahmen dieser Arbeit dagegen nicht eingegangen werden.

Beobachtungsintervall und Bearbeitungszeit

Eine Diskussion der Zeitspanne, die für eine Verhaltensbeobachtung zu veranschlagen ist findet im eigentlichen Sinne nicht statt. Die Bearbeitungsdauer der Instrumente wird weitestgehend nur im Zusammenhang mit der klinischen Nützlichkeit bzw. Alltagspraktikabilität diskutiert, und bezieht sich auf das (nachträgliche) Ausfüllen des Bogens (van Herk et al., 2007). Eine inhaltliche Auseinandersetzung dahingehend, welche Zeiträume zur Beobachtung verschiedener Schmerzindikatoren oder auch verschiedener Schmerzqualitäten besonders angemessen sein könnten, wird dagegen nicht geleistet.

Bei den meisten der vorgeschlagenen Skalen wird in den Instruktionen kein Hinweis auf die Dauer der Beobachtung gegeben. Entsprechend wurden in den berichteten empirischen Arbeiten die Beobachtungsintervalle an die Erfordernisse der Studie oder die Art der beobachteten Situation angepasst. Eine Pflegesituation beispielsweise, in die der Beobachter direkt eingebunden ist, kann nicht unbedingt nach einer definierten Zeit abgebrochen werden. Die Bearbeitung des Bogens erfolgt dabei nachträglich und in der Regel ist davon auszugehen, dass die Verhaltensindikatoren während der Beobachtung nicht *mitgelesen* werden können. Wurden die Beobachtungen dagegen nicht-teilnehmend durchgeführt (z.B. Beobachtung des Bewohners in Ruhe, bei Alltagsaktivitäten, oder bei einem nachträglichen Rating von Videomaterial), so kann der Beobachtungszeitraum stringenter bestimmt werden, und das Mitlesen und unmittelbare Markieren schmerzbezogener Verhaltensweisen ist möglich. Husebo und Kollegen (2009) berichten von systematisch gesteigerten Raten beobachteten Schmerzverhaltens in den Videoratings gegenüber der in-situ-Einschätzung am Pflegebett.

Morello und Kollegen (2007) wählten für ihre Untersuchungen zur EPCA-2 jeweils zehnminütige Beobachtungsintervalle. Leider berichten die Autoren nicht, ob dieser Zeitraum sich lediglich auf die Verhaltensbeobachtung des Bewohners in Ruhe (Items 1-4) bezog, oder ebenso die Beobachtung des Bewohners während der Pflege bzw. Mobilisation (Items 5-8) mit einschloss.

Die Dauer für die standardisierte Mobilisation im Rahmen der MOBID wird im Proto-

koll nicht genau festgelegt, dennoch berichtet die Autorengruppe um Husebo (2007, 2009) von Videoaufnahmen mit einer Länge zwischen 3 und 8 Minuten.

Zwakhalen, Hamers und Berger (2006) legten bei ihrem empirischen Vergleich der Schmerzskalen PAINAD, PACSLAC und DOLOPLUS-2 für die beiden Erfassungssituationen Ruhe und Grippeimpfung ein Beobachtungsintervall von jeweils 2 Minuten fest. Für die Einschätzung schmerzbezogenen Verhaltensausdrucks in den zu einem späteren Zeitpunkt erhobenen bewohnerspezifischen Schmerzsituationen (z.B. Waschen, Wundversorgung) berichten die Autoren kein bestimmtes Zeitintervall, so dass davon auszugehen ist, dass der Beobachtungszeitraum die gesamte Dauer der spezifischen Pflegehandlung umfasste.

Für ihren Vergleich der PAINAD, CNPI und OPBAI wählten Cohen-Mansfield (2008) bzw. Cohen-Mansfield und Lipson (2008) ein Beobachtungsintervall von 5 Minuten.

Diese deutlichen Unterschiede im Beobachtungszeitraum, der Eingebundenheit der Rater und Unmittelbarkeit der Bearbeitung zwischen verschiedenen betrachteten Situationen machen den direkten Vergleich der über konkret beobachtetes Schmerzverhalten erfassten Schmerzbelastung schwierig.

Mit der Forderung, auch Veränderungen in den Verhaltens- und Erlebensweisen der Betroffenen gegenüber den zuvor erkennbaren Verhaltensgewohnheiten, Vorlieben und befindensbezogenen Eigenschaften (z.B. Temperament, Persönlichkeit) mit in die Schmerzerfassung einzubeziehen (immerhin 3 von 6 der durch das Panel for Persistent Pain in Older Adults vorgeschlagenen Indikatorkategorien; AGS, 2002), stellt sich die Frage nach einem angemessenen Beobachtungsintervall in besonderer Weise. Längerfristige, und über mehrere Einzelsituationen konstante Verhaltensveränderungen können selbstverständlich nicht mehr in einer einzelnen Beobachtungssituation eingeschätzt werden, und erfordern entweder die wiederholte Beobachtung des Betroffenen oder zumindest eine retrospektive Fremdauskunft durch Personen, die mit dem Betroffenen und seinen Gewohnheiten vertraut sind. Die zu diesem Zweck möglichen Auskunftspersonen jedoch können in ihren Motivationen und in ihrem (pflegerischen) Ausbildungsstand sehr unterschiedlich sein (s. 3.4.6.3). Der zeitliche Rahmen, den beispielsweise Angehörige für die Einschätzung von demenz- und schmerzbedingten Veränderungen der Betroffenen heranziehen können, mag dabei in der Regel bedeutend weiter gefasst sein als der Erfahrungsausschnitt, den Pflegende in stationären Einrichtungen gewöhnlich haben.

Die Skalen ECPA und BISAD geben als Referenzrahmen für die einzuschätzenden Veränderungen im Bewegungsverhalten und der sozialen Interaktion den bzw. die vorangegangenen Tage an, sprechen jedoch gleichzeitig von Verhaltensweisen entgegen der Gewohnheiten. Eine Abschätzung von Gewohnheiten über einen so kurzen Zeitraum hinweg erscheint zumindest schwierig. Zudem bergen kürzere Bestimmungsintervalle für Gewohnheiten und *normales* Verhalten die Gefahr, dass schleichende Veränderungen, die sich beispielsweise aufgrund der häufigen chronischen Schmerzbelastung dieser Klientel ergeben könnten, unbemerkt bleiben bzw. nicht mit erlebtem Schmerz in Beziehung gesetzt werden. Das ist vor allem auch deshalb problematisch, weil der Schmerzausdruck in der Gruppe chronisch schmerzkranker Menschen oftmals wenig auffällig ist, da typische

Schmerzreaktionen (z.B. Grimassieren, Stirnrunzeln) und -anzeichen (z.B. physiologische Marker) häufig fehlen (Carr & Mann, 2002, S. 58).

Schmerzveränderung als Effekt zeitbezogener Prozesse

Um den inhaltlichen Charakter der Situation als Bestimmungsgröße für das Schmerzgeschehen ausweisen zu können, muss die Eigendynamik (akuten) Schmerzerlebens berücksichtigt werden. Annahmen zum Verlauf des Schmerzerlebens könnten die Wahl eines geeigneten zeitlichen Abstandes zwischen zwei Schmerzmessungen wesentlich mitbestimmen. Allerdings findet sich in der aktuellen Diskussion der Güte verschiedener Verfahren kaum ein Hinweis auf eine systematische Berücksichtigung dieser Dynamik. Ein Schmerzerleben, dass beispielsweise in einer Aktivitätssituation durch eine bestimmte Bewegung angestoßen wurde, könnte auch eine unmittelbar folgende Ruhesituation noch deutlich kennzeichnen. Nicht zuletzt darum wird gewöhnlich eine Ruhesituation als Baseline *vor* potenziell schmerzbelasteten Situationen erfasst. Dennoch können Merkmalsveränderungen, die sich über verschieden lange Zeiträume ergeben, den Blick auf die Wirkung der interessierenden Situationscharakteristik verstellen. Eine Identifikation spezifischer Situationen und Erfassungsintervalle zur verhaltensgestützten Abbildung chronischer Schmerzzustände wurde bislang nicht geleistet, liegt jedoch auch außerhalb des Zielbereiches der vorliegenden Arbeit.

Eine weitere nur selten berücksichtigte zeitbezogene Veränderung betrifft nicht das Schmerzerleben per se, sondern die über mehrere Schmerzbeobachtungen hinweg ansteigende Expertise und Deutungssicherheit der Beobachter selbst. Einen solchen Trainingseffekt der Beobachter konnte beispielsweise für das kürzlich vorgestellte Instrument MOBID nachgewiesen werden (Husebo et al., 2009). Sicherlich sind entsprechende Effekte zunächst zu begrüßen; dennoch können sie gegebenenfalls die Unterschiede in der Beziehung zwischen interessierenden Situationsmerkmalen und dem Schmerzerleben maskieren. Eine höhere professionelle Expertise im Umgang mit schmerzbelasteten Menschen kann aus psychometrischer Sicht jedoch auch Probleme für die Schmerzmessung mit sich bringen. So weisen Sjöström und Kollegen darauf hin, dass mit der Dauer der Beschäftigung im Versorgungssektor von Schmerzpatienten seitens der Pflegenden eine Adaptation des Referenzrahmens zur Schmerzeinschätzung stattfindet, der einem empathischen Nachvollziehen individuell erlebter Schmerzen unter Umständen im Weg steht und häufiger zu einer Unterschätzung der durch die Betroffenen berichteten Schmerzbelastung führt (vgl. Bergh, Jakobsson & Sjöström, 2008; Bergh & Sjöström, 1999; Sjöström, 1995). Verändern sich jedoch der Referenzrahmen und das Konzept von Schmerzverhalten, liegen also zu verschiedenen Zeitpunkten unterschiedlich ausdifferenzierte Schmerzkonstrukte vor, dann erscheint auch die für deren Vergleich zu fordernde Invarianz der Messstruktur nicht gegeben. Sowohl für Veränderungen in der Sensibilität bei der Identifikation von Schmerzmarkern, als auch für konzeptuelle Veränderungen des Schmerzkonstruktes werden im methodischen Teil dieser Arbeit verfügbare statistische Prüfverfahren diskutiert.

3.4.6.3 Beobachtermerkmale

Vergleichsweise wenig Aufmerksamkeit wird bei der Beurteilung der psychometrischen Güte verschiedener Verfahren der Schmerzmessung durch Verhaltensbeobachtung den Merkmalen der einschätzenden Personen gewidmet. Sollen Verfahren der Schmerzbeobachtung im Klinik- oder Pflegealltag praktikabel sein, müssen sie von denjenigen Personen eingesetzt werden, die tatsächlich einen substanziellen Anteil der Betreuungsarbeit leisten und damit nicht nur über den besten Einblick in das Befinden und die Gewohnheiten der Betroffenen, sondern auch über Möglichkeiten einer gezielten Schmerzintervention verfügen. Im Praxisalltag ist die Kombination dieser Kompetenzen allerdings nicht immer einfach zu finden.

Mahoney und Peters (2008) beispielsweise merken an, dass der hohe Anteil der in ihrer Studie zur empirischen Überprüfung der MPS beteiligten Pflege*helfer* die externe Validität der Untersuchung steigern sollte, da diese un- oder angelernten Kräfte einen Großteil der eigentlichen alltäglichen Betreuung der Bewohner übernehmen, während die formalpflegerisch besser ausgebildeten Fachkräfte (in Australien besitzen *registered nurses* eine mit einem Universitätsabschluss gleichwertige Qualifikation) häufiger für die Organisation der Pflege verantwortlich seien.

Dem zum Trotz stützen beispielsweise Husebo und Kollegen (2009) ihre Analysen zur Abschätzung der psychometrischen Güte der MOBID ausschließlich auf die Videoratings von Pflegekräften, die die betroffenen Bewohner nicht kannten. Auch in der Arbeit von Zwakhalen, Hamers und Berger (2006) wird den Einschätzungen der Autoren selbst als erfahrenen externen Ratern ein vergleichsweise großes Gewicht eingeräumt. Dabei ist klar, dass sowohl videogestützte Ratings, als auch die Schmerzeinschätzung durch externe Experten für die alltägliche Versorgung demenzkranker Menschen weder repräsentativ noch praktikabel sind.

Daneben sprechen Hadjistavropoulos und Kollegen den Umstand an, dass gezeigtes Schmerzverhalten auch davon abhängig sein könnte, welche Personen (z.B. Pflegekräfte, Angehörige) in der entsprechenden Beobachtungssituation anwesend sind (Hadjistavropoulos et al., 2007). Auf die insbesondere von Hadjistavropoulos und Craig hervorgehobene Bedeutung des Schmerzausdruckes als einem kommunikativen Akt in einer sozialen Situation wurde bereits in Kapitel 3.1 (Schmerzkommunikation) hingewiesen (Hadjistavropoulos & Craig, 2002; Hadjistavropoulos, Fuchs-Lacelle & Craig, 2004; Hadjistavropoulos, von Baeyer & Craig, 2001).

In ihrer Zusammenschau verschiedener Schmerzassessments spricht das Expertenpanel um Hadjistavropoulos (2007) nur wenige der zuvor genannten Unterschiede in den umgesetzten Beobachtungssituationen, die eine Vergleichbarkeit der Messwerte einschränken, an. Dabei wird erkannt, dass die Reliabilität von Skalen (i.S. von Retest-Reliabilität) systematisch unterschätzt wird, wenn Unterschiede in den Erfassungsbedingungen (z.B. Aktiviertheitsgrad, Länge des Beobachtungsintervalles, anwesende Personen) nicht berücksichtigt werden (Hadjistavropoulos et al., 2007). Die Bestimmung der psychometrischen Güte der Schmerzbeobachtung leidet jedoch nicht nur unter einer *unabsichtlichen Nicht-*

berücksichtigung tatsächlicher Unterschiede zwischen verschiedenen Beobachtungssituationen. Selbst wenn die Kontextfaktoren hinreichend kontrolliert scheinen, sind die Möglichkeiten für einen Vergleich von Messwerten an spezifische psychometrische Voraussetzungen geknüpft, die gewöhnlich nicht überprüft werden.

So kommt insbesondere bei der Abschätzung der Konstruktvalidität eines Instrumentes dem Vergleich der in verschiedenen Situationen mit a priori unterschiedlich angenommener Schmerzbelastung erhobenen Schmerzwerte eine wichtige Bedeutung zu. Die Voraussetzung dafür, dass solche erwarteten Unterschiede in den Messwerten als Hinweis auf die diskriminante Validität eines Verfahrens gewertet werden können, ist jedoch, dass die Messung zu beiden Zeitpunkten bzw. in beiden Untersuchungsbedingungen äquivalent ist.

Für einige Schmerzinventare wurden in aktivierten, pflegerischen, unangenehmen oder post-operativen Situationen erwartungskonform höhere Skalenscores ermittelt als in ruhigen, angenehmen oder pre-operativen Situationen (Abbey et al., 2004; Feldt, 2000; Hadjistavropoulos et al., 2002; Schuler, 2008; Schuler et al., 2007; Warden et al., 2003). Die gesteigerte Beobachtungsrate für viele der Einzelindikatoren in der vermeintlichen Schmerzsituation führt im Sinne der klassischen Testtheorie zu einer reduzierten Itemschwierigkeit, d.h. der Test auf Schmerzerleben wird in solchen Beobachtungssituationen insgesamt leichter. Solange aber nicht belegt ist, dass die einzelnen Verhaltensindikatoren in beiden Situationen gleich reliable und schwierige Schmerzindikatoren darstellen, stellt sich die Frage, ob die Bewohner tatsächlich mehr Schmerz erleben, oder ob die höheren Testscores nicht eher auf die Situationssensitivität des Tests an sich zurückzuführen sind. Test- und Bewohnercharakteristika lassen sich auf der Grundlage der Annahmen der klassischen Testtheorie also nicht stringent trennen.

Für keines der bisher vorgeschlagenen Instrumente zur Schmerzbeobachtung wurde jedoch ein angemessener Nachweis der trans-situativen Invarianz der Verhaltensindikatoren geführt. Damit aber bleibt unklar, ob die konsistent berichteten höheren Skalenwerte in Situationen, die durch Bewegung und Aktivität gegenüber Ruhe und Entspannung gekennzeichnet sind auf eine tatsächlich höhere Schmerzbelastung zurückgeführt werden können oder ob die Verhaltensindikatoren bei Aktivität lediglich ‚anders funktionieren'.

Eine detaillierte Diskussion der statistischen Verfahren zur Überprüfung verschiedener Aspekte der Äquivalenz bzw. Invarianz einer Messung über verschiedene Zeitpunkte oder unterschiedliche Situationen hinweg findet sich im methodischen Teil dieser Arbeit (Kapitel 4.3.3.5).

3.4.7 Demenzsspezifität

Beeinträchtigungen des Denkens, des Gedächtnisses und der Urteilsfähigkeit sind Leitsymptome der Demenz, die eine schmerzbezogene Selbstauskunft unmöglich machen können. Über weite Teile des Krankheitsverlaufes hinweg scheinen jedoch noch schmerzbezogene Selbstauskünfte möglich (vgl. Kapitel 3.2.3). In ihrer Konsenserklärung zur Schmerzerfassung bei älteren Menschen kommen Hadjistavropoulos und Kollegen (2007) zusammenfassend zu dem Urteil, dass Demenzkranke mit einem Mini Mental State Exam

(MMSE)-Wert über 18 (aus maximal 30) Punkten (geringe bis mittlere kognitive Beeinträchtigung) sehr wahrscheinlich Auskunft auf einer angemessenen standardisierten Skala geben können (z.B. Chibnall & Tait, 2001; Weiner et al., 1999). Zwar gibt es Hinweise darauf, dass auch Menschen mit stärkerer Beeinträchtigung noch zu einer Selbstauskunft in der Lage sind (z.B. Gibson et al., 2001), dennoch ist davon auszugehen, dass die Reliabilität der Selbstauskunft spätestens bei MMSE-Werten unter 12 Punkten gravierend beeinträchtigt ist. Da die schmerzbezogene Selbstauskunft als Goldstandard gilt, sollten alle Möglichkeiten ausgeschöpft werden, solche Selbstauskünfte auch bei Menschen mit kognitiven Beeinträchtigungen einzuholen.

Spätestens jedoch wenn die kommunikativen Fähigkeiten des demenzkranken Menschen so stark beeinträchtigt sind, dass auch einfache Verfahren der Befragung bzw. Selbstauskunft nicht mehr durchgeführt werden können, wird die verhaltensbezogene Schmerzeinschätzung als Alternative empfohlen. Die speziell für diesen Zweck vorgeschlagenen Beobachtungsverfahren wurden zuvor mit Blick auf ihre psychometrische Güte eingehend dargestellt. Auf Merkmale, die als für demenzkranke Menschen spezifisch gelten können, wurde dabei allerdings genau betrachtet überhaupt kein Bezug genommen. Tatsächlich stellt sich dann aber die Frage, anhand welcher Besonderheiten im Schmerzerleben und non-verbalen Schmerzausdruck sich demenzkranke Menschen überhaupt von anderen Pflegeheimbewohnern ohne kognitive Beeinträchtigung unterscheiden lassen sollten.

3.4.7.1 Experimentelle Befunde

Um diese Frage zu klären, sind vorrangig solche Untersuchungen nützlich, bei denen der Schmerzreiz selbst hinreichend gut kontrolliert wurde. Bei der Zusammenschau der experimentellen Befunde zu demenzkorrelierten Veränderungen im Schmerzerleben älterer Menschen (Kapitel 2.2.4.1) wurde dargelegt, dass demenzkranke Menschen im Vergleich zu altersgleichen kognitiv Unbeeinträchtigten seltener über Schmerzen klagen, eine deutlich reduzierte vegetative, aber deutlich gesteigerte mimische Schmerzreaktion zeigen, und eine erhöhte Toleranzgrenze besitzen.

Für die Verhaltensbeobachtung des Schmerzausdruckes bei Demenz bedeutet das aber, dass für die vorgeschlagenen Verfahren lediglich im Bereich mimischen Schmerzausdruckes eine (mindestens) genauso sensitive Schmerzidentifikation für demenzkranke Menschen angenommen werden kann wie für nicht-demenzkranke Personen. Die Verhaltensindikatoren aus anderen Ausdrucksbereichen können dagegen ein vergleichbares Schmerzempfinden bei Demenzkranken weniger gut zur Abbildung bringen als bei kognitiv unbeeinträchtigten Älteren. Trotz dieser Befunde werden gelegentlich auch physiologische Marker für die Schmerzmessung schwer beeinträchtigter Demenzkranker empfohlen (Stolee et al., 2005).

Auf der Grundlage der bisherigen experimentellen Forschungsergebnisse erscheint keine belastbare weitergehende Aussage über einen spezifischen Schmerzausdruck bei verschiedenen Graden kognitiver Beeinträchtigung oder bestimmten Demenzätiologien

möglich, an dem sich die Entwicklung eines besser auf die Zielgruppe demenzkranker Menschen ausgerichteten Verfahrens der Schmerzbeobachtung orientieren könnte. Die Generalisierbarkeit der Befunde der referierten experimentellen Studien auf die Gruppe schwer beeinträchtigter oder an einer anderen Demenzform als der Alzheimer Demenz erkrankter Menschen muss bezweifelt werden, da überwiegend Selbstauskünfte erhoben und DAT-Patienten untersucht wurden (vgl. Kunz, 2006).

3.4.7.2 Schmerzkennzeichen im Pflegealltag

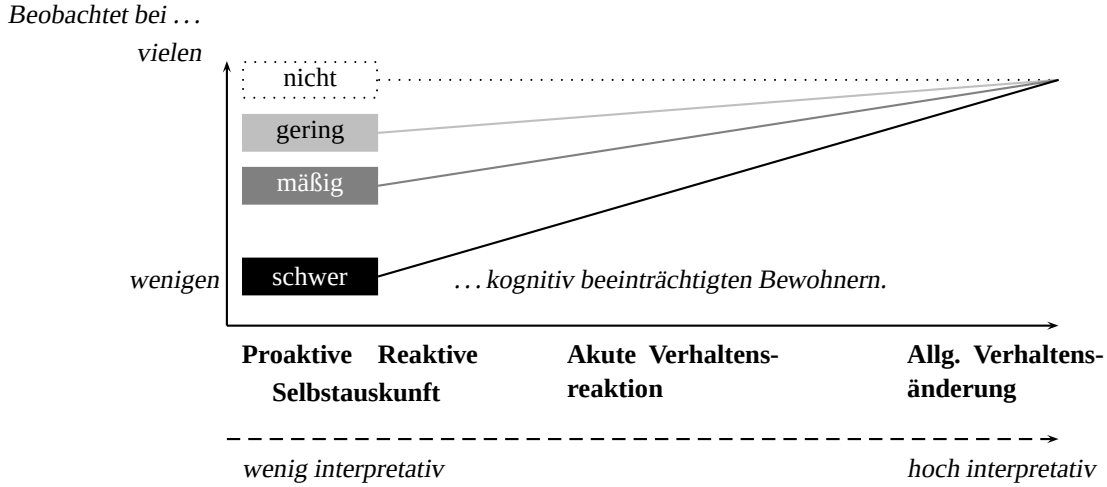
Mittlerweile wurden einige Arbeiten vorgelegt die untersuchten, welche Verhaltensweisen bei individuellen Bewohnern (teilweise mit unterschiedlichem Grad kognitiver Beeinträchtigung) von den beteiligten Pflegenden als Hinweis auf mögliches Schmerzerleben genutzt werden (Closs, Cash, Barr & Briggs, 2005; Fuchs-Lacelle & Hadjistavropoulos, 2004; Kovach et al., 1999; Weiner, Peterson & Keefe, 1999). Die Erkenntnisse aus diesen Interviewstudien flossen direkt in die Erstellung klinisch relevanter Schmerzskalen wie beispielsweise der PACSLAC, DS-DAT oder PAINAD ein.

Trotz der Vielzahl vorgeschlagener Schmerzindikatoren sehen Closs und Kollegen (2005) einen sich abzeichnenden Konsens dahingehend, dass Gesichtsgrimassen, negative Vokalisationen, und angespannte Körperhaltung oder Schonhaltung bei Schmerzen wahrscheinliche Verhaltensweisen demenzkranker Menschen sind. In einer eigenen Studie untersuchten die Autoren den Zusammenhang zwischen den von 65 Pflegekräften und 36 pflegenden Angehörigen verwendeten behavioralen Schmerzanzeichen und dem Grad der kognitiven Beeinträchtigung von 113 demenzkranken Menschen in insgesamt 15 Pflegeheimen. Aufgrund ihres Mini Mental-Wertes wurden die Bewohner den Gruppen unbeeinträchtigt (n=22, MMSE=24-30), gering beeinträchtigt (n=24, MMSE=18-23), mäßig beeinträchtigt (n=31, MMSE=10-17) und schwer beeinträchtigt (n=28, MMSE=0-9) zugeordnet. Die Autoren beschreiben proaktive verbale und körpersprachliche Schmerzindikatoren, reaktive verbale und körpersprachliche Schmerzindikatoren, akute Verhaltensindikatoren als Reaktion auf Bewegung, und allgemeine Veränderungen im Verhalten oder der Stimmung des Bewohners als übergeordnete Kategorien für die berichteten Schmerzanzeichen. Die Anzahl aller verschiedener genutzter Schmerzindikatoren steigt im Vergleich der Demenzgruppen bis zur Gruppe mittlerer kognitiver Beeinträchtigung deutlich an, während in der Gruppe der schwer Demenzkranken wieder weniger Schmerzindikatoren genutzt werden. Körperbewegungen und mimische Schmerzindikatoren wurden in allen Gruppen am häufigsten genannt, wobei in der Gruppe der schwer Demenzkranken der Gesichtsausdruck wieder weniger häufig, Körperbewegungen und Verbalisation dagegen vermehrt zur Identifikation von Schmerzen genutzt werden. Verbale Schmerzäußerungen sind dabei bei schwerer Demenz stärker durch Schreien und eine inkohärente Sprache bestimmt, als durch einen gezielten Schmerzbericht. Daneben nehmen die Autoren an, dass die geringere Anzahl berichteter verhaltensbezogener Schmerzanzeichen in den Gruppen nicht und gering kognitiv beeinträchtigter Bewohner als Hinweis darauf gewertet werden könnte, dass die Kommunikation von Schmerzen hier noch weitestgehend im Gespräch

erfolgt, so dass der Rückgriff auf Verhaltensmerkmale weniger notwendig sein könnte.

Auf der Grundlage ihrer Befunde schlagen Closs und Kollegen (2005) ein vorläufiges theoretisches Modell vor, das den Zusammenhang zwischen dem Grad der kognitiven Beeinträchtigung und der Nutzung verschiedener Typen von Schmerzanzeichen mit unterschiedlichen Voraussetzungen hinsichtlich der Kenntnis des Betroffenen und dem Ausmaß nötiger Interpretation beschreibt (s. Abbildung 5).

Abbildung 5: In verschiedenen Schweregraden der Demenz genutzte Verhaltensindikatoren (nach Closs et al., 2005, p. 10).



Mit dem Grad der kognitiven Beeinträchtigung nimmt nach diesem Modell die Verfügbarkeit konkreterer Verhaltensindikatoren (v.a. der verbalen Selbstauskunft), die wenig Interpretation vom Beobachter fordern ab, während die Bedeutung von weniger eindeutigen potenziell schmerzbezogenen Verhaltensäußerungen zunimmt. Für die am stärksten beeinträchtigten Bewohner erscheint die Schmerzeinschätzung durchgängig mit ungleich mehr Unsicherheit belastet zu sein als für die weniger beeinträchtigten Bewohnergruppen. In dieser Gruppe kommt Veränderungen in den gewohnten Verhaltensweisen trotz ihres hohen Anspruches bezüglich der Kenntnis des Bewohners und einer umsichtigen Interpretation eine gesteigerte Bedeutung zu.

Zur Identifikation spezifischer Verhaltensweisen, die in einer konkreten Beobachtungssituation verlässlich Auskunft über die Schmerzbelastung verschieden stark kognitiv beeinträchtigter Personen geben könnten (wenn sie nur gezielt berücksichtigt würden), leistet dieses – die gegenwärtige Alltagspraxis beschreibende – Rahmenmodell allerdings keinen wesentlichen Beitrag.

Auch die Studien zur Entwicklung und Überprüfung verhaltensgestützter Schmerzassessments bei Demenz machen kaum eine präzise Aussage darüber, in welcher Hinsicht die vorgeschlagenen Assessments für (verschieden stark) kognitiv beeinträchtigte Men-

schen spezifisch sein sollten. Die Möglichkeiten der bisherigen Arbeiten, Demenzspezifität überhaupt detailliert zu diskutieren, bleiben aufgrund der häufig geringen Stichprobengröße bzw. Anzahl demenzkranker Studienteilnehmer (z.B. Baiardi et al., 2002; Davies et al., 2004a,b; Sign & Orrell, 2003; Snow et al., 2004; Warden et al., 2003) oder aber der bewusst gewählten Homogenität der Probanden (z.B. nur schwer Demenzkranke; Cohen-Mansfield, 2008; Cohen-Mansfield & Lipson, 2008; Pautex et al., 2006; Villanueva et al., 2003) tatsächlich sehr beschränkt.

Diejenigen Studien, die Probandengruppen mit unterschiedlichem kognitiven Status unterscheiden, scheinen sich ausschließlich für Gruppenunterschiede in den Gesamtskalenscores zu interessieren, und setzen damit voraus, dass die Verhaltensinventare den Schmerz in beiden Untersuchungsgruppen gleich gut abbilden (z.B. Baiardi et al., 2002; Feldt, 2000).

Aber auch stärker psychometrisch ausgerichtete Arbeiten übergehen mögliche Unterschiede in der Angemessenheit der Schmerzinventare in verschiedenen Demenzschweregraden. Bei ihrer psychometrischen Überprüfung der PACSLAC, DOLOPLUS-2 und PAINAD berichten beispielsweise Zwakhalen, Hamers und Berger (2006) zwar verschiedene Beeinträchtigungsgrade und Demenzätiologien der Studienteilnehmer, diskutieren aber weder die Auftretenshäufigkeit der Einzelindikatoren, noch die weiteren Aspekte der Reliabilität und Validität dieser Skalen mit Blick auf die Unterschiede im kognitiven Status der Bewohner.

Insgesamt drängt sich der Eindruck auf, dass es vielfach im wesentlichen allein der *Verzicht auf die Erfragung einer Selbstauskunft* ist, der die ‚Demenzspezifität' der vorgestellten Beobachtungsverfahren begründet. Anstelle der *besonderen Eignung* der vorgeschlagenen Verhaltensinventare für die Schmerzerfassung bei demenzkranken Menschen scheint gegenwärtig vielmehr noch deren *prinzipielle Anwendbarkeit* diskutiert zu werden.

Tatsächlich können für die Gruppe demenzkranker Menschen – neben der kognitiven Symptomatik, die eine Selbstauskunft erschwert – einige nicht-kognitive Merkmale beschrieben werden, die eine besondere Herausforderung auch für die verhaltensgestützte Schmerzmessung darstellen. Als für eine Verhaltensbeobachtung bei demenzkranken alten Menschen schwierige Umstände werden vorrangig die Belastung mit Verhaltensauffälligkeiten (i.S. nicht-kognitiver Demenzsymptome), die Heterogenität der betroffenen Hirnareale bei unterschiedlichen Demenzätiologien und die zunehmende Disintegration auch körperlicher Steuerungsmechanismen über den Krankheitsverlauf hinweg, sowie die ausgeprägte Multimorbidität der Klientel demenzkranker Menschen erkannt und diskutiert.

3.4.7.3 Nicht-kognitive Demenzsymptome

Der überwiegende Teil aller demenzkranken Menschen zeigt zu irgendeinem Zeitpunkt im Krankheitsverlauf neuropsychiatrische Symptome wie Wahnvorstellungen, Halluzinationen, Erregung bzw. Aggression, Depression bzw. Dysphorie, Angst, Hochstimmung bzw. Euphorie, Gleichgültigkeit bzw. Apathie, Enthemmung, Reizbarkeit bzw. La-

bilität oder abweichendes motorisches Verhalten (Cummings et al., 1994). Störungen der Affektkontrolle, des Antriebs und des Sozialverhaltens stellen dabei wichtige Diagnosekriterien für die Demenz dar (siehe Kapitel 2.1.1). Einige der genannten Verhaltensauffälligkeiten, beispielsweise Aggression, Angst, Depression, Reizbarkeit oder abweichendes motorisches Verhalten, können auch als Reaktion auf Schmerzen beobachtet werden, was die Interpretation der verhaltensgestützten Schmerzbeobachtung bei Menschen mit nicht-kognitiven Demenzsymptomen schwierig macht.

Aufgrund der nahezu ausschließlichen Konzentration auf den kognitiven Status der Probanden liegen zum Zusammenhang zwischen Schmerzbeobachtung und weiteren Facetten des Demenzsyndromes gegenwärtig nur wenige empirische Befunde vor.

Eine in dieser Hinsicht herauszuhebende Ausnahme stellt die vor kurzem von Mahoney und Peters (2008) vorgestellte Arbeit zur MPS dar, die den Versuch unternimmt, Schmerzerleben und Agitation voneinander zu unterscheiden. Als Kernmerkmale der Agitation nennen die Autoren störendes Verhalten wie beispielsweise Widerstand gegen Pflegehandlungen, Schreien, anhaltendes Schaukeln mit dem Oberkörper, Nesteln bzw. Zappeln oder nach Anderen Ausschlagen. Herausforderndes Verhalten resultiert dabei nicht selten aus der Überforderung, die mit abnehmender kognitiver Leistungsfähigkeit einhergeht, oder stellt einen Ausdruck unbefriedigter Bedürfnisse dar (Barton, Findlay & Blake, 2005; Fischer, Spahn & Kovach, 2007; Kovach, Cashin & Sauer, 2006). Eine Abgrenzung solcher Unbehagenszustände von erlebten Schmerzen ist schwierig, und es besteht die Gefahr, dass Pflegende alle agitierten Verhaltensweisen als unvermeidliche Demenzsymptome interpretieren, anstatt auch potenziell behandelbare Ursachen, beispielsweise unerkannte Schmerzen in Betracht zu ziehen.

Um Schmerz und Agitation voneinander abzugrenzen, schlagen die Autoren der MPS vor, das konkret beobachtete potenzielle Schmerzverhalten in den Ausdrucksbereichen Mimik, Atmung, Lautäußerung und Körpersprache (Itemset 1) mit dem gewöhnlichen (agitierten) Verhalten des Betroffenen, Veränderungen in den Schlaf- und Essgewohnheiten, aktuellem vegetativen Zustand, sowie mit (chronisch) schmerzhaften Krankheiten in der Anamnese (Itemset 2) abzugleichen. Liegen dabei viele Hinweise darauf vor, dass der konkret beobachtete Zustand ungewöhnlich ist, spricht das eher für Schmerz und gegen Agitation. Dabei muss jedoch bedacht werden, dass auch ein häufig bzw. üblicherweise gezeigtes agitiertes Verhalten durch lang anhaltende oder wiederkehrende unerkannte Schmerzzustände bestimmt sein kann.

In ihrer Studie verglichen Mahoney und Peters (2008) Bewohner mit mehreren bekannten aktuellen schmerzhaften Erkrankungen (Schmerzgruppe, n=52) ohne Agitation und Bewohner ohne bekannte akute Schmerzbelastung aber mit regelmäßigem agitierten Verhalten (Agitationsgruppe, n=35) in einer angenehmen und einer unangenehmen Situation. In der unangenehmen Situation wurden für beide Gruppen vergleichbare Werte für das konkret beobachtete Schmerzverhalten (Itemset 1) beobachtet, das jedoch erwartungsgemäß für die Schmerzgruppe als ungewöhnlicher eingeschätzt wurde als für die Agitationsgruppe (Itemset 2). Die Autoren werten diesen Befund als ersten Hinweis auf die Möglichkeit, mit dem MPS schmerzbezogenen und nicht-schmerzbezogenen agitier-

ten Verhaltensausdruck differenzieren zu können. Prinzipiell trifft das Problem der Konfundierung von agitiertem und schmerzbezogenem Verhalten nicht nur auf die Schmerzeinschätzung, sondern auch auf die Agitationseinschätzung selbst zu. Leider verwenden die Autoren für die unterschiedenen Gruppen eine abweichende Zusammenstellung von Einzelitems und machen keine genaueren Angaben zur Art der unangenehmen Situationen, so dass die Interpretation der im Vergleich zur angenehmen Situation gesteigerten Werte für Itemset 1 als Schmerzausdruck fraglich bleibt. Trotz ihrer konzeptuellen und methodischen Schwächen trägt diese Arbeit sicherlich dazu bei, das Bewusstsein für die nicht unerheblichen Herausforderungen zu schärfen, die auf dem Weg zu einem tatsächlich demenzspezifischen Verhaltensinventar zur Schmerzmessung noch überwunden werden müssen.

3.4.7.4 Demenzätiologie

Der Zusammenhang zwischen dem Schmerzerleben bzw. Schmerzausdruck und der Schädigung spezifischer Hirnareale bei verschiedenen Demenzformen wird gegenwärtig nur unzureichend verstanden. Die verlässlichsten Aussagen können dabei zur Demenz vom Alzheimer-Typus (AD) getroffen werden, bei der aufgrund massiver Schädigungen im medialen Schmerzsystem vorrangig Beeinträchtigungen der motivational-affektiven, kognitiv-evaluativen und autonomen Schmerzreaktion bei – aufgrund des weniger beeinträchtigten lateralen Schmerzsystems – erhaltener sensorisch-diskriminativer Schmerzempfindung erwartet werden können (vgl. Kapitel 2.2.4). Wenn erlebte Schmerzen aber nicht antizipiert werden, länger toleriert werden, und führen sie zu weniger ausgeprägten vegetativen Reaktionen, dann kann erwartet werden, dass Verhaltensindikatoren die Angst und Vermeidung, Abwehr und Klagen, oder Atmung und Puls erfassen das tatsächliche Schmerzerleben unterschätzen. Neben den unmittelbar relevanten Veränderungen im Schmerzerleben und Schmerzausdruck ist jedoch auch die Alzheimer-Demenz spätestens in den letzten Krankheitsphasen durch massive körperlich-funktionale Beeinträchtigungen gekennzeichnet (s. Kapitel 2.1.2.1). Verhaltensindikatoren, die ein vergleichsweise hohes funktionales Niveau bzw. unbeeinträchtigte Mobilität voraussetzen (z.B. gesteigertes Umherwandern, Schaukeln mit dem Oberkörper, Knie anziehen, nach Anderen Schlagen oder sich entziehen) können bei demenzkranken Menschen in schweren Erkrankungsstadien häufig nicht mehr beobachtet werden. Die Angemessenheit eines Verfahrens variiert somit nicht nur in Abhängigkeit von der Art der demenziellen Erkrankung, sondern auch über den Verlauf der Erkrankung hinweg. Insbesondere bei der Forderung nach solchen Verhaltensindikatoren, die eine potenzielle Schmerzbelastung durch subtile Veränderungen wie sozialen Rückzug oder Veränderungen in Befinden und Emotionalität zu erfassen suchen, muss der progrediente, je nach Ätiologie verschieden schnell und diskontinuierlich erfolgende Verlauf der Demenzerkrankung (z.B. AD vs. VD) berücksichtigt, und müssen angemessene Referenzzeiträume definiert werden.

Durch die Schädigung einzelner vergleichsweise eng umrissener Hirnareale können sich vor allem bei Demenzen mit einer zerebrovaskulären Ätiologie sehr individuelle Be-

einträchtigungen in solchen Verhaltensweisen ergeben, die sich auch in Verhaltensinventaren zur Schmerzmessung finden lassen. In Kapitel 2.1.2.2 wurde bereits auf die Extrapyramidalsymptomatik hingewiesen, die sich in einer Tonussteigerung bis hin zur Akinese ausdrückt, und durch Verhaltensindikatoren wie angespannte Körperhaltung, oder starre Körpersprache als Schmerzausdruck verkannt werden könnte. Spezifische auffällige Verhaltensweisen, die eine Schmerzerfassung erschweren, können auch für die Gruppe der Demenzen mit fokal-atrophischer Ätiologie, insbesondere für die frontotemporale Demenz (FTLD) vom pseudoneurasthenischen Typus beschrieben werden. Die hier beschriebene klinische Symptomatik einer veränderten Persönlichkeit, des sozialen Rückzugs oder des Interessenverlustes (Apathie) sollte insbesondere bei einer Schmerzeinschätzung auf der Grundlage veränderungsbezogener AGS-Kategorien (d.h. Änderungen der Gewohnheiten, des Sozialverhaltens und des Befindens; AGS, 2002) berücksichtigt werden.

Die Schwierigkeit, angesichts dieser Heterogenität der Demenzen allgemeingültige Erwartungen hinsichtlich des normalen (i.S. eines nicht schmerzbezogenen) und schmerzbezogenen Verhaltensausdrucks zu machen wird in allen Übersichtsarbeiten beklagt, beispielsweise konstatieren Herr und Kollegen, dass die "[...] effects of dementia on the brain can be quite variable, depending on the part of the brain affected, such that patient's pain responses can be unique" (Herr et al., 2006, p. 187). Gegenwärtig erscheinen Verfahren, die diese Heterogenität strukturieren und spezifische Schmerzabbildungen bei verschiedenen Ätiologien oder über den Verlauf der Erkrankung hinweg erlauben jedoch noch in weiter Ferne.

3.4.7.5 Hohes Lebensalter und Komorbidität

Während zuvor auf Faktoren hingewiesen wurde, die in einem ursächlichen Zusammenhang mit der Demenz stehen, können jedoch auch Herausforderungen der Schmerzmessung bei Demenz beschrieben werden, die einen weniger engen Bezug zur Ätiologie und Symptomatik der Demenz besitzen. Da Demenzen eindeutig alterskorrelierte Erkrankungen darstellen, weisen die Betroffenen allesamt auch typische psychische und körperliche Merkmale eines hohen oder sehr hohen Lebensalters auf, die für die verhaltensbezogene Schmerzerfassung bedeutsam sein können. Dabei steht die hohe Multimorbidität demenzkranker alter Menschen häufig im Mittelpunkt der Betrachtungen. Bei der Diskussion der Beobachtungsverfahren für Demenz finden sich nur vereinzelt Hinweise auf nicht-schmerz- und nicht-demenzbezogene Veränderungen bei den Betroffenen und deren Bedeutung für die Verhaltensbeobachtung. Darum, und auch, weil sich ihre Übersichtsarbeit nicht auf demenzkranke Menschen beschränkt, heben Hadjistavropoulos und Kollegen (2007) in ihren Empfehlungen hervor:

> "The physical evaluation requires an understanding of the diversity in range of function that is considered normal in an older population. False attribution of age-related changes on physical examination may lead to an incorrect diagnosis. For example, some degree of muscle atrophy, muscle weakness, or

> decreased range of motion of major joints is expected in very old patients." (ebd., p. S8)

Zu den normalen altersbedingten Veränderungen, die Einfluss auf die Schmerzerfassung haben können, sind neben solchen muskulo-skellettalen Degenerationen aber beispielsweise auch sensorische Beeinträchtigungen (z.B. Hören, Sehen), die Faltenbildung im Gesicht oder der Verlust der Zähne zu zählen. In Verfahren zur Beobachtung und Interpretation des mimischen Ausdrucksverhaltens sind diese Veränderungen gewöhnlich nicht berücksichtigt (Re, 2003). Schmerzindikatoren wie lautes Rufen oder Veränderungen des Blickes erscheinen bei starken sensorischen Einbußen uneindeutig.

Bedeutend schwerwiegender in ihrer Folge für das Schmerzassessment als diese normalen Alternsprozesse ist die Kumulation krankhafter Bedingungen im hohen bzw. sehr hohen Lebensalter. Mit größerer Multimorbidität sind auch die Möglichkeiten, Schmerzen auszudrücken stärker beeinträchtigt (Zwakhalen, Hamers & Berger, 2006).

Aufgrund der durch die Demenzätiologie, den Krankheitsverlauf, Altersprozesse und Multimorbidität mitbestimmten Heterogenität des individuellen Verhaltensausdruckes demenzkranker Menschen muss davon ausgegangen werden, dass sich nicht nur der erfasste Schmerzausdruck selbst, sondern bereits die für dessen Einschätzung ausgewählten Situationen deutlich voneinander unterscheiden. Erfahrungen aus weiteren HILDE-Projektphasen und Kooperationsprojekten mit teilweise sehr schwer beeinträchtigten Bewohnern machen deutlich, dass beispielsweise eine Aktivitätssituation für die Gruppe schwer beeinträchtigter bettlägriger Bewohner auch konzeptionell unterschiedlich gefasst werden muss als bei mobilen Bewohnern mit weniger schweren Beeinträchtigungen (Projektbericht MeDiA in Cura; Kaspar, Becker & Kruse, 2007). Die Möglichkeiten eines direkten Vergleiches der Schmerzbelastung verschieden schwer beeinträchtigter Bewohner könnten damit selbst in einer vermeintlich homogenen Beobachtungssituation eingeschränkt sein.

3.4.7.6 Zusammenfassung

Der Erfolg der Entwicklung eines demenzspezifischen Verfahrens wird zusammenfassend davon abhängen, ob es gelingt, zukünftig nicht ausschließlich den kognitiven Status, sondern mehrere parallel vorliegende Kernmerkmale demenzkranker Menschen zu berücksichtigen. Ein Vorschlag für eine ganzheitlichere Betrachtung des Musters der erhaltenen Kompetenzen demenzkranker Heimbewohner wurde kürzlich von Kruse und Kollegen gemacht (Becker, Kaspar & Kruse, 2006; Kruse & Schmitt, 2008). Für eine Zuordnung der Bewohner zu den vier beschriebenen Prägnanztypen der Demenz werden – einem Verständnis der Demenz als einem multidimensionalen Syndrom folgend – neben dem kognitiven Status auch erhaltene Alltagsfähigkeiten sowie die Belastetheit mit nichtkognitiven Demenzsymptomen berücksichtigt. Eine detaillierte Beschreibung dieser stark an der Pflegepraxis ausgerichteten Binnendifferenzierung demenzkranker Heimbewohner findet sich in Kapitel 5.4.1.

Unabhängig davon, woraus sich die vielfach berichteten idiosynkratischen Muster des Schmerzausdruckes demenzkranker Menschen im Einzelnen ergeben, stellt sich die Frage danach, wie ‚breit' ein entsprechendes Verhaltensinventar konzipiert werden sollte, um einerseits viele Arten individuellen Schmerzausdruckes berücksichtigen zu können, andererseits jedoch auch noch hinreichend schmerzspezifisch und praktikabel zu sein. Die Frage der optimalen Balance von Sensitivität und Spezifität einer verhaltensgestützten Schmerzeinschätzung kann allerdings nicht ohne eine Berücksichtigung der mit dem Instrumenteneinsatz verbundenen Ressourcen, Ziele oder Folgen beantwortet werden. Die Expertengruppe um Herr (2006) spricht sich beispielsweise dafür aus, dass Verhaltensinventare zum Schmerzscreening umfassend sein sollten, um keine potenzielle Schmerzbelastung unerkannt zu lassen, weist jedoch darauf hin, dass sich an die Verhaltensbeobachtung weitere Schritte zur genaueren Diagnose des Schmerzes anschließen sollten.

3.5 Desiderate der zukünftigen Skalenentwicklung und -beurteilung

Bei der Darstellung des gegenwärtig verfügbaren Arsenals von Verfahren zur (verhaltensgestützten) Messung von Schmerzen bei Demenz und den referierten empirischen Befunden zu Potenzialen und Grenzen ihres Einsatzes in der anspruchsvollen Zielpopulation demenzkranker Menschen wurde an vielen Stellen deutlich, dass der Großteil der vorgeschlagenen Instrumente die Differenziertheit der zur Zeit geleisteten theoretischen Diskussion in keinster Weise abzubilden vermag.

Diese Diskrepanz wird wesentlich auch dadurch bestimmt und perpetuiert, dass die verwendete Methodik für die Zusammenstellung, Überprüfung und Optimierung der Verhaltensinventare nur unzureichend Auskunft über die psychometrischen Eigenschaften der einzelnen in Frage stehenden verhaltensbezogenen Schmerzindikatoren gibt.

Auf der Grundlage der nachgewiesenen begrenzten Anwendungs- und Aussagebereiche der bislang entwickelten Verfahren zur Schmerzerfassung durch Verhaltensbeobachtung und der konzeptionellen Unzulänglichkeiten der für ihre Erstellung und Evaluation üblicherweise herangezogenen Methoden können die nachfolgenden Anforderungen an eine Messung der Schmerzbelastung demenzkranker Menschen abgeleitet werden.

3.5.1 Orientierung an konkret beobachtbarem Verhalten

Das wesentlichste, und in nahezu allen referierten Arbeiten aufscheinende Hindernis für die konsequente Weiterentwicklung des Forschungsfeldes der Schmerzbeobachtung ist der mangelnde Kenntnisstand zur psychometrischen Funktionsweise konkret beobachtbarer Verhaltenseinheiten als dem gewissermaßen grundlegenden Baustein aller Beobachtungsinstrumente. Eine Klärung der Beziehung zwischen dem Schmerzerleben und potenziellen Verhaltensindikatoren wird dabei nicht selten bereits konzeptionell dadurch erschwert, dass lediglich übergeordnete Verhaltenskategorien oder ein beispielhafter Schmerzausdruck beschrieben werden. Darüber hinaus trägt der gewöhnlich für die Skalenanalyse

gewählte methodische Ansatz kaum dazu bei, die theoretischen Annahmen zur Funktionsweise einzelner Verhaltensweisen zu prüfen.

Die in den bisher vorgelegten Studien realisierten Stichproben demenzkranker Menschen, anhand derer die Verfahren entwickelt und evaluiert worden sind, unterscheiden sich beträchtlich im anzunehmenden Niveau ihrer Schmerzbelastung. Die Beobachtungsraten für einzelne schmerzbezogene Verhaltensweisen und die daraus abgeleiteten Itemschwierigkeiten können aber nicht über die realisierte Stichprobe hinaus generalisiert, und entsprechend schlecht miteinander verglichen werden. Wenig Aufmerksamkeit wird gegenwärtig der Überprüfung der (implizit) getroffenen Annahmen zum Zusammenhang zwischen Schmerzintensität und spezifischen Verhaltensweisen geschenkt, selbst wenn einige Skalen (für bestimmte Ausdrucksbereiche) konkrete Verhaltensweisen beschreiben, die verschiedene Schmerzintensitäten anzeigen sollen. Auch scheint die Zusammenstellung der Schmerzitems sich nicht bewusst an klaren Zielvorgaben beispielsweise mit Blick auf die ‚Breite' des abgebildeten Schmerzintervalles oder einer möglichst differenzierten Abbildung eines bestimmten (z.B. für klinische Entscheidungen besonders relevanten) Ausschnittes von Schmerzintensitäten zu orientieren.

Je nachdem, ob die Skalenitems auf der Ebene konkret zu beobachteten Verhaltens (z.B. Stirnrunzeln, Anklammern an Gegenstände) oder abstrakterer Ausdruckskategorien (z.B. Mimik) beschrieben sind, unterscheiden sich die vorgeschlagenen Verhaltensinventare mitunter beträchtlich hinsichtlich ihres Umfanges. Die Frage, wie sehr sich ein einzelnes Verhaltensmerkmal für die Abbildung von Schmerzen eignet, wird – wenn überhaupt – nur im Hinblick auf die Gesamtheit der Testbatterie diskutiert. Der dringend notwenigen Identifikation eines maximal informativen Subsets von Schmerzindikatoren aus dem Gesamtpool der in mehreren Dutzend Inventaren vorgeschlagenen potenziell schmerzbezogenen Verhaltensweisen sind damit entsprechend enge Grenzen gesetzt.

Aus einer differenzierten Betrachtung der durch ein Schmerzverhalten angezeigten Schmerzintensität und dem Ausmaß, in dem ein Verhalten durch Schmerzerleben (und nicht durch andere Faktoren) bestimmt ist, erwachsen auch erweiterte Möglichkeiten, den Aussagebereich eines Gesamtinventares und die über diesen Bereich unterschiedlich gute Schmerzabbildung zu bestimmen. Erst unter diesen Rahmenbedingungen ist zukünftig eine Entwicklung tatsächlich effizienter Verfahren zur Schmerzmessung denkbar.

3.5.2 Berücksichtigung des Erfassungskontextes

Einige Hindernisse für die Weiterentwicklung des Instrumentariums zur Schmerzerfassung bei Demenz ergeben sich aus der nur unzureichend, und wenn überhaupt meistens unangemessen erfolgenden Berücksichtigung verschiedener Kontextmerkmale der Schmerzmessung. Ein handhabbares Verfahren zur Schmerzmessung bei demenzkranken Menschen muss sich stark am Praxisalltag der Pflege ausrichten. In diesem Anwendungsfeld bestehen kaum Chancen, bekannte potenzielle Störfaktoren (experimentell) zu kontrollieren, so dass Informationen zu Art und Ausmaß solcher situativen Effekte für die Interpretation der Verhaltensbeobachtungen besonders wichtig erscheinen.

Das Konzept der Reliabilität, das der Evaluation der vorgeschlagenen Verfahren zugrundeliegt, erscheint allerdings nicht geeignet, die offenkundig verschiedenen Bedingungen, unter denen die Schmerzeinschätzung de facto erfolgt angemessen zu adressieren. Beobachtete Variabilität, beispielsweise zwischen Beobachtern, Untersuchungsgruppen, Zeitpunkten oder Situationen, wird nahezu ausschließlich als Messfehler begriffen, wenn es darum geht die Reliabilität eines Verfahrens nachzuweisen, jedoch unkritisch als systematische Merkmalsvariation interpretiert, wenn Fragen nach der Validität der Instrumente beantwortet werden sollen.

Um zukünftig belastbare Aussagen zur Reliabilität und Validität der Verfahren zur Schmerzerfassung bei demenzkranken Menschen machen zu können, muss die Äquivalenz der Schmerzabbildung über die variierenden Erfassungskontexte und -bedingungen hinweg systematisch nachgewiesen werden. Nur so kann sichergestellt werden, dass beispielsweise verschiedene Beobachter ein Schmerzerleben auf dem Hintergrund eines vergleichbaren Konzeptes und Referenzrahmens von Schmerzen einschätzen, dass die durch eine analgetische Intervention reduzierten Skalenwerte tatsächlich als Hinweis auf die prädiktive Validität des Instrumentes gelten können, oder die im Vergleich zu schwer demenzkranken Menschen beobachtete geringere Anzahl potenziell schmerzbezogener Verhaltensweisen tatsächlich als vergleichsweise geringere Schmerzbelastung gewertet werden kann.

3.5.3 Vergleich konkurrierender Verhaltensinventare

Aufgrund der mitunter erheblichen Unterschiede im Umfang der vorgeschlagenen Verfahren, dem veranschlagten Abstraktionsniveau der Verhaltensindikatoren (bzw. -bereiche) und dem Skalenformat bleiben die Möglichkeiten einer direkten Gegenüberstellung der psychometrischen Eigenschaften verschiedener Skalen selbst dann noch außerordentlich eingeschränkt, wenn sie in derselben Stichprobe parallel eingesetzt wurden. Der mit einem parallelen Einsatz konkurrierender Verfahren verbundene *Mehraufwand* erscheint vor diesem Hintergrund nur dann zu rechtfertigen, wenn es zukünftig besser gelingt, diese Verfahren und ihre Indikatoren auch tatsächlich auf die angenommene *gemeinsame Merkmalsdimension Schmerz* zu beziehen.

Einige der damit beschriebenen methodischen Herausforderungen der Entwicklung und Bewertung von Instrumenten zur verhaltensgestützten Schmerzmessung bei demenziell erkrankten alten Menschen können voraussichtlich nur dann überwunden werden, wenn neuere testtheoretische Ansätze und die auf dieser Grundlage entwickelten psychometrischen statistischen Verfahren systematisch für dieses – gegenwärtig noch stark klinisch bestimmte – Anwendungsfeld nutzbar gemacht werden.

Die vorliegende Arbeit setzt sich darum zum Ziel, aktuelle Methoden der probabilistischen und latenten Modellierung empirischer Daten herauszustellen und die Potenziale bzw. den Mehrgewinn dieser Verfahren am Beispiel der verhaltensbezogenen Schmerzerfassung bei demenzkranken Menschen zu veranschaulichen. Der detaillierten Beschrei-

bung der konzeptionellen Grundlagen und Aussagebereiche verschiedener testtheoretischer Grundpositionen, sowie ihrer wechselseitigen Bezüge ist dementsprechend ein eigenes Kapitel gewidmet. Die in diesem zentralen Kapitel diskutierten theoretischen Potenziale für die Bewertung und Weiterentwicklung von Verfahren zur verhaltensbezogenen Schmerzmessung in der Population demenzkranker alter Menschen sollen anschließend anhand der beiden im Rahmen der zweiten HILDE-Projektphase eingesetzten Verhaltensinventare BESD und CNPI auch empirisch nachvollzogen werden.

4 Methodische Optionen für die Schmerzmessung

Im einleitenden Kapitel (2.2) wurde Schmerz als ein latentes, also nicht direkt beobachtbares Merkmal definiert. In der Diskussion um eine adäquate Messung von Schmerzen insbesondere in der Population Demenzkranker wurde deutlich, dass verschiedene Merkmalsqualitäten oder -zustände darum nur aufgrund verbaler Beschreibungen, über Analogien oder durch die Beobachtung von Korrelaten des Schmerzes erschlossen werden können.

Dieser Abschnitt stellt dar, welche messtheoretischen Modelle gegenwärtig die Praxis und Forschung zum Themenfeld Schmerz bei Demenz bestimmen, und welche Aussagen zum Merkmalsraum auf dieser Grundlage getroffen werden können. Dazu werden zunächst die Grundannahmen einer an der Klassischen Testtheorie (KTT; Lord & Novick, 1968) orientierten Interpretation von Messwerten dargestellt, die gegenwärtig in nahezu allen Bereichen der Schmerzforschung die Verfahren der Instrumentenentwicklung und -evaluation dominiert. Die Diskussion der konzeptionellen Grenzen dieses Ansatzes, z.B. mit Blick auf die Bestimmung verschiedener Gütekriterien, macht den Mehrwert alternativer, einer probabilistischen Messtheorie verpflichteten Interpretation schmerzbezogener Messwerte deutlich.

Obschon sich probabilistische Messmodelle über die letzten 50 Jahre hinweg weitgehend unabhängig – und weitestgehend unbemerkt – vom Mainstream der Schmerzforschung entwickelten, sind sie in den letzen Jahren durch ihre Eingliederung in eine übergeordnete Struktur statistischer Modellierung (Generalized Latent Variable Modeling; Skrondal & Rabe-Hesketh, 2004) einem breiteren Kreis von Forschenden zugänglich gemacht worden. Die in den letzten Jahrzehnten nahezu paradigmatische Kontrastierung der beiden testtheoretischen Ansätze weicht zunehmend einer integrierenden Perspektive, die wechselseitig die spezifischen Vorteile der verschieden restriktiven Orientierungen zu würdigen weiß. Nicht zuletzt unterstützt auch die Verfügbarkeit aktueller statistischer Softwarepakete, die zunehmend auch die Analyse nicht-metrischer (ordinaler und nominaler) Daten erlauben, den Transfer komplexerer Messmodelle aus Disziplinen wie der Psychometrie, Biostatistik oder Ökonomie in stärker anwendungsorientierte Forschungsfelder wie dem des geriatrischen Schmerzmanagements.

4.1 Schmerzmessung im Kontext der Zufallsstichprobentheorie

Für die Abbildung selbst erlebter oder bei anderen Personen angenommener Schmerzzustände und deren Intensität in ein numerisches Relativ (Messung) kommen eine ganze Reihe von Einzelitems mit unterschiedlichem Skalenniveau zum Einsatz. Zur Abschätzung und Steigerung der Messgenauigkeit werden darüber hinaus häufig mehrere Items als Skala vorgegeben und die Rohwerte zu einem Gesamtscore verrechnet. Insbesondere Verhaltensinventare umfassen mehrere potenziell unmittelbar beobachtbare Schmerzindikatoren aus verschiedenen Ausdrucksbereichen (z.B. Mimik, Gestik, Lautäußerungen) oder Verhaltensänderungen, die sich als Abweichungen vom bisher bekannten oder situativ erwartbaren Verhalten der betroffenen Personen dokumentieren lassen.

4.1.1 Grundmodell der Klassischen Testtheorie

Die klassische Testtheorie lässt sich als ein Spezialfall der Zufallsstichprobentheorie (Random Sampling Theory) verstehen, bei welcher der nicht direkt beobachtete, wahre Merkmalswert η_j (Truescore) als der Erwartungswert der beobachteten Messwerte y_j (Rohscores)

$$\eta_j \equiv \mathrm{E}(y_j) \tag{1}$$

über potenzielle Replikationen der Messung hinweg definiert wird.[2] In diesem Zusammenhang wird darum auch von der *Erwartungswert-Definition* des latenten Merkmals gesprochen. Der durch eine Messung für eine Person j gewonnene Wert y_j wird als fehlerbehaftete Messung der *wahren* Merkmalsausprägung η_j (d.h. des Truescores) der Person

$$y_j = \eta_j + \epsilon_j, \tag{2}$$

betrachtet und ein linearer Zusammenhang zwischen den konkret beobachteten Verhaltensweisen und den eigentlich interessierenden latenten Merkmalen einer Person postuliert. Die beschriebene Zerlegung von beobachtetem Wert in einen wahren und einen Fehleranteil stellt im eigentlichen Sinne kein Modell dar, das empirisch überprüft werden kann (vgl. Steyer et al., 1997), sondern beschreibt grundsätzliche Eigenschaften der Messfehler, weswegen häufig auch von einer Messfehlertheorie gesprochen wird.

Die klassische Testtheorie ist in ihrer ursprünglichen Form durch einige mitunter sehr restriktive *Annahmen* gekennzeichnet. Die mit der konkreten Messung für eine Person j verbundenen Messfehler ϵ_j werden dabei üblicherweise als mit $N(0, \theta_j)$ normalverteilt angenommen, d.h. der Erwartungswert des Messfehlers über potenzielle Replikationen der Messung hinweg entspricht

$$E(\epsilon_j) = 0. \tag{3}$$

Weiterhin wird angenommen, dass die Messfehler vom Truescore unabhängig sind, und eine Messung darum nicht in bestimmten Abschnitten des zu untersuchenden wahren Merkmalskontinuums mit einem geringeren oder größeren Messfehler behaftet ist als bei anderen Ausprägungsgraden. Wenn ρ eine Produkt-Moment-Korrelation und P eine (bedingte)

[2] Eine Übersicht der in dieser Arbeit verwendeten Notation ist als Anhang A beigefügt.

Wahrscheinlichkeit repräsentieren, kann dieser Zusammenhang durch

$$\rho_{\epsilon_j \eta_j} = 0 \qquad \text{oder} \qquad P(\epsilon_j|\eta_j) = P(\epsilon_j) \cdot P(\eta_j) \tag{4}$$

formal dargestellt werden. Es ist offensichtlich, dass diese Annahme wohl durch die wenigsten bekannten Instrumente der sozialwissenschaftlichen Forschung erfüllt sein dürfte. Deutlich wird der schwer einzulösende Anspruch dieser Forderung auch an folgendem Beispiel aus der physikalischen Messung. Während es mit einem gewöhnlichen 30cm Bürolineal problemlos möglich ist, die Länge eines Bleistiftes zu bestimmen, eignet sich dieses Messinstrument weniger gut, die Stärke eines Blattes oder die Länge des Büroraumes zu messen.

Eine weitere Annahme der klassischen Testtheorie betrifft die wechselseitige Unabhängigkeit der Messfehler der einzelnen Replikationen selbst,

$$\rho_{\epsilon\epsilon'} = 0, \tag{5}$$

die sich sowohl auf die wiederholte Erfassung eines Merkmales bei ein-und-derselben (fixen) Person, als auch auf die Messung zufällig ausgewählter Einheiten einer definierten Population von Merkmalsträgern beziehen können. Im Bereich der Schmerzmessung werden beide Perspektiven regelmäßig vertreten, auch wenn man annehmen darf, dass die stärker psychologisch orientierten Untersuchungen häufiger an der Belastung bestimmter (Sub-)Populationen interessiert sind, während der an einzelnen Individuen ausgerichteten Bestimmung von Schmerzen in der klinischen Schmerzforschung und -therapie eine besondere Bedeutung zukommt.

Wird ein interessierendes latentes Merkmal anhand von mehreren vergleichbaren Tests, Subtests oder Items $i = 1, 2, 3, \ldots k$ erhoben, so ergibt sich als Messstruktur das in Abbildung 6 dargestellte Faktoren-Modell.[3]

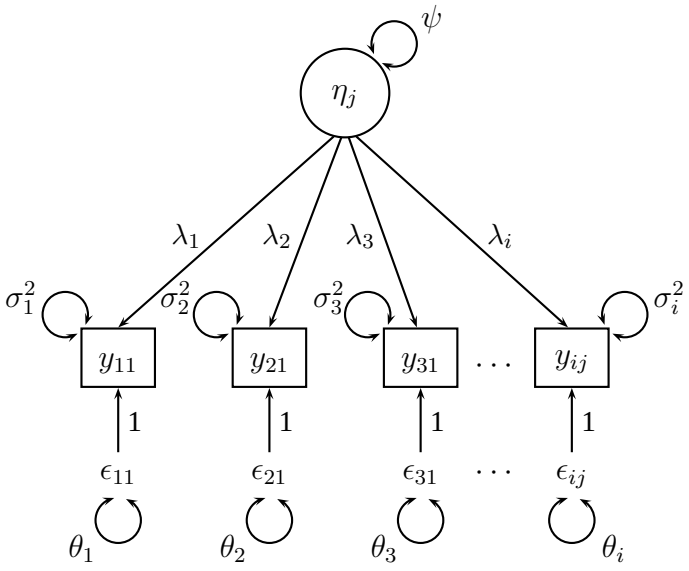
Die Richtung und Enge des linearen Zusammenhanges zwischen den einzelnen Indikatoren und dem Truescore werden dabei durch den Regressionsparameter λ_i repräsentiert, so dass sich die Grundgleichung der klassischen Testtheorie (2) zu

$$y_{ij} = \lambda_i \eta_j + \epsilon_{ij} \tag{6}$$

verallgemeinern lässt. Dabei ist der Grundgedanke der, dass die empirisch beobachteten Zusammenhänge zwischen den einzelnen Items durch das allen Einzelindikatoren gemeinsame latente Merkmal bestimmt werden. Würde man nur solche Personen betrachten,

[3] Die hier gewählte Darstellung des Messmodelles folgt dem Reticular Action Model (RAM; McArdle & Boker, 1990). Nach den im Kontext von Strukturgleichungsmodellen üblichen Konventionen werden beobachtete Variablen dabei durch Rechtecke, latente Variablen durch Kreise, und lineare Beziehungen durch gerichtete Pfeile repräsentiert. Kovarianzen werden durch gebogene Doppelpfeile zwischen Variablen angezeigt. Da (Residual-)Varianzen gewissermaßen als Kovarianz einer Variablen mit sich selbst verstanden werden können, werden diese ebenfalls durch einen kreisförmigen Doppelpfeil dargestellt. Zusätzlich können Konstanten mit einem Wert von 1 für alle Beobachtungen als Dreieck dargestellt werden, um über entsprechende Pfade zu Variablen Mittelwertsstrukturen grafisch zu veranschaulichen. Der Vorteil dieser Darstellungsweise gegenüber anderen Schematisierungen liegt in der Vollständigkeit der Spezifikation, die eine direkte Übersetzung in algebraische Gleichungssysteme erlaubt.

Abbildung 6: Messung eines latenten Merkmals durch mehrere beobachtete Indikatoren.



die bezüglich ihrer wahren Merkmalsausprägung vergleichbar sind, sollten keine Zusammenhänge mehr zwischen den beobachteten Variablen bestehen, weshalb auch von der Definition latenter Variablen über die *lokale stochastische Unabhängigkeit* ihrer Indikatoren gesprochen werden kann.

Die für dieses faktorielle Messmodell beschriebenen Parameter sind im Kontext der ursprünglichen Testtheorie in ihren Ausprägungen jedoch stark restringiert, da die klassische Testtheorie konzeptionell lediglich zwischen einem wahren Wert und einem nicht näher bestimmbaren Messfehler differenziert. Jede Variabilität in den beobachteten Variablen, die nicht bei allen Messungen gemeinsam vorzufinden ist, muss daher als Messfehler gelten. Alle Einzelmessungen des beschriebenen einfaktoriellen Messmodells stellen nach dieser Vorstellung strikt parallele Tests dar. Zusätzlich zu den vorgenannten Annahmen über die Verteilung der Fehlerwerte und deren Zusammenhang mit dem wahren Merkmalswert müssen darum weitere Bedingungen erfüllt sein, um von strikt parallelen Tests bzw. Items sprechen und mithilfe der mehrfachen Erfassung die Güte der Messung abschätzen zu können.

Da im Kontext der klassischen Testtheorie die Einzelitems als äquivalent, also beliebig untereinander austauschbar angenommen werden, soll an dieser Stelle auf die in anderen Publikationen gebräuchliche Differenzierung der Nomenklatur für durch Einzelitems, Subtests und den Gesamttest erhobenen Messwerte verzichtet werden. Sub- oder Gesamttestwerte errechnen sich als eine Linearkombination (Summe, Mittelwert oder jede andere lineare Transformation) der zugeordneten Einzelitems $y \equiv \sum_{i=1}^{k} w_i y_i$, wobei durch w_i eine spezifische Gewichtung der Einzelkomponenten repräsentiert werden kann.

4.1.2 Reliabilität

Im Kontext der klassischen Testtheorie stellt sich die Frage, wie gut der gemessene Wert y die wahre Merkmalsausprägung (Truescore η) repräsentiert. Dieser Zusammenhang kann theoretisch durch den Pearson'schen Korrelationskoeffizienten ρ zwischen Roh- und Truescore

$$\rho_{y_i\eta} = \frac{\psi_{y_i\eta}}{\sqrt{\sigma_i^2\psi}} = \frac{\psi_{(\eta+\epsilon_i)\eta}}{\sqrt{\sigma_i^2\psi}} = \frac{\psi + \psi_{\eta\epsilon_i}}{\sqrt{\sigma_i^2\psi}} \tag{7}$$

abgeschätzt werden und wird als *Reliabilitäts-Index* bezeichnet. Da angenommen wird, dass Truescore und Messfehler unkorreliert sind, entfällt der letzte Term des Zählers. Das Quadrat des Reliabilitäts-Index wird als *Reliabilitätskoeffizient* bezeichnet und stellt den Anteil der wahren Merkmalsvariation an der beobachteten Variabilität der Messwerte dar:

$$\rho_{y_i\eta}^2 = \frac{\psi^2}{\sigma_i^2\psi} = \frac{\psi}{\sigma_i^2} = \frac{\psi}{\psi + \theta_i}. \tag{8}$$

Im Gegensatz zum Reliabilitätsindex, der theoretisch Werte zwischen -1 und +1 annehmen könnte, liegt der theoretische Wertebereich des Reliabilitätskoeffizienten zwischen 0 und 1.

Da die wahren Merkmalswerte als latente Variablen nicht direkt beobachtet werden können, ist eine Abschätzung der Reliabilität auf der Grundlage einer einzelnen Messung nicht möglich. Zur Bestimmung der Güte einer Messung muss diese in vergleichbarer Weise, d.h. durch parallele Erfassungen wiederholt durchgeführt werden. Der wahre Merkmalswert einer Person wird über die beiden Testungen hinweg als konstant angenommen, und die Korrelation der beiden Messwertreihen y_i und $y_{i'}$ lautet

$$\rho_{y_iy_{i'}} = \frac{\psi_{y_iy_{i'}}}{\sqrt{\sigma_i^2\sigma_{i'}^2}} = \frac{\psi_{(\eta+\epsilon_i)(\eta+\epsilon_{i'})}}{\sqrt{\sigma_i^2\sigma_{i'}^2}} = \frac{\psi + \psi_{\eta\epsilon_i} + \psi_{\eta\epsilon_{i'}} + \psi_{\epsilon_i\epsilon_{i'}}}{\sqrt{\sigma_i^2\sigma_{i'}^2}}, \tag{9}$$

wobei durch die zuvor beschriebene Annahme der Unkorreliertheit von Truescore und Messfehler die beiden mittleren Terme des Zählers gleich null sind, so dass sich die Gleichung zu

$$\rho_{y_iy_{i'}} = \frac{\psi + \psi_{\epsilon_i\epsilon_{i'}}}{\sqrt{\sigma_i^2\sigma_{i'}^2}} \tag{10}$$

vereinfacht.

4.1.2.1 Voraussetzung paralleler Tests

Damit aufgrund dieser Korrelation die Reliabilität der Messung abgeschätzt werden kann, müssen die Tests die Voraussetzungen strikt paralleler Tests erfüllen. Zum ersten werden dabei die Messfehler ϵ_i der parallelen Einzelitems als voneinander unabhängig angenommen

$$\rho_{\epsilon_i\epsilon_{i'}} = 0, \tag{11}$$

wodurch der zweite Term des Zählers in Gleichung 10 entfällt. Zum zweiten müssen die parallelen Items y_i und $y_{i'}$ identische Varianzen

$$\sigma_i^2 = \sigma_{i'}^2 = \sigma^2 \tag{12}$$

aufweisen, womit sich Gleichung (10) zu

$$\rho_{y_i y_{i'}} = \frac{\psi}{\sigma_i^2} = \frac{\psi}{\psi + \theta_i} = \rho_{y_i \eta}^2, \tag{13}$$

und damit also zum gesuchten Reliabilitätskoeffizienten vereinfachen lässt.

Durch die sehr restriktiven Paralleltest-Annahmen der klassischen Testtheorie wird die in Abbildung 6 beschriebene allgemeine Messstruktur auf einen empirisch eher unwahrscheinlichen Spezialfall reduziert. Da die Messfehler der Einzelmessungen als wechselseitig unabhängig angenommen werden, sind in der grafischen Veranschaulichung keine Kovarianzen (Doppelpfeile) zwischen den Residualvarianzen enthalten. Eine weitere Parameterrestriktion ergibt sich aus der Annahme identischer Varianzen σ^2 der beobachteten Variablen. Da es im Kontext der klassischen Testtheorie nur einen einzigen wahren Reliabilitätswert geben kann, und der Truescore für beide Messungen identisch ist, müssen parallele Messungen schließlich einen gleich engen Zusammenhang zwischen beobachtetem und wahren Wert anzeigen. Die in Abbildung 6 dargestellten Regressionsgewichte λ_i werden darum für alle Einzelindikatoren auf den gleichen Wert $\lambda_i = \lambda_{i'}$ restringiert, damit für beide parallele Messungen identische Reliabilitäten

$$\begin{aligned} \rho_{y_i \eta}^2 &= \rho_{y_{i'} \eta}^2 \\ \frac{\psi}{\sigma_i^2} &= \frac{\psi}{\sigma_{i'}^2} \\ \frac{\psi}{\lambda_i^2 \psi + \theta_i} &= \frac{\psi}{\lambda_{i'}^2 \psi + \theta_{i'}} \end{aligned} \tag{14}$$

geschätzt werden.

Zur Gewinnung paralleler Messungen wurden im Kontext der klassischen Testtheorie verschiedene Verfahrensweisen vorgeschlagen. Eine Messung kann z.B. dann als parallel angenommen werden, wenn der identische Test an denselben Merkmalsträgern wiederholt eingesetzt wird. Selbstverständlich ist der Schluss von diesem *Retest*-Reliabilitätskoeffizienten auf die Enge des Zusammenhanges zwischen Truescore und Messwert nur insoweit gültig, wie sich der wahre Merkmalswert der Personen zwischen beiden Messungen nicht verändert. Wenn man verhindern will, dass Probanden bei einer zweiten Erhebung höhere Werte erreichen, da sie sich an die Items erinnern, können equivalente, aber eben nicht identische Testformen eingesetzt werden. Auch wenn diese *equivalenten* Testversionen sorgfältig entwickelt wurden, kann deren tatsächliche Parallelität lediglich unterstellt, nicht aber empirisch bewiesen werden. Besteht ein Test aus mehreren Items, werden häufig alle Einzelitems als parallele Messungen desselben Merkmales aufgefasst und ihre wechselseitigen Korrelationen zur Abschätzung der Reliabilität des Gesamttestwertes herangezogen.

Der bekannteste Kennwert der so abgeschätzten *internen Konsistenz* einer Skala ist dabei der von Cronbach vorgestellte Koeffizient α für metrische und der als Kuder-Richardsons KR-20 bekannte Kennwert für dichotome Daten:

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma^2}), \tag{15}$$

$$\text{KR-20} = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k} pq}{\sigma^2}) \tag{16}$$

Auch der Split-Half-Reliabilitätskoeffizient wird nach diesem Grundschema berechnet. Jeder dieser Kennwerte stellt dabei lediglich einen mehr oder minder plausiblen Schätzer für die *eine* wahre Reliabilität der Messung dar.

Die Quadratwurzel der Fehlervarianz θ_i kann als Standardfehler der Messung ($S.E.$) zur Konstruktion von Konfidenzintervallen

$$(y_{ij} - z_\alpha \, S.E.) < \eta_j < (y_{ij} + z_\alpha \, S.E.) \tag{17}$$

um einen beobachteten Testwert herum verwendet werden. Da für diesen Test theoretisch nur eine einzige Reliabilität existiert, bleibt das Konfidenzband dieser Intervallschätzung über alle anzunehmenden wahren Merkmalswerte hinweg identisch breit, der Messfehler ist also in Bereichen extrem geringer und hoher Merkmalsausprägung genauso groß wie im mittleren Bereich. Dabei ist leicht nachzuvollziehen, dass ein gegebener Test zwar verschieden schwierige Items umfassen mag, die Personen mit unterschiedlichen wahren Merkmalswerten voneinander zu unterscheiden erlauben; dennoch erscheint es unrealistisch, dass mit einem konkreten Test über alle interessierenden Bereiche des latenten Merkmalskontinuums hinweg eine gleich präzise Differenzierung wahrer Merkmalswerte geleistet werden kann.

4.1.2.2 Möglichkeiten zur Beschreibung weiterer Gütekriterien

In der psychometrischen Literatur werden häufig Haupt- und Nebengütekriterien einer Messung unterschieden. Als Hauptgütekriterien gelten dabei Objektivität, Reliabilität und Validität, während Aspekte der Ökonomie, Normierung, Vergleichbarkeit und Nützlichkeit einer Messung als Nebengütekriterien beschrieben werden. Zuvor wurde eingehend dargelegt, dass Merkmalen des Kontextes (z.B. der institutionellen Versorgungssituation) bei der Schmerzerfassung bei demenzkranken Menschen eine besondere Bedeutung zukommt. Entsprechend wichtig ist es darum, methodische Verfahren zur Abschätzung der Effizienz und Praktikabilität eines Verfahrens zur Verfügung zu haben. Um klären zu können, welchen Aufwand (beispielsweise durch Schulungen oder die Sicherung bestimmter Beobachtungsvoraussetzungen) Einrichtungen in Kauf nehmen müssen, um zu einer hinreichend reliablen Abbildung der Schmerzbelastung der Bewohner zu gelangen, müssen Verfahren darum aber insbesondere den alltagspraktischen Kontext der Schmerzmessung berücksichtigen.

Die klassische Testtheorie im engeren Sinn kann über die Bestimmung der Reliabilität einer Messung (genauer gesagt *strikt paralleler Messungen*) hinaus nur sehr begrenzt Auskunft über weitere Eigenschaften der Messung geben.

Die als Reliabilitätsindex bezeichnete Korrelation zwischen den beobachteten und den wahren Merkmalswerten $\rho_{y_i\eta}$ (siehe Gleichung 7) kann als obere Grenze der Korrelation des Testwertes mit einem beliebigen Kriterium verstanden werden, stellt also das Maximum nachweisbarer Validität des Messwertes dar.

Eine explizite Diskussion weiterer Bedingungen der Messung, wie sie beispielsweise für die Abschätzung der Objektivität eines Messverfahrens notwendig sind, wird erst durch die Lockerung einiger restriktiver Annahmen der klassischen Testtheorie im Rahmen der Generalisierungstheorie (Generizability Theory; Brennan, 2001) ermöglicht.

4.1.3 Erweiterung der KTT durch die Generalisierungstheorie

Da die Messfehler der klassischen Testtheorie konzeptionell lediglich ein unspezifisches Residuum repräsentieren, das keine weitere Differenzierung verschiedener Fehlerquellen erlaubt, ist es in ihrem Kontext zunächst auch nicht möglich, weitere Bedingungen der Messung (z.B. verschiedene Untersuchungsbedingungen oder Rater) zu berücksichtigen. Insofern bleiben die Annahmen dieser Konzeption auf standardisierte, also mit Blick auf die Untersuchungsbedingungen (Instruktion, Beleuchtung, Aktiviertheit, Rater etc.) konstant gehaltene Erfassungssituationen beschränkt, wie sie vor allem bei Paper-and-pencil-Tests angestrebt werden. Im Rahmen der Schmerzmessung jedoch sind die Abhängigkeiten der ermittelten Werte vom Kontext der Erfassung von zentraler Bedeutung. Da angenommen werden kann, dass die Schmerzbelastung aufgrund von körperlichen Vulnerabilitäten bei Bewegung bzw. gesteigerter Aktivität stärker ist als in Ruhesituationen, sehen einige der verfügbaren Instrumente eine wiederholte Schmerzerfassung in Situationen unterschiedlicher Aktiviertheit vor. Weitere sinnvolle Differenzierungen des Erfassungskontextes betreffen verfügbare Informanten (z.B. Pflegende und Angehörige), unterschiedliche Beobachtungszeitpunkte (z.B. morgens und abends) oder verschiedene Ausdrucksbereiche (z.B. Mimik, Gestik, Vokalisation).

Durch eine Lockerung der Annahme strikt paralleler Tests der KTT kann die in Abbildung 6 dargestellte Messstruktur als eine allgemeine (einfaktorielle) Messstruktur begriffen werden, bei dem die Rohwerte nicht wie zuvor (bis auf den Messfehler) vollständig und in gleichem Maße den wahren Merkmalswert repräsentieren, sondern nur zu einem bestimmten, über verschiedene Items üblicherweise variierenden Anteil durch ein gemeinsames latentes Konstrukt bestimmt sind. Die Richtung und Enge des linearen Zusammenhanges zwischen Indikator und Konstrukt

$$y_{ij} = \lambda_i \eta_j + \epsilon_{ij} \qquad \text{Wdh. (6)}$$

werden dabei hier durch einen itemspezifisch variierenden Regressionsparameter λ_i abgebildet, der unter der Annahme paralleler Tests im Rahmen der ursprünglichen klassi-

schen Testtheorie auf jeweils den gleichen Wert $\lambda_i = \lambda_{i'}$ restringiert wurde. Die einzelnen Indikatoren werden dabei nun nicht mehr als strikt parallele Tests begriffen, sondern als eine Zufallsauswahl aus der Gesamtheit der möglichen Testfragen oder Verhaltensindikatoren, warum Nunnaly (1978) diesbezüglich auch von einem „domain-sampling"-Ansatz spricht. Da die Einzelitems damit unterschiedlich gute Repräsentanten der wahren Merkmalswerte sind, kann für jede Einzelmessung eine eigene Reliabilität beschrieben werden. Diese Unterschiedlichkeit der Indikatoren kann nun aber selbst Gegenstand psychometrischer Analysen werden. Während zuvor alle Variation in den beobachteten Werten, die nicht wahre Merkmalsvariation anzeigte, als unsystematische Residualvarianz begriffen wurde, können die Fehlerwerte ϵ_{ij} im Kontext der Generalisierungstheorie neben der eigentlichen Ungenauigkeit der Messung nun konzeptionell zusätzlich einen eigenständigen, vom indizierten latenten Konstrukt unabhängigen (uniquen) Anteil umfassen. Je nach Fragestellung und Untersuchungsdesign können diese systematischen Anteile in den Residualvarianzen inhaltlich (z.B. als Variation zwischen Indikatorengruppen oder Situationen) näher bestimmt werden.

Dieses Messmodell liegt dem wohl größten Teil der psychologischen Forschung zugrunde, da die beobachteten Verhaltensindikatoren das in Frage stehende latente Konstrukt in den meisten Fällen nur zu jeweils einem bestimmten Teil anzeigen, und die Beziehung zwischen den beobachteten und den wahren Werten ganz wesentlich durch weitere Personen- oder Situationsmerkmale bestimmt sind, die ihrerseits einen weiteren Untersuchungsgegenstand darstellen *können*. Die Entscheidung darüber, welche potenziellen Einflussfaktoren auf die Messung als Messfehler oder aber als inhaltlich relevante Facetten der (situativen) Bedingtheit der Messung interpretiert werden sollen, kann nicht durch eine Testtheorie, sondern muss vom Forschenden geleistet werden.

4.1.3.1 Bestimmung des Messgegenstandes

Das Ziel der Messung im Kontext der KTT ist die Bestimmung der wahren Personenmerkmale η_j, die somit den einzigen Gegenstand der Messung darstellen. Alle Variation in den beobachteten Werten, die sich nicht auf wahre Merkmalsunterschiede der Personen zurückführen lässt, wird entsprechend als Fehlervarianz interpretiert. Um den Fehleranteil der Messungen möglichst gering zu halten, werden nach Möglichkeit sämtliche potenzielle situativen Einflußfaktoren eliminiert oder zumindest konstant gehalten.

Werden im Rahmen einer Generalisierungsstudie bestimmte Bedingungen der Messung, wie beispielsweise die Informationsquelle für ein Schmerzrating explizit berücksichtigt, ergeben sich mehrere Möglichkeiten, den Gegenstand der Messung zu bestimmen. Wird z.B. die aktuelle Schmerzbelastung eines einzelnen Heimbewohners durch Pflegende und Angehörige als sehr unterschiedlich eingeschätzt, so trägt diese Unterschiedlichkeit bei der Bestimmung der wahren Schmerzbelastung des Bewohners zu einer hohen Fehlervarianz bei. Ist man aber daran interessiert, wie verschiedene Informantengruppen das Schmerzerleben der Bewohner einschätzen, werden die Abweichungen in den Einschätzungen als sinnvolle Variabilität interpretiert. Ein weiteres Beispiel aus dem Be-

reich der Schmerzerfassung ist die wiederholte Beobachtung von Schmerzindikatoren in Situationen geringer und hoher Aktiviertheit, in denen häufig sehr verschiedene Schmerzbelastungen dokumentiert werden.

4.1.3.2 Facetten des Erhebungsdesigns

Eine Differenzierung des Messgegenstandes nach einem bestimmten Operationalisierungsmerkmal wird als Facette der Messung bezeichnet. Werden zur Abbildung der Schmerzbelastung von Pflegeheimbewohnern beispielsweise fünf Schmerzitems eingesetzt, spricht man von einem Messmodell mit einer fünfstufigen Facette, werden zusätzlich zwei unterschiedliche Informanten (z.B. zwei Pflegepersonen) als weitere Messbedingung berücksichtigt, ergibt sich ein 2-Facetten-Design und so weiter. Wird der Messgegenstand von den im Erfassungsdesign konkret umgesetzten Bedingungen der Messung auf das individuelle Schmerzerleben der Bewohner verallgemeinert, so fließen beispielsweise die Unterschiede zwischen den Items und den Beurteilern in den Messfehler mit ein.

Stufen einer Messfacette als Ausprägung einer Zufallsvariable

Die in einem konkreten Untersuchungsplan umgesetzten Stufen der Facetten eines Messmodelles können in einem ersten Fall als zufällige Realisationen aus der Gesamtheit aller möglicher Stufen des Differenzierungsmerkmales konzeptualisiert sein, sodass bei zukünftigen Messungen andere Ausprägungen zu erwarten sind. Als Realisation einer Zufallsvariablen kann der Einfluss des Differenzierungsmerkmales auf die Messung damit als „random effect“ bezeichnet werden. Durch die Zufallsauswahl von Personen aus einer Grundgesamtheit oder über zufällige Replikationen der Messungen einer einzelnen Person ist der durch die klassische Testtheorie bestimmte Messgegenstand (Merkmalswerte einer Population oder einer Person) als Zufallsvariable bestimmt, auch wenn mit Blick auf die vollständige Vernachlässigung des Kontextes der Messung hier gewöhnlich nicht von einer Facette i.S. der Generalisierungstheorie gesprochen wird. Als Repräsentanten eines Universums vergleichbarer Indikatoren werden auch die Einzelitems eines Tests i.d.R. als zufällig betrachtet.

Stufen einer Messfacette als fixe Variable

Im anderen Fall stellen die konkret umgesetzten Stufen der Facette der Messung die einzigen interessierenden Ausprägungsgrade bzw. Realisationen des Kontextmerkmales der Messung dar. Im Kontext der schmerzbezogenen Verhaltensbeobachtung stellt z.B. der Aktivierungsstatus der Zielperson ein theoretisch bedeutsames Differenzierungsmerkmal der Messsituation dar. Aus allen denkbaren Graden von Aktiviertheit werden dabei gewöhnlich eine Situation mit keiner bzw. möglichst geringer körperlicher Aktivität und eine mit einem möglichst hohen Niveau von Bewegung und Anstrengung als Beobachtungsbedingungen festgelegt. Damit wird die Schmerzbeobachtung gewissermaßen

bezüglich des konfundierenden Merkmales der Aktiviertheit standardisiert, und der abgeschätzte Einfluss der betrachteten Situationsmerkmale auf die Messung kann als „fixed effect“ bezeichnet werden. Da durch die Standardisierung der Erfassungsmodalitäten ein Teil der nicht durch die wahren Merkmalsunterschiede zwischen einzelnen Personen bestimmten Variabilität in den Messwerten (d.h. der Fehlervarianz) verhindert (eine Stufe) bzw. gebunden wird (mehrere Stufen), trägt eine solche fixe Facette zur Erhöhung der Reliabilität und Validität der Messung bei. Die so gewonnenen Schmerzbeurteilungen können dann jedoch auch nur auf entsprechende Situationen sehr geringer und sehr hoher Aktiviertheit generalisiert werden, wodurch der Messgegenstand selbst gegebenenfalls unerwünscht stark eingeschränkt und die inhaltliche Validität der Messung beeinträchtigt wird. Eine Analyse vor dem Hintergrund der Generalisierungstheorie macht selbstverständlich nur dann Sinn, wenn zumindest eine der für eine Messung spezifizierten Bedingungen als Zufallsvariable angenommen wird. Als Beispiel kann hier wieder die Einschätzung des Schmerzverhaltens von Bewohnern in einer Ruhesituation durch die Pflegenden einer Einrichtung herangezogen werden. Würde man hier alle drei die Messung bestimmenden Faktoren, also Merkmalsträger, Beurteiler und Ruhesituation als fixe Erfassungsbedingungen annehmen, so wäre der für einen bestimmten Bewohner *Herrn Muster* durch die bestimmte Pflegeperson *Schwester Hilde* eingeschätze Schmerzwert eine für diese bestimmte Aktivierungssituation *Ruhe* per definitionem perfekt reliable Messung, ließe jedoch keine Art von Generalisierung auf andere Bewohner, Rater oder Aktivierungsgrade zu. Ein wesentliches Ziel einer psychometrisch fundierten Instrumentenentwicklung liegt damit in einer möglichst optimalen Abwägung zwischen der Präzision und Generalisierbarkeit der Messung.

Es ist offensichtlich, dass sich auch die Definition des Truescores als der unter Berücksichtigung verschiedener Erfassungsbedingungen erwartbare Wert in Abhängigkeit vom gewählten Messmodell und Untersuchungsdesign verändert. Im Gegensatz zur klassischen Testtheorie gibt es im Kontext der Generalisierungstheorie also nicht nur einen einzigen wahren Merkmalswert, sondern so viele verschiedene Truescores, wie Kontextmerkmale der Messung berücksichtigt und spezifische Fragestellungen formuliert werden.

4.1.3.3 Varianzdekomposition (G-Studie)

Nachdem aus der Sichtung der bisherigen Forschung und der Theorie ein zunächst tentatives Messmodell konzeptualisiert wurde, bei dem solche Merkmale, von denen angenommen werden kann, dass sie die Messung konfundieren könnten, als Facetten des Erhebungsdesigns mit berücksichtigt sind, ergibt sich die Aufgabe, das Ausmaß dieser Fehlervariation abzuschätzen und mit der anzunehmenden wahren Merkmalsvariation in Beziehung zu setzen. Auf der Grundlage dieser geschätzten Varianzkomponenten können Aussagen zur Präzision und Generalisierbarkeit verschiedener Operationalisierungen der Messung getroffen werden, weshalb dieser Analyseschritt auch als *Generalisierungs-* oder *G-Studie* bezeichnet wird. Beispielsweise könnte man die Frage stellen, inwieweit die Präzision einer Schmerzeinschätzung gesteigert werden kann, wenn nicht eine einzelne

Pflegeperson, sondern zwei Pflegende einen Bewohner parallel einschätzen, und diese unabhängigen Urteile anschliessend miteinander verrechnet werden. Die Facette *Beurteiler* hätte dann zwei zufällig ausgewählte (random) Pflegepersonen als ihre Stufen. Gelangen die Pflegenden zu sehr unterschiedlichen Einschätzungen, dürfte man weniger Vertrauen in die Urteile einzelner Pflegenden haben.

Das statistische Verfahren zur Abschätzung der potenziellen Messfehler, die sich durch die Vernachlässigung verschiedener die Messung möglicherweise beeinflussender Merkmale ergeben können, ist die Varianzanalyse (ANOVA). Die Facetten der Messung entsprechen dabei den mehrfach gestuften Faktoren eines Analysemodelles, das Objekt der Messung stellt die abhängige Variable dar. Die interessierenden Varianzkomponenten in der Population können durch die mit den Haupt- und Interaktionseffekten verknüpften mittleren Quadratsummen anhand der Stichprobe geschätzt werden. Im Gegensatz zur herkömmlichen Varianzanalyse ist die Konstruktion von Tests auf Haupt- und Interaktionseffekte bzw. Mittelwertsunterschiede zwischen einzelnen Stufen der Faktoren (z.B. zwischen verschiedenen Items) in der G-Studie von nachrangiger Bedeutung.

Im einfachsten Fall wird ein einzelner Erfassungsaspekt (z.B. der Einsatz mehrerer Items in einer Skala) in seiner Bedeutung für die Messung des interessierenden Konstruktes betrachtet, und alle Merkmalsträger unter allen Stufen des Erfassungsdesigns erhoben, d.h. alle Personen $j = 1, 2, 3 \ldots N$ bearbeiten alle $i = 1, 2, 3 \ldots k$ Items. Die durch dieses gekreuzte Erhebungsdesign gewonnene Datenstruktur (*engl.* fully crossed design) ist in Abbildung 7 schematisch dargestellt.

Abbildung 7: Vollständig gekreuzte Datenstruktur

	Items					
Personen	Item 1	Item 2	Item 3	$\cdots$	Item i	
Person 1	y_{11}	y_{21}	y_{31}	$\cdots$	y_{i1}	$\bar{y}_1$
Person 2	y_{12}	y_{22}	y_{32}	$\cdots$	y_{i2}	$\bar{y}_2$
Person 3	y_{13}	y_{23}	y_{33}	$\cdots$	y_{i3}	$\bar{y}_3$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Person j	y_{1j}	y_{2j}	y_{3j}	$\cdots$	y_{ij}	$\bar{y}_j$
	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\cdots$	$\bar{y}_i$	$\bar{y}$

Der mittlere Testscore $\bar{y}_j$ einer Person über alle Items hinweg, der mittlere Wert eines Items $\bar{y}_i$ über alle Personen hinweg, und schließlich der Gesamtmittelwert $\bar{y}$ über alle Personen und Items hinweg ergeben sich zu

$$\bar{y}_j = (\sum_{i=1}^{k} y_{ij})/k, \quad \bar{y}_i = (\sum_{j=1}^{N} y_{ij})/N \quad \text{und} \quad \bar{y} = (\sum_{j=1}^{N} \sum_{i=1}^{k} y_{ij})/Nk. \tag{18}$$

Die Unterschiedlichkeit in den beobachteten Werten resultiert aus den drei möglichen Varianzquellen Person, Item und Interaktion zwischen Person und Item. Diese additiven Va-

rianzkomponenten können berechnet werden als

$$SS_{Person} = SS_j = k\sum_{j=1}^{N}\bar{y}_j^2 - Nk\bar{y}^2, \tag{19}$$

$$SS_{Item} = SS_i = N\sum_{i=1}^{k}\bar{y}_i^2 - Nk\bar{y}^2 \quad \text{und} \tag{20}$$

$$SS_{Interaktion} = SS_{ij} = \sum_{j=1}^{N}\sum_{i=1}^{k}y_{ij}^2 - N\sum_{i=1}^{k}\bar{y}_i^2 - k\sum_{j=1}^{N} + Nk\bar{y}^2. \tag{21}$$

Werden diese Quadratsummen (SS) an ihren jeweiligen Freiheitsgraden relativiert, lassen sich die resultierenden mittleren Quadratsummen (MS)

$$MS_j = SS_j/(N-1), \quad MS_i = SS_i/(k-1) \quad \text{und} \quad MS_{ij} = SS_{ij}/(N-1)(k-1) \tag{22}$$

als Stichprobenkennwerte für die Schätzung der Varianzkomponenten

$$\sigma_j^2 = (MS_j - MS_{ij})/k, \quad \sigma_i^2 = (MS_i - MS_{ij})/N \quad \text{und} \quad \sigma_{ij}^2 = MS_{ij} \tag{23}$$

in der Population heranziehen. Die für dieses Messdesign beschriebene Varianzkomponente σ_i^2 beschreibt die Variabilität der Gesamtheit aller erfassten Werte, die durch die Verwendung verschiedener Items erklärt werden kann und darf deshalb nicht mit der für ein Einzelitem i beobachteten Varianz σ_i^2 verwechselt werden.

Mit der Bestimmung der Varianzanteile für bestimmte Komponenten des Messdesigns ist die G-Studie abgeschlossen. Weiterführende Aussagen zur Güte der Messung unter bestimmten Voraussetzungen werden erst durch zusätzliche Spezifikationen dessen möglich, was im Rahmen der jeweiligen Untersuchung überhaupt als Messfehler begriffen werden soll. Diese Bestimmung und der Vergleich alternativer Messmodelle sind Gegenstand der sich an die G-Studie anschließenden D-Studie.

4.1.3.4 Vergleich von Szenarien (D-Studie)

Im beschriebenen Beispiel wurden mehrere Items zur Erfassung eines latenten Merkmals verwendet. Jedes Item stellt dabei eine Realisation aus dem Gesamtpool möglicher vergleichbarer Indikatoren dar. Im Gegensatz zur klassischen Testtheorie zeigen die einzelnen Items nur zu einem bestimmten Teil ein gemeinsames latentes Konstrukt an. Darüber hinaus aber weist jedes Item einen eigenständigen (uniquen) Varianzanteil auf. Die in diesem Szenario beobachteten Unterschiede der erhobenen Werte y_{ij} können auf Merkmalsunterschiede zwischen verschiedenen Personen (Person j ist stärker schmerzbelastet als Person j'), auf die unterschiedlichen Items als Facetten der Messung (Item i ist ein besserer Indikator für Schmerz als Item i') und auf die Interaktion von Items und Personenmerkmalen (Item i ist für die Person j ein besonders guter Schmerzindikator, nicht aber für Person j') zurückgeführt werden.

Die Bezeichnung D-Studie (aus dem Englischen: Decision Study) deutet darauf hin, dass es das Ziel dieses Analyseschrittes ist, eine empirisch gestützte Entscheidung bezüglich des für eine Fragestellung optimalen Messverfahrens treffen zu können. Dabei ist es notwendig, den Einfluß abzuschätzen, den eine Messfacette (z.B. eine spezifische Auswahl von Items, Beurteilern oder Erfassungskontexten) auf den Gegenstand der Messung besitzt, indem beispielsweise mehrere potenzielle Erfassungsdesigns (Szenarien) miteinander verglichen werden.

Die Unterschiedlichkeit der einzelnen Merkmalsträger σ_j^2 ist im Zuge der Messung üblicherweise die erwünschte Zielgröße, und kann als wahre Merkmalsvariation ψ begriffen werden. Die *relative Fehlervarianz* wird im Rahmen der Generalisierungstheorie mit $\sigma^2(\delta)$ bezeichnet, und kann je nach betrachtetem D-Studien-Szenario unterschiedlich zusammengesetzt sein. Nachfolgend sollen zwei solcher Szenarien beschrieben werden, anhand derer die Güte der durch das Messmodell mit einer Facette (Items) erfassten Item- oder Gesamttestwerte abgeschätzt werden kann.

Itemscore als Messung des wahren Merkmalswertes

Im einfachsten Fall möchte man abschätzen, wie gut ein einzelnes zufällig herausgegriffenes Einzelitem den wahren Merkmalswert repräsentiert. Der Kennwert wird in Analogie zum Reliabilitätskoeffizienten der klassischen Testtheorie als Anteil der wahren Merkmalsvarianz an der beobachteten Varianz der Messwerte

$$\rho^2 = \frac{\sigma_j^2}{\sigma_j^2 + \sigma^2(\delta)} \equiv \frac{\psi}{\psi + \theta} \tag{24}$$

konzeptualisiert. Um deutlich zu machen, dass es sich dabei – anders als in der klassischen Testtheorie – um einen von mehreren möglichen Reliabilitätskoeffizienten für ein-und-denselben empirischen Datensatz handelt, wird dieser Kennwert als Generalisierungskoeffizient oder einfach G-Koeffizient bezeichnet. Wenn wir davon ausgehen, dass für unser einfaches Beispiel kein externes Kriterium (wie beispielsweise eine medizinische Diagnose o.ä.) zur Bestimmung der Güte der Messung verfügbar ist, können die erhobenen Werte nur norm-basiert, also relativ zueinander interpretiert werden. In diesem Fall wird die in der G-Studie geschätzte Varianzkomponente für die Interaktion von Items und Merkmalsträgern σ_{ij}^2 als relative Fehlervarianz $\sigma^2(\delta)$ interpretiert. Ein Teil der beobachteten Variabilität in den Messwerten, nämlich derjenige Varianzanteil, der aus den Mittelwertsunterschieden zwischen einzelnen Items resultiert (σ_i^2), spielt also für die norm-basierte Abschätzung der Reliabilität der Messung keine Rolle, da nur die relativen, nicht aber die absolut erreichten Messwerte berücksichtigt werden.

Testscore als Messung des wahren Merkmalswertes

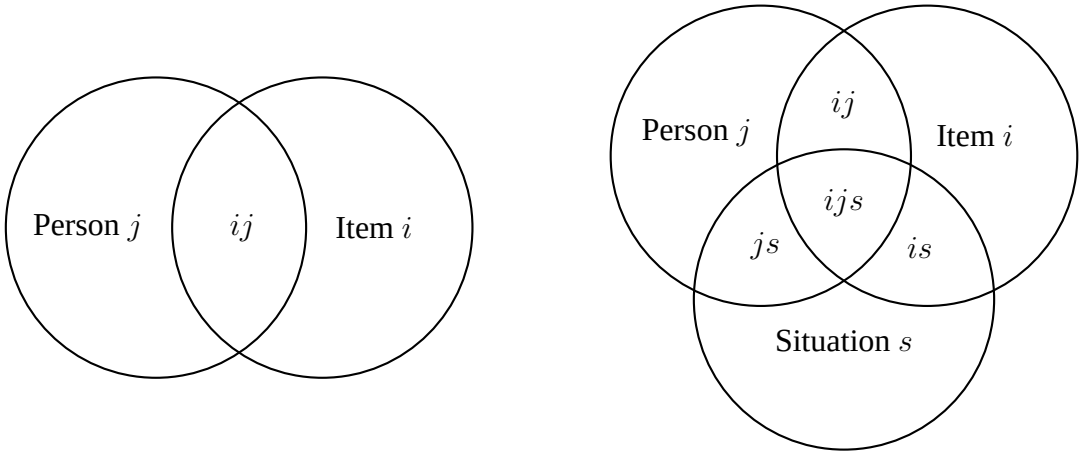
Werden wie in unserem Beispiel mehrere gleichsinnige Items zur Messung eines gemeinsamen latenten Merkmales eingesetzt, dann ist vor allem die Reliabilität des Gesamttestscores von Interesse, der sich als Summe der Einzelitems oder Subtests (oder einer

anderen linearen Transformation, wie z.B. dem Mittelwert) ergibt. Die relative Fehlervarianz $\sigma^2(\delta)$ besteht in diesem Szenario aus dem an der Itemanzahl relativierten Varianzanteil für die Item-Person-Interaktion σ^2_{ij}/k. Aufgrund der Division durch die Anzahl der Items im Test reduziert sich der relative Fehlervarianzanteil entsprechend, und als Folge ergibt sich eine im Vergleich zum Einzelitem höhere Messgenauigkeit bzw. Reliabilität des Testscores. Der für dieses 1-Facetten-Messmodell berechnete G-Koeffizient ist mit dem Reliabilitätskoeffizienten Cronbach Alpha der klassischen Testtheorie numerisch identisch. Genauer gesagt lassen sich die Reliabilitätskoeffizienten der Klassischen Testtheorie (Cronbach's α bzw. Kuder-Richardson's KR-20) als Spezialfall einer D-Studie verstehen, bei welcher der mit einem Testscore verbundene Messfehler in einer normbasierten Perspektive geschätzt wird.

4.1.3.5 Messdesigns mit mehr als einer Facette

Der konzeptionelle Mehrgewinn der Generalisierungstheorie gegenüber der klassischen Testtheorie wird sicherlich erst dann deutlich erkennbar, wenn in einem mehrfaktoriellen Erhebungsdesign zwei oder mehr Facetten der Messung berücksichtigt werden. Beispielsweise könnte eine Itembatterie zur Schmerzmessung in zwei Situationen unterschiedlicher Aktivierung ($s = 1, 2, 3...m$, z.B. Ruhe und Bewegung) wiederholt eingesetzt werden. Während für ein Design mit nur einer Facette lediglich drei Varianzkomponenten geschätzt werden, ergeben sich für ein Design mit zwei Facetten der Messung bereits sieben Varianzkomponenten (3 Haupteffekte, 3 Interaktionseffekte erster Ordnung und ein Interaktionseffekt zweiter Ordnung).

Abbildung 8: Venn-Diagramme für vollständig gekreuzte Erhebungsdesigns mit einer bzw. zwei Facetten.



Zur besseren Veranschaulichung der mit einem bestimmten Erfassungsdesign verbundenen Varianzkomponenten werden häufig sogenannte *Venn-Diagramme* herangezogen,

in denen einzelne Varianzkomponenten grafisch als Flächen dargestellt werden. Kreise repräsentieren dabei die Facetten und den Gegenstand der Messung. Überschneidung der Kreise repräsentieren Interaktionskomponenten, die nicht-überlappenden Kreissegmente stellen die Haupteffekte der Messfacetten dar. Abbildung 8 zeigt Venn-Diagramme für ein vollständig gekreuztes Erhebungsdesign mit einer (Items) und zwei Facetten der Messung (Items und Situationen).

Im Kontext einer norm-orientierten Interpretation der erhobenen Messwerte werden nur jene Varianzkomponenten zur Bestimmung der Reliabilität berücksichtigt, die sich innerhalb des Kreises befinden, der den Messgegenstand (gewöhnlich die Merkmalsunterschiede zwischen Personen) repräsentiert. Die Interaktionseffekte innerhalb dieses Kreises werden dabei in der Regel als relativer Messfehler betrachtet, der den besonderen Umständen der Erfassungssituation (z.B. der Verwendung heterogener Items oder verschiedener Beurteiler) geschuldet ist.

Die Populationsvarianzen der beschriebenen Komponenten des 2-Facetten-Erfassungsdesigns können aufgrund der mittleren Quadratsummen der Stichprobe wie folgt geschätzt werden:

$$\begin{aligned}
\sigma_j^2 &= (MS_j - MS_{ij} - MS_{js} + MS_{ijs})/km \\
\sigma_i^2 &= (MS_i - MS_{ij} - MS_{is} + MS_{ijs})/Nm \\
\sigma_s^2 &= (MS_s - MS_{js} - MS_{is} + MS_{ijs})/Nk \\
\sigma_{ij}^2 &= (MS_{ij} - MS_{ijs})/m \\
\sigma_{js}^2 &= (MS_{js} - MS_{ijs})/k \\
\sigma_{is}^2 &= (MS_{is} - MS_{ijs})/N \\
\sigma_{ijs}^2 &= MS_{ijs}.
\end{aligned} \tag{25}$$

Aufgrund des komplexeren Erfassungsdesigns stehen für eine sich anschließende D-Studie entsprechend mehr Möglichkeiten zum Vergleich verschiedener Szenarien offen.

Zufällige vs. feste Effekte von Messbedingungen

Bisher sind wir stets davon ausgegangen, dass die einzelnen Stufen der berücksichtigten Facette der Messung eine Zufallsauswahl aus der Gesamtmenge vergleichbarer Messbedingungen ist, die Facette also eine Zufalls- bzw. Randomkomponente darstellt. In einem Messdesign mit mehreren Facetten lässt sich – über die zuvor berichteten D-Studien-Szenarien hinaus – untersuchen, welche Steigerung an Präzision bzw. Reliabilität für eine Messung resultiert, wenn eine der Facetten nicht als Zufallskomponente (random effect), sondern als standardisierte Bedingung der Messung (fixed effect) begriffen wird. Im angeführten Beispiel waren wir davon ausgegangen, dass die beiden Situationen in denen Schmerzbeurteilungen anhand einer Itemliste stattfanden, eine zufällige Auswahl aus den vielen möglichen/gültigen Situationen geringerer und gesteigerter Aktiviertheit darstellen. Bei der Bestimmung des G-Koeffizienten des Gesamttestwertes fließen in diesem Szenario all jene Varianzkomponenten in den relativen Messfehler $\sigma^2(\delta)$ ein, die mit dem

eigentlichen Messgegenstand (hier: den Merkmalswerten der Personen) interagieren

$$\rho^2 = \sigma_j^2/[\sigma_j^2 + \sigma^2(\delta)] = \sigma_j^2/(\sigma_j^2 + \sigma_{ij}^2/k + \sigma_{js}^2/m + \sigma_{ijs}^2/km), \tag{26}$$

womit auch diejenigen Schmerzunterschiede, die sich aufgrund individueller Reaktionsneigungen auf Aktiviertheit und Bewegung ergeben (σ_{js}^2), als Fehlervariation interpretiert werden.

Die implizite Dichotomisierung des Aktivitätskontinuums und der Blick auf die beiden Pole absoluter Ruhe und absoluter Aktiviertheit legen jedoch nahe, diese Bedingungsvariation als fixe Facette der Messung zu begreifen. Durch diese Standardisierung der Schmerzmessung auf zwei komplementäre Erfassungssituationen wird die Generalisierbarkeit der erhobenen Schmerzwerte auf genau diese *Extremsituationen* beschränkt. Die Interaktion zwischen der Aktiviertheit und der Schmerzbelastung einer Person wird in der Folge nicht mehr als Komponente des Messfehlers, sondern als Bestandteil der wahren Merkmalsvariation begriffen, sodass sich für dieses Szenario ein anderer G-Koeffizient

$$\rho^2 = (\sigma_j^2 + \sigma_{js}^2/m)/(\sigma_j^2 + \sigma_{ij}^2/k + \sigma_{js}^2/m + \sigma_{ijs}^2/km) \tag{27}$$

mit einer im Vergleich zum zuvor beschriebenen Szenario gesteigerten Reliabilität ergibt. Wie zuvor bereits beschrieben, wird diese Reliabilitätsteigerung jedoch durch einen Verlust an Generalisierbarkeit erkauft, da sich die Aussagen nunmehr lediglich auf das Schmerzerleben in den beiden beschriebenen Erfassungssituationen beziehen können.

4.1.4 Konventionelle Bestimmung von Itemeigenschaften

Bei der bisherigen Darstellung der klassischen Test- und Generalisierungstheorie lag das Hauptaugenmerk auf der Abschätzung der Reliabilität des Gesamttestwertes einer mehrere Items umfassenden Skala. Die Einzelitems dieser Skala werden dabei in aller Regel vor dem Hintergrund publizierter Erkenntnisse aus inhaltsverwandten Instrumenten entlehnt oder neu konstruiert. Die einzelnen Indikatoren bzw. Items werden in der Logik der klassischen Testtheorie dabei als strikt parallele Messungen begriffen, oder im Kontext der Generalisierungstheorie zumindest als eine Zufallsauswahl von Indikatoren aus einem gemeinsamen Universum vergleichbarer Items. Trotz dieser angenommenen weitgehenden Äquivalenz werden bei der konkreten Entwicklung eines Testverfahrens aus einem größeren Pool möglicher Indikatoren solche Items identifiziert und selektiert, die als die besten Repräsentanten des anzuzeigenden Konstruktes erscheinen. Auch bei einer konzeptionellen Ausrichtung auf die Reliabilität des Gesamttestwertes ist es zu diesem Zweck notwendig, Aussagen zu den psychometrischen Eigenschaften einzelner Items machen zu können.

4.1.4.1 Itemschwierigkeiten

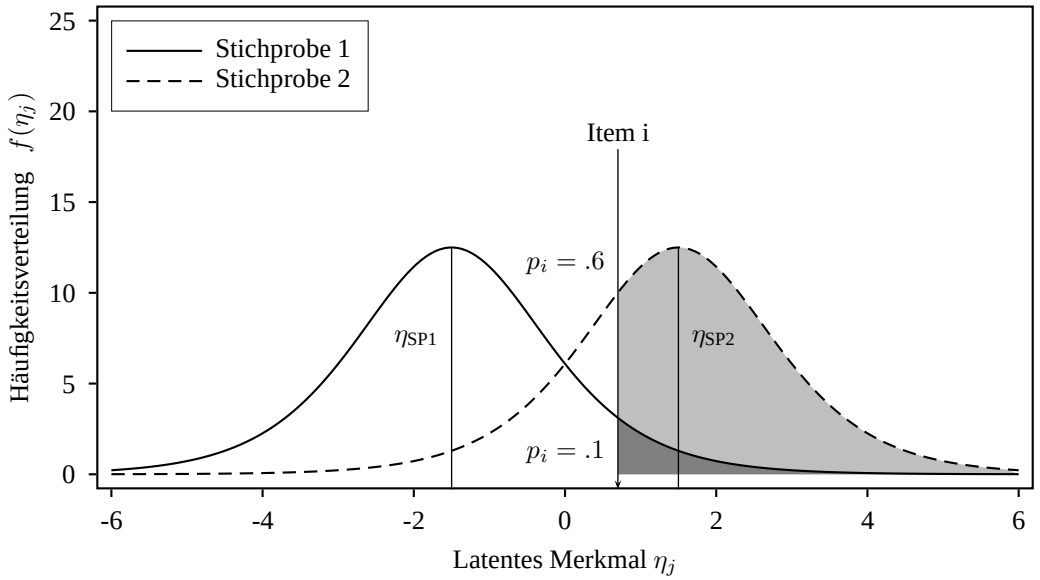
Bei der herkömmlichen Itemanalyse werden die interessierenden Itemeigenschaften auf der Grundlage des in der Stichprobe beobachteten Antwortverhaltens abgeschätzt. Im

einfachen Fall dichotomer Items, wie z.B. Testfragen, die entweder gelöst oder nicht gelöst werden können, wird der Anteil der richtigen Lösungen in der Stichprobe als Maß für die Schwierigkeit eines Items betrachtet. Wenn ein großer Teil der Probanden die Testaufgabe löst, wird diese als leichter angenommen als ein Item, das nur von wenigen Personen richtig beantwortet werden kann. Der Kennwert p_i für die Itemschwierigkeit lässt sich damit als Wahrscheinlichkeit interpretieren, mit der eine zufällig ausgewählte Person in der Stichprobe das betreffende Item zu lösen in der Lage ist. Im Falle mehrfach gestufter Antwortalternativen (z.B. bei Ratingskalen) ergibt sich die Itemschwierigkeit als Anteil der erreichten Punktwerte an den maximal möglichen Punktwerten

$$p_i = \frac{\sum_{j=1}^{N} y_{ij}}{NC_i} \quad \text{mit} \quad c = 1, 2, 3 \ldots C \quad \text{und} \quad j = 1, 2, 3 \ldots N. \tag{28}$$

Um verlässliche Rückschlüsse auf die Schwierigkeit eines Indikators in der Zielpopulation machen zu können, muss die betrachtete Stichprobe für die in Frage stehende Grundgesamtheit repräsentativ sein. Wenn beispielsweise in einer Stichprobe Personen mit hohen Merkmalsausprägungen überrepräsentiert sind, erscheint ein Item aufgrund der hohen Lösungsrate leichter, als es für die Zielpopulation eigentlich ist. Umgekehrt erscheint ein Item als unangemessen schwierig, wenn in der gewählten Stichprobe verhältnismäßig viele Personen mit niedriger Merkmalsausprägung vorhanden sind. Die Abbildung 9 zeigt diese Abhängigkeit der Itemschwierigkeit vom mittleren Merkmalsniveau für zwei z.B. verschieden stark schmerzbelastete Patientengruppen.

Abbildung 9: Stichprobenabhängigkeit der Itemschwierigkeiten.



4.1.4.2 Itemdiskrimination

Eine weitere bedeutende psychometrische Eigenschaft ist die Diskriminationsfähigkeit eines einzelnen Items oder Subtests, die häufig auch als Trennschärfe beschrieben wird. Ein Item lässt sich dann als besonders trennscharf beschreiben, wenn Personen mit ähnlichen wahren Merkmalsausprägungen aufgrund ihres Antwortverhaltens bezüglich des betreffenden Items gut voneinander unterschieden werden können. Als Kennwert für die Diskriminationskraft eines Items wird im Kontext der klassischen Testtheorie die Korrelation eines Einzelitems y_i mit dem Gesamttestwert y herangezogen. Da die Variabilität des Einzelitems in die Gesamtvariation des Tests mit einfließt, ist diese Item-Total-Korrelation systematisch höher als die Korrelation des Items mit einem aus allen verbleibenden Items gebildeten Gesamtwert y_{-i}. Dieser korrigierte Kennwert für die Diskriminationskraft eines Items wird häufig auch als Item-Rest-Korrelation bezeichnet. Strukturell sind diese beiden Trennschärfeindizes mit den zuvor beschriebenen Ansätzen zur Bestimmung der Reliabilität durch parallele Tests (siehe Gleichung 10ff.) vergleichbar, und wie diese in ihrer Interpretierbarkeit an die Voraussetzung paralleler Tests gebunden.

Im Falle von dichotomen Verhaltensindikatoren kann die Trennschärfe eines bestimmten Items $y_i \in (0, 1)$ durch dessen *punkt-biseriale Korrelation* mit dem Gesamttestwert

$$r_{pb} = \frac{(\mu_p - \mu_q)}{\sigma} \sqrt{\frac{n_p n_q}{(n_p + n_q)^2}} \tag{29}$$

abgeschätzt werden, wobei mit p diejenigen Fälle angezeigt sind, bei denen der entsprechende Verhaltensindikator beobachtet wurde ($y_i = 1$), und q die Fälle ohne das entsprechende Verhaltensmerkmal ($y_i = 0$) beschreibt. Im Allgemeinen muss davon ausgegangen werden, dass für ein- und dasselbe Item im Kontext verschiedener Subtests oder Itemzusammenstellungen jeweils verschiedene Trennschärfen geschätzt werden.

Die angeführten Beispiele machen deutlich, dass im Kontext der Itemanalyse der klassischen Testtheorie nicht zwischen Stichprobenmerkmalen und Itemcharakterisiken unterschieden werden kann. Sowohl die Bestimmung der Testgüte, als auch die Aussagen zu den psychometrischen Eigenschaften der Einzelindikatoren, und schließlich auch zu den Fähigkeiten bzw. Eigenschaften der Merkmalsträger bleiben auf die konkret realisierte Stichprobe (Stichprobenabhängigkeit) und Testzusammensetzung (Testabhängigkeit) beschränkt.

Der entscheidende Nachteil der herkömmlichen Itemanalyse ist darin zu sehen, dass die geschätzten Kennwerte die postulierten individuellen Unterschiede in der wahren Ausprägung des Merkmals nicht berücksichtigen können. Personen mit hohen wahren Merkmalswerten haben im Vergleich zu solchen Personen mit geringen wahren Merkmalsausprägungen natürlich eine sehr viel höhere Wahrscheinlichkeit, ein entsprechendes Item zu lösen; die Itemschwierigkeit sollte nach dieser Auffassung also von der jeweiligen individuellen Position einer Person auf dem latenten Merkmalskontinuum abhängen, und folglich nicht als über das gesamte Spektrum wahrer Merkmalswerte identisch angenommen werden. In ähnlicher Weise wird ein sehr schwieriges Item kaum eine Differenzierung

von Personen mit einem ähnlich geringen Merkmalsniveau erlauben, während ein Test, der sehr geringe Anforderungen stellt in Bereichen höherer Merkmalsausprägung nur wenig diskriminativ sein wird.

Für die Erstellung eines reliablen Tests ist es wichtig, dass die Variabilität der Messwerte möglichst hoch ist. Ein Test, der kaum Varianz aufweist, da alle Einzelitems von den entsprechenden Merkmalsträgern gelöst werden können, oder ein interessierendes Verhalten so gut wie nie beobachtet werden kann, muss in seiner Reliabilität beschränkt sein. Optimal ist ein Test in dieser Hinsicht dann, wenn die Schwierigkeit der Einzelitems ungefähr im Bereich der mittleren Merkmalsausprägungen bzw. Fähigkeiten der Stichprobe liegen, also ungefähr jeweils gleich viele Personen ein Item lösen bzw. nicht lösen können. Durch diese Konzentration auf das durchschnittliche Leistungsniveau der Stichprobe lassen sich um so mehr in den Randbereichen der Fähigkeitsverteilung weniger diskriminative und reliable Messungen erwarten.

Als stichprobenabhängige Schätzungen repräsentieren die Test- und Itemkennwerte der klassischen Testtheorie also lediglich die Zuverlässigkeit, Diskriminationskraft und Schwierigkeit der Items für einen durchschnittlichen Merkmalsträger.

4.2 Schmerzmessung im Kontext der Probabilistischen Testtheorie

Im Kontext der Leistungsforschung wurden, weitgehend unabhängig vom Mainstream psychologischer Skalenentwicklung und Instrumentenevaluation, alternative Messmodelle zur Abschätzung des Zusammenhanges zwischen einem nicht direkt beobachtbaren Personenmerkmal, also latenten Leistungsparameter wie z.B. der Intelligenz, und dem konkreten Verhalten bei der Beantwortung von leistungsbezogenen Testfragen entwickelt. Während der in einem Test erreichte Gesamttestwert im Rahmen der klassischen Testtheorie als ein direktes, wenngleich auch fehlerbelastetes Maß für die Ausprägung des dahinterstehenden latenten Merkmales interpretiert wird, liegt der Fokus der sogenannten probabilistischen Testtheorie auf dem in Abhängigkeit vom wahren Merkmalswert zu erwartenden Verhalten bei der Lösung bzw. Beantwortung *einzelner* Testitems, weshalb diese Konzeption auch als *Item Response Theorie* (IRT) bezeichnet wird. Die psychometrischen Eigenschaften eines Tests ergeben sich aus den unabhängig vom Gesamttestwert geschätzten psychometrischen Eigenschaften seiner Einzelitems.

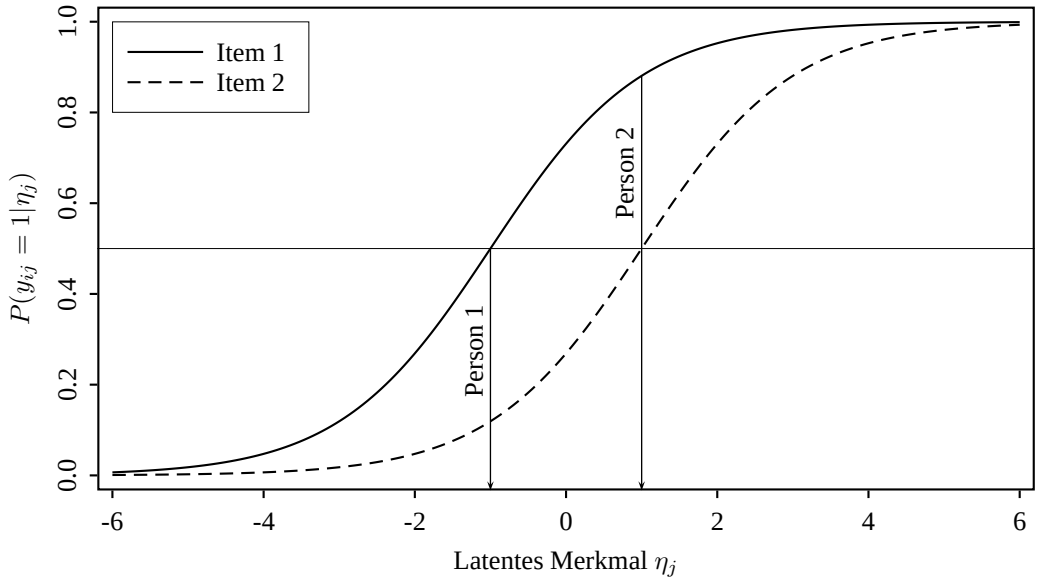
Indem die Item Response Theorie die Wahrscheinlichkeit abzuschätzen erlaubt, mit der in Abhängigkeit von einem wahren Merkmalswert η ein bestimmtes Antwortverhalten erwartet werden kann, werden viele der konzeptionellen Grenzen der KTT hinsichtlich der Bestimmung von Reliabilität und Itemcharakteristiken überwunden.

4.2.1 Grundmodell der Item-Response-Theorie

Der Zusammenhang zwischen einem zu messenden latenten Merkmal und konkreten Verhaltensäußerungen (zu denen auch die Beantwortung von Testfragen gehört) wird bei dieser Konzeption von Messung als s-förmige Funktion beschrieben. In Abbildung 10

sind diese sogenannten *itemcharakteristischen Funktionen* (ICCs) des probabilistischen Messmodelles für zwei verschieden schwer zu lösende dichotome Items dargestellt. Während Personen mit geringen wahren Merkmalswerten eine geringe Wahrscheinlichkeit besitzen, ein interessierendes Verhalten (z.B. die Lösung einer Leistungsaufgabe) im Test zu zeigen, steigt die bedingte Wahrscheinlichkeit für dieses Verhalten mit dem wahren Merkmalsniveau kontinuierlich an. Personen mit hohen wahren Merkmalswerten sollten die Items eines Tests mit großer Wahrscheinlichkeit lösen können. Die wahren Merkmalswerte der betrachteten Personen und die Schwierigkeit der Items werden im Kontext der Item-Response-Theorie auf demselben latenten Merkmalskontinuum abgebildet.

Abbildung 10: Bedingte Wahrscheinlichkeiten für zwei dichotome Items als Funktion der wahren Merkmalswerte.



Auch die herkömmliche Itemanalyse der KTT geht implizit von einer solchen systematischen Beziehung zwischen wahren Merkmalswerten und der Lösungswahrscheinlichkeit bzw. Schwierigkeit von Einzelitems aus. Da der Kennwert für die Schwierigkeit eines Items aber durch dessen Lösungswahrscheinlichkeit in der Stichprobe bestimmt ist, können lediglich Aussagen zum mittleren Schwierigkeitsniveau gemacht werden. Ähnlich verhält es sich mit dem Kennwert für die Diskriminationskraft eines Items, bei dem ebenfalls keine individuellen Merkmalsunterschiede berücksichtigt werden können.

Da im Rahmen der probabilistischen Testtheorie die Lösungswahrscheinlichkeiten eines Items in Abhängigkeit von den zugrundeliegenden Merkmalswerten geschätzt werden, wird die übliche Darstellung von IRT-Modellen auch als *Conditional Probability (CP)*

Formulierung bezeichnet, und allgemein durch die Funktion

$$P(y_{ij} = 1|\eta_j) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\eta_j - b_i)} \frac{1}{\sqrt{2\pi}} e^{.5(-z^2)} dz \tag{30}$$

beschrieben. Da es sich dabei um ein Integral über eine Normalverteilung, also eine kumulative Normalverteilungsfunktion handelt, sind die notwendigen Rechenoperationen aufwendig.

Eine technisch leichter zu handhabende Approximation des in Gleichung (30) ausgedrückten Zusammenhanges kann durch die Verwendung der logistischen Funktion

$$P(y_{ij} = 1|\eta_j) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\eta_j - b_i)}} \tag{31}$$

geleistet werden, wobei D eine Konstante mit dem Wert 1,7 ist, durch welche die logistische Funktion insbesondere in den Randbereichen geringer und hoher latenter Merkmalswerte an jene der kumulativen Normalverteilungsfunktion optimal angenähert wird.

Gleichung (31) unterscheidet die drei Modellparameter a, b und c und wird darum auch als dreiparametrisches logistisches (3PL) Modell bezeichnet. Die Definition und inhaltliche Interpretation dieser Modellparameter soll im Folgenden kurz dargestellt werden.

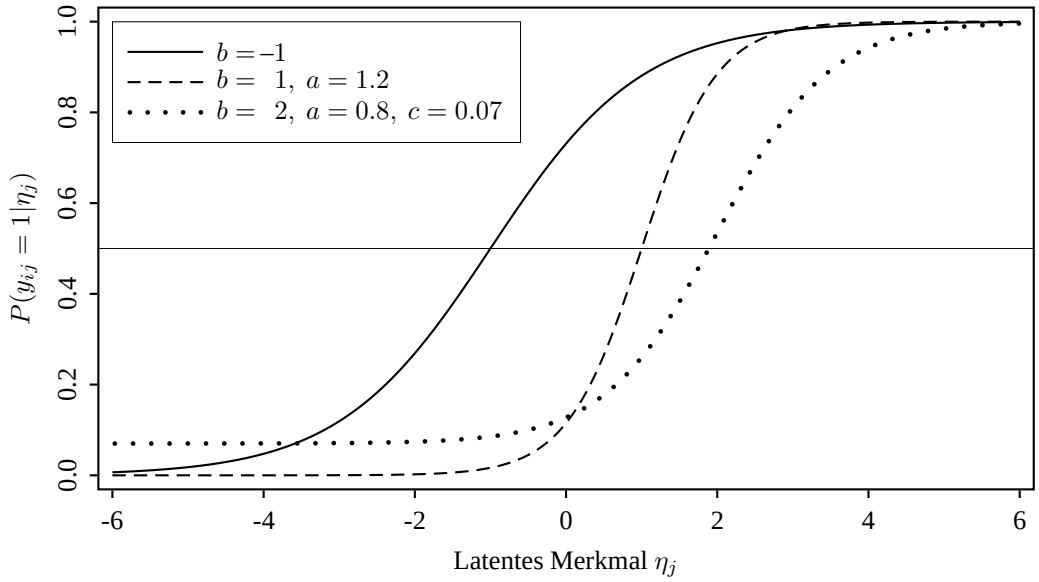
4.2.1.1 Itemschwierigkeiten

Im Rahmen der probabilistischen Testtheorie kann ein einzelnes Item je nach vorliegendem wahren Fähigkeits- bzw. Merkmalsniveau als unterschiedlich schwierig beschrieben werden. Der Kennwert für die Schwierigkeit eines Items b ist darum als derjenige latente Merkmalswert definiert, bei dem die Wahrscheinlichkeit für die Lösung des Items

$$P(y_{ij} = 1|\eta_j) = 0.5 \tag{32}$$

beträgt. Die Schwierigkeit eines Items b_i wird in IRT-Modellen also auf derselben Skala ausgedrückt wie die wahren Merkmalswerte η_j der Personen und ist nicht mit der – durch die in der Stichprobe beobachteten Lösungsrate definierte – Itemschwierigkeit p_i der klassischen Testtheorie vergleichbar. Die beiden in Abbildung 10 dargestellten Items weisen identische Funktionen auf, sind jedoch unterschiedlich schwierig. Während die Funktion des ersten Items eine Wahrscheinlichkeit von 0.5 bei einem wahren Merkmalswert von $\eta_j = b_1 = -1$ erreicht, beträgt die Schwierigkeit für das zweite Item $\eta_j = b_2 = 1$. Da die in der Abbildung dargestellte Person 1 ein geringeres Fähigkeitsniveau besitzt als die Person 2, sind auch ihre erwartbaren Lösungswahrscheinlichkeiten für beide Items geringer. Dieser Zusammenhang wird auch unmittelbar aus der Differenz zwischen Itemschwierigkeit und individuellem wahren Merkmalswert $(\eta_j - b_i)$ in den Gleichungen (31) und (30) erkennbar.

Abbildung 11: Bedingte Beobachtungswahrscheinlichkeiten dichotomer Schmerzindikatoren nach 1-, 2- und 3-parametrischen IRT-Modellen.



4.2.1.2 Itemdiskriminationen

Während die beiden Items in Abbildung 10 abgesehen von ihrer Lokalisation identische Funktionsverläufe aufwiesen, unterscheiden sich die drei in Abbildung 11 dargestellten Items zusätzlich in ihrer Trennschärfe. Die Diskriminationsfähigkeit wird in der üblichen IRT-Notation mit a bezeichnet und technisch als die Steigung der Funktion im Wendepunkt, also dem Merkmalswert der Itemschwierigkeit definiert. Wie diese ist auch der Kennwert für die Diskriminationskraft eines Items nicht mit den Trennschärfeindizes der KTT vergleichbar, da er unabhängig von der spezifischen Zusammensetzung von Einzelitems in einem Test und in Abhängigkeit von den wahren Merkmalswerten geschätzt wird. Die Lösungswahrscheinlichkeiten steigen für das Item 2 stärker an als für die beiden anderen Items; Personen mit ähnlichen wahren Merkmalswerten können aufgrund ihres Antwortverhaltens mit diesem Item darum besser voneinander unterschieden werden als es durch die Items 1 und 3 möglich ist. Der Parameter a kann als Gewicht verstanden werden, das der Differenz zwischen Itemschwierigkeit und individuellem wahren Merkmalswert bei der Schätzung der Lösungswahrscheinlichkeit zukommt, was in den Gleichungen (31) und (30) durch den Term $a(\eta_j - b_i)$ deutlich wird. Ein optimal diskriminatives Item sollte danach einen möglichst sprunghaften Anstieg der bedingten Lösungswahrscheinlichkeiten über ein möglichst kurzes Intervall wahrer Merkmalswerte aufweisen. Hoch diskriminative Items leisten eine Differenzierung damit gleichzeitig auch nur über einen vergleichsweise engen Abschnitt des angenommenen latenten Merkmalskontinuums, können also

kaum zur Unterscheidung von Personen beitragen, die wahre Merkmalswerte aufweisen, die weit unter oder über dem Merkmalswert für die Schwierigkeit des Items liegen.

4.2.1.3 Ratewahrscheinlichkeit

Insbesondere im Kontext eines Assessments von Leistungsmerkmalen, und hier insbesondere bei Multiple-Choice-Antwortformaten stellt sich die Frage, ob die richtige Lösung eines Items nicht auch unabhängig vom Fähigkeitsniveau (z.B. zufällig durch Raten) zustande gekommen sein könnte. Werden beispielsweise bei einer Multiple-Choice-Aufgabe vier Antwortalternativen vorgegeben, so beträgt die Wahrscheinlichkeit, bei der Wahl einer beliebigen Antwortalternative zufällig richtig zu liegen nominell p=0.25. Die Studienlage zur Bedeutung spezifischer Itemcharakteristiken und des Fähigkeitsniveaus der Probanden für Ratewahrscheinlichkeiten belegt jedoch, dass die empirisch gefundenen zufälligen Lösungsraten nicht unbedingt auch mit der nominellen Wahrscheinlichkeit für eine rein zufällige Lösung übereinstimmen. In der Einschätzung von Suen (1990) kann gegenwärtig noch nicht von einer konsistenten und empirisch gut belegten Theorie zur Ratewahrscheinlichkeit gesprochen werden.

Probabilistische Testmodelle können eine unabhängig vom individuellen Merkmalsniveau gegebene Lösungswahrscheinlichkeit eines Indikators durch die Konstante c berücksichtigen, um welche die geschätzte itemcharakteristische Funktion in Richtung einer höheren Beobachtungswahrscheinlichkeit verschoben ist. In Abbildung 11 nähern sich die Lösungswahrscheinlichkeiten der ersten beiden Items mit sinkendem Merkmalsniveau einer Wahrscheinlichkeit von 0 an. Die untere Asymptote der ICC für das dritte Item jedoch nähert sich einem Wert von 0.07 an. Vom wahren Merkmalsniveau unabhängige Einflüsse auf die Beobachtungswahrscheinlichkeit eines Indikators werden in den Gleichungen (31) und (30) durch den Term $c_i + (1 - c_i)$ ausgedrückt.

Für den Kontext der Schmerzmessung kann angenommen werden, dass Ratewahrscheinlichkeiten im zuvor beschriebenen Sinn eine zu vernachlässigende Rolle spielen. Da es sich nicht um eine Leistungaufgabe handelt, bei der bestimmte Antwortalternativen zwingend gewählt werden müssen, kann zumindest bei der Beobachtung von spezifischen Verhaltensindikatoren des Schmerzerlebens, die den für diese Arbeit wichtigsten Zugang zur Schmerzmessung bei nicht-kommunikativen demenzkranken Menschen darstellen, auf den Modellparameter c verzichtet werden.

Dennoch ist es allgemein sicherlich vorstellbar, dass demenzkranke Menschen in frühen Phasen ihrer Erkrankung, in denen sie prinzipiell noch auskunftsfähig erscheinen, die konkrete Befragungs- bzw. Erfassungssituation als Anforderung begreifen und darum eventuell ein nicht direkt auf ihr tatsächliches Schmerzerleben bezogenes Antwortverhalten zeigen. Zusätzlich werden Selbstauskünfte zum erlebten Schmerz sinnvoller Weise mit möglichst wenigen Antwortkategorien (ja/nein oder schwach/mäßig/stark) erhoben, was zu einer hohen nominellen Ratewahrscheinlichkeit führt. Durch die direkte Orientierung auf Schmerzen hin (Haben Sie zur Zeit Schmerzen?) kommt zusätzlich möglicherweise Tendenzen zum bestätigenden Antwortverhalten eine substanzielle Bedeutung zu.

4.2.1.4 Informationsgehalt von Einzelitems und Testbatterien

Aus der bisherigen Darstellung der itemcharakteristischen Funktion und ihrer Parameter wurde deutlich, dass jedes Einzelitem lediglich einen bestimmten Ausschnitt möglicher Ausprägungen eines latenten Merkmals anzeigen kann. Der damit über das Merkmalskontinuum hinweg variierende Informationswert I_{ij} eines Indikators ist ganz wesentlich durch dessen Diskriminationskraft bestimmt

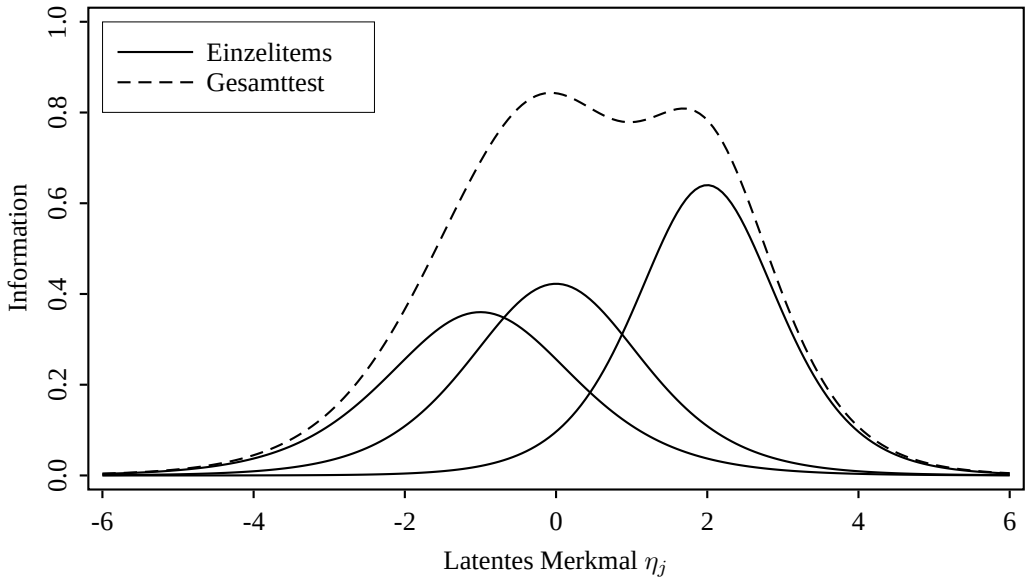
$$I_{ij}(\eta_j) = D^2 a^2 P(y_{ij} = 1|\eta_j) P(y_{ij} = 0|\eta_j) \qquad \text{mit} \quad P(y_{ij} = 0|\eta_j) = 1 - P(y_{ij} = 1|\eta_j), \tag{33}$$

und wird als Item-Information-Function (IIF) bezeichnet. Der Informationswert eines Gesamttests ergibt sich aus der Summe

$$I_j(\eta_j) = \sum_{i=1}^{k} I_{ij}(\eta_j) \tag{34}$$

der itemspezifischen Informationsfunktionen. In Abbildung 12 sind die IIFs für drei Einzelitems eines Tests, sowie die Test-Information-Function (TIF) für den Gesamttest abgebildet.

Abbildung 12: Funktionen für den Informationsgehalt von Einzelitems und Gesamttest.



Der mit einer Messung verbundene Fehler (Error of Measurement Function, EMF) verhält sich reziprok zum Informationswert eines Indikators oder Tests

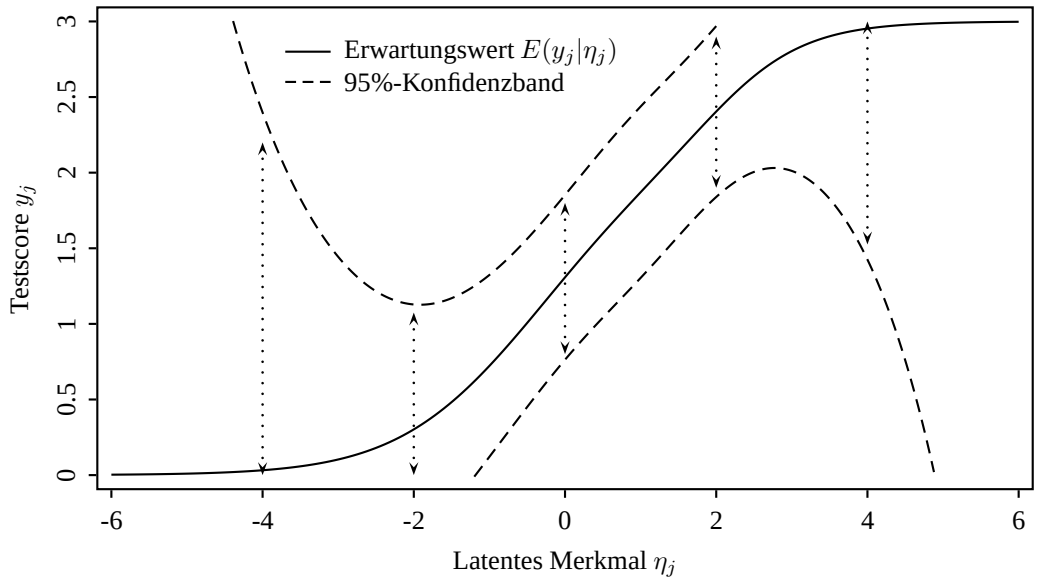
$$\epsilon_{ij}(\eta_j) = 1/I_{ij}(\eta_j) \tag{35}$$

und ist also für verschiedene wahre Merkmalsausprägungen unterschiedlich groß. In Abbildung 12 würde der Messfehler beispielsweise in den Randbereichen hoher und niedriger Merkmalsausprägungen ansteigen. Der Standardfehler $S.E.$ der Messung

$$S.E._{ij}(\eta_j) = 1/\sqrt{I_{ij}(\eta_j)} \tag{36}$$

kann zur Bestimmung eines Konfidenzbandes $y_j \pm z_\alpha S.E.$ um den gemessenen Testwert herum verwendet werden, wobei dieses Konfidenzband über die wahren Merkmalswerte hinweg nun nicht mehr, wie zuvor im Kontext der klassischen Testtheorie und Generalisierungstheorie angenommen, konstant breit ist. Das Konzept eines einzelnen Kennwertes für die Reliabilität eines Tests wird dadurch durch eine am wahren Merkmalswert ausgerichteten variablen Bestimmung der Güte der Messung ersetzt. In Abbildung 13 sind die für einen Tests mit drei dichotomen Items erwarteten Gesamtscores (Wertebereich 0-3 Punkte) in Abhängigkeit vom latenten wahren Merkmalswert dargestellt. Während die KTT einen linearen Zusammenhang zwischen beobachtetem und wahren Merkmalswert postuliert, wird dieser Zusammenhang im Kontext der IRT als nicht-linear angenommen.

Abbildung 13: Erwarteter Gesamttestwert und Konfidenzintervall für einen Test mit drei dichotomen Items.



Personen mit hohen Merkmalswerten sollten danach eher hohe Gesamttestwerte erreichen als Personen mit geringen Merkmalswerten. Der erwartete Testscore ergibt sich als Summe der Lösungswahrscheinlichkeiten der Einzelitems

$$E(y_j|\eta_j) = \sum_{i=1}^{k} P(y_{ij} = 1|\eta_j). \tag{37}$$

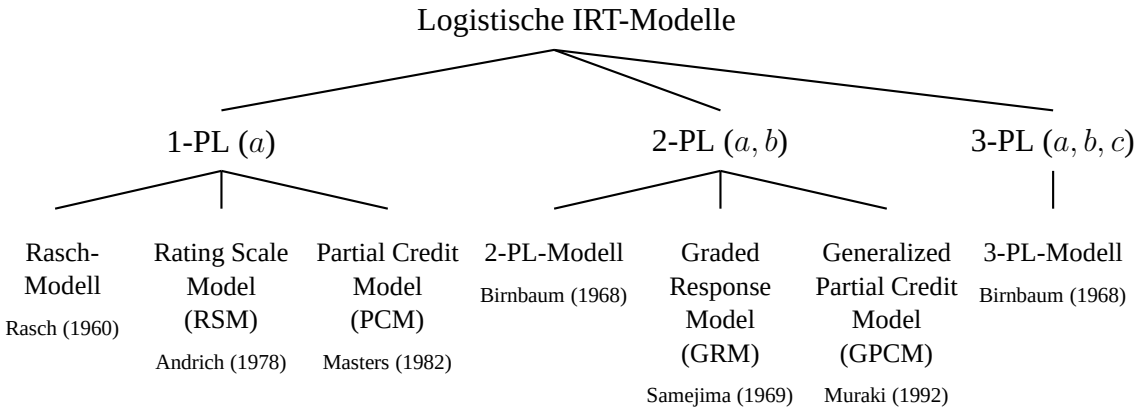
Aus Abbildung 13 wird ersichtlich, dass die Bestimmung des wahren Merkmalswertes auf der Grundlage des verwendeten Tests in diesen Extrembereichen mit einer sehr viel größeren Unsicherheit belastet ist als im Bereich mittlerer Merkmalsausprägungen. Jedes Item, und jede Zusammenstellung von Einzelitems zu einem Gesamttest haben also einen bestimmten Merkmalsbereich, der besonders gut abgebildet werden kann. Sind die Charakteristiken der Einzelitems bekannt, können so unterschiedliche Tests für beispielsweise Personen mit sehr unterschiedlichem Merkmalsniveau zusammengestellt werden.

Eng verbunden mit der Frage nach dem Informationswert und der Messfehlerbehaftetheit von Testbatterien ist auch die Diskussion um die relative Effizienz alternativer Testverfahren oder Itemzusammenstellungen. Eine optimale Zusammenstellung von Einzelitems sichert die Zuverlässigkeit bzw. Genauigkeit der Schmerzabbildung gerade in solchen Bereichen, die für die klinische oder pflegerische Versorgung von besonderer Relevanz sind, beispielsweise weil sie die Indikation für die Medikamentengabe oder ähnliche Entscheidungen darstellen. Für Verfahren, die Cut-off-Werte vorsehen, ist darum zu fordern, dass sie im Bereich um diesen Skalenwert besonders gut differenzieren.

4.2.1.5 Klassifizierung von IRT-Modellen

Bei der bisherigen Darstellung der Modellannahmen und -spezifikationen wurde auf eine Beschreibung der historischen Entwicklungslinien probabilistischer Testmodelle und der Beiträge einzelner herausragender Stellvertreter bestimmter Modelltypen bewusst verzichtet. Einen kurzen Abriss über die historische Entwicklung von IRT geben beispielsweise Embretson und Reise (2000). Ebenso wurden aufgrund ihrer für den Kontext der Verhaltensbeobachtung herausragenden Bedeutung lediglich eindimensionale probabilistische Messmodelle für dichotome Items berücksichtigt.

Abbildung 14: Klassifikation grundlegender Item-Response-Modellfamilien (nach Becker, 2004).



In den vergangenen 20 Jahren wurde dieses Grundmodell für die Modellierung ordinaler Antwortformate erweitert (Rating Scale Model RSM Andrich, 1978a,b; Partial Credit

Model PCM Masters, 1982; Graded Response Model GRM Samejima, 1969; Generalized Partial Credit Model GPCM Muraki, 1992). Die genannten Modelle für mehrfach gestufte (polytome) Antwortkategorien unterscheiden sich beispielsweise in ihren Annahmen zur Diskriminationsfähigkeit oder Äquidistanz der Schwellenwerte der einzelnen Antwortkategorien.

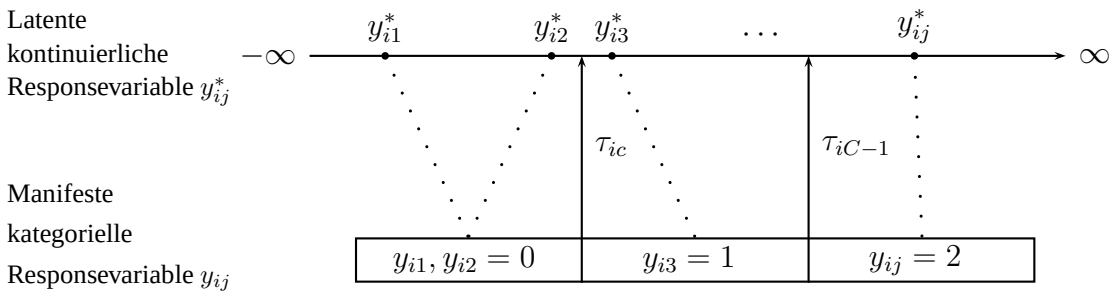
Neben dem Format der Indikatoren wurden eine Reihe weiterer Klassifikationsmerkmale, wie beispielsweise die Dimensionalität des latenten Merkmalsraumes, die verwendete logistische oder Normalverteilungs-Funktion, die Anzahl der berücksichtigten Modellparameter oder der verwendete Schätzalgorithmus vorgeschlagen. In der obigen Übersicht 14 sind die wesentlichsten eindimensionalen Grundtypen von IRT-Modellen beschrieben. Eine praxisorientierte Darstellung verschiedener Modelltypen und ihrer Anwendungsmöglichkeiten in den Sozialwissenschaften geben Rost und Langeheine (1997).

4.2.2 Latent Response Variable Formulierung von IRT-Modellen

In den letzten Jahrzehnten wurden verstärkt Bemühungen unternommen, verschiedene statistische Traditionen und Ansätze in ein übergeordnetes allgemeines Modell zu integrieren. Gegenwärtig kann die beispielsweise von Skrondal und Rabe-Hesketh (2004) vorgestellte Konzeption eines allgemeinen linearen Modells latenter Variablen (Generalized Latent Variable Modeling) als Status-Quo der Modellbildung gelten. Die zunächst auf die Analyse kontinuierlicher (metrischer) Variablen beschränkten Möglichkeiten von Strukturgleichungsmodellen (Structural Equation Modeling, SEM) wurden in den letzten Jahren um die Möglichkeit erweitert, auch Variablen mit geringerem, also kategoriellem oder nominalem Messniveau einzubeziehen.

Diese Entwicklung ist eng mit der Vorstellung verknüpft, dass beobachtete kategorielle oder dichotome Variablen y_{ij} eine gewissermaßen vergröberte Messung eines latenten kontinuierlichen Merkmales y^*_{ij} darstellen, das prinzipiell auch differenzierter erfasst und in eine metrische Skala abgebildet werden könnte (siehe Abbildung 15).

Abbildung 15: Verhältnis von latenter und beobachteter Responsevariable.



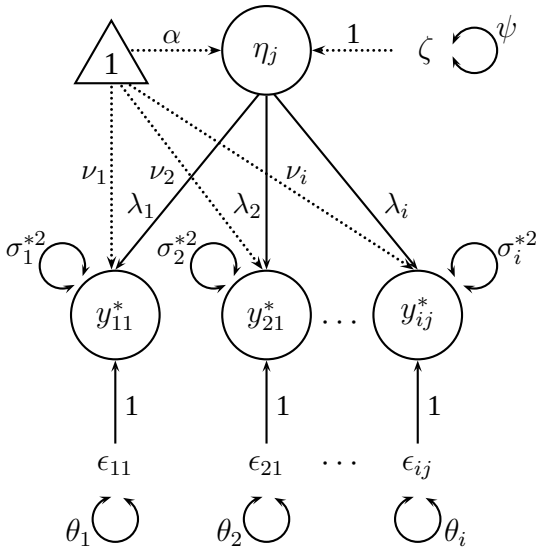
Die Beziehung zwischen der angenommenen latenten Antwortvariable (Latent Re-

sponse Variable, LRV) und der im einfachsten Fall dichotomen beobachteten Variable wird dabei über einen kritischen Schwellenwert τ_c des kontinuierlichen Merkmales definiert, ab dem eine bestimmte manifeste kategorielle Merkmalsausprägung $c = 0, 1, 2, \ldots, C-1$ beobachtet werden kann:

$$y_{ij} = c \quad \text{wenn} \quad \tau_{ic} < y^*_{ij} \leq \tau_{ic+1} \quad \text{mit} \quad \tau_{i0} = -\infty \quad \text{und} \quad \tau_{ic} = \infty. \tag{38}$$

Auch wenn die Bedeutsamkeit solcher latenten kontinuierlichen Verhaltensantworten als statistisches Werkzeug leicht zu erkennen ist, wurde aus einer stärker inhaltlichen Perspektive heraus Kritik an dieser Grundannahme geäußert. Im Kontext der Verhaltensbeobachtung ist es einleuchtend, bestimmte Verhaltensparameter, wie z.B. Atmung, Vokalisation oder Haltung als im Grunde kontinuierliche Merkmale zu begreifen, und für beschriebene Kategorien wie z.B. Hyperventilation, Schreien oder Rigidität spezifische Schwellenwerte anzunehmen. Für andere, genuin nominell angenommene Merkmale wie beispielsweise Geschlecht oder Tod, bleibt die Sinnhaftigkeit entsprechender Modellierungen jedoch fraglich.

Abbildung 16: Grundmodell IRT-Analyse mit Mplus.



Latent Response Variable (LRV):

$$y^*_{ij} = \nu_i + \lambda_i \eta_j + \epsilon_{ij}$$
$$\mu^*_i = \nu_i + \lambda_i \alpha$$
$$\sigma^{*2}_i = \lambda^2_i \psi + \theta_i$$

Standardisierung:

$$E(\eta_j) = \alpha = 0 \quad \longrightarrow \quad \mu^*_i = \nu_i$$
$$\nu_i = 0 \quad \text{und} \quad \sigma^{*2}_i = 1 \quad \longrightarrow \quad \theta_i = 1 - \lambda^2_i \psi$$

In Abbildung 16 ist ein einfaches Messmodell mit einem einzelnen latenten Merkmal und mehreren dichotomen Indikatoren dargestellt. Der zuvor beschriebenen RAM-Notation folgend sind die postulierten kontinuierlichen Antwortvariablen y^*_{ij}, die durch die beobachteten dichotomen Indikatoren y_{ij} nur vergröbert abgebildet werden können, ebenfalls als latente Variablen spezifiziert.

Die ursprünglichen bzw. klassischen Anwendungen von Strukturgleichungsmodellen beschränken sich auf auf eine Analyse der beobachteten Kovarianzen zwischen den Items.

Die Mittelwerte der Einzelitems werden dabei nicht berücksichtigt, da in die Berechnung der Kovarianzen bzw. Korrelationen nur die Abweichungsbeträge der Einzelpersonen vom Itemmittelwert einfließen, die Items damit also *zentriert* werden. Soll im LVM-Framework ein IRT-Modell spezifiziert werden, müssen darüber hinaus auch die Itemmittelwerte in die Analyse mit einbezogen werden. Eine Berücksichtigung der Mittelwertsstruktur der Daten stellt auch gegenwärtig noch einen Spezialfall dar und kann nicht zu den Standardanwendungen des LVM gezählt werden. Im hier betrachteten Fall dichotomer Schmerzindikatoren geben die Mittelwerte direkt Auskunft über die Beobachtungsraten der Verhaltensindikatoren in der Stichprobe.

4.2.2.1 Kovarianzstruktur

Algebraisch kann das in Abbildung 16 dargestellte Messmodell durch

$$y_{ij}^{*} = \nu_i + \lambda_i \eta_j + \epsilon_{ij} \tag{39}$$

beschrieben werden, wobei ν_i ein itemspezifisches Intercept der latenten Responsevariablen anzeigt. Die Varianz der unbeobachteten kontinuierlichen Antwortvariablen ergibt sich damit zu

$$\sigma_i^{*2} = \lambda_i^2 \psi + \theta_i. \tag{40}$$

4.2.2.2 Mittelwertsstruktur

Für die Schätzung der threshold-Parameter τ_{ic} der beobachteten dichotomen Variablen wird das Modell der Kovarianzen um eine Mittelwertsstruktur erweitert. Die Pfadkoeffizienten der Mittelwertsstruktur sind zur besseren Übersichtlichkeit als gepunktete Pfeile darsgestellt. Der Erwartungswert des latenten Schmerzfaktors $E(\eta)$ soll mit α bezeichnet und wie in der SEM-Notation üblich durch eine Konstante (in Abb. 16 als Dreieck dargestellt) mit dem Wert von 1 vorhergesagt werden, so dass der Mittelwert der latenten Responsevariablen durch ein itemspezifisches Intercept ν_i und gegebenenfalls verschieden stark (unterschiedliche Faktorladungen λ_i) durch die wahre Schmerzbelastung der Population α bestimmt ist

$$\mu_i^{*} = \nu_i + \lambda_i \alpha, \tag{41}$$

wobei wie üblich angenommen wird, dass die mittleren Fehlerterme für den latenten Schmerzfaktor und die beobachteten Variablen gleich 0 sind ($E(\zeta) = 0$ und $E(\epsilon) = 0$).

4.2.2.3 Standardisierung

Da es keinen natürlichen Maßstab für den latenten Faktor gibt, wird dessen Erwartungswert üblicherweise auf $E(\eta_j) = \alpha = 0$ standardisiert.

Auch die latente Responsevariable y_i^{*} besitzt keine festgelegte Skala. Eine häufig verwendete Standardisierung besteht darin, das Intercept der kontinuierlichen unbeobachteten Antwortvariable auf $\nu_i = 0$ zu setzen. Zusammen mit der Zentrierung des angezeigten

latenten Merkmalswertes (hier: Schmerz) ist der Erwartungswert μ_i^* der latenten Responsevariablen damit ebenfalls gleich Null.

Die Varianz der LRV σ_i^{*2} wird häufig auf einen Wert von 1 standardisiert bzw. gesetzt, wodurch die Fehlervarianz θ_i keinen frei zu schätzenden Modellparameter mehr darstellt, sondern sich als Residuum

$$\theta_i = 1 - \lambda_i^2 \psi \tag{42}$$

zur durch die wahren Merkmalswerte η_j erklärten Varianz der latenten Responsevariablen y_{ij}^* ergibt. Durch die Restriktionen der Mittelwerte und Varianzen der LRV wird die Skalierung der zu schätzenden Threshold-Parameter τ_{ic} und Faktorladungen λ_i festgelegt. Eine allgemeinere Formulierung dieses Skalierungsfaktors ist

$$\Delta_i^{-2} = \lambda_i^2 \psi + \theta_i \quad \text{mit} \quad \Delta_i = 1/\sqrt{\sigma^{*2}}, \tag{43}$$

wodurch beispielsweise auch die Residualvarianz auf einen bestimmten Wert (i.d.R. 1) standardisiert werden kann.

4.2.2.4 Beobachtungswahrscheinlichkeiten

Für die latenten Responsevariablen müssen Verteilungsannahmen getroffen werden, um die Wahrscheinlichkeit für einen bestimmten y_i^* ermitteln und mit dem Grenzwert τ_{ic} in Bezug setzen zu können. Gewöhnlich werden die latenten Responsevariablen als mit $N(0,1)$ multivariat standardnormalverteilt angenommen. Die Wahrscheinlichkeit für die Beobachtung eines dichotomen Merkmals y_i kann unter Berücksichtigung dieses Grenzwertes dann durch

$$P(y_i = 1) = \int_{\Delta_i(\tau_i - \mu_i^*)}^{\infty} \phi_1(y_i^*)\, dy_i^* \tag{44}$$

beschrieben werden, wobei ϕ_1 die univariate Standardnormalverteilung repräsentiert. Der Schwellenwert τ_{ic} ist dadurch als ein z-Wert definiert. Die kombinierte Wahrscheinlichkeit für ein bestimmtes beobachtetes Responsemuster ($y_i = 1, y_{i'} = 1$) ergibt sich auf der Grundlage einer bivariaten Standardnormalverteilung ϕ_2 und der jeweils univariat geschätzten Thresholds τ_{ic} und $\tau_{i'c}$ zu

$$P(y_i = 1, y_{i'} = 1) = \int_{\Delta_i(\tau_i - \mu_i^*)}^{\infty} \int_{\Delta_{i'}(\tau_{i'} - \mu_{i'}^*)}^{\infty} \phi_2(y_i^*, y_{i'}^*)\, dy_i^* y_{i'}^*. \tag{45}$$

Die Kovarianz zweier latenter Responsevariablen y_i^* und $y_{i'}^*$ ist durch die Enge ihres Zusammenhanges λ_i mit dem latenten Merkmalsfaktor η (hier Schmerz) bestimmt

$$\sigma_{ii'}^* = \lambda_i \psi \lambda_{i'}, \tag{46}$$

und durch die beschriebene Standardisierung der Verteilung der LRV mithilfe des Skalierungsfaktors Δ folgt für die Korrelation zwischen den kontinuierlichen unbeobachteten Antwortvariablen

$$\rho_{ii'}^* = \Delta_i \sigma_{ii'}^* \Delta_{i'}. \tag{47}$$

4.2.2.5 Identifikation

Durch die Standardisierung der latenten Responsevariablen können die korrespondierenden Thresholdparameter τ als z-Werte der kumulativen Standardnormalverteilung (44) bestimmt werden. Auf der Grundlage der univariat geschätzten τ-Werte können die bivariaten Korrelationen der latenten Responsevariablen durch (45) bestimmt werden.

Damit die Regressionsgewichte λ_i und die Varianz des latenten Merkmales ψ durch diese Korrelationen geschätzt werden können, muss – wie in faktoranalytischen Modellen üblich – das Regressionsgewicht eines (beliebigen) Indikators auf $\lambda_i = 1$ restringiert werden.

4.2.2.6 IRT-Itemparameter (b, a) in LRV-Parametrisierung

Im Gegensatz zur zuvor beschriebenen herkömmlichen Parametrisierung des Messmodelles über die bedingten Wahrscheinlichkeiten der beobachteten Kategorien (CP-Formulierung) anhand der Itemparameter a und b wird ein entsprechendes 2-parametrisches Item-Response-Modell im Kontext der LRV-Formulierung gewöhnlich durch

$$P(y_{ij} \geq c|\eta_j) = F[-(\tau_{ic} - \nu_i - \lambda_i\eta_j)\theta_i^{-1/2}] \tag{48}$$

als Funktion zu schätzender Varianzen, Pfadkoeffizienten und Threshold-Parameter beschrieben. Mit F wird dabei die verwendete logistische oder Normalverteilungsfunktion bezeichnet, wobei sich die Wahl an den Annahmen zur Verteilung der Messfehler ϵ_i ausrichtet. Um die Interpretation der ermittelten Parameter zu erleichtern, werden diese gewöhnlich in die herkömmlichen IRT-Metriken für Itemdiskriminationen

$$a_i = \frac{\lambda_i}{\theta_i^{1/2}} \tag{49}$$

und Itemschwierigkeiten

$$b_{ic} = \frac{\tau_{ic}}{\lambda_i} \tag{50}$$

überführt (Muthén & Asparouhov, 2002; Muthén & Muthén, 2006).

Mithilfe der Latent Response Variable (LRV)-Formulierung und entsprechenden Parameterrestriktionen können mittlerweile die meisten klassischen IRT-Modelle auch mit moderner LVM-Software spezifiziert und geschätzt werden. Ausführlichere Darstellungen zur Äquivalenz der herkömmlichen CP-IRT-Formulierung und der alternativen Spezifizierung im LVM-Framework des hier verwendeten Analyseprogramms *Mplus* findet der an einer technischen Vertiefung interessierte Leser bei Muthén und Asparouhov (2002) und Muthén und Muthén (2006).

Darüber hinaus können herkömmliche IRT-Modelle um weitere strukturelle Bestandteile, z.B. Regressoren erweitert, nach Subgruppen differenziert oder – wie später noch eingehend dargestellt wird – in eine längsschnittliche Analyse eingebunden werden. Die in Frage stehenden Zusammenhänge zwischen den interessierenden Konstrukten und den

verwendeten Indikatoren können somit auch mit Blick auf verschiedene Kontextbedingungen einer Messung deutlich präzisiert, und vergleichsweise komplexe Messprobleme bearbeitet werden.

4.2.3 Voraussetzungen von Item-Response- und Latent Variable Modellen

Auch wenn probabilistische Verfahren im Vergleich zur klassischen Testtheorie im Allgemeinen schwächere Voraussetzungen formulieren, sind die beschriebenen Vorteile (z.B. die Stichproben- und Testunabhhängigkeit der Itemparameter) dieses Messansatzes dennoch an eine Reihe von Bedingungen geknüpft.

Die nachfolgende Diskussion der Voraussetzungen stellt zuerst die spezifischen Anforderungen probabilistischer Messmodelle dar. Dabei nimmt die Annahme der *lokalen stochastischen Unabhängigkeit* einen zentralen Stellenwert ein.

Da, wie durch die bisherige Darstellung deutlich geworden ist, itemresponseanalytische Modelle in den übergeordneten Kontext des Latent Variable Modeling (LVM) eingegliedert werden können, teilt diese Modellfamilie selbstverständlich auch die allgemeinen Voraussetzungen, die für Strukturgleichungsmodelle formuliert werden. Ein wesentlicher Aspekt ist dabei das verwendete *Schätzverfahren* mit seinen Ansprüchen an Stichprobenumfang, Anzahl von Indikatoren und zu schätzender Parameter oder die Verteilung der Modellvariablen. Eine zweite Annahme betrifft die *korrekte Spezifikation des Modelles*, also dessen Gültigkeit in der Population. Konkret lässt sich auch der Aspekt der angemessenen Repräsentation des latenten Merkmalsraumes, also die Fragen nach der Dimensionalität des Modelles und der lokalen stochastischen Unabhängigkeit der Indikatoren aus diesem Blickwinkel betrachten. Ein weiterer eng an die IRT-Logik angelehnter Aspekt potenzieller Fehlspezifikation betrifft die Annahme einer monotonen (z.B. durch eine logistische oder NV-Funktion) Abbildung des Verhältnisses zwischen den latenten Merkmalswerten und der Antwortwahrscheinlichkeit der Items.

Über diese grundsätzlichen Voraussetzungen für die Gültigkeit des Modelles in der Population, die sich daraus ergebende Passung mit den empirischen Daten und die Aussagekraft der geschätzten Einzelparameter hinaus, lassen sich natürlich auch die Angemessenheit der spezifischen Annahmen des umgesetzten Modells (z.B. Gleichheit der Itemdiskriminationen in Rasch-Modellen) als Bedingung für die Anwendbarkeit von IRT-Modellen beschreiben.

4.2.3.1 Dimensionalität und lokale stochastische Unabhängigkeit

Bei der bisherigen Diskussion messtheoretischer Modelle wurde stets davon ausgegangen, dass nur ein einziges latentes Merkmal η_j abgebildet werden soll, es sich also um eine eindimensionale Messstruktur handelt. Zuvor wurde das Konzept der *Lokalen stochastischen Unabhängigkeit* der Indikatoren bereits kurz beschrieben: die Beziehungen, die zwischen den Einzelitems festgestellt werden, sollen ausschließlich auf ihre gemeinsame Bestimmtheit durch das dahinterstehende latente Merkmal zurückgeführt werden können.

Variabilität, die nicht auf unterschiedliche wahre Merkmalswerte zurückgeführt werden kann, wird als unsystematischer Messfehler begriffen. Personen mit den gleichen wahren Merkmalswerten haben dieselbe Wahrscheinlichkeit, ein gegebenes Item zu lösen. Eventuell störende Einflüsse auf die Beantwortung von Testfragen führen zu unsystematischen Verschätzungen, die über mehrere Personen hinweg unabhängig voneinander sind.

Prinzipiell bleiben IRT-Modelle selbstverständlich nicht auf Messstrukturen mit einer einzelnen Dimension beschränkt. Im Kontext der Schmerzmessung wäre es beispielsweise denkbar, dass neben der gewöhnlich abgebildeten Schmerzintensität auch noch weitere Personenmerkmale (z.B. Offenheit, Apathie oder der Grad funktioneller Beeinträchtigung) den beobachtbaren Schmerzausdruck mitbestimmen. Die Voraussetzung der lokalen stochastischen Unabhängigkeit der Indikatoren ist auch im Falle mehrdimensionaler Modelle ($d = 1, 2, 3 \ldots, l$) dann erfüllt, wenn *alle relevanten Faktoren, die das Antwortverhalten systematisch mitbestimmen* durch einen entsprechenden latenten Merkmalsvektor $\boldsymbol{\eta'} = (\eta_1, \eta_2, \ldots, \eta_l)$ berücksichtigt sind. Formal kann die lokale stochastische Unabhängigkeit der Indikatoren durch

$$P(Y_i = y_i, \ldots, Y_k = y_k|\boldsymbol{\eta}) = \prod_{i=1}^{k} P(Y_i = y_i|\boldsymbol{\eta}) \tag{51}$$

ausgedrückt werden. Die Annahme der Eindimensionalität kann dabei als Spezialfall mit $d = 1$ beschrieben werden.

Die bedingte stochastische Unabhängigkeit der Indikatoren ist eine für die Schätzung der Modellparameter aufgrund der Maximum-Likelihood-Methode (ML) notwendige Voraussetzung, da nur in diesem Fall verlässlich von der Wahrscheinlichkeit eines bestimmten gefundenen Antwortmusters auf die bedingende Merkmalsausprägung geschlossen werden kann.

Verfahren zur Überprüfung der Dimensionalität

Für den Fall, dass die Dimensionalität einer Itembatterie noch nicht hinreichend gut theoretisch und durch vorangegangene empirische Studien begründet erscheint, können explorative Verfahren wie Faktoren- und Hauptkomponentenanalyse (EFA) Hinweise auf die einer Itembatterie zugrundeliegende Merkmalsstruktur geben.

Als mögliche Kriterien für die Faktorenextraktion finden die visuelle Analyse der Scree-Plots, das Kaiser-Guttman-Kriterium (K1), sowie die inhaltliche Interpretierbarkeit der potenziellen Faktoren weitreichende Anwendung. Eine weitere Entscheidungshilfe für die Bestimmung der Dimensionalität ist die bereits 1965 durch Horn beschriebene Parallelanalyse (PA, s.a. Allen & Hubbard, 1986). Dabei werden die für die empirischen Daten berechneten Eigenwerte mit solchen Eigenwerten, die für eine analoge Matrix mit zufälligen Daten berechnet werden können, verglichen. Eine von Velicer (1976) vorgeschlagene weitere Entscheidungshilfe stellt der Minimum Average Partial (MAP)-Test dar. Bei dieser Form der Residualanalyse werden die nach der sukzessiven Auspartialisierung der in Frage stehenden Varianzkomponenten (mittels einer Hauptkomponentenanalyse) die jeweils

verbleibenden mittleren Abweichungsquadrate zwischen allen Einzelindikatoren berechnet und ihr Verlauf grafisch analysiert. Der Extraktionsschritt mit den geringsten verbleibenden mittleren Itemzusammenhängen (Residual-Korrelationen) gibt dabei die anzunehmende Anzahl zugrunde liegender Faktoren an.

Kennwerte der lokalen Abhängigkeit von Items. Aufgrund der Bedeutung der Annahme lokaler stochastischer Unabhängigkeit für alle IRT-Modelle, wurden mittlerweile verschiedene Verfahren zur Abschätzung des Ausmaßes lokaler Abhängigkeiten der Items (*engl.* local item dependence LID) vorgeschlagen und Simulationsstudien zu den Auswirkungen auf die Schätzung von Item- und Personenparametern durchgeführt (Ackerman, 1987; Chen & Thissen, 1997; Dawadi, 1998; Fennessy, 1995; Huynh, Michaels & Ferrara, 1995; Yen, 1993; Reese, 1995; Zenisky, Hambleton & Robin, 2003a,b). Für die Abschätzung der LID nimmt die von Yen (1984) vorgeschlagene Q_3*-Statistik,* ein gewissermaßen IRT-spezifisches Verfahren der Residualanalyse, eine führende Rolle ein.

Dennoch besteht in der Fachliteratur Uneinigkeit darüber, wie stark eventuelle Verletzungen der lokalen stochastischen Unabhängigkeit bzw. der Dimensionalität die Gültigkeit der geschätzten Personen- und Itemparameter beeinträchtigen. Belege für eine relative Robustheit der Parameterschätzungen wurden beispielsweise von Lord und Novik (1968), Loyd (1988), Ward (1986) und Harrison (1986) angeführt. Dem stehen beispielsweise Arbeiten von Cook, Eignor und Taft (1985), Loyd und Hoover (1980) und Slinde and Linn (1978) gegenüber.

Mehrere Autoren haben vorgeschlagen, die vergleichsweise unrealistische Annahme perfekter lokaler stochastischer Unabhängigkeit durch die schwächere Annahme einer *essenziellen Unabhängigkeit* zu ersetzten (Holland, 1981; Rosenbaum, 1984; Stout, 1987, 1990). Demnach wäre die Voraussetzung essenzieller lokaler Unabhängigkeit dann gegeben, wenn die nach Kontrolle von η verbleibenden Kovarianzen der Items hinreichend gering sind.

4.2.3.2 Modellgültigkeit

Kann der in Frage stehende bedeutsame Merkmalsraum aufgrund vorangegangener Untersuchungen oder theoretischer Modellbildungen jedoch bereits gut beschrieben werden, sprich, ist die Dimensionalität einer Itembatterie bereits bekannt, bietet sich die konfirmatorische Faktorenanalyse (Confirmatory Factor Analysis, CFA) zur Überprüfung der angenommenen Skalenstruktur an. Obwohl diese Modellfamilie zunächst im Rahmen von Strukturgleichungsmodellen für kontinuierliche Indikatoren entwickelt wurde, können mittlerweile auch Faktorenmodelle mit dichotomen Indikatoren hinsichtlich ihrer Passung mit den empirischen Daten getestet werden. Diese Anpassungstest sind selbstverständlich auch für die Überprüfung der implizierten Dimensionalität von IRT-Modellen bedeutsam (Alagumalai & Keeves, 1996; Hattie, 1985; McDonald & Mok, 1995).

Allerdings können sich die Abweichungen zwischen einem postulierten Messmodell und der Empirie aus einer nahezu endlosen Anzahl verschiedener Quellen speisen. *Fehlspezifikationen* können zum Einen die Struktur des latenten Merkmalsraumes betreffen,

beispielsweise indem zu wenige latente Dimensionen berücksichtigt oder unzutreffende Annahmen zum ((nicht)-kompensatorischen) Zusammenspiel dieser Faktoren gemacht werden. Auch die verwendete funktionale Beziehung zwischen latentem Merkmalswert und der Antwortwahrscheinlichkeit (Link-Funktion und Parameteranzahl) mag die Realität nur unzureichend abzubilden erlauben und einen Model-Data-Misfit bedingen.

Im Folgenden werden verschiedene Kriterien zur Beurteilung der allgemeinen Modellanpassung vorgestellt und bewertet. Allerdings merken Skrondal und Rabe-Hesketh (2004) an, dass für einen festgestellten Misfit im allgemeinen eine ganze Reihe möglicher Quellen verantwortlich sein können. Keiner der in der Literatur diskutierten Fit-Indizes für Latent Variable Modelle erlaubt gegenwärtig eine *spezifische* Bestimmung des Misfits zulasten eines falsch definierten latenten Merkmalsraumes, aus dem sich Verletzungen der bedingten Unabhängigkeit der Indikatoren der Messstruktur ergeben, oder identifiziert *spezifisch* übermäßig restriktive Modellannahmen beispielsweise hinsichtlich der Invarianz von Itemparametern (und damit die Unangemessenheit eines z.B. 2- vs. 1-PL-IRT-Modells; vgl. Tanaka, 1993, S. 35). Hinweise auf die Nützlichkeit des von McDonald und Marsh (1990) vorgeschlagenen Relative Fit Index (RFI) zur Bestimmung der Dimensionalität gibt die Arbeit von Alagumalai und Keeves (1996).

Erlauben die aus der Theorie abgeleiteten spezifizierten Modelle eine nur unzureichende Abbildung der empirischen Verhältnisse, bleibt ihr Erkenntnis- bzw. wissenschaftlicher Wert selbstverständlich fraglich. Als idealtypische Vorstellung der wechselseitigen Zusammenhänge zwischen den Indikatoren darf andererseits keine exakte Deckung mit realweltlichen Daten erwartet und muss ein gewisses Maß an Misfit als vertretbar angenommen werden.

Maximum-Likelihood-Schätzung der Modellparameter

Da gegenwärtig nur unzureichende Informationen zu den messtheoretischen Eigenschaften einzelner Verhaltensindikatoren zur Schmerzbeobachtung vorliegen, besteht das Ziel von IRT-Anwendungen gegenwärtig zumeist in der parallelen Schätzung von Person- und Itemeigenschaften (d.h. in der gleichzeitigen Kalibrierung der Items und Bestimmung des Schmerzniveaus der Probanden). Bei einer Stichprobengröße von n Personen, und einer Itembatterie von k Einzelindikatoren sind für ein zweiparametrisches IRT-Modell also $n \times 1$ individuelle Merkmalswerte η_j (z.B. Schmerzniveaus) und $k \times 2$ Itemparameter (a_i und b_i) zu schätzen.

Das grundlegende Verfahren dieser Parameterschätzung ist dabei die *Maximum-Likelihood-Methode (ML)*, bei der die frei schätzbaren Modellparameter so gewählt werden, dass das Muster tatsächlich beobachteter Antworten bzw. Verhaltensindikatoren maximal wahrscheinlich wird. Unter der Annahme lokaler stochastischer Unabhängigkeit ergibt sich die Wahrscheinlichkeit für das individuelle Antwortmuster $\mathbf{u}'_j = (0, 1, 0, \ldots, y_{kj})$ einer Person aufgrund der intraindividuellen Konstanz des wahren Merkmalswertes η_j als Produkt der bedingten Wahrscheinlichkeiten für die k Einzelindikatoren (s. Gleichung 51). Im hier beschriebenen Fall der Schmerzmessung durch dichotome Verhaltensindikatoren

soll die Wahrscheinlichkeit, einen Schmerzindikator zu beobachten als P_i^u, und die entsprechende Gegenwahrscheinlichkeit als Q_i^{1-u} bezeichnet werden. Für ein 2PL-Modell ergibt sich damit die kombinierte Wahrscheinlichkeit für einen individuellen Messwertvektor zu

$$\mathbf{L}[\mathbf{u}'_j = (0, 1, 0, \ldots, y_{kj})|\eta_j, \mathbf{a}, \mathbf{b}] = Q_1 P_2 Q_3 \ldots P_k^u Q_k^{1-u} = \prod_{i=1}^{k} P_i^u Q_i^{1-u}, \tag{52}$$

und wird als *Likelihood* bezeichnet. Die Gesamtwahrscheinlichkeit für alle empirisch beobachteten Antwortmuster in der Stichprobe ergibt sich aufgrund der Unabhängigkeit zwischen den Personen als Summe der individuellen Likelihoods zur *verbundenen (*engl. *joint) Likelihood*-Funktion

$$\mathbf{L}(\mathbf{U}|\eta, \mathbf{a}, \mathbf{b}) = \prod_{j=1}^{n} \prod_{i=1}^{k} P_{ij}^u Q_{ij}^{1-u}. \tag{53}$$

Zur leichteren Handhabbarkeit wird dieser Ausdruck gewöhnlich logarithmiert,

$$\ln \mathbf{L}(\mathbf{U}|\eta, \mathbf{a}, \mathbf{b}) = \sum_{j=1}^{n} \sum_{i=1}^{k} [u \ln P_{ij} + (1 - u) \ln Q_{ij}] \tag{54}$$

wodurch an die Stelle des Produkt-Operators der weniger aufwendige Summenoperator tritt. Alternativ zur Maximierung der Likelihood-Funktion kann auch die sog. *Fitting*-Funktion minimiert werden, die sich in der für die Darstellung von LVM-Modellen üblichen matrix-algebraischen Form durch

$$F_{ML} = \log |\mathbf{\Sigma}| + \text{tr}\,(\mathbf{S}\mathbf{\Sigma}^{-1}) - \log |\mathbf{S}| - k \tag{55}$$

ausdrücken lässt, wobei $\mathbf{\Sigma}$ die durch das Modell implizierte Kovarianzmatrix und $\mathbf{S}$ die empirisch beobachtete Kovarianzmatrix darstellen (Skrondal & Rabe-Hesketh, 2004).

Durch die Minimierung der Anpassungsfunktion mit Hilfe eines mehrstufigen, iterativen Algorithmus (z.B. Newton-Raphson, Fisher Scoring oder Expectation Maximization (EM), vgl. Kale, 1962; Skrondal & Rabe-Hesketh, 2004) wird eine optimale Schätzung der Personen- und Itemparameter möglich. Manche Autoren geben zu bedenken, dass bei der gleichzeitigen Schätzung von Personen- und Itemparametern (hier als *Joint Maximum Likelihood Methode* bezeichnet) die Gefahr relativ groß ist, inkonsistente Schätzungen zu erhalten (Samejima, 1973).

Bock und Lieberman (1970) und Bock und Aitkin (1981) entwickelten mit der *Marginal Maximum Likelihood Methode* ein Verfahren, bei dem die Verteilung der Personenmerkmale bei der Schätzung der Itemparameter gewissermaßen herausgerechnet (bzw. -integriert) werden. Mit den so erhaltenen asymptotisch konsistent geschätzten Itemparametern werden in einem zweiten Schritt die endgültigen Personenparameter bestimmt (*Conditional Maximum Likelihood*).

Ein Nachteil des ML-Algorithmus ist die notwendige *multivariate Normalverteiltheit* der beobachteten Daten. Im allgemeinen kann die ML-Schätzung als verhältnismäßig robust gegenüber Verletzungen der multivariaten Normalverteiltheit gelten (Hoyle & Panter, 1995). Eine Verletzung der NV wirkt sich vorallem auf die Schätzungen der Standardfehler und des χ^2-Wertes aus, während die eigentlichen Modellparameter auch mit nichtnormalverteilten Daten korrekt geschätzt werden (Muthén, 1993). Andere Autoren berichten, dass auch die Parameterschätzungen bei starken Abweichungen von einer multivaraten Normalverteilung deutlich verzerrt sein können (Nevitt & Hancock, 2001). Mittlerweile sind eine Reihe robuster und/oder verteilungsfreier Maximum Likelihood Schätzer vorgeschlagen worden (MLR: Chou, Bentler & Satorra, 1991; Muthén, 1992; Satorra, 1992; Satorra & Bentler, 1994; ADF/WLS: Browne, 1984).

Browne (1984) schlug einen alternativen Kleinst-Quadrate-Schätzer vor, bei dem die k(k+1)/2 nichtredundanten Elemente (σ bzw. $\mathbf{s}$) der implizierten Kovarianzmatrix und empirischen Kovarianzmatrix mit einer Matrix $\mathbf{W}$ gewichtet werden. Da dieser *Weighted Least Squares (WLS)-Schätzer*

$$F_{WLS} = [\sigma - \mathbf{s}]' \mathbf{W}^{-1} [\sigma - \mathbf{s}] \tag{56}$$

keine multivariate Normalverteilung der Daten voraussetzt, aber asymptotisch (also bei sehr großen Stichproben) zu den ML äquivalente Schätzungen erlaubt, wird er auch als *Asymptotic Distribution Free (ADF)-Schätzer* bezeichnet.

Gegen die Verwendung verteilungsfreier Schätzverfahren spricht jedoch, dass diese im Allgemeinen nur für vergleichsweise sehr große Stichproben (N$\geq$5000; Nevitt & Hancock, 2001; Olsson et al., 2000) konsistente Schätzungen erlauben, und die Schätzungen für umfangreiche Modelle mit ML-Schätzern vergleichbar oder sogar schlechter sind (Muthén, 1993).

Im Kontext der hier betrachteten IRT-Anwendungen für beobachtete Schmerzindikatoren ist die Voraussetzung multivariater Normalverteiltheit durch das dichotome Skalenniveau gewissermaßen modellimanent verletzt. Entsprechend wird im Kontext von Strukturgleichungsmodellen mit kategoriellen endogenen Variablen nicht die phi- oder tetrachorische Korrelationsmatrix der beobachteten dichotomen Indikatoren direkt, sondern die Kovarianzmatrix der latenten Responsevariablen y^* (LRV) verwendet (Jöreskog & Sörbom, 1993; Muthén, 1993). Geht man davon aus, dass die beobachten kategoriellen Daten vergröberte Abbildungen einer kontinuierlichen latenten Responsevariable darstellen, stellt die tetrachorische Korrelation die theoretisch angemessene Statistik für die beobachteten Daten dar. Die latenten Responsevariablen werden dabei als multivariat standardnormalverteilt angenommen. Im Gegensatz zu Modellen mit kontinuierlichen Indikatoren muss hier also eine zusätzliche Annahme zum zugrundeliegenden Verteilungsmodell getroffen und – auch wenn die wenigsten Forscher dieser Forderung nachkommen – auf ihre Gültigkeit hin überprüft werden (Muthén, 1993). Die Überprüfung der vorrangig interessierenden strukturellen Beziehungen zwischen endogenen und exogenen Modellvariablen durch Anpassungstests, wie sie für kontinuierliche Daten beschrieben ist, stellt im Kontext der Modellierung kategorieller Daten erst eine zweite Ebene der Modellbeurteilung dar,

nachdem zuvor die Angemessenheit des zugrundegelegten Verteilungsmodells bestätigt werden konnte.

Beurteilung der Modellanpassung

Die im folgenden beschriebenen allgemeinen Kriterien zur Beurteilung der Modellanpassung bzw. die zur Abschätzung des Model-Fits vorgeschlagenen Kennwerte beziehen sich im Wesentlichen auf die Analyse kontinuierlicher Variablen und ihrer Kovarianzstruktur. Einer knappen allgemeinen Übersicht gängiger Fit-Indizes soll sich darum eine Diskussion der Besonderheiten bei der Modellierung kategorieller, und hier besonders dichotomer Daten mit angenommener einfaktorieller Messstruktur anschließen.

Die Möglichkeiten eines Schätzalgorithmus, ein postuliertes IRT-Modell an die empirischen Antwortmuster anzupassen, hängen wesentlich von der Anzahl frei schätzbarer Modellparameter (wie z.B. Itemschwierigkeiten und -diskriminationen) ab. Je mehr Modellparameter berücksichtigt werden (z.B. 2PL vs. 1PL-IRT), desto besser kann die theoretisch implizierte Datenstruktur durch geeignete Parameterschätzungen an die empirisch gefundenen Variablenzusammenhänge angepasst werden (Browne & Cudeck, 1993, S. 136). Statistische Modelle mit vielen frei schätzbaren Parametern verletzen jedoch das Leitprinzip der *Parsimonität* und besitzen gegenüber restriktiveren Messstrukturen mit vergleichbarem Model-Data-Fit einen geringeren wissenschaftlichen Informationsgehalt.

Falls alle Indikatoren eines Messmodelles tatsächlich eine vergleichbare Diskriminationskraft ($a_i = a_{i'}$) besitzen und sich also lediglich hinsichtlich ihrer Schwierigkeiten unterscheiden ($b_i \neq b_{i'}$), können die durch ein einparametrisches Rasch-Modell geschätzten Parameter als sog. *erschöpfende Statistik* gelten. Alle zu schätzenden Personen- und Itemparameter können in diesem Fall auch aus den Kennwerten der klassischen Itemanalyse der KTT errechnet werden, womit der spezifische Gewinn probabilistischer Messmodelle wieder relativiert würde. Für die meisten sich in der Realität stellenden Messprobleme darf jedoch angenommen werden, dass die identifizierten Indikatoren das in Frage stehende Merkmal tatsächlich unterschiedlich gut anzeigen, und zweiparametrische Messmodelle darum eine auch substanziell begründete bessere Abbildung der empirischen Realität zu leisten in der Lage sind.

Die für die Beurteilung der Passung zwischen theoretisch impliziter und empirischer Kovarianzstruktur vorgeschlagenen Kriterien lassen sich für eine grobe Gliederung in *absolute* und *relative* bzw. *inkrementelle Fit-Indizes* einteilen (s.a. Byrne, 2001). Tanaka (1993) schlägt darüber hinaus eine Differenzierung nach Normierung, Stichproben- vs. Populationsorientierung, sowie Unabhängigkeit von Schätzverfahren und Stichprobenumfang vor.

Absolute Fit-Indizes

Absolute Kriterien der Modellanpassung betrachten das Ausmaß der Abweichungen zwischen der implizierten und empirischen Kovarianzmatrix (=*„lack of fit“*). Ein direktes Maß für diesen Misfit stellt der aus dem Wert der minimierten Anpassungsfunktion (s. Gl.

55 oder Gl. 56) berechnete χ^2- oder *Devianz-Wert*

$$\chi^2 = (n-1)F_{ML} \tag{57}$$

dar. Da diese Statistik mit $df = (k(k-1))/2 - p$ Freiheitsgraden χ^2-verteilt ist (p=Anzahl zu schätzender Parameter), kann das implizierte Modell auf seine exakte Übereinstimmung mit den empirischen Daten hin getestet werden (Bollen, 1989). Zwei Kritikpunkte haben jedoch dazu geführt, dass dieses Fitkriterium in den letzen Jahrzehnten deutlich an Bedeutung verloren hat. Zum ersten erscheint die Nullhypothese perfekten Modell-Fits in der Population *unrealistisch* und für reale Forschungsfragen kaum sinnvoll. Zweitens ist der χ^2-Wert stark von der Fallzahl abhängig (s. Gl. 57), wodurch bei hinreichend großer Stichprobe auch ein geringer Misfit zur Zurückweisung eines praktisch gut passenden Modells führt.

Jöreskog und Sörbom (1993) schlugen vor, den Devianzwert weniger als strikte Teststatistik, sondern eher als globales Gütemaß zu interpretieren. In diesem Sinne entwickelte Indizes relativieren den χ^2-Wert beispielsweise an der Anzahl der Freiheitsgrade (z.B. als CMIN/DF-Statistik im Programmpaket AMOS, Arbuckle & Wothke, 1999). Empfehlungen dahingehend, ab wann ein guter Modell-Fit erreicht ist, schwanken für diesen Index zwischen 3 und 5 (Arbuckle & Wothke, 1999). Im Gegensatz dazu spricht sich Muthén (1993) dafür aus, die χ^2-Statistik durchaus als Teststatistik für einen direkten Vergleich genesteter Daten zu verwenden.

Weitere aus dem Diskrepanzwert abgeleitete Indizes sind beispielsweise der Adjusted Goodness of Fit Index (AGFI), das Akaike Information Criterion (AIC, vgl. Tanaka, 1993), das Bayesian Information Criterion (BIC), der Standardized Root Mean Residual Index (SRMR) und der Root Mean Square Error of Approximation (RMSEA, Browne & Cudeck, 1993). Dabei wurde in den letzten Jahren vor allem der *RMSEA* als Kennwert der Modellgüte empfohlen (Hair et al., 2006; Hu & Bentler, 1998, 1999; MacCallum, Browne & Sugawara, 1996). Der *RMSEA*-Wert bezieht in die Beurteilung des hypothetisierten Modells die Schätzfehler der Populationsparameter aller Modellvariablen mit ein, und leistet somit eine Einschätzung der Passung des postulierten Modells auf die Verhältnisse *in der Population*. Der theoretische Wertebereich dieses Indikators liegt zwischen 0.0 (für perfekten Fit) und 1.0, wobei als Grenzen für eine akzeptable Repräsentation der Populationsverhältnisse durch das spezifizierte Modell ein Wert von 0.08, und für ein gutes Modell ein Wert von 0.05 empfohlen werden (Browne & Cudeck, 1993). Hu und Bentler (1998) berichten für den SRMR, dass dieser sensibel auf Fehlspezifikationen im Strukturmodell reagiert, während die anderen Fit-Indizes am sensibelsten auf Fehler im Messmodell zu reagieren scheinen.

Einige Fit-Indizes (Relative Noncentrality Index, RNI; Centrality Index, CI; McDonald & Marsh, 1990) tragen dem Umstand Rechnung, dass beim üblichen χ^2-Anpassungstest die gängige Testlogik umgekehrt ist, indem als Nullhypothese (H_0) eine optimale Passung des spezifizierten Modells mit den Populationsverhältnissen angenommen wird. Die Verteilung dieser χ^2-Werte wird als *zentral* bezeichnet. Es würde der Logik Forschung

eher entsprechen, als Nullhypothese anzunehmen, dass das postulierte Modell in der Population nicht gültig ist. Der χ^2-Wert für ein perfekt fittendes Modell würde der Anzahl der jeweiligen Freiheitsgrade entsprechen. Die an der Stichprobengröße relativierte Differenz zwischen dem Devianzwert und den Freiheitsgraden eines Modells $d = (\chi^2 - df)/(N-1)$ wird als Rescaled Non-Centrality Parameter bezeichnet und seine Testverteilung wird als *nonzentrale* χ^2-Verteilung bezeichnet. Auch der RMSEA und der CFI (s. nächster Abschnitt) beruhen auf dem Nonzentralitätsparameter.

Inkrementelle Fit-Indizes

Die vorgeschlagenen *inkrementellen* Fit-Indizes schätzen ab, inwieweit das postulierte Modell die Daten besser zu erklären in der Lage ist als ein stärker restringiertes Alternativmodell (für einen allgemeinen Überblick siehe Marsh, Balla & Hau, 1996). Als Vergleichs- oder Nullmodell wird häufig ein Modell gewählt, bei dem alle Kovarianzen zwischen den beobachteten Variablen auf einen Wert von Null restringiert sind (d.h. keine latenten Variablen und keine Zusammenhänge zwischen Einzelitems). Für dieses *independence model* kann eine große Abweichung zu den empirischen Verhältnissen angenommen werden. Damit beurteilen diese Fit-Indikatoren, von denen Hoyle und Panter (1995) im einzelnen den *Incremental Fit Index, IFI* (Bollen, 1989) und den *Comparative Fit Index, CFI* (Bentler, 1990) empfehlen, die *relative* Anpassungsgüte (=*"goodness of fit"*) des postulierten Modells. Weitere relative Fitindizes sind beispielsweise der Normed Fit Index (NFI; Bentler & Bonnet, 1980), der Relative Fit Index (RFI) und der Tucker-Lewis Index (TLI). Allerdings unterschätzen manche dieser Indikatoren bei Nichtnormalverteiltheit der Daten die tatsächliche Modellgüte systematisch oder sind stark durch die Stichprobengröße beeinflusst.

Eine Reihe dieser Indizes berücksichtigen auch die *Sparsamkeit (Parsimonität)* eines Modelles mit und bestrafen komplexe Modelle mit vielen Parametern. Wenig restriktive Modelle, bei denen ein großer Teil der Variablenzusammenhänge frei geschätzt werden kann, führen einerseits zwar zu einer guten Modellanpassung (geringer Misfit), sind andererseits aber vergleichsweise wenig informativ. Modelle, die die beobachteten Kovarianzen durch nur wenige zu schätzende Parameter beschreiben können (also sparsam sind), sind in der Regel mit größerem Misfit verbunden, aber inhaltlich weniger trivial. Sivo und Kollegen (2006, S. 285) bezweifeln jedoch die Interpretierbarkeit entsprechender Indizes, da verbindliche Optimalwerte für die Modellparsimonität fehlen.

Anpassungsgüte für Modelle mit kategoriellen Indikatoren

Die angeführten Kriterien zur Beurteilung der Modellanpassung sind nicht für Item-Response-Modelle spezifisch, sondern wurden im Wesentlichen im Kontext herkömmlicher Kovarianzstrukturanalysen (Structural Equation Modeling, SEM), und damit vorrangig für Messmodelle mit kontinuierlichen beobachteten Indikatorvariablen entwickelt und evaluiert. Eine Abschätzung der Anpassungsgüte von IRT-Modellen muss berücksichtigen, dass die beobachteten Daten in den meisten Fällen lediglich kategoriell, also keinesfalls (multivariat) normalverteilt sind, die in der Regel einfaktorielle Messstruktur durch

vergleichsweise viele Einzelindikatoren bestimmt ist und bestimmte IRT-Modellfamilien systematisch mit einer Reihe von Parameterrestriktionen verbunden sind.

Zur Beurteilung der Modellanpassung von Strukturgleichungsmodellen mit kategoriellen Indikatoren wurde von Muthén (1993) ein zweistufiges Verfahren vorgeschlagen. In einem ersten Schritt wäre danach zu überprüfen, ob die Annahme der multivariaten Normalverteiltheit der den beobachteten dichotomen Indikatoren y zugrundegelegten latenten Responsevariablen y^* angemessen ist (*First Level Fit*). Der Test des gewählten Verteilungmodells (das nicht notwendigerweise eine Normalverteilung sein muss) wird dabei gegen ein unrestringiertes Alternativmodell durchgeführt, das beispielsweise eine multinomiale Verteilung der Daten annimmt. Sollen beispielsweise zwei vierstufige Indikatoren betrachtet werden, ergibt sich ein multinomiales bivariates Verteilungsmodell mit 2×4=8 Zellen bzw. Parametern. Da sich die Zellwahrscheinlichkeiten insgesamt zu 1 addieren müssen, besitzt der χ^2-Test für dieses unrestringierte Modell 8-1=7 Freiheitsgrade. Legt man den beobachteten Variablen hingegen ein Modell multivariat normalverteilter latenter Responsevariablen zugrunde, müssen insgesamt 6 Threshold-Parameter τ, sowie die Korrelation zwischen den beiden Responsevariablen y^* geschätzt werden, womit sich für das LRV-Modell 8-7=1 Freiheitsgrade ergeben. Da beide Modelle als genestet aufgefasst werden können, erlaubt der χ^2-Differenzen-Test mit 7-1=6 Freiheitsgraden eine Überprüfung der Passung der NV-Annahme auf die Daten.

Liegen jedoch dichotome Daten vor, kann dieses Verfahren nicht angewendet werden, da das binomiale Verteilungsmodell (2×2)-1=3 Freiheitsgrade besitzt und durch das LRV-NV-Modell 2×1+1=3 Parameter restringiert werden, womit letzteres gerademal identifiziert ist. Muthén und Hofacker (1988) schlagen als eine mögliche Alternative die Testung aufgrund von Itemtripeln vor, deren Binomialverteilungsmodell $2\times2\times2$ Zellen und entsprechend 7 Freiheitsgrade besitzt. Das Modell für tetrachorische Korrelationen schätzt drei Threshold-Parameter und drei Korrelationen und erlaubt damit einen χ^2-Differenzen-Test mit einem einzelnen Freiheitsgrad.

Nach der Einschätzung von Muthén (1993) finden die verfügbaren Verfahren zur Überprüfung des für kategorielle Daten angenommenen Verteilungsmodells zu selten Anwendung. Auch zum Verhältnis zwischen der Passung des Verteilungsmodells mit den Daten und dem Test bzw. den Schätzungen der strukturellen Modellparameter (Second Level Fit) liegen noch kaum Befunde vor. Auf der Grundlage simulierter Modelle mit 5 und 15 dichotomen Indikatoren konnte Muthén nachweisen, dass die von ihm (1992) vorgeschlagenen robusten Maximum-Likelihood-Schätzer insbesondere bei größeren Modellen eine genauere Testung des Modellfits (robuster χ^2) und der Modellparameter (robuste Standardfehler S.E.) erlauben.

Auf eine weitere häufige Problematik bei der Abschätzung der Anpassungsgüte von Modellen mit vielen kategoriellen Indikatoren weist von Davier (1997) hin: selbst bei großen Stichproben werden viele Antwortmuster nur bei einzelnen Personen und ein großer Teil der potenziellen Antwortmuster überhaupt nicht beobachtet werden können. Der übliche χ^2-Anpassungstest ist im Falle geringer Zellbesetzungen und fehlender Kombinationen nicht anwendbar. Die vorgeschlagene Permutation der beobachteten Daten (*Boot-*

strapping) erweist sich in der Simulationsstudie des Autors als ein angemessenes alternatives Verfahren zur Bestimmung des χ^2-Wertes.

Trotz der Vielzahl verfügbarer diagnostischer Verfahren zur Identifikation von Verletzungen der Modellvoraussetzungen und einer ganzen Reihe vorgeschlagener Indizes für die Anpasungsgüte wurde bislang kein breiter Konsens zur Bewertung der Modellanpassung und insbesondere zum Umgang mit schlecht fittenden Modellen (i.S. eines allgemeinverbindlichen Verfahrens) gefunden. Im Extremfall wollte man darauf verzichten, die geschätzten Modellparameter überhaupt zu interpretieren, falls das postuliere Modell nicht eine zumindest hinreichend gute Abbildung der beobachteten Daten erlaubt. Orientiert man sich am hypothesentestenden konfirmatorischen Paradigma, sollten potenzielle Modellveränderungen, die sich aus einer Inspektion des Misfits ergeben, an weiteren Stichproben auf ihre Gültigkeit hin untersucht werden. Nicht zuletzt aufgrund des mit latenten Variablenmodellen im allgemeinen verbundenen hohen Aufwandes (v.a. Beobachtungsanzahl) erscheint tatsächlich jedoch eine stärker auch empiriegeleitete Modellbewertung und -entwicklung vertretbar.

Im Kontext der Neuentwicklung und Optimierung von Skalen mit IRT-Modellen ist es eine weit verbreitete Praxis, nicht-modellkonforme Items aus dem Datenpool zu entfernen und so *vermeintlich* die implizierten Antwortcharakteristiken der Indikatoren zu garantieren (insbesondere für das Rasch-Modell). Skrondal und Rabe-Hesketh (2004) bezweifeln allerdings den wissenschaftlichen Wert eines solchen datengeleiteten Vorgehens der Modelltestung.

Reckase (1997) weist darauf hin, dass die Beurteilung der Modellgüte hier ganz deutlich auch durch die doch sehr unterschiedlichen Philosophien faktoranalytischer und Item-Response-analytischer Forschungstraditionen mitbestimmt ist. Während es das Ziel der IRT ist, „to accurately reproduce the probability of correct response to an item for individuals at a particular point in the θ [hier: η] space“, betrachtet die Faktorenanalyse die Modellanpassung als „a global measure of the hypothesized model to reproduce a variance/covariance matrix for a group of individuals, rather than for single variables and selected subgroups of individuals.“ (Reckase, 1997, p. 31).

4.2.3.3 Stichprobenumfang und Itemanzahl

Eine stabile Schätzung von Item- und Personenparametern durch IRT-basierte Verfahren erfordert eine vergleichsweise hohe Anzahl beobachteter Fälle (Personen) und Items. Der Gesamtumfang empirisch verfügbarer Informationseinheiten entspricht den $N \times k$ Bestandteilen der Rohwertematrix, und je weniger Modellparameter auf dieser empirischen Grundlage geschätzt werden müssen, umso zuverlässiger sollten die Parameterschätzungen im Allgemeinen sein.

Da Maximum-Likelihood Verfahren asymptotisch erwartungstreue Schätzungen liefern, sollten für Kovarianzanalysen möglichst viele Beobachtungseinheiten zur Verfügung stehen. Eine genaue Grenze bezüglich der Anzahl von Personen oder Items ist nicht verfügbar. Boomsma (1982) schlägt einen Stichprobenumfang von mindestens N=100 vor.

Für verteilungsfreie Schätzverfahren wie ADF werden dagegen beispielsweise Stichproben von N≥5000 gefordert (Nevitt & Hancock, 2001).

Neben minimalen Stichprobengrößen für verschiedene konkrete Modelltypen oder Modellfamilien (z.B. IRT-Modelle) wurden eine Reihe von Faustregeln vorgeschlagen. Scholderer und Balderjahn (2005) beispielsweise schlagen ein Verhältnis von Personen zu Modellparametern zwischen 5:1 bis 10:1 bei einer Stichprobengröße von mindestens N=250 vor.

Entgegen der naheliegenden Annahme, dass eine größere Stichprobe bei gleichem Umfang einer Itembatterie zu besseren Schätzungen von Personen- und Itemparametern führen sollte, konnte in Simulationsstudien nachgewiesen werden, dass nur durch eine gleichzeitige Erhöhung von Stichproben- und Itemumfang eine bessere Schätzung der wahren Personen- und Itemparameter erreicht werden kann (Andersen, 1973; Haberman, 1975; Lord, 1980; Swaminathan & Gifford, 1983). Lord (1980) rät zur Berücksichtigung von mindestens 20 Items, um die Gefahr lokaler Maxima der Likelihood-Funktion zu minimieren. Für das einparametrische Raschmodell wurden mindestens 20 Items und 200 Personen gefordert (Wright & Stone, 1979), für zweiparametrische Birnbaum-Modelle mindestens 30 Items und 500 Personen, und für 3-PL-Modelle bereits 60 Items und 1000 Personen (Hulin, Lissak & Drasgow, 1982). Unabhängig von der Parameteranzahl berichtete Hambleton (1989) gute Schätzergebnisse bei Verwendung von 80 Items.

4.3 Methoden für Kernprobleme der Schmerzmessung bei Demenz

Die realweltliche Schmerzmessung im klinischen und pflegerischen Kontext ist mit einer Reihe von Wünschen und Anforderungen an die einzusetzenden Instrumente verbunden. So müssen beispielsweise für die Akzeptanz einer Schmerzmessung im Praxisfeld neben der Objektivität, Reliabilität und Validität auch weitere Gütekriterien wie Ökonomie und Nützlichkeit gefordert werden. Methodischen Verfahren zur Itemselektion kommt für die Optimierung von Schmerzinventaren darum eine besondere Bedeutung zu.

Auch eine zielgruppengerechte Adaptation von Messinstrumenten, z.B. durch eine Zusammenstellung von Indikatoren mit bekannten besonders relevanten Eigenschaften stellt eine solche Forderung dar. Schmerzmessungen werden sowohl zur Entscheidung darüber herangezogen, ob eine Schmerztherapie indiziert ist, als auch zur Erfolgskontrolle umgesetzter Therapiemaßnahmen, womit Wünsche an Cut-off-Werte und Aussagen zur Änderungssensitivität verknüpft sind. In den nachfolgenden Abschnitten werden wichtige bestehende Problemfelder der Schmerzerfassung näher beschrieben und die Potenziale einer methodischen Bearbeitung auf der Grundlage verschieden restriktiver testtheoretischer Konzeptionen einander gegenübergestellt.

4.3.1 Optimierung der Schmerzmessung

In Kapitel 3.5 wurden die Problemfelder der Schmerzerfassung durch Verhaltensbeobachtung beschrieben und Kriterien für eine optimale Schmerzerfassung bei nicht-kom-

munikativen demenzkranken Menschen skizziert. In einem weiten Verständnis tragen alle Strategien und Verfahren, die zur Überwindung der beschriebenen theoretischen und (alltags-)praktischen Probleme beitragen, zu einer Optimierung der Schmerzmessung bei. Methodische Ansätze zur Bearbeitung zentraler psychometrischer Fragen, beispielsweise nach der Vergleichbarkeit von Schmerzbeobachtungen mit konkurrierenden Instrumenten, in unterschiedlichen Beobachtungssituationen, mit Blick auf verschiedene Subgruppen demenzkranker Heimbewohner etc. werden in den nachfolgenden Kapiteln detailliert dargestellt. Sicherlich können die durch diese Ansätze und Analysen gewonnenen Informationen auch zur Optimierung der Erfassungsverfahren selbst genutzt werden, indem sie z.B. auf kontextsensitive Items hinweisen oder deren unterschiedliche Funktionsweise gar erklären können.

Zuvor jedoch sollen grundsätzlichere Strategien zur Optimierung von Instrumenten beschrieben werden, die zunächst keine verschiedenen (Sub-)Populationen, Erfassungszeitpunkte oder sonstige Kontextbedingungen unterscheiden. Die wichtigste Strategie der Instrumentenoptimierung ist die *Itemselektion*, wenngleich daneben auch die inhaltliche und sprachliche Überarbeitung von Testitems oder Veränderungen im Antwortformat zur Verbesserung der Instrumenteneigenschaften beitragen können.

4.3.1.1 Kriterien zur Optimierung der Messung

Die Kriterien, die zur Bestimmung eines optimalen Tests herangezogen werden können, sind wesentlich durch die Charakteristiken der Merkmalsträger, die Rahmenbedingungen der Testanwendung und den eigentlichen Zweck der Testung bestimmt. Auch wenn damit eine theoretisch unbegrenzte Anzahl von sich z.T. widersprechenden Qualitäten für ein konkretes Messinstrument impliziert sind, soll an dieser Stelle davon ausgegangen werden, dass eine über einen interessierenden Merkmalsausschnitt hinweg maximal informative Messung mit möglichst minimalem Aufwand eine allgemeingültige Zielvorstellung für die Erstellung und Optimierung von Assessmentinstrumenten darstellt.

4.3.1.2 Verfahren der Itemselektion

Würden die Einzelitems eines Tests im Rahmen der *konventionellen Itemanalyse* als strikt parallel begriffen, könnte sich eine Selektion bestimmter Items sinnvoll nur an deren Schwierigkeitsgrad orientieren, da strikt parallele Items gleiche Reliabilitäten aufweisen. Gewöhnlich aber werden die Einzelitems entsprechend der weniger restriktiven Annahmen der Generalisierungstheorie als Zufallsauswahl aus einem Universum vergleichbarer Indikatoren begriffen, die unterschiedlich eng mit dem latenten Merkmal verknüpft, und damit verschieden reliabel sind. Gewöhnlich werden solche Items für eine Skala zusammengestellt, die eine möglichst hohe Diskriminationskraft besitzen (vgl. Rost, 1996, Kap. 6).

Um einen möglichst großen Merkmalsbereich messen zu können, sollte die Schwierigkeit der einbezogenen Items möglichst breit streuen. Wie in Kapitel 4.1.4.1 bereits dar-

gelegt, ist die Schwierigkeit eines Items in der klassischen Testtheorie durch die Lösungsrate bzw. Beobachtungswahrscheinlichkeit in der Stichprobe definiert (vgl. Gl. 28). Sehr leichte und sehr schwierige Items weisen darum eine schiefere Merkmalsverteilung bzw. ein unausgeglicheneres Besetzungverhältnis ihrer Kategorien auf als Items von mittlerer Schwierigkeit. Damit aber müssen leichte und schwere Items weniger reliabel sein als mittelschwere Items, da letztere eine generell höhere Varianz (maximal bei p_i=.5) aufweisen und darum auch höher mit dem Gesamttestscore korrelieren (vgl. Gl. 29).

Dieses Dilemma wird im Kontext *probabilistischer Messmodelle* dadurch vermieden, dass die Itemschwierigkeit und -diskrimination in Abhängigkeit vom zugrunde liegenden wahren Merkmalswert geschätzt werden. Iteminformation, -reliabilität und -messfehler sind damit für verschiedene Abschnitte des latenten Merkmalskontinuums unterschiedlich. Lord (1977) schlug ein einfaches grundsätzliches Verfahren zur Itemauswahl bei der IRT-gestützten Skalenerstellung vor. In einem ersten Schritt sollte danach auf der Grundlage theoretischer Überlegungen zur Merkmalsverteilung in der Population und dem vorgesehenen Zweck des Testeinsatzes eine gewünschte bzw. optimale Test-Informations-Funktion (vgl. TIF Abb. 12) spezifiziert werden. Anschließend werden diejenigen Items ausgewählt, die gemeinsam die beste Annäherung an diese Zielfunktion ermöglichen.

Liegen aus vorangegangenen Untersuchungen vergleichbare Informationen zur Funktionsweise verschiedener Einzelitems in einer *Itembank* vor, können für konkrete Anwendungsziele und Populationen gewissermaßen „maßgeschneiderte“ Assessmentinstrumente zusammengestellt werden. Voraussetzung dafür ist jedoch, dass alle in Frage stehenden Itemparameter auf ein und derselben latenten Merkmalsskala abgebildet und damit untereinander vergleichbar sind (s. auch den nachfolgenden Abschnitt).

Die Kenntnis invarianter Itemparameter ist auch für das *computergestützte adaptive Testen* („Computerized Adaptive Testing“; CAT) als dem modernsten und effizientesten Verfahren der Itemselektion die Grundvoraussetzung. Adaptive Tests berücksichtigen die Antworten auf bereits bearbeitete Items für die Auswahl weiterer Testfragen mit. Personen, die beispielsweise eine schwierige Testfrage lösen können erhalten solange sowohl etwas leichtere als auch etwas schwerere Fragen vorgelegt, bis mit großer Sicherheit davon ausgegangen werden kann, dass die gezeigte Leistung nicht durch zufällige Effekte (z.B. der Itemformulierung) bedingt ist, sondern ein konstantes wahres Merkmalsniveau abbildet. Durch die flexible Zusammenstellung eines individuellen Tests in Echtzeit wird eine optimale Diskrimination auch von Personen mit ähnlichen wahren Merkmalswerten durch eine vergleichsweise geringe Anzahl von Einzelitems ermöglicht. Da von diesem Verfahren im Rahmen der vorliegenden Arbeit aus naheliegenden Gründen jedoch kein Gebrauch gemacht werden kann, sollen die Einzelheiten des CAT hier nicht weiter ausgeführt werden. Für ein anschauliches Beispiel der Potenziale eines solchen computergestützten adaptiven Verfahrens im Kontext der Schmerzerfassung sei der interessierte Leser auf die Arbeiten von Ware und Kollegen verwiesen (Ware, Bjorner & Kosinski, 2000; Ware, 2001).

4.3.2 Vergleich verschiedener Schmerzinstrumente

Da mittlerweile eine ganze Reihe ähnlicher Verhaltensinventare zur Erfassung von Schmerzen bei nicht-kommunikationsfähigen Menschen entwickelt worden sind (vgl. die in Kapitel 3.4 referierten Übersichtsarbeiten und die Tabellen 2 und 4), stellt sich zusehens die Frage, welches der vorgeschlagenen Instrumente die beste Abbildung latenter Schmerzzustände bei diesem Klientel erlaubt.

Der direkteste und informativste Zugang zum Vergleich verschiedener Schmerzassessments ist der parallele Einsatz konkurrierender Instrumente. In Anbetracht der Vulnerabilität der betrachteten Personengruppe und der knappen zeitlichen, finanziellen und auch personellen Ressourcen in diesem Praxisfeld müssen Untersuchungen, die zeitgleich mehrere Alternativverfahren zur Schmerzerfassung bei denselben Personen einsetzen, um diese anschließend direkt miteinander vergleichen zu können, sicherlich als die Ausnahme gelten. Ohne Zweifel ist ein solchermaßen kombiniertes Schmerzassessment auch theoretisch nicht unproblematisch und kann beispielsweise aufgrund der gesteigerten Testlänge oder der sich ergebenden inhaltlichen Redundanzen nur bedingt mit einem separaten Assessment verglichen werden.

Jede Form eines indirekten Vergleiches zweier Instrumente hingegen leidet unter der Konfundierung der durch die verschiedenen Instrumente bedingten Unterschiedlichkeit in den Messwerten mit Merkmalen des unterschiedlichen Kontextes der Erfassung (Zeit, Rater, Stichprobe, etc.). Mit der Entwicklung probabilistischer Messkonzepte, die eine vom Gesamttest und der konkret realisierten Stichprobe unabhängige Bestimmung von Itemkennwerten erlauben, wurden auch die Möglichkeiten eines Vergleiches von Instrumenten deutlich erweitert, die aus teilweise unterschiedlichen Items zusammengesetzt oder in unterschiedlichen Stichproben eingesetzt worden sind (*Test-Score-Equating und Item Banking*).

4.3.2.1 Direkter Vergleich mehrerer Schmerzmessungen

Selbst dann, wenn – vorrangig zum Zwecke der Skalenvalidierung – mehrere konkurrierende Schmerzassessments an der selben Stichprobe parallel eingesetzt werden, bleibt deren Kontrastierung häufig auf die Darstellung korrelativer Zusammenhänge zwischen den Gesamttestscores beschränkt. Der direkte Vergleich verschiedener Instrumente wird durch eine ganze Reihe struktureller Unterschiede im Skalenaufbau (z.B. hinsichtlich Beobachtungskontext, Anzahl von Ausdrucksbereichen bzw. Indikatoren, oder Itemscoring) erschwert.

Häufig stellen die aufgeführten Indikatoren lediglich *Beispiele* für potenziell schmerzbezogenes Verhalten in einem der berücksichtigten Ausdrucksbereiche dar. Daraus resultieren jedoch Probleme für die Bewertung eines Verhaltensinventars, da nicht klar ist, welches konkrete Verhaltensmerkmal denn nun de facto beobachtet wurde.

Da die Diskriminationsfähigkeit der Einzelindikatoren einer Skala, und damit gewissermaßen auch deren „Reliabilität“ im Kontext der herkömmlichen Itemanalyse immer

an die spezifische Testzusammensetzung gebunden ist, kann die relative Trennschärfe zweier Items aus verschiedenen Instrumenten lediglich bezogen auf den kombinierten Gesamttestscore beider Instrumente abgeschätzt werden. Eine Vergleichbarkeit mit den Trennschärfekoeffizienten in der Originalskala ist dann jedoch nicht gewährleistet.

Bessere Möglichkeiten eines direkten Vergleiches ergeben sich für die Itemschwierigkeiten gleichzeitig an derselben Stichprobe eingesetzter Instrumente, da diese nicht von den Beobachtungsraten der anderen Indikatoren abhängig sind.

Prinzipiell bleiben die Möglichkeiten eines Vergleiches verschiedener Assessments im Kontext der klassischen Testtheorie stets durch die Stichproben- und Testabhängigkeit der Itemkennwerte, die Annahme einer einzigen wahren Reliabilität der Messung über das gesamte latente Merkmalskontinuum hinweg, und die Voraussetzung paralleler Tests beeinträchtigt.

Deutlich weiterreichende Möglichkeiten eines direkten Vergleiches verschiedener Instrumente zur Schmerzmessung und ihrer Verhaltensindikatoren bieten probabilistische Messmodelle. Wenn davon ausgegangen werden kann, dass die von mehreren konkurrierenden Instrumenten vorgeschlagenen Verhaltensindikatoren allesamt ein gemeinsames latentes Schmerzkonstrukt anzeigen, wenn also für diesen kombinierten Itempool die Voraussetzung der lokalen stochastischen Unabhängigkeit erfüllt scheint, können die Einzelitems gemeinsam auf dem latenten Merkmalsfaktor skaliert und ihre geschätzten Itemkennwerte direkt miteinander verglichen werden.

Aufgrund des im IRT-Framework über das latente Merkmalskontinuum variabel angenommenen Informationsgehaltes und die Additivität der Iteminformation können damit auch konkurrierende Schmerzassessments in ihrer Gesamtheit hinsichtlich des abgebildeten Schmerzbereiches und dem erwartbaren Messfehler verglichen werden (vgl. IIFs und TIF in Abb. 12).

4.3.2.2 Indirekter Vergleich von Schmerzassessments

In einem anderen Fall sollen Instrumente miteinander hinsichtlich ihrer Itemkennwerte verglichen werden, die jedoch nicht gemeinsam erhoben wurden. Da die durch beide Messmodelle angezeigten latenten Merkmalsfaktoren keine natürliche Skalierung besitzen (Inderterminiertheit des latenten Faktors), muss diese durch die in Kapitel 4.2.2.3 beschriebenen Standardisierungen festgelegt werden. Um Item- und Personenkennwerte miteinander vergleichen zu können, müssen diese in derselben Metrik ausgedrückt sein. Eine Gleichskalierung beider geschätzter Merkmalsfaktoren kann beispielsweise durch Items erreicht werden, die in jeweils beiden konkurrierenden Instrumenten enthalten sind („anchor items“; vgl. Von Davier & Wilson, 2007). Durch dieses Prinzip können Items aus verschiedenen Instrumenten, mit denen unterschiedliche Populationen untersucht wurden, sukzessive in der gleichen Metrik ausgedrückt, und beispielsweise zu einer *Itembank* zusammengestellt werden, aus der Items mit beschriebenen Eigenschaften zum Zwecke einer zielgruppenspezifischen Instrumentenentwicklung ausgesucht werden können. Einen alternativen Weg zur Gleichskalierung bei nicht-überlappenden Itembatterien skizziert Su-

en (1990, S. 199).

Aufgrund des in der vorliegenden Studie gewählten Erfassungsdesigns kommt den Möglichkeiten des Test Equating eine nur nachgeordnete Bedeutung zu. Für eine vertiefende untechnische Darstellung der Voraussetzungen des Testequating sei der interessierte Leser an dieser Stelle auf Von Davier und Wilson (2007) verwiesen.

4.3.3 Veränderungsmessung latenter Schmerzzustände

Im Praxisalltag der stationären Versorgung demenzkranker Menschen ist die Schmerzerfassung gewöhnlich in einen kontinuierlichen Prozess des Monitorings und der bewussten Steuerung der körperlichen und seelischen Verfassung der Bewohner durch die Pflegenden eingebettet. Um eine kontinuierliche Schmerzfreiheit zu gewährleisten, dient die Schmerzerfassung sowohl der Identifizierung von (medikamentösen oder nicht-medikamentösen) Interventionsbedarfen als auch der Abschätzung des Erfolges einer entsprechenden Intervention. Wird die Diskussion der Eigenschaften von Verhaltensinventaren zur Schmerzerfassung damit auf mehrere Zeitpunkte oder verschiedene Situationen ausgeweitet, so stellt sich die Frage, welches Potenzial ein Schmerzassessment besitzt, potenzielle wahre Merkmalsveränderungen abzubilden.

In den nachfolgenden Abschnitten wird zunächst der Begriff der Veränderung bzw. Stabilität konkretisiert und verschiedene Möglichkeiten für dessen statistische Abbildung beschrieben. Während eine Fokussierung auf Veränderungen über die Zeit hinweg häufig den Erfassungskontext vernachlässigt, soll dieser aufgrund der theoretisch herausragenden Bedeutung der Beobachtungsbedingungen für die Schmerzerfassung durch Außenstehende ins Zentrum der vorliegenden Bearbeitung gestellt werden, auch wenn sich für die Art der statistischen Analyse dadurch keine substanziellen Veränderungen ergeben (vgl. Embretson, 1991). Anschließend wird der Stand der Entwicklung längsschnittlicher Anwendungen von probabilistischen Testmodellen dargestellt und Unterschiede bzw. Gemeinsamkeiten mit Strukturgleichungsmodellen für Längsschnittdaten herausgearbeitet. Schließlich wird ein Analysemodell für die vorliegende Fragestellung entwickelt, das die Stärken beider Forschungstraditionen zu kombinieren, und damit einen Großteil der formulierten Forschungsfragen in einem konsistenten Auswertungs- und Interpretationsrahmen zu beantworten erlaubt.

4.3.3.1 Merkmalsstabilität

Der Begriff der Veränderung bzw. Stabilität soll hier aus zwei verschiedenen Blickwinkeln beleuchtet werden. Zum Einen stellt sich die Frage, in welcher Hinsicht ein interessierendes Merkmal eigentlich als stabil bzw. unverändert angenommen werden soll (z.B. hinsichtlich seines Niveaus, seiner Variabilität in einer Population oder mit Blick auf seine zeitliche Entwicklung) und welche statistischen Kennwerte Auskunft über verschiedene *Typen* oder *Ausprägungsgrade von Stabilität und Veränderung* geben können. Für eine ausführliche Diskussion längsschnittlicher Analysen zur Merkmalsstabilität im Kon-

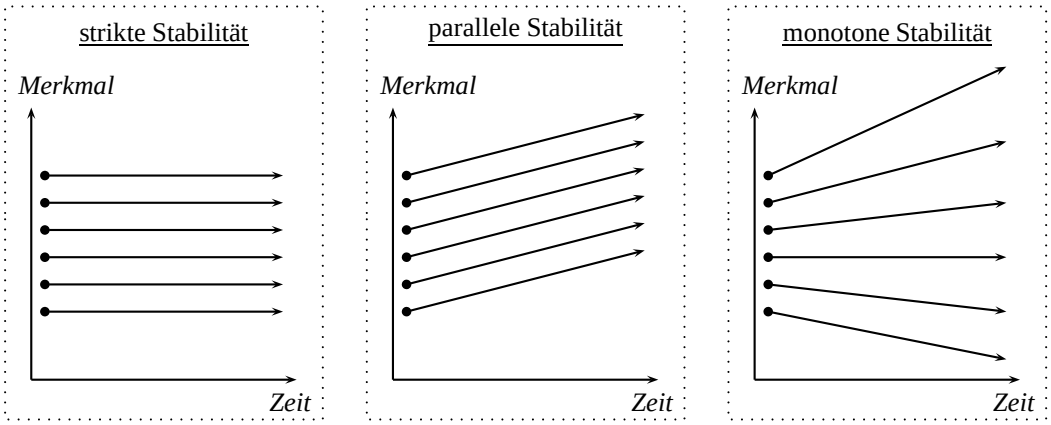
text gerontologischer Forschung sei der Leser an dieser Stelle auf die Arbeit von Schilling (2004) zur Lebenszufriedenheit verwiesen. Zum Zweiten sollen verschiedene *Prozesse* aufgezeigt werden, die für eine zwischen zwei Zeitpunkten beobachtete *Messwertänderung* theoretisch verantwortlich gemacht werden können. Dabei wird deutlich werden, dass die eigentlich interessierende Veränderung der Merkmalsträger auf dem latenten Faktor gegebenenfalls durch eine Reihe nicht beabsichtigter Urteilsprozesse überlagert werden. Eine inhaltliche Aufarbeitung solcher potenzieller Fehlereinflüsse erscheint insbesondere für die Überprüfung der Invarianz der Messung über die Zeit oder verschiedene Erfassungssituationen hinweg von allergrößter Wichtigkeit.

Typen von Merkmalsstabilität

Mit Blick auf die interessierenden wahren Merkmalsveränderungen können mit Tisak und Meredith (1990) prinzipiell eine *strikte*, eine *parallele* und eine zumindest *monotone* Merkmalsstabilität unterschieden werden.

Ein Merkmal ist in einer Population oder Stichprobe dann strikt stabil, wenn das individuelle Niveau für alle Merkmalsträger identisch bleibt. Da es in diesem Fall keine intra-individuellen Veränderungen gibt, bleiben sowohl die inter-individuelle Variabilität der Stichprobe, als auch die (relativen) Positionen der Merkmalsträger gleich (s. Abb. 17). Die Korrelation beider Messwertreihen ist perfekt ($\rho_{ss'} = 1$) und der Stichprobenmittelwert bleibt unverändert.

Abbildung 17: Typen der Merkmalsstabilität.



Von paralleler Stabilität kann dann gesprochen werden, wenn alle Merkmalsträger die gleichen Änderungen erfahren, wenn also keine inter-individuelle Variabilität in den intra-individuellen Veränderungen gibt. Durch die identischen Veränderungen bleibt die relative Position der Merkmalsträger zueinander erhalten, sodass die Korrelation der Messwertreihen wiederum 1 beträgt. Auch die Varianz der Messwerte bleibt zu beiden Zeitpunkten

gleich. Allerdings verändert sich durch die uniformen Gewinne oder Verluste der Stichprobenmittelwert entsprechend.

Die monotone Merkmalsstabilität als schwächste Stabilitätsform ist dann erfüllt, wenn die relative Position der Merkmalsträger zueinander erhalten bleibt, die intra-individuellen Veränderungen jedoch für die Merkmalsträger verschieden stark ausfallen (inter-individuelle Variabilität der intra-individuellen Veränderungen). Durch die heterogenen Veränderungsbeträge sind die Varianzen der beiden Messwertreihen nun nicht mehr unbedingt identisch, und möglicherweise (wenn auch nicht notwendigerweise) sind auch Stichprobenmittelwerte voneinander verschieden.

Als Bestimmungsstücke des Konzeptes von Merkmalsstabilität bzw. -veränderung wurden damit (a) die Merkmalsausprägungen, (b) die inter-individuelle Variabilität der Merkmalsausprägungen, (c) die relative Positionen der Merkmalsträger zueinander und (d) die inter-individuelle Variabilität der intra-individuellen Merkmalsveränderungen bestimmt. Eine Berücksichtigung der zentralen Tendenz oder Streuung der Merkmalswerte zu verschiedenen Zeitpunkten allein erlaubt somit nur eingeschränkt Aussagen über den Typus zugrundeliegender Merkmalsveränderung bzw. -stabilität.

Rudinger und Rietz (2001) machen darauf aufmerksam, dass eine Stabilitätsanalyse auf der Grundlage beobachteter Messwerte aufgrund der zu erwartenden Messfehler kaum verlässliche Schlüsse auf die (Art der) Stabilität der zugrundeliegenden latenten Merkmale zulässt. Die Korrelation beobachteter Messwertreihen unterschätzt die Stabilität der wahren Merkmalswerte

$$\rho_{y12} = \rho_{\eta 12}\sqrt{Rel_{y1} \cdot Rel_{y2}} \tag{58}$$

da die beobachteten Werte in der Regel nicht perfekt reliabel sind. Um die wahre Merkmalsstabilität abschätzen zu können bieten sich darum LVM-Verfahren an, die zwischen Mess- und Strukturmodell unterscheiden und somit eine messfehlerbereinigte Analyse der Zusammenhänge zwischen Truescores erlauben.

Da im Kontext der Schmerzmessung nicht nur die Beziehungen der Indikatoren zueinander, sondern auch das absolute Niveau erlebten Schmerzes interessieren, müssen die zur Beantwortung der hier gestellten Forschungsfragen herangezogenen Strukturgleichungsmodelle sowohl die Kovarianz- als auch die Mittelwertsstruktur der beobachteten Daten berücksichtigen. Auch wenn die Modellierung latenter Merkmale mittlerweile als Status Quo auch der sozialwissenschaftlichen Methodik anzusehen ist, bleibt eine Einbeziehung der Mittelwerte noch immer die Ausnahme (Rudinger & Rietz, 2001).

Typen von Veränderungsprozessen

Bereits vor über dreißig Jahren wiesen Golembiewski und Kollegen darauf hin, dass für beobachtete Merkmalsveränderungen theoretisch eine Reihe verschiedener Prozesse verantwortlich gemacht werden können, von denen gewöhnlich jedoch nur manche - wenn überhaupt - berücksichtigt werden (Golembiewski, Billingsley & Yeager, 1976).

Die vorrangig interessierende Veränderung ist diejenige im latenten Merkmal selbst, die von den Autoren als „alpha change“ bezeichnet wird. Wie bereits mehrfach angemerkt,

kann eine valide Abbildung solcher wahrer Merkmalsveränderungen nur dann gelingen, wenn die Messstruktur zu beiden Zeitpunkten vergleichbar ist. Golembiewski und Kollegen (1976) beschreiben zwei Veränderungen in der kognitiven Repräsentation eines Tests, die bei Messwiederholung zu einem veränderten Antwortverhalten führen und die Invarianz bzw. Äquivalenz der Messtruktur („measurement invariance/equivalence " MI/E) in Frage stellen können.

Sollen die Items eines Tests beispielsweise auf einer mehrstufigen Likert-Ratingskala (z.B. 1=„kein Schmerz", 2=„geringer", 3=„mäßiger", 4=„starker Schmerz") beantwortet werden, so ist es die Aufgabe des Befragten, die vorgegebenen Ausprägungskategorien auf dem latenten Kontinuum individuell erlebten Schmerzes zu positionieren. Wenn nun bei einer wiederholten Befragung die Antwortkategorien in anderer Weise mit dem latenten Schmerzkontinuum verknüpft werden, kann von den beobachteten Unterschieden in den Antworten nicht valide auf Änderungen des wahren Schmerzniveaus geschlossen werden. Schmerzrelevante Erfahrungen, die in der Zeit zwischen zwei Schmerzerfassungen gemacht worden sind, können durchaus eine solche *Verschiebung der Referenz- oder Ankerpunkte* bewirken. Sjöström (1995) konnte nachweisen, dass das Pflegepersonal eines Akutkrankenhauses seine Schmerzeinschätzung an die professionellen Erfahrungen mit Schmerzpatienten anpasst und die von den Patienten erlebten Schmerzen darum häufig unterschätzt (Bergh, Jakobsson & Sjöström, 2008; Sjöström, 1995). Auch eine eigene akute schwerwiegende Verletzung wie z.B. ein Knochenbruch könnte beispielsweise dazu führen, dass die bisherige mentale Repräsentation erlebten Schmerzes um bisher unbekannte *Quantitäten* ergänzt wird. Blieben die Antwortoptionen weiterhin gleichmäßig über das Merkmalskontinuum verteilt, würde das zu einer Streckung der Kategorienabstände führen. Vorstellbar wäre natürlich auch jede andere Verschiebung der Schwellenwerte („thresholds" τ_{ic}) für einzelne Antwortkategorien. Eine solche Veränderung im Antwortverhalten bezeichnen die Autoren als „beta change". Bei der Fremdbeobachtung schmerzrelevanten Verhaltens könnten sich solche Veränderungen beispielsweise in herauf- oder herabgesetzten Schwellenwerten für die Wahrnehmung und Dokumentation behavioraler Schmerzsymptome äußern. Werden also bestimmte Verhaltensweisen als *im Kontext einer gegebenen Beobachtungssituation relativ* auffällig bewertet, erscheint die transsituationale Invarianz der Messstruktur fraglich. Meade, Lautenschlager und Hecht (2005) arbeiten die besonderen Potenziale von probabilistischen Messmodellen für eine direkte Abschätzung von Messinvarianz heraus, die durch beta change bestimmt ist.

Eine weitere Beeinträchtigung der Invarianz zweier Messungen besteht dann, wenn sich die *inhaltliche Bedeutung* des zu messenden latenten Konstruktes über die Messzeitpunkte hinweg verändert. Dieser als „gamma change" bezeichnete Effekt ist dann zu erwarten, wenn Lern- oder Sensibilisierungsprozesse eine in wiederholten Messungen zunehmend differenzierte Beurteilung des erfassten Merkmalsbereiches mit sich bringen. Änderungen in der inhaltlichen Bedeutungsstruktur eines latenten Konstruktes äußern sich technisch als verändertes Ladungsmuster einzelner Indikatoren auf dem indizierten Faktor oder sogar einem veränderten latenten Merkmalsraum (z.B. durch zusätzliche Merkmalsdimensionen).

Im Kontext der Verhaltensbeobachtung erscheint es schwer abzuschätzen, ob eine (erwünschte) Sensibilisierung der Rater allein zu einem im Sinne des gamma change veränderten Interpretations- und Dokumentationsverhalten führt. Schließlich wird die Verhaltensbeobachtung ja nicht zuletzt deshalb als vergleichsweise objektives Verfahren bewertet, weil sie von einer mentalen Repräsentation des Schmerzes gewissermaßen absieht und an deren Statt beobachtbare konkrete Verhaltensäußerungen in den Mittelpunkt rückt. Ein bestimmtes Verhalten sollte demnach, falls es salient wird, auch dann dokumentiert werden, wenn es keinen zentralen Stellenwert im Schmerzkonzept des Raters einnimmt. Andererseits stellt die Schmerzbeobachtung ein in der Regel konkret motiviertes und „gerahmtes“ Verhalten dar, bei dem Vorstellungen über typischen oder potenziellen Schmerzausdruck sowohl die Aufmerksamkeit für als auch die Interpretation von bestimmten Verhaltensweisen mitbestimmen.

In der Einschätzung von Meade und Kollegen (2005) sind die Möglichkeiten von IRT-Verfahren, eine auf gamma change zurückgehende Invarianz der Messung zu identifizieren vergleichsweise begrenzt, da die korrekte Spezifizierung des latenten Merkmalsraumes (bspw. im Sinne der Eindimensionalität) eine Grundvoraussetzung aller IRT-Verfahren darstellt. Aussagen zur Dimensionalität einer Itembatterie jedoch stützen sich weitestgehend auf Verfahren der explorativen oder konfirmatorischen Faktorenanalyse, deren Aussagebereich erst seit vergleichsweise kurzer Zeit auch auf kategorielle und dichotome Daten ausgeweitet worden ist. Lediglich ein Teil der aktuellen Analyseprogramme für Latent Variable Modeling (z.B. M*plus*, Muthén & Muthén, 1998-2006) erlaubt gegenwärtig die gleichzeitige Untersuchung von invarianten Thresholds und Faktorladungen bei der längsschnittlichen oder Mehr-Gruppen-Analyse kategorieller Daten (Muthén & Asparouhov, 2002).

Statistische Verfahren zur Abschätzung der Invarianz der Messung über verschiedene Zeitpunkte oder Situationen hinweg werden in Kapitel 4.3.3.5 eingehend dargestellt.

Zeit versus Situation

Auf den besonderen Stellenwert, den die Frage nach der zeitlichen und trans-situativen Stabilität von Schmerzen im Kontext der Schmerzmessung einnimmt, wurde im Kapitel 3.4.2 bereits hingewiesen. In Abhängigkeit davon, ob Schmerzen aufgrund einer akuten Noxe erlebt werden, oder aber bereits chronifiziert sind, lassen sich unterschiedliche Veränderungen der gemessenen Schmerzzustände über bestimmte Zeiträume oder verschiedene Situationen hinweg erwarten. Auch die spezifischen Auslösebedingungen des nozizeptiven Schmerzreizes lassen Unterschiede im Schmerzerleben zu bestimmten Zeitpunkten oder in bestimmten Situationen (z.B. morgens/abends oder bei Bewegung) erwarten.

In der vorliegenden Arbeit kommt dem Faktor Zeit als solchem trotz seiner unbestrittenen Bedeutung eine nur nachgeordnete Rolle zu. Obwohl zwischen zwei unterschiedlich charakterisierten Beobachtungssituationen selbstverständlich auch immer eine gewisse Zeitspanne vergeht, liegt das Augenmerk der vorliegenden Untersuchung doch auf der

systematischen Bedingungsvariation geringer und hoher Aktiviertheitszustände der Bewohner.

Um eine sinnvolle Abschätzung der zeitbezogenen Veränderungen des Schmerzerlebens und -ausdruckes leisten zu können, sollte die Datenerfassung dagegen unter möglichst konstanten Bedingungen, in theoretisch bedeutsamen zeitlichen Abständen zu deutlich mehr als nur zwei Zeitpunkten erfolgen.

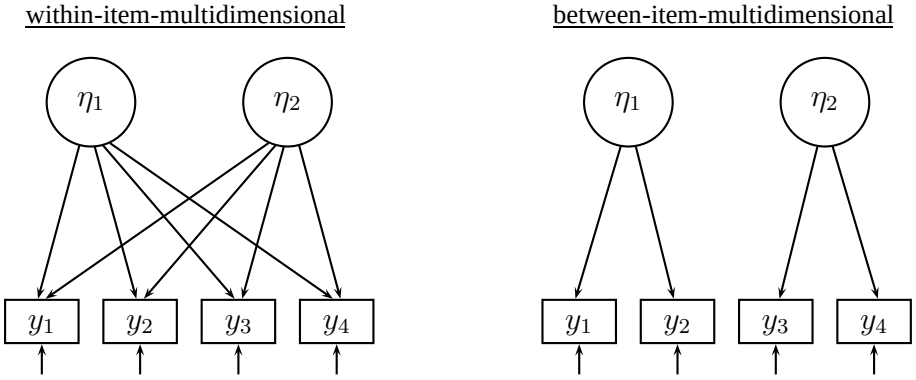
4.3.3.2 Längsschnittliche IRT-Modelle

Die Vorteile aktueller psychometrischer Verfahren (IRT und LVM) sollen auch für die Klärung der Frage nach Veränderungen in der Schmerzbelastung in unterschiedlichen Situationen genutzt werden.

Obwohl sich der sicherlich größte Teil der theoretischen und anwendungsbezogenen Veröffentlichungen zu IRT-Modellen auf einen singulären Erfassungszeitpunkt bezieht, wurde die Frage nach einer IRT-basierten Abschätzung von Entwicklungsverläufen (z.B. über die Schullaufbahn hinweg) bereits von Beginn der Methodenenentwicklung an mitbearbeitet (vgl. Fischer, 1974, Kap. 18). Aufgrund der starken Fokussierung der Verfahren an der Messstruktur wurden längsschnittlichen und situationsbezogenen Anwendungen jedoch weit weniger Beachtung geschenkt als beispielsweise dem Testverhalten in verschiedenen Bevölkerungsgruppen (s. DIF). Einen aktuellen Überblick über die wichtigsten der bisher vorgeschlagenen längsschnittlichen IRT-Modelle geben te Marvelde und Kollegen (2006).

Auch wenn die meisten IRT-Modelle lediglich ein einziges latentes Merkmal berücksichtigen, werden diese unidimensionalen Modelle bei wiederholten Messungen häufig auch als multidimensional beschrieben. Zur Wahrung der sprachlichen und konzeptionellen Konsistenz soll darum auf eine Definition multidimensionaler IRT-Modelle zurückgegriffen werden, die von Adams und Kollegen (1997) vorgeschlagen wurde (s. Abb. 18).

Abbildung 18: Schematische Darstellung verschiedener Typen von Multidimensionalität.



Sind zumindest einzelne Items eines Tests durch mehrere latente Merkmale gleichzeitig bestimmt, soll dieser Test als „within-item-multidimensional“ (WIMD) bezeichnet werden. Misst ein Test hingegen mehrere latente Merkmale durch jeweils verschiedene Items oder Subtests, die ihrerseits durch jeweils nur ein einziges latentes Merkmal bestimmt sind, kann von „between-item-multidimensional models“ (BIMD) gesprochen werden. Mehrdimensionale Modelle mit merkmalsspezifischen Items (BIMD) können als ein Spezialfall der WIMD-Modelle – bei denen alle Items in unterschiedlichem Maße durch den gesamten latenten Merkmalsraum bestimmt sind – betrachtet werden. Die Diskriminationen derjenigen Dimensionen des latenten Merkmalsraumes, die durch ein Item nicht abgebildet werden, wären dann jeweils auf einen Wert von 0 zu restringieren. Damit kann auch der wiederholte Einsatz eines Instrumentariums auf verschiedene Weise als mehrdimensionales Modell beschrieben werden.

Ein einfaches BIMD-Rasch-Modell für die wiederholte Anwendung derselben Itembatterie beschreibt Andersen (1985). Dabei werden die Itemantworten zu jedem Erfassungszeitpunkt bzw. in jeder Erhebungssituation $s = 1, 2, \ldots, m$ durch einen eigenen latenten Merkmalsfaktor η_{sj} bestimmt modelliert

$$P(Y_{sij} = 1|\eta_{sj}, b_i) = \frac{1}{1 + e^{-(\eta_{sj} - b_i)}}, \tag{59}$$

wobei wie üblich für die Einzelitems lokale stochastische Unabhängigkeit angenommen wird. Die in der gemeinsamen Kovarianzmatrix enthaltenen Abhängigkeiten zwischen den zu verschiedenen Zeitpunkten gegebenen Antworten auf ein und das selbe Item werden durch die Matrix der Kovarianzen der latenten Merkmalswerte abgebildet, die somit als direkte Kennwerte der wahren Merkmalsstabilitäten zwischen den Zeitpunkten interpretiert werden können.

Um auf der Grundlage dieser Modellierung jedoch sinnvolle Aussagen zur Stabilität oder Veränderung der wahren Merkmalswerte machen zu können, dürfen sich die Eigenschaften des Messmodelles über die Zeit hinweg natürlich nicht verändern und müssen die Itemschwierigkeiten (und -diskriminationen) zu allen Zeitpunkten gleich bleiben. Auf verschiedene Möglichkeiten, die angenommene Invarianz der Messtruktur zu testen soll im Anschluss an die Vorstellung der längsschnittlichen Modellierungen gesondert detailliert eingegangen werden.

Ein entscheidender Nachteil des durch Andersen vorgestellten Modelles liegt darin, dass für verschiedene Messzeitpunkte unterschiedliche Merkmalsfaktoren η_{sj} postuliert werden, anstatt die (interindividuellen Unterschiede in der) Veränderung wahrer Merkmalswerte direkt zu modellieren.

Im Gegensatz dazu werden im von Embretson (1991) beschriebenen „multidimensional Rasch model for learning and change“ (MRMLC) die Truescoreveränderungen für aufeinander folgende Erfassungszeitpunkte oder verschiedene Situationen ($s = 1, 2, 3 \ldots, m$) explizit in das statistische Modell eingebunden. Wenn man davon ausgeht, dass nur ein einziges latentes Merkmal wiederholt gemessen wird, entspricht die Dimensionalität des alle Messzeitpunkte berücksichtigenden Gesamtmodelles ($d = 1, 2, 3 \ldots, l$) der Anzahl

von Messzeitpunkten. Zum ersten Zeitpunkt ist das Verhalten durch eine einzige Merkmalsdimension bestimmt, während die Einflussgewichte der weiteren Dimensionen bzw. Messzeitpunkte jeweils auf einen Wert von $\lambda_{s>1} = 0$ restringiert werden. Für jeden weiteren Erfassungszeitpunkt wird nun eine zusätzliche latente Merkmalsdimension in das Modell eingebunden, die sich inhaltlich als wahre Merkmalsveränderung zwischen sukzessiven Messzeitpunkten begreifen lässt. Das Testverhalten bei der zweiten Messung beispielsweise ist damit durch einen zweidimensionalen latenten Merkmalsraum bestimmt, das Testverhalten zum dritten Erfassungszeitpunkt durch drei latente Dimensionen und so weiter. Da die faktorielle Komplexität der Messung mit der Anzahl von Messwiederholungen steigt und die Merkmalswerte mit wachsendem zeitlichen Abstand zur Ersterhebung als zunehmend unähnlicher angenommen werden können, ähnelt die implizierte Struktur der von Jöreskog (1970) beschriebenen Wiener-Simplex-Struktur. Analog zur von Jöreskog für die Schätzung von Kovarianzstrukturen mit sukzessiv ansteigender Komplexität vorgeschlagenen Gewichtungsmatrix beschreibt Embretson (1991) die nachfolgende (in diesem Fall quadratische) $m \times l$-Matrix von Faktorladungen

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ a_{21} & a_{22} & \dots & a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{s1} & a_{s2} & \dots & a_{sd} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}, \tag{60}$$

mit deren Hilfe sukzessive weitere Random-Komponenten (sog. „modifiabilities“) in das Messmodell eingebunden werden. Der Parameter a_{sd} beschreibt hier das zu einem bestimmten Zeitpunkt s für alle Items einheitliche Einflussgewicht der latenten Merkmalsdimension d, und darf nicht mit dem itemspezifisch variierenden Diskriminationsparameter a_i in zweiparametrischen Birnbaummodellen verwechselt werden.

Die Wahrscheinlichkeit des (Antwort-)Verhaltens für ein bestimmtes Item i unter der Bedingung bzw. zum Zeitpunkt s kann dann durch das einparametrische logistische Modell

$$P(Y_{sij} = 1 | \boldsymbol{\eta}_{sj}, b_i, \boldsymbol{a}_s) = \frac{1}{1 + e^{-(\sum_{d=1}^{l} a_{sd}\eta_{sdj} - b_i)}} \tag{61}$$

ausgedrückt werden, wobei $\boldsymbol{\eta}_{sj}$ den l-dimensionalen Vektor (von latentem Merkmal zu Erhebungsbeginn und sukzessiven Differenzfaktoren) und $\boldsymbol{a}_s$ den entsprechenden Vektor der Einflussgewichte dieser latenten Merkmale zum Zeitpunkt s darstellt. Das Testverhalten zum ersten Erfassungszeitpunkt ist somit eine Funktion von $(1\eta_{11j} + 0\eta_{12j} + \dots - b_i)$, dasjenige zum Folgezeitpunkt bzw. in einer zweiten Situation eine Funktion von $(1\eta_{11j} + 1\eta_{12j} + 0\eta_{13j} + \dots - b_i)$, und so fort. Der zu einem beliebigen Zeitpunkt m bestehende wahre Merkmalswert wird also in additive Differenzkomponenten zerlegt. Dieses Komposit aus initialer Merkmalsausprägung und $m-1$ sukzessiven wahren Merkmalsänderungen entspricht, sofern man die Invarianz der Messstruktur annehmen kann, der zu diesem Zeitpunkt durch ein unidimensionales Modell geschätzten latenten Merkmalsausprägung

$$\eta^{c}_{mdj} = \sum_{s=1}^{m} \eta_{sdj}. \tag{62}$$

Skalierung. Um die Skala für die latenten Merkmale festzulegen, kann die mittlere Schwierigkeit der Items auf $\sum_{i=1}^{k} b_i = 0$ und, wie bereits dargestellt, die Einflussgewichte der für einen bestimmten Zeitpunkt relevanten latenten Dimensionen auf $a_{sd} = a = 1$ restringiert werden. Durch diese einheitliche Gewichtung der latenten Merkmalskomponenten sind nicht nur die Diskriminationen über die Einzelitems hinweg konstant, sondern auch zu allen Messzeitpunkten und bezüglich aller (Differenz-)Komponenten identisch. Dadurch erhalten theoretisch alle latenten Komponenten die gleiche Bedeutung für das Antwortverhalten zu einem bestimmten Zeitpunkt. Sind einzelne wahre Merkmalskomponenten für das Testverhalten de facto unterschiedlich bedeutsam, fließen diese Unterschiede in deren Varianzschätzungen mit ein. Für einen wenig relevanten Faktor resultiert also eine geringere Varianz und – geht man von gleichbeibenden Messfehlern aus – werden geringere Reliabilitäten geschätzt.

Eine weniger restriktive Formulierung des MRMLC setzt lediglich die Faktorladungen der latenten Komponenten zu allen Zeitpunkten $a_{sd} = a_{s'd}$ gleich, sodass die einzelnen Komponenten das Testverhalten jeweils verschieden stark bestimmen können. Die Skalierung der latenten Faktoren kann dann dadurch erreicht werden, dass der Mittelwert der wahren Merkmalswerte zum ersten Messzeitpunkt α_1 auf Null und die Varianzen ψ_d aller latenten Merkmals(differenz)faktoren auf Eins restringiert werden. Diese Skalierung erlaubt die direkte Analyse mittlerer Truescoreveränderungen. Die relative Bedeutung der verschiedenen (Differenz-) Komponenten für das zu einem Zeitpunkt m beobachtete Testverhalten ergibt sich durch die unterschiedlich geschätzten Dimensionsgewichte zu

$$\psi_{\eta^c_{md}} = \sum_{d=1}^{l} a_d^2 \psi_{\eta_d}, \tag{63}$$

sodass im zuvor beschriebenen Fall einheitlicher Faktorladungen ($a_{sd} = 1$) ein hoher Einfluss einer Komponente durch eine größere Varianz geschätzt werden könnte.

Wie in Kapitel 3.4.1 dargelegt wurde, erscheint die Annahme gleicher Diskriminationsfähigkeiten für den hier betrachteten Kontext behavioraler Schmerzindikatoren nicht angemessen, womit Modellierungen auf der Grundlage von einparametrischen Rasch-Modellen unbefriedigend bleiben müssen. Für eine Erweiterung des zuvor beschriebenen MRMLC auf ein 2PL-Modell kann das von McKinley und Reckase (1982) vorgeschlagene allgemeine multidimensionale logistische Modell auf die Situation wiederholter Messungen der gleichen Itembatterie bezogen werden:

$$P(Y_{sij} = 1|\boldsymbol{\eta}_{sj}, \boldsymbol{a}_{si}, b_{si}) = \frac{1}{1 + e^{-(\sum_{d=1}^{l} a_{sdi}\eta_{sdj} - b_{si})}}. \tag{64}$$

Da in der vorliegenden Arbeit nur ein einziges latentes (Schmerz-)Merkmal betrachtet werden soll, entsprechen sich die Anzahl zu berücksichtigender Komponenten $d \in (1, 2)$ und Beobachtungsbedingungen $s \in (1, 2)$ jeweils, so dass sich nun für jedes Einzelitem eine quadratische Matrix der für dieses Item spezifischen Diskriminationen

$$\mathbf{A}_i = \begin{bmatrix} a_{11} & a_{1d} \\ a_{s1} & a_{sd} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{11} & a_{22} \end{bmatrix} \tag{65}$$

ergibt. Um auch im zweiparametrischen Fall ein zum MRMLC analoges Modell mit einem einzelnen Merkmalsfaktor zum ersten Erfassungszeitpunkt und einem Komposit aus diesem Initialfaktor und einer latenten Differenzkomponente zu spezifizieren, muss die Matrix $\boldsymbol{A}_i$ der Itemdiskriminationen der implizierten Simplex-Struktur entsprechend restringiert werden. Eine erste Restriktion betrifft die Konstanz der individuell geschätzten Itemdiskriminationen für die Einzelkomponenten in beiden Situationen $a_{sdi} = a_{s'di}$. Eine zweite Restriktion setzt den Einfluss der Differenzkomponente für das Testverhalten zum ersten Messzeitpunkt auf $a_{12i} = 0$ (s. Gl. 65).

Eine weitere bereits im MRMLC getroffene Modellannahme betrifft die zeitliche bzw. situative Konstanz der itemspezifischen Schwierigkeitsparameter $b_{si} = b_{s'i}$. Im Kontext der Verhaltensbeobachtung entspricht diese Setzung einer konstanten Ausprägung eines Verhaltens, ab dem ein entsprechender Indikator salient bzw. auffällig und entsprechend als beobachtet dokumentiert wird.

Die Wahrscheinlichkeit für die Beobachtung eines schmerzrelevanten Verhaltensindikators ergibt sich damit in der ersten Erfassungssituation (Ruhe) zu

$$P(Y_{1ij} = 1|\boldsymbol{\eta}_{1j}, \boldsymbol{a}_{1i}, b_i) = \frac{1}{1 + e^{-(a_{11i}\eta_{11j} - b_i)}}, \tag{66}$$

und wird in der zweiten Beobachtungsbedingung (Aktivität) durch das Komposit aus Ausgangswert und wahrer Merkmalsänderung

$$P(Y_{2ij} = 1|\boldsymbol{\eta}_{2j}, \boldsymbol{a}_{2i}, b_i) = \frac{1}{1 + e^{-(a_{11i}\eta_{11j} + a_{22i}\eta_{22j} - b_i)}} \tag{67}$$

bestimmt.

Da zum zweiten Erfassungszeitpunkt die beteiligten Merkmalskomponenten additiv verknüpft sind, kann die Beobachtungswahrscheinlichkeit eines Indikators sowohl von einem hohen Ausgangsniveau (Ruheschmerz), als auch durch eine hohe Vulnerabilität für aktivitätsbezogenen Schmerz bestimmt sein. Damit lässt sich das beschriebene Modell als *kompensatorisch* charakterisieren (vgl. Reckase, 1997; te Marvelde et al., 2006). Eine Alternative dazu stellen *nicht-kompensatorische* Modelle dar, die von der Annahme eines multiplikativen Zusammenspiels von Ausgangsschmerz und Schmerzänderung durch Aktivierung ausgehen. Allerdings muss hier berücksichtigt werden, dass durch die Schätzung einer Differenzkomponente für einen Teil der Bewohner negative Veränderungsbeträge i.S. eines reduzierten Schmerzniveaus zu erwarten sind. Da in einem solchen Falle ($a_{11i}\eta_{11j} \cdot a_{22i}(-\eta_{22j})$) die Wirkrichtung des geschätzten Ruheschmerzes in sein Gegenteil verkehrt würde, erscheinen nicht-kompensatorische multidimensionale Konzeptionen als für das vorliegende Modellierungsproblem weniger angemessen.

4.3.3.3 Längsschnittliche Latent Variable Modelle

Während die im Kontext der IRT vorgeschlagenen Modelle für Messwertreihen sich auf wenige und im wesentlichen einparametrische Modelltypen beschränken (vgl. Ponocny, 2002; te Marvelde et al., 2006) wurden im Kontext des LVM eine mittlerweile kaum

mehr überschaubare Anzahl verschiedener Modelle zur Längsschnittsanalyse vorgeschlagen. Eine gängige Einteilung unterscheidet dabei zwischen autoregressiven und faktoranalytischen Modellierungen von Zeitreihenwerten (Rudinger & Rietz, 2001). McArdle und Aber (1990) unterscheiden innerhalb der letztgenannten Modellfamilie nochmals „difference components“, „growth curve“ und „factor analysis change models“.

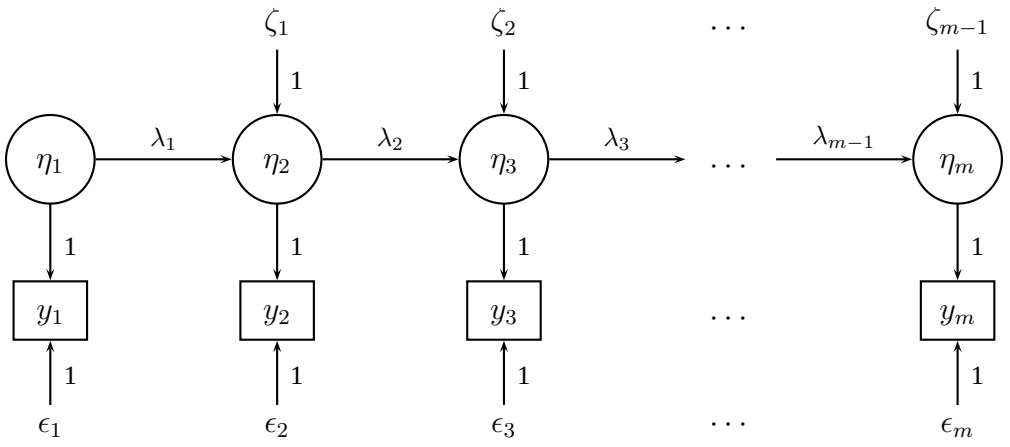
Autoregressive Modelle

In autoregressiven Längsschnittmodellen werden die zu einem bestimmten Zeitpunkt gemessenen Merkmalswerte durch die Werte aus vorangegangenen Messungen vorhergesagt. Werden auf diese Weise jeweils die Messungen von direkt aufeinander folgenden Zeitpunkten zueinander in Beziehung gesetzt, so spricht man von einer *Autokorrelation erster Ordnung*. Werden für die Bestimmung eines Merkmalswertes zu einem bestimmten Zeitpunkt mehrere vorausgegangene Messungen berücksichtigt, wird von einer Autoregression höherer Ordnung gesprochen. Der durch eine solche Verkettung sukzessiver Regressionen auf frühere Zustände abgebildete stochastische Prozess der Entwicklung von Merkmalen über die Zeit wird auch als *Markov-Prozess* bezeichnet.

In vielen Anwendungsbereichen konnte beobachtet werden, dass die Prädiktionskraft einer Messung für eine Folgemessung umso stärker nachlässt, je größer der zeitliche Abstand oder die Zahl intermediärer Messungen (*engl.* lag) wird. Diese Beobachtung, die durch das sukzessive Anwachsen unsystematischer bzw. nicht nachvollziehbarer Einflüsse erklärt werden kann, wurde von Jöreskog (1970) als *Simplex-Struktur* beschrieben.

In Abbildung 19 ist ein Markov-Simplex-Modell für die messfehlerbereinigten wahren Merkmalswerte beschrieben, das entsprechend als *Quasi-Markov-Simplex-Modell* bezeichnet wird.

Abbildung 19: Quasi-Markov-Simplex-Modell.



Das Messmodell und die strukturellen Beziehungen zwischen den wahren Merkmals-

werten können dabei durch

$$\begin{aligned} y_{si} &= \eta_s + \epsilon_{si} \\ \eta_s &= \lambda_{s-1}\eta_{s-1} + \zeta_{s-1} \end{aligned} \tag{68}$$

beschrieben werden. Da in diesem Modell eine strukturelle Beziehung zwischen latenten Konstrukten postuliert ist, werden die bislang lediglich als exogene (d.h. nicht durch das Modell erklärten) Variablen beschriebenen wahren Merkmalswerte (bis auf den ersten Messzeitpunkt) zu endogenen (also vorhergesagten) Modellvariablen. Neben dem systematischen Vorhersageanteil muss darum für alle Folgezeitpunkte $s > 1$ ein Residualwert ζ_s eingebunden werden.

Die Parallelen zwischen dieser Modellierung und dem von Andersen (1985) vorgeschlagenen probabilistischen Modell für wiederholte Testungen sind offensichtlich, auch wenn sich letzteres auf die Abschätzung der Korrelationen zwischen den Truescores zu den verschiedenen Messzeitpunkten beschränkt statt die Regressionsparameter direkt zu schätzen. Damit teilen beide Konzeptionen aber auch die Schwäche, zwar die *Verteilung der Truescores* zu allen Meszeitpunkten, nicht aber die zwischen diesen Zeitpunkten stattfindenden individuellen *wahren Merkmalsveränderungen* zu beschreiben.

Eine weitere zentrale Eigenschaft autoregressiver Modelle ist die direkte Berücksichtigung der *zeitlichen Ordnung* der verschiedenen Messungen. Im Gegensatz zur direkten Vorhersage sukzessiver beobachteter oder latenter Merkmalswerte in autoregressiven Modellen, ist die spezifische Reihenfolge der Messzeitpunkte bei der von Andersen vorgeschlagenen lediglich korrelativ verbundenen Messstruktur irrelevant. Ähnliches gilt auch für weitere faktoranalytische Modelle für Längsschnittdaten, bei denen die zeitliche Struktur entweder als zusätzliche Information in die Modellspezifikation integriert werden muss (z.B. in Growth Curve Models), oder durch den Forschenden erst post hoc bei der Interpretation der Ergebnisse berücksichtigt wird (z.B. bei RM-ANOVA).

Für die hier bearbeitete Fragestellung ist der Beitrag autoregressiver Modelle im Wesentlichen in der Veranschaulichung der Simplex-Struktur und der Kontrastierung mit Modellen zu sehen, die eine explizite Schätzung individueller wahrer Merkmalsveränderungen vorsehen.

Faktoranalytische Modelle

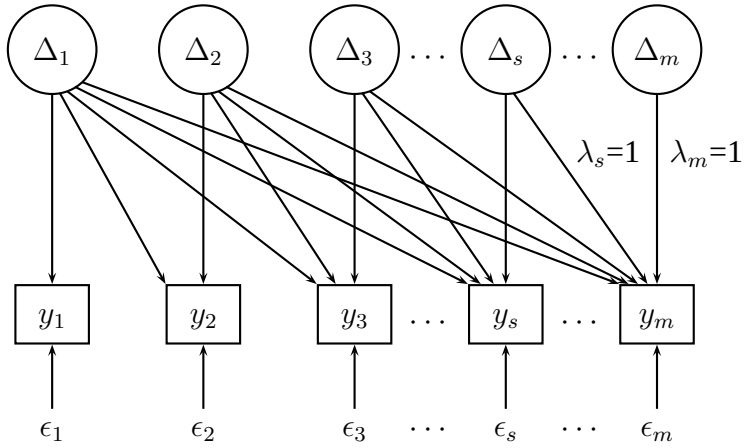
Während autoregressive Verfahren eine Modellierung der Verteilungen der wahren Werte selbst vorsehen und Stabilität bzw. Veränderung durch den Vergleich dieser Truescoreverteilungen abschätzen, werden latente Merkmalsveränderungen in faktoranalytischen Zugängen als eigenständige Komponenten modelliert. Eine solche direkte Schätzung wahrer Merkmalsveränderung hat den großen Vorteil, dass die Veränderung selbst durch weitere Variablen vorhergesagt werden kann. Durch die Möglichkeit der Erklärung wahrer Schmerzveränderung durch spezifische Merkmale der Person oder der Situation können genauere Anhaltspunkte für die Abschätzung der Validität der Schmerzmessung

gewonnen und konkrete Handlungsanleitungen für den Umgang mit entsprechend vulnerablen Bewohnern entwickelt werden.

Für die vorliegende Arbeit erscheinen beide mit dieser faktoriell orientierten Konzeption von Veränderungsmessung verbundenen Möglichkeiten relevant. Körperlich-funktionelle Beeinträchtigungen (z.B. des Bewegungsapparates) könnten als Prädiktor für eine Schmerzsteigerung beim Wechsel von Ruhe zu Aktivität zur Entwicklung von zielgruppenspezifischen Verhaltensinventaren zur Schmerzmessung beitragen. Informationen zur Kontextgebundenheit der psychometrischen Eigenschaften vorgeschlagener Schmerzindikatoren ermöglichen nicht nur eine reliablere Erfassung von Schmerzänderungen, sondern auch spezifischere (nicht-)medikamentöse Interventionen, und schließlich eine rational geleitete Standardisierung von Beobachtungsbedingungen.

Das ursprünglich von McArdle und Aber (1990) als „difference components change model“ vorgestellte *Modell latenter Differenzkomponenten* (LDCM; s. Abb. 20) begreift den zu einem bestimmten Zeitpunkt vorliegenden Truescore als Summe sukzessiver wahrer Merkmalsveränderungen, die sich seit dem ersten Messzeitpunkt ereignet haben.

Abbildung 20: Modell latenter Differenzkomponenten.



Die wahre Veränderung im latenten Merkmalswert, die sich zu einem Folgezeitpunkt $s > 1$ ergeben hat, wird als latente Differenzkomponente

$$\Delta_{sj} = \eta_{sj} - \eta_{(s-1)j} \tag{69}$$

modelliert, so dass sich der Truescore η_2 zum zweiten Messzeitpunkt als Summe des Ausgangsniveaus η_1 und der geschätzten Truescoreveränderung Δ_1 ergibt, der Truescore zum dritten Messzeitpunkt als Summe des Ausgangswertes und der beiden sukzessiven Veränderungsbeträge Δ_1 und Δ_2, und so weiter:

$$\begin{aligned}
\eta_{1j} &= \Delta_{1j} \\
\eta_{2j} &= \eta_{1j} + \Delta_{2j} = \Delta_{1j} + \Delta_{2j} \\
\eta_{3j} &= \eta_{2j} + \Delta_{3j} = \Delta_{1j} + \Delta_{2j} + \Delta_{3j}.
\end{aligned}$$

Das Messmodell für die zu einem bestimmten Zeitpunkt beobachteten Verhaltensweisen kann damit zum einen als Funktion

$$y_{sij} = \eta_{sj} + \epsilon_{sij} \tag{70}$$

des latenten Truescores selbst, oder aber als eine Funktion von Ausgangswert und sukzessiven wahren Veränderungsbeträgen

$$y_{sij} = \sum_{s=1}^{m} \Delta_{sj} + \epsilon_{sij} \tag{71}$$

konzeptualisiert werden (vgl. Steyer et al., 1997, S. 24). In der Nomenklatur der zuvor beschriebenen IRT-Modelle entspricht diese Definition des LDCM durch ein latentes Merkmalskomposit einem mehrdimensionalen Modell für wiederholte Messungen. Die Varianz des geschätzten Differenzfaktors ψ_{Δ_1} lässt sich als inter-individuelle Unterschiedlichkeit in der intra-individuellen Merkmalsveränderung interpretieren.

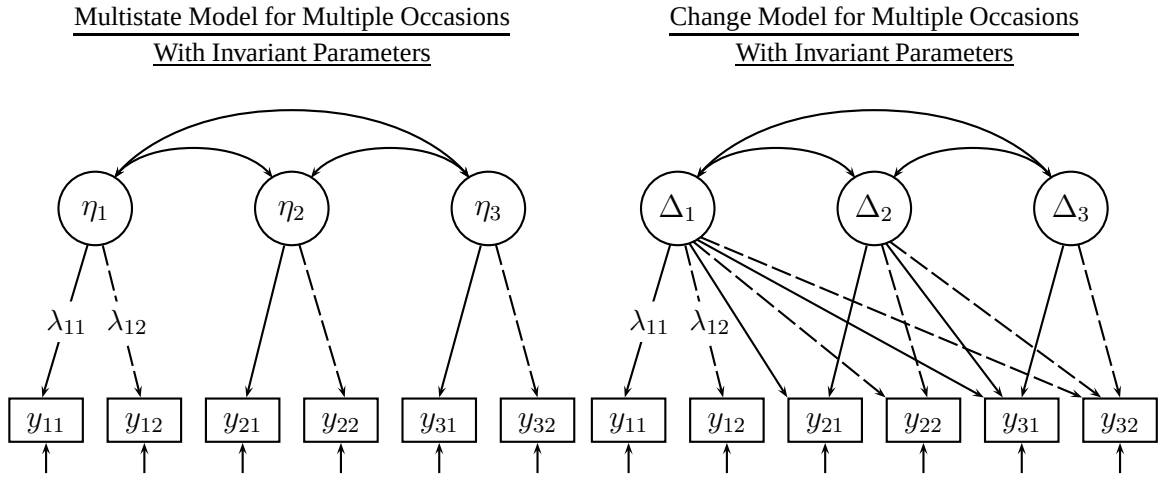
In seiner ursprünglichen Formulierung nimmt das „difference components change model“ von McArdle und Aber (1990) die latenten Differenzkomponenten als wechselseitig unabhängig an (vgl. Abb. 20). Diese Beschränkung erscheint jedoch eher dem Umstand geschuldet, dass das beschriebene Ein-Indikator-Modell nicht identifiziert wäre, wenn die Kovarianz zwischen den Differenzkomponenten ebenfalls geschätzt werden sollte. Steyer und Kollegen zeigen, dass für eine vergleichbare Modellierung, die sie als „true intraindividual change models“ (TIC_2) bezeichnen, eine freie Schätzung dieser Parameter möglich wird, wenn zu jedem Zeitpunkt mehrere Indikatoren verfügbar sind (Steyer, Eid & Schwenkmezger, 1997; Steyer, Partchev & Shanahan, 2000). Allerdings unterscheidet sich die vorgestellte Konzeption doch deutlich vom hier betrachteten Modell, da zunächst nur parallele (also essentiell τ-equivalente) Indikatoren und keine Mittelwertsstruktur berücksichtigt wurden (vgl. auch Abb. 20). In einer weiteren Arbeit generalisieren die Autoren dieses Modell vor dem Hintergrund des „multistate model (for multiple occasions of measurement) with invariant parameters“ (MISP) auf kongenerische Indikatoren (mit unterschiedlichen Faktorladungen) und beschreiben Möglichkeiten einer Modellierung latenter Merkmalsniveaus (Steyer et al., 2000). Einen Vergleich der äquivalenten Definitionen von MISP und dem von den Autoren als „Neighbor change model with invariant parameters“ bezeichneten Latenten Differenzkomponentenmodell erlaubt die nachfolgende schematische Veranschaulichung (Abb. 21).

Steyer, Partchev und Shanahan (2000) schlagen zwei alternative Skalierungen der nicht determinierten latenten (Differenz-)Komponenten vor. Eine direkte Skalierung wird durch eine z-Standardisierung der Truescoreverteilung zum ersten Erhebungszeitpunkt erreicht, indem

$$E(\eta_1) = \alpha_1 = 0 \quad \text{und} \quad Var(\eta_1) = \psi_{\eta_1} = 1 \tag{72}$$

gesetzt werden. Eine indirekte Skalierung der latenten Komponenten wird erreicht, wenn zu jedem Erfassungszeitpunkt s beispielsweise das Intercept des ersten Indikators jeweils

Abbildung 21: Schematische Darstellung MISP und TIC.



auf Null und die entsprechende Faktorladung jeweils auf einen Wert von Eins restringiert werden:

$$\nu_{1s} = 0 \quad \text{und} \quad \lambda_{1s} = 1. \tag{73}$$

In Tabelle 5 sind die Ausdrücke für die Identifikation aller Modellparameter unter beiden vorgeschlagenen Skalierungen zusammengestellt (vgl. Steyer et al., 2000, S. 116f).

Die in Tabelle 5 aufgeführten Formeln zur Identifikation der Modellparameter bei mindestens zwei beobachteten Indikatoren pro Messzeitpunkt beschreiben ein Messwiederholungsmodell mit äquivalenter Messstruktur (MISP, s. Steyer et al., 1997, 2000). Da das Veränderungsmodell mit latenter Differenzkomponente jedoch lediglich eine Re-Parametrisierung dieses Modells darstellt, sind mit

$$\alpha_{\Delta_2} = \alpha_{(\eta_2 - \eta_1)} = \alpha_2 - \alpha_1, \tag{86}$$

$$\psi_{\Delta_2} = \psi_{(\eta_2 - \eta_1)} = \psi_2 + \psi_1 - 2\psi_{2,1} \quad \text{und} \tag{87}$$

$$\psi_{\Delta_2,\Delta_4} = \psi_{(\eta_2 - \eta_1)(\eta_4 - \eta_3)} = \psi_{1,3} - \psi_{1,4} - \psi_{2,3} + \psi_{2,4} \tag{88}$$

auch die Modellparameter der wahren Merkmalsveränderungen identifiziert.

4.3.3.4 LDCM mit dichotomen Indikatoren

Die konzeptionellen Parallelen zwischen LDCM und MRMLC sind offensichtlich. Durch die mit der Anzahl der Messungen zunehmende faktorielle Komplexität ist auch hier eine Simplex-Struktur für die Messwerte impliziert. Durch die Dekomposition des wahren Merkmalswertes zu einem Zeitpunkt in additive (Differenz-)Komponenten kann auch hier von einem kompensatorischen Modell gesprochen werden. De facto lassen sich

Tabelle 5: Identifikation des Latent Difference Component Modells

Skalierung durch $\alpha_1 = 0$ und $\psi_1 = 1$	Skalierung durch $\nu_{1s} = 0$ und $\lambda_{1s} = 1$
$\nu_i = \mu^*_{i1}$ (74)	$\nu_i = \mu^*_{is} - \lambda_{is}\mu^*_{1s}, \quad i > 1$ (80)
$\lambda_i = \sqrt{\frac{\sigma^*_{i1,i'1}}{\sigma^{*2}_{i's}/\sigma^*_{is,is'}}}, \quad i \neq i', s \neq s'$ (75)	$\lambda_i = \frac{\sigma^*_{is,1s}}{\psi_s}, \quad i > 1$ (81)
$\alpha_s = \frac{\nu_{is} - \nu_{i1}}{\lambda_i}, \quad s \geq 1$ (76)	$\alpha_s = \nu_{1s}$ (82)
$\psi_s = \frac{\sigma^*_{is,i's'}}{\lambda_i \lambda_{i'}}, \quad s \geq 1$ (77)	$\psi_s = \frac{\sigma^*_{1s,is}}{\sqrt{\sigma^*_{is,is'}/\sigma^*_{1s,1s'}}}, \quad i > 1, s \neq s'$ (83)
$\psi_{s,s'} = \frac{\sigma^*_{is,i's'}}{\lambda_i \lambda_{i'}}$ (78)	$\psi_{s,s'} = \frac{\sigma^*_{is,i's'}}{\lambda_i \lambda_{i'}}$ (84)
$\theta_{is} = \sigma^{*2}_{is} - \lambda^2_i \psi_s$ (79)	$\theta_{is} = \sigma^{*2}_{is} - \lambda^2_i \psi_s$ (85)

ein LDCM mit mehreren dichotomen Indikatoren und die zuvor beschriebene zweiparametrische Erweiterung des MRMLC als unterschiedliche Formulierungen des selben Modells verstehen.

In Abbildung 22 ist das auf die vorliegende Fragestellung bezogene Analysemodell als latentes Differenzkomponentenmodell für unter verschiedenen Beobachtungsbedingungen beobachtete dichotome Schmerzindikatoren grafisch veranschaulicht (vgl. auch Abb. 16, 15 und 20).

Bezogen auf das hier bearbeitete Problem der Schmerzerfassung durch dichotome Verhaltensindikatoren, und in der beschriebenen Formulierung des IRT-Modelles mit latenten Responsevariablen, entspricht das Messmodell für den wahren Merkmalswert zum initialen Erfassungszeitpunkt $(s = 1)$ also einem gewöhnlichen eindimensionalen zweiparametrischen Messmodell

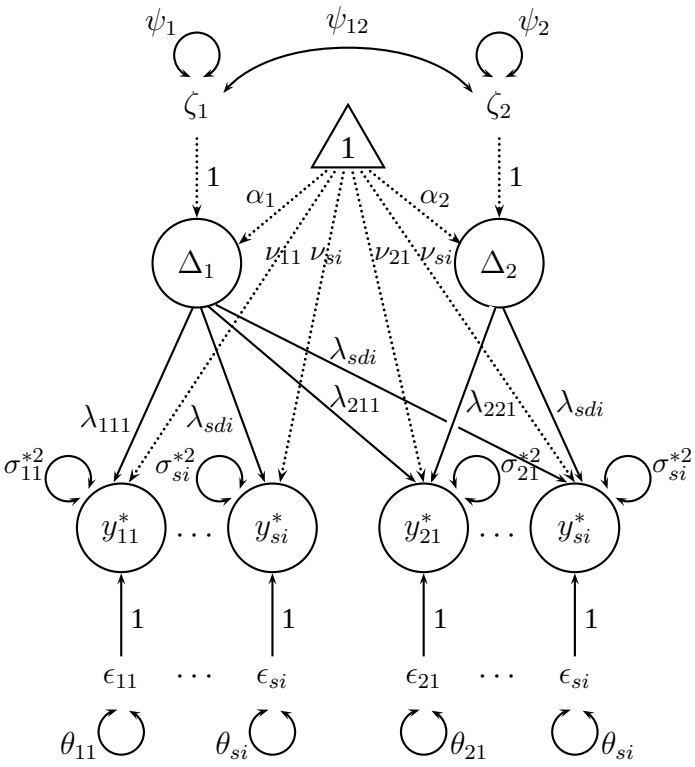
$$y^*_{1ij} = v_{1i} + \lambda_{11i}\Delta_{1j} + \epsilon_{1ij} \tag{89}$$

wie es in den vorangegangenen Kapiteln dargestellt wurde. Bei der Bestimmung der wahren Merkmalsveränderungen zwischen beiden Beobachtungssituationen dient der Truescore der Ruhesituation als Baseline-Messung. Das Ausmaß der latenten schmerzbezogenen Verhaltensreaktion y^* auf die Aktivierung – und damit die Wahrscheinlichkeit dafür, ob ein individueller Indikator y in der Aktivitätssituation beobachtet werden kann – hängt somit zum einen von der generellen, unbedingten Schmerzbelastetheit der Bewohner ab, und zum zweiten davon, wie sensibel er auf Bewegung oder Aktivierung reagiert:

$$y^*_{2ij} = v_{2i} + \lambda_{21i}\Delta_{1j} + \lambda_{22i}\Delta_{2j} + \epsilon_{2ij}. \tag{90}$$

Prinzipiell wäre auch eine umgekehrte Kontrastierung von Ruhe- und Aktivitätssituation mit einer entsprechenden Interpretation des zu erwartenden reduzierten Schmerzni-

Abbildung 22: LDCM für wiederholte Schmerzbeobachtung mit dichotomen Indikatoren.



veaus in Ruhe als Vermeidungs- bzw. Linderungseffekt denkbar. Da für die meisten der vorgeschlagenen Instrumente jedoch keine oder nur vage Aussagen zur Bedeutung von Aktivierung gemacht werden, und ein möglichst uneingeschränkter/universeller Anwendungsbereich eines Instrumentariums zur Schmerzerfassung sicherlich ein sinnvolles allgemeines Zielkriterium darstellt, wird im Rahmen dieser Arbeit die weniger spezifische Beobachtungssituation Ruhe als Baseline definiert.

Damit der zusätzliche Faktor in der zweiten Beobachtungsbedingung als wahre Merkmalsveränderung durch Aktivität interpretiert werden kann, muss die Messstruktur des interessierenden latenten Merkmals zu allen Zeitpunkten invariant sein. Aus diesem Grunde werden alle itemspezifischen Faktorladungen für beide Zeitpunkte und latente Merkmalskomponenten auf jeweils den gleichen Wert $\lambda_{sdi} = \lambda_{s'd'i}$ restringiert (vgl. Abb. 21).

Aus Gründen der Übersichtlichkeit wurden in Abbildung 22 die Pfadkoeffizienten der Mittelwertsstruktur als gepunktete Linien dargestellt (vgl. Abb. 16). Der zu schätzende Interceptparameter ν_{si} entspricht durch die gewählte Skalierung des latenten Merkmalsfaktors ($\alpha_s = 0$) dem Erwartungswert μ^*_{si} der latenten Responsevariablen (s. auch nächster Abschnitt). Da auch die LRV keine natürliche Skala besitzt, wird deren Erwartungswert gewöhnlich ebenfalls auf Null restringiert.

Im Kontext von Skalen zur Verhaltensbeobachtung kann der zu schätzende Threshold-

Parameter τ_{sic} als derjenige Schwellenwert auf einem gedachten latenten Verhaltenskontinuum verstanden werden, ab dem ein entsprechendes Verhalten 'salient' bzw. auffällig, und entsprechend als Beobachtung gekennzeichnet wird. In diesem Sinne könnte beispielsweise die Atmung als kontinuierlich zwischen Atemstillstand und Hyperventilation variierend angenommen werden, wobei Atempausen, angestrengte Atmung und Hyperventilation verschiedene Positionen auf diesem Responsefaktor repräsentieren. Um die inhaltliche Äquivalenz der Indikatoren zu beiden Erfassungszeitpunkten zu gewährleisten, werden darum auch ihre Threshold-Parameter zu beiden Zeitpunkten auf den gleichen Wert $\tau_{s\,i\,c} = \tau_{s-1\,i\,c}$ restringiert.

Wie zuvor bereits für das MRMLC beschrieben, können die latenten Merkmalsdimensionen, die prinzipiell keine natürliche Skalierung besitzen, je nach Forschungsinteresse unterschiedlich definiert werden. Die von Muthén und Asparouhov (2002) beschriebenen üblichen Restriktionen zur Skalierung des Merkmalsfaktors bei der Modellierung kategorieller Daten wurden in Kapitel 4.2.2.3 bereits dargestellt. Dabei wird der Mittelwert des latenten Merkmalsfaktors α auf Null restringiert und dessen Varianz ψ als freier Parameter geschätzt.

Wird das interessierende Merkmal mehrfach gemessen, so könnte jedoch auch die Frage nach dem wahren Merkmalsniveau bzw. dessen Änderung in den Vordergrund treten. Eine entsprechende alternative Skalierung würde die Varianz der Merkmalsfaktoren zu allen Zeitpunkten jeweils auf einen Wert von $\psi_s = 1$, und den Mittelwert des latenten Merkmales zum ersten Erfassungszeitpunkt auf $\alpha_1 = 0$ restringieren, damit die frei zu schätzenden mittleren Veränderungsbeträge für die Folgezeitpunkte identisch skaliert sind und somit direkt miteinander verglichen werden können.

4.3.3.5 Invarianz der Messstruktur

Wird beispielsweise ein Verhaltensinventar zur Schmerzerfasung wiederholt eingesetzt, so wird gewöhnlich davon ausgegangen, dass die Messstruktur über diese Zeitpunkte hinweg konstant bleibt, die Parameter der berücksichtigten Items also invariant sind.

Wie in Kapitel 3.4.6 bereits ausführlich dargelegt, erscheint es fraglich, ob die betrachteten Verhaltensinventare in den unterschiedenen Beobachtungsbedingungen (Ruhe und Aktivität) tatsächlich identisch funktionieren. Sowohl die Diskriminationskraft einzelner Verhaltensweisen, als auch deren Beobachtungshäufigkeiten (und somit ihre Schwierigkeit) können als substanziell durch den Erfassungskontext bestimmt angenommen werden.

Bei der empirischen Abschätzung von Invarianz beschritten faktoranalytische und item-response-theoretische Konzeptionen lange Zeit unterschiedliche, in weiten Teilen komplementäre Wege; die im Wesentlichen auf Kovarianzen beruhenden faktoranalytischen Zugänge interessierten sich hauptsächlich für Abweichungen im Ladungsmuster von Einzelindikatoren auf den angenommenen Faktoren bzw. veränderten Faktorstrukturen (Meade, Lautenschlager & Hecht, 2005). Da alle IRT-Anwendungen die korrekte Spezifikation des latenten Merkmalsraumes zu allen Erfasungszeitpunkten voraussetzen, scheinen diese

Verfahren weniger geeignet Veränderungen in der Bedeutungsstruktur der latenten Merkmale selbst, also einen potenziellen gamma-change, zu identifizieren.

Durch die Berücksichtigung der Responseraten (und damit gewissermaßen der Mittelwertsstruktur kategorieller Daten) standen dagegen bei der IRT-gestützten Abschätzung von Messinvarianz unterschiedliche Itemschwierigkeiten, und damit der beta-change, im Vordergrund (vgl. Reckase, 1997). Diese Ausrichtung ist sicherlich auch dadurch zu erklären, dass frühe Anwendungen dieser doch sehr datenintensiven Methodik hauptsächlich im Bildungs- und Beschäftigungssektor zur Leistungsmessung genutzt wurden, wo besondere Ansprüche auch an die Testfairness (i.S. der Vermeidung einer systematischen Benachteiligung bestimmter Personengruppen, i.e. *item bias*) formuliert werden müssen. Die Logik der Testverfahren, die zur Bestimmung eines solchen „differential item functioning" (DIF; Zumbo, 2007) in bestimmten Gruppen vorgeschlagen wurden, lässt sich prinzipiell auch auf die Fälle unterschiedlicher Testbedingungen (Kontext) oder wiederholter Testungen („item parameter drift"; Donoghue & Isham, 1998) übertragen (s. auch Muthén & Asparouhov, 2002).

Allgemeines Testverfahren

Verfahren zur empirischen Überprüfung invarianter Modellparameter sehen im Allgemeinen einen Vergleich der Anpassungsgüte einer Reihe von genesteten, d.h. unterschiedlich stark (gleich-)restringierten Modellen vor (*Likelihood Ratio (LR)*- bzw. χ^2-*Differenzen Test*). Das maximal restriktive Referenzmodell setzt dabei alle zu schätzenden (Schwierigkeits- und Diskriminations-)Parameter aller Indikatoren für die unterschiedenen Messbedingungen jeweils gleich. Wird eine freie Schätzung von Itemparametern (und damit diesbezügliche Invarianz) zwischen Gruppen, Situationen oder Zeitpunkten erlaubt, kann erwartet werden, dass sich der Model-Fit verbessert. Ist diese Verbesserung hinreichend deutlich (bzw. statistisch signifikant), liegt DIF vor und sollten die entsprechenden Itemparameter in der Population als unterschiedlich interpretiert werden.

Muthén und Asparouhov (2002) beschreiben verschiedene Möglichkeiten der DIF-Analyse im Kontext von Strukturgleichungsmodellen mit kategorialen Indikatoren. Obgleich die dort beschriebenen Verfahren zur Überprüfung der Parameterinvarianz *in verschiedenen Gruppen* nicht im Zentrum der vorliegenden Arbeit stehen, sollen die wesentlichen Linien ihrer Argumentation hier dargestellt werden, da diese gewissermaßen die Basis für eine Übertragung auf den hier bearbeiteten Fall wiederholter Messungen unter verschiedenen Erfassungsbedingungen bilden.

Welche Modellparameter zwischen den Gruppen verglichen werden können, hängt von der gewählten Standardisierung der latenten Responsevariable y_{ij}^* ab. Mplus bietet hier zwei Alternativen, die als *Delta-* und *Theta-Standardisierung* bezeichnet werden.

In Kapitel 4.2.2 wurden die Grundzüge einer Modellierung kategorialer Indikatoren mit LRV bereits beschrieben. In Analysen mit einer einzigen Population bzw. Gruppe werden die Varianzen σ_i^{*2} der unbeobachteten Responsevariablen gewöhnlich auf einen Wert von Eins standardisiert. Sollen jedoch mehrere Gruppen $s = 1, 2, 3 \ldots, m$ betrach-

tet werden, die sich untereinander hinsichtlich Itemdiskriminationen λ_{si}, Merkmalsvarianz ψ_s und Fehlervarianz θ_{si} poteniell unterscheiden, können auch gruppenspezifische Varianzen der y^*_{ij}-Werte erwartet werden, die dieser einheitlichen Standardisierung auf $\sigma^{*2}_i = 1$ widersprächen.

Delta-Standardisierung

Eine flexiblere Standardisierung wird durch die Einführung des Skalierungsfaktors Δ (siehe Gl. 43) ermöglicht, so dass für jede Gruppe eine eigene Variabilität der y^*_{ij}-Werte durch

$$\Delta^{-2}_{si} = \sigma^{*2}_{si} \quad \text{bzw.} \quad \Delta_{si} = 1/\sqrt{\sigma^{*2}_{si}} \tag{91}$$

geschätzt werden kann. Wie zuvor beschrieben, stellt die Residualvarianz θ_{si} bei dieser Parametrisierung keinen eigens zu schätzenden Modellbestandteil dar, sondern ergibt sich als Residuum

$$\theta_{si} = \Delta^{-2}_{si} - \lambda^2_{si}\psi_s. \tag{92}$$

Zur Kontrastierung mehrerer Gruppen wird der Skalierungsfaktor Δ_{si} und damit die Varianz der LRV in einer Referenzgruppe jeweils auf Eins standardisiert und in den verbleibenden Gruppen frei geschätzt.

Theta-Standardisierung

Ergaben sich die Fehlervarianzen der y^*_{ij}-Werte zuvor als Residuum aus den im Modell geschätzten λ-, ψ- und Δ-Parametern, werden sie in der Theta-Standardisierung direkt geschätzt

$$\Delta^{-2}_{si} = \lambda^2_{si}\psi_s + \theta_{si}. \tag{93}$$

Dabei stellt nun jedoch der der Skalierungsfaktor Δ keinen frei schätzbaren, sondern durch die anderen Modellbestandteile determinierten Parameter dar. Der Theta-Ansatz erlaubt damit die freie Schätzung der Residualvarianzen für alle Gruppen außer der Referenzgruppe (hier wird θ_{1i}=1 restringiert) oder einen Test gleicher Residualvarianzen in verschiedenen Gruppen.

Sowohl die Wahl der geeigneten Parametrisierung, als auch die Auswahl und Überprüfung potenziell unterschiedlich funktionierender Indikatoren oder einzelner Itemparameter sollte dabei durch die theoretischen Vorannahmen und das Erkenntnisinteresse des Forschenden bestimmt werden. Ein Standardalgorithmus, der ein postuliertes Messmodell gewissermaßen auf Verletzungen der Invarianz „abklappert“, ist im hier verwendeten Softwarepaket M*plus* nicht vorgesehen.

Ein solches permutatives Verfahren für die Identifizierung von Items mit „lack of invariance“ (LOI) der Parameter ist jedoch beispielsweise mit dem Programmpaket IRTLRDIF verfügbar (Thissen, 2001). Hierbei wird in einem ersten Schritt für jedes Einzelitem überprüft, ob sich der Model Fit signifikant verbessert, wenn dessen Parameter in beiden

Bedingungen frei geschätzt werden könnten. In einem anschließenden Set von Modellvergleichen wird die Quelle der Invarianz (z.B. a oder b-Parameter in 2PL-Modellen) genauer bestimmt (s. Meade et al., 2005).

Die Modellierung wiederholt erfasster dichotomer Indikatoren und die Abschätzung der Invarianz der Messstruktur zu beiden Beobachtungszeitpunkten ist zu den für unabhängige Gruppen vorgestellten DIF-Analysen analog (Muthén & Asparouhov, 2002, S. 11). Das für die Beantwortung der vorliegenden Fragestellung vorgeschlagene LD-CM für dichotome Indikatoren selbst ist in seinem potenziellen Beitrag zur Klärung der Frage nach der Invarianz der Messstruktur in beiden Beobachtungssituationen ziemlich beschränkt: die Spezifizierung der latenten Differenzkomponente setzt per definitionem invariante threshold-Parameter τ_i in beiden Beobachtungssituationen voraus und auch die Faktorladungen resp. Itemdiskriminationen müssen für alle Zeitpunkte und beteiligte latente Merkmalskomponenten identisch sein. Da im Kontext dieses Modells damit aber keine Überprüfung von Messinvarianz möglich ist, soll diese in vorgeschalteten Analysen mit der in Abb. 21 (links) beschriebenen Spezifikation als „Multistate Model“ (MISP) untersucht werden.

4.3.4 Demenzspezifität

In Kapitel 3.4.7 wurde auf Personenmerkmale demenzkranker Menschen hingewiesen, welche die Möglichkeit der Schmerzerfassung durch Verhaltensindikatoren beeinflussen könnten. Die Spezifität eines Messinstrumentes für eine bestimmte Subpopulation lässt sich allgemein beschreiben als optimale Performance der berücksichtigten Indikatoren, die häufig durch gezielte Zusammenstellungen von Itembatterien für Personengruppen mit bestimmten Merkmalseigenschaften (z.B. hohes Schmerzniveau, geringe Kommunikationsfähigkeit etc.) sicherzustellen versucht wird.

Im Kontext des Latent Variable Modeling können solche Kovariaten direkt in das Modell eingebunden und ihr direkter oder durch die latenten Merkmalswerte vermittelter (also indirekter) Effekt auf die Indikatoren einer Messung abgeschätzt werden (s. Muthén & Asparouhov, 2002). Um direkte Effekte darzustellen, wird das Messmodell der latenten Responsevariablen y^*_{ij} (vgl. Gl.39) zu

$$y^*_{ij} = \nu_i + \lambda_i \eta_j + \kappa_i x_j + \epsilon_{ij} \tag{94}$$

erweitert, wobei x_j die individuelle Merkmalsausprägung der Kovariaten (z.B. die Zugehörigkeit zu einer bestimmten Gruppe) und κ_i ihren direkten Effekt auf die latente Indikatorvariable darstellen. Die Werte des in Frage stehenden latenten Merkmals selbst können ebenfalls durch die Kovariate bestimmt sein

$$\eta_j = \gamma x_j + \zeta_j, \tag{95}$$

wobei γ das Regressionsgewicht der individuellen Kovariate x_j und ζ_j das Residuum der nun endogenen Modellvariable η_j repräsentieren. Da in diesem Fall in den unterschiedenen Gruppen tatsächlich unterschiedliche Merkmalsniveaus vorliegen, kann hier – im

Gegensatz zu den direkten Effekten auf die Indikatoren – nicht von einem Bias gesprochen werden.

Bei einer Berücksichtigung direkter Effekte einer Kovariaten auf die latenten Responsevariablen wird das Vorhersagemodell für die Wahrscheinlichkeit der Beobachtung dichotomer Indikatorvariablen y_{ij} entsprechend erweitert

$$P(y_{ij} = 1|\eta_j, x_j) = F[-(\tau_i - \lambda_i\eta_j - \kappa_i x_j)\theta_{ij}^{-1/2}], \tag{96}$$

sodass sich durch den Ausdruck für den Thresholdparameter $\tau_i - \kappa_i x_j$ je nach Auspägung der Kovariate unterschiedliche Itemschwierigkeiten ergeben.

Wie aus (96) ersichtlich, bleibt eine lediglich über den latenten Merkmalsfaktor vermittelte *indirekte* Beeinflussung der latenten Responsevariablen durch eine berücksichtigte Kovariate für das eigentliche Messmodell selbst folgenlos.

Der prinzipielle Nachteil dieser Modellierung ist, dass nur die Invarianz der Thresholdparameter, nicht jedoch gruppenspezifisch variierende Faktorladungen untersucht werden können. Für die Abschätzung invarianter Diskriminationen stellt eine Mehrgruppenanalyse darum den potenteren Ansatz dar. In Anbetracht der geringen Stichprobengrößen, die in empirischen Untersuchungen gewöhnlich zur Repräsentation der Population Demenzkranker Menschen gewonnen werden können, erscheint ein Mehrgruppenansatz mit einer Splittung der Analysestichprobe jedoch wenig attraktiv.

4.3.5 Zusammenfassung

Dieses Kapitel beschrieb die Möglichkeiten und Grenzen verschiedener messtheoretischer Konzepte, sowohl das nicht direkt beobachtbare Personenmerkmal Schmerzbelastetheit, als auch die Funktionsweise und Güte der verwendeten Instrumente (BESD, CNPI) und ihrer Indikatoren abzuschätzen.

Eine maximal informative Verwertung der erfassten Daten begreift Schmerz danach als latentes Merkmal, das unterschiedlich gut durch die beobachteten Verhaltensindikatoren angezeigt wird. Die grundsätzliche Äquivalenz faktoranalytischer und probabilistischer Zugänge zur Abbildung latenter Merkmale wurde ausführlich dargestellt.

Theoretische Annahmen zur Funktionsweise der vorgeschlagenen, und im Praxisalltag häufig verwendeten Schmerzinventare (beispielsweise hinsichtlich der Diskriminationskraft oder dem angezeigten Schmerzniveau einzelner Indikatoren) können so einer expliziten empirischen Überprüfung zugeführt werden.

Die durch konkurrierende Schmerzinventare konkret beschriebenen Verhaltensindikatoren können als zufällige Realisationen aus einem Universum vergleichbarer behavioraler Schmerzindikatoren begriffen, und damit gemeinsam zur Messung des latenten Schmerzmerkmals verwendet werden. Die Skalierung alternativer Indikatoren auf einem gemeinsamen Schmerzfaktor erlaubt einen direkten Vergleich der psychometrischen Eigenschaften einzelner Verhaltensindikatoren, aber auch ganzer konkurrierender Schmerzinventare.

Der Einfluss von Merkmalen der Beobachtungssituation auf die Möglichkeit der Erfassung von Schmerzen durch Verhaltensinventare kann durch eine Überprüfung der In-

varianz der Messung in Situationen geringer und hoher Aktiviertheit abgeschätzt werden. Theoretisch stehen dabei nicht allein die zu erwartenden Beobachtungsraten (und damit die Itemschwierigkeiten) der Verhaltensmerkmale, sondern auch deren Reliabilität (resp. Diskriminationskraft) in Frage. Wo die Annahme der Invarianz der Messstruktur erfüllt erscheint, bieten sich verschiedene Möglichkeiten, Veränderungen in den wahren Merkmalswerten abzubilden. Eine attraktive Option stellt dabei die direkte Schätzung latenter Veränderungsbeträge dar. Wie dargelegt wurde, finden probabilistische Modelle zur Abschätzung wahrer Merkmalsveränderung in entsprechend spezifizierten Modellen latenter Differenzkomponenten ihre konzeptionelle Entsprechung.

Die methodischen Analyseverfahren zur Abschätzung der Spezifität der vorgeschlagenen Inventare für die Zielgruppe demenzkranker Menschen sind in weiten Teilen mit den bereits beschriebenen Verfahren zur Abschätzung von Messinvarianz vergleichbar. Sowohl Merkmale der Demenz selbst (z.B. Ätiologie und Verlauf), als auch nicht-demenzielle Merkmale des hohen Lebensalters (z.B. Multimorbidität) wurden in Kapitel 3.4.7 als Faktoren beschrieben, die den hochgradig individuellen Schmerzausdruck demenzkranker Menschen mitbestimmen und darüber sowohl die Anwendbarkeit als auch die Interpretierbarkeit einer beobachtungsgestützten Schmerzerfassung kompromittieren könnten. Die beschriebenen statistischen Verfahren erlauben eine Untersuchung dieser Einflüsse nicht nur auf die gemessenen Schmerzniveaus in unterschiedlichen Beobachtungssituationen, sondern auch auf die tatsächlichen Merkmalsveränderungen selbst.

Die in Kapitel 3.5 erarbeiteten Forschungsfragen sollen im empirischen Teil dieser Arbeit durch die Analyse der schmerzbezogenen Erfassungsinhalte der zweiten HILDE-Feldphase beantwortet werden. Zu diesem Zweck sind mehrere, hinsichtlich der verwendeten statistischen Methoden aber auch mit Blick auf die Bedeutung für die psychometrische Schmerzforschung, aufeinander aufbauende Auswertungsschritte vorgesehen.

1. Deskription der schmerzbezogenen Daten der zweiten HILDE-Feldphase und Analyse der Item- und Skaleneigenschaften der BESD und CNPI auf der Grundlage der KTT

2. Schätzung der Item- und Personenparameter auf der Grundlage der IRT und Vergleich der konkurrierenden Verhaltensinventare BESD und CNPI

3. Überprüfung der Invarianz der Schmerzerfassung in Ruhe und Aktivität und Schätzung der Vulnerabilität für aktivitätsinduzierten Schmerz

4. Abschätzung der Spezifität des Assessments für Bewohner mit unterschiedlichem Muster erhaltener Kompetenzen (Demenzspezifität)

5 Datenbasis – das Forschungsprojekt HILDE

Zur empiriegestützten Beantwortung der formulierten Forschungsfragen werden schmerzrelevante Daten eines breit angelegten Forschungsprojektes am Institut für Gerontologie

der Universität Heidelberg (Leitung: Prof. A. Kruse) herangezogen, an dem der Autor als wissenschaftlicher Mitarbeiter beteiligt ist. Das Projekt „Heidelberger Instrument zur Erfassung von Lebensqualität“ (HILDE) beschäftigt sich mit Fragen der Erfassung und qualitätsbezogenen Interpretation von objektiven und subjektiv erfahrenen Lebensumständen demenzkranker Menschen in Einrichtungen der stationären Altenhilfe.

In der Anlage dieses Forschungsvorhabens ist das Schmerzerleben der Bewohner als inhaltliche Dimension von Lebensqualität und moderierende Variable im Prozess der Ausbildung und Förderung von Lebensqualität ausgewiesen.

Vorerfahrungen aus früheren HILDE-Projektphasen und die langjährige Kooperation mit der Sektion Gerontopsychiatrie der Psychiatrischen Universitätsklinik Heidelberg konnten im Sinne inhaltlicher, methodischer und personeller Synergien bei der Operationalisierung und Erfassung von Schmerzen genutzt werden.

Im Folgenden wird das Forschungsprojekt HILDE hinsichtlich seiner Zielsetzungen, Projektphasen, Untersuchungsdesigns, Stichproben und für die vorliegende Arbeit relevanten schmerz- und lebensqualitätsbezogenen Erfassungsbereiche detailliert dargestellt.

5.1 Zielsetzungen des Projektes

Das Forschungsprojekt HILDE ist ein seit Juni 2003 vom Bundesministerium für Familie, Senioren, Frauen und Jugend gefördertes Projekt (Fördernummer: BMBFSJ 311-1700-3/1), das in einem iterativen Prozess Möglichkeiten einer allgemeinverbindlichen, im Pflegealltag handhabbaren Einschätzung der aktuellen Lebensqualität von Bewohnern stationärer Pflegeeinrichtungen entwickelt und evaluiert (Becker et al., 2005).

Wie bereits aus dem Titel ersichtlich, besteht das vorrangige Ziel des Projektes darin, ein Instrumentarium zu entwickeln, das eine wissenschaftlich fundierte und im Pflegealltag umsetzbare Erfassung und auf nachvollziehbare Qualitätskriterien bezogene Bewertung der Lebensumstände von stationär versorgten Menschen mit demenzieller Erkrankung ermöglicht. Über die Dokumentation der aktuellen Lebensqualität hinaus sollen sich aus dem Instrument bzw. Verfahren auch Potenziale zur Förderung der individuellen Lebensqualität Demenzkranker ableiten lassen.

Die besonderen Herausforderungen, die mit der Konzeption eines entsprechenden Instrumentes und der Abschätzung seiner Handhabbarkeit verbunden sind, wurden bei der Charakterisierung der Zielpopulation, der theoretischen Aufarbeitung des Begriffes der Lebensqualität und der Diskussion der Messproblematik bereits dargestellt. Der gegenwärtige Entwicklungsstand des HILDE-Instrumentes stellt das Ergebnis eines systematischen iterativen Prozesses der Gestaltung, empirischen Überprüfung, und Modifikation von Inhalten, Erfassungsform und Interpretationsgrundlage dar. Die genannten Anforderungen wurden dabei in sukzessiven Projektphasen mit jeweils unterschiedlicher Gewichtung adressiert.

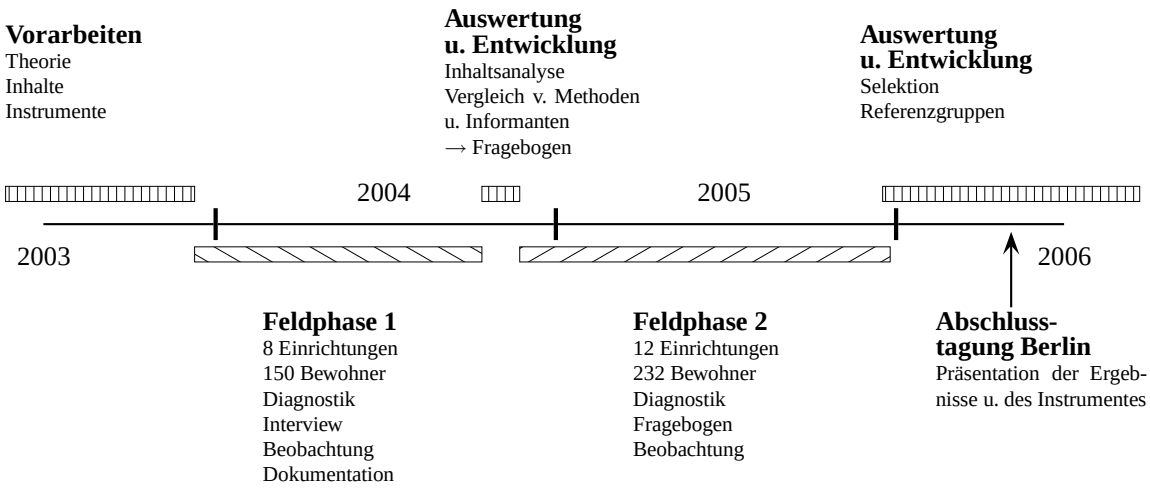
5.2 Phasen der Instrumentenentwicklung

Um dem Leser die Einordnung dieser Arbeit in den Gesamtkontext der Projektarbeit zu erleichtern, werden nachfolgend die wichtigsten entweder bereits abgeschlossenen oder zum Zeitpunkt der Erstellung laufenden konzeptionellen und empirischen Phasen des Projektes HILDE kurz beschrieben. Aufgrund der hier gewählten spezifischen Fragestellung kann diese Projektbeschreibung selbstverständlich nicht erschöpfend sein. Ebenso sollen nicht alle durch das Projekt erfassten prinzipiell schmerzbezogenen Informationen in gleichem Ausmaß dargestellt und für die empirische Bearbeitung der Fragestellung herangezogen werden. Eine Übersicht der verschiedenen Phasen und Meilensteile des Gesamtprojekts geben die nachfolgenden Abbildungen 23 und 24.

5.2.1 Förderphase 1 – Inhalte und Erhebungsmethodik

Im Rahmen der ersten Förderphase (Juni 2003 bis Mai 2006) wurden zunächst relevante inhaltliche Dimensionen der Lebensqualität demenzkranker Menschen aus der gerontologischen Literatur identifiziert. Diese theoretisch abgeleiteten bedeutsamen Lebensumstände und Kennzeichen erlebter Lebensqualität, sowie die in der Literatur beschriebenen Methoden und Instrumente ihrer Erfassung wurden daraufhin in zwei sukzessiv aufeinander aufbauenden Feldphasen in Einrichtungen der stationären Altenhilfe erhoben und analysiert.

Abbildung 23: Meilensteine der ersten Förderphase des Projektes HILDE



5.2.1.1 Prüfung theoretischer Inhalte und Erfassungsverfahren

Der Großteil der für die erste Feldphase (November 2003 bis September 2004, HILDE 1) zusammengestellten Untersuchungsmaterialien erforderte eine umfangreiche Beteiligung der sozialwissenschaftlichen und medizinischen Projektmitarbeiter bei der Erfassung lebensqualitätsrelevanter Marker. Die Erfassung der demenziellen Symptomatik und die Überprüfung der Einschlusskriterien wurden in diagnostischen Interviews mit den Bewohnern und Pflegenden durch einen Gerontopsychiater der Universitätsklinik geleistet. In teilstandardisierten Interviews mit den Bewohnern selbst, ihren Angehörigen und den Pflegenden in den Einrichtungen wurden theoretisch abgeleitete Chancen und Outcomes von Lebensqualität mit Blick auf den konkreten Bewohner erfragt und ggfs. durch die Befragten um weitere Aspekte ergänzt. Zur Abschätzung der Anwendbarkeit und Güte bestimmter theoretisch geeignet erscheinender Erhebungsinstrumente, wie beispielsweise der Verhaltensbeobachtung zur Identifizierung von Schmerzen oder verschiedener emotionaler Zustände, wurden parallele Erfassungen mit mehreren Instrumenten (in-vivo-Beobachtung von Emotionen und videogestützte Mimikanalyse), verschiedenen Beurteilern (Pflegende und Angehörige), oder in unterschiedlichen Situationen (Ruhe-, Aktivitäts- und Pflegesituationen) durchgeführt. Weitere Aspekte, z.B. zur (baulichen) Gestaltung der Wohnbereiche, oder bestimmte Informationen aus der Pflegedokumentation wurden direkt durch die Projektmitarbeiter eingeschätzt bzw. erfasst.

5.2.1.2 Erfassung von Lebensqualität durch Pflegende

Um die Akzeptanz und Handhabbarkeit des Instrumentes für die Pflegepraxis sicherzustellen, wurde für die zweite Feldphase (Oktober 2004 bis Dezember 2005, HILDE 2) das Untersuchungsmaterial soweit komprimiert, vereinfacht und standardisiert, dass die Einschätzung der Lebensqualität der Bewohner nunmehr weitgehend selbständig durch Pflegende erfolgen konnte. Die Erkenntnisse aus der ersten empirischen Phase flossen – beispielsweise durch Kategorisierungen offener Antworten – in die Überarbeitung des Instruments ein. Die diagnostische Untersuchung der Bewohner hingegen erfolgte weiterhin durch die medizinischen Mitarbeiter des Projekts. Auf der Grundlage der in diesem zweiten empirischen Studienabschnitt gesammelten Informationen wurde das Erfassungsmaterial für den Einsatz in der zweiten Projektförderphase nochmals kompakter und stringenter gestaltet.

5.2.1.3 Medizinische Diagnostik und Versorgung

Die zukünftige eigenständige Anwendung in den Einrichtungen durch Pflegemitarbeiter erscheint nur dann realisierbar, wenn eine zielgruppengerechte Einschätzung der Lebensqualität auch ohne die zwingende Einbindung von externen Sachverständigen (z.B. den Forschungspartnern oder Fachärzten) sinnvoll möglich ist. Damit stellte sich dem Projektteam die Aufgabe, solche demenzspezifischen Merkmale aus der bisher recht umfangreichen medizinisch-diagnostischen Untersuchung durch einen Gerontopsychiater, die in

besonderem Ausmaß mit den Möglichkeiten zur Realisierung von Lebensqualität verknüpft sind, zu identifizieren und für eine Bearbeitung durch Pflegende zugänglich zu machen. Eine kontinuierliche allgemeinmedizinische und gerontopsychiatrische Betreuung der Bewohner – Grundvorraussetzung für die möglichkeitsgerechte Realisierung von Lebensqualität – sollen und können solche Einschätzungen selbstverständlich nicht ersetzen. Diagnosestellung und Behandlungsanweisung müssen durch entsprechend ausgebildete Mediziner erfolgen. Im Rahmen der empirischen Untersuchungen der zweiten HILDE-Förderphase wird daher die Regelmäßigkeit bzw. Aktualität der medizinischen Begleitung erfragt und bei größeren Zeitabständen zur letzten Untersuchung eine erneute Vorstellung angeregt. Solche Merkmale der Demenzerkrankung, die für eine angemessene Einschätzung und Bewertung realisierter und potenzieller Lebensqualität unabdingbar scheinen, wurden detailliert anhand von Fallbeispielen beschrieben, so dass die Pflegenden das Muster individuell verfügbarer Bewohnerkompetenzen zuverlässig einschätzen können (siehe auch nächsten Absatz).

5.2.1.4 Qualitätskriterien zur Beurteilung von Lebensumständen

Eine zweite weitreichende Entscheidung des Projektteams betrifft die geforderte Verbindlichkeit von Qualitätsurteilen und daraus abgeleiteten Förderbedarfen und Potentialen. Eine Einteilung der demenziell erkrankten Bewohner lediglich anhand der Schwere ihrer kognitiven Beeinträchtigungen greift zu kurz und entspricht nicht immer auch dem Symptombild, das Pflegende erfahren und an dem sich auch die Art und der Umfang einer angemessenen Betreuung ausrichtet. Anhand der diagnostischen Kriterien Selbständigkeit im Alltagsleben, kognitiver Status und Belastung durch nicht-kognitive Demenzsymptome (z.B. Depression oder Apathie) wurden darum vier Bewohnergruppen mit jeweils spezifischem Muster erhaltener Kompetenzen identifiziert (Becker et al., 2006). Die in den bisherigen HILDE-Untersuchungen erfassten Lebensumstände dieser vier Bewohnergruppen werden als sozial-normativer Referenzstandard für die Interpretation der für einen individuellen Bewohner erfassten Lebensumstände bereitgestellt. Individuelle Bedürfnislagen, spezifische Förderpotenziale, aber auch Grenzen der Gestaltung bzw. Förderung von Lebensqualitäten können so einfacher herausgearbeitet und begründet werden, als es bei rein idiosynkratisch orientierten Interpretationen möglich scheint.

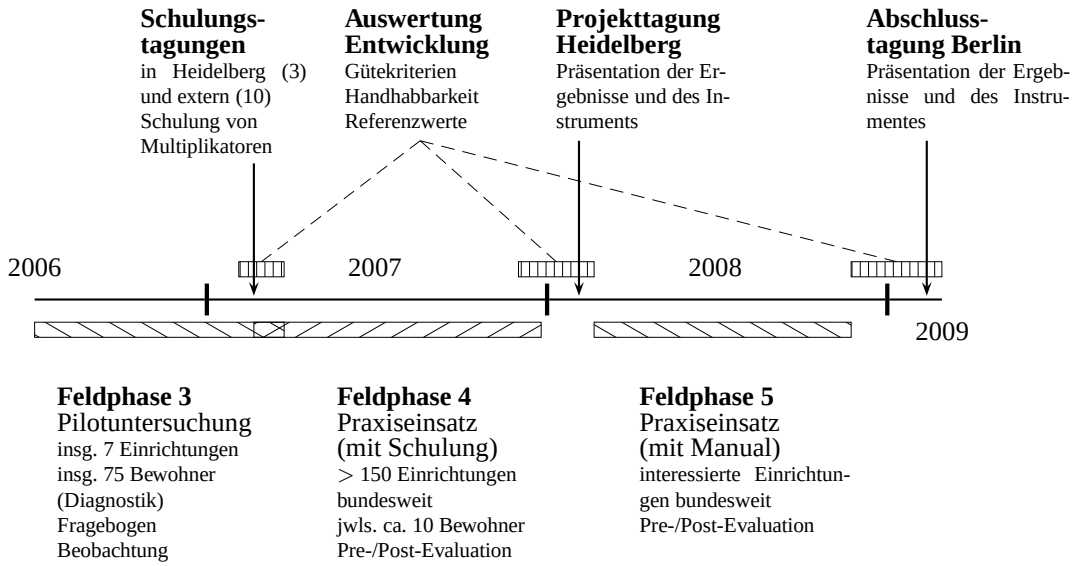
Zum Ende der ersten Förderphase konnte damit ein Instrument zur Erfassung von Lebensqualität bei Demenz vorgelegt werden, dass durch die enge Praxisanbindung über die gesamte Entwicklungsphase hinweg sowohl inhaltlich, als auch formal den im Pflegealltag gestellten Anforderungen entsprechen und darum eine weite Verbreitung und Anwendung finden kann.

5.2.2 Förderphase 2 – Optimierung der Praxistauglichkeit

Im Rahmen der zweiten Förderphase (Juni 2006 bis Februar 2009) wird die Handhabbarkeit des HILDE-Instruments in der Pflegepraxis gegenwärtig in möglichst vielen Einrich-

tungen im gesamten Bundesgebiet überprüft. Einen Überblick über die empirischen und konzeptionellen Aufgaben dieses Projektabschnittes gibt Abbildung 24.

Abbildung 24: Meilensteine der zweiten Förderphase des Projekts HILDE



Um auch für solche Erfassungsbereiche des aktuellen Instruments, die auf der Grundlage der Ergebnisse aus HILDE 2 maßgeblich modifiziert wurden, empirische Referenzwerte bereitstellen zu können und um umfangreichere Aussagen zu den psychometrischen Gütekriterien dieser Instrumentenversion zu ermöglichen, wurde HILDE zunächst in zwei unabhängigen Pilotprojekten mit unterschiedlichen Untersuchungsdesigns eingesetzt (Juli 2006 bis April 2007, HILDE 3). Gleichzeitig dienten diese Untersuchungen als Pilotstudien für den neu zu entwickelnden Evaluationsfragebogen, mit dem die in der Hauptphase selbständig mit HILDE arbeitenden Einrichtungen ihre Erfahrungen zur Handhabbarkeit des Instrumentes berichten sollen (seit Februar 2007, HILDE 4). Bis zum Ende der Projektlaufzeit sollen in einer weiteren fünften empirischen Erhebungsphase weitere Einrichtungen gewonnen werden, die den selbständigen Einsatz des Instrumentes (mit Manual und Erhebungsunterlagen) testen und evaluieren. Schließlich werden die Endergebnisse des Gesamtprojektes gegen Anfang 2009 der Praxis und Fachöffentlichkeit vorgestellt. Verhandlungen bezüglich der Publikation des HILDE-Instrumentes sind bereits eingeleitet.

5.2.2.1 Aussagemöglichkeiten

Das an der Fachhochschule Frankfurt am Main angesiedelte Projekt „MeDiA in Cura“ (Projektleitung: Prof. Ruth Schwerdt) realisiert wiederholte Messungen der Lebensqua-

lität mit HILDE an denselben Bewohnern vor und nach einer insgesamt 6 Monate umfassenden pflegewissenschaftlichen Intervention.

In einer weiteren durch die Projektgruppe initiierten Pilotstudie wurde in einer kooperierenden Einrichtung (Kooperationspartner: Prof. Dr. Hans Georg Nehen) der stationären Altenhilfe im Raum Essen die Lebensqualität der Bewohner zu einem ersten Erfassungszeitpunkt von zwei Pflegenden parallel eingeschätzt. Während die erste Pflegeperson dieses Tandems nach 5 Tagen eine erneute HILDE-Einschätzung vornahm, wiederholte die zweite Pflegeperson ihre Einschätzung erst nach 8 Tagen.

5.2.2.2 Bewertung durch die Praxis

Um trotz der knappen personellen Ressourcen des Projekts möglichst viele Einrichtungen aus dem gesamten Bundesgebiet für die Untersuchung der praktischen Handhabbarkeit des Instrumentes (HILDE 4) zu gewinnen, wurden in zentralen Schulungstagungen Multiplikatoren mit der Anwendung des Instruments vertraut gemacht, die diese Informationen wiederum an die zu beteiligenden Pflegenden aus ihren Einrichtungen weitergeben. Im Gegensatz zu früheren Untersuchungsphasen wurde nun lediglich ein verbindlicher Zeithorizont von ca. 10 Wochen für den Instrumenteneinsatz und die Evaluation vereinbart, Anzahl und Auswahl einzuschätzender Bewohner oder zu beteiligender Pflegekräfte jedoch können die interessierten Einrichtungen je nach den personellen und strukturellen Möglichkeiten frei bestimmen. Da der Schwerpunkt dieser Phase nun nicht mehr primär auf den Erfassungsinhalten, sondern auf der Anwendbarkeit des Instrumentes in der Praxis liegt, werden dem Projektteam nun nur noch die aggregierten Kennwerte der individuellen Lebensqualitätsprofile der Bewohner aus den Einrichtungen rückgemeldet. Auch wenn damit die Binnenstruktur einzelner Lebensbereiche nicht mehr nachvollzogen werden kann, sollen diese bewohnerspezifischen Daten zur weiteren Präzisierung der Referenzwerte genutzt werden.

Aufgrund dieser Verschiebung des Erkenntnisinteresses weg von individuellen Bewohnerdaten hin zur Fragen des Erhebungsverfahrens, der Implementierbarkeit in den Pflegealltag, der Akzeptanz durch die in der Praxis Tätigen sowie der Potenziale für die Verbesserung oder Entwicklung können die Daten der aktuellen Forschungsphase nur einen vergleichsweise kleinen Beitrag zur Beantwortung der im Rahmen dieser Arbeit gestellten Fragen leisten.

5.3 Schmerzbezogene Erfassungsinhalte

Aufgrund der theoretischen Vorüberlegungen zur Bedeutung des Schmerzerlebens für die Lebensqualität wurde diesem Aspekt individueller Vulnerabilität bei der Entwicklung des Gesamtverfahrens sowohl inhaltlich als auch bezüglich der Erfassung besondere Aufmerksamkeit gewidmet. Bei der Anlage des Gesamtprojektes wurden verschiedene Beurteiler, Alltagssituationen, Skalentypen und Instrumente berücksichtigt, die einen multiperspektivischen Blick auf das Schmerzerleben der einzuschätzenden Bewohner erlauben.

5.3.1 Selbstauskunft des Bewohners zu Schmerzen im klinischen Interview

Im ersten Projektabschnitt (HILDE 1 und HILDE 2) wurden alle Bewohner, die von den Einrichtungen als potenzielle Studienteilnehmer vorselektiert wurden, im Rahmen eines diagnostischen Gespräches dem gerontopsychiatrischen Projektmitarbeiter vorgestellt. Bewohner, die an der zweiten Feldphase von HILDE teilnehmen sollten, wurden während dieses Gespräches unter Anderem gefragt, ob sie zur Zeit unter Schmerzen litten. Reaktionen, die keine eindeutige Interpretation zuließen, wurden gesondert vermerkt[4]. Diejenigen Bewohner, die akute Schmerzen angaben, sollten zusätzlich auf einer dreistufigen Ratingskala einschätzen, wie stark diese Schmerzen wären[5].

5.3.2 Fremdratings der Schmerzbelastung durch Angehörige und Pflegende

In der umfangreichen ersten HILDE Erfassungsphase konnten schmerzbezogene Informationen von Angehörigen, Pflegenden, sowie in einigen Fällen auch von den behandelnden Hausärzten gewonnen werden. Die Angehörigen der Studienteilnehmer wurden gebeten, während eines Besuches auf mögliche Schmerzen zu achten und ein entsprechendes Erfassungsheft zu bearbeiten. Die erbetenen Informationen bezogen sich auf bekannte chronische Schmerzleiden vor dem Beginn der Demenzsymptomatik[6] und das Vorliegen[7], die Lokalisation[8] und die Intensität[9] akuter Schmerzen während des Besuches.

Für die Einschätzung von Schmerzbelastung beim Bewohner durch die Pflegenden wurden bis zu drei Pflegende, die den Studienteilnehmer betreuen, gemeinsam befragt. Eventuell abweichende Einschätzungen zu den Inhalten chronische und aktuelle Schmerzbelastung, Schmerzlokalisation und Schmerzintensität wurden dabei im Erfassungsheft ebenfalls festgehalten.

Schließlich wurde versucht, schmerzrelevante Informationen von den Hausärzten der Bewohner einzuholen. Die chronische und aktuelle Schmerzbelastung, -lokalisation und -intensität wurde wie zuvor beschrieben erfragt. Zusätzlich wurde nach einer länger andauernden Einnahme von Schmerzmedikamenten vor dem Heimeinzug gefragt[10].

Im Rahmen der zweiten Feldphase von HILDE – das Instrument war mittlerweile zu einem *Fragebogen* weiterentwickelt worden – wurden auch die schmerzbezogenen Informationen von den mit der Einschätzung der Lebensqualität eines Bewohners betrauten Pflegenden erhoben. Die Inhalte und Frageformulierungen wurden größtenteils un-

[4] „Haben Sie momentan Schmerzen? Tut Ihnen momentan etwas weh?“; 0=„nein“, 1=„ja“, 9=*keine klar interpretierbare Antwort.*

[5] „Wie stark sind Ihre Schmerzen?“; 1=„schwach“, 1=„mäßig“, 3=„stark“.

[6] „Litt Ihr(e) Angehörige(r) vor seiner/ihrer Demenz an chronischen Schmerzen (z.B. Kopfschmerzen, Rückenschmerzen etc.)?“; 0=„nein“, 1=„ja“, 9=„nicht bekannt“.

[7] „Hat Ihr(e) Angehörige(r) aus Ihrer Sicht bei Ihrem Besuch Schmerzen?“; 0=„nein“, 1=„ja“.

[8] „Wenn ja, in welcher(n) Region(en) vermuten Sie die Schmerzen hauptsächlich?“; Angabe auf Liste.

[9] „Wenn ja, wie stark waren aus Ihrer Einschätzung die Schmerzen?“; 0=„kein Schmerz“ bis 5=„unerträglich starker Schmerz“.

[10] „Nahm Ihr(e) Patient(in) vor der Heimeinweisung über längere Zeit Schmerzmedikamente ein?“; 0=„nein“, 1=„ja“, 9=„nicht bekannt“.

Tabelle 6: Schmerzratings verschiedener Informantengruppen der beiden ersten Feldphasen

	Angehörige	*Pflegende*	*Hausärzte*
Akute Schmerzbelastung	HILDE 1	HILDE 1+2	HILDE 1
Schmerzlokalisation	HILDE 1	HILDE 1+2	HILDE 1
Schmerzintensität	HILDE 1	HILDE 1+2	HILDE 1
Chronische Schmerzen	HILDE 1	HILDE 1+2	HILDE 1
Frühere Schmerzmitteleinnahme	–	–	HILDE 1

verändert aus der ersten Projektphase übernommen. Lediglich die Abstufung der Skala zur Intensität vorliegender Schmerzen wurde von fünf auf nunmehr drei Antwortkategorien vereinfacht[11].

5.3.3 Verhaltensbeobachtung durch verschiedene Informantengruppen

Die erste Feldphase der Instrumentenentwicklung wurde dazu genutzt, um die Möglichkeiten einer standardisierten Beobachtung schmerzrelevanten Ausdrucksverhaltens durch verschiedene Informantengruppen eingehend zu untersuchen. Zu diesem Zweck wurde ein kürzlich von der Projektgruppe „Schmerz im Alter“ der Deutschen Gesellschaft zum Studium des Schmerzes (DGSS) ins Deutsche übertragenes Instrument zur Beobachtung von schmerzbezogenem Ausdrucksverhalten eingesetzt (PAINAD; Warden, Hurley & Volicer, 2003). Über den gesamten Prozess der Entwicklung des HILDE-Instrumentariums hinweg wurden verschiedene Versionen dieser Skala zur ’BEurteilung von Schmerzen bei Demenz’ (BESD; Schuler et al., 2007) erarbeitet und getestet. Neben der Originalskala, die jeden der einbezogenen Verhaltensbereiche nach der Schmerzintensität mit 0 bis 2 Punkten gewichtet, wurde eine Version erstellt, bei der alle beschriebenen Verhaltensindikatoren dichotom abgefragt werden. Aufgrund der Erfahrungen aus der ersten Erfassungsphase wurden der Wortlaut oder die logische Binnenstruktur mancher Indikatoren bzw. -bereiche für die Verwendung in der zweiten Erfassungsphase präzisiert.

Das Zeitfenster für die Beobachtungen mimischer, gestischer und vokaler Verhaltensweisen, die auf Schmerz hinweisen, wurde – unabhängig vom Beurteiler oder der Situation, in der diese Beobachtung stattfinden sollte – auf zwei Minuten beschränkt. Damit wird das Beobachtungsintervall kürzer gesetzt als für die Pilotstudien der Originalskala berichtet (5 Minuten). Der Vorteil eines vergleichsweise engen Intervalls liegt für diese Untersuchungen u.E. in der größeren Konstanz der situativen Lage der Bewohner (Ruhe, Aktivität) und der leichteren Implementierbarkeit in den Pflegealltag.

In der Literatur scheint Einigkeit darüber zu bestehen, dass eine schmerzbezogene Ver-

[11] „Wie stark würden Sie die Schmerzen Ihrer Meinung nach einschätzen?“; 1=„leichter“, 2=„mäßiger“, 3=„starker Schmerz“.

haltensbeobachtung die situativen Bedingungen, in denen sich der Einzuschätzende befindet, berücksichtigen sollte. In Situationen körperlicher Aktivierung (insbesondere Bewegung) kann beispielsweise mehr schmerzbezogenes Verhalten erwartet werden, wenn die normale Funktion des Bewegungsapparates durch arthritische Veränderungen geschädigt ist. Aber auch die soziale Rahmung einer Situation (z.B. soziale Erwünschtheit) kann dazu beitragen, dass schmerzbezogenes Verhalten situationsadäquat gesteuert wird und damit häufiger oder seltener beobachtet werden kann. Wie bereits ausgeführt, können bestimmte Charakteristika der Beobachtungssituationen die Messung von Schmerzen mit Hilfe von behavioralen Indikatoren erschweren, falls diese neben ihrer Spezifität für Schmerzen in solchen Situationen auch unabhängig vom Schmerzniveau gehäuft oder seltener beobachtet werden können (z.B. angestrengte Atmung, angezogene Knie). In diesem Projekt wurden darum Bewohner sowohl in Ruhe als auch bei Aktivitäten beobachtet. Eine weitere Differenzierung erfolgte nach der Tageszeit der Beobachtung, wobei die Bewohner sowohl morgens (8-10 Uhr) als auch abends (18-20 Uhr) beobachtet wurden.

Neben den oben bereits beschriebenen allgemeinen Ratings zur Schmerzbelastung der Bewohner wurden die Angehörigen gebeten, während eines Besuches schmerzrelevante Indikatoren anhand der BESD-Skala (Einzelitemversion) zu dokumentieren. Die entsprechenden Erfassungsbögen wurden den teilnehmenden Angehörigen im Vorfeld zugeschickt. Zusätzlich wurden die Angehörigen telefonisch darum gebeten, diese Unterlagen bei ihrem nächsten Besuch mitzubringen und zu bearbeiten. Die jeweils 2-minütigen Beobachtungsintervalle sollten dabei eine Ruhesituation und eine Situation gesteigerter Aktivität umfassen.

Weitere Schmerzbeobachtungen wurden während der beiden Bewohnerinterviews von den Mitarbeitern des Projektes durchgeführt. Jeweils zwei Projektmitarbeiter dokumentierten unabhängig voneinander die beobachteten Schmerzindikatoren anhand des Verhaltensinventares in seiner ursprünglichen kategoriellen Form. Dabei wurden die Bewohner in Ruhe und in einer Aktivitätssituation eingeschätzt.

Schließlich nahmen für die erste Feldphase auch jeweils zwei Pflegende Schmerzbeobachtungen am Bewohner vor (BESD Einzelitemversion). Die wiederum 2-minütigen Beobachtungsintervalle umfassten eine Situation am Morgen und eine am Abend.

Zur weiteren konkordanten Validierung der BESD-Schmerzerfassung wurde für den zweiten empirischen Studienteil zusätzlich die *Checklist of Nonverbal Pain Indicators* (*CNPI*; Feldt, 2000) in die Schmerzbeobachtung integriert. Da die Einschätzung der Tröstbarkeit innerhalb der BESD-Skala eine explizite Intervention vorsieht, wurde dieser Itemblock nach den CNPI-Items ans Ende der kombinierten Indikatorliste gesetzt.

Da die Einschätzung der Lebensqualität während der zweiten Feldphase des HILDE-Projektes mit Ausnahme der medizinisch-diagnostischen Informationen durch Mitarbeiter der Pflege vorgenommen wurde, liegen auch die Schmerzbeobachtungen nun nur noch von der Informantengruppe der Pflegenden vor. Die Beobachtungsintervalle wurden weiterhin bei zwei Minuten festgelegt, und als Bedingungsvariation wurden die Bewohner in einer Ruhe- und einer Aktivitätssituation eingeschätzt. Einen Überblick der eingesetzten Beobachtungsinstrumente, realisierten Beobachtungssituationen und Informantengruppen in

beiden empirischen Studienteilen der ersten HILDE-Förderphase gibt Tabelle 7.

Tabelle 7: Schmerzbeobachtungen durch verschiedene Informantengruppen in den beiden ersten Feldphasen

		Angehörige	*Pflegende*	*Projektmitarbeiter*
HILDE 1	BESD Originalversion[1]	–	–	Ruhe/Aktivität
	BESD Einzelitems[2]	Ruhe/Aktivität	morgens/abends	–
HILDE 2	BESD Einzelitems	–	Ruhe/Aktivität	–
	CNPI Einzelitems[2]	–	Ruhe/Aktivität	–

[1] Scoring für alle 5 Indikatorbereiche: 0-2 Punkte.
[2] Antwortformat für Einzelindikatoren: 0=„nicht beobachtet“, 1=„beobachtet“.

5.3.4 Die Skala *BEurteilung von Schmerzen bei Demenz* (BESD)

Die im Original von Warden, Hurley und Volicer (2003) vorgestellte *Pain Assessment in Advanced Dementia Scale* (PAINAD) wurde auf der Grundlage der *Discomfort Scale - Dementia of the Alzheimer Type* (DS-DAT; Hurley et al., 1992) und des *Face, Legs, Activity, Cry and Consolability Pain Assessment Tools* (FLACC; Merkel et al., 1997) entwickelt und umfasst die fünf Indikatorbereiche (1) Atmung, (2) Lautäußerung, (3) Mimik, (4) Körperhaltung und (5) Trost. Für jeden Ausdrucksbereich sind mehrere Verhaltensmerkmale beschrieben, die unterschiedliche Schmerzintensitäten anzeigen (z.B. gelegentlich angestrengt atmen vs. lautstark angestrengt atmen). In Abhängigkeit davon ob, und falls ja, welche Verhaltensweisen beobachtet werden können, werden die unterschiedenen Ausdrucksbereiche mit 0, 1 oder 2 Punkten geratet (siehe Abb. 25). Eine verbale Umschreibung dieser Punktwerte bzw. Intensitätskategorien geben die Autoren nicht. Auch die empirische Grundlage für diese implizite Gewichtung der Einzelindikatoren und deren Beziehung zueinander bleiben unkommentiert. Der Gesamtscore der Originalversion errechnet sich aus den Subscores der 5 Indikatorbereiche und umfasst somit einen Wertebereich von 0 bis 10 Punkten[12]. Cut-off-Kriterien für das Vorliegen von Schmerzen oder die Indikation von Interventionsmaßnahmen werden leider nicht bereitgestellt.

5.3.4.1 Befunde aus dem anglo-amerikanischen Raum

Die Pilottestung umfasste im Rahmen einer ersten Studie die parallele Beobachtung von 19 schwer beeinträchtigten Demenzkranken durch Pflegende in Situationen angenehmer Aktivität, Ruhe und potenziell aversiv erlebter Pflegehandlungen. In einer zweiten Studie wurden Beobachtungen für weitere 25 Heimbewohner vor und nach einer medikamentösen Schmerztherapie durchgeführt. Für diese Stichprobe konnten jedoch keine

[12] 0=„keine Schmerzen“, 10=„starke Schmerzen“.

Abbildung 25: Originalskala *Pain Assessment in Advanced Dementia* (PAINAD)

Item/point value	**0**	**1**	**2**
Breathing, independent of vocalization	Normal	Occasional laboured breathing; short period of hyperventilation	Noisy, laboured breathing; long period of hyperventilation; Cheyne-Stokes respirations
Negative vocalization	None	Occasional moan or groan; low-level speech with a negative or disapproving quality	Repeated, troubled calling out; loud moaning or groaning; crying
Facial expression	Smiling or inexpressive	Sad; frightened; frowning	Facial grimacing
Body language	Relaxed	Tense; distressed; pacing; fidgeting	Rigid; fists clenched; knees pulled up; pulling or pushing away; striking out
Consolability	No need to console	Distracted or reassured by voice or touch	Unable to be consoled, distracted, or reassured

soziodemographischen oder demenzbezogenen Daten berichtet werden. Die Pflegenden erhielten eine Schulung auf das Instrument und ein Merkblatt, auf dem die potenziellen Schmerzindikatoren eingehend beschrieben waren. Die Beobachtungsintervalle betrugen jeweils 5 Minuten. Die Skala muss für die gepoolte Gesamtstichprobe in allen unterschiedenen Erfassungssituationen als wenig homogen gelten (Cronbach's α zwischen .50 und .65). Dementsprechend konnte die Annahme der Eindimensionalität der Skala durch eine Hauptkomponentenanalyse nur eingeschränkt bestätigt werden. Neben einem ersten Faktor (Eigenwert 2,51), der ungefähr die Hälfte der Variabilität in den PAINAD-Scores an sich binden konnte, wurde ein weiterer Faktor mit einem Eigenwert über 1 und 20,6 Prozent Varianzaufklärung berichtet. Die Inter-Rater-Reliabilität für die parallelen Einschätzungen der ersten Teilstichprobe (n=19) betrug in der angenehmen Situation .97, in der potenziell negativen Pflegesituation .82. Die Korrelationen der PAINAD mit der DS-DAT, einer visuell-analogen schmerzbezogenen Einschätzung der Pflegenden und einer entsprechenden auf Missempfinden (discomfort) bezogenen VAS-Beurteilung wurden für die Ruhesituation mit .75, .76 und .76 berichtet. In den stärker durch Aktivität charakterisierten positiv und negativ-valenten Situationen wurden Zusammenhänge zwischen der VAS-Schmerzeinschätzung und dem PAINAD-Score von .82 und .95 gefunden. Für die unterschiedenen Beobachtungssituationen werden mittlere Skalenscores von 1,0$\pm$1,3 (angenehme Aktivität), 1,3$\pm$1,3 (Ruhe) und 3,1$\pm$1,7 Punkten für die potenziell stärker mit Schmerzen verbundene Pflegesituation berichtet, was als Hinweis auf die diskriminante Validität des Instrumentes gewertet werden kann. Im zweiten Studienteil (n=25) konnten wie erwartet 30 Minuten nach einer Schmerzmittelgabe signifikant reduzierte mittlere Skalenscores (von 6,7$\pm$1,8 auf 1,8$\pm$2,2 Punkte; t=9,6, df=24, p$<$.001) gefunden werden. Die hohe anfängliche Schmerzbelastung dieses Klientels scheint jedoch kaum mit derjenigen der Demenzkrankengruppe vergleichbar zu sein, wodurch sich die berichteten Hin-

weise zur Änderungssensitivität des Instrumentes nicht ohne weiteres auf die eigentliche Zielpopulation verallgemeinern lassen.

Eine italienische Übersetzung der PAINAD-Skala wurde von der Arbeitsgruppe um Costardi (Costardi et al., 2007) entwickelt und getestet. Die Stichprobe umfasste 20 Altenheimbewohner mit einem mittleren MMST-Wert von 13,4±6,8 Punkten. Verhaltensbeobachtungen wurden von einem speziell geschulten Beurteiler (expert rater) zu Beginn der Studie und nach 15 Tagen wiederholt eingesetzt. Zur Abschätzung der Inter-Rater-Reliabilität erfolgte eine unabhängige PAINAD-Einschätzung durch einen zweiten Beobachter am ersten Tag der Erhebungen. Die Skalenkonsistenz wird mit Cronbach's α=.74 angegeben. Die Inter-Rater-Reliabilität betrug .87. Die Test-Retest-Reliabilität wird mit .88 berichtet. Die Übereinstimmung mit der auf einer Verbalen Rating Skala (VRS) durch den Bewohner eingeschätzten Schmerzintensität betrug .65. Aufgrund der knappen Darstellung der Autoren bleiben jedoch einige Fragen zur Datengrundlage für diese Kennwerte, beispielsweise die Rate nicht-kommunikativer Bewohner, offen.

Eine weitere Studie von Leong, Chong und Gibson (2006) überprüfte retrospektive Schmerzeinschätzungen durch Pflegende anhand der PAINAD-Skala (letzte Woche) für 88 mittelgradig und schwer demenziell beeinträchtigte Altenheimbewohner in Singapur. Daneben wurden auch retrospektive Selbstauskünfte (Schmerz in der letzten Woche) von den Bewohnern auf einer 4-stufigen Ratingskala (Verbal Descriptor Scale, VDS) und analoge Fremdeinschätzungen durch die Pflegenden erfasst. Die Zusammenhänge der PAINAD-Einschätzungen korrelierten mit .84 deutlich höher mit der retrospektiven Intensitätseinschätzung der Pflegenden als mit der Selbstauskunft der Bewohner (.30). Die signifikante Korrelation von .29 der PAINAD mit der *Cornell Scale for Depression in Dementia (CSDD)* wird als Bestätigung der divergenten Validität der Schmerzbeobachtung interpretiert. Die Autoren schlagen im Abgleich mit der durch die Pflegenden eingeschätzten Ratingskala eine Kategorisierung der erhobenen PAINAD-Scores in die Intervalle 0-1 Punkte (=kein Schmerz), 2-3 Punkte (=geringer Schmerz) und 4+ Punkte (=mittlerer und stärkerer Schmerz) vor. Sicherlich kann der Vorteil einer solchen Interpretation der PAINAD-Skala in Anbetracht des im Vergleich zur VDS höheren Bearbeitungsaufwandes nur in einer gesteigerten Präzision der retrospektiven Urteile durch die Vorgabe konkreter Schmerzindikatoren liegen. Eine solche Präzisionssteigerung konnte jedoch bei der gewählten Anlage der Untersuchung nicht schlüssig nachgewiesen werden. Das Argument der Autoren, dass eine Momentaufnahme akuten Schmerzes durch eine 5-10minütige Verhaltensbeobachtung im Kontext der stationären Altenpflege und der häufig chronischen Schmerzbelastungen dieser Klientel nur bedingt sinnvoll sei, ist prinzipiell nachvollziehbar. Zumindest aber sollte hier versucht werden, den Zusammenhang von wiederholten in-vivo-Verhaltensbeobachtungen und retrospektiven Einschätzungen weiter zu klären.

5.3.4.2 Befunde aus dem deutschen Sprachraum

In den deutschen Sprachraum wurde die PAINAD-Schmerzskala durch den Arbeitskreis *Schmerz im Alter* der Deutschen Gesellschaft zum Studium des Schmerzes (DGSS,

Vorsitz: Prof. Dr. H.D. Basler) eingeführt. Die Übersetzung wurde von den Autoren der Originalskala autorisiert und folgte dem international üblichen Prozedere von Übersetzung, Rückübersetzung und Abgleich durch einen Muttersprachler. Die deutsche Fassung wurde als Skala zur *Beurteilung von Schmerzen bei Demenz* (BESD) bezeichnet (Basler et al., 2006). Im Rahmen der Kooperation des Arbeitskreises mit dem Institut für Gerontologie an der Universität Heidelberg konnte die PAINAD-Skala in dieser deutschen Adaptation erstmalig im Kontext der stationären Versorgung Demenzkranker eingesetzt und hinsichtlich ihrer psychometrischen Gütekriterien untersucht werden (Becker et al., 2005).

Zeitgleich wurde das Instrument in der Akutgeriatrie auf seine Fähigkeit getestet, die Wirksamkeit einer medikamentösen Therapie durch einen reduzierten schmerzbezogenen Verhaltensausdruck abzubilden (Basler et al., 2006). Dazu wurden 12 verbal nicht mehr auskunftsfähige, multimorbide Demenzpatienten durch geschulte Pflegende während verschiedener Routinepflegetätigkeiten, die eine Mobilisierung der Patienten erforderten, für jeweils 2 Minuten beobachtet. Unmittelbar im Anschluss an diese Situationen wurden die entsprechenden BESD-Dokumentationsbögen ausgefüllt. Bei mehr als 5 Punkten im Gesamtscore wurden die Patienten dem WHO-Stufenschema entsprechend analgetisch behandelt. Die Wiederholungsmessungen nach 2 (T2) und 24 (T3) Stunden wurden durch dieselben Pflegenden in denselben Pflegesituationen durchgeführt, die für den Bewohner zuvor als schmerzhaft eingeschätzt wurden. Die mittleren BESD-Scores von 7,5$\pm$1,6, 4,7$\pm$1,9 und 5,2$\pm$2,6 Punkten für die Ersterfassung und die beiden Folgezeitpunkte sprechen für die Konstruktvalidität der Skala. Bei 5 der 12 Patienten wurde die Medikation zwischen T2 und T3 ausgesetzt, woraufhin eine erneute Zunahme schmerzbezogenen Ausdrucksverhaltens beobachtet werden konnte (T1: 8,4$\pm$2,1, T2: 4,8$\pm$1,6, T3: 7,0$\pm$2,8), während bei fortgeführter Medikation nochmals leicht reduzierte BESD-Scores ermittelt wurden.

Auch zu den in der ersten empirischen Untersuchungsphase des HILDE-Projektes erfassten Schmerzdaten liegen mittlerweile veröffentlichte Befunde vor (Basler et al., 2006; Becker et al., 2005; Schuler et al., 2007). Eine detaillierte Analyse schmerzbezogener Erfassungsinhalte des HILDE-Instrumentes in seiner ersten Fassung wurde als interner Arbeitsbericht für die kooperierenden Kollegen des DGSS-Arbeitskreises vom Autor dieser Arbeit mitverfasst. Da sich die vorliegende Arbeit im Wesentlichen auf die Schmerzerfassung der zweiten empirischen HILDE-Untersuchungsphase konzentriert, sollen die wichtigsten in der ersten Projektphase erarbeiteten Ergebnisse gewissermaßen als Vorarbeiten verstanden und bereits an dieser Stelle als Teil des gegenwärtigen Kenntnisstandes zu den psychometrischen Eigenschaften der BESD-Skala dargestellt werden.

Schmerzeinschätzung durch Angehörige

Insgesamt konnte mit Angehörigen von 105 Bewohnern zumindest telefonisch Kontakt aufgenommen werden. Diesen Angehörigen wurde der BESD-Erfassungsbogen mit der Bitte, bei den nächsten Besuchen verstärkt auf potenzielle Schmerzen der Bewohner

zu achten und dem Projekt diese schmerzbezogenen Informationen rückzumelden, postalisch zugesandt. Insgesamt entsprachen 56 mit den Bewohnern verwandte oder nahe stehende Personen dieser Bitte (53,3% Rücklaufquote). Aus den Angaben der Angehörigen wurde ersichtlich, dass einige Bewohner aufgrund von Mobilitätseinschränkungen (Bettlägerigkeit oder Angewiesensein auf Rollstuhl) nicht in einer Aktivitätssituation im Sinne der BESD-Definition beobachtet werden konnten (18 vs. 2 unbearbeitete Bögen in Aktivität und Ruhesituation). Für eine möglichst gute Vergleichbarkeit mit bisherigen psychometrischen Befunden zur PAINAD-Skala, wurden die 24 Einzelitems der 5 übergeordneten BESD-Kategorien Atmung, Lautäußerung, Gesichtsausdruck, Körpersprache und Trost der Originalversion entsprechend gewichtet (Intensitätswert jwls. 0-2 Punkte) und anschließend über alle 5 Kategorien hinweg zu einem Gesamtscore aggregiert (möglicher Wertebereich 0-10 Punkte). Die Summe schmerzbezogenen Ausdrucks liegt für die Ruhesituation mit 3,7$\pm$2,5 Punkten geringfügig, nicht aber statistisch bedeutsam über derjenigen der Aktivitätssituation (3,4$\pm$2,2 Punkte). Die Skalenkonsistenz beträgt für die Ruhesituation α=.70, in der Aktivitätssituation lediglich .62. Von allen fünf Indikatorbereichen erscheint insbesondere die Möglichkeit, den Bewohner zu trösten, als vom Rest der Skala deutlich unterscheidbare Qualität. Als Maß transsituationaler Konstanz, ergibt sich für die 36 in beiden Bedingungen beobachteten Bewohner ein mittlerer positiver Zusammenhang beider Schmerzeinschätzungen (r_{ss} =.45). Für Bewohner, die nach Meinung der Angehörigen während ihres Besuches unter akuten Schmerzen litten, wurden signifikant höhere Niveaus schmerzbezogenen Ausdrucksverhaltens sowohl in der Ruhe- als auch der Aktivitätssituation berichtet als bei schmerzfreien Bewohnern (Ruhe: 3,0$\pm$2,4 vs. 4,5$\pm$2,3; Aktivität: 2,5$\pm$2,0 vs. 4,1$\pm$2,2 Punkte). Nicht differenzieren konnte der BESD-Score jedoch zwischen Bewohnern mit und ohne chronischem Schmerzleiden vor der demenziellen Erkrankung. Die Korrelationen der BESD-Summen mit den Fremdratings der Schmerzintensität betrugen für die schmerzbelasteten Bewohner in der Ruhesituation .46 ($p<.024$), jedoch lediglich .13 ($p<.585$) in der Aktivitätssituation.

Schmerzeinschätzung durch Pflegende

Insgesamt konnten Einschätzungen der Schmerzbelastetheit aus der Sicht des Pflegepersonals für 99 Bewohner gewonnen werden. Während aktuelle Schmerzen (Vorliegen, Intensiät und Lokalisation) der Bewohner im Pflegeinterview durch in der Regel zwei Pflegende gemeinsam eingeschätzt wurden, sollten die BESD-Beobachtungen der Pflegekräfte möglichst unabhängig voneinander in üblichen Situationen des Bewohnerkontaktes dokumentiert werden. Wie zuvor für die Angehörigendaten berichtet, wurden die als Einzelitems erfassten Schmerzindikatoren auch hier im Sinne der Originalskala gewichtet und zu einem Gesamtscore verrechnet.

Die Übereinstimmung der BESD-Werte aus den beiden Beobachtungen unterschiedlicher Pflegender betrug r_{ii}=.82 am Morgen und .72 am Abend. Eine Interpretation dieser Kennwerte als Objektivitätskriterium für die BESD-Einschätzungen bleibt jedoch durch die unterschiedlichen Beobachtungssituationen beschränkt. Für die weiteren Analysen

wurden beide Pflegekrafteinschätzungen auf der Ebene der kategoriellen Indikatorbereiche durch Mittelwertsbildung miteinander kombiniert. Das durch die Pflegenden beobachtete Niveau schmerzbezogenen Ausdrucksverhaltens in zufällig herausgegriffenen Situationen am Morgen und abends erwies sich als miteinander vergleichbar (morgens: 3,1±3,1; abends: 3,3±3,1 Punkte). Für die 76 Personen, die sowohl morgens als auch abends von (mindestens) einer Pflegefachkraft eingeschätzt wurden, lässt sich für den BESD-Gesamtscore eine Situationskonstanz von r_{ss}=.94 errechnen. Auch die Skalenkonsistenzen sind mit α=.86 und .85 zu beiden Tageszeiten ähnlich zufriedenstellend. Für die 33 Bewohnerinnen bzw. Bewohner, die nach Meinung der Pflegekräfte unter akuten Schmerzen litten, wurden nahezu doppelt so hohe Niveaus schmerzbezogenen Ausdrucksverhaltens sowohl morgens als auch abends berichtet wie für schmerzfrei eingeschätzte Bewohner (morgens: 4,4±3,0 vs. 2,2±2,8; abends: 4,5±3,1 vs. 2,3±2,7 Punkte). Dennoch korrelieren die BESD-Werte weder morgens noch abends mit den Ratings für die Intensität aktueller Schmerzen und differenzieren nicht zwischen Bewohnern mit und ohne chronischen Schmerzen.

Tabelle 8: Psychometrische Befunde zur Schmerzbeobachtung der ersten HILDE-Feldphase

	Angehörige	*Pflegende*	*Projektmitarbeiter*	
			T1	T2
Reliabilität				
Inter-Rater-Reliabilität	–	morgens: .82 abends: .72	Ruhe: .75 Aktivität: .82	.77 .72
Interne Konsistenz	Ruhe: .70 Aktivität: .62	morgens: .86 abends: .85	Ruhe: .73 Aktivität: .53	.62 .59
Situations-Spezifität	.45	.94	.84	.73
Retest-Reliabilität	–	–	Ruhe: .60 Aktivität: .76	
Validität				
Akute Schmerzbelastung[1]	ja	ja	–	
VDS-Fremdrating	Ruhe: .46 Aktivität: .13	morgens: -.01 abends: -.05	–	
Chronische Schmerzen[1]	nein	nein	–	

[1] Möglichkeit, anhand des BESD-Scores als nicht (chronisch) schmerzbelastet eingeschätzte Bewohner von solchen Bewohnern mit Schmerzen zu differenzieren.

Schmerzeinschätzung durch Projektmitarbeiter

Insgesamt 99 Bewohnerinnen und Bewohner wurden im Abstand von ca. 2 bis 4 Wochen (in Abhängigkeit vom gesundheitlichen Zustand der Betreffenden) von zwei wissenschaftlichen Mitarbeiterinnen des HILDE-Projekts gleichzeitig interviewt. Zu jedem

Beobachtungszeitpunkt schätzten die beiden Projektmitarbeiter in den Rollen von Interviewerin und Kamerafrau das schmerzbezogene Ausdrucksverhalten der Bewohnerinnen und Bewohner jeweils anhand der BESD-Skala in ihrer 5-Item-Originalform ein. Auch hier wurden eine Ruhe- und eine Aktivitätssituation beobachtet. Anders als in den Angehörigen- und Pflegeinterviews schätzten die Projektmitarbeiter weder das Vorliegen von chronischen oder akuter Schmerzen, noch deren Intensität oder dominante Schmerzregionen ein. Die Übereinstimmung der beiden parallelen BESD-Beobachtungen betrug für den ersten Besuch r_{ii}=.75 in der Ruhe- und .82 in der Aktivitätssituation, für den zweiten Untersuchungstermin wurden Inter-Rater-Reliabilitäten von .77 und .72 berechnet. Um den unsystematischen Fehleranteil möglichst gering zu halten, wurden die Urteile beider Projektmitarbeiter analog zu den parallelen Pflegekrafteinschätzungen für jede Situation gemittelt.

Im Bewohnerinterview werden im Vergleich zu den Beobachtungen von Pflegekräften und Angehörigen in beiden Untersuchungsbedingungen deutlich geringere Niveaus von schmerzbezogenem Ausdrucksverhalten dokumentiert. Zum ersten Interviewtermin wurden BESD-Scores von 0,4$\pm$1,0 Punkten in Ruhe und 0,4$\pm$0,8 Punkten bei Aktivität ermittelt, und auch zum zweiten Interview wurden vergleichsweise geringe mittlere Schmerzausprägungen dokumentiert (Ruhe: 0,4$\pm$0,8; Aktivität: 0,4$\pm$0,8 Punkte). Eine mögliche Erklärung hierfür könnten strengere Interpretationskriterien der wissenschaftlichen Mitarbeiter, z.B. bezüglich des Beobachtungszeitraumes von jeweils 2 Minuten oder im Hinblick auf die Eindeutigkeit der gezeigten diskreten Verhaltenseinheiten, sein. Eine ebenfalls naheliegende Erklärung für diese systematisch erscheinende Abweichung liegt im kategoriellen Erhebungsformat selbst begründet, bei dem durch die implizite Intensitätsgewichtung bestimmte Verhaltensbeobachtungen vielleicht zurückhaltender interpretiert werden als bei einer Einzelvorgabe ohne diesen Referenzrahmen. Der korrelative Zusammenhang der BESD-Summen für Ruhe- und Aktivitätssituation beträgt für die kombinierten Urteile von Interview- und Kameraführenden r_{ss}=.84 zum ersten und .73 zum zweiten Interviewzeitpunkt. Die Retest-Reliabilitäten ergeben sich zu r_{tt}=.60 für die Ruhesituation und .76 für die Einschätzungen in der Aktivitätsbedingung. Mit Cronbach's α-Werten von .73 (Ruhe) und .53 (Aktivität) zum ersten und .62 bzw. .59 zum zweiten Messzeitpunkt müssen die interne Konsistenz der Skala (Originalform) bei einer Bearbeitung durch Projektmitarbeiter als wenig befriedigend und zeitlich nicht stabil gelten.

Vor dem Hintergrund der bisher veröffentlichten internationalen Arbeiten zur PAINAD und den Erfahrungen aus der eigenen Projektarbeit mit der deutschen BESD wurden – trotz der heterogenen Untersuchungsdesigns und Befundlage – die folgenden Aspekte als Anforderungen für weitere Forschungen identifiziert.

Die berichteten Konsistenzkoeffizienten erscheinen insgesamt zu gering für eine Skala, die vorgibt, ein einziges Merkmal abzubilden. Die Annahme, der aus den intensitätsgewichteten Einzelindikatoren berechnete Skalenscore bilde tatsächlich Teile eines Kontinuums verschieden intensiven Schmerzerlebens ab, konnte zumindest auf der Grundlage des Vergleiches mit selbst- und fremdeingeschätzten Intensitätsskalen bislang nicht einheitlich

bestätigt werden. Damit bliebe natürlich auch die Angemessenheit der internen Strukturierung der Skala zu überprüfen. Zu diesem Zwecke erschien es notwendig, zukünftig alle einzubeziehenden Verhaltensindikatoren dichotom zu dokumentieren.

Es wurde vorgeschlagen, einzelne Indikatoren und Ausdrucksbereiche umzuformulieren und zu präzisieren. Der Begriff der *Cheyne Stoke Atmung* erwies sich in den Voruntersuchungen als für die Pflegenden schwer verständlich, und wurde für die zweite Erfassungsphase durch die umfassendere Beschreibung *tiefer werdende und wieder abflachende Atemzüge mit Atempausen* ersetzt.

Die beiden Items, die sich auf den Erfolg des *Tröstens* beziehen, wurden ebenfalls umformuliert. Die Logik der ursprünglichen Items beurteilte solche Schmerzen, die man nicht durch Trösten lindern kann, als stärker als solche, bei denen der Versuch des Tröstens Erfolg hat. Damit wird hier implizit ein aktives Eingreifen des Beobachters in die Beurteilungssituation gefordert, das zusätzlich noch implizit reaktiv auf die beobachtete Situation bezogen ist (d.h. sind die Schmerzäußerungen deutlich genug, als dass aktives Trösten überhaupt angezeigt erscheint?). Um diese Konnotationen zumindest deutlich zu machen, wurde in der zweiten Erhebungsphase zunächst gefragt, ob die beobachtende Pflegeperson den Wunsch verspürt, den beobachteten Bewohner zu trösten. Nur in diesem speziellen Fall soll tatsächlich aktiv tröstend in die Beobachtungssituation eingegriffen und der Erfolg des Trostspendens eingeschätzt werden. Das neuformulierte Item „Ist das auffällige Verhalten durch Stimme oder Berührung abzulenken oder zu beruhigen?“ muss selbstverständlich vor der Aggregation in einen Skalenscore entsprechend gespiegelt werden. Auch wenn dadurch die implizite Filterführung expliziert wurde, bleiben die konzeptionellen und methodischen Probleme dieses Erfassungsbereiches aber prinzipiell bestehen.

Als weitgehend ungeklärt wurde auch der Einfluss situativer Charakteristika auf die Schmerzbeobachtung eingeschätzt. In den bisherigen Untersuchungen wurden keine höheren Niveaus schmerzbezogenen Ausdrucksverhaltens in Aktivitäts- gegenüber Ruhesituationen beobachtet. Deutlich wurden dagegen die praktischen Einschränkungen, die sich für die Anwendbarkeit einer Schmerzbeobachtung dann ergeben, wenn die Instruktionen ein Ausmaß von Mobilität fordern, das vom Bewohner nicht mehr erbracht werden kann. Es bleibt zu fragen, inwiefern sich ein solches Instrument dann noch für den Einsatz im normalen Pflegealltag (v.a. außerhalb von mobilisierenden Pflegehandlungen) demenzkranker Menschen eignet.

5.3.5 Deutsche Adaptation der *Checklist of Nonverbal Pain Indicators* (CNPI)

Die im Original 2000 von Feldt vorgestellte *Checklist of Nonverbal Pain Indicators* (CNPI) wurde auf der Grundlage der *University of Alabama Birmingham Pain Behavior Scale* (UAB-PBS; Richards, Nepomuceno, Riles & Suer, 1982) zur Erfassung von postoperativem Schmerz bei älteren Menschen entwickelt. Solche Items der UAB-PBS, die von mobilen, verbal nicht beeinträchtigten Patienten ausgingen (z.B. Gehen, Körperhaltung beim Stehen, wegen Schmerzen im Bett verbrachte Tageszeit, Bitten um Schmerzmedikamente oder Ruhelosigkeit) wurden nicht in das neue Instrument übernommen oder für den

Einsatz in der potenziell verbal und bezüglich der Mobilität eingeschränkten Population kognitiv beeinträchtigter Menschen entsprechend angepasst. Zusätzlich wurden verbale Schmerzäußerungen, wie beispielsweise Klagen, Fluchen oder Protestrufe, die von Raway (1994) als Indikatoren postoperativen Schmerzes bei kognitiv beeinträchtigten Älteren identifiziert wurden, in das neue Instrument eingebunden. Die CNPI umfasst damit die sechs Ausdrucksbereiche (1) vokaler Schmerzausdruck, (2) Gesichtsgrimassen, (3) verbale Schmerzäußerungen, (4) Reiben, (5) (An-)Klammern und (6) Körperhaltung, die nur zum Teil durch mehrere konkrete Verhaltensmerkmale beschrieben werden. Im Gegensatz zur BESD sieht die CNPI-Skala keine Intensitätsabstufung innerhalb der Indikatorbereiche vor. So trägt beispielsweise die Mimik, unabhängig von der Art oder Anzahl verschiedenen beobachteten mimischen Schmerzausdruckes, mit maximal einem Punkt zum Gesamtscore der Originalskala bei, womit sich ein Wertebereich von 0 bis 6 Punkten ergibt (s. Abbildung 26).

Abbildung 26: Originalskala *Checklist of Nonverbal Pain Indicators* (CNPI)

(Write a 0 if the behavior was not observed, and a 1 if the behavior occurred even briefly during activity or rest.)		
	Movement	Rest
1. Vocal complaints: Non-verbal (Expression of pain, not in words, moans, groans, grunts, cries, gasps, sighs)		
2. Facial grimaces/Winces (Furrowed brow, narrowed eyes, tightened lips, jaw drop, clenched teeth, distorted expressions)		
3. Bracing (Clutching or holding onto side rails, bed, tray table, or affected area during movement)		
4. Restlessness (Constant or intermittent shifting of position, rocking, intermittent or constant hand motions, inability to keep still)		
5. Rubbing (Massaging affected area)		
6. Vocal complaints: Verbal (In addition, record verbal complaints.) (Words expressing discomfort or pain, “ouch”, “that hurts”, cursing during movement, or exclamations of protest (e.g., stop, that’s enough))		
Subtotal Scores		
Total Score		

Die Pilottestung umfasste 88 ältere Patienten (83,2±7,7 Jahre) mit unterschiedlicher kognitiver Beeinträchtigung (53 kognitive beeinträchtigte Patienten mit MMST-Werten unter 24 Punkten), die nach einer Hüftfraktur im Krankenhaus behandelt wurden. Die Beobachtungssituationen umfassten dabei jeweils eine Ruhesituation im Krankenzimmer und eine Transfersituation vom Bett zum Stuhl oder vom Stuhl zum Bett (unterstützt durch eine Pflegekraft). Genauere Angaben zum zeitlichen Umfang der Beobachterschulungen und der Verhaltensbeobachtungen selbst werden nicht gemacht. In der Ruhesituation zeigten über die Hälfte der Patienten (56%) keine der berücksichtigten schmerzbezogenen Verhaltensweisen, während die Beobachtungsraten bei Bewegung deutlich höher

waren. Hier konnte für nahezu zwei Drittel der Patienten (62%) mindestens ein behavioraler Schmerzindikator beobachtet werden. Insgesamt wurden in beiden Beobachtungssituationen jedoch nur vergleichsweise geringe mittlere Skalenwerte (0,7$\pm$1,1 Punkte in Ruhe und 1,3$\pm$1,3 Punkte bei Transfer) erreicht. Die interne Konsistenz der Skala wird mit KR-20=.54 und .64 für Ruhe- und Transfersituation berichtet. Auch wenn die Höhe dieser Koeffizienten bekanntermaßen durch die Anzahl der Einzelitems beeinflusst wird, muss der Vorschlag der Autorin, durch die Hinzunahme weiterer Items die Skalenkonsistenz zu erhöhen, kritisch bewertet werden. Schließlich sollte eine Einbeziehung weiterer Indikatoren vorrangig an inhaltlichen Überlegungen orientiert sein und nicht bloß die statistische Unschärfe entsprechender Kennwerte ausnutzen. Die Inter-Rater-Reliabilität der Skala kann mit einer mittleren Übereinstimung von 93 Prozent hinsichtlich der dichotomen Indikatorbereiche als gut bewertet werden. Allerdings wird der Geltungsanspruch dieser Überprüfung dadurch eingeschränkt, dass insgesamt nur 12 Patienten durch zwei Rater eingeschätzt wurden und ein Teil der Indikatoren in diesen parallelen Einschätzungen überhaupt nicht beobachtet werden konnten. Die Korrelationen zwischen den Gesamtscores für die Schmerzbeobachtung und der durch die auskunftsfähigen Patienten eingeschätzten Schmerzintensität betrugen r_S=.37 für den Ruhe-CNPI und r_S=.43 für die Beobachtung der Transfersituation. Aufgrund der geringen Beobachtungsraten der Indikatoren in der Ruhesituation empfiehlt die Autorin einen kombinierten Einsatz des CNPI sowohl in Ruhe, als auch bei potenziell schmerzverbundener Aktivität bzw. Bewegung. Hinweise darauf, wie die CNPI-Scores für beide Beobachtungbedingungen zueinander in Beziehung gesetzt werden sollten, oder hinsichtlich möglicher Cut-off-Werte für die Ableitung konkreter Interventionen werden jedoch nicht gegeben.

Eine Überprüfung der Einsatzmöglichkeiten des CNPI im Bereich der stationären Langzeitpflege wurde kürzlich von einer norwegischen Forschungsgruppe geleistet (Nygaard & Jarland, 2006). Um den realen Pflegealltag möglichst angemessen widerzuspiegeln, dokumentierten sowohl Altenpfleger als auch Altenpflegehelfer und angelernte Hilfskräfte das schmerzbezogene Verhalten mithilfe des CNPI. Insgesamt 46 Altenheimbewohner wurden an zwei aufeinander folgenden Tagen im Bett liegend und während der morgendlichen Grundpflege und Mobilisierung durch dieselbe Pflegeperson eingeschätzt. Eine zweite Pflegeperson schätzte unabhängig davon das Schmerzverhalten der Bewohner am dritten Tag ein. Leider werden keine Angaben zur Dauer des Beobachtungsintervalles gemacht. In der Pflegesituation wurde signifikant mehr schmerzbezogener Ausdruck beobachtet als in der Ruhesituation. Die interne Konsistenz der Skala wird als moderat und mit früheren Arbeiten vergleichbar beschrieben. Aufgrund der geringen Beobachtungsraten in der Ruhesituation berichten die Autoren lediglich die psychometrischen Kennwerte der Skala für die Verwendung in der Pflegesituation. Die Test-Retest-Übereinstimmung betrug für die einzelnen Ausdrucksbereiche zwischen 73,9 und 89,1 Prozent. Allerdings sind die berechneten κ-Koeffizienten mit Werten zwischen 0,23 und 0,66 als eher moderat zu beurteilen. Für die Inter-Rater-Übereinstimmung werden vergleichbare Kennwerte berichtet. In ihrer sehr knappen Darstellung scheinen die Autoren jedoch nicht alle beschriebenen erhobenen Informationen zur Bestimmung der Validität

der CNPI-Einschätzungen (wie z.B. die Selbstauskunft der Bewohner) auszunutzen. Die Spearman-Korrelationen zwischen den CNPI-Beobachtungen der Pflegenden und ihren Einschätzungen der Schmerzintensität auf einer visuell-analogen Skala (VAS) betrugen zwischen .69 und .88.

In einer Studie von Cohen-Mansfield und Lipson (2008) zur Nützlichkeit verschiedener Formen der Schmerzerfassung zur Indikation und Erfolgskontrolle medikamentöser Schmerzbehandlung wurden unter anderem auch PAINAD und CNPI mit einbezogen. Alle Beobachtungsinstrumente wurden in einer Ruhe- und einer Aktivitätssituation (Bewegung oder Transfer durch Pflegekraft) und einem Zeitintervall von jeweils 5 Minuten eingesetzt. Alle Beobachtungen wurden durch wissenschaftliche Projektmitarbeiter durchgeführt. Da der Fokus der Studie auf einer Abschätzung der Änderungssensitivität der Einzelinstrumente bei erfolgter Schmerztherapie lag, wurden jedoch lediglich wenige basale Gütekriterien für die einzelnen Beobachtungsinstrumente berichtet. Die Inter-Rater-Reliabilität der CNPI-Skala wird für drei parallele Beurteiler mit einer Übereinstimmungsrate von 94 Prozent und einer Intra-Class-Korrelation (ICC) von .92 als hoch berichtet. Genauere Angaben zur Grundlage der Verrechnung der in beiden Beobachtungssituationen gewonnenen Skalenscores werden jedoch nicht gemacht. Auch die Frage, ab welchem Cut-off-Wert eine Schmerzindikation gegeben ist, wird nicht beantwortet. Keines der berücksichtigten Beobachtungsinstrumente schien in der Lage, eine anzunehmende unterschiedliche Schmerzentwicklung in den Treatment- und Kontrollgruppen abzubilden (Interaktionseffekt Treatment$\times$Zeitpunkt).

Scherder und van Manen (2005) verglichen die durch die Autoren selbst eingeschätzten CNPI-Werte von 20 Altenheimbewohnern mit wahrscheinlich vorliegender Alzheimer Demenz und 17 nicht demenziell erkrankten Heimbewohnern nach einer Bewegungssituation (Gehen). Entgegen ihren Erwartungen konnte für die Gruppe der Bewohner mit AD kein reduziertes Schmerzniveau bestätigt werden, was die Autoren auf die mögliche mangelnde Differenzierungskraft des Instrumentes bei geringeren Schmerzintensitäten zurückführen.

Mit diesen neueren Studien wurde zumindest ein Teil der in den beiden 2006 veröffentlichten Reviews zur schmerzbezogenen Verhaltensbeobachtung von Herr und Kollegen (Herr et al., 2006) und Zwakhalen und Kollegen (Zwakhalen et al., 2006) angesprochenen weiteren Forschungsbedarfe mittlerweile erfüllt. Die Forderung, die Skala um subtilere Schmerzindikatoren, wie beispielsweise Veränderungen im Sozialverhalten, zu erweitern, bleibt jedoch bestehen. Unbeantwortet bleiben bisher auch die Fragen, inwiefern die konkreten Verhaltensbeobachtungen (i.S. der enthaltenen Einzelindikatoren) differenziell zur Schmerzbestimmung beitragen (Skalenkonstruktion) und welche Rolle dabei der körperlichen Aktivierung bzw. Bewegung zukommt.

Um diese Fragen zu adressieren und auch in Bezug auf die CNPI-Indikatoren den größtmöglichen Auflösungsgrad zu erreichen, wurden wie bei der BESD-Skala auch hier die beschriebenen Schmerzindikatoren der sechs Ausdrucksbereiche extrahiert und einzeln als 15 dichotome Verhaltensmerkmale zur Beobachtung vorgegeben.

5.4 Kompetenzmerkmale der Bewohner

Während chronische und akute Schmerzbelastungen ganz allgemein als Vulnerabilitätsfaktoren für die Lebensqualität der Bewohner verstanden werden müssen, bemüht sich das Projekt um eine möglichst differenzierte Abbildung individuell erhaltener Kompetenzen bei verschiedenen Schweregraden oder inhaltlichen Syndromlagerungen der demenziellen Erkrankung.

5.4.1 Demenzsymptomatik

Durch die psychometrische Ausrichtung der vorliegenden Arbeit und den hohen Differenzierungsgrad der direkt schmerzbezogenen Informationen ist eine Konzentration auf die in Frage stehenden Beobachtungsinstrumente sicherlich zu rechtfertigen. Um dennoch weder die Bodenhaftung (i.S. der Besonderheiten des untersuchten Klientels) noch den nötigen Weitblick (i.S. der Diskussion von Schmerzfreiheit als einer Dimension von Lebensqualität bei Demenz) zu verlieren, werden im Folgenden ausgesuchte diagnostische Erfassungsbereiche des Projektes und Zielkriterien einer vom Bewohner selbst erlebten Lebensqualität vorgestellt.

Während der ersten beiden HILDE-Feldphasen wurden im Rahmen eines diagnostischen Gespräches mit dem Bewohner selbst und den betreuenden Mitarbeitern der Pflege eine Reihe unterschiedlich komplexer funktionaler Bewohnerkompetenzen durch einen Gerontopsychiater erfasst. Dem von Lawton (1991) eingebrachten Modell hierarchischer behavioraler Kompetenzen folgend wurde versucht, ein möglichst breites Spektrum von auch bei einer vorliegenden demenziellen Erkrankung noch (teilweise) verfügbaren Verhaltensressourcen für die Ausbildung von Lebensqualität abzubilden.

Die fünf Bereiche Orientierung, Sprache, konstruktive Praxis und Gedächtnis wurden durch das vom *Consortium to Establish a Registry for Alzheimer's Disease* (CERAD; Morris et al., 1989) vorgeschlagene neuropsychologische Demenzsscreening erhoben (zur praktischen Anwendung im deutschen Sprachraum siehe Satzger et al., 2001; Thalmann & Monsch, 1997; Thalmann et al., 1998). In dieser Testbatterie sind die Einzeltests (1) verbale Flüssigkeit, (2) basale Benennungsleistung (*Modified Boston Naming Test*, MBNT), (3) *Mini-Mental State Examination* (MMSE; Folstein et al., 1975), (4) Lernen einer Wortliste, (5) konstuktive Praxis, (6) freies Abrufen der gelernten Wörter aus dem Gedächtnis und (7) Wiedererkennen der Wörter in einer größeren Menge enthalten.

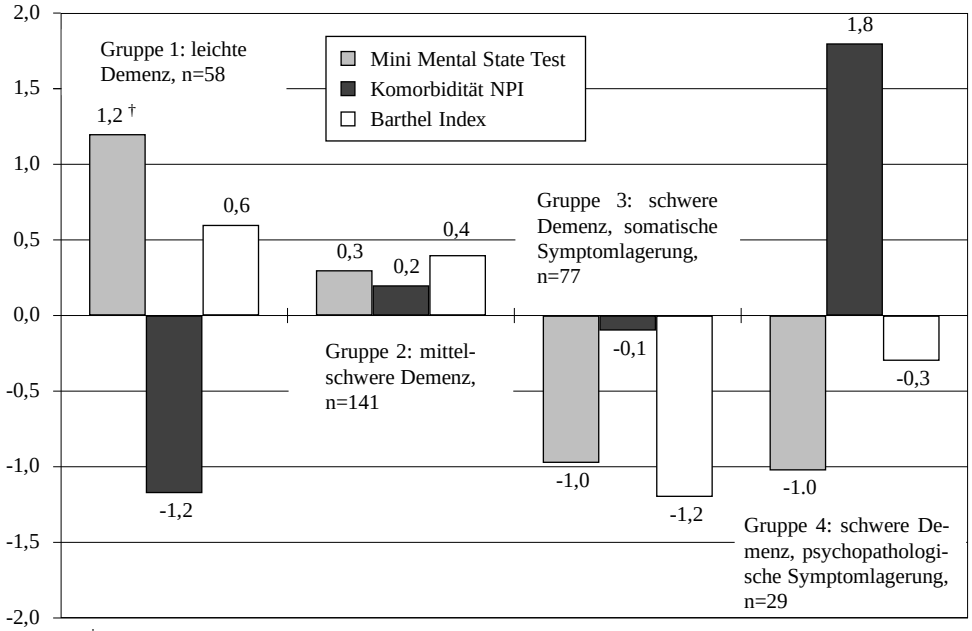
Zusätzlich wurde die nicht-kognitive Demenzsymptomatik anhand des von Cummings und Kollegen (1994) vorgestellten *Neuropsychiatrischen Inventars – Pflegeheimversion* (NPI-NH) im Gespräch mit den Pflegenden abgeklärt. Im Rahmen des späteren Pflegeinterviews schätzten die Pflegenden auch die erhaltene Selbstständigkeit der Bewohner bei der Verrichtung alltäglicher Aufgaben mit dem *Barthel-Index* ein (Mahoney & Barthel, 1965).

Weitere neurologische, psychiatrische und demenzdiagnostische Informationen, die für die vorliegende Arbeit jedoch von nachrangiger Bedeutung sind, wurden mit Hilfe

der *Global Deterioration Scale* (GDS, auch bekannt als *Reisberg-Skala*), dem *Uhrentest*, dem *Bielefelder Autobiografischen Gedächtnisinventar* (BAGI) und der *Apathy Evaluation Scale* (AES) erfasst. Daneben wurden weitere klinische Bewohnermerkmale, wie beispielsweise der Allgemein- und Ernährungszustand oder die wahrscheinliche Demenzätiologie durch den gerontopsychiatrischen Facharzt beurteilt.

Wie bereits bei der Darstellung des Projektverlaufes und der Instrumentenentwicklung beschrieben, wurden für eine alltagsnahe Charakterisierung der sich im Pflegealltag darstellenden Demenzsymptomatik auf der Grundlage von kognitivem Status (MMSE), Alltagskompetenzen (Barthel-Index) und Verhaltensauffälligkeiten (NPI) vier Bewohnergruppen mit unterschiedlichen Mustern erhaltener Kompetenzen identifiziert. Eine ausführliche Beschreibung des methodischen Vorgehens und der Implikationen dieser Differenzierung für die Erfassung und Interpretation der Lebensqualität mit dem HILDE-Instrument wurde bereits an anderer Stelle geleistet (Becker et al., 2006). Einen schematischen Überblick über die relativen Ausprägungen der berücksichtigten Kompetenzen in den vier Bewohnergruppen gibt Abbildung 27.

Abbildung 27: Bestimmungsmerkmale der vier identifizierten Kompetenzgruppen demenzkranker Heimbewohner in HILDE



† Die Skalenwerte sind zur einfacheren Veranschaulichung z-standardisiert angegeben.

Zur Erläuterung der individuellen Charakteristika dieser prototypischen Bewohnergruppen wird im Folgenden auf die Beschreibungen zurückgegriffen, anhand derer die Pflegenden beim Einsatz der aktuellen Instrumentenversion die informativste Referenzgruppe für eine sozial-normative vergleichende Interpretation der beim individuellen Bewohner erfassten Lebensumstände wählen sollen.

Jeder Bewohner ihrer Einrichtung ist in seinen Möglichkeiten, am Leben teilzunehmen, einzigartig. Seine geistigen und körperlichen Fähigkeiten, seine Selbständigkeit in alltäglichen Verrichtungen, aber auch spezifische (z.T. vielleicht auffällige) Verhaltensweisen bestimmen mit, wie der Bewohner sich und seine Umwelt erlebt. Bei der Entwicklung dieses Instruments konnten, trotz dieser großen Individualität der Bewohner, vier verschiedene Bewohnergruppen ausgemacht werden, die sich mit Blick auf ihre *körperlichen Fähigkeiten* und ihre *Selbständigkeit in Alltagsaktivitäten* (Alltagspraktische, funktionale Kompetenzen), ihr *Gedächtnis* und *Denken* (Kognitive Fähigkeiten), sowie in ihrer Belastetheit durch *spezifische Verhaltensauffälligkeiten* (Psychopathologische Symptome, nicht-kognitive Auffälligkeiten) sehr ähnlich sind, und deshalb als vier unterschiedliche Kompetenzgruppen beschrieben werden können:

1. Leicht demenzkranke Bewohner (LD) mit weitgehend erhaltenen alltagspraktischen Kompetenzen bei beginnender Demenz
2. Mittelgradig demenzkranke Bewohner (MD) mit in Teilen erhaltenen alltagspraktischen Kompetenzen bei mittelgradigen kognitiven Einbußen und erkennbaren nicht-kognitiven Symptomen
3. Schwer demenzkranke Bewohner mit somatischen Einschränkungen (SD-S) mit stark eingeschränkten alltagspraktischen Kompetenzen bei schweren kognitiven Einbußen
4. Schwer demenzkranke Bewohner mit psychopathologischen Verhaltensauffälligkeiten (SD-P) mit eingeschränkten alltagspraktischen Kompetenzen bei schweren kognitiven Einbußen und einer Häufung verschiedener nicht-kognitiver Symptome.

Bitte überlegen Sie, welcher dieser vier Kompetenzgruppen Sie den Bewohner Ihrer Einrichtung am ehesten zuordnen würden. Um Ihnen die Zuordnung zu erleichtern, werden diese vier Gruppen nachfolgend genauer beschrieben. Bitte beachten Sie bei Ihrer Beurteilung, dass nicht immer alle Merkmale einer Kompetenzgruppe auf den von Ihnen beschriebenen Bewohner zutreffen müssen. Die beschriebenen typischen Merkmale sollen Ihnen die Einordnung erleichtern, werden aber sicher nie ganz exakt auf den zu beurteilenden Bewohner passen.

Leicht demenzkrank (LD) – Weitgehend erhaltene alltagspraktische Kompetenzen bei beginnender Demenz

Körperliche Fähigkeiten und Selbstständigkeit in Alltagsaktivitäten. Der Bewohner kann viele alltagspraktische Aktivitäten selbständig, d.h. ohne die Hilfe Dritter (z.B. Pflegepersonen oder Angehörige) ausführen. Er kann sich z.B. weitgehend selbständig waschen (z.B. Zähne putzen), sowie weitgehend selbständig essen (Brot streichen, Fleisch schneiden, Besteck benutzen). Der Bewohner ist weitgehend mobil, d.h. er kann selbständig vom Bett oder Stuhl aufstehen oder gehen bzw. sich mit dem Rollstuhl fortbewegen. Der Bewohner benutzt selbständig die Toilette oder einen Nachtstuhl. Er ist in seinen kommunikativen Fähigkeiten (sprachlich, nonverbal) kaum ein-

geschränkt, und falls Verständnis- bzw. Verständigungsschwierigkeiten vorkommen, sind diese häufig auf nicht-demenzielle Ursachen (z.B. Seh- oder Hörbeeinträchtigungen) zurückzuführen.

Gedächtnis und Denken. Der Bewohner findet sich an fremden Orten nicht gut zurecht oder er vergisst früher gut bekannte Namen und/oder kann sich Namen bei der Vorstellung unbekannter Personen schlechter merken. Vielleicht verlegt er manchmal Gegenstände oder hat Probleme bei der Wahl der richtigen Worte. Wird er auf seine Defizite angesprochen oder darauf hingewiesen, werden diese häufig verleugnet. Der Bewohner kann Fakten und Erlebnisse, die seine Person angehen noch relativ gut erinnern z.B. weiß er seinen letzten Wohnsitz oder seinen früheren Beruf und kann einzelne Episoden aus seiner Schulzeit oder auch aus der jüngeren Vergangenheit (letzte fünf Jahre) relativ detailreich erzählen.

Verhaltensauffälligkeiten. Der Bewohner zeigt kaum besondere Verhaltensauffälligkeiten. Am ehesten erscheint er manchmal etwas traurig, niedergeschlagen oder in mutloser bzw. gedrückter Stimmungslage.

Mittelgradig demenzkrank (MD) – In Teilen erhaltene alltagspraktische Kompetenzen bei mittelgradigen kognitiven Einbußen und erkennbaren nicht-kognitiven Symptomen

Körperliche Fähigkeiten und Selbständigkeit in Alltagsaktivitäten. Der Bewohner kann manche Alltagsaktivitäten nur noch mit Unterstützung durch Andere (z.B. mit entsprechenden Anweisungen oder Hilfestellungen) durchführen. Er kann jedoch häufig noch selbständig essen, von einem Stuhl oder dem Bett aufstehen, sich Waschen, sowie die Toilette mit nur wenig Unterstützung benutzen. Die Kommunikation mit dem Bewohner kann aufgrund von Wortfindungs- oder Verständnisschwierigkeiten nur noch eingeschränkt möglich sein.

Gedächtnis und Denken. Der Bewohner hat zunehmend Schwierigkeiten, sich an jüngere Ereignisse zu erinnern. Häufig fällt es ihm schwer, sich zu Zeit (z.B. Datum, Wochentag) oder räumlich (z.B. an bekannten Orten) zu orientieren. Das autobiografische Gedächtnis des Bewohners ist bereits deutlich eingeschränkt, d.h. er ist zwar häufig noch in der Lage, etwa seinen ehemaligen Beruf oder Wohnsitz zu erinnern, jedoch können diese nur selten detailreich und lebendig berichtet werden. Das Wissen um jüngere Ereignisse wie z.B. den Heimeinzug oder Besuche von Verwandten/Bekannten ist bereits deutlich beeinträchtigt.

Verhaltensauffälligkeiten. Der Bewohner fällt insgesamt wenig durch besonderes oder extremes Verhalten auf, trotzdem können nicht selten Zustände gedrückter Stimmung oder Traurigkeit beobachtet werden. Teilweise scheint er auch das Interesse an seiner Umgebung (Personen oder Aktivitäten) verloren zu haben und ihm fehlt die Motivation, etwas (Neues) zu beginnen. Von sich aus beginnt der Bewohner kaum noch ein Gespräch und es ist (zunehmend) schwer, ihn in ein Gespräch zu verwickeln. In anderen Situationen kann der Bewohner auch schon einmal leicht erregbar sein, d.h. er weigert sich vielleicht, sich bei verschiedenen Aktivitäten helfen zu lassen, zu kooperieren oder regt sich über Andere (Mitbewohner, Pflegekräfte, Besuch) auf. Möglich

sind auch spezifische Stereotypien bzw. Angewohnheiten im Verhalten, d.h. er wiederholt bestimmte Aktivitäten immer wieder auf die gleiche Art und Weise und das scheinbar ohne erkennbares Ziel (z.B. Auf und ab gehen, Nesteln an Knöpfen, Kleidung oder Tischdecken, Schubladen oder Schränke aufräumen, Tische abwischen).

Schwer demenzkrank mit somatischen Einschränkungen (SD-S) – Stark eingeschränkte alltagspraktische Kompetenzen bei schweren kognitiven Einbußen und häufig auffälliger Teilnahmslosigkeit

Körperliche Fähigkeiten und Selbständigkeit in Alltagsaktivitäten. Der Bewohner ist in nahezu allen Bereichen der Alltagsaktivitäten auf die Hilfe Anderer angewiesen, so ist er z.B. in seiner Mobilität stark eingeschränkt, benötigt aber auch bei der Körperhygiene oder bei seiner Ernährung Unterstützung. Der Bewohner ist meist sowohl urin-, als auch stuhlinkontinent. Der Bewohner ist meist auch in seinen kommunikativen (sprachlichen und nonverbalen) Fähigkeiten stark eingeschränkt, so dass es in vielen Fällen zu Verständigungsschwierigkeiten kommt. Meistens gehört der Bewohner eher zu den hochaltrigen Bewohnern der Einrichtung.

Gedächtnis und Denken. Der Bewohner kann sich häufig nicht mehr an die Namen seiner Familienangehörigen erinnern und kann auch kurz zurückliegende Ereignisse nicht behalten. Seine Erinnerungen an eigene lebensgeschichtliche Ereignisse sind stark reduziert, er hat das lebendige Wissen über die eigene Biografie praktisch vollständig verloren, so dass auch Fakten aus der eigenen Vergangenheit (z.B. Ort des Schulbesuchs oder Namen von Lehrern oder Mitschülern) nur vereinzelt wiedergegeben werden können.

Verhaltensauffälligkeiten. Der Bewohner fällt insgesamt eher wenig durch besonderes oder extremes Verhalten auf. Wenn solche Symptome auftreten, scheint der Bewohner häufiger das Interesse an seiner Umgebung (Personen und/oder Aktivitäten) verloren zu haben, und wirkt teilnahmslos und wenig vital. Ihn dann in ein Gespräch zu verwickeln fällt schwer. Gelegentlich kann der Bewohner aber auch erregt wirken, d.h. er kann auch mal wütend werden und fluchen. Möglicherweise weigert er sich auch, sich bei verschiedenen Aktivitäten helfen zu lassen bzw. zu kooperieren. Auch über Andere (Mitbewohner, Besuch) regt sich der Bewohner häufiger auf. Seltener lassen sich beim Bewohner auch Traurigkeit, Niedergeschlagenheit oder Mutlosigkeit beobachten.

Schwer demenzkrank mit psychopathologischen Verhaltensauffälligkeiten (SD-P) – Eingeschränkte alltagspraktische Kompetenzen bei schweren kognitiven Einbußen und einer Häufung verschiedener nicht-kognitiver Symptome

Körperliche Fähigkeiten und Selbständigkeit in Alltagsaktivitäten. Der Bewohner kann eine Reihe von alltagspraktischen Tätigkeiten nur noch mit Unterstützung von Anderen (z.B. mit entsprechenden Anweisungen oder Hilfestellungen) durchführen. Er kann jedoch oft noch selbständig essen (z.B. bei entsprechender Vorbereitung der Mahlzeiten), ist noch relativ mobil (kann z.T. noch selbständig gehen) und kann mit

Hilfe auch die Toilette noch benutzen, wobei er häufig (urin-/stuhl-) inkontinent ist. Der Bewohner ist meist auch in seinen kommunikativen (sprachlichen und nonverbalen) Fähigkeiten stark eingeschränkt, so dass es in vielen Fällen zu Verständigungsschwierigkeiten kommt.

Gedächtnis und Denken. Der Bewohner kann sich häufig nicht mehr an die Namen seiner Familienangehörigen erinnern und kann auch kurz zurückliegende Ereignisse nicht behalten. Seine Erinnerung an eigene lebensgeschichtliche Ereignisse ist stark reduziert, er hat das lebendige Wissen über die eigene Biografie praktisch vollständig verloren, so dass auch Fakten aus der eigenen Vergangenheit nur vereinzelt wiedergegeben werden können (z.B. Ort des Schulbesuchs oder Namen von Lehrern oder Mitschülern).

Verhaltensauffälligkeiten. Der Bewohner fällt durch eine Häufung von außergewöhnlichen oder extremen Verhaltensweisen auf. Er ist häufig erregt, ist leicht durch bestimmte Situationen oder Personen aufzubringen und zeigt aggressives Verhalten. Er schreit vielleicht wütend, flucht oder wird gegen andere Personen (z.B. Bewohner, Pflegekräfte) handgreiflich. Er weigert sich häufig, sich bei verschiedenen Aktivitäten helfen zu lassen oder zu kooperieren. Häufig befindet sich der Bewohner aber auch in gedrückter Stimmung und wirkt niedergeschlagen. Der Bewohner scheint daneben in manchen Momenten das Interesse an seiner Umgebung (Personen und/oder Aktivitäten) verloren zu haben und wirkt dann teilnahmslos und wenig vital. Auch abweichendes motorisches Verhalten, z.B. immer wieder in der gleichen Art und Weise wiederholte Bewegungen scheinbar ohne erkennbares Ziel (z.B. Auf und ab gehen, Nesteln an Knöpfen, Kleidung oder Tischdecken, Schubladen oder Schränke aufräumen, Tische abwischen) lässt sich beim Bewohner beobachten. Auch seltene Verhaltensauffälligkeiten wie Halluzinationen, Wahnvorstellungen, aber auch Euphorie und Enthemmung können beim Bewohner vorkommen. Sein Verhalten wird häufig auch von Anderen (z.B. Bewohnern, Besuchern) als störend oder belastend empfunden.

Bitte ordnen Sie den Bewohner jener Kompetenzgruppe zu, die Ihrer Einschätzung nach am ehesten das individuelle Kompetenzprofil des Bewohners wiedergibt. Für die Auswertung des vorliegenden Instruments ist die bestmöglich treffende Einordnung in eine der vier Gruppen sehr wichtig. Sollten Sie daher bezüglich eines Merkmals unsicher sein, klären Sie dies bitte mit Ihrem Team, bevor Sie eine Entscheidung treffen. Denken Sie auch daran, Kollegen und Kolleginnen einzubeziehen, die zu unterschiedlichen Tageszeiten Kontakt mit dem Bewohner haben.

Für die unterschiedenen Demenzgruppen konnten in den bisherigen HILDE-Untersuchungen deutlich verschiedene Lebensqualitäten im Sinne verfügbarer und genutzter Merkmale der sozialen, infrastrukturellen und räumlichen Lebenswelt beschrieben werden. Auch stärker auf das individuelle emotionale Erleben oder die persönliche Beurteilung eigener Lebensumstände bezogene Qualitäten scheinen in den Kompetenzgruppen deutlich verschieden ausgestaltet zu sein (Becker et al., 2006).

Eine besondere Bedeutung für die vorliegende Arbeit gewinnt eine solche holistische Perspektive auf die demenzielle Symptomatik dadurch, dass innerhalb des Bereiches stärkerer kognitiver Beeinträchtigung zwischen Bewohnern mit unterschiedlichem körperlichem Funktionsniveau differenziert wird, für die eine sowohl qualitativ als auch quantitativ verschiedene Schmerzbelastung angenommen werden könnte.

5.4.2 Sprachverständnis und Kommunikationsfähigkeit

Neben der Demenzsymptomatik als solcher erscheinen für die Frage der Schmerzmessung konkrete Beeinträchtigungen im Sprachverständnis sowie der verbalen und non-verbalen Kommunikationsfähigkeiten der Bewohner zentral. Alle drei Aspekte wurden hinsichtlich ihres Schweregrades und der vermuteten Ursache (durch die Demenz oder weitere Beeinträchtigungen bedingt) durch die Pflegenden eingeschätzt.

5.4.3 Selbständige Aktivitäten

Als weitere komplexere personseitige Ressource kann eine sinnvolle Zeitverwendung im Sinne einer Teilnahme an angebotenen Aktivitäten, mehr noch aber im Sinne einer selbständigen Tagesstrukturierung und Beschäftigung gelten. Selbständige, also nicht wesentlich durch Pflegende initiierte und betreute Aktivitäten der Bewohner wurden in der ersten Untersuchungsphase in offener Form erfragt, und nach einer inhaltsanalytischen Kategorisierung der Antworten in der zweiten Untersuchungsphase als Kategorienliste vorgegeben. Die Kategorien umfassen (1) Fernsehen, Radio hören, (2) Bücher oder Zeitschriften lesen, (3) Bilder anschauen (Bildbände, Fotos, Zeitschriften), (4) Hausarbeiten (Ordnung machen, Putzen, Pflanzen pflegen), (5) Kreatives (Hand-)Werken (Malen, Stricken, Basteln), (6) Bewegung (Spazierengehen/-fahren), (7) Spiele spielen (mit anderen oder alleine), (8) Selbstpflege (Kosmetik etc.), sowie (9) Schreiben (Postkarten, Briefe, Tagebuch).

5.5 Durchführung

Im folgenden wird das Prozedere der Erfassung der zuvor beschriebenen schmerzbezogenen Projektinhalte in der zweiten Feldphase des HILDE-Projektes nur knapp beschrieben. Detailreichere Informationen zur Praxiskooperation können dem ausführlichen Abschlussbericht für diese Forschungsphase entnommen werden.

5.5.1 Kontaktaufnahme

Die Heim- und Pflegedienstleitungen der interessierten Einrichtungen wurden gebeten, aufgrund der vorab kommunizierten Einschlusskriterien der HILDE-Untersuchung geeignete Bewohner zu identifizieren und deren prinzipielle Teilnahmebereitschaft abzuklären. Falls für den Bewohner ein gesetzlicher Vertreter benannt war, wurde dieser ebenfalls vorab über die Studie informiert. Materialien zur Studienaufklärung und Einverständniserklärungen wurden den Einrichtungen im Vorfeld der Untersuchung zugesandt.

5.5.2 Schulung der Mitarbeiter

Die zu beteiligenden Pflegenden der Einrichtung erhielten eine gemeinsame In-house-Schulung durch die Projektmitarbeiter. In diesen mindestens zweistündigen Veranstaltungen wurden die bisherige Projektarbeit, die gegenwärtigen Forschungsaufgaben und natürlich die Erhebungsmaterialien vorgestellt. Besonderer Wert wurde dabei auf eine alltagsnahe Vermittlung der theoretischen Grundlagen des Instrumentes gelegt. Die Anleitungen zum Arbeiten mit dem HILDE-Instrument hoben stets die Möglichkeiten und Grenzen der Abbildung der konkreten Lebenswelt eines individuellen Bewohners hervor. Insbesondere dort, wo aus der ersten Untersuchung entwickelte Antwortkategorien vorgegeben wurden (z.B. im Bereich der Alltagsaktivitäten), wurden die für die Kategorienbildung wesentlichen Merkmale erläutert und mit Beispielen veranschaulicht. Die Kriterien für die Verhaltensbeobachtungen zu den Bewohnermerkmalen Schmerz und Emotionalität in verschiedenen Alltagssituationen wurden vorgestellt und diskutiert. Damit sowohl die Pflegenden als auch die Bewohner selbst den größtmöglichen Nutzen aus der Teilnahme am Projekt ziehen können, kam der Diskussion verschiedener Möglichkeiten die eingeschätzten Lebensverhältnisse zu interpretieren und konkrete Maßnahmen zur Förderung der individuellen Lebensqualität der Bewohner abzuleiten große Bedeutung zu. Die Einrichtungen erhielten eine ausreichende Anzahl vorbereiteter gehefteter Erfassungsmappen, um den Beteiligten die fehleranfällige Kopierarbeit zu ersparen.

5.5.3 Basisdiagnostik und Studieneinschluss

Bevor die Pflegenden mit der Bearbeitung der Erhebung der Lebensumstände und Erlebenswelt der Bewohner begannen, wurden die von der Einrichtung zur Teilnahme vorgeschlagenen Bewohner dem gerontopsychiatrischen Projektmitarbeiter vorgestellt, der in einem strukturierten Gespräch die Einschlusskriterien überprüfte und weitere diagnostische Informationen erhob. Zu diesem Vor-Ort-Termin konnten aufgrund von beispielsweise Krankenhausaufenthalten leider nicht immer alle vorab ausgesuchten Bewohner erreicht werden (siehe Stichprobenbeschreibung). In die Studie konnten – unabhängig von der Demenzätiologie und dem Schweregrad der Erkrankung – alle Bewohner aufgenommen werden, sofern nicht zeitgleich eine Substanzabhängigkeit (Alkohol, Medikamente), primäre psychiatrische Erkrankung (z.B. Schizophrenie) oder ein sehr schlechter Allgemeinzustand vorlagen.

5.5.4 Erhebungen durch die Pflegenden

Im Anschluss an die Diagnostik arbeiteten die Pflegenden selbständig mit dem HILDE-Erfassungsmaterial. Die interne Organisation der Erhebungen lag vollständig in den Händen der Heim- bzw. Pflegedienstleitungen. Die Lebensumstände und die erlebte Lebensqualität der einbezogenen Bewohner sollten dabei jeweils durch diejenige Pflegeperson eingeschätzt werden, die den Bewohner am besten kannte bzw. den besten Zugang zum Bewohner hatte (Bezugspflegeperson). Die HILDE-Erfassungen für einen einzelnen Bewoh-

ner wurden von der zugeordneten Pflegeperson eigen- und alleinverantwortlich durchgeführt. Dennoch wurde, falls der Pflegende sich in seiner Beurteilung mancher Lebensbereiche unsicher fühlte, eine Rücksprache mit bzw. Klärung im Pflegeteam empfohlen. Insbesondere an der Interpretation der erfassten Lebensumstände und der Entwicklung von Ideen zur weiteren Förderung der Bewohner konnte das gesamte Pflegeteam sich beteiligen. Die Bearbeitung des Instrumentariums dauerte insgesamt im Schnitt ca. zwei Stunden, wobei sich der erforderliche Zeitaufwand nach Aussage der Pflegenden bei einem mehrfachen Einsatz deutlich reduziert. Während der gesamten Erfassungsphase war das Projektteam für Rückfragen telefonisch oder per e-mail erreichbar.

5.5.5 Abschlussgespräche

Nach ungefähr 8 bis 10 Wochen – dieser Zeitraum richtete sich nach der Anzahl einzubeziehender Bewohner und der Arbeitsbelastung der beteiligten Mitarbeiter – konnten in allen Einrichtungen abschließende Gespräche mit möglichst allen Beteiligten, zumindest aber mit den verantwortlichen Pflegedienstleitungen geführt und die bearbeiteten Unterlagen entgegengenommen werden. Als Anerkennung ihres Engagements erhielt jede Pflegekraft, die mit HILDE gearbeitet hatte, ein Teilnahmezertifikat, das ihren besonderen Einsatz für die Bewohner und das Forschungsprojekt dokumentiert. Die enge Kooperation mit den Einrichtungen wurde darüber hinaus in einigen Fällen durch hausinterne Schulungen zu demenzrelevanten Themen oder durch die Mitarbeit der Einrichtungen im Projektbeirat und die Beteiligung an den vom Projekt ausgerichteten Veranstaltungen fortgeführt.

6 Ergebnisse

Zur Veranschaulichung des Potenzials der zuvor beschriebenen neueren statistischen Verfahren zur Überwindung der gegenwärtigen Probleme bei der Beurteilung der verhaltensbezogenen Schmerzmessung bei Demenz wird auf Daten zurückgegriffen, die im Rahmen der zweiten empirischen HILDE-Projektphase erhoben wurden. Schmerzrelevante Informationen wurden dabei sowohl bereits beim gerontopsychiatrischen Screening der potenziellen Studienteilnehmer durch einen Gerontopsychiater des Projektteams erhoben, als auch im Hauptteil der Studie durch die beteiligten Mitarbeiter der Pflege. Der Schwerpunkt der vorliegenden Arbeit liegt auf der Schmerzbeobachtung durch Pflegemitarbeiter mit den BESD- und CNPI-Instrumenten, weshalb die an diesem Erfassungsteil beteiligten Bewohner die für die vorliegende Arbeit relevante Stichprobe darstellt.

6.1 Stichprobenmerkmale

Die stationäre Langzeitpflege demenzkranker Menschen stellt für Forschende der Sozial- und Verhaltenswissenschaften in vielerlei Hinsicht ein anspruchsvolles Praxisfeld dar. Die Entscheidungsstrukturen für eine Beteiligung an einem wissenschaftlichen Vorhaben können beispielsweise je nach Trägerschaft oder hausinternen Organisation (Heim-

beirat) sehr unterschiedlich sein. Da gewöhnlich keine finanziellen Entschädigungen oder Incentives gegeben werden können, muss der Aufwand einer Beteiligung für die Entscheidungsträger und Mitwirkenden abschätzbar sein und der Mehrwert für die Einrichtungen unmittelbar einsichtig kommuniziert werden. Auch das Engagement der Pflegenden für Schulungen, Befragungen und selbständige Datenerhebungen muss mit einer häufig straff organisierten Pflegeplanung koordiniert werden. Durch die Verteilung der Pflegenden auf Früh-, Spät- und Nachtschichten ist es darüberhinaus schwer, alle am Pflegeprozess beteiligten Mitarbeiter zu einem gemeinsamen Termin zu erreichen. Mit Blick auf die im Zuge der demenziellen Erkrankung häufig gestörte zirkadiane Rhythmik bleiben wichtige Aspekte der Lebenswelt der Bewohner darum nicht selten unbeachtet. Da die Bewohner selbst nur noch zum Teil auskunftsfähig sind, müssen Betreuungspersonen über die Studie aufgeklärt und deren Einverständnis zur Teilnahme eingeholt werden. Daneben sind auch räumliche (Wohnbereiche) oder strukturelle (Pflegeteams) Gegebenheiten bei der Zuteilung zu Treatmentgruppen zu berücksichtigen. Auch muss durch die hohe Vulnerabilität demenzkranker Menschen mit einem schwer abschätzbaren mortalitätsbedingten Verlust von Probanden gerechnet werden. Sowohl die Akquise der Kooperationspartner, als auch die Betreuung der Umsetzung der Untersuchung in den Einrichtungen verlangen daher besondere Flexibilität und Sensibilität des Projektteams.

Die zuvor genannten Umstände sind sicherlich nicht nur in diesem Forschungsfeld anzutreffen, sollten aber bei der Diskussion und Würdigung der Feldarbeit und Datengrundlage berücksichtigt werden. Eine unreflektierte Übertragung von Ansprüchen an Stichprobenumfang oder Randomisierung aus anderen Forschungsfeldern (z.B. Medikamentenstudien oder bevölkerungsrepräsentativen Surveys) birgt die Gefahr, dass das Potenzial neuerer methodischer Ansätze in praxisnahen Feldern nicht ausgeschöpft oder inhaltlich substanzielle Effekte nicht veröffentlicht werden können. Im hier vorgestellten Projekt konnte eine – nicht nur an der inhaltlichen Breite und dem Erfassungsaufwand gemessene – substanzielle Anzahl unterschiedlicher Pflegeeinrichtungen und demenziell erkrankter Bewohner einbezogen werden.

6.1.1 Einrichtungen

Die im Verteiler des gerontologischen Instituts enthaltenen Einrichtungen der stationären Altenhilfe und verschiedene Träger wurden schriftlich über die Studie informiert und um ihre Kooperation gebeten. Mit den Heim- bzw. Pflegedienstleitungen der interessierten Einrichtungen wurden persönliche Gespräche geführt, in denen eingehender über bisherige Projektarbeiten und weitere Planungen informiert wurde. Schließlich konnten 12 Einrichtungen unterschiedlicher Trägerschaft (kirchlich, gemeinnützig und privat) für die Studie gewonnen werden. Etwas über die Hälfte der Einrichtungen liegen im süddeutschen Raum (Rhein-Neckar-Region: 4, Raum Stuttgart: 3), drei im Großraum Köln/Bonn (Nordrhein-Westfalen), eine in Trier (Rheinland-Pfalz) und zwei weitere Einrichtungen konnten aus Weimar in Thüringen gewonnen werden. Insgesamt konnten Bewohner aus 33 verschiedenen Wohnbereichen in die Studie eingeschlossen werden. Vier der 12 Einrich-

tungen wiesen Wohnbereiche mit seggregativer Betreuung demenzkranker Bewohner auf, während der überwiegende Teil der beteiligten Einrichtungen integrative oder teilintegrative Wohnbereiche unterhielt. Die Anzahl der Plätze in den untersuchten Wohnbereichen variierte zwischen 9 und 52 Betten. Die Auslastung der Wohnbereiche (max./belegte Heimplätze) war mit im Mittel 98,3 Prozent durchgängig erwartungsgemäß hoch (min=88,9%).

6.1.2 Mitarbeiter

In den teilnehmenden Einrichtungen konnten jeweils zwischen 3 und 17 Mitarbeiter an der HILDE-Untersuchung beteiligt werden. Insgesamt konnten 107 Pflegende in In-House-Schulungen mit dem Projekt sowie dem Instrument vertraut gemacht und für die Beurteilung der Lebensqualität mit HILDE gewonnen werden. Davon waren 87,9 Prozent Frauen (n=94) und 12,2 Prozent Männer (n=13). Das Lebensalter der Mitarbeiter variierte zwischen 18 und 61 Jahren (M=40,2$\pm$11,9 Jahre). Nur 5 Prozent der einbezogenen Pflegenden waren zur gerontopsychiatrischen Fachkraft weitergebildet. Insgesamt 64,5 Prozent der Pflegenden (n=67) konnten eine dreijährige Ausbildung zur Alten- bzw. Krankenpflegerin vorweisen, wovon eine Pflegende zum Erhebungszeitpunkt gerade die Weiterbildung zur gerontopsychiatrischen Fachkraft durchlief. Knapp 14 Prozent hatten eine einjährige Altenpflegehelfer-Ausbildung absolviert oder wurden gerade ausgebildet. Weitere 11 Prozent der Befragten waren angelernte Hilfskräfte. Weitere 7 Personen schließlich gaben keine für die Altenpflege spezifische Berufsausbildung an. Die Mitarbeiter waren im Durchschnitt seit knapp 6 Jahren in den teilnehmenden Einrichtungen beschäftigt (range: weniger als 1 Jahr bis 24 Jahre). Die am Projekt beteiligten Pflegenden arbeiteten überwiegend in Vollzeit (n=64, 59,8%) oder in Drei-Viertel-Stellen (n=32, 29,9%). Nur 11 der an HILDE beteiligten Mitarbeiter waren halbtags (n=10, 9,4%) oder in geringerem Umfang beschäftigt (n=1, 1%).

Bezüglich der Mitarbeitermerkmale Geschlecht, Lebensalter und Dauer der Tätigkeit in der Einrichtung gab es zwischen den einbezogenen Häusern keine systematischen Unterschiede. Die formale Qualifikation und der Tätigkeitsumfang der kooperierenden Pflegenden erscheinen in einzelnen Einrichtungen dagegen deutlich uneinheitlich. Während die HILDE-Erfassungen in allen Heimen hauptsächlich durch Pflegende mit mindestens einer einjährigen Ausbildung zum Alten- bzw. Krankenpflegehelfer durchgeführt wurden, bezogen einzelne Einrichtungen (trotz der in den Schulungen angesprochenen anders lautenden Empfehlungen des Projektteams) auch angelernte Kräfte (6 Häuser) und weitere Berufsgruppen (5 Häuser) mit in die Untersuchung ein. In einem Drittel der Heime konnten gerontopsychiatrisch weitergebildete Fachkräfte an der Arbeit mit HILDE beteiligt werden. In vier Häusern wurde (fast) ausschließlich auf ganztags arbeitende Pflegende zurückgegriffen, der Anteil der an der Studie beteiligten halbtags und geringfügiger beschäftigten Personen betrug zwischen 0 Prozent (6 Häuser) und knapp 30 Prozent.

6.1.3 Bewohner

Im Rahmen der zweiten HILDE-Erhebungsphase wurden von der Heim- oder Pflegedienstleitung der 12 teilnehmenden Einrichtungen insgesamt 234 Bewohner als mögliche Studienteilnehmer vorgeschlagen. Mit 20 dieser Bewohner konnte jedoch entweder kein diagnostisches Gespräch geführt werden, da die entsprechenden Personen zum Untersuchungszeitpunkt im Krankenhaus waren oder die Teilnahme spontan verweigerten, oder mussten aufgrund von Substanzabhängigkeiten oder schwerwiegenden psychiatrischen Erkrankungen aus der Untersuchung ausgeschlossen werden. Damit konnten insgesamt 214 Bewohner eingehend diagnostisch untersucht werden. Insgesamt 199 dieser Bewohner wurden in den darauf folgenden Tagen von den Mitarbeitern der Einrichtungen beobachtet und mit HILDE erfasst. Eine Übersicht über die wesentlichen soziodemographischen Merkmale und den Gesundheitszustand der in die Studie eingeschlossenen Bewohner gibt Tabelle 9.

Tabelle 9: Charakteristika der Bewohnerstichprobe der zweiten HILDE-Feldphase

N % bzw. N M±SD	*Diagnostisches Interview*		*Pflege-interview*	
Geschlecht				
– Frauen	179	83,6%	168	84,4%
– Männer	35	16,4%	31	15,6%
Alter	212	86,1±6,5	198	86,1±6,6
Heimaufenthaltsdauer (in Jahren)	212	3,7±5,0	198	3,7±5,1
Allgemeinzustand				
– gut	149	69,6%	139	69,9%
– mäßig	50	23,4%	46	23,1%
– schlecht	15	7,0%	14	7,0%
Ernährungszustand				
– adipös	42	19,6%	39	19,6%
– gut	126	58,9%	117	58,8%
– mäßig	36	16,8%	34	17,1%
– schlecht	10	4,7%	9	4,5%
Barthel-Index Gesamt	-	-	199	48,8±28,2
Barthel-Index Mobilität[1]	-	-	198	21,3±13,8
Mini Mental Status Test (MMST)	213	12,3±10,1	198	12,5±10,0
Neuropsychiatrisches Inventar (NPI-NH)	214	11,4±8,7	199	11,3±8,7
NPI-Komorbidität[2]	214	2,0±1,1	199	2,0±1,1
Apathy Evaluation Scale (AES-D)	214	27,0±16,5	199	27,0±16,4

[1] Summe mobilitätsbezogener Items des Barthel-Index: Aufstehen, Gehen, Treppen steigen.
[2] Summe gleichzeitiger Belastungen durch Depression, Apathie, Erregung, abweichendem motorischem Verhalten und Sonstigem (Wahn, Halluzination, etc.), jwls. 0=„nein“, 1=„ja“.
Datenbasis: HILDE2 2006; N=214.

6.1.3.1 Soziodemographie

Erwartungsgemäß war der Anteil einbezogener Männer mit ca. 16 Prozent der Gesamtstichprobe vergleichsweise gering. Die eingeschätzten Bewohner waren zum Untersuchungszeitpunkt zwischen 59 und 103 Jahre, im Durchschnitt etwas über 86 Jahre alt. Die Heimaufenthaltsdauer betrug im Mittel 3,7 Jahre. Dabei wohnte ungefähr jeder vierte untersuchte Bewohner erst seit einem Jahr oder kürzer in der Einrichtung. Insgesamt wohnten knapp 20 Prozent der Bewohner der Stichprobe bereits länger als 5 Jahre in der Einrichtung, wobei fünf Bewohner bereits seit über 20 bis maximal 36 Jahren in der Einrichtung betreut wurden.

6.1.3.2 Körperlicher Zustand und Alltagskompetenz

Der durch den psychiatrischen Facharzt eingeschätzte Allgemeinzustand wurde für knapp 70 Prozent der Bewohner als gut beschrieben. Das klinische Urteil des körperlichen Zustandsbildes fiel dagegen für knapp jeden vierten Bewohner als nur mäßig aus, und insgesamt 15 Bewohner befanden sich zum Zeitpunkt des diagnostischen Interviews in einer als schlecht beurteilten Gesamtverfassung. Trotz der naheliegenden Erwartung höherer Ausfallraten in der Gruppe der Bewohner mit schlechtem Allgemeinzustand konnte für 14 Bewohner auch das Pflegeinterview durchgeführt werden.

Insgesamt wurden für etwas mehr als 20 Prozent der Bewohner Hinweise auf einen mangelhaften oder gar schlechten Ernährungszustand beschrieben. Weitere 20 Prozent der Bewohner wurden als übergewichtig eingeschätzt, sodass insgesamt nur etwas mehr als jeder zweite Bewohner einen guten Ernährungszustand aufwies.

Die Einschätzungen der alltagspraktischen Fähigkeiten anhand des Barthel-Index erfolgte nicht bereits im Rahmen des diagnostischen Interviews, sondern wurde erst bei der Befragung der Mitarbeiter erhoben. Von den bei der Einschätzung von zehn Alltagsverrichtungen zur Selbstpflege maximal zu vergebenden 100 Barthel-Punkten, erreichten die untersuchten Bewohner im Mittel insgesamt knapp 50 Punkte, was den substanziellen Hilfebedarf dieses Klientels unterstreicht. Lediglich zwei Bewohner wurden in ihrer Selbstpflege als vollständig unabhängig von fremder Hilfe beschrieben, während 15 Bewohner bei allen erfragten Tätigkeiten vollständig auf die Unterstützung durch Andere angewiesen waren.

Insbesondere mit Blick auf die im höheren Lebensalter zu erwartenden degenerativen Veränderungen des Haltungs- und Bewegungsapparates und die damit verbundenen Schmerzen bei Bewegung, sollen die drei direkt mobilitätsbezogenen Items Aufstehen von Stuhl oder Bett, Gehen und Treppensteigen des Barthel-Index in dieser Arbeit zusätzlich als separater Indikator selbständiger Lebensführung berücksichtigt werden. Die Gesamtstichprobe der zweiten HILDE-Feldphase erreicht im Mittel etwas über 21 der für dieses Komposit maximal 40 erreichbaren Barthel-Punkte. Dabei erscheint der Unterstützungsbedarf bei mobilitätsbezogenen Aktivitäten mit jeweils ca. 28 Prozent gering (0-5 Punkte) und stark (35-40 Punkte) hilfebedürftiger Bewohnern symmetrisch verteilt.

6.1.3.3 Kognitiver Status und demenzielle Symptomatik

Mit durchschnittlich 12,3 erreichten Punkten im Mini-Mental-Status Test kann die realisierte Stichprobe als erwartungsgemäß in ihrer Denk- und Gedächtnisleistung substanziell beeinträchtigt beschrieben werden. Insgesamt 50 Bewohnern (23,5%) war die Bearbeitung des MMST nicht mehr möglich bzw. erreichten keinen einzigen Punkt, während (trotz des Einschlusskriteriums Demenz) immerhin noch 6 Bewohner (2,8%) die volle Punktzahl erreichten. Legt man als Cut-off-Score für das Vorliegen einer Demenz einen MMST-Wert von unter 24 Punkten zugrunde (O'Briant et al., 2008), lässt sich für insgesamt 81,7 Prozent der eingeschlossenen Bewohner eine krankheitswertige Beeinträchtigung konstatieren.

Der spezifischen Ätiologie der demenziellen Erkrankung kommt im Rahmen dieses Projektes und der vorliegenden Arbeit aufgrund der Orientierung am klinischen und pflegerischen Gesamtbild der Beeinträchtigungen in verschiedenen alltagspraktischen Merkmalsbereichen eine nur nachgeordnete Rolle zu. Dennoch wurden die Bewohner durch den gerontopsychiatrischen Facharzt des Projektteams auch bezüglich der vermutlich vorliegenden Erkrankungsursache bzw. dem Demenztyp klinisch beurteilt. Die Verteilung verschiedener Demenztypen in der realisierten Stichprobe entspricht nach diesem Urteil weitestgehend den auch in anderen Studien berichteten und in der ersten Feldphase des HILDE-Projektes in den kooperierenden Einrichtungen vorgefundenen Prävalenzraten. Mit 150 Bewohnern litt nach dieser Einschätzung der überwiegende Teil (70,1%) der eingeschlossenen Bewohner (wahrscheinlich) an einer Demenz vom Alzheimer-Typ, während für 10 weitere Bewohner (4,7%) eine vaskuläre Demenz, und für sechs weitere Bewohner (2,8%) eine Mischform dieser beiden Demenztypen festgestellt wurden. Seltener wurden Demenz bei Parkinson (N=3, 1,4%) oder sonstige Demenzursachen (z.B. Korsakow oder Morbus Pick; N=9, 4,2%) konstatiert. Eine leichte kognitive Beeinträchtigung (Mild Cognitive Impairment, MCI) wurde bei 22 Bewohnern (10,3%) festgestellt, während 14 an der Untersuchung teilnehmende Bewohner (6,5%) keine Anzeichen für eine Demenz aufwiesen und als gesunde Kontrolle dienen sollten.

Lediglich für 18 (8,4%) der 214 untersuchten Bewohner wurde kein einziges der insgesamt zehn durch das Neuropsychiatrische Inventar erfassten nicht-kognitiven Demenzsymptome berichtet. Da einige der erfragten Verhaltensauffälligkeiten, beispielsweise Wahn, Halluzinationen, Angst, Euphorie, Enthemmung und Reizbarkeit in der ersten HILDE-Erfassungsphase nur sehr selten genannt wurden, wurden diese Symptome in eine gemeinsame NPI-Kategorie (Sonstiges) zusammengefasst. Für die Bewohner wurden im Mittel zwei dieser nunmehr fünf unterschiedenen Verhaltensauffälligkeiten berichtet, wobei am häufigsten Depression/Dysphorie (62,1%) und Erregung/Aggression (47,9%) auftraten. Im Durchschnitt betrug die NPI-Gesamtbelastung der Betroffenen (Summe des Produkts aus Häufigkeit und Schwere der Symptome) 11,4 von maximal 60 möglichen Punkten.

Die Apathie als sozial-emotionale Responsivität der demenzkranken Menschen stellt insbesondere mit Blick auf die Schmerzkommunikation und -erfassung ein bedeutsames

Bewohnermerkmal dar, und erfuhr in diesem Forschungsprojekt durch NPI-NH und AES-D eine besonders eingehende Abbildung. Detaillierte Informationen zu den psychometrischen Eigenschaften der durch das Projektteam ins Deutsche übertragenen Apathy Evaluation Scale wurden bereits an anderer Stelle veröffentlicht (Lueken et al., 2007; Lueken et al., 2006; Seidl et al., 2007). Die einbezogenen Bewohner nutzten den theoretisch möglichen Wertebereich von 0 bis maximal 54 Punkten voll aus, und erreichten im Mittel $27 \pm 16{,}5$ Punkte. Lediglich für 13 Bewohner (6,1%) wurden überhaupt keine Anzeichen eines entsprechenden Motivationsverlustes beschrieben.

Wie bei der Beschreibung der Anlage und Durchführung des Projektes in Kapitel 5.4.1 bereits ausgeführt wurde, kommt einer am Gesamtbild erhaltener Kompetenzen orientierten, und damit gewissermaßen holistischen Perspektive auf die kognitiven und alltagspraktischen Fähigkeiten der Bewohner im Kontext einer alltagsnahen Erfassung und Diskussion von Lebensqualität bei Demenz eine besondere Bedeutung zu. Durch die Identifizierung von vier Prägnanztypen demenzkranker Bewohner in stationären Einrichtungen der Altenhilfe (sog. Kompetenzgruppen) wird dieser Forderung sowohl bei der Erfassung, als auch bei der Interpretation realisierter Lebensqualität bis hin zu der Entwicklung potenzieller Fördermöglichkeiten mit HILDE nachgekommen (vgl. Abbildung 27 auf Seite 178). Aufgrund der verschiedenen Syndromlagerungen in den Gruppen der schwer demenzkranken Bewohner mit massierten Verhaltensauffälligkeiten einerseits und massiven körperlichen Kompetenzeinbußen andererseits kommt dem für das Gesamtprojekt entwickelten gruppenspezifischen Ansatz auch mit Blick auf die Erfassung von Schmerzzuständen Bedeutung zu.

In der Einschätzung der Pflegenden war mit insgesamt knapp 59 Prozent (N=116 Bewohner) der überwiegende Teil der einbezogenen Bewohner in ihrer verbalen Kommunikationsfähigkeit eingeschränkt. Als Ausmaß der Einschränkungen wurden für annähernd 40 Prozent der verbal kommunikationsbeeinträchtigten Bewohner geringe Schwierigkeiten angegeben, während für knapp jeden vierten (24,1%) Demenzkranken massive Kommunikationseinbußen berichtet wurden und für etwas mehr als 5 Prozent der Bewohner überhaupt keine verbale Kommunikation mehr möglich war. Für den weitaus überwiegenden Teil der verbalen Einschränkungen (n=92, 89,3%) wurde die demenzielle Erkrankung als Ursache genannt. In insgesamt 26 Fällen (28,3%) wurde jedoch eine weitere, zusätzliche Ursache für die Beeinträchtigung der verbalen Kommunikation genannt.

Auch die Möglichkeiten, non-verbal mit anderen Menschen zu kommunizieren, wurden für knapp 45 Prozent (N=85) der einbezogenen Bewohner als beeinträchtigt eingeschätzt. Das Ausmaß der jeweiligen Kommunikationseinbußen wurde für wiederum ca. 40 Prozent der Betroffenen als eher leicht bzw. gering angegeben, und auch hier erscheint ungefähr jeder vierte Bewohner (24,1%) schwer beeinträchtigt. Keinerlei Möglichkeiten zu non-verbaler Kommunikation wurden für einen Fall (1,2%) berichtet. Auch für die Beeinträchtigungen der nicht-sprachlichen Ausdrucksfähigkeit machten die Pflegenden größtenteils (N=69, 92%) die demenzielle Erkrankung verantwortlich, wobei für insgesamt ein Drittel der Betroffenen (N=23, 32,9%) alternativ oder zusätzlich eine Beeinträchtigung des nonverbalen Ausdrucks durch nicht-demenzielle, z.B. neurologische Er-

krankungen berichtet wurde.

Damit schmerzbezogene Auskünfte demenzkranker Menschen als verlässlich gelten können, muss sichergestellt sein, dass diese die an sie gerichteten Fragen verstehen. Für etwas mehr als jeden zweiten Bewohner (N=107, 54,9%) gaben die Pflegenden jedoch an, dass dieser im Allgemeinen Schwierigkeiten hätte, zu verstehen, was man ihm sagt. Dabei erscheint das Ausmaß dieser Verständnisschwierigkeiten unter den betroffenen Personen recht symmetrisch verteilt, wobei jeweils ungefähr einer aus zehn Bewohnern als entweder nur geringfügig, oder aber vollständig beeinträchtigt beschrieben wird, und sich die restlichen Bewohner einheitlich auf die mittleren Kategorien weniger extremer Verständnisschwierigkeiten verteilen. Wie bereits für die Kommunikationsprobleme, werden Probleme des Verständnisses in den allermeisten Fällen (N=83, 86,5%) auf die demenzielle Erkrankung zurückgeführt. Nicht-demenzielle Ursachen für die Verständnisschwierigkeiten, z.B. Schwerhörigkeit, wurden für (alternativ oder zusätzlich) für insgesamt nahezu 40 Prozent (N=33) der betroffenen Bewohner berichtet.

Insgesamt muss das berücksichtigte Klientel bzw. die realisierte Stichprobe damit als in seinen Möglichkeiten, über innere Zustände Auskunft zu geben, substanziell beeinträchtigt gelten, was die Notwendigkeit für eine ‚evidenzbasierte' Fremdbeurteilung durch Verfahren der Verhaltensbeobachtung unterstreicht.

6.2 Selbstauskunft Demenzkranker zu aktuellen Schmerzen

Erste schmerzbezogene Informationen wurden bereits während der klinischen Begutachtung durch die medizinischen Projektmitarbeiter erfasst. Dabei wurde der Bewohner direkt nach dem Vorliegen aktueller Schmerzen gefragt und – falls akute Schmerzen berichtet wurden – um eine Einschätzung der Stärke dieser Schmerzen gebeten. Wenn der Bewohner weder verbal noch durch nonverbale Zeichen der Zustimmung (z.B. Nicken) auf die Frage reagierte oder Zweifel am Verständnis der Frage bestanden, wurde dies als nicht klar interpretierbare Antwort des Bewohners gesondert kodiert. In Tabelle 10 sind die schmerzbezogenen Antworten der Bewohner im diagnostischen Gespräch dargestellt.

Für nahezu 14% der Befragten war keine eindeutige Antwort auf die Frage nach dem Vorliegen akuten Schmerzes (mehr) möglich. In der Gruppe der Bewohner, für die durch die Pflegenden Verständnisschwierigkeiten berichtet wurden, ist die Rate nicht eindeutig interpretierbarer Antworten mit 19,1 Prozent fast dreimal so hoch wie bei den übrigen Bewohnern (6,8%). Deutlich schwieriger war das direkte Erfragen der aktuellen Schmerzbelastung natürlich auch bei Personen, die als in ihrer verbalen oder nonverbalen Kommunikation beeinträchtigt beschrieben wurden (23,7% bzw. 27,7% nicht klar interpretierbarer Antworten). Diejenigen Bewohner, die keine Antwort auf die Frage nach aktuellen Schmerzen geben konnten, zeichnen sich darüberhinaus erwartungsgemäß durch im Vergleich zu den auskunftsfähigen Bewohnern insgesamt bedeutend stärkere kognitive und nicht-kognitive Symptombelastungen aus (MMST, GDS, Apathie (AES), psycho-pathologische Komorbidität (NPI), Allgemeinzustand, Ernährungszustand, alltagspraktische Fähigkeiten) und finden sich entsprechend überzufällig häufiger in den beiden

Tabelle 10: Selbst- und Fremdeinschätzung akuter und chronischer Schmerzen

N % bzw. N M±SD	*Selbstauskunft*		*Fremdeinschätzung*	
Schmerzbelastung				
keine klar interpretierbare Antwort	29	13,7	–	–
keine akuten Schmerzen	129	60,8	112	57,1
akute Schmerzbelastung	54	25,5	84	42,9
Schmerzintensität				
schwach	10	20,4	8	9,5
mäßig	28	57,1	57	67,9
stark	11	22,5	19	22,6
Mittelwert	49	2,0±0,7	84	2,1±0,6
chronisches Schmerzleiden				
nicht bekannt/weiss nicht			32	17,2
nein			87	46,8
ja			67	36,0

Datenbasis: HILDE2 2006; N=214 (Diagnostik), N=199 (Pflegeinterview).

Kompetenzgruppen der somatisch und psychopathologisch auffälligen schwerdementen Bewohner (Auskunftsrate: LD 100%, MD 93%, SD-S 69%, SD-P 64%; χ^2=27,3, df=3, p<.001).

Obgleich der Großteil der Bewohner sich als zum Erfassungszeitpunkt schmerzfrei beschreibt (ca. 70%), leiden insgesamt 54 Bewohner (ca. 30%) nach eigenen Angaben unter akuten Schmerzen. In der überwiegenden Zahl der Fälle handelt es sich dabei um Schmerz von mäßiger Intensität (57%), nahezu jeder vierte Bewohner aber schätzt seine Schmerzen als stark ein (23%). Bewohner, die im diagnostischen Gespräch vollständige Auskunft zum Vorliegen und der Intensität ihrer Schmerzen machen konnten, weisen einen nur geringfügig besseren kognitiven Status auf als Bewohner, die trotz geäußerter Schmerzen keine Intensitätseinschätzung leisten können (MMST: 14,0±10,5 vs. 17,8±9,8; t=-0.8, df=51, p<.409). Die selbstberichtete Intensität der erlebten Schmerzen korreliert weder mit kognitiven noch mit nicht-kognitiven Demenzmerkmalen der Bewohner.

6.3 Einschätzung der Schmerzbelastung durch Pflegende

Die im Rahmen des Pflegefragebogens durch die Pflegenden eingeschätzten Schmerzinformationen berücksichtigen neben dem Vorliegen und der Intensität aktueller Schmerzen auch Angaben zur dominierenden Schmerzregion und zur chronischen Schmerzbelastetheit der Bewohner. Aufgrund dieser weitgehend parallelen Schmerzerfassung in beiden Studienteilen sind die Ergebnisse der Fremdbeurteilung des zum Untersuchungszeitpunkt beim Bewohner vorliegenden Schmerzes durch die Pflegenden mit der Selbstauskunft der Bewohner gemeinsam in Tabelle 10 auf Seite 193 dargestellt.

6.3.1 Fremdauskunft zu akuten Schmerzen

In der Einschätzung der Pflegenden leiden nahezu 43 Prozent aller teilnehmenden Bewohner, und damit überzufällig mehr Bewohner an akuten Schmerzen, als per Selbstauskunft erfragt wurden (NcNemar's S=7,56, df=1, $p<.006$). Die insgesamt als etwas stärker eingeschätzte Schmerzintensität unterscheidet sich hingegen nicht signifikant von der selbsteingeschätzten Schmerzstärke (Wilcoxon's Vorzeichen Rang Test S=12, $p<.607$). Solche Bewohner, die nach Auskunft der Pflegenden momentan unter Schmerzen litten, waren signifikant weniger kognitiv beeinträchtigt, jedoch in ihren mobilitätsbezogenen Alltagsfähigkeiten (Barthel-Items: Gehen, Treppen steigen, Aufstehen) eingeschränkter als die nach Ansicht der Pflegenden aktuell schmerzfreien Bewohner. Auch in der Fremdauskunft werden häufiger Schmerzen in der Kompetenzgruppe der leicht Demenzkranken berichtet (LD: 63,6%, MD: 39,4%, SD-S: 38,8%, SD-P: 28,6%), wenngleich der statistische Testwert das konventionelle alpha-Niveau von 5% knapp verfehlt. Für die Schmerzintensität können jedoch wie zuvor für die Selbstauskunft berichtet keine systematischen Bezüge zu den berücksichtigten kognitiven und nicht-kognitiven Bewohnermerkmalen nachgewiesen werden.

6.3.2 Fremdauskunft zu chronischen Schmerzbelastungen

Für den weitaus größten Teil der eingeschätzten Bewohner konnten die Pflegenden Angaben zum Vorliegen eines chronischen Schmerzleidens machen (ca. 83%). Danach litt nahezu die Hälfte der einschätzbaren Bewohner bekanntermaßen unter chronischen Schmerzen (ca. 44%). Ob die Pflegenden hinsichtlich chronischer Schmerzerkrankungen informiert waren oder nicht steht in keinem nachweisbaren Zusammenhang mit der Verständnis- oder Auskunftsfähigkeit des Bewohners oder dessen Heimaufenthaltsdauer. Die formale Berufsqualifikation der Pflegenden scheint für dieses Wissen von geringerer Bedeutung zu sein als der Umfang der Tätigkeit; Auszubildende und Pflegende mit geringerem Stundenkontingent entscheiden sich häufiger für die Kategorie „nicht bekannt". Erwartungsgemäß wurden solche Bewohner, für die ein chronisches Schmerzleiden berichtet wurde, auch in der Untersuchungssituation selbst häufiger als akut schmerzbelastet eingeschätzt als die nicht chronisch Schmerzbelasteten (67% vs. 25%; χ^2=27,0, df=2, $p<.001$).

6.3.3 Fremdauskunft zur Schmerzlokalisation

Als die dominierende Schmerzregion wurde am häufigsten der Bereich der Hüfte, des Beines bzw. der Füße angegeben (s. Tabelle 11). Für jeden zweiten Bewohner wurden hier die hauptsächlichen Schmerzen lokalisiert. Aber auch Schmerzen der oberen Extremitäten (Schulter, Arm, Hand) oder Rückenbeschwerden (untere Rückenhälfte, Kreuz, Gesäß) wurden häufiger genannt. Schmerzen im Brustkorb oder im Bereich des Mundes wurden hingegen vergleichsweise selten als hauptsächliche Schmerzregion gewählt.

Tabelle 11: Hauptsächliche Schmerzregion in der Einschätzung durch die Pflegenden

	N	%
Mund/Zähne/Zahnprothese	1	1,4
Kopf/Gesicht	5	6,9
Schulter/Arm/Hand	12	16,7
Brustkorb	2	2,8
Obere Rückenhälfte	3	4,2
Bauchbereich	4	5,5
Untere Rückenhälfte/Kreuz/Gesäß	9	12,5
Hüfte/Bein/Fuß	36	50,0

Datenbasis: HILDE2 2006; N=199.

6.4 Verhaltensbeobachtung mit der BESD-Skala

Wie bei der Beschreibung der Erfassungsinstrumente und des methodischen Vorgehens bereits eingehend beschrieben, liegt dieser Arbeit eine Version der BESD-Skala zugrunde, bei der die in den fünf Indikatorbereichen Atmung, negative Lautäußerung, Gesichtsausdruck, Körpersprache und Trost der Originalversion beschriebenen Verhaltensweisen explizit als zu beobachtende Einzelindikatoren vorgegeben wurden. Insgesamt konnten so 24 diskrete Verhaltens- bzw. Ausdrucksweisen aus dem BESD-Instrument extrahiert werden.

Der ursprünglichen Form der BESD liegt die Annahme zugrunde, dass die einzelnen Verhaltensweisen unterschiedlich starke Schmerzzustände anzuzeigen in der Lage sind, weswegen für manche Verhaltensweisen ein einzelner Punktwert, und für andere zwei Punktwerte vergeben werden, die anschließend als fünf Subskalenwerte in den Gesamtskalenscore einfließen. Auf der Grundlage der einzelnen beobachteten Verhaltensindikatoren ist es möglich, die ursprünglich vorgesehene jeweils dreistufige kategorielle Struktur für die fünf Indikatorbereiche „nachzubauen“. Selbstverständlich bestehen im Rahmen der ursprünglichen Skalenversion, bei der die beschriebenen Verhaltensweisen als Beispiele begriffen werden sollen, für die Beobachter mehr Freiheitsgrade als bei der in diesem Projekt vorgesehenen Beobachtung konkreten Schmerzausdrucks. Es ist darum zu erwarten, dass durch diese engeren Kriterien geringere Subskalen- und Gesamtscores erreicht werden als mit der Originalskala. Die Befunde aus der ersten Feldphase des HILDE-Projektes bestätigen diese Erwartung (s. auch den Arbeitsbericht an den AK Schmerz im Alter der DGSS, 2006).

Um eine optimale Bewertung des Indikationspotenziales der BESD-Skala zu erreichen, muss die Ebene beispielhafter schmerzbezogener Verhaltensweisen überwunden und eine Diskussion möglichst klarer diskreter Verhaltenseinheiten geleistet werden. Darum schließen sich an die Darstellung der Ergebnisse einer an der klassischen Testtheorie orientierten Skalenanalyse vertiefende psychometrische Analysen der Einzelindikatoren auf der Folie einer probabilistischen Testtheorie an (Itemanalyse).

6.4.1 Beobachtete Situationen

Die Bewohner sollten in jeweils zwei für den Pflegealltag relevanten Situationen beobachtet und ihr Schmerzverhalten dokumentiert werden. Ausgewählt wurden hierfür eine Ruhesituation ohne weitere Ansprache durch Pflegekräfte, Mitbewohner oder Besucher, sowie eine Situation mit gesteigerter Aktivität.

Die zu beobachteten Situationen wurden nicht eigens hergestellt oder durch die Pflegenden angeregt, sondern stellen Ausschnitte aus dem alltäglichen Lebensvollzug des Bewohners dar. Um eine größtmögliche Praktikabilität der Schmerzbeobachtung für die Pflegenden bei möglichst ungestörten Alltagsabläufen für die Bewohner sicherzustellen, wurde die Reihenfolge der Beobachtung nicht a priori festgelegt.

Aufgrund der variierenden körperlichen Kompetenzen der Bewohner wurde der Aktivierungsgrad nicht an eine bestimmte vorgegebene Tätigkeit (z.B. Gehen) gebunden vorgegeben, sondern konnte durch die Pflegenden in Abhängigkeit vom körperlichen Status des Bewohners gewählt werden. Aktivierung konnte demnach neben Situationen der Bewegung und körperlichen Aktivität auch solche Situationen einschließen, die stärker durch geistige Anspannung und Konzentration gekennzeichnet sind. Die Pflegekräfte wurden gebeten, die jeweils beobachtete Situation anhand vorgegebener Kategorien genauer zu spezifizieren.

Für vier von fünf Bewohnern wurden Ruhesituationen ausgewählt, in denen sich der Bewohner ohne Ansprache durch Andere in einer sitzenden Position befand, während jeder fünfte Bewohner in einer liegenden Ruhesituation beobachtet wurde.

Die am häufigsten beobachteten Aktivitätssituationen bezogen sich auf das Aufstehen (32%) und Gehen (29%), etwas seltener wurden dagegen Situationen eingeschätzt, in denen der Bewohner die Toilette besucht (14%) oder gelagert wird (6%). Weitere näher dokumentierte Aktivitätssituationen (19%) waren beispielsweise Grund- bzw. Körperpflege (10 Bewohner), Essen (6), Singen (2) oder Malen (2 Bewohner).

6.4.2 Zeitlicher Abstand zwischen Schmerzbeobachtungen

Alle Beobachtungen fanden im Zeitraum von 7 Uhr morgens bis 21 Uhr abends statt, und verteilen sich ungefähr hälftig auf den Vor- und Nachmittag. Die Reihenfolge der durchgeführten Ruhe- und Aktivitätsbeobachtung erscheint ebenfalls annähernd gleichverteilt (Aktivität nach Ruhe: 36%; Ruhe nach Aktivität: 40%). Für knapp ein Viertel der Bewohner konnten beide Beobachtungen innerhalb derselben Stunde durchgeführt werden, insgesamt zwei Drittel aller Beobachtungen liegen nicht weiter als zwei Stunden auseinander.

6.4.3 Klassische Skalenanalyse des BESD-Inventars

Nahezu alle publizierten Untersuchungen zur Güte verschiedener Schmerzassessments gründen auf statistischen Verfahren der klassischen Testtheorie. Der überwiegende Teil

der Arbeiten beschränkt sich dabei auf Aussagen zur Qualität der Gesamtinstrumente, ohne die psychometrischen Eigenschaften der einzelnen Skalenbausteine im Detail zu diskutieren. Als test- und stichprobenabhängiger Ansatz bleiben darüber hinaus viele Fragen der Vergleichbarkeit einzelner Instrumente oder der Repräsentativität der Befunde unbeantwortet. In den folgenden Abschnitten werden eine Deskription der erfassten Daten und eine erste Diskussion der wichtigsten Skalen- und Itemkennwerte der klassischen Testtheorie geleistet.

6.4.3.1 Beobachtungsraten für Einzelindikatoren

Tabelle 12 gibt einen Überblick über die durch die Pflegenden dokumentierten Einzelindikatoren der BESD-Skala unter beiden Beobachtungsbedingungen. Neben der üblichen personenbezogenen Prozentuierung wurden die beobachteten Indikatoren in Tabelle 12 zusätzlich an der Gesamtzahl der innerhalb der einzelnen Beobachtungssituationen beobachteten Indikatoren und mit Blick auf die unterschiedliche Auftretenshäufigkeit in Ruhe- und Aktivitätssituation relativiert. Bereits bei der einführenden Beschreibung der PAINAD bzw. BESD-Skala wurde dargelegt, dass die Informationen zum Wunsch einer (tröstenden) Intervention und deren Erfolg sowohl inhaltlich als auch methodisch nicht mit den ansonsten registrierten Schmerzkennzeichen vergleichbar sind. Darum bezieht sich bei der Ergebnisdarstellung auch die Relativierung der Einzelindikatoren an der Gesamtheit beobachteter Verhaltensmerkmale ausschließlich auf die verbleibenden 22 Items zu mimischem, verbalem und gestischem Ausdrucksverhalten sowie der Atmung.

Ruhesituation

Für die 195 in einer Ruhesituation beobachteten Bewohner wurden insgesamt 620 diskrete schmerzbezogene Verhaltensweisen dokumentiert. Im Durchschnitt konnten pro Bewohner also 3,2 (verschiedene) BESD-Schmerzindikatoren beobachtet werden (vgl. nachfolgende Tabelle 13). Alle im Instrument angesprochenen 22 behavioralen Schmerzkennzeichen konnten in der vorliegenden Stichprobe in der Ruhebedingung beobachtet werden, auch wenn die Beobachtungsraten für nahezu die Hälfte der BESD-Indikatoren weniger als 10 Prozent, und auch insgesamt im Mittel nur knapp 15 Prozent betragen.

Am häufigsten wurden ein trauriger Gesichtsausdruck, ein sorgenvoller Blick, gelegentliches Stöhnen und Ächzen sowie eine angespannte Körperhaltung dokumentiert. Selten wurden in der Ruhesituation hingegen solche Indikatoren angegeben, die eine gesteigerte Aktivierung bzw. Anstrengung implizieren, z.B. Hyperventilation, lautes Stöhnen oder Ächzen oder aggressive Verhaltensweisen wie geballte Fäuste, sich entziehen/wegstoßen oder Schlagen. Deutlich häufiger als in der Aktivitätssituation lassen sich in der Ruhesituation die (insgesamt nicht sehr häufige) Cheyne Stoke Atmung (tieferwerdend/-abflachend/Atempausen), ein trauriger Gesichtsausdruck, Nesteln, sowie starre Körpersprache und (allein durch die häufig sitzende Position bereits erwartbar) angezogene Knie beobachten.

Tabelle 12: Beobachtete Indikatoren der BESD bei geringer und hoher Aktivierung

		Ruhe				*Aktivität*			
Nr.	Item	N	$\%_P$	$\%_K$[1]	$\%_S$	N	$\%_P$	$\%_K$[1]	$\%_S$
Atmung									
1	gelegentlich angestrengt atmen	47	24,1	7,6	31,5	102	52,0	12,8	68,5
2	lautstark angestrengt atmen	5	2,6	0,8	16,7	25	12,8	3,1	83,3
3	kurze Phasen von Hyperventilation	8	4,1	1,3	34,8	15	7,7	1,9	65,2
4	lange Phasen von Hyperventilation	3	1,5	0,5	42,9	4	2,0	0,5	57,1
5	Cheyne Stokes Atmung	30	15,4	4,8	62,5	18	9,2	2,3	37,5
Lautäußerungen									
6	gelegentlich stöhnen oder ächzen	59	30,3	9,5	38,8	93	47,4	11,7	61,2
7	leise missbilligend o. negativ äußern	39	20,0	6,3	47,0	44	22,4	5,5	53,0
8	wiederholt beunruhigt rufen	27	13,9	4,4	47,4	30	15,3	3,8	52,6
9	laut stöhnen oder ächzen	10	5,1	1,6	28,6	25	12,8	3,1	71,4
10	Weinen	17	8,7	2,7	53,1	15	7,7	1,9	46,9
Mimik									
11	trauriger Gesichtsausdruck	69	35,4	11,1	63,3	40	20,4	5,0	36,7
12	ängstlicher Gesichtsausdruck	31	15,9	5,0	34,4	59	30,1	7,4	65,6
13	sorgenvoller Blick	59	30,3	9,5	51,8	55	28,1	6,9	48,2
14	Grimassieren	22	11,3	3,5	48,9	23	11,7	2,9	51,1
Körperhaltung									
15	angespannte Körperhaltung	52	26,7	8,4	31,5	113	57,7	14,2	68,5
16	nervös hin- und hergehen	23	11,8	3,7	37,7	38	19,4	4,8	62,3
17	Nesteln	33	16,9	5,3	61,1	21	10,7	2,6	38,9
18	Körpersprache starr	38	19,5	6,1	65,5	20	10,2	2,5	34,5
19	geballte Fäuste	12	6,2	1,9	46,2	14	7,1	1,8	53,8
20	angezogene Knie	22	11,3	3,5	62,9	13	6,6	1,6	37,1
21	sich entziehen oder wegstoßen	8	4,1	1,3	34,8	15	7,7	1,9	65,2
22	Schlagen	6	3,1	1,0	31,6	13	6,6	1,6	68,4
Gesamt		**620**		**100**		**795**		**100**	
Trost									
38	Bedürfnis zu Trösten	94	48,2		47,2	105	53,6		52,8
39	Trösten nicht möglich[2]	16	17,0			21	20,0		

$\%_P$=bewohnerbezogene -; $\%_K$=kennzeichenbezogene -; $\%_S$=situationsbezogene Beobachtungsrate.
[1] Bezogen auf alle beobachteten BESD-Verhaltensindikatoren mit Ausnahme der Trostitems.
[2] Bezogen auf Bewohner, für die Bedürfnis zu trösten angegeben wurde.
Datenbasis: HILDE2 2006; N=195 (Ruhe), N=196 (Aktivität).

Die Mimik scheint bei der Schmerzerfassung die dominante Indikatorkategorie zu sein. Für durchschnittlich 23,2 Prozent der Bewohner wurde ein Schmerzindikator aus dem Bereich Mimik dokumentiert. Deutlich seltener wurden dagegen im Mittel die Indikatoren der Ausdrucksbereiche Lautäußerungen (15,6%), Körperhaltung (12,4%) oder Atmung (9,5%) beobachtet.

Bei nahezu der Hälfte aller Beobachtungen in der Ruhebedingung verspürten die Pflegenden den Wunsch, das Verhalten der Bewohner durch Berührung oder verbal zu beruhigen bzw. abzulenken. In der überwiegenden Zahl der Fälle konnte das auffällige Verhalten so gemindert werden (83%), bei 16 Bewohnern konnte jedoch keine Ablenkung oder Beruhigung durch die Pflegenden erreicht werden.

Aktivitätssituation

In der Aktivitätssituation konnten erwartungsgemäß deutlich mehr auffällige Verhaltensweisen als potenzielle Schmerzkennzeichen beobachtet werden. Insgesamt wurden hier nahezu 800 (verschiedene) Verhaltensäußerungen dokumentiert, womit durchschnittlich 4,1 der BESD-Indikatoren gezeigt wurden. Auch in der Aktivitätsbedingung konnten alle im Instrument angesprochenen Verhaltensweisen beobachtet werden, wobei jedoch wiederum ein substanzieller Anteil Beobachtungsraten von unter 10 Prozent aufwies. Die durchschnittliche Beobachtungswahrscheinlichkeit für einen BESD-Indikator betrug in der Aktivitätssituation knapp 1:5 (18,6%).

Eine angespannte Körperhaltung, gelegentliches angestrengtes Atmen, und gelegentliches Ächzen oder Stöhnen wurden bei ungefähr der Hälfte aller Bewohner dokumentiert. Mit ungefähr 30 Prozent wurden auch ein ängstlicher Gesichtsausdruck oder sorgenvoller Blick vergleichsweise häufig beobachtet. Auch die insgesamt selten beobachtbaren Schmerzindikatoren wie Schlagen, nervös hin und her gehen, lautes Stöhnen oder lautstark angestrengtes Atmen werden in der Aktivitätssituation deutlich häufiger beobachtet als zuvor in der Ruhesituation.

Betrachtet man die durchschnittlichen Beobachtungsraten der unterschiedenen Ausdrucksbereiche, wurden die Indikatoren aus dem Bereich der Mimik (22,6%) am häufigsten beobachtet. Die Nutzungsrate für diese Ausdruckskategorie erscheint damit mit derjenigen für die Ruhesituation vergleichbar. Die Indikatoren der Kategorien Lautäußerung und Atmung wurden mit nunmehr 21,1 bzw. 16,7 Prozent bei aktivierten Bewohnern deutlich häufiger beobachtet als zuvor in der Ruhesituation. Eine nur mäßige allgemeine Steigerung ist für die Indikatoren aus dem Bereich der Körperhaltung festzustellen (15,8%), wobei hier offensichtlich ein auch inhaltlich gut begründbarer Wechsel der Beobachtungsraten für einzelne Indikatoren deutlich ist. Eine starre Körpersprache und angezogene Knie sind erwartungsgemäß in Aktivität seltener zu beobachten als in der Ruhesituation, während eine angespannte Körperhaltung mehr als doppelt so häufig festgehalten wurde.

Der Wunsch, den Bewohner zu Trösten oder zu Beruhigen ist mit ca. 54 Prozent aller Beobachtungen gegenüber der Ruhesituation leicht erhöht, die Erfolgswahrscheinlichkeit dieser Intervention liegt mit ca. 82 Prozent nur leicht unter derjenigen in Ruhe.

6.4.3.2 Itemschwierigkeit, Trennschärfe und interne Konsistenz

Die Überprüfung der Skalengüte des BESD-Instrumentes wird für alle 22 beobachteten Einzelindikatoren (d.h., exklusive der beiden Trost-Items) für die Ruhe- und Aktivitätssituation in einer gemeinsamen Tabelle dargestellt (s. Tabelle 13). Als Kennwerte der klassischen Skalenanalyse werden dabei jeweils die relativen Beobachtungsraten (als Itemschwierigkeiten), die Item-Gesamtscore-Korrelationen (als Indikator der Diskriminationsfähigkeit), sowie die Konsistenz der Gesamtbatterie der Items (Kuder/Richardson's KR-20) berichtet. Zusätzlich wird für jedes Item die erwartete Skalenkonsistenz angegeben, die mit dessen Elimination aus dem Instrument verbunden wäre. Insbesondere dieser letzte Aspekt, der eng mit der Diskriminationskraft der Items verbunden ist und häufig zur Itemselektion und Skalenoptimierung verwendet wird, ist in den letzten Jahren jedoch zunehmend in die Kritik geraten (Raykov 1997, 2007).

Logische Abhängigkeit zwischen Einzelitems

Mit der Extraktion verschiedener Indikatoren aus den ursprünglich ordinal angenommenen Indikatorbereichen stellt sich allerdings natürlich die Frage, ob denn die in verschiedenen Qualitäten (wie beispielsweise der Lautstärke, Dauer oder Frequenz) gesteigerten Items, die unterschiedliche Intensitäten des Schmerzerlebens anzeigen sollen, prinzipiell überhaupt unabhängig voneinander beobachtet werden können, bzw. ob sich diese nicht gegenseitig ausschließen. Besonders kritisch sind dabei die Itempaare kurze vs. lange Phasen der Hyperventilation, gelegentlich vs. lautstark angestrengt atmen und gelegentlich vs. laut stöhnen bzw. ächzen zu bewerten. Genaugenommen stellt dabei jedoch nur die Hyperventilation ein auf der selben Qualität – nämlich der Dauer – gesteigertes Verhaltensmerkmal dar. Allerdings können auch den ausführlichen Definitionen der BESD-Schmerzindikatoren, die dem Instrument als Material zur Vorbereitung beiliegen, keine konkreteren Anhaltspunkte für die Bestimmung dessen entnommen werden, was hinsichtlich der Hyperventilation als kurz oder lang gelten soll.

In der Ruhesituation werden für knapp 95 Prozent der Bewohner weder kurze noch lange Phasen der Hyperventilation berichtet, für sieben Bewohner (3,7%) wurden nur kurze Phasen und für zwei Bewohner (1,1%) nur lange Phasen von Hyperventilation berichtet. Bei einem Bewohner (0,5%) wurden sowohl kurze als auch lange Phasen hyperventilierender Atmung dokumentiert. In Aktivität wurde bei insgesamt 19 Bewohnern (10%; nur kurz: 15 7,8%; nur lang: 4 2,1%) Hyperventilation beobachtet. Der vermutete gerichtete negative Zusammenhang zwischen beiden Atmungsformen konnte statistisch jedoch nicht bestätigt werden.

Bei den verbleibenden Itempaaren sind die Qualitäten Frequenz und Lautstärke vermischt, weswegen sich diese Verhaltensindikatoren auch theoretisch nicht unbedingt ausschließen sollten. Auch hier konnte die Hypothese eines gerichteten negativen Zusammenhanges weder für die Ruhe-, noch in der Aktivitätssituation bestätigt werden. Vielmehr weisen die gefundenen in der Tendenz positiven Phi-Koeffizienten darauf hin, dass die

Wahrscheinlichkeiten für die Beobachtungen eher gleichgerichtet sind, gelegentliche angestrengte Atmung und Stöhnen häufiger also auch mit lautstarker Atmung bzw. Stöhnen beobachtet werden kann.

Schwierigkeit der Einzelitems

Die in Tabelle 13 dargestellten Itemschwierigkeiten entsprechen aufgrund des dichotomen Skalenniveaus der Indikatoren den zuvor bereits beschriebenen Beobachtungsraten. Mit einer durchschnittlichen Itemschwierigkeit von $\bar{p}_i$=.14 muss das durch BESD erfasste Schmerzverhalten in der Ruhesituation insgesamt als vergleichsweise schwierig eingeschätzt werden. Lediglich drei Indikatoren erreichen Itemschwierigkeiten von .30 und adressieren damit einen Bereich geringeren Schmerzes, während die verbleibenden Items sich nur für die Abbildung stärkerer Schmerzen zu eignen scheinen.

In der Aktivitätsbedingung konnten viele der Indikatoren wie bereits berichtet häufiger beobachtet werden. Entsprechend liegen die Itemschwierigkeiten im Bereich von p_i=.02 und p_i=.58, mit einer mittleren Itemschwierigkeit von $\bar{p}_i$=.18 etwas höher als in Ruhe. Dennoch sind auch hier nur vergleichsweise wenige Items in einem mittleren Schwierigkeitsbereich lokalisiert, so dass die BESD-Skala auch in Aktivität eher geeignet zu sein scheint, stärkere Schmerzzustände anzuzeigen bzw. zu differenzieren.

Auf der Grundlage der für die Einzelitems der BESD geschätzten Itemschwierigkeiten kann die Angemessenheit des durch die Originalskala vorgesehenen Itemscorings innerhalb der Ausdrucksbereiche direkt überprüft werden.

Im Bereich Atmung werden die diskreten Verhaltensweisen gelegentlich angestrengt atmen und tiefer werdende und wieder abflachende Atemzüge mit Atempausen (Cheyne Stokes Atmung) im Vergleich zu den verbleibenden Atmungsitems niedrigere Itemschwierigkeiten geschätzt. Danach erscheint es gerechtfertigt, dass in der kategoriell organisierten Orginalversion für gelegentlich angestrengte Atmung ein Subskalenpunkt und für lautstarke angestrengte Atmung zwei Punkte vergeben werden. Weniger nachvollziehbar ist jedoch, dass die Cheyne Stokes Atmung mit zwei Punkten in den Gesamtscore einfließt. Die für die Verhaltensmerkmale kurze und lange Phasen von Hyperventilation geschätzten Itemschwierigkeiten bestätigen zwar die Rangreihe der Indikatoren in der Tendenz, dennoch erscheint die Vergabe unterschiedlicher Punktwerte aufgrund der minimalen absoluten Schwierigkeitsunterschiede nicht gerechtfertigt. In der Aktivitätssituation kann ein vergleichbares Muster beobachtet werden.

Im Bereich Lautäußerung scheint die Punktevergabe der BESD schon eher mit den geschätzten Itemschwierigkeiten kongruent zu sein. Gelegentliches Stöhnen bzw. Ächzen und leise missbilligende Äußerungen, die jeweils mit einem Punkt gewertet werden, erscheinen tatsächlich weniger schwierig als die drei verbleibenden Schmerzäußerungen, für die zwei Punkte vergeben werden. Dennoch sind die Unterschiede in den Schwierigkeiten eher gering und die Itemschwierigkeiten auch innerhalb der 1- und 2-Punkte-Indikatoren recht verschieden, so dass die jeweilige Gleich- bzw. Doppeltgewichtung auch hier für beide Durchführungsbedingungen insgesamt wenig begründbar erscheint.

Tabelle 13: Kennwerte (KTT) der Indikatoren der BESD bei geringer und hoher Aktivierung

	Ruhe			*Aktivität*		
Item[1]	p_i	r_{it}	KR20$_c$	p_i	r_{it}	KR20$_c$
BESD: Atmung						
gelegentlich angestrengt atmen	.24	.19	.68	.52	.17	.72
lautstark angestrengt atmen	.03	.16	.68	.13	.23	.71
kurze Phasen von Hyperventilation	.04	.27	.67	.08	.32	.70
lange Phasen von Hyperventilation	.02	.10	.68	.02	.06	.71
tiefer werdende u. abflachende Atemzüge/Atempausen	.15	.02	.69	.09	.10	.72
BESD: Lautäußerungen						
gelegentlich stöhnen oder ächzen	.30	.26	.67	.47	.35	.70
sich leise missbilligend oder negativ äußern	.20	.25	.67	.22	.24	.71
wiederholt beunruhigt rufen	.14	.28	.67	.15	.22	.71
laut stöhnen oder ächzen	.05	.25	.67	.13	.33	.70
Weinen	.09	.41	.66	.08	.30	.70
BESD: Mimik						
trauriger Gesichtsausdruck	.35	.28	.67	.20	.21	.71
ängstlicher Gesichtsausdruck	.16	.50	.64	.30	.47	.68
sorgenvoller Blick	.30	.28	.67	.28	.29	.70
Grimassieren	.11	.27	.67	.12	.25	.70
BESD: Körperhaltung						
angespannte Körperhaltung	.27	.42	.65	.58	.47	.68
nervös hin- und hergehen	.12	.34	.66	.19	.17	.71
Nesteln	.17	.18	.68	.11	.21	.71
Körpersprache starr	.20	.12	.69	.10	.26	.70
geballte Fäuste	.06	.25	.67	.07	.33	.70
angezogene Knie	.11	.17	.68	.07	.33	.70
sich entziehen oder wegstoßen	.04	.36	.67	.08	.37	.70
Schlagen	.03	.23	.67	.07	.27	.70
Kuder/Richardson's KR-20		.68			.71	
Gesamtscore BESD (ohne Trost)[2]		195 3,2±2,7			196 4,1±3,0	
Gesamtscore BESD (mit Trost)		195 3,7±3,0			196 4,7±3,3	

[1] Dichotome Antwortskala: 0=„nein"; 1=„ja"; [2] Summe beobachteter Indikatoren: N M±SD.
p_i=Itemschwierigkeit (Beobachtungsrate); r_{it}=Itemdiskrimination (punkt-biserielle Item-Total-Korrelation); KR20$_c$=Skalenkonsistenz bei Reduzierung um Item.
Datenbasis: HILDE2 2006; N=195 (Ruhe), N=196 (Aktivität).

Im Bereich der Mimik sind nur für einen grimassierenden Gesichtsausdruck zwei Punkte vorgesehen. Tatsächlich wird dieses Schmerzitem als vergleichsweise schwierig geschätzt. Doch auch das Merkmal ängstlicher Gesichtsausdruck wird vergleichsweise bzw. ähnlich selten beobachtet, erhielte in der BESD-Wertung dagegen nur einen Punkt. Die Beobachtungsraten in der Aktivitätssituation bestätigen das vorgesehene Scoring etwas stärker.

Insgesamt fünf der acht extrahierten Indikatoren des Ausdrucksbereiches Körperhaltung sollten vergleichsweise stärkere Schmerzniveaus anzeigen, und würden mit zwei Punkten gewertet. Insbesondere eine starre Körpersprache wird in der Ruhesituation jedoch als vergleichsweise einfaches Item geschätzt. Auch angezogene Knie können vergleichsweise häufig beobachtet werden und dieses Item erscheint zumindest in der Ruhesituation nicht schwieriger als beispielsweise nervös- hin und hergehen, für das lediglich ein Punkt vergeben würde. In der Aktivitätssituation dagegen verändern sich die Schwierigkeiten der Einzelitems in unterschiedliche Richtungen, sodass das nun vergleichsweise seltene Nesteln mit einem Punkt zu gering gewichtet erscheint, während angezogene Knie nun tatsächlich vergleichsweise schwierig zu beobachten sind.

Trennschärfe der Einzelitems

Als Kennwerte für die Trennschärfe bzw. Diskriminationskraft sind in Tabelle 13 die punkt-biseriellen Korrelationen der Einzelitems mit dem Gesamtskalenscore dargestellt. In der Ruhesituation erreichen diese Zusammenhangsmaße lediglich Werte zwischen r_{it}=.02 und .50, und müssen mit einer mittleren Item-Total-Korrelation von nur $\bar{r}_{it}$=.24 als wenig diskriminativ gelten.

Die Diskriminationsparameter der Einzelitems der BESD liegen in der Aktivitätssituation mit Werten zwischen r_{it}=.06 und .47, und einer mittleren Item-Total-Korrelation von $\bar{r}_{it}$=.27 zwar etwas höher; dennoch können Bewohner mit ähnlichen Schmerzausprägungen auf der Grundlage der BESD-Items auch in der Aktivitätssituation kaum voneinander unterschieden werden.

Skalenkonsistenz

Da die Einzelindikatoren jeweils nur zu einem kleinen Teil Schmerzen anzeigen, daneben jedoch viel unerklärte Varianz besitzen, muss auch die Gesamtskala als wenig homogen beurteilt werden. Entsprechend liegen die geschätzten Skalenkonsistenzen mit Werten von Kuder/Richardson's KR-20=.68 für die Ruhe und .71 für die Aktivitätssituation deutlich niedriger als für ein Instrument zu erwarten, das für sich beansprucht, ein einziges gemeinsames Merkmal Schmerz zu erfassen.

Zur Optimierung von Messinstrumenten wird häufig analysiert, wie sich die Skalenhomogenität bzw. -konsistenz verändern sollte, wenn ein einzelnes Item aus der Testbatterie gestrichen würde. Nach diesem Kriterium könnte sowohl in der Ruhe- als auch in der Aktivitätssituation eine geringfügig bessere Skalenkonsistenz erreicht werden, wenn

das BESD-Inventar um das offensichtlich überhaupt nicht diskriminierende Item Cheyne-Stokes-Atmung (tiefer werdende u. wieder abflachende Atemzüge mit Atempausen) reduziert würde. Gleichermaßen erscheint auch das Item gelegentlich angestrengte Atmung in der Aktivitätssituation als verzichtbar. Für alle anderen Indikatoren jedoch wäre ihre Eliminierung für die Skalenkonsistenz nicht von Belang oder geringfügig nachteilig. Die Aussagekraft dieses Optimierungsverfahrens ist jedoch ganz offensichtlich sehr gering, da stets nur ein einzelnes Item berücksichtigt wird.

Gesamtskalenscore

Lediglich ein vergleichsweise kleiner Bereich des in der Einzelitem-Version möglichen Wertebereichs von 0 bis maximal 22 Punkten wurde in den gegebenen Beobachtungsbedingungen und der untersuchten Bewohnerstichprobe belegt (Ruhe: 0-12 Punkte; Aktivität: 0-14 Punkte). Lediglich 13,5 Prozent der Bewohner zeigten in der Ruhesituation, und weniger als 10 Prozent in der Aktivitätssituation überhaupt keine schmerzbezogenen Ausdrucks- und Verhaltensweisen. Ungefähr vier aus fünf Bewohnern erreichten einen Wert von unter 6 Punkten und liegen damit im untersten Quartil des theoretischen Wertebereiches. Die empirischen Verteilungen des Gesamtscores sind damit jeweils rechtsschief und weichen deutlich von einer Normalverteilung ab.

Für die Aktivitätssituation wurden im Mittel höhere BESD-Skalenscores erreicht als in der Ruhesituation. Dieser systematische Unterschied konnte auch auf Bewohnerebene durch eine längsschnittliche Varianzanalyse (Repeated Measures ANOVA) bestätigt werden (BESD: 4,7$\pm$3,3 vs. 3,7$\pm$3,0, Wilk's λ=0,899, df=1, 193, p$<$.001).

Trostitems

Der Indikatorbereich Trost wurde aus inhaltlichen und formal-methodischen Gründen aus der Analyse der Konsistenz der Itembatterie ausgeschlossen. Es ist naheliegend und von den Entwicklern der PAINAD-Skala wohl auch erwünscht, dass die beobachtenden Pflegenden sich in ihrem Wunsch oder der Beurteilung der Notwendigkeit, den Bewohner zu Trösten, am in der Beobachtungssituation vom Bewohner geäußerten Schmerzverhalten orientieren. Sowohl in der Ruhesituation wie auch bei Aktivität wurden für Bewohner, die nach Meinung der Pflegenden getröstet werden sollten mehr schmerzbezogene Verhaltensäußerungen beobachtet als für Bewohner, für die Trösten nach Meinung der Pflegenden nicht notwendig war (Ruhe: 94 4,4$\pm$2,8 vs. 101 2,0$\pm$2,0; t=-6,9, df=169, p$<$.001; Aktivität: 105 5,2$\pm$2,8 vs. 91 2,7$\pm$2,7; t=-6,4, df=194, p$<$.001). Keine systematischen Unterschiede im BESD-Skalescore konnten hingegen in der Ruhesituation zwischen denjenigen Bewohnern bestätigt werden, bei denen Trösten möglich war und solchen, die nicht getröstet werden konnten. Ein deutlich anderes Bild ergibt sich für die Aktivitätssituation, in der diejenigen 21 Bewohner, die nicht durch Pflegende beruhigt bzw. getröstet werden konnten deutlich höhere BESD-Scores aufwiesen (21 6,4$\pm$2,7 vs. 84 4,9$\pm$2,7; t=2,2, df=103, p$<$.027).

6.4.3.3 Limitationen

Wie bei der methodischen Diskussion der Vor- und Nachteile verschiedener testtheoretischer Konzeptionen bereits ausführlich diskutiert, ergeben sich bei der Interpretation der durch die klassische Skalenanalyse gewonnenen psychometrischen Kennwerte einige Einschränkungen. Im konkreten Fall der Schmerzmessung durch die Items der BESD-Skala zeigen sich die folgenden Grenzen der klassischen Testtheorie.

Testabhängigkeit

Die beschriebenen Itemdiskriminationen bleiben jeweils auf den Gesamtscore der Skala bezogen, und können somit immer nur mit Blick auf die von allen Items geteilte Merkmalsvarianz interpretiert werden. Aufgrund der hierarchischen Organisation der BESD-Einzelindikatoren im Rahmen dieser Studie ergibt sich darüber hinaus die Möglichkeit und Aufgabe, auch Gemeinsamkeiten zwischen Einzelitems eines Ausdrucksbereiches (z.B. der Körperhaltung) zu beschreiben, die ihrerseits nur zum Teil mit anderen Ausdrucksbereichen (z.B. der Mimik) kovariieren. Eine Optimierung der Skala durch eine Selektion der trennschärfsten oder Eliminierung wenig diskriminativer Items bleibt wenig nützlich, solange die herangezogenen Kennwerte nicht unabhängig von einem spezifischen Skalenkomposit geschätzt werden können.

Stichprobenabhängigkeit

Im Rahmen der bisherigen Skalenanalyse kann nicht entschieden werden, ob die geringen Beobachtungsraten und folglich die hohen geschätzten Itemschwierigkeiten für die meisten der beschriebenen Indikatoren eine Folge des in der realisierten Stichprobe demenzkranker Heimbewohner mäßigen Schmerzniveaus sind, oder ob die geschätzte geringe Schmerzbelastung besser als Folge zu schwieriger Verhaltensindikatoren interpretiert werden sollte.

Heterogenes Skalenniveau der Indikatoren

Die beiden Indikatorstufen des Ausdrucksbereiches Trost bzw. Tröstbarkeit der Originalskala wurden wie die für die anderen Ausdrucksbereiche beschriebenen Indikatoren als dichotome Indikatoren zur Beobachtung vorgegeben. Durch die explizite Filterführung und die sich dadurch ergebenden linearen Abhängigkeiten in den Daten musste der Bereich Trost aus den zuvor beschriebenen Analysen ausgeschlossen werden. Formal bleibt der Ausdrucksbereich Trost damit trotz der Vorgabe einzelner Items ein dreistufiger Indikator. Die herkömmliche Skalenanalyse erlaubt jedoch keine Berücksichtigung gemischter (d.h. hier dichotomer und dreistufiger) Itemformate.

Vergleich beider Beobachtungsbedingungen

In der Aktivitätssituation wurde deutlich mehr schmerzbezogener Verhaltensausdruck beobachtet als in Ruhe. Entsprechend sollte die Aktivitätssituation als die mit stärkerem Schmerzerleben verbundene Beobachtungsbedingung interpretiert werden. Die gesteigerte Beobachtungsrate für viele der Einzelindikatoren führt im Sinne der klassischen Testtheorie zu einer reduzierten Itemschwierigkeit, d.h. der Test auf Schmerzerleben wird in der Aktivitätssituation insgesamt leichter, womit sich wiederum die zuvor beschriebenen Interpretationsschwierigkeiten ergeben. Der Einfluss verschiedener Beobachtungsbedingungen als Facette einer Messung wird im Kontext der herkömmlichen Skalenanalyse nicht berücksichtigt. Von der Möglichkeit, im Rahmen der Generalisierungstheorie verschiedene Szenarien der Schmerzerfassung systematisch miteinander zu vergleichen, ist zumindest mit Blick auf die in Frage stehenden Schmerzinventare bislang kein Gebrauch gemacht worden. Die Frage, ob das BESD-Inventar in Situationen geringer und hoher Aktiviertheit gleich bzw. gleich gut funktioniert, wird im Kapitel 6.7 bearbeitet. Zuvor jedoch soll die Struktur der BESD-Skala mit Methoden der probabilistischen Testtheorie eingehend untersucht werden. In Anbetracht des hohen Auflösungsgrades sollen sich diese Analysen zunächst auf die in der Ruhebedingung erhobenen Informationen beschränken.

6.4.4 IRT-Analyse der BESD-Schmerzindikatoren

Die Ergebnisse der deskriptiven Auswertungen und klassischen Skalenanalysen der BESD wurden zuvor – im Gegensatz zur Bearbeitung in der ersten HILDE-Projektphase (s. Kap. 5.3.4.2) – für die dichotomen Einzelitems dargestellt. Bereits im Zuge der klassischen Itemanalyse wurden Hinweise auf die nur in Teilen gerechtfertigte implizite Zuordnung der einzelnen Verhaltensweisen zu verschiedenen Schmerzintensitäten im Originalaufbau der PAINAD gefunden. Diesen Hinweisen soll im Folgenden mit einer Methodologie weiter nachgegangen werden, mit der die Itemparameter unabhängiger (von Testzusammensetzung und Schmerzniveau der Stichprobe) geschätzt werden können.

Die Untersuchung der psychometrischen Eigenschaften der BESD-Indikatoren (und in paralleler Weise auch der CNPI-Items) auf der Folie der probabilistischen Testtheorie gliedert sich in zwei aufeinander aufbauende Analyseschritte.

1. Da aufgrund der Erkenntnisse der klassischen Skalenanalyse davon ausgegangen werden muss, dass manche der in den Inventaren enthaltenen Verhaltensweisen nur eine schlechte Abbildung von Schmerzen erlauben, und auch die reliabler erscheinenden Items in ihrer Indikationskraft durchaus heterogen sind, werden in einem ersten Schritt zweiparametrische logistische (2-PL) Messmodelle separat für jeden Ausdrucksbereich geschätzt, der hinreichend viele (mindestens drei) Indikatoren für die Bestimmung eines latenten bereichsspezifischen Schmerzfaktors umfasst. Auf der Grundlage dieser Schätzungen werden die bisher getroffenen Aussagen zu Itemschwierigkeiten und -diskriminationen aus einer probabilistischen Perspektive

heraus vervollständigt und Überlegungen für eine angemessenere statistische Repräsentation des erfassten schmerzbezogenen Ausdrucksverhaltens angestellt.

2. Auf der Grundlage der Erkenntnisse zur Messstruktur innerhalb der Indikatorbereiche soll in einem zweiten Auswertungsschritt ein BESD-Gesamtmodell geschätzt werden, das die implizite hierarchische faktorielle Struktur der Skala berücksichtigt und auch diejenigen Ausdrucksbereiche mit einschließt, für die aufgrund ihrer Binnenstruktur oder in Ermangelung mehrerer Indikatoren im ersten Auswertungsschritt kein eigener bereichsspezifischer latenter Schmerzfaktor geschätzt werden konnte (z.B. BESD: Trost; CNPI: verbale Beschwerden, Klammern und Reiben).

6.4.4.1 Messstruktur der einzelnen BESD-Ausdrucksbereiche

Die Ergebnisse der separaten Modellierung der empirischen Zusammenhänge zwischen den Indikatoren der vier unterschiedenen BESD-Ausdrucksbereiche Atmung, Lautäußerung, Mimik und Körperhaltung sind für die Ruhesituation gemeinsam in Tabelle 14 dargestellt. Für eine bessere Anschaulichkeit der berichteten Modellparameter sind die Ergebnisse jeweils mit einem Piktogramm der spezifizierten Messstruktur hinterlegt.

Die Modellparameter der jeweils einfaktoriellen Messstruktur mit dichotomen Indikatoren wurden mit der Latent Variable Modeling (LVM) Software M*plus* (Muthén & Muthén, 1998-2006) geschätzt. Wie in der einführenden Diskussion der methodischen Zugänge zur Messung von Schmerzen durch kategorielle Verhaltensindikatoren ausgeführt, definieren die an der Logik der Strukturgleichungsmodelle orientierten Analyseverfahren latente Responsevariablen y_i^* (LRVs), die über Schwellenparameter mit den empirisch erfassten kategoriellen Indikatoren y_i verknüpft sind. Die durch ein gegen Verletzungen der Modellvoraussetzungen robustes Maximum-Likelihood (MLR)-Verfahren geschätzten Modellparameter sind dabei die Varianz des latenten Merkmals ψ, die Einflussgewichte λ_i dieses Faktors auf die latenten Responsevariablen, sowie schließlich die itemspezifischen Schwellenparameter τ_i, ab denen sich eine latente Verhaltensausprägung in einer konkreten Verhaltensbeobachtung realisiert.

Die logische und formale Äquivalenz der so modellierten eindimensionalen Messstruktur mit einem zweiparametrischen logistischen Item Response Modell wurde ebenfalls bereits im Methodenteil dieser Arbeit detailliert dargestellt. Um trotz der unterschiedlichen geschätzten Modellparameter dieser LRV-Formulierung auch bei der Ergebnisdarstellung eine an der probabilistischen Testtheorie orientierte Interpretation zu erleichtern, werden in Tabelle 14 neben den M*plus* Modellparametern auch die Itemschwierigkeiten b und Diskriminationsparameter a in der üblichen IRT-Metrik berichtet.

Zur Abschätzung der Güte der Repräsentation der empirischen Datenstruktur durch das spezifizierte Modell wurde der bei der MLR-Schätzung verfügbare χ^2-Anpassungstest herangezogen. Weitere Hinweise auf Bereiche potenzieller Fehlspezifikation lieferten darüber hinaus äquivalente – hier nicht dargestellte – konfirmatorische Faktorenanalysen mit einer adjustierten Weighted-Least-Squares (WLSMV)-Parameterschätzung.

Tabelle 14: Kennwerte (IRT) der Indikatoren der BESD Ausdrucksbereiche in Ruhe

		Separate 2PL-Modelle für Indikatorbereiche						
Nr.	Item	θ	τ_s	λ_s	ψ	b	a	Anpassung
	Atmung							
1	gelegentlich angestrengt atmen	0,65	0,67	.59		1,13	0,79	χ^2=28,3
2	lautstark angestrengt atmen	0,79	1,98	.46	1,79	4,27	0,56	*df*=21
3	kurze Phasen von Hyperventilation	0,29	1,73	.85		2,05	1,68	p <.131
4	lange Phasen von Hyperventilation	0,69	2,22	.56		3,97	0,72	
5	Cheyne Stokes Atmung	0,54	0,99	−.68		−1,47	−0,98	
	Lautäußerungen							
6	gelegentlich stöhnen oder ächzen	0,86	0,47	.38		1,26	0,43	χ^2=24,6
7	leise missbilligend/negativ äußern	0,79	0,79	.46	0,54	1,71	0,55	*df*=21
8	wiederholt beunruhigt rufen	0,87	1,02	.36		2,87	0,41	p <.266
9	laut stöhnen oder ächzen	0,19	1,63	.90		1,82	2,16	
10	Weinen	0,45	1,34	.74		1,83	1,16	
	Mimik							
11	trauriger Gesichtsausdruck	0,83	0,34	.41	0,68	0,83	0,49	χ^2=6,3
12	ängstlicher Gesichtsausdruck	0,58	0,97	.65		1,48	0,92	*df*=7
13	sorgenvoller Blick	0,75	0,48	.50		0,96	0,62	p <.511
14	Grimassieren	0,94	1,14	.24		4,80	0,26	
	Körperhaltung							
15	angespannte Körperhaltung	0,77	0,85	.48		1,22	0,58	χ^2=290,7
16	nervös hin- und hergehen	0,94	1,12	.25		4,38	0,28	*df*=239
17	Nesteln	0,85	0,90	.39	0,97	2,33	0,45	p <.012
18	Körpersprache starr	0,90	0,79	.32		2,52	0,35	
19	geballte Fäuste	0,28	1,54	.85		1,80	1,75	
20	angezogene Knie	0,70	1,17	.55		2,12	0,71	
21	sich entziehen oder wegstoßen	0,15	1,73	.92		1,88	2,54	
22	Schlagen	0,24	1,86	.87		2,14	1,87	
	Trost[1]							
38	Bedürfnis zu Trösten	–	–	–	–	–	–	–
39	Trösten nicht möglich	–	–	–	–	–	–	–

θ=Residualvarianz von y^*; τ_s=Threshold (standardisiert); λ_s=Regressionsparameter (standardisiert).
ψ=Faktorvarianz (unstandardisiert); b=Itemschwierigkeit; a=Itemdiskrimination.
[1] Der Indikatorbereich Trost ist mit nur zwei Indikatoren nicht separat schätzbar, da nicht identifiziert.
Datenbasis: HILDE2 2006; N=194 (Ruhe).

BESD-Ausdrucksbereiche – Atmung

Die wechselseitigen Zusammenhänge der atmungsbezogenen Schmerzitems in der Ruhesituation können danach angemessen durch ein einfaktorielles Messmodell repräsentiert werden. Dennoch wird der Indikator Cheyne Stokes Atmung als negativ mit dem zugrundeliegenden Atmungsfaktor verknüpft geschätzt. Dieser Befund erscheint nachvollziehbar, da in der ausführlichen Formulierung dieses Items explizit auch flacher werdende Atmung und Atempausen als charakteristisch beschrieben werden, während die verbleibenden Items allesamt eine Steigerung der Atmung anzeigen. Zusätzlich dazu erscheint dieser Indikator im Gegensatz zu den bisher vorgestellten Skalenanalysen als substanziell mit dem gemeinsamen latenten Faktor verknüpft, kann also als durchaus reliabler Indikator zumindest für den Teilbereich atmungsbezogenen potenziellen Schmerzausdruckes angesehen werden. Der jeweilige Anteil gemeinsamer Variation innerhalb der fünf Einzelitems des Indikatorbereiches Atmung kann durch eine Analyse der itemspezifischen Residuen θ_i abgeschätzt werden. Am vollständigsten erscheint mit 29 Prozent unerklärter Varianz das Item kurze Phasen von Hyperventilation repräsentiert, während die verbleibenden angenommenen latenten Verhaltensindikatoren jeweils nur mit weniger als der Hälfte ihrer geschätzten Variabilität den gemeinsamen Faktor schmerzbedingter Atmungsveränderung anzeigen. Die Transformation der geschätzten Modellparameter zu IRT-Itemdiskriminationen zeichnet erwartungsgemäß das gleiche Bild. Mit einem Wert von a=1,7 besitzt das Item kurze Phasen der Hyperventilation die größte Diskriminationskraft, während lautstark angestrengtes Atmen mit a=0,6 als am wenigsten diskriminativ gelten muss. Nicht zuletzt aufgrund des negativen Regressionsgewichtes für den Indikator Cheyne Stokes Atmung muss angenommen werden, dass eine Berücksichtigung unterschiedlicher Diskriminationen im Rahmen eines zweiparametrischen Modells eine deutlich bessere Repräsentation der Daten erlaubt als eine 1-PL-Schätzung mit als invariant angenommenen Itemdiskriminationen. Tatsächlich weist der für die Nicht-Normalität der Daten korrigierte skalierte χ^2-Differenzentest für diese genesteten Modelle das einparametrische Modell als die signifikant schlechter fittende Modellvariante aus ($-2ll_\Delta$=13,7, df_Δ=4, p$<$.008). Allerdings können nach der Eliminierung des offensichtlich abweichenden Cheyne Stokes Items die verbleibenden vier Atmungsindikatoren ohne substanziellen Model-Fit-Verlust auch mit invarianten Diskriminationsparametern geschätzt werden ($-2ll_\Delta$=0,65, df_Δ=3, p$<$.884).

Die Transformation der geschätzten Schwellenparameter τ_i zu Itemschwierigkeiten zeigt, dass die einzelnen Atmungsitems ihren maximalen Informationswert an deutlich unterschiedlichen Positionen auf dem latenten Schmerzfaktor erreichen. Als vergleichsweise schwierig können mit Werten von b=4,3 und 4,0 in der Ruhesituation die Items lautstark angestrengt atmen und lange Phasen von Hyperventilation gelten. In einem mittleren Bereich liegt dagegen der Verhaltensausdruck kurze Phasen von Hyperventilation (a=2,1), während gelegentliches angestrengtes Atmen mit einem Schwierigkeitsparameter von a=1,1 am ehesten auch Bereiche geringer Schmerzausprägungen anzuzeigen in der Lage ist. Die für den Einzelindikator Cheyne Stokes Atmung geschätzte negative Itemdis-

krimination verletzt offensichtlich die in der Item-Response-Theorie formulierte Grundvoraussetzung monoton ansteigender itemcharakteristischer Funktionen, weshalb von einer inhaltlichen Interpretation des geschätzten Schwierigkeitswertes von a=-1,5 abgesehen werden soll. Mit Ausnahme der Cheyne Stokes Atmung bestätigen die geschätzten Schwierigkeitsparameter weitestgehend die durch die BESD gewählte Staffelung der zu vergebenden Punktwerte, auch wenn das Ausdrucksverhalten Hyperventilation generell schwieriger zu sein scheint als der Bereich angestrengter Atmung.

BESD-Ausdrucksbereiche – Lautäußerung

Die Zusammenhänge zwischen den fünf Schmerzindikatoren des Ausdrucksbereiches Lautäußerungen scheinen ebenfalls angemessen durch einen gemeinsamen latenten Faktor abbildbar. Die LRV für den Verhaltensausdruck laut stöhnen oder ächzen ist mit einem standardisierten Regressionsgewicht von λ=.90 am engsten mit dem dahinterstehenden schmerzbezogenen Bereichsfaktor verknüpft, und auch Weinen scheint zum überwiegenden Teil (θ=0,45) durch das gemeinsame latente Merkmal bestimmt. Deutlich geringer sind jedoch die Regressionsgewichte (.36 bis .46) und dementsprechend die Determinationsanteile (.13 bis .21) für die verbleibenden Items gelegentlich stöhnen oder ächzen, leise missbilligende Äußerungen und wiederholtes beunruhigtes Rufen. Die Berechnung der IRT-Diskriminationen akzentuiert diese Unterschiede zusätzlich, so dass sich für die drei letztgenannten Items vergleichsweise flache (a=0,4 bis a=0,6) itemcharakteristische Kurven (ICCs) ergeben. Mit Itemdiskriminationen von a=2,2 und 1,2 erscheinen lautes Stöhnen oder Ächzen und Weinen dagegen sehr viel informativer. Obgleich die geschätzten Regressionsgewichte und Itemdiskriminationen doch recht heterogen scheinen, muss der Zuwachs an Misfit, der mit einer Gleichrestriktion der Regressionsgewichte verbunden wäre, als relativ unerheblich betrachtet werden ($-2ll_{\triangle}$=7,6, $df_{\triangle}$=4, p$<$.109). In Anbetracht der zu fordernden Modellparsimonität sollten die schmerzbezogenen Verhaltensindikatoren des Ausdrucksbereiches Lautäußerung danach als vergleichbar diskriminativ angenommen und durch ein 1PL-Modell repräsentiert werden.

Der durch die vokalen Indikatoren abgedeckte Bereich auf dem gemeinsamen Schmerzbereichsfaktor erscheint mit geschätzten Itemschwierigkeiten zwischen b=1,3 (gelegentlich stöhnen oder ächzen) und 2,9 (wiederholtes beunruhigtes Rufen) vergleichsweise eng. Die durch das BESD vorgesehene Gewichtung der Ausdrucksweisen erfährt durch die Rangreihe der Schwierigkeiten eine gewisse Bestätigung, auch wenn das Weinen oder laute Stöhnen oder Ächzen, für die jeweils zwei Punkte vorgesehen sind, sich in ihrer Schwierigkeit (jwls. b=1,8) nicht wesentlich von leisen missbilligenden Äußerungen unterscheiden (b=1,7), für die nur ein Punkt vergeben würden.

BESD-Ausdrucksbereiche – Mimik

Auch die vier BESD-Indikatoren aus dem Bereich potenziell schmerzbezogenen mimischen Ausdrucks zeigen einen gemeinsamen Faktor an. Allerdings erscheint die Aus-

drucksform Grimassieren (λ=.24) nur wenig eng an den gemeinsamen Faktor gebunden. Auch für die verbleibenden Items ängstlicher Gesichtsausdruck, sorgenvoller Blick und trauriger Gesichtsausdruck können jedoch nur vergleichsweise flache ICCs geschätzt werden (a=0,9, 0,6 und 0,5). Vergleicht man wiederum das dargestellte 2PL-Modell mit einer einparametrischen Spezifikation, wäre das stärker restringierte Modell mit einheitlichen Itemdiskriminationen aus Gründen der Modellsparsamkeit vorzuziehen ($-2ll_{\Delta}$=2,4, df_{Δ}=3, p$<$.496).

Die geschätzte Itemschwierigkeit des Schmerzausdrucks Grimassieren liegt um ein Vielfaches höher als diejenigen der verbleibenden mimischen Schmerzreaktionen, womit zumindest das BESD-Scoring bestätigt scheint. Trotzdem erscheint diese zweigeteilte Lokalisierung der unterschiedenen mimischen Indikatoren insgesamt ungünstig, da Bereiche mittlerer Schmerzausprägungen offensichtlich nicht gut abgedeckt werden.

BESD-Ausdrucksbereiche – Körperhaltung

Der signifikante χ^2-Anpassunstest für diesen Ausdrucksbereich lässt vermuten, dass die wechselseitigen Itemzusammenhänge zwischen den BESD-Indikatoren aus dem Bereich der Körperhaltung durch einen einzigen gemeinsamen Faktor nur unzureichend aufgeklärt werden können. Tatsächlich weisen die drei Indikatoren nervös hin- und hergehen, Nesteln und starre Körpersprache jeweils Regressionsgewichte von unter λ=.40 auf und sind damit lediglich zu einem kleinen Teil (max. 15%) durch den gemeinsamen gestischen Schmerzfaktor bestimmt. Dem stehen die drei Items geballte Fäuste, sich entziehen oder wegstoßen und Schlagen entgegen, die mit Regressionsgewichten deutlich über .80 auch die inhaltliche Interpretation der geschätzten Schmerzausdruckskomponente dominieren. Die verbleibenden Items angespannte Körperhaltung und angezogene Knie nehmen dagegen hinsichtlich ihrer Diskriminationskraft eine mittlere Position ein. Dementsprechend variieren auch die transformierten Diskriminationsparameter a_i deutlich zwischen 0,28 (nervös hin- u. hergehen) und 2,5 für Entzug und Wegstoßen. Wie bereits angemerkt, kann die Matrix der empirischen Itemzusammenhänge durch das einfaktorielle 2-PL-Messmodell nur unvollständig repräsentiert werden. Wie aufgrund der sehr heterogen geschätzten Regressionsparameter erwartet werden konnte, verschlechtert sich der Model-Fit für den Bereich Körperhaltung nochmals bedeutend, wenn die λ-Parameter jeweils auf den gleichen Wert restringiert werden (1PL; $-2ll_{\Delta}$=15,5, df_{Δ}=7, p$<$.030). Auch für diesen Bereich ist zu erwarten, dass ein Subset diskriminativer Items sich ohne substanziellen Verlust an Modellpassung auch durch ein weniger komplexes einparametrisches Messmodell repräsentieren lässt.

Die geschätzten itemspezifischen Thresholdparameter, und die daraus ermittelten Itemschwierigkeiten bestätigen die zuvor bereits angesprochene kontextuelle Gebundenheit mancher Schmerzindikatoren. So kann die Verhaltensweise nervös hin- und hergehen in der Ruhesituation nur vergleichsweise selten beobachtet werden und wird mit b=4,4 entsprechend auch als sehr schwieriges Schmerzitem eingeschätzt. Für angespannte Körperhaltung, geballte Fäuste und sich entziehen oder wegstoßen wurden dagegen vergleichs-

weise geringe Schwierigkeiten geschätzt. Insgesamt finden die theoretischen Annahmen der BESD zum Indikationsbereich einzelner Items, und demnach auch die BESD-Punktevergabe in den hier vorliegenden empirischen Daten keine hinreichende Bestätigung. Zwei der drei mit einem Punkt verknüpften Indikatoren weisen mit die höchsten Itemschwierigkeiten auf, und zwei der fünf mit zwei Punkten verknüpften Items haben mit die geringsten Schwierigkeitsschätzungen.

Zusammenfassend kann für die Ruhesituation angemerkt werden, dass einige der in der Originalskala vorgestellten Schmerzitems wenig Varianz mit den verbleibenden Items teilen und darum einen nur geringen Beitrag zur Abbildung des jeweiligen bereichsspezifischen Schmerzausdrucksfaktors leisten können. Eine Reduktion um diese Items könnte eine reliablere Erfassung des Schmerzniveaus in verschiedenen Kommunikations- bzw. Ausdrucksbereichen ermöglichen. Hinweise auf deutliche Abweichungen des Itemverhaltens von der angenommenen Funktionsweise wurden für die Cheyne Stokes Atmung festgestellt, für die eine negative Diskrimination geschätzt wurde. Für zwei der vier eingehend untersuchten BESD-Ausdrucksbereiche wurde festgestellt, dass eine Modellierung von variablen Itemschwierigkeiten und Itemdiskriminationen unnötig komplex ist und die empririschen Zusammenhänge ähnlich gut auch durch ein sparsameres einparametrisches Messmodell mit vergleichbaren Diskriminationsparametern repräsentiert werden könnten. Für Bereiche, in denen die Zusammenhänge zwischen Einzelitems und Schmerzfaktor wenig homogen oder gar invertiert sind (Atmung, Körperhaltung) erscheinen frei schätzbare Itemdiskriminationen dagegen eher notwendig. Es darf jedoch angenommen werden, dass auch diese Ausdrucksbereiche nach der Bereinigung um einzelne besonders wenig trennscharfe Indikatoren ebenfalls als 1PL-Modell geschätzt werden könnten.

Die ermittelten Itemschwierigkeiten stellen die durch die BESD-Originalskala implizierte gewichtete Verrechnung zu einem Summenscore in den meisten Ausdrucksbereichen in Frage. Die Verteilung einfacher und schwerer Items innerhalb der einzelnen Ausdrucksbereiche erscheint darüberhinaus vielfach unausgewogen. Selbst dort, wo die Items offensichtlich in einen niedrigen und einen hohen Abschnitt des latenten Schmerzmerkmals fallen, wie durch die Punktevergabe impliziert, klaffen entsprechende Lücken im Indikationsbereich mittlerer Schmerzausprägung.

6.4.4.2 Messstruktur der BESD-Gesamtskala

In einem engeren Sinne interessieren die zuvor separat modellierten – für bestimmte Ausdrucksbereiche spezifischen – Schmerzfaktoren nur mittelbar, da gegenwärtig der weitaus größte Teil der Theoriebildung und Testentwicklung als Zielmerkmal lediglich ein vergleichsweise unspezifisches Schmerzerleben berücksichtigt. Im Fokus der Aufmerksamkeit steht demnach ein latentes Merkmal Schmerz, das seinen Ausdruck ggfs. unterschiedlich deutlich in allen unterschiedenen Verhaltensbereichen findet. Im folgenden Auswertungsschritt wurde darum eine hierarchische Messstruktur für die BESD-Skala geschätzt.

Die bisher gewählten Modellierungen zur BESD lassen den Ausdrucksbereich Trost jeweils unberücksichtigt. Die Einschätzung der Tröstbarkeit geschieht im Rahmen der BESD-Erfassung in drei Stufen. Dabei soll zunächst das schmerzbezogene Ausdrucksverhalten in der beschriebenen Situation und über die definierte Zeitspanne nicht-teilnehmend beobachtet werden. Auf der Grundlage des beobachteten Verhaltens bzw. Schmerzausdruckes soll in einem zweiten Schritt entschieden werden, ob ein Trösten des Bewohners nötig ist. Nur für diejenigen Bewohner, die nach Einschätzung der Pflegenden getröstet werden sollten, geschieht anschließend eine Minimalintervention (z.B. Berühren oder andersweitig Ablenken), deren Erfolg oder Misserfolg als weiterer Indiaktor gewertet wird. Aufgrund dieser formal-logisch deutlich anderen Struktur des Erfassungsbereiches Trost und die sich daraus ergebenden linearen Abhängigkeiten sind die Möglichkeiten für eine gemeinsame Berücksichtigung aller BESD-Einzelindikatoren eingeschränkt. Um dennoch eine Abschätzung des Vorhersage bzw- Indikationsbeitrages dieses Ausdrucksbereiches leisten zu können, soll der Erfassungsbereich Trost im zu schätzenden BESD-Gesamtmodell als einzelnes dreistufiges kategorielles Item berücksichtigt werden.

Dabei reicht allerdings die Rechenleistung der zur Verfügung stehenden PCs offenbar nicht aus, um alle fünf Ausdrucksbereiche unter Beibehaltung aller Spezifikationen für das verwendete Schätzverfahren gemeinsam zu modellieren. Der Rechenaufwand steigt mit der Anzahl berücksichtigter latenter Merkmalsdimensionen mit dichotomen oder kategoriellen Indikatoren, über die hinweg Integrale gebildet werden müssen, expotenziell an. Um den Rechenaufwand für die notwendige numerische Integration zu senken, wurden darum den Empfehlungen von Muthén und Muthén folgend, die Anzahl der im Programmpaket M*plus* berechneten Integrationspunkte von 15 (Standard) auf 10 reduziert, womit sich insgesamt 5^{10}=9.765.625 Integrationspunkte ergeben (Muthén & Muthén, 1998-2006, p. 296). Damit werden bei der Abschätzung der Parameter in jedem Iterationsschritt weniger Punkte auf den jeweiligen latenten Merkmalsfaktoren (Dimensionen der Integration) differenziert, wodurch die Parameterschätzungen unter Umständen weniger präzise ausfallen könnten als bei den vorangegangenen separaten Analysen für die einzelnen BESD-Ausdrucksbereiche.

Die Modellierung des Second Order Factor Modells dient vorrangig der Abschätzung des Potenzials einzelner Ausdrucksbereiche, einen gemeinsamen generellen Schmerzfaktor abzubilden, sowie der schmerzspezifischen (i.S. des indirekten Einflusses dieses Generalfaktors) und verschiedenen schmerzunspezifischen Varianzanteile der Einzelitems dieser Indikatorbereiche. Da hierbei die empirischen Zusammenhänge zwischen allen 22 Einzelindikatoren modelliert werden, sind auch innerhalb der Indikatorbereiche Veränderungen in den itemspezifischen Threshold- und Regressionsparametern möglich. Anstatt die Diskussion des vorangegangenen Analyseschrittes in allen Einzelheiten zu wiederholen, sollen hier lediglich die neuen Modellparameter der zweiten Ebene der Messtruktur (Regressionsgewichte der Einzelbereiche für den Generalfaktor) und ggfs. substanzielle Veränderungen in den zuvor bereits beschriebenen Modellparametern berichtet werden.

BESD-Gesamtskala – Atmung

Der Ausdrucksbereich veränderter Atmung erscheint mit einem geschätzten standardisierten Regressionsgewicht von λ=.67 am wenigsten gut das gemeinsame Schmerzmerkmal abbilden zu können, und dementsprechend muss über die Hälfte der mit BESD erfassten Variabilität im Atmungsverhalten (56%) als entweder schmerzunspezifisch, oder aber zumindest hochgradig bereichsspezifisch gelten. Der Charakter des gemeinsam modellierten Atmungsfaktors verändert sich gegenüber der zuvor berichteten separaten Modellierung nur geringfügig. Deutlich geringer wird nun das Regressionsgewicht für die Cheyne Stokes Atmung (λ=–.13) geschätzt. Lautstark angestrengtes Atmen ist nun mit λ=.58 etwas stärker mit der schmerzbezogenen Atmungskomponente assoziiert als gelegentlich angestrengtes Atmen (.53). Die Möglichkeit der unterschiedenen Atmungsindikatoren, eine unspezifische Schmerzbelastung anzuzeigen, wie sie durch den Generalfaktor der zweiten Ebene der Messstruktur repräsentiert wird, kann als indirekter Effekt anhand des Produkts der beiden beteiligten Regressionsparameter abgeschätzt werden. Während weniger als 1 Prozent der Varianz der Cheyne Stokes Atmung durch eine allgemeine Schmerzbelastung bestimmt scheint, die auch durch die anderen Ausdrucksbereiche gemessen werden kann, beträgt dieser Anteil bei kurzen Phasen von Hyperventilation immerhin 39 Prozent.

Die Thresholdparameter der einzelnen Atmungsindikatoren unterscheiden sich nicht wesentlich von den zuvor berichteten Schwierigkeiten. Dieser Befund trifft auch auf alle weiteren Ausdrucksbereiche zu, weswegen diese weitestgehend äquivalent geschätzten Parameter an dieser Stelle nicht nochmals besprochen werden sollen.

BESD-Gesamtskala – Lautäußerung

Der durch die fünf Items aus dem Bereich der Lautäußerungen angezeigte Ausdrucksfaktor der BESD-Skala ist mit einem geschätzten Regressionsparameter von λ=.93 sehr eng mit dem gemeinsamen generellen Schmerzmerkmal verknüpft. Nahezu alle (d.h. genaugenommen 86,5 Prozent der) Unterschiedlichkeit in der Lautäußerung kann damit auf unterschiedliche allgemeine Schmerzniveaus zurückgeführt werden. Die Einzelindikatoren erscheinen damit vergleichsweise wenig spezifisch für eine bestimmte Form oder Qualität von Schmerz, die ihren Ausdruck nicht auch in den verbleibenden Indikations- bzw. Kommunikationsbereichen finden könnte. Die Bestimmung der Itemschwierigkeiten und -diskriminationen im Rahmen des Gesamtinstrumentes führt zu im Vergleich zur zuvor berichteten separaten Modellierung für die Einzelitems wiederholt beunruhigt rufen (λ=.69 vs. .36) und laut stöhnen oder ächzen (λ=.69 vs. .90) gegenläufig abweichenden Parameterschätzungen für die Itemdiskrimination. Offensichtlich verändert sich in der Abgrenzung gegenüber den weiteren BESD-Einzelindikatoren und -Indikationsbereichen auch der inhaltliche Charakter des Ausdrucksbereiches Lautäußerung etwas.

BESD-Gesamtskala – Mimik

Wenn man einen messbaren allgemeinen Schmerzfaktor annehmen kann, so erscheint dieser am deutlichsten durch die beobachtete Mimik der demenzkranken Bewohner angezeigt. Der durch die BESD-Mimikitems abgedeckte Verhaltensausdruck lässt sich rechnerisch mit dem durch alle Ausdrucksbereiche gemeinsam angezeigten Erlebenszustand gleichsetzen (λ=1). Als bester Einzelindikator für das allgemeine Schmerzerleben demenzkranker Bewohner in einer Ruhesituation kann damit ein ängstlicher Gesichtsausdruck gelten. Die Binnenstruktur des Ausdrucksbereiches Mimik der BESD erscheint deutlich verändert, wenn alle Einzelindikatoren parallel berücksichtigt werden. Während ein sorgenvoller Blick nunmehr weniger eng an den gemeinsamen Faktor gebunden scheint (λ=.39), werden die Regressionsparameter der Items ängstlicher Gesichtsausdruck und Grimassieren mit .87 bzw. .46 deutlich höher geschätzt.

BESD-Gesamtskala – Körperhaltung

Der in sich vergleichsweise heterogene Schmerzausdrucksbereich der Körperhaltung ist mit einem standardisierten Pfadkoeffizienten von λ=.99 ebenfalls nahezu mit dem gemeinsam angezeigten latenten Schmerzfaktor identisch. Lediglich ein verschwindend geringer Anteil der durch den Bereichsfaktor abgebildeten Variation in der Gestik erscheint damit als bereichsspezifischer (Schmerz-)Ausdruck, während ein Großteil der bei den Bewohnern beobachteten Unterschiedlichkeit in der Körperhaltung durch ein Schmerzmerkmal bestimmt scheint, dass seinen Ausdruck auch in den verbleibenden Verhaltenskategorien finden sollte. Im direkten Vergleich zur vorangegangenen separaten Analyse erscheint die knappe Mehrzahl der Indikatoren nunmehr enger an den gemeinsamen Bereichsfaktor geknüpft (z.B. nervös hin und hergehen .65 vs. .25; angespannte Körperhaltung .59 vs. .48; Nesteln .48 vs. .39). Zu denjenigen Indikatoren, die nunmehr weniger stark den gemeinsamen Gestikfaktor bestimmen, sind eine starre Körpersprache, angezogene Knie und geballte Fäuste zu zählen. Die Bedeutsamkeit des Indikators sich entziehen oder wegstoßen konnte in der gemeinsamen Analyse aufgrund von rechnerischen Schätzproblemen nicht nachvollzogen werden. Aufgrund der dominierenden Rolle, den dieser Indikator bei den vorangegangenen separaten Analysen jedoch zu spielen schien, ist davon auszugehen, dass Widerstand und Entzug basale schmerzbezogene Verhaltensreaktionen darstellen.

BESD-Gesamtskala – Trost

Aufgrund der linearen Abhängigkeiten zwischen den dichotomen Indikatoren Trostwunsch und Trosterfolg wurde dieser Bereich durch einen dreistufigen Verhaltensindikator angezeigt, für den ebenfalls ein entsprechender latenter Responsefaktor mit entsprechend zwei Schwellenparametern geschätzt wurde. Der latente Responsefaktor Schmerz ist mit einem Regressionsgewicht von λ_s=.63 mit dem generellen latenten Schmerzfaktor in Ruhe verknüpft. Sowohl der Wunsch, den Bewohner zu trösten, als auch der Erfolg einer entsprechenden Intervention sind also substanziell auch durch die wahre Schmerzbelastung

der Bewohner bestimmt. Im Vergleich zu den anderen Bereichen des Schmerzausdruckes muss Trost jedoch als weniger reliabler Indikatorbereich angesehen werden. Die für den Bereich Trost geschätzten Thresholdparameter liegen mit τ_1=0,06 und τ_2=1,4 weit auseinander. Dies spricht zum einen dafür, dass Pflegende vergleichsweise früh das Bedürfnis entwickeln, den Bewohner zu trösten, und dass eine Ablenkung von den Schmerzen oder eine Linderung der Schmerzen selbst nur dann wenig Erfolg zeigt, wenn eine vergleichsweise schwere Schmerzbelastung vorliegt.

Zusammenfassend sind alle innerhalb der unterschiedenen Ausdrucksbereiche durch diskrete Verhaltensweisen angezeigte bereichsspezifischen Dimensionen des Schmerzausdrucks eng mit dem anzunehmenden gemeinsamen latenten Schmerzmerkmal verknüpft. Dabei erscheinen insbesondere mimische und vokale bzw. verbale Ausdrucksformen einen herausragenden Stellenwert in der Schmerzerkennung einzunehmen, während die Ausdrucksbereiche der Gestik, vor allem aber die Bereiche der Atmung und des Trostes offensichtlich auch größere spezifische Varianzanteile besitzen. Die inhaltliche Interpretation dieser spezifischen Anteile innerhalb der Körperhaltung, Atmung und Trost ist jedoch schwierig, da das eingesetzte Instrumentarium auf ein hinsichtlich Lokalisation oder sonstiger Qualitäten nicht näher spezifiziertes Schmerzerleben ausgerichtet ist. Auf bereits publizierte konkrete Hypothesen zu spezifischen Schmerzqualitäten, die sich beispielsweise selektiv nur in manchen BESD-Verhaltensbereichen äußern, nicht jedoch über andere Kommunikationswege mitgeteilt oder erfasst werden können, kann gegenwärtig leider nicht zurückgegriffen werden.

6.5 Verhaltensbeobachtung mit der CNPI-Skala

Auch für die *Checklist of Nonverbal Pain Indicators* (CNPI) sollen im Folgenden die zuvor für die BESD-Skala berichteten Analyseschritte nachvollzogen werden, bevor weitere psychometrische Spezialprobleme (Skalenvergleich, Ruhe vs. Aktivität und Demenzspezifität) adressiert werden.

6.5.1 Skalenaufbau und abgeleitete Deskriptoren

Auch die CNPI-Skala wurde in von der Originalversion abweichender Weise erfasst, indem alle in den Originalpublikationen beschriebenen Beispielitems als konkret zu beobachtende Verhaltensindikatoren vorgegeben wurden. Entsprechend kann auch hier der Originalscore der CNPI-Skala „nachgebaut“ werden, indem auf der Grundlage der beobachteten 15 Einzelindikatoren entsprechende Bereichsscores (0 bzw. 1 Punkt) gebildet werden. Im Rahmen dieser Arbeit sollen die Analysen dagegen auf Einzelitemebene dargestellt werden, um möglichst differenzierte Aussagen machen zu können.

6.5.2 Klassische Skalenanalyse des CNPI-Inventars

Ein erster deskriptiver Überblick über die empirischen Auftretenshäufigkeiten der aus der Originalversion der CNPI extrahierten Schmerzitems wird gefolgt von einer klassischen Skalenanalyse. Im Anschluss werden wiederum die Möglichkeiten einer detaillierteren Analyse der CNPI-Messstruktur mithilfe der Item-Response-Theorie ausgelotet und genutzt. Wie zuvor werden die komplexeren Analysen nurmehr für die Ruhesituation beschrieben.

6.5.2.1 Beobachtungsraten für Einzelindikatoren

Tabelle 15 gibt – analog zur zuvor für die BESD-Skala geleistete Ergebnisdarstellung – einen ersten Überblick über die dokumentierten Einzelindikatoren des CNPI-Instrumentariums in beiden Beobachtungssituationen.

In der Ruhesituation wurden für alle Bewohner insgesamt 292 schmerzbezogene Verhaltensweisen beobachtet und dokumentiert. Von den 15 unterschiedenen CNPI-Schmerzindikatoren konnten in der Ruhesituation damit durchschnittlich lediglich 1,5 Ausdrucksweisen beobachtet werden. Am häufigsten (bei mindestens jeder zehnten Person) wurden die Indikatoren zusammengekniffene Lippen, gerunzelte Augenbrauen, verzerrter Gesichtsausdruck, Stöhnen bzw. Ächzen, verbale Äußerungen von Unbehagen oder Schmerz und die Unfähigkeit, still zu halten, festgehalten. Am seltensten wurden dagegen zusammengebissene Zähne oder Schaukeln beobachtet.

Differenziert man die sechs in der Originalskala vorgesehenen Ausdrucksbereiche, wurden die Kategorie verbale Beschwerden und Gesichtsgrimassen mit 12,8 bzw. 12,2 Prozent im Durchschnitt am häufigsten beobachtet, gefolgt von nicht-verbalen Lautäußerungen (9,6%), Ruhelosigkeit (8,2%), Klammern (7,7%) und Reiben (6,7%). Auch wenn der direkte Vergleich beider Schmerzinventare Gegenstand eines eigenen Analyseschrittes ist, und hier nicht vorweggenommen werden soll, erscheinen die zwischen den Instrumenten einigermaßen vergleichbaren Ausdrucksbereiche Lautäußerung vs. verbale und nonverbale Beschwerden, Mimik vs. Gesichtsgrimassen und Körperhaltung vs. Klammern, Ruhelosigkeit und Reiben für die CNPI-Skala deutlich seltener beobachtet worden zu sein als für die BESD-Skala.

In der Aktivitätssituation wurden auch mit der CNPI im Vergleich zur Ruhebedingung mehr potenziell schmerzbezogene Ausdrucksweisen festgehalten (N=409 dokumentierte Verhaltensweisen). Doch auch in dieser Beobachtungsbedingung konnten mehr als ein Drittel der vorgeschlagenen Schmerzkennzeichen nur selten (bei weniger als 10 Prozent der Bewohner) beobachtet werden. Verbale Äußerungen von Unbehagen bzw. Schmerz, Stöhnen bzw. Ächzen und zusammengekniffene Lippen können hier bei jedem vierten Bewohner beobachtet werden, für jeden fünften Bewohner wurden gerunzelte Augenbrauen bzw. enggestellte Augen dokumentiert. Besonders kontextsensitiv sind dabei die CNPI-Einzelitems Stöhnen bzw. Ächzen, vokale Beschwerden und Klammern, deren Beobachtungsraten sich von der Ruhe zur Aktivitätssituation jeweils verdoppeln. Die Items Schau-

Tabelle 15: Beobachtete Indikatoren der CNPI bei geringer und hoher Aktivierung

		Ruhe				*Aktivität*			
Nr.	Item	N	$\%_P$	$\%_K$	$\%_S$	N	$\%_P$	$\%_K$	$\%_S$
	Vokale Beschwerden (nicht-verbal)								
1	Keuchen, Seufzen (keine verständl. Worte)	16	8,2	5,5	36,4	28	14,3	6,8	63,6
2	Jammern, Schreien (keine verständl. Worte)	14	7,2	4,8	38,9	22	11,2	5,4	61,1
3	Stöhnen, Ächzen	26	13,3	8,9	33,3	52	26,5	12,7	66,7
	Vokale Beschwerden (verbal)								
4	Worte, die Unbehagen oder Schmerz ausdrücken; Fluchen oder Protestrufe	25	12,8	8,6	30,9	56	28,6	13,7	69,1
	Gesichtsgrimassen								
5	gerunzelte Augenbrauen, enggestellte Augen	30	15,4	10,3	42,3	41	20,9	10,0	57,7
6	zusammengekniffene Lippen	42	21,5	14,4	46,7	48	24,5	11,7	53,3
7	heruntergefallener Kiefer	16	8,2	5,5	69,6	7	3,6	1,7	30,4
8	zusammengebissene Zähne, Zähneknirschen	8	4,1	2,7	34,8	15	7,7	3,7	65,2
9	verzerrter Gesichtsausdruck	23	11,8	7,9	41,1	33	16,8	8,1	58,9
	Klammern								
10	Krampfhaftes Anklammern	15	7,7	5,1	33,3	30	15,3	7,3	66,7
	Ruhelosigkeit								
11	ständiger oder immer wieder Lagewechsel	12	6,2	4,1	40,0	18	9,2	4,4	60,0
12	Schaukeln	5	2,6	1,7	71,4	2	1,0	0,5	28,6
13	Konstante/wiederkehr. Handbewegungen	17	8,7	5,8	58,6	12	6,1	2,9	41,4
14	Unfähigkeit, still zu halten	30	15,4	10,3	46,2	35	17,9	8,6	53,8
	Reiben								
15	Massage eines bestimmten Körperbereiches	13	6,7	4,5	56,5	10	5,1	2,4	43,5
	Gesamt	**292**		**100**		**409**		**100**	

$\%_P$=bewohnerbezogene -; $\%_K$=kennzeichenbezogene -; $\%_S$=situationsbezogene Beobachtungsrate.
Datenbasis: HILDE2 2006; N=195 (Ruhe), N=196 (Aktivität).

keln, konstante bzw. wiederkehrende Handbewegungen und Massage eines bestimmten Körperbereiches werden dagegen in der Aktivitätssituation etwas seltener beobachtet. Insgesamt kann in der Aktivitätsbedingung eine veränderte Rangfolge der durchschnittlichen Beobachtungsraten für die sechs unterschiedenen Ausdrucksbereiche festgestellt werden: Die prominenteste Ausdruckskategorie ist danach mit 28,6 Prozent der Bewohner verbale Beschwerden, gefolgt von nicht-verbalen Beschwerden (17,3%), Klammern (15,3%), Gesichtsgrimassen (14,7%), Ruhelosigkeit (8,6%) und schließlich Reiben mit einer Beobachtungsrate von 5,1 Prozent.

6.5.2.2 Itemschwierigkeit, Trennschärfe und interne Konsistenz

Um eine möglichst differenzierte Grundlage für die Diskussion der angestrebten vertiefenden Itemanalysen zu schaffen, werden auch für die CNPI-Items wie zuvor für die BESD-Indikatoren die Kennwerte der klassischen Skalenanalyse unter beiden Beobachtungsbedingungen auf der Ebene der konkret vorgegebenen Einzelitems dargestellt. Auch wenn dieses Durchdeklinieren verschiedener Erfassungsbereiche dem Leser ein nicht unerhebliches Maß an Konzentrationsvermögen abverlangt, werden so am ehesten die Vorteile einer verschiedene Beobachtungsbedingungen, Erfassungsbatterien und hierarchische Messstrukturen integrierenden Analysestrategie unmittelbar einsehbar.

Schwierigkeiten der Einzelitems

Da nahezu alle vorgegebenen CNPI-Einzelindikatoren sowohl in der Ruhesituation, als auch in Situationen höherer Bewohneraktivierung selten beobachtet werden konnten, bewegen sich auch die entsprechenden Itemschwierigkeiten mit Werten zwischen p_i=.01 und .29 in einem sehr niedrigen Bereich. Mit einer mittleren Itemschwierigkeit von $\bar{p}_i$=.10 (Ruhe) bzw. .14 (Aktivität) muss auch die CNPI-Skala als sehr schwierig beurteilt werden. Die Variabilität der Itemschwierigkeiten ist dabei vergleichsweise gering, sodass es auf der Grundlage dieses Instruments schwer fallen sollte, geringere oder mittlere Schmerzniveaus differenziert abzubilden.

Anhand der für die diskret erfassten Indikatoren geschätzten Schwierigkeiten kann auch die Angemessenheit der impliziten Gleichgewichtung der für alle sechs CNPI-Ausdrucksbereiche beschriebenen Beispielitems überprüft werden.

Von den drei in dieser Studie diskret erfassten Indikatoren des Bereiches nicht-verbale Beschwerden erscheint das Item Stöhnen bzw. Ächzen in beiden Beobachtungsbedingungen ungefähr doppelt so einfach wie die verbleibenden Items Keuchen bzw. Seufzen und Jammern bzw. Schreien. Damit aber ist zu erwarten, dass Personen mit zwar jeweils vergleichsweise starken, aber dennoch unterschiedlichen Schmerzausprägungen in diesem Bereich als belastet eingeschätzt werden.

Etwas deutlichere Schwierigkeitsunterschiede lassen sich für die insgesamt fünf differenzierten Indikatoren des Ausdrucksbereiches Gesichtsgrimassen feststellen. Zusammengekniffene Lippen, gerunzelte Augenbrauen bzw. enggestellte Augen und ein verzerrter

Gesichtsausdruck können dabei unter beiden Beobachtungsbedingungen etwas häufiger beobachtet werden als ein heruntergefallener Kiefer und zusammengebissene Zähne bzw. Zähneknirschen. Der hier zu vergebende Subskalenpunkt kann also ebenfalls von Personen mit leicht unterschiedlichen, im Allgemeinen aber stärkeren Schmerzausprägungen erreicht werden.

Zu einer ähnlichen Gesamteinschätzung gelangt man auch mit Blick auf die vier diskret erhobenen Verhaltensindikatoren des CNPI-Ausdrucksbereiches Ruhelosigkeit. Zwar sind hier alle Items sowohl in Ruhe als auch bei Aktivität vergleichsweise schwierig, doch auf diesem geringen Niveau können doch deutliche Schwierigkeitsunterschiede beschrieben werden. Beispielsweise war es in der Ruhesituation fünfmal, in der Aktivitätssituation sogar achtzehnmal so schwer, den Indikator Schaukeln zu beobachten als eine Unfähigkeit, still zu halten. Auch wenn diese drastischen Unterschiede im Wesentlichen auf die allgemein sehr hohen Itemschwierigkeiten zurückzuführen sind, bleibt doch festzuhalten, dass dem gleichen Skalenscore gegebenenfalls recht unterschiedliche Schmerzbelastungen zugrundeliegen.

Diskrimination und Skalenkonsistenz

Die 15 CNPI-Einzelitems müssen mit Trennschärfen zwischen r_{it}=.16 und .53 als nur wenig diskriminativ beurteilt werden. Lediglich jeweils ca. ein Drittel der Indikatoren erreicht in den unterschiedenen Beobachtungsbedingungen eine Item-Total-Korrelation von über .30 und liegt damit in einem mittleren Bereich. Entsprechend bleibt auch die Skalenreliabilität der CNPI mit Konsistenzkoeffizienten von KR-20=.67 bzw. .68 unter beiden Durchführungsbedingungen deutlich hinter den Erwartungen an das Instrumentarium zum Schmerzscreening zurück. Nicht zuletzt aufgrund der niedrigen Zusammenhänge zwischen Einzelitem und Gesamtscore sind auch bei der Eliminierung einzelner Items keine substanziellen Steigerungen der Skalenkonsistenz zu erwarten.

Gesamtskalenscore

Von den 15 vorgegebenen schmerzbezogenen Verhaltensweisen konnten im Mittel insgesamt 1,5$\pm$1,9 bzw. 2,1$\pm$2,1 Ausdrucksweisen in Ruhe bzw. Aktivität beobachtet werden. Der für die Aktivitätssituation gesteigerte Skalenscore kann auch für die CNPI-Skala als systematisch ausgewiesen werden (Wilk's λ=0,911, df=1, 190, p$<$.001). Der Anteil von Bewohnern ohne erkennbare Schmerzäußerungen betrug in der Ruhesituation knapp 40 Prozent, in der Aktivitätssituation dagegen nur knapp 30 Prozent. In beiden Situationen wurde das obere Drittel des theoretisch möglichen Wertebereiches von 0 bis 15 Punkten nicht besetzt, und muss der Skalenscore aufgrund der deutlichen Rechtsschiefe als nichtnormalverteilt betrachtet werden.

Eine kurze Rekapitulation der wesentlichen konzeptionellen Einschränkungen bei der Interpretation und Verallgemeinerung der Ergebnisse der an der klassischen Testtheorie orientierten Skalenanalyse wurde bereits zuvor im Anschluss an die Ergebnisdarstel-

Tabelle 16: Kennwerte (KTT) der Indikatoren der CNPI bei geringer und hoher Aktivierung

	Ruhe			*Aktivität*		
Item[1]	p_i	r_{it}	KR20$_c$	p_i	r_{it}	KR20$_c$
CNPI: Vokale Beschwerden (nicht-verbal)						
Keuchen, Seufzen (keine verständlichen Worte)	.08	.24	.66	.14	.22	.67
Jammern, Schreien (keine verständlichen Worte)	.07	.33	.66	.11	.32	.66
Stöhnen, Ächzen	.13	.39	.64	.27	.41	.64
CNPI: Vokale Beschwerden (verbal)						
Worte, die Unbehagen oder Schmerzen ausdrücken, Fluchen oder Protestrufe	.13	.23	.67	.29	.27	.66
CNPI: Gesichtsgrimassen						
gerunzelte Augenbrauen, eng gestellte Augen	.15	.33	.65	.21	.28	.66
zusammengekniffene Lippen	.22	.25	.67	.25	.24	.67
heruntergefallener Kiefer	.08	.24	.66	.04	.18	.67
zusammengebissene Zähne, Zähneknirschen	.04	.26	.66	.08	.38	.65
verzerrter Gesichtsausdruck	.12	.53	.62	.17	.33	.65
CNPI: Klammern						
Krampfhaftes Anklammern ans Bettgitter o.ä.	.08	.31	.66	.15	.35	.65
CNPI: Ruhelosigkeit						
ständiger oder immer wieder Lagewechsel	.06	.26	.66	.09	.41	.65
Schaukeln	.03	.19	.67	.01	.16	.67
Konstante oder wiederkehrende Handbewegungen	.09	.18	.67	.06	.21	.67
Unfähigkeit, still zu halten	.15	.33	.65	.18	.35	.65
CNPI: Reiben						
Massage eines bestimmten Körperbereiches	.07	.22	.67	.05	.23	.67
Kuder/Richardson's KR-20		.67			.68	
Gesamtscore CNPI[2]		195 1,5±1,9			196 2,1±2,1	

[1] Dichotome Antwortskala: 0=„nein“; 1=„ja“; [2] Summe beobachteter Indikatoren: N M±SD.
p_i=Itemschwierigkeit (Beobachtungsrate); r_{it}=Itemdiskrimination (punkt-biserielle Item-Total-Korrelation); KR20$_c$=Skalenkonsistenz bei Reduzierung um Item.
Datenbasis: HILDE2 2006; N=195 (Ruhe), N=196 (Aktivität).

lung der BESD-Skala geleistet und soll an dieser Stelle nicht wiederholt werden, da sich auch für die CNPI-Indikatoren selbstverständlich die selben prinzipiellen Interpretationsschwierigkeiten ergeben.

6.5.3 IRT-Analyse der CNPI-Schmerzindikatoren

Auch für die CNPI-Skala sollen zwei aufeinander aufbauende Analyseschritte nachvollzogen werden. Zunächst werden diejenigen Ausdrucksbereiche der CNPI, die durch mehrere Einzelindikatoren angezeigt werden, in separaten Modellen geschätzt und vor dem Hintergrund einer probabilistischen Testtheorie interpretiert. Im Anschluss daran sollen alle Einzelindikatoren in einem gemeinsamen hierarchischen Messmodell berücksichtigt, und wie zuvor die Indikationsgüte der Einzelitems und Bereichsfaktoren bestimmt werden.

Auch mit Blick auf die CNPI-Skala soll die Frage, ob das Inventar in Situationen geringer und hoher Aktiviertheit gleich bzw. gleich gut funktioniert in einem späteren eigenständigen Kapitel bearbeitet werden. Die folgende Darstellung der Ergebnisse der IRT-Analysen bezieht sich aufgrund des hohen Auflösungsgrades wiederum nur auf den in der Ruhebedingung beobachteten Schmerzausdruck.

6.5.3.1 Messstruktur der einzelnen CNPI-Ausdrucksbereiche

In Tabelle 17 sind - analog zur Ergebnistabelle 14 auf Seite 208 für die BESD-Skala - die wichtigsten psychometrischen Kennwerte der CNPI-Ausdrucksbereiche sowohl in der durch das Analysepaket M*plus* verwendeten LRV-Formulierung, als auch in der herkömmlichen IRT-Metrik dargestellt.

Mehrere diskrete Verhaltensweisen konnten für die vorliegende Untersuchung aus den Beschreibungen der Ausdrucksbereiche nichtverbale vokale Beschwerden, Gesichtsgrimassen und Ruhelosigkeit der CNPI-Originalskala gewonnen werden, während die verbleibenden drei Ausdrucksbereiche verbale vokale Beschwerden, Klammern und Reiben durch jeweils nur ein einzelnes dichotomes Item angezeigt werden. Letztere werden folglich erst im nächsten Analyseschritt, der gleichzeitigen Modellierung aller CNPI-Indikatoren, berücksichtigt.

CNPI-Ausdrucksbereiche – Nicht-verbale Beschwerden

Das mit einem Freiheitsgrad gerade identifizierte Messmodell weist insgesamt eine angemessene Passung an die empirischen Daten auf. Mit geschätzten Regressionsgewichten zwischen λ=.61 (Stöhnen/Ächzen) und .78 (Jammern/Schreien) erscheinen alle Indikatoren dieses Bereiches substanziell mit dem geschätzten Bereichsfaktor verknüpft, und die für die latenten Responsevariablen geschätzte Variabilität kann zu Anteilen zwischen 61 und 37 Prozent gebunden werden. Die IRT-Parametrisierung weist alle Items mit maximalen Steigungskoeffizienten zwischen a=0,8 und 1,3 als ähnlich trennscharf aus. Aufgrund der relativ einheitlich geschätzten Regressionsparameter kann erwartet werden, dass die

Itemzusammenhänge auch durch ein entsprechendes einparametrisches Messmodell ähnlich gut abgebildet werden können. Tatsächlich ist der χ^2-Differenzentest für beide genestete Modellversionen nicht signifikant ($-2ll_\Delta$=0,55, df$_\Delta$=2, p$<$.762).

Hinsichtlich der Schwierigkeit einzelner beschriebener CNPI-Indikatoren wird in der Originalskala keine Aussage getroffen. Die Beobachtung jedes einzelnen aufgeführten Verhaltensmerkmales kann zu einem Bereichsscoring von einem Punkt beitragen. Für die in dieser Studie gesammelten Daten können vergleichsweise einheitliche Itemschwierigkeiten zwischen b=1,8 und 2,0 nachgewiesen werden. Ihr Auftreten kennzeichnet danach also jeweils Schmerzen in einem vergleichsweise ähnlichen Ausprägungsgrad.

Tabelle 17: Kennwerte (IRT) der Indikatoren der CNPI Ausdrucksbereiche in Ruhe

		Separate 2PL-Modelle für Indikatorbereiche						
Nr.	Item	θ	τ_s	λ_s	ψ	b	a	Anpassung
	Vokale Beschwerden (nicht-verbal)							
1	Keuchen, Seufzen	0,51	1,37	.70	3,07	1,97	1,03	χ^2=2,5
2	Jammern, Schreien	0,39	1,45	.78		1,86	1,32	*df*=1
3	Stöhnen, Ächzen	0,63	1,07	.61		1,77	0,82	$p <$.113
	Vokale Beschwerden (verbal)[1]							
4	Worte, die Unbehagen oder Schmerz ausdrücken; Fluchen oder Protestrufe	–	–	–	–	–	–	–
	Gesichtsgrimassen							
5	gerunzelte Augenbrauen, enge Augen	0,72	0,97	.53		1,86	0,66	χ^2=20,9
6	zusammengekniffene Lippen	0,55	0,76	.67	1,25	1,15	0,95	*df*=21
7	heruntergefallener Kiefer	0,73	1,35	.52		2,61	0,65	$p <$.466
8	zus.gebissene Zähne, Zähneknirschen	0,23	1,73	.88		1,98	1,94	
9	verzerrter Gesichtsausdruck	0,33	1,17	.82		1,43	1,54	
	Klammern[1]							
10	Krampfhaftes Anklammern	–	–	–	–	–	–	–
	Ruhelosigkeit							
11	ständiger Lagewechsel	0,79	1,51	.46	0,87	3,31	0,55	χ^2=10,3
12	Schaukeln	0,42	1,95	.76		2,55	1,27	*df*=7
13	Konstante Handbewegungen	0,26	1,36	.86		1,57	1,81	$p <$.173
14	Unfähigkeit, still zu halten	0,52	1,00	.69		1,45	1,02	
	Reiben[1]							
15	Massage eines Körperbereiches	–	–	–	–	–	–	–

θ=Residualvarianz von y^*; τ_s=Threshold (standardisiert); λ_s=Regressionsparameter (standardisiert).
ψ=Faktorvarianz (unstandardisiert); b=Itemschwierigkeit; a=Itemdiskrimination.
[1] Die Indikatorbereiche Verbale Beschwerden, Klammern und Reiben sind mit nur einem einzelnen Indikator nicht separat schätzbar, da nicht identifiziert.
Datenbasis: HILDE2 2006; N=194 (Ruhe).

CNPI-Ausdrucksbereiche – Gesichtsgrimassen

Auch für die fünf unterschiedenen CNPI-Indikatoren aus dem Bereich der Mimik scheint ein eindimensionales Messmodell die wechselseitigen Zusammenhänge zwischen den Items angemessen zu repräsentieren. Die Regressionskoeffizienten werden zwischen λ=.52 (heruntergefallener Kiefer) bzw. .53 (gerunzelte Augenbrauen/eng gestellte Augen) und λ=.88 (zusammengebissene Zähne/Zähneknirschen) bzw. .82 (verzerrter Gesichtsausdruck) geschätzt, und entsprechend variieren die in der IRT-Metrik ausgedrückten Itemdiskriminationen zwischen a=0,6 und 1,9 deutlich. Trotz der z.T. um ein Mehrfaches steileren itemcharaktristischen Kurven für einzelne Items erscheint für diesen Ausdrucksbereich auch eine Repräsentation durch ein einfacheres einparametrisches Modell mit äquivalenten Diskriminationen ohne substanzielle Einbußen der allgemeinen Anpassungsgüte möglich ($-2ll_{\Delta}$=7,6, df$_{\Delta}$=4, p$<$.106).

Mit geschätzten Itemschwierigkeiten zwischen b=1,2 und 2,6 decken die Einzelitems zum mimischen Schmerzausdruck einen vergleichsweise breiten Bereich auf dem angenommenen latenten Schmerzkontinuum ab. Das wenig trennscharfe Item heruntergefallener Kiefer scheint dabei der schwierigste Indikator zu sein, während zusammengekniffene Lippen bereits in geringeren Ausprägungsgraden des Schmerzes erwartet werden können. Die doch deutlich heterogen geschätzten Schwierigkeiten lassen die Frage aufkommen, ob für den Bereich der Mimik nicht vielleicht eine aussagekräftigere bzw. präzisere Schmerzabschätzung erreicht werden könnte, wenn beim Scoring unterschiedliche Punktzahlen vergeben würden. Beim bisher vorgeschlagenen Scoring kann erwartet werden, dass Personen mit faktisch unterschiedlichen Schmerzausprägungen denselben Subskalenscore erreichen.

CNPI-Ausdrucksbereiche – Ruhelosigkeit

Auch die für den gestischen Ausdrucksbereich gefundenen empirischen Merkmalszusammenhänge scheinen durch das eindimensionale zweiparametrische Messmodell zufriedenstellend repräsentiert zu sein. Der geringste Regressionsparameter und entsprechend die unvollständigste Determination der latenten Responsevariable wird für das Item ständiger Lagewechsel geschätzt (λ=.46, θ=0,79). Die drei verbleibenden Items erscheinen deutlich enger mit dem angezeigten Bereichsausdrucksfaktor verknüpft, der am stärksten durch den Indikator konstante Handbewegungen charakterisiert ist (λ=.86). Die Diskriminationen liegen in einem Bereich zwischen a=0,6 und 1,8. Trotz der für einzelne Items geringer geschätzten Trennschärfe erscheint ein 1PL-Modell eine mit der vorgestellten zweiparametrischen Schätzung vergleichbar gute Abbildung der empirischen Datenstruktur zu erlauben ($-2ll_{\Delta}$=2,3, df$_{\Delta}$=3, p$<$.520).

Nicht zuletzt wegen der variierenden Diskriminationsfähigkeit der Items werden auch deutlich variierende Schwierigkeitsparameter geschätzt. Als leichte Items erscheinen dabei die Unfähigkeit, still zu halten und konstante Handbewegungen (b=1,5 bzw. 1,6). Ständiger Lagewechsel und Schaukeln können dabei erst bei einem zugrundeliegenden

höheren Schmerzniveau erwartet werden (b=3,3 bzw. 2,6). Damit aber stellt sich wiederum die Frage der optimalen Verwertung der zur Schmerzintensität gewonnenen Information durch das vorgeschlagene Bereichsscoring.

Zusammenfassend scheinen die durch die CNPI-Originalskala berücksichtigten diskreten Verhaltensweisen tatsächlich jeweils eine gemeinsame Ausdrucksqualität anzuzeigen. Die gefundenen mäßigen Unterschiede in der Diskriminationskraft der Einzelindikatoren ließen auch eine einfachere 1PL-Modellierung im Sinne eines Rasch-Modelles möglich erscheinen. Nicht in allen drei hier betrachteten Indikationsbereichen können alle Items als vergleichbar schwierig betrachtet werden, sodass das herkömmliche CNPI-Scoring verfeinert werden könnte.

6.5.3.2 Messstruktur der CNPI-Gesamtskala

Selbstverständlich kann sich eine Beurteilung der psychometrischen Güte der CNPI-Skala nicht ausschließlich auf die Hälfte der vorgesehenen Indikatorbereiche beschränken. Neben den Bereichen, für die jeweils durch mehrere Einzelitems ein gemeinsamer Ausdrucksbereichsfaktor angezeigt wird, sollen darum in einer weiterführenden Analyse auch solche Indikatorbereiche in das Messmodell integriert werden, die in der CNPI-Originalversion durch nur ein diskretes Verhaltensmerkmal beschrieben sind. Damit ergibt sich parallel zur BESD-Analyse auch für diese Skala eine hierarchische Messstruktur mit insgesamt sechs latenten Bereichsfaktoren und einem gemeinsamen Second Order Schmerzfaktor.

Die Ergebnisse dieses Analyseschrittes werden wie zuvor nicht tabellarisch dargestellt, sondern mit Blick auf die bereits berichteten bereichsspezifischen Analysen diskutiert.

CNPI-Gesamtskala – Nicht-verbale Beschwerden

Der durch alle sechs Indikatorbereiche gemeinsam angezeigte latente Schmerzfaktor kann mit einem standardisierten Regressionsgewicht von λ=.82 ungefähr zwei Drittel der im Ausdrucksbereich nicht-verbaler Beschwerden geschätzten Variabilität aufklären. Entsprechend gering ist der bereichsspezifische Anteil des schmerzbezogenen nicht-verbalen Ausdrucksverhaltens der CNPI-Skala. Die zuvor bereits berichtete konsistent hohe Bestimmtheit der drei Indikatoren nicht-verbaler Beschwerden mit dem gemeinsamen Faktor bestätigt sich auch in der gemeinsamen Analyse aller CNPI-Indikatoren. Dennoch findet sich ein leicht von der vorangegangenen Analyse abweichendes Einflussmuster. Während zuvor Jammern und Schreien die den gemeinsamen Faktor dominierenden Verhaltensweisen waren (λ=.72 vs. .78), erscheint nun das Stöhnen bzw. Ächzen als ähnlich bedeutsam (λ=.75 vs. .61), während Keuchen und Seufzen nun etwas weniger eng an den gemeinsamen Faktor gebunden scheinen (λ=.61 vs. .70).

Wie bereits zuvor für die entsprechende Analyse der BESD-Skala berichtet, werden auch bei einer alle CNPI-Items berücksichtigenden Analyse im wesentlichen identische

Schwellenwerte geschätzt. Um unnötige Redundanz zu vermeiden, sollen darum im Weiteren nur für die bislang noch nicht berücksichtigten Einzelitems entsprechende Schätzungen der Thresholdparameter bzw. Itemschwierigkeiten berichtet werden.

CNPI-Gesamtskala – Verbale Beschwerden

Sinnhaft verbale Lautäußerungen wurden in der vorliegenden Studie lediglich als Einzelitem berücksichtigt. Der für die dichotomen beobachteten Daten geschätzte latente Responsefaktor ist dabei direkt mit einem Regressionskoeffizienten von λ=.56 mit dem allgemeinen Second Order Schmerzfaktor verknüpft. Damit werden ca. 31 Prozent der Varianz dieser LRV als allgemeiner Schmerzausdruck interpretiert, während die verbleibende Untersciedlichkeit im so erfassten verbalen Ausdrucksverhalten als spezifischer (Schmerz)-Ausdruck gelten soll. Im Kanon aller sechs CNPI-Ausdrucksbereiche muss diese Facette potenziellen Schmerzausdrucks jedoch als vergleichsweise wenig reliabel gelten. Relativiert man den geschätzten Regressionskoeffizienten an der Quadratwurzel der Residualvarianz, kann mit a=0,68 eine vergleichsweise flache itemcharakteristische Funktion dieses Indikators über das latente allgemeine Schmerzkontinuum hinweg angenommen werden.

Der geschätzte Thresholdparameter liegt für diesen Indikator mit τ=1,1 im unteren Drittel aller Einzelitems. Relativiert man diesen Schwellenwert am geschätzten Einflussgewicht des Gesamtschmerzfaktors, so liegt die IRT-Itemschwierigkeit dieses Items mit b=1,95 in einem mit den verbleibenden Items vergleichbaren mittleren Bereich.

CNPI-Gesamtskala – Gesichtsgrimassen

Der durch insgesamt fünf Indikatoren angezeigte Faktor des mimischen Schmerzausdruckes erscheint mit einem geschätzten Regressionskoeffizienten von λ=.78 seinerseits ein guter Indikator für den letztendlich zu messenden generellen Schmerzfaktor zu sein. Ungefähr 61 Prozent der mit den CNPI-Items abgebildeten Variabilität im mimimischen Schmerzausdruck können durch diesen Faktor gebunden werden, während die verbleibenden 39 Prozent der Varianz des Bereichsfaktors offenbar einen mimikspezifischen Schmerzausdruck darstellen. Dabei wird im CNPI-Gesamtmodell ein merkbar verändertes Muster von Regressionskoeffizienten gefunden als für die separate Schätzung dieses Einzelbereiches berichtet. Dabei werden für das Item gerunzelte Augenbrauen bzw. enggestellte Augen geringfügig (λ=.57 vs. .53), für das Item verzerrter Gesichtsausdruck (.93 vs. .82) höhere Diskriminationen geschätzt. Die Bedeutung der verbleibenden Indikatoren zusammengekniffene Lippen, heruntergefallener Kiefer und zusammengebissene Zähne bzw. Zähneknirschen für den im Ensemble geschätzten Mimikfaktor dagegen ist z.T. deutlich reduziert (λ=.57, .49 und .77).

CNPI-Gesamtskala – Klammern

Der dichotome Einzelindikator krampfhaftes Anklammern kann mit einem Regressionskoeffizienten von λ=.65 durch den generellen Second Order Schmerzfaktor vorher-

gesagt werden, wodurch ungefähr 42 Prozent der geschätzten Varianz der für diesen kategoriellen Indikator modellierten latenten Responsevariable erklärt werden können. Der nachträglich berechnete IRT-Diskriminatonskoeffizient beträgt a=1,93 und liegt damit im oberen Bereich der zuvor für die Einzelbereiche gefundenen Trennschärfeindizes.

Für die latente Responsevariable des Skalenitems Klammern wurde ein Schwellenparameter von τ=1,40 geschätzt, der in der üblichen IRT-Metrik einer Itemschwierigkeit von b=2,16 entspricht. Krampfhaftes Anklammern gehört damit zu den in der Ruhesituation relativ schwierigen Schmerzindikatoren.

CNPI-Gesamtskala – Ruhelosigkeit

Der durch die Gestikitems angezeigte Ausdrucksfaktor erscheint mit einem geschätzten Regressionskoeffizient von λ=.63 mäßig eng mit dem letzten Endes interessierenden allgemeinen Schmerzfaktor verknüpft. Nahezu 60 Prozent der in diesem Bereichsfaktor abgebildeten Unterschiedlichkeit in der Gestik können bestenfalls als bereichsspezifischer Schmerzausdruck interpretiert werden, der nicht auch durch Verhalten aus anderen Ausdrucks- bzw. Kommunikationsbereichen nachvollzogen werden kann.

Die Binnenstruktur des Ausdrucksbereiches Ruhelosigkeit ist gegenüber der separaten Schätzung des letzten Analyseschrittes dahingehend verändert, dass die Items ständiger Lagewechsel und Unfähigkeit still zu halten nunmehr deutlich wichtiger für die Charakterisierung des Gestikfaktors sind (λ=.57 vs. .46 bzw. .79 vs. .69), während konstante Handbewegungen an Bedeutung verlieren (λ=.73 vs. .86).

CNPI-Gesamtskala – Reiben

Die geschätzte latente Responsevariable für den dichotomen CNPI-Einzelindikator Massage eines Körperbereiches ist der mit einem Regressionskoeffizienten von λ=.55 auf Bereichsebene am wenigsten diskriminative Indikator für den übergeordneten generellen Schmerzfaktor. Nahezu 70 Prozent der für dieses Verhaltensmerkmal geschätzten Variabilität kann demnach nicht systematisch mit dem in den anderen Ausdrucksbereichen beobachteten Schmerzausdruck in Verbindung gebracht werden. Der nachträglich in die IRT-Metrik transformierte Diskriminationsindex beträgt a=0,66 und zeigt eine entsprechend flache itemcharakteristische Kurve über den latenten Schmerzfaktor hinweg an.

Der ermittelte Thresholdparameter liegt mit τ=1,47 im Vergleich mit den verbleibenden Einzelindikatoren in einem mittleren Bereich. Durch die Relativierung am vergleichsweise niedrigen Regressionskoeffizienten kann der IRT-Parameter für die Itemschwierigkeit mit b=2,66 jedoch durchaus als hoch gelten.

Zusammenfassend bestätigen die Ergebnisse dieses Analyseschrittes die theoretischen Annahmen zur hierarchischen Messstruktur der CNPI-Indikatoren. Sowohl die durch mehrere diskrete Verhaltensindikatoren beschriebenen Ausdrucksbereiche, als auch die berücksichtigten Einzelindikatoren weisen substanzielle Zusammenhänge auf, die als genereller

Schmerzfaktor interpretiert werden sollen. Die differenzierter erfassten Ausdrucksbereiche, und hier vor allem nicht-verbale Beschwerden und Gesichtsgrimassen, erscheinen als reliablere Indikatorbereiche als die lediglich durch ein Einzelitem erfassten Ausdrucksqualitäten.

Durch die unterschiedliche Zuordnung einzelner konkreter Verhaltensweisen zu unterschiedlichen Hierarchieebenen des Messmodelles wird jedoch ein direkter Vergleich der Itemparameter erschwert, insbesondere wenn dieser in der herkömmlichen IRT-Metrik erfolgen soll. Der Einfluss des generellen Schmerzfaktors auf die Einzelindikatoren eines Ausdrucksbereiches der durch einen Bereichsfaktor repräsentiert wird, kann als indirekter Effekt, und damit als Produkt der entsprechenden Regressionskoeffizienten auf beiden Hierarchieebenen berechnet werden. Die jeweils unvollständigen Zusammenhänge auf diesen Ebenen führen entsprechend zu vergleichsweise geringen Diskriminationen und damit Reliabilitäten dieser Einzelitems für den interessierenden übergeordneten Schmerzfaktor.

Nicht alle Erkenntnisse des vorangegangenen Analyseschrittes, insbesondere hinsichtlich der essentiellen Gleichheit der Trennschärfen der Einzelitems in den Ausdrucksbereichen, wurden für die Spezifikation und Schätzung des CNPI-Gesamtmodelles genutzt. Sowohl inhaltlich als auch technisch stellen die vorgestellten Analysen gewissermaßen einen Grenzpunkt dar, ab dem weiterführende Fragestellungen (beispielsweise ein direkter Vergleich beider Instrumente durch eine gemeinsame second-order-factor Modellierung) nur bei einer entsprechenden Reduktion des Auflösungsgrades (hinsichtlich der Modellparameter oder der faktoriellen Struktur) handhabbar erscheinen.

6.6 Vergleich der BESD- und CNPI-Schmerzskalen

Ein direkter Vergleich der für die BESD- und CNPI-Inventare ermittelten herkömmlichen Reliabilitätsindikatoren (siehe Tabelle 13 und 16) wird durch die unterschiedliche Testlänge (24 vs. 15 Einzelitems) erschwert. Auch wenn die Itemanzahl der jeweiligen Originalformate sich nicht ganz so stark unterscheiden, blieben auch hier die Konsistenzkoeffizienten aufgrund der sehr unterschiedlichen Binnenstruktur der Skalen (fünf dreistufige Indikatorbereiche mit impliziter Ungleichgewichtung der dichotomen Einzelindikatoren vs. sechs dichotome Indikatorbereiche mit impliziter Gleichgewichtung der dichotomen Einzelitems) nur schwer vergleichbar. So bleibt unklar, ob die für die jeweiligen Gesamtscores beider Skalen ähnlich ausfallenden Konsistenzkoeffizienten als Indikatoren vergleichbarer Testverlässlichkeit interpretiert werden sollten. Insofern als in die Berechnung der BESD-Scores mehr Items (und für die Originalfassung in einer differenzierteren Weise) einfließen als in die der CNPI-Scores, ließen sich die für die BESD-Skala ermittelten etwas höheren Konsistenzkoeffizienten bereits a priori erwarten.

In den vorangegangenen Kapiteln wurde die Angemessenheit der impliziten Annahmen zur jeweiligen Binnenstruktur und Funktionsweise beider Inventare unabhängig voneinander empirisch überprüft. Der prinzipiellen Logik dieser detaillierten Einzelanalysen folgend sollten für einen direkten Vergleich beider Schmerzinventare alle Einzelitems ge-

meinsam modelliert werden, um die Itemparameter auf einer gemeinsamen Skala interpretieren zu können. Die auf einem gemeinsamen latenten Merkmalsfaktor skalierten Inventare können anschließend mit Blick auf ihren jeweiligen Abbildungsbereich, d.h. die Lokalisation und Streuung der Itemschwierigkeiten und den über diesen Bereich hinweg anzunehmenden Informationsgehalt, der durch die Itemanzahl und deren Diskriminationskraft bestimmt ist, miteinander verglichen werden.

In der nachfolgenden Tabelle 18 sind die Parameterschätzungen eines gemeinsamen eindimensionalen 2PL-Modells für die Einzelindikatoren der BESD- und CNPI-Skala dargestellt. Zur leichteren Orientierung ist auch hier ein Piktogramm des geschätzten Analysemodells in grauer Farbe in die Tabelle eingebunden. Wie bereits bei den separaten BESD-Auswertungen diskutiert, lässt sich der Indikatorbereich Trost am besten als dreistufiger Indikator in die Messstruktur einbinden. Berichtet werden neben den für das Strukturgleichungsmodell mit kategoriellen Indikatoren geschätzten LRV-Parametern wiederum die üblichen IRT-Itemparameter Diskriminationsfähigkeit (a) und Schwierigkeit (b).

6.6.1 Modellanpassung

Aufgrund der vorangegangenen Analysen ist zu erwarten, dass nicht alle Einzelitems gute Indikatoren eines gemeinsamen latenten Konstruktes darstellen, so dass das postulierte Messmodell in dieser Hinsicht nicht optimal auf die Daten passen kann. Da aber ein Vergleich der beiden Instrumente möglichst alle in der jeweiligen Itembeschreibung angesprochenen konkreten Verhaltensindikatoren mit einbeziehen sollte, wird diese bereits zuvor beschriebene Fehlpassung hier bewusst in Kauf genommen.

Aufgrund der großen Zahl dichotomer Einzelindikatoren bzw. Indikatorstufen und der sich daraus ergebenden Matrix von Kombinationen ist ein χ^2-Anpassungstest, wie er für die einzelnen Indikationsbereiche der Skalen zuvor noch berichtet wurde, nun nicht mehr verfügbar.

6.6.2 Ein- vs. zweiparametrische Modellierung

In den vorgeschalteten Analysen für die Indikatorbereiche der Skalen konnten Hinweise darauf gefunden werden, dass die Zusammenhänge zwischen den Indikatoren auch durch ein einfacheres Rasch-Modell mit lediglich unterschiedlichen Itemschwierigkeiten, jedoch gleich zu schätzenden Diskriminationen repräsentiert werden könnten.

Ein Vergleich der Modellanpassung der kombinierten Messstruktur beider Skalen (unter Vernachlässigung der mittleren Ebene bereichsspezifischen Schmerzausdruckes) jedoch macht deutlich, dass die geschätzten Regressionsparameter und damit die Itemdiskriminationen so unterschiedlich sind, dass eine Gleichrestriktion nicht angemessen erscheint ($-2ll_{\Delta}$=70,5, df_{Δ}=36, $p<.001$).

Für eine möglichst anschauliche Diskussion der Analyseergebnisse, und insbesondere um die direkte Gegenüberstellung der BESD- und CNPI-Skalen zu erleichtern, werden die in

Tabelle 18: Kennwerte (IRT) der BESD- und CNPI-Indikatoren bei gemeinsamer Analyse in Ruhe

Skala			*2PL-Modell BESD+CNPI*				
Ausdrucksbereich	Nr.	Item	θ	τ_s	λ_s	a	b
BESD							
Atmung	1	gelegentlich angestrengt atmen	0,93	0,6	.27	0,28	2,37
	2	lautstark angestrengt atmen	0,84	2,0	.40	0,44	4,92
	3	kurze Phasen von Hyperventilation	0,52	1,7	.69	0,96	2,45
	4	lange Phasen von Hyperventilation	0,83	2,2	.41	0,46	5,40
	5	Cheyne Stokes Atmung	1,00	0,9	.04	0,04	22,31
Lautäußerungen	6	gelegentlich stöhnen oder ächzen	0,81	0,5	.44	0,49	1,10
	7	leise missbilligend/negativ äußern	0,73	0,8	.52	0,60	1,55
	8	wiederholt beunruhigt rufen	0,56	1,1	.66	0,88	1,60
	9	laut stöhnen oder ächzen	0,43	1,6	.76	1,16	2,09
	10	Weinen	0,46	1,3	.73	1,07	1,81
Mimik	11	trauriger Gesichtsausdruck	0,84	0,4	.40	0,43	0,87
	12	ängstlicher Gesichtsausdruck	0,41	1,0	.77	1,20	1,29
	13	sorgenvoller Blick	0,84	0,5	.40	0,43	1,20
	14	Grimassieren	0,68	1,2	.56	0,68	2,08
Körperhaltung	15	angespannte Körperhaltung	0,73	0,6	.52	0,61	1,13
	16	nervös hin- und hergehen	0,64	1,2	.60	0,75	1,91
	17	Nesteln	0,74	0,9	.51	0,59	1,79
	18	Körpersprache starr	0,93	0,8	.27	0,28	2,97
	19	geballte Fäuste	0,61	1,5	.62	0,79	2,42
	20	angezogene Knie	0,91	1,1	.29	0,31	3,89
	21	sich entziehen oder wegstoßen	0,19	1,7	.90	2,07	1,85
	22	Schlagen	0,43	1,8	.75	1,15	2,40
Trost[1]	38	Bedürfnis zu Trösten	0,59	0,1	.64	0,84	0,10
	39	Trösten nicht möglich	0,59	1,4	.64	0,84	2,11
CNPI							
Nonverb. Beschwerde	1	Keuchen, Seufzen	0,70	1,4	.55	0,65	2,47
	2	Jammern, Schreien	0,46	1,4	.74	1,09	1,93
	3	Stöhnen, Ächzen	0,59	1,1	.64	0,84	1,68
Verbale Beschwerde	4	Worte des Unbehagens o. Schmerzes	0,62	1,1	.62	0,78	1,79
Gesichtsgrimassen	5	gerunzelte Augenbrauen, enge Augen	0,77	1,0	.48	0,55	2,02
	6	zusammengekniffene Lippen	0,75	0,8	.50	0,58	1,49
	7	heruntergefallener Kiefer	0,86	1,3	.37	0,40	3,60
	8	zus.gebissene Zähne, Zähneknirschen	0,66	1,7	.58	0,72	2,93
	9	verzerrter Gesichtsausdruck	0,41	1,2	.77	1,20	1,52
Klammern	10	Krampfhaftes Anklammern	0,40	1,4	.77	1,22	1,80
Ruhelosigkeit	11	ständiger Lagewechsel	0,60	1,5	.63	0,81	2,38
	12	Schaukeln	0,59	1,9	.64	0,84	3,00
	13	Konstante Handbewegungen	0,81	1,3	.43	0,48	3,00
	14	Unfähigkeit, still zu halten	0,64	1,0	.60	0,74	1,65
Reiben	15	Massage eines Körperbereiches	0,79	1,5	.46	0,51	3,20

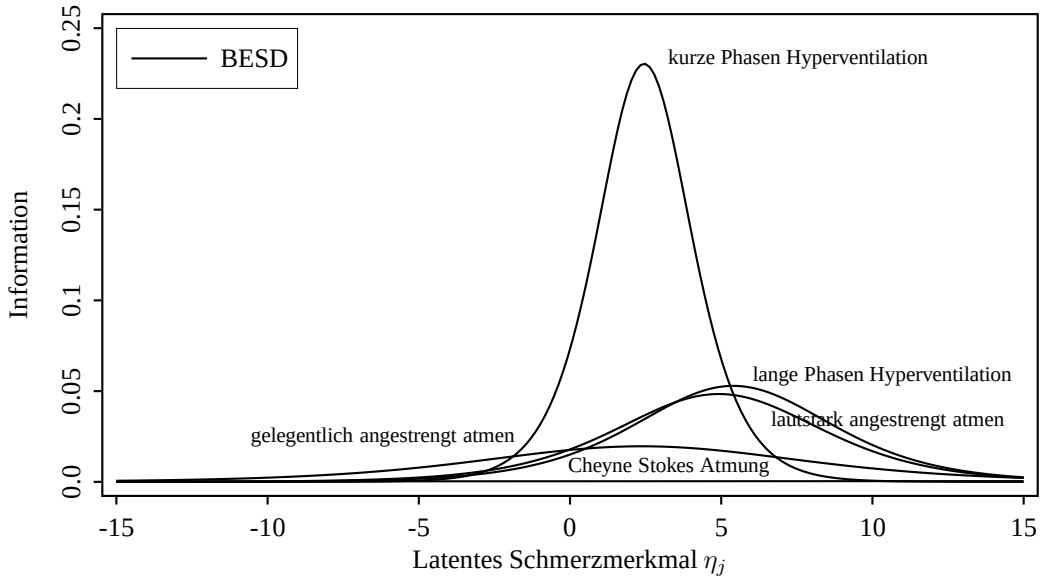
θ=Residualvarianz von y^*; τ_s=Threshold (standardisiert); λ_s=Regressionsparameter (standardisiert).
b=Itemschwierigkeit; a=Itemdiskrimination. Die Varianz des latenten Faktors beträgt ψ=0,258.
[1] Der Indikatorbereich Trost wurde als dreistufiges kategorielles Item berücksichtigt.
Datenbasis: HILDE2 2006; N=194 (Ruhe).

Tabelle 18 aufgeführten Itemkennwerte graphisch aufbereitet. Für inhaltlich vergleichbare Indikationsbereiche, z.B. Mimik im BESD und Gesichtsgrimassen im CNPI, werden darum jeweils die Informationsfunktionen (vgl. Abb. 12) der zugeordneten Items gemeinsam geplottet. Der Informationsgehalt eines Items erreicht sein Maximum dabei in demjenigen Punkt auf dem latenten Merkmalskontinuum, an dem die Wahrscheinlichkeit für die Beobachtung dieses Items p=0.5 beträgt, also im Wendepunkt seiner itemcharakteristischen Funktion bzw. bei seiner Itemschwierigkeit. Wie bereits in Gleichung 33 dargestellt, ergibt sich der maximale Informationswert jeweils als ein Viertel des Quadrates des Diskriminationsparameters. Die grafische Gegenüberstellung der inhaltsähnlichen Verhaltensindikatoren beider Inventare ergänzt den tabellarischen Ergebnisbericht und bereitet den Vergleich des Aussagebereiches und der Effizienz der beiden Schmerzinventare vor.

6.6.3 Skalenvergleich – Ausdrucksbereich Atmung

Der Ausdrucksbereich der Atmung wird lediglich vom BESD-Instrument abgedeckt, weswegen alle in Abbildung 28 dargestellten Informationsfunktionen mit durchgezogenen Linien dargestellt sind. Im Bereich der Atmung wird mit a=0,96 die höchste Diskrimination für den Indikator kurze Phasen von Hyperventilation geschätzt, weswegen für die Ordinate ein Bereich von Informationswerten zwischen I=0,0 und 0,25 gewählt wurde.

Abbildung 28: Informationsgehalt der Einzelitems des BESD-Skalenbereiches Atmung.



Das geschätzte latente Merkmal Schmerz, das hier auf der Abszisse abgetragen ist, besitzt keine natürliche Skalierung. Der Nullpunkt repräsentiert darum die mittlere wahre Schmerzbelastetheit der untersuchten HILDE-Bewohnerstichprobe.

Zunächst ist leicht ersichtlich, dass alle Indikatoren den Großteil ihrer Informativität in Bereichen gesteigerten Schmerzes entwickeln, insgesamt also relativ schwierig sind. Während gelegentlich angestrengtes Atmen und kurze Phasen der Hyperventilation auch unter der Bedingung leicht überdurchschnittlichen Schmerzes ihren (wenngleich sehr verschieden großen) Beitrag zur Schmerzmessung leisten, zeigen lange Phasen von Hyperventilation und lautstark angestrengtes Atmen stärkere Schmerzzustände an.

Der Informationsgehalt des Indikators Cheyne Stokes Atmung ist bei einer gemeinsamen Schätzung des latenten Schmerzfaktors durch alle BESD- und CNPI-Items praktisch gleich Null, was an einer nahezu geraden Informationsfunktion abgelesen werden kann. Der Indikator gelegentliches angestrengtes Atmen trägt ebenso fast nichts zur Schmerzmessung bei, und auch die Indikatoren lange Phasen von Hyperventilation und lautstark angestrengtes Atmen sind selbst über Bereiche stärkeren Schmerzes hinweg nur wenig informativ. Mit einem Diskriminationsparameter von nahezu 1 und einer demnach engeren und höheren Informationsfunktion erscheint aus dem Bereich der BESD-Atmung lediglich der Indikator kurze Phasen von Hyperventilation von Nutzen für die Schmerzmessung zu sein. Im Bereich hoher Schmerzbelastungen dagegen erscheint dieses Item wieder weniger informativ als die zuvor genannten schwierigeren, jedoch insgesamt wenig informativen Atmungsindikatoren.

6.6.4 Skalenvergleich – Ausdrucksbereich Lautäußerung

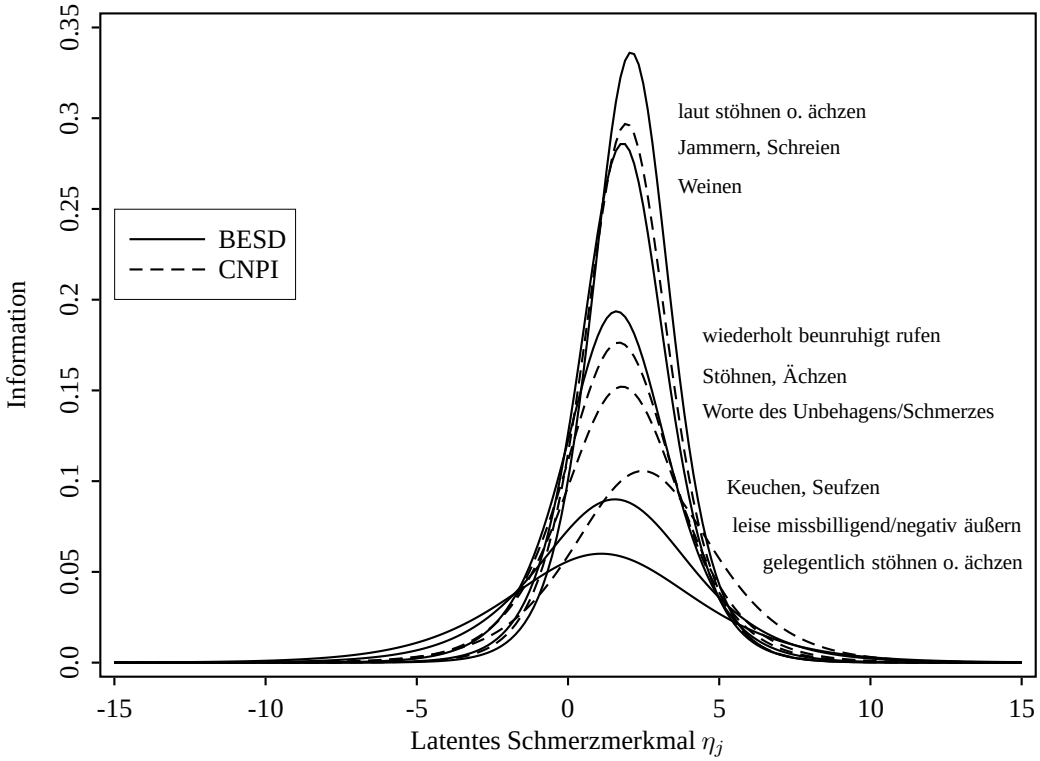
Auch wenn in Tabelle 18 die einzelnen Indikatorbereiche der BESD und CNPI nacheinander dargestellt sind, bietet sich eine direkte Gegenüberstellung der Einzelindikatoren aus vergleichbaren Ausdrucksbereichen in der grafischen Aufbereitung der Ergebnisse an. Die BESD-Indikatoren zum Ausdrucksbereich der Lautäußerung sind darum in Abbildung 29 gemeinsam mit den Indikatoren der CNPI-Ausdrucksbereiche verbale und nonverbale vokale Beschwerden dargestellt.

Auch hier fällt zunächst auf, dass alle Indikatoren ihren maximalen Informationswert im Bereich überdurchschnittlicher Schmerzbelastung erreichen, auch wenn die lautbezogenen potenziellen Schmerzindikatoren im Vergleich zu den Atmungsindikatoren etwas weniger schwierig, und in ihren Schwierigkeiten auch deutlich einheitlicher sind. Bei sehr ähnlich geschätzten Bereichen auf dem latenten Schmerzkontinuum, die durch die insgesamt neun Items dieses Bereiches abgedeckt werden, lassen sich drei Gruppen deutlich verschieden informativer Items unterscheiden.

Die größte Schmerzrelevanz weisen dabei die Items laut stöhnen oder ächzen, Jammern bzw. Schreien und Weinen auf. Dabei liefert BESD die Mehrheit der informativeren Indikatoren. Etwas weniger informativ erscheinen dagegen die Items wiederholt beunruhigt rufen, Stöhnen bzw. Ächzen und Worte des Unbehagens bzw. Schmerzes.

Als vergleichsweise uninformativ und darum weniger nützlich erscheinen dagegen die Items Keuchen bzw. Seufzen, sich leise missbilligend bzw. negativ äußern und gelegentlich stöhnen oder ächzen. Dabei ist interessant, dass die drei Indikatoren zum Stöhnen und Ächzen, zum einen zwar die erwartete Rangfolge ihrer Schwierigkeiten bestätigen, dane-

Abbildung 29: Informationsgehalt der Einzelitems des BESD-/CNPI-Bereiches Lautäußerung.



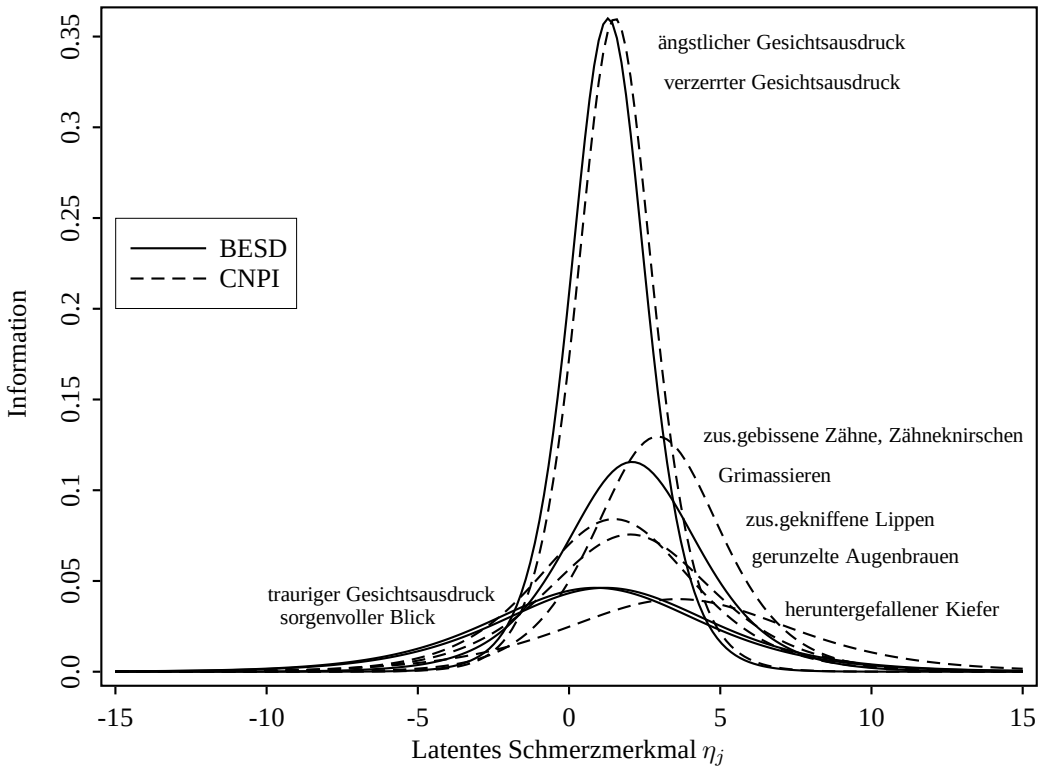
ben jedoch auch bedeutende Unterschiede in der Diskriminationsgüte und im Informationsgehalt bestehen, je nachdem, ob, und welche zusätzlichen Qualitäten (d.h. Lautstärke vs. Frequenz) in die Formulierung einbezogen werden. Im Vergleich zum atmungsbezogenen Schmerzausdruck erscheint der Indikatorbereich schmerzbezogener Lautäußerungen insgesamt deutlich informativer zu sein.

6.6.5 Skalenvergleich – Ausdrucksbereich Mimik

Insgesamt neun Indikatoren erfassen potenziellen Schmerz auf der Grundlage mimischen Ausdrucksverhaltens. Dabei erscheinen die vier BESD-Indikatoren trauriger und ängstlicher Gesichtsausdruck, sorgenvoller Blick und Grimassieren stärker an einer holistischen Interpretation des mimischen Ausdruckes orientiert zu sein als die etwas molekulareren mimischen Verhaltensqualitäten des CNPI-Inventars.

Die Indikatoren ängstlicher Gesichtsausdruck der BESD-Skala und verzerrter Gesichtsausdruck der CNPI-Skala erscheinen in ihrem Informationsgehalt und ihrer Lokalisierung auf dem latenten Schmerzkontinuum miteinander vergleichbar. Interessant ist, dass der im BESD-Inventar enthaltene Indikator Grimassieren etwas schwieriger scheint (was vielleicht durch die im Praxisalltag ungewöhnlichere Bezeichnung mitbedingt sein könnte),

Abbildung 30: Informationsgehalt der Einzelitems des BESD-/CNPI-Bereiches Mimik.



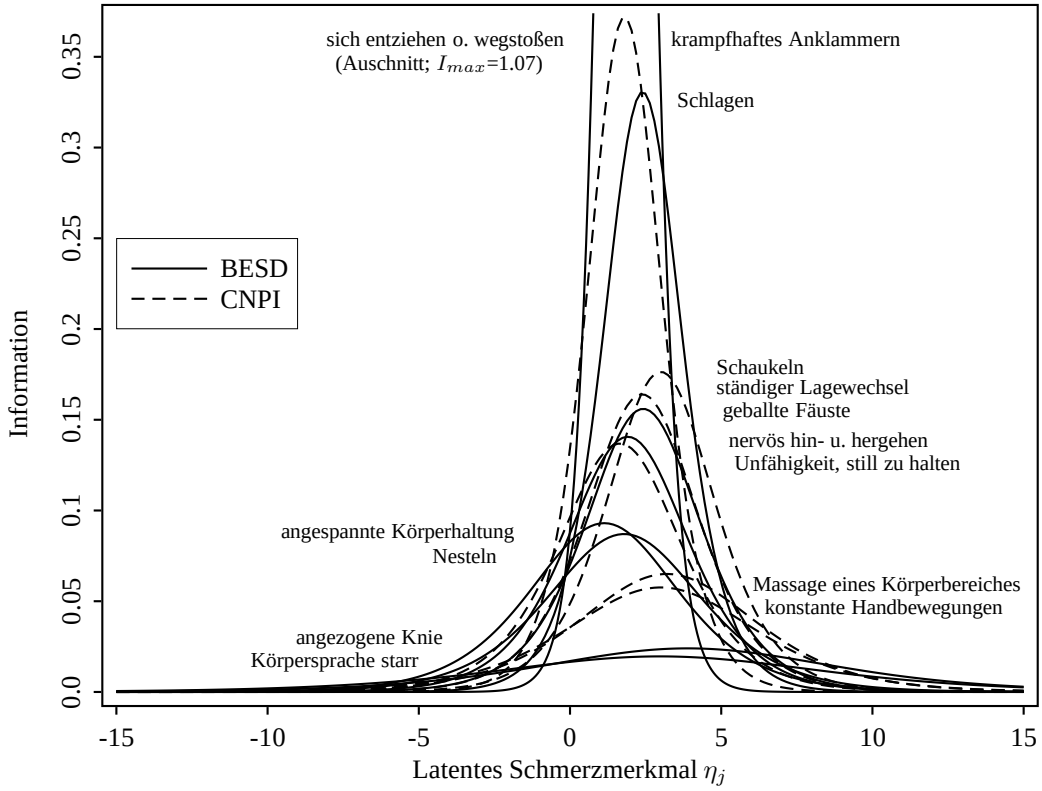
vor allem aber deutlich weniger diskriminativ geschätzt wird als ein verzerrter Gesichtsausdruck. Dennoch gehört neben zusammengebissenen Zähnen bzw. Zähneknirschen, zusammengekniffenen Lippen und gerunzelten Augenbrauen auch das Grimassieren noch zu denjenigen Indikatoren, die im Vergleich zumindest mittelmäßig gut zur Abschätzung von Schmerz beitragen können. Wenig spezifisch für das Schmerzerleben scheinen hingegen die Indikatoren trauriger Gesichtsausdruck, sorgenvoller Blick und ein heruntergefallener Kiefer zu sein, die entsprechend flache Informationsfunktionen aufweisen. Im Bereich der Mimik erscheinen die durch das CNPI definierten Indikatoren insgesamt etwas aussagekräftiger zu sein als der durch das BESD beschriebene mimische Schmerzausdruck.

6.6.6 Skalenvergleich – Ausdrucksbereich Körperhaltung

Die BESD-Skala beschreibt insgesamt acht potenziell schmerzbezogene Verhaltensweisen, wohingegen bei den insgesamt sechs gestischen Verhaltensindikatoren der CNPI-Skala zusätzlich die Einzelbereiche Anklammern, Ruhelosigkeit und Reiben unterschieden werden. Wie zuvor sollen alle in einem weiteren Sinn auf die Körperhaltung bezogenen Indikatoren beider Skalen in Abbildung 31 gemeinsam dargestellt und verglichen werden.

Zunächst fällt auf, dass der BESD-Indikator sich entziehen bzw. wegstoßen mit einem Diskriminationswert von a=2,07 als im Vergleich zu allen anderen potenziellen Schmerzindikatoren beider Inventare außergewöhnlich trennscharfes Item geschätzt wurde. Bereits in den vorangegangenen bereichs- und skalenspezifischen Auswertungen schienen diese Verhaltensäußerungen als dominierender Schmerzindikator auf. Der vergleichsweise steilen itemcharakteristischen Funktion dieses Items entsprechend wird für dessen Informationsfunktion ein Maximalwert von I=1,07 geschätzt, der das zumindest Dreifache der durch die verbleibenden Items bereitgestellten Schmerzinformation beträgt. In Abbildung 31 wurde darum nur ein Ausschnitt der Informationsfunktion dieses Items dargestellt.

Abbildung 31: Informationsgehalt der Einzelitems des BESD-/CNPI-Bereiches Körperhaltung.



Von den verbleibenden Items erscheinen krampfhaftes Anklammern und Schlagen als besonders informative Verhaltensindikatoren für den von den Bewohnern in Ruhe erlebten Schmerz, wobei letzterer Indikator etwas stärkere Schmerzzustände anzeigt bzw. schwieriger ist.

Auch wenn kein sinnvoller Cut-off-Wert für die Iteminformation bestimmt werden kann, so erscheinen die Items Schaukeln, ständiger Lagewechsel, geballte Fäuste, nervös hin- und hergehen sowie Unfähigkeit, still zu halten mit Maximalwerten zwischen I=0,18 und 0,14 als immerhin noch mittelmäßig informativ, während die verbleibenden Verhal-

tensweisen angespannte Körperhaltung, Nesteln, Massage eines Körperbereiches, konstante Handbewegungen, angezogene Knie und eine starre Körpersprache wenig Aussagekraft für das Schmerzerleben zu besitzen scheinen.

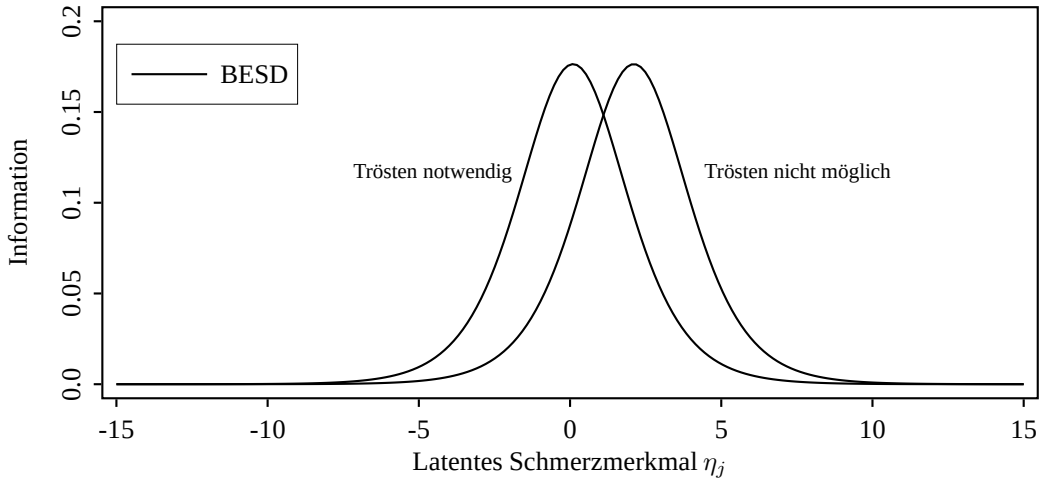
Insbesondere aufgrund des hochdiskriminativen Items sich entziehen oder wegstoßen ist der Anteil der durch die BESD-Indikatoren erfassten Schmerzinformationen sehr viel größer als der Gesamtinformationswert der CNPI-Gestikitems.

6.6.7 Skalenvergleich – Ausdrucksbereich Trost

Wie die Atmung ist auch der Indikationsbereich Trost nur in der BESD-Skala vertreten. Eine separate Modellierung beider Einzelitems ist durch den logischen Ablauf bei der Einschätzung dieses Bereiches und der sich daraus ergebenden Filterung bzw. vollständigen Abhängigkeit nicht möglich. Statt der dichotomen Einzelitems wurde darum wie zuvor beschrieben ein dreistufiger Indikator mit in die Gesamtschätzung einbezogen.

Damit wird das Item Trost durch eine einzige latente Responsevariable (LRV) repräsentiert, für die entsprechend zwei Thresholdparameter geschätzt werden. Da somit nur ein einziges Regressionsgewicht geschätzt wird sind auch die Diskriminationen für die beiden unterschiedenen Stufen des Trostitems identisch.

Abbildung 32: Informationsgehalt der Einzelitems des BESD-Skalenbereiches Trost.

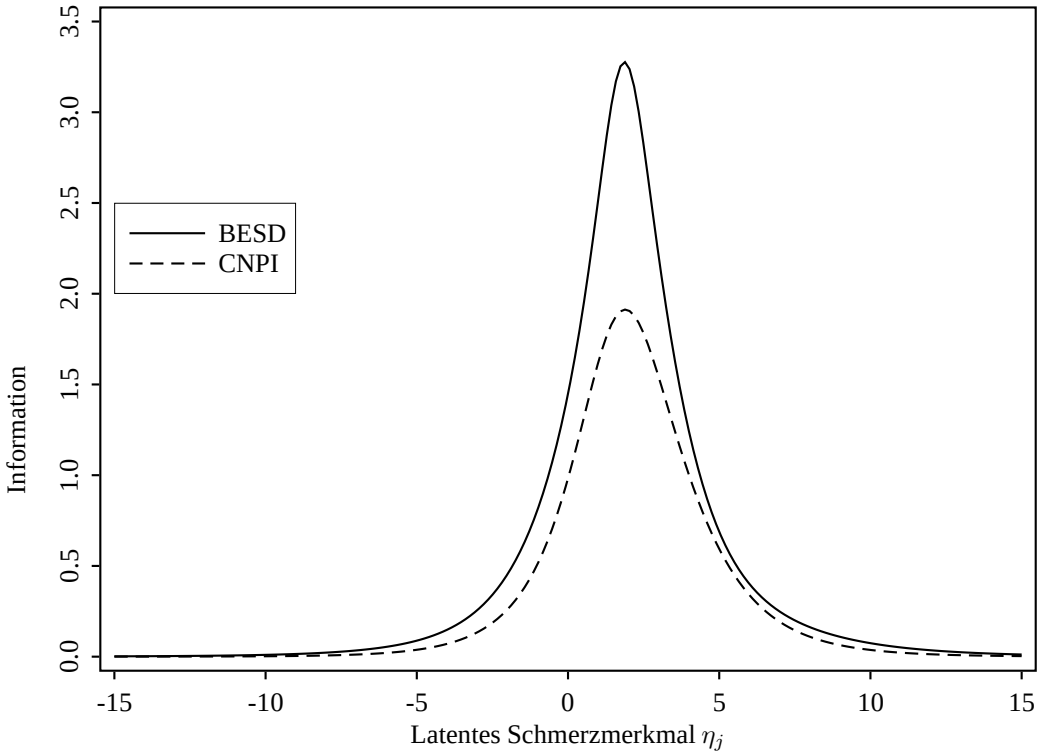


Die in Abbildung 32 geplotteten Informationsfunktionen sind darum lediglich gegeneinander verschoben, wobei selbstverständlich die Frage, ob der Bewohner getröstet werden sollte oder nicht für den Bereich mittleren oder geringeren Schmerzes informativ erscheint, wohingegen der Erfolg des Tröstens zusätzliche Informationen in Bereichen höherer Schmerzausprägungen beisteuert.

6.6.8 Informationsgehalt der Gesamtskalen

Ein wesentliches Ziel dieses Analyseschrittes besteht darin, das Potenzial aller durch die BESD vorgestellten Schmerzindikatoren mit demjenigen des CNPI-Gesamtinventars zu vergleichen. Die insgesamt durch die Itembatterien verfügbare Schmerzinformation ergibt sich als Summe aller einzelnen itembezogenen Informationsfunktionen. Die so ermittelten Test Information Functions (TIFs) können anschließend direkt hinsichtlich des abgedeckten Intervalles auf dem latenten Schmerzkontinuum und der über diese Bereiche hinweg geleisteten Güte der Messung miteinander verglichen werden.

Abbildung 33: Informationsgehalt der BESD- und CNPI-Gesamtskalen.



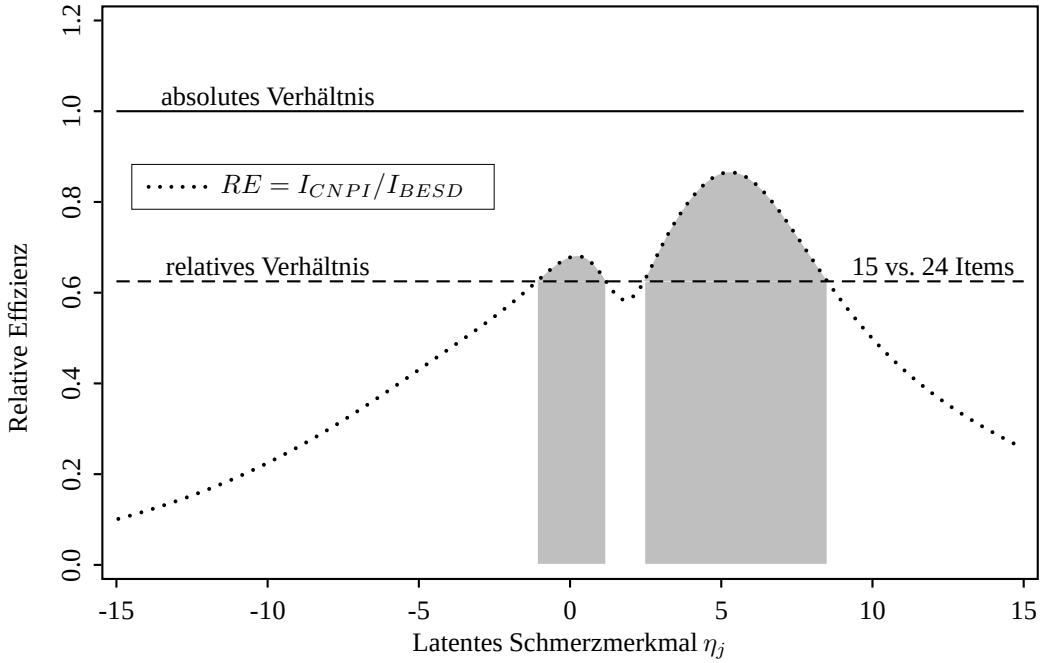
Die durch die insgesamt 23 (22 dichotom, 1 dreistufig) Einzelitems der BESD-Skala und die insgesamt 15 dichotomen Einzelitems der CNPI-Skala erfasste Gesamtinformation über das latente Schmerzkontinuum hinweg ist in Abbildung 33 abgetragen. Der Maximalwert der Testinformationsfunktion der BESD-Skala liegt mit 3,2 Punkten deutlich höher als derjenige für die CNPI-Skala (1,9 Punkte). Dabei muss berücksichtigt werden, dass für die BESD-Skala mehr Einzelindikatoren berücksichtigt wurden, und der Einzelindikator sich entziehen oder wegstoßen mit einem überproportional großen Informationsanteil in diese Gesamtfunktion einfließt.

Im Gegensatz zur Annahme der klassischen Testtheorie, dass eine Skala über den gesamten Bereich des zu messenden Merkmals hinweg gleich reliabel sei, wird aus der Abbildung ersichtlich, dass beide Inventare in den Randbereichen extrem hohen und geringen bzw. fehlenden Schmerzes weniger Information erfassen und die Messung dieser Schmerzzustände darum mit einer höheren Ungenauigkeit belastet ist als in Bereichen mittleren und höheren Schmerzes.

Insgesamt unterscheiden sich beide Inventare nur unwesentlich mit Blick auf diejenige wahre Schmerzausprägung, die jeweils am zuverlässigsten gemessen werden kann (jwls. ca. η=2). Durch die generell höhere Informationsfunktion können mit der BESD offensichtlich auch geringere Schmerzausprägungen besser abgebildet werden, während der Unterschied zwischen beiden Inventaren im Bereich hohen Schmerzes insgesamt geringer ausfällt.

Die nachfolgende Abbildung 34 setzt die durch die beiden Schmerzinventare erfasste Information direkt zueinander ins Verhältnis. Die Referenzlinie bei einem Wert von RE=1 kennzeichnet ein gleiches *absolutes* Ausmaß der durch die zu vergleichenden Instrumente erfassten Schmerzinformation. Absolut gesehen wird mit den BESD-Items über das gesamte latente Schmerzkontinuum hinweg mehr Information gesammelt als mit dem CNPI-Inventar. Wie bereits im vorherigen Abschnitt angemerkt, leistet das CNPI insbesondere um einen Merkmalswert von η=5 herum eine annähernd vergleichbar gute Schmerzerfassung wie die umfangreichere BESD-Skala.

Abbildung 34: Relative Effizienzfunktionen der BESD- und CNPI-Gesamtskalen.

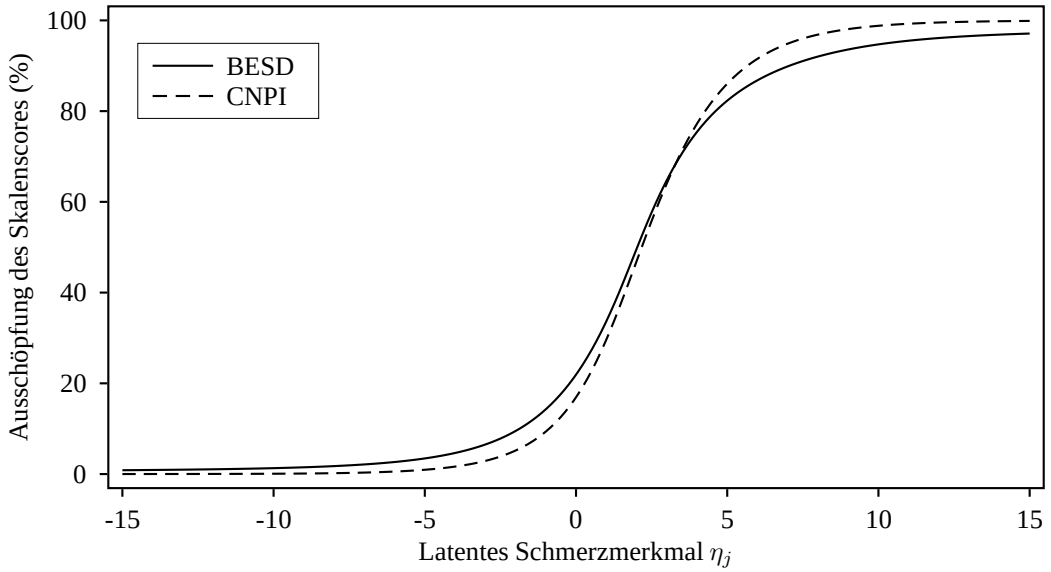


Für einen aussagekräftigeren Vergleich der *Effizienz* beider Skalen muss die unterschiedliche Anzahl verwendeter Indikatoren berücksichtigt werden. Bei der CNPI-Skala wurden 15 Verhaltensweisen als dichotome Einzelindikatoren extrahiert und vorgegeben, während aus der Original BESD-Skala insgesamt 24 dichotome Schmerzindikatoren gewonnen wurden. Dieses Verhältnis von 15 zu 24 (=0,625) Items ist als eine alternative Referenzlinie ebenfalls in Abbildung 34 dargestellt. Diejenigen Abschnitte der Relative Efficiency Function (REF), die über diese Referenzlinie hinausreichen markieren Abschnitte auf dem latenten Schmerzkontinuum (grau dargestellt), die durch einen typischen CNPI-Indikator besser abgebildet werden können als durch einen durchschnittlichen BESD-Indikator. Diese Relativierung an der Itemanzahl der Inventare zeigt, dass die CNPI-Skala sowohl im Bereich mittleren Schmerzes (bezogen auf die Schmerzbelastung der Bewohnerstichprobe) als auch über einen breiten Bereich gesteigerten Schmerzes hinweg eine effizientere Schmerzmessung erlaubt als die BESD-Skala.

6.6.9 Ausschöpfung des Wertebereiches

Für den praktischen Einsatz sind solche Skalen besonders effektiv, die den gesamten vorgesehenen Punktwertebereich für eine Differenzierung verschiedener Schmerzzustände ausnutzen. Sowohl für die BESD als auch CNPI haben die vorangegangenen Analysen gezeigt, dass zumindest bei der Einschätzung der hier betrachteten demenzkranken Heimbewohner, beide Skalen ihren theoretischen Wertebereich lange nicht voll ausschöpfen. Aufgrund der für die Einzelitems der Testbatterien geschätzten psychometrischen Kennwerte kann nun näher analysiert werden, welche Punktwerte für beide Instrumente über das latente wahre Schmerzkontinuum hinweg erwartet werden können.

Abbildung 35: Testcharakteristische Funktionen der BESD- und CNPI-Gesamtskalen.



Beim Vergleich der BESD- und CNPI-Skalen muss wiederum die unterschiedliche Itemanzahl beider Inventare berücksichtigt werden. In Abbildung 35 sind daher die Funktionsplots für die relative Ausschöpfung des Wertebereiches beider Instrumente über das latente Schmerzkontinuum hinweg dargestellt.

In Bereichen geringeren Schmerzes weist die BESD gegenüber dem CNPI eine höhere Ausschöpfung auf, d.h. hier schlagen sich leichtere Schmerzen früher in höheren erreichten Punktwerten nieder. Die BESD kann somit in diesem Bereich als das sensitivere Schmerzinstrument gelten. Andererseits erscheint es auch bei starken Schmerzen unwahrscheinlich, dass die volle mögliche BESD-Punktzahl erreicht wird, weswegen der Funktionsplot hier deutlich unter 100 Prozent Ausschöpfung bleibt. Für die CNPI-Skala kann dagegen über einen vergleichsweise engeren Bereich des latenten Schmerzkontinuums hinweg eine stärker beschleunigte Ausschöpfung des Wertebereichs angenommen werden. Auch höhere Skalenscores können noch in einem Bereich erlebter Ruheschmerzen erwartet werden, die für die Klientel nicht-klinischer demenzkranker Altenheimbewohner realistisch scheint.

6.7 Ruhe und Aktivität als Beobachtungskontext

Die vorangegangenen Auswertungsschritte konzentrierten sich darauf, die interne Struktur der beiden zur Diskussion stehenden Inventare zur beobachtungsgestützten Schmerzerfassung zu beleuchten und die psychometrischen Eigenschaften der diskret erfassten Verhaltensmerkmale detailliert abzuschätzen. Sowohl für diese bereichs- und skalenspezifischen Analysen als auch für den anschließenden direkten Vergleich der beiden Inventare wurden lediglich die in einer Ruhesituation erfassten Schmerzdaten berücksichtigt. Da keines der Instrumente explizit vorgibt, eine Schmerzqualität zu messen, die spezifisch für bestimmte Bewegungen oder Aktiviertheitsgrade wäre, macht eine solche Beschränkung auf die gewissermaßen *unbedingtere* Erfassungssituation zunächst sicherlich Sinn.

Dennoch verweisen beide Instrumente darauf, dass eine umfassende Schmerzeinschätzung sowohl Beobachtungen des Bewohners in Ruhe als auch bei Bewegung bzw. in Aktivität einschließen sollte. Die Originalanweisung der CNPI-Skala beispielsweise sieht darüber hinaus sogar eine Verrechnung der in beiden Situationen gewonnenen Skalenwerte zu einem Gesamttestwert vor.

Dabei wird offensichtlich unterstellt, dass die Instrumente unter beiden Beobachtungsbedingungen gleich (gut) funktionieren, also dieselbe latente Schmerzqualität gleich reliabel abzubilden in der Lage sind. Das folgende Kapitel geht darum der Frage nach, ob die Messstruktur der Schmerzinventare über beide Situationen hinweg als grundsätzlich vergleichbar angenommen werden kann. Insoweit, wie eine äquivalente Abbildung von Schmerzen unter beiden Beobachtungsbedingungen geleistet werden kann, bieten sich durch die realisierte Messwiederholung an denselben Bewohnern weiterführende längsschnittliche Analysen der individuellen Schmerzveränderung an. Werden in Situationen höherer Aktiviertheit höhere Schmerzbelastungen ermittelt, wäre es von großem Interesse, Prädiktoren für diese individuelle Vulnerabilität für bewegungs- bzw. aktivitätskorrelier-

ten Schmerz zu ermitteln.

Erste Hinweise auf eine in den unterschiedenen Beobachtungsbedingungen nicht oder nur teilweise äquivalente Schmerzmessung konnten bereits im Zuge der klassischen Skalenanalyse der BESD- und CNPI-Instrumente gewonnen werden.

Aufgrund der für die meisten potenziell schmerzbezogenen Verhaltensweisen in Aktivität gesteigerten Beobachtungsraten wurden die Einzelitems in Aktivität jeweils als durchschnittlich weniger schwierig eingeschätzt als in Ruhe (BESD: $\bar{p}_i$=.18 vs. .14; CNPI: $\bar{p}_i$=.14 vs. .10). Eine Minderheit von Schmerzindikatoren konnten jedoch in der Aktivitätssituation seltener beobachtet werden als in Ruhe, so dass diese Items in Aktivität als schwieriger gelten müssen. Dies sind im Einzelnen die Verhaltensweisen Cheyne Stokes Atmung, laut stöhnen oder ächzen, trauriger Gesichtsausdruck, sorgenvoller Blick, Nesteln, starre Körpersprache und angezogene Knie der BESD-Skala und heruntergefallener Kiefer, Schaukeln, konstante oder wiederkehrende Handbewegungen und Massage eines bestimmten Körperbereiches aus der CNPI-Skala. Dabei sind die absoluten Unterschiede in den Beobachtungsraten allerdings in den meisten Fällen sehr gering.

Für die Abschätzung der Invarianz der Messstruktur ebenso bedeutsam sind die gefundenen Unterschiede der Diskriminationsparameter für beide Beobachtungsbedingungen. Für einzelne Items können Veränderungsbeträge im Kennwert der Item-Gesamttest-Korrelation von bis zu $r_{it\Delta}$=.20 berechnet werden. In Aktivität erscheinen im Einzelnen die Indikatoren starre Körpersprache, angezogene Knie, zusammengebissene Zähne und ständiger Lagewechsel enger ($r_{it\Delta}$ mindestens +.10) mit dem latenten Schmerzmerkmal verknüpft als in Ruhe, während die Indikatoren verzerrter Gesichtsausdruck, nervös hin- und hergehen und laut stöhnen oder ächzen als weniger trennscharf ($r_{it\Delta}$ mindestens -.10) geschätzt werden. Insgesamt erscheinen die Itemdiskriminationen jedoch unter beiden Beobachtungsbedingungen gut miteinander vergleichbar.

Die Möglichkeiten, diese Unterschiede in den geschätzten Itemparametern als tatsächliche Abweichungen in der Messstruktur zu interpretieren sind im Rahmen der klassischen Testtheorie jedoch sehr eingeschränkt. Der Kontext einer Messung findet im engeren Sinne überhaupt keine Beachtung. Die höheren Itemschwierigkeiten in der Aktivitätssituation könnten aufgrund der Stichprobenabhängigkeit der Messung sowohl auf eine in dieser Beobachtungssituation tatsächlich gesteigerte Schmerzbelastung, oder aber auf eine nicht schmerzbezogene, prinzipiell höhere Beobachtungswahrscheinlichkeit der Einzelindikatoren zurückgeführt werden.

Im Rahmen der Generalisierungstheorie könnten verschiedene Beobachtungsbedingungen als Stufen einer Messfacette begriffen und ihr Einfluss auf die Variabilität der Messwerte abgeschätzt werden. Doch auch hierbei würde die Invarianz der Messstruktur zu beiden Zeitpunkten vorausgesetzt, nicht jedoch überprüft werden.

6.7.1 Invarianz der Messstruktur

Die vorangegangenen Analysen machten deutlich, dass beide Schmerzinstrumente neben psychometrisch zufriedenstellenden Indikatoren auch solche Verhaltensweisen beschrei-

ben, die nur wenig mit dem latenten Schmerzmerkmal verknüpft sind und damit keine reliable Schmerzmessung erlauben. Selbstverständlich macht eine Überprüfung der Invarianz einer Messung nur dann Sinn, wenn die zugrunde gelegte Messstruktur prinzipiell (d.h. in einer Referenzgruppe oder Durchführungsbedingung) eine angemessen reliable Abbildung des interessierenden Merkmals erlaubt. Für einzelne Indikatoren kann erwartet werden, dass diese erst in bestimmten Situationen (z.B. starre Körperhaltung in Aktivität; ständiger Lagewechsel in Ruhe) einen substanziellen Anzeigewert für erlebten Schmerz entwickeln. Diese Items können selbstverständlich wenig zu einer Abbildung der wahren Schmerzveränderung beim Wechsel zwischen Ruhe und Aktivität beitragen. Aus diesem Grunde erscheint die Beibehaltung aller Einzelindikatoren für diesen Analyseschritt weniger bedeutsam als für die zuvor beschriebenen psychometrischen Analysen.

Insofern, als die Überprüfung der Messinvarianz abschätzen soll, ob unter beiden Beobachtungsbedingungen jeweils dasselbe Merkmal vergleichbar reliabel erfasst wird, und damit die Voraussetzungen für eine inhaltliche Interpretation der abgebildeten Verhaltensunterschiede als wahre Schmerzänderung erfüllt sind, soll hierbei auf eine Differenzierung der Zugehörigkeit der Indikatoren zu den beiden Schmerzinventaren BESD und CNPI verzichtet werden. Entsprechend sollen für beide Durchführungsbedingungen strukturell äquivalente eindimensionale Messmodelle mit allen zuvor (d.h. in der Ruhesituation) als hinreichend reliabel bewerteten dichotomen (bzw. dreistufigen) Einzelindikatoren gegenübergestellt werden.

Die im Gesamtmodell geschätzten Diskriminationsparameter der Einzelindikatoren variieren wie bereits dargestellt in der Ruhesituation zwischen Werten von a=0,04 und a=2,07, mit einer mittleren Trennschärfe von $\bar{a_i}$=0,74. Wird, in Anlehnung an das Vorgehen der klassischen faktoranalytisch gestützten Skalenkonstruktion, als Cut-off Kriterium für die Trennschärfe jeweils ein Regressionsgewicht von $\lambda_i \geq .50$ veranschlagt, wird die Gesamtbatterie um 13 Indikatoren (also um ein Drittel) reduziert. Neun der insgesamt 24 BESD-Indikatoren und vier der 15 CNPI-Indikatoren sind demnach zu eliminieren. Besonders deutlich verringert sich die relative Anzahl zu berücksichtigender Verhaltensweisen in den BESD Bereichen Atmung (-80%) und Mimik (-50%). Ähnlich hoch fällt die Reduktion für den CNPI-Bereich Gesichtsgrimassen aus (-40%). Der körperbezogene Schmerzindikatorbereich Reiben, der durch nur ein Verhaltensmerkmal angezeigt wird, entfällt vollständig.

Wie bei der methodischen Diskussion der Voraussetzungen längsschnittlicher Veränderungsmessung bereits dargelegt, wurden zur Überprüfung der Invarianz einer Messstruktur eine Reihe alternativer Analyseverfahren und -strategien vorgeschlagen. Um in optimaler Weise an die bereits geleisteten Auswertungen anzuschließen und sowohl die Gleichheit der Diskriminationsparameter als auch der Schwellenparameter der latenten Responsevariablen überprüfen zu können, sollen die Daten auch hier als Strukturgleichungsmodell mit entsprechender IRT-Parametrisierung analysiert werden. Da theoretisch von einem in der Aktivitätssituation gesteigerten wahren Schmerzniveau ausgegangen werden soll, wird der Mittelwert der latenten Schmerzkomponente in dieser Bedingung nicht länger auf α=0 restringiert, sondern frei geschätzt.

Zur Überprüfung der Invarianz der Schmerzmessung in Situationen geringer und hoher Aktivierung wurden die Anpassungsstatistiken verschiedener Modelle mit unterschiedlich starken Gleichheitsrestriktionen der Messstruktur für beide Beobachtungsbedingungen per Likelihood-Ratio-Test (LR-Test) miteinander verglichen. Als Baselinemodell (H_0) dient dabei das in beiden Situationen maximal unrestringierte Modell. Die Invarianz der Itemparameter wurde nach den Empfehlungen von Golembiewski, Billingsley und Yeager (1976) in zwei Stufen überprüft. Zunächst wurden alle Faktorladungen als über die betrachteten Situationen hinweg invariant spezifiziert, um die Äquivalenz der inhaltlichen Bedeutungsstruktur des abzubildenden latenten Schmerzmerkmales zu überprüfen (gamma change). In einem weiteren Analysemodell sind zusätzlich auch die itemspezifischen Thresholdparameter jeweils für beide Beobachtungsbedingungen gleichgesetzt, um zu überprüfen, ob die Schmerzindikatoren in beiden Situationen vergleichbar schwierig sind (beta change). Da alle Analysen auf robusten Maximum-Likelihood-Schätzungen (MLR) beruhen, wurde anstelle des üblichen χ^2-Differenzentests der hierfür vorgeschlagene korrigierte LR-Test (Satorra & Bentler, 1999) verwendet.

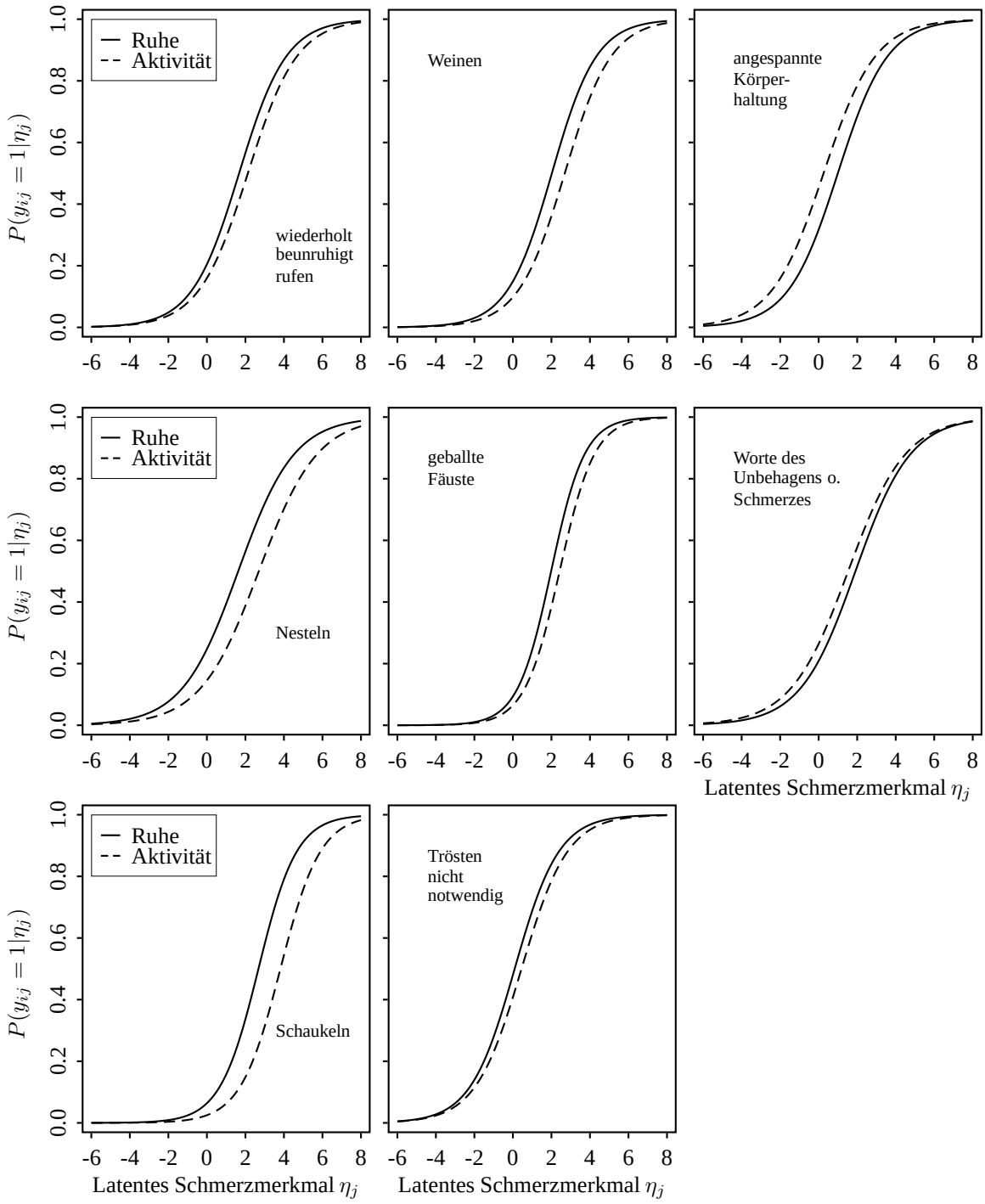
Der Verlust an Modellanpassung, der mit einer Gleichsetzung der faktoriellen Struktur, d.h. der Diskriminationsparameter, verbunden ist, kann auf der Grundlage dieses Tests als nicht substanziell gewertet werden ($-2ll_{\Delta c}$=20,72, df_{Δ}=24, $p<.655$). Die Indikatoren lassen sich damit sowohl in Situationen geringer als auch gesteigerter Aktivität als vergleichbar reliable Schmerzindikatoren beschreiben. Mit einer weiteren Gleichrestriktion der itemspezifischen Schwellenparameter in beiden Beobachtungsbedingungen geht hingegen ein signifikanter Verlust von Anpassungsgüte einher ($-2ll_{\Delta c}$=129,69, df_{Δ}=26, $p<.001$). Offensichtlich sind die veränderten Beobachtungsraten in der Aktivitätssituation nicht allein dem in diesen Analysen signifikant höher geschätztem wahren Schmerzniveau geschuldet, sondern auch durch zumindest für manche Items reduzierte Itemschwierigkeiten mitbedingt. Um diejenigen Items zu identifizieren, die bezüglich ihrer Thresholds als nicht invariant gelten müssen, wurden die entsprechenden Restriktionen für alle Einzelindikatoren separat getestet.

Insgesamt acht Einzelindikatoren weisen in den verschiedenen Beobachtungsbedingungen deutlich verschiedene Itemschwierigkeiten auf. Die itemcharakteristischen Kurven dieser Items mit “lack of invariance” (LOI) sind in nachfolgender Abbildung abgetragen.

Aus dem Indikatorbereich Lautäußerungen weisen die BESD-Items wiederholt beunruhigt rufen und Weinen, aus dem CNPI das Item Worte des Unbehagens bzw. Schmerzes in beiden Situationen unterschiedliche Schwellenwerte auf. Während die BESD-Items in Aktivität seltener zu beobachten waren, und darum hier als schwerer gelten müssen, können Worte des Unbehagens und Schmerzes in Situationen mit höherer Aktiviertheit als etwas leichter gelten.

Im Bereich der Körperhaltung sind die BESD-Indikatoren angespannte Körperhaltung, Nesteln, und geballte Fäuste, sowie das CNPI-Item Schaukeln in beiden Situationen deutlich verschieden schwierig. Eine angespannte Körperhaltung kann aufgrund der Analyseergebnisse in Situationen, die durch Aktivierung und Bewegung gekennzeichnet sind, eher

Abbildung 36: Items mit in Ruhe und Aktivität deutlich invarianten Thresholdparametern.



erwartet werden als in der u.a. als entspannt charakterisierten Ruhesituation. Insbesondere die körperbezogenen Verhaltensindikatoren Nesteln und Schaukeln weisen jeweils ein hohes Maß an fehlender Invarianz auf, wobei beide Verhaltensweisen eher in Ruhe beobachtet werden können.

Das Bedürfnis, den Bewohner aktiv zu trösten, scheint in der Ruhesituation etwas einfacher zu entstehen als in Situationen, die durch gesteigerte Aktivität gekennzeichnet sind. Im Gegensatz zu den verbleibenden analysierten dichotomen Items wurde dieser Ausdrucksbereich dreistufig modelliert. Aufgrund der nur teilweisen Threshold-Invarianz dieses dreistufigen Indikators soll im Weiteren auf den Indikationsbereich Trost vollständig verzichtet werden.

Zusammenfassend verbleiben für eine Analyse der wahren Merkmalsveränderung über die beobachteten Situationen hinweg insgesamt 17 ausreichend reliable und über die beiden Beobachtungsbedingungen hinweg invariante Verhaltensindikatoren. Da es sich dabei um ein Subset des ursprünglich vorgeschlagenen Gesamtpools von Verhaltensindikatoren handelt, der durch die vorangegangenen Analysen mit Blick auf das abgebildete Schmerzmerkmal in Teilen als durchaus heterogen identifiziert wurde, sind die Itemkennwerte für die getroffene Auswahl reliabler und transsituational konsistenter Items in einer eigenständigen Schätzung nochmals bestimmt worden. Wie zu erwarten war, ist der Verlust an Modellanpassungsgüte durch die Gleichrestriktion sowohl der Diskriminationen als auch der Thresholdparameter für dieses entsprechend selektierte Modell auch insgesamt nicht substanziell (-2ll$_{\Delta c}$=24,9, df$_{\Delta}$=32, p$<$.810). In Tabelle 19 sind die unter MI-Spezifikation geschätzten psychometrischen Kennwerte für diese Schmerzindikatoren dargestellt.

Aufgrund der in beiden Situationen unterschiedlich geschätzten wahren Merkmalsvarianzen ψ weichen die dargestellten standardisierten Regressionskoeffizienten und Thresholdparameter für beide Situationen gegebenenfalls voneinander ab, obgleich die unstandardisierten Modellparameter für die Faktorladungen und Schwellenwerte jeweils auf den gleichen Wert restringiert wurden. Für die Aktivitätssituation ergeben sich daraus konsistent etwas geringere standardisierte Regressionskoeffizienten und folglich leicht höhere Residualvarianzen. Diese Abweichungen übertragen sich entsprechend auch auf die nachträgliche Berechnung der Diskriminations- und Schwierigkeitsparameter a und b in der herkömmlichen IRT-Metrik (s. a. Muthén & Asparouhov, 2002), stellen jedoch nicht die prinzipielle Äquivalenz der Messstruktur über beide Situationen hinweg in Frage.

Im Vergleich zur in Tabelle 18 beschriebenen Messstruktur für alle Einzelitems der BESD- und CNPI-Skalen ergeben sich für die abschließende Itemauswahl erwartungsgemäß Veränderungen im Muster der geschätzten Itemladungen. Dabei erscheinen die Items sich leise missbilligend bzw. negativ äußern und nervös hin- und hergehen, die zuvor Regressionsgewichte von jeweils λ=.60 aufwiesen, mit dem nunmehr gemeinsam angezeigten Schmerzmerkmal nur noch vergleichsweise gering verknüpft. Alle weiteren Items weisen mit geschätzten Koeffizienten zwischen λ=.50 und .93 erwartungsgemäß weiterhin eine substanzielle Bestimmtheit durch den latenten Faktor auf. Der Stellenwert, den einzelne Items bei der Abbildung des Schmerzes in Ruhe und Aktivität nun jeweils

Tabelle 19: Kennwerte (IRT) situations-invarianter BESD- und CNPI-Indikatoren

		Ruhe					*Aktivität*				
Nr.[1]	Item	θ	τ_s	λ_s	a	b	θ	τ_s	λ_s	a	b
Atmung											
B-3	kurze Phasen v. Hyperventilation	0,55	1,7	.67	0,90	2,5	0,61	1,8	.62	0,80	2,8
Lautäußerungen											
B-7	leise missbilligend/neg. äußern	0,84	0,9	.40	0,44	2,1	0,86	0,9	.37	0,39	2,4
B-9	laut stöhnen oder ächzen	0,42	1,5	.76	1,17	2,0	0,48	1,6	.72	1,04	2,2
C-1	Keuchen, Seufzen	0,51	0,9	.70	0,99	1,3	0,56	1,0	.66	0,88	1,4
C-2	Jammern, Schreien	0,71	1,3	.54	0,64	2,4	0,75	1,3	.50	0,57	2,7
C-3	Stöhnen, Ächzen	0,71	1,1	.54	0,65	2,0	0,75	1,1	.50	0,57	2,3
Mimik											
B-12	ängstlicher Gesichtsausdruck	0,13	1,7	.93	2,57	1,8	0,16	1,9	.92	2,28	2,1
B-14	Grimassieren	0,26	1,8	.86	1,69	2,1	0,31	1,9	.83	1,51	2,3
C-6	zusammengekniffene Lippen	0,75	1,3	.50	0,58	2,6	0,79	1,3	.46	0,52	2,9
C-8	zus.gebissene Zähne, Knirschen	0,56	1,5	.67	0,89	2,2	0,61	1,5	.62	0,79	2,4
C-9	verzerrter Gesichtsausdruck	0,57	1,0	.66	0,87	1,5	0,62	1,0	.61	0,78	1,7
Körperhaltung											
B-16	nervös hin- und hergehen	0,82	0,8	.43	0,48	1,9	0,85	0,8	.39	0,42	2,1
B-21	sich entziehen o. wegstoßen	0,51	1,7	.70	0,99	2,4	0,56	1,8	.66	0,88	2,7
B-22	Schlagen	0,55	1,2	.67	0,90	1,8	0,61	1,3	.63	0,80	2,0
C-10	Krampfhaftes Anklammern	0,44	1,4	.75	1,12	1,8	0,50	1,4	.71	1,00	2,0
C-11	ständiger Lagewechsel	0,50	1,6	.71	1,00	2,2	0,56	1,6	.67	0,89	2,5
C-14	Unfähigkeit, still zu halten	0,60	1,1	.63	0,81	1,8	0,66	1,2	.59	0,72	2,0
Latenter Schmerzfaktor		ψ=2,641					α=0,584, ψ=2,093				

[1] B=BESD, C=CNPI; θ=Residualvarianz von y^*; τ_s=Threshold; λ_s=Regressionsparameter; b=Itemschwierigkeit; a=Itemdiskrimination.
Datenbasis: HILDE2 2006; N=194 (Ruhe und Aktivität).

einnehmen, ist dabei von der Bedeutung im Gesamtpool jedoch häufig verschieden. Höhere Gewichte wurden insbesondere für die Items ängstlicher Gesichtsausdruck (λ=.93 vs. .77), Grimassieren (.86 vs. .56) und Keuchen bzw. Seufzen (.70 vs. .55) gefunden. Etwas reduziert erscheint dagegen im ausgesuchten Inventar die Indikationsgüte der Items sich entziehen oder wegstoßen (λ=.70 vs. .90), Schlagen (.67 vs. .75) und Jammern bzw. Schreien (.54 vs. .74). Der Charakter des abgebildeten Schmerzerlebens trägt damit deutlich emotionale (v.a. angstbezogene) Züge.

6.7.2 Modellierung wahrer Merkmalsveränderung

Steht nun eine Auswahl verschiedener Indikatoren zur Verfügung, die das gemeinsam zu messende Schmerzmerkmal hinreichend präzise und über die berücksichtigten Beobachtungsbedingungen hinweg in vergleichbarer Weise abzubilden erlauben, kann der Frage nachgegangen werden, welche tatsächlichen Veränderungen sich im Schmerzerleben demenzkranker Menschen durch den betrachteten Wechsel im Aktiviertheitsgrad

ergeben. Verschiedene Ansätze zur Modellierung wahrer Merkmalsveränderung wurden im methodischen Teil dieser Arbeit bereits detailliert beschrieben. Dabei wurde für eine Konzeption dieser Merkmalsveränderung argumentiert, welche die individuelle Vulnerabilität demenzkranker Menschen, bei Aktivierung und Bewegung eine Schmerzsteigerung zu erfahren, in den Mittelpunkt stellt. Entsprechend soll diese nicht direkt erfassbare wahre Schmerzveränderung als latente Differenzkomponente eines längsschnittlichen IRT-Modells spezifiziert und abgeschätzt werden. Wird somit an Stelle der Merkmalsverteilungen des in beiden Situationen erlebten Schmerzes die Verteilung der individuellen Schmerzveränderung direkt modelliert, ergibt sich darüber hinaus die Möglichkeit, potenzielle Prädiktoren dieser individuellen Vulnerabilität zu formulieren und zu prüfen.

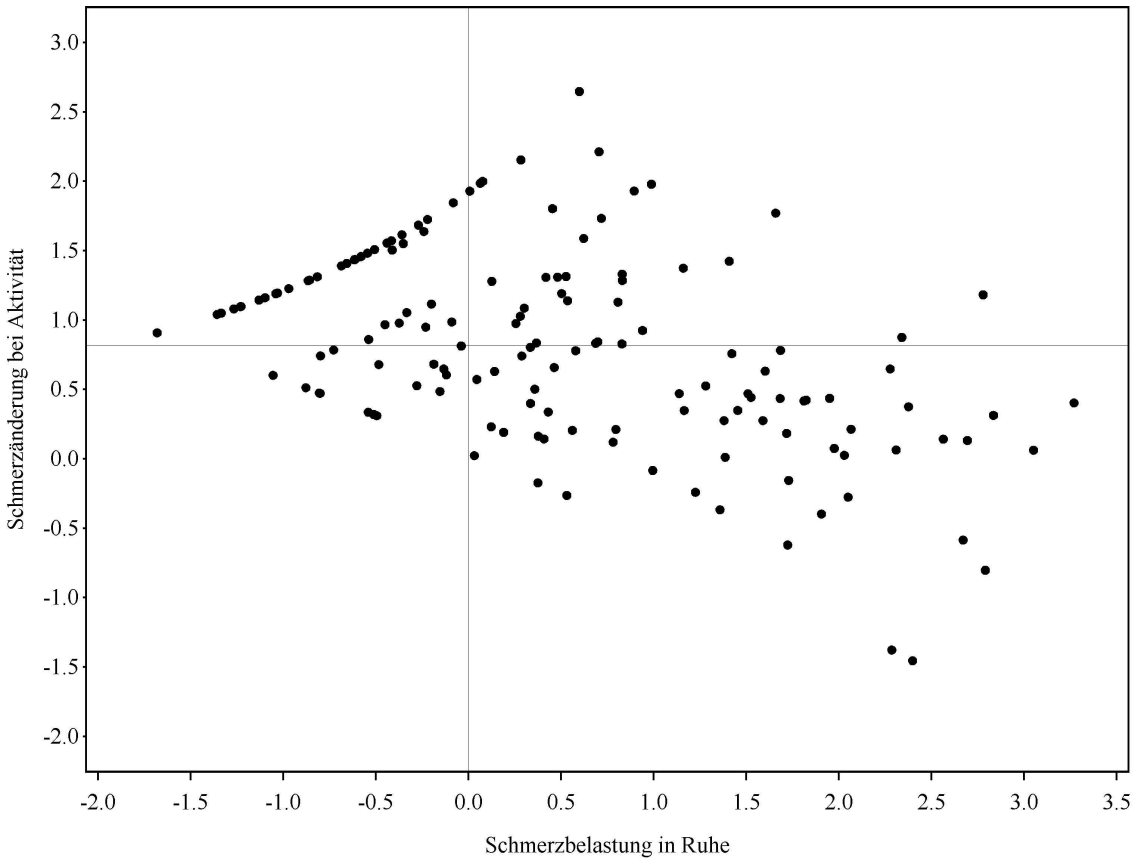
Wie bereits in Kapitel 4.3.3.5 dargelegt, sind das MISP und das LDCM äquivalente Parametrisierungen desselben Modells wiederholter Messungen. Entsprechend sind auch die Parameterschätzungen und die Modellanpassung beider Modelle identisch. In Aktivität sind durch die direkte Modellierung der latenten Merkmalsveränderung nun zwei latente Schmerzkomponenten (Ausgangsniveau und Veränderung) beteiligt.

In Abbildung 37 sind die für die demenzkranken Heimbewohner geschätzten Faktorwerte auf den modellierten Dimensionen Ruheschmerz und Schmerzveränderung bei Aktivierung gegeneinander abgetragen.

Für die wahre Schmerzveränderung wird ein Mittelwert von α=0,802 geschätzt, d.h. im Mittel liegt die Schmerzbelastung der Bewohner in der Aktivitätssituation eine halbe Standardabweichung höher als in der Ruhesituation. Die Varianz der wahren Schmerzveränderung wird auf ψ_{Δ}=1,037 geschätzt. Damit erscheinen die Bewohner hinsichtlich ihrer individuellen Vulnerabilität für bewegungs- oder aktivitätsbezogenen Schmerz deutlich homogener zu sein als hinsichtlich ihrer individuellen Schmerzbelastung in der Ruhesituation (ψ=2,474). Die Korrelation zwischen dem wahren Schmerzniveau und der wahren Schmerzveränderung bei Aktivierung wird auf $\psi_{\eta\Delta}$=-.49 geschätzt. Für Bewohner, die in Ruhesituationen geringen Schmerz erleben, kann also mit deutlicheren Schmerzveränderungen bei Aktivierung gerechnet werden als für Bewohner, die bereits in Ruhe unter Schmerzen leiden.

Das jeweilige Niveau und die Variabilität beider Faktoren, sowie der Zusammenhang von Schmerzniveau und Schmerzveränderung sind wie beschrieben auch in Abbildung 37 zu ersehen. Auffallend ist darüber hinaus, dass Personen mit einer geringeren basalen Schmerzbelastung in Ruhe auf Aktivierung vergleichsweise einheitlich mit einer mäßigen Schmerzsteigerung reagieren. Mit höheren Grundbelastungen (bis zu einem Schmerzwert von ungefähr η=1 wächst auch die Heterogenität der in Aktivität erlebten Schmerzveränderung kontinuierlich an. Bewohner, für die in der Beobachtungssituation Ruhe jedoch ein gesteigertes oder hohes Maß an Schmerzbelastung angenommen werden kann, erfahren bei Aktivierung offensichtlich deutlich seltener eine weitere Schmerzsteigerung. Für diese Bewohner kann im Gegenteil in Aktivität sogar häufiger eine geringere Schmerzbelastung angenommen werden als in Ruhe.

Abbildung 37: Geschätzte Faktorwerte der Personen auf beiden Dimensionen des LDCM.



6.7.3 Prädiktoren des Schmerzniveaus und der Schmerzveränderung

Eine abschließende Analyse setzt die geschätzten wahren Schmerzbelastungen und individuellen Vulnerabilitäten, bei Aktivierung Schmerz zu erfahren, mit persönlichen Kompetenzmerkmalen der berücksichtigten Bewohner in Bezug. Dabei sollen die Kompetenzen der demenzkranken Menschen nicht allein auf die kognitive Denk- und Gedächtnisleistung beschränkt bleiben, sondern darüberhinaus auch Aspekte der selbständigen Lebensführung (Barthel Index) und der nicht-kognitiven Demenzsymptomatik (NPI; herausforderndes Verhalten) umfassen. Zuvor wurde bereits auf diese im Gesamtprojekt vertretene holistische, am praktischen pflegerischen Bild orientierte Perspektive auf die Demenz hingewiesen und vier disjunkte Bewohnergruppen mit unterschiedlichem Muster erhaltener Kompetenzen beschrieben. Für die Beurteilung des Effektes verschiedener Beeinträchtigungen auf die Schmerzbelastetheit der Bewohner bietet sich eine Nutzung der drei differenzierten Kompetenzmaße eher an als die gemeinsame Betrachtung im Rahmen eines gruppenspezifischen Ansatzes. Während alle drei Aspekte das prinzipielle Niveau erlebter Schmerzen beeinflussen sollten, wird angenommen, dass insbesondere mobilitiäts-

bezogene körperliche Beeinträchtigungen für ein gesteigertes Schmerzerleben in Aktivität verantwortlich sind.

Durch die Prädiktoren MMSE, Barthel Index (Mobilitätsscore) und NPI können zusammengenommen knapp 23 Prozent der Variabilität im geschätzten latenten Merkmal Schmerzbelastung und etwas über 12 Prozent der Variabilität der Schmerzänderung bei Aktivität vorhergesagt bzw. erklärt werden. Die Denk- und Gedächtnisfähigkeit erscheint mit einem Einflußgewicht von λ_1=-.21 für das Schmerzniveau bedeutsam, d.h. für stärker kognitiv beeinträchtigte Bewohner werden höhere Schmerzbelastungen geschätzt. Dagegen kann die Schmerzänderung bei Aktivität mit λ_2=-.04 nicht durch den kognitiven Status vorhergesagt werden. Geradezu umgekehrt stellt sich der Einfluss des Prädiktors körperlicher Alltagskompetenzen dar. Für die im wesentlichen auch durch körperliche Inaktivität gekennzeichnete Situation der Ruhe bzw. für den als Basismessung geschätzten Schmerz hat der Barthel Index keine Erklärungskraft (λ_1=-.05), während der Faktor der Schmerzänderung bei Aktivierung hypothesenkonform deutlich mit diesem Prädiktor verknüpft ist (λ_2=-.31), ein höherer mobilitätsbezogener Hilfebedarf also häufiger mit einer gesteigerten aktivitätsbezogenen Schmerzvulnerabilität einherzugehen scheint. Der dritte theoretisch postulierte Prädiktor für die Schmerzbelastung demenzkranker Menschen, die nicht-kognitiven Verhaltensauffälligkeiten (NPI), erweist sich mit Blick auf beide Merkmalsdimensionen als bedeutsam. Dabei ist eine höhere Belastetheit mit Verhaltensauffälligkeiten deutlich positiv mit gesteigertem Schmerzausdruck bzw. dem durch diesen angezeigten Schmerzerleben in der Ruhesituation verbunden (λ_1=.33). Je stärker aber diese Verhaltensauffälligkeiten ausgeprägt sind, desto geringer fallen die tatsächlichen Schmerzveränderungen über beide Beobachtungssituationen hinweg aus (λ_2=-.21).

Trotz des sinnfälligen Parametermusters können die geschätzten Effekte für die Prädiktoren MMST und BI lediglich unter Vorbehalt interpretiert werden, da das konventionelle alpha-Niveau von 5 Prozent knapp verpasst wird.

6.8 Demenzspezifität der Schmerzmessung

Die bisher für die Gesamtgruppe aller berücksichtigten demenzkranken Heimbewohner vorgestellten Analysen belegen die prinzipielle Anwendbarkeit eines beobachtungsgestützten Schmerzassessments durch professionell Pflegende in verschiedenen Alltagssituationen mit unterschiedlichem Aktiviertheitsgrad der Bewohner.

Selbstverständlich sind auch solche Erfassungszugänge zum Schmerzerleben demenzkranker Menschen, die stärker auf Fremdurteile oder konkrete Verhaltensbeobachtung setzen, nicht von den mit der Erkrankung verbundenen Veränderungen im Erleben und Verhalten der Bewohner unabhängig.

Eines der von den Befürwortern standardisierter Verhaltensbeobachtungen am häufigsten vorgebrachten Argumente betrifft die mangelnde Einsichts- und Auskunftsfähigkeit der an einer Demenz erkrankten Menschen. Dabei stellt insbesondere die Kombination eines potenziell veränderten Schmerzerlebens mit einer deutlich veränderten bzw. reduzierten Kommunikations- bzw. Ausdrucksfähigkeit sowohl Praktiker als auch Testentwickler

vor große Probleme. Eine systematische Verknüpfung der Bemühungen um eine Schmerzabbildung bei verschiedenen ‚Problemgruppen' nicht-kommunikationsfähiger Menschen (z.B. Aphasiker, Menschen mit Locked-In-Syndrom, beatmete Patienten) könnte dazu beitragen, die demenzspezifischen Schwernisse für eine reliable und valide Schmerzerfassung deutlicher herauszustellen und Möglichkeiten für deren Entwicklung bzw. Anpassung zu identifizieren.

6.8.1 Verbale Auskunftsfähigkeit

Im Zuge der Charakterisierung der Bewohnerstichprobe wurden verschiedene demenzspezifische Beeinträchtigungen des Denkvermögens und Verhaltens, sowie der demenziellen Erkrankung und anderen Faktoren geschuldete Einbußen im kommunikativen Bereich bereits beschrieben. In Tabelle 20 sind über diese auf die Gesamtstichprobe bezogenen Kompetenzbeschreibungen hinaus die in den vier unterschiedenen Prägnanztypen der Demenz gefundenen kommunikativen Beeinträchtigungen aufgeführt.

In der Einschätzung der Bezugspflegenden leidet ungefähr jeder vierte leicht demenzkranke Bewohner und jeder zweite mittelgradig beeinträchtigte Bewohner an einer Minderung seiner verbalen Kommunikationsfähigkeiten, während fast alle als schwer demenzkrank eingestuften Bewohner entsprechende Kompetenzeinbußen aufweisen. Neben diesem häufigeren Auftreten erscheint auch die Schwere der verbalen Kommunikationsprobleme in den beiden schwer demenzkranken Bewohnergruppen deutlich gesteigert.

Tabelle 20: Verbale Kommunikationsfähigkeit und schmerzbezogene Selbstauskunft in vier Bewohnergruppen mit unterschiedlichem Muster erhaltener Kompetenzen

	Kompetenzgruppe			
N % bzw. N M±SD	LD	MD	SD-S	SD-P
verbal-kommunikative Einschränkungen[1]				
Vorliegen v. Beeinträchtigungen	8 24%	48 49%	45 92%	14 93%
Schwere der Beeinträchtigung	8 2,1±1,0	48 2,2±0,9	45 3,4±1,0	14 3,4±1,0
Ursache: Demenz	4 67%	40 95%	35 85%	13 100%
Ursache: andere Erkrankung	3 60%	8 20%	13 41%	2 15%
Einschränkungen des Verständnisses[1]				
Vorliegen v. Beeinträchtigungen	9 25%	47 48%	38 79%	13 87%
Schwere der Beeinträchtigung	9 2,1±0,8	47 2,7±1,1	38 3,2±1,1	13 3,9±1,0
Ursache: Demenz	3 50%	36 84%	31 91%	13 100%
Ursache: andere Erkrankung	6 86%	12 32%	13 45%	2 17%
Selbstauskunft zu Schmerzen[2]				
keine klar interpretierbare Antwort	–	7 7%	15 31%	5 36%

[1] Urteil der Bezugspflegeperson; Nennung mehrerer Ursachen möglich.

[2] Einschätzung durch Gerontopsychiater im diagnostischen Interview.

Datenbasis: HILDE2 2006; N=214 (Diagnostik), N=199 (Pflegeinterview).

Als Ursache für die Kommunikationsbeeinträchtigungen wurde in allen Demenzgraden vor allem die Demenz selbst angegeben. Daneben werden aber insbesondere in den Gruppen der beginnend Demenzkranken und der körperlich stark beeinträchtigten Schwerdementen häufiger (zusätzlich) auch weitere Erkrankungen bzw. Funktionsverluste als Ursache für vorhandene Kommunikationsprobleme angegeben.

6.8.2 Verständnisfähigkeit

Betrachtet man die Häufigkeit berichteter Verständnisschwierigkeiten als einer weiteren Voraussetzung für eine gelingende Kommunikation auch über Schmerzen, ergibt sich in den beiden Gruppen der leicht und mittelschwer demenzkranken Bewohner ein ähnliches Bild wie zuvor; die beiden Gruppen schwer demenzkranker Bewohner unterscheiden sich nun hingegen etwas deutlicher voneinander, wobei der Anteil von in ihrem Verständnis beeinträchtigten Bewohnern in der Gruppe der psychopathologisch auffälligen Bewohner nochmals höher liegt als in der Gruppe der somatisch stark beeinträchtigten Schwerdementen. Die kontinuierliche Abnahme der Verständnisfähigkeit über die vier Kompetenzgruppen hinweg zeigt sich auch bezüglich des Schweregrades der Verständnisschwierigkeiten, und wird durch einen kontinuierlich steigenden Anteil von Demenz als Ursache hierfür unterstützt. Interessanterweise zeigen sich jedoch auch mit Blick auf das allgemeine Verständnis der Bewohner die zuvor bereits angesprochenen höheren Anteile zusätzlicher oder alternativer verursachender Konditionen in den Gruppen der leicht Demenzkranken und sowohl körperlich als auch kognitiv stark beeinträchtigten Bewohner.

6.8.3 Schmerzbezogene Selbstauskunft

Legt man die doch häufig berichteten verbalen und Verständnisschwierigkeiten zugrunde, so erscheint es unerwartet, dass auch in den Gruppen der Schwerstbeeinträchtigten von noch ungefähr jeweils zwei Dritteln der Bewohner offenbar klar interpretierbare Auskünfte zur akuten Schmerzbelastung erfragt werden konnten. Von den mittelschwer demenzkranken Bewohnern konnten lediglich 7 Prozent keine nach dem Ermessen des erfassenden Gerontopsychiaters sinnvolle bzw. aussagekräftige Antwort mehr geben. In der Gruppe der leicht Demenzkranken konnten alle Bewohner Auskunft über ihre momentane Schmerzbelastung geben. Offensichtlich ist die Frage nach aktuellen Schmerzen für die Pflegeheimbewohner vergleichsweise einfach zu beantworten, denn für die etwas komplexere Frage nach der allgemeinen Lebenszufriedenheit anhand einer 11-stufigen Skala (von 0 bis 10 mit den verbalen Ankern „sehr unzufrieden“ und „sehr zufrieden“) lagen die Auskunftsraten deutlich niedriger (97%, 68%, 42% und 14%). Jedem zehnten Bewohner, der angab unter Schmerzen zu leiden, war es nicht mehr möglich, die Intensität dieser Schmerzen auf einer dreistufigen Skala (schwach-mäßig-stark) einzuschätzen. Für diese Bewohner konnten, nicht zuletzt auch wegen der sehr geringen Gruppengröße, keine substanziellen Unterschiede hinsichtlich des Kompetenzgrades belegt werden.

6.8.4 Non-verbale Kommunikationsfähigkeit

Während für einen informativen verbalen Austausch Verständnis- und verbale Kommunikationsfähigkeiten im Vordergrund stehen, mögen für die nicht-teilnehmende Verhaltensbeobachtung (die nicht als in gleichem Maße reziproke bzw. gerichtete Kommunikation verstanden werden kann) Einschränkungen in der non-verbalen Kommunikationsfähigkeit von vergleichsweise größerer Bedeutung sein. Um mögliche demenzspezifische Verluste in der gestischen, mimischen oder vokalen (non-verbalen) Ausdrucksfähigkeit nachzuvollziehen, sind in nachfolgender Tabelle 21 die diesbezüglichen Einschätzungen der Pflegenden wiederum getrennt für die unterschiedenen HILDE-Kompetenzgruppen dargestellt. Um darüberhinaus Anhaltspunkte dafür zu erhalten, wie sich diese Kommunikationseinbußen auf die Schmerzerfassung durch Verhaltensbeobachtung auswirken, wurden zusätzlich die Anteile derjenigen Bewohner dargestellt, für die bei der Schmerzbeobachtung keine der für die Ausdrucksbereiche der BESD- und CNPI-Skalen beschriebenen Verhaltensweisen beobachtet werden konnten. Damit soll gewissermaßen die demenzspezifische Nutzung bzw. Nicht-Nutzung bestimmter Ausdruckskanäle bei der nonverbalen Schmerzkommunikation nachvollzogen werden, auch wenn die Nichtnutzung dabei selbstverständlich nicht automatisch auch als eine Nichtverfügbarkeit interpretiert werden darf.

Tabelle 21: Non-verbale Kommunikationsfähigkeit und Schmerzausdruck in vier Bewohnergruppen mit unterschiedlichem Muster erhaltener Kompetenzen

	Kompetenzgruppe			
N % bzw. N M±SD	LD	MD	SD-S	SD-P
Non-verbale Kommunikationseinschränkungen[1]				
Vorliegen v. Beeinträchtigungen	6 18%	37 38%	32 70%	10 71%
Schwere der Beeinträchtigung	5 2,0±1,0	37 2,2±1,0	31 3,1±1,0	10 3,3±0,5
Ursache: Demenz	3 75%	32 94%	25 89%	9 100%
Ursache: andere Erkrankung	2 50%	6 19%	12 48%	3 33%
Nicht beobachtete Verhaltenskategorie[2]				
BESD - Atmung	44%	50%	29%	53%
BESD - Lautäußerung	52%	43%	26%	20%
BESD - Mimik	38%	45%	30%	20%
BESD - Körperhaltung	56%	42%	19%	20%
CNPI - vokale Beschwerden	81%	73%	55%	57%
CNPI - Gesichtsgrimassen	68%	66%	45%	33%
CNPI - Klammern	100%	93%	81%	57%
CNPI - Ruhelosigkeit	85%	78%	76%	53%
CNPI - Reiben	92%	94%	96%	93%

[1] Urteil der Bezugspflegeperson; Nennung mehrerer Ursachen möglich. [2] Anteil v. Bewohnern ohne Beobachtungen in diesen Verhaltensbereichen der BESD bzw. CNPI-Skala (beide Situationen).
Datenbasis: HILDE2 2006; N=199 (Pflegeinterview).

In der Einschätzung der Bezugspflegenden sind die Beeinträchtigungen in der nonverbalen Kommunikation im Vergleich zur verbalen Kommunikation in allen Gruppen etwas weniger stark ausgeprägt. Allerdings lässt sich auch für diesen Funktionsbereich eine für die beiden schwer demenzkranken Gruppen in Häufigkeit und Ausmaß deutlich gesteigerte Beeinträchtigung feststellen als für die beiden weniger beeinträchtigten Bewohnergruppen. Auch mit Blick auf die non-verbale Kommunikation wird in allen Gruppen zumeist (75% bis 100%) die demenzielle Erkrankung als Ursache der Kommunikationsprobleme genannt. Wie zuvor liegen dabei insbesondere für die Gruppen der leicht demenzkranken und körperlich stark beeinträchtigten schwer demenzkranken Bewohner häufiger auch (zusätzlich) nicht-demenzielle Ursachen für Einschränkungen im Verhaltensausdruck vor.

In der Kompetenzgruppe der körperlich und kognitiv stark beeinträchtigten Demenzkranken (SD-S) ist der mittlere Anteil der Bewohner, bei denen in den beiden Situationen keine auffällige Atmung beobachtet werden konnte vergleichsweise gering. Auch die durchschnittlich mit steigendem Demenzschweregrad höheren Beobachtungsraten für die Indikatorbereiche Lautäußerungen, Mimik, Körperhaltung, Gesichtsgrimassen, Klammern und Ruhelosigkeit weisen eher nicht darauf hin, dass schwerer demenziell beeinträchtigte Menschen in ihren non-verbalen Möglichkeiten, Schmerzen auszudrücken besonders stark eingeschränkt wären.

6.8.5 Wahl von Beobachtungssituationen

Eine mögliche Einschränkung in der universellen Anwendbarkeit oder zumindest der Vergleichbarkeit beobachtungsgestützter Schmerzassessments über verschiedene Schweregrade oder Syndromgruppen der Demenz hinweg könnte sich daraus ergeben, dass in Abhängigkeit von den noch erhaltenen Kompetenzen der Bewohner systematisch unterschiedliche Beobachtungssituationen gewählt werden bzw. verfügbar sind. Im Kontext der vorgesehenen Ruhesituation mögen die Unterschiede mit Blick auf ihr jeweiliges Schmerzpotenzial noch vergleichsweise gering sein; der Charakter der Aktivitätssituation erscheint demgegenüber jedoch sehr wahrscheinlich vom Kompetenzgrad der Bewohner abhängig.

Eine grobe Klassifizierung der Situationen, die von den Pflegenden für die Schmerzbeobachtung in Ruhe- bzw. Aktivität gewählt wurden, ist in Tabelle 22 nach Kompetenzgruppen dargestellt. Wie ihr vergleichsweise schlechter körperlicher Zustand erwarten lässt, wurde für Bewohner der Kompetenzgruppe SD-S häufiger eine Ruhesituation gewählt, in der sich der Bewohner in liegender Position befindet. Auch die Wahl der Aktivitätssituation scheint sich wie erwartet an den verbliebenen Bewohnerkompetenzen zu orientieren. Während Aufstehen oder Gehen für die Gruppe der körperlich stark eingeschränkten Bewohner entsprechend selten als Beobachtungssituation gewählt wurde, boten sich solche Situationen zur Beobachtung der kognitiv ebenso stark beeinträchtigten Gruppe psychopathologisch auffälliger Bewohner offensichtlich eher an. Der Toilettengang oder Transfer wurde für die Gruppe der SD-S häufiger (letzterer ausschließlich hier) als Beobachtungssituation gewählt, was sicherlich auch auf den erhöhten Unterstützungsbedarf dieser

Tabelle 22: Gewählte Beobachtungssituationen zur Schmerzerfassung in vier Bewohnergruppen mit unterschiedlichem Muster erhaltener Kompetenzen

	Kompetenzgruppe			
N %	LD	MD	SD-S	SD-P
Ruhesituation				
sitzend ohne Ansprache	23 92%	60 82%	27 66%	8 80%
liegend ohne Ansprache	2 8%	13 18%	14 34%	2 20%
Aktivitätssituation				
Aufstehen oder Gehen	16 59%	57 68%	20 45%	9 75%
Toilettengang	3 11%	8 10%	10 23%	2 17%
Transfer	–	–	10 23%	–
sonstige Aktivität	8 30%	19 23%	4 9%	1 8%

Datenbasis: HILDE2 2006; N=195 (Ruhe), N=196 (Aktivität).

Bewohner bei diesen Tätigkeiten zurückgeführt werden kann. In den weniger stark beeinträchtigten Demenzgruppen wurden dagegen eine Vielzahl weiterer angeleiteter (z.B. Malen in der Beschäftigungstherapie) und selbständiger Aktivitäten (z.B. Erzählen, Lesen) beobachtet.

Einschränkend muss angemerkt werden, dass die vorgestellten Analysen bestenfalls als ein erster Schritt in Richtung einer tatsächlich demenzspezifischen Schmerzerfassung gelten können, jedoch durch weitere Forschungsarbeiten ergänzt werden müssen, welche die schmerzbezogenen Erlebens- und Ausdrucksweisen stärker auf die Ätiologie der demenziellen Erkrankung oder spezifischer Kompetenzverluste beziehen. Auf eine psychometrische Analyse der Messinvarianz der im Fokus dieser Arbeit stehenden BESD und CNPI-Skalen, die dem in Kapitel 4.3.4 kurz angerissenen Multi-Group-Modeling Ansatz folgt, wurde an dieser Stelle aufgrund der beschränkten Stichprobengröße verzichtet.

6.8.6 Schmerzbelastung von Prägnanztypen des Demenzsyndroms

Die vorstehenden Ergebnisse weisen darauf hin, dass die Schmerzmessung mit keinem der gewählten Zugänge unabhängig vom Schweregrad der Demenz oder der spezifischen Syndromlagerung der Krankheit geleistet werden kann. Trotz dieser potenziellen Unschärfe sollen die mit den im Rahmen dieser Arbeit eingesetzten Assessments abgeschätzten Schmerzbelastungen abschließend auch nach Kompetenzgruppen differenziert dargestellt und diskutiert werden (s. Tabelle 23).

Im diagnostischen Interview gab jeder zweite Bewohner aus der Kompetenzgruppe der leicht Demenzkranken an, unter Schmerzen zu leiden, während sich von den auskunftsfähigen Bewohnern aus den verbleibenden Gruppen nur ungefähr jeder fünfte (MD, SD-P) bzw. jeder vierte Demenzkranke (SD-S) als gegenwärtig schmerzbelastet beschrieb (χ^2=8,7, df=3, p$<$.034). Für die selbsteingeschätzte Schmerzintensität konnten jedoch keine Unterschiede zwischen den Demenzgruppen bestätigt werden (F=0,4, df=3/41, p$<$.753).

Tabelle 23: Verschiedene Maße der Schmerzbelastung der vier Bewohnergruppen mit unterschiedlichem Muster erhaltener Kompetenzen

	Kompetenzgruppe			
N % bzw. N M±SD	LD	MD	SD-S	SD-P
Selbstauskunft				
akute Schmerzen	17 50%	22 24%	9 27%	2 22%
Intensität der akuten Schmerzen	15 1,9±0,7	21 2,1±0,8	8 1,9±0,4	1 2 ± 0
Fremdurteil (Pflegende)				
akute Schmerzen	21 64%	39 39%	19 39%	4 29%
Intensität der akuten Schmerzen	21 2,1±0,6	39 2,0±0,6	19 2,3±0,5	4 2,3±0,5
chronisches Schmerzleiden	17 61%	32 43%	14 37%	4 33%
Schmerzbeobachtung (Pflegende) - Ruhesituation				
Gesamtscore BESD (mit Trost)	32 2,8±2,1	100 3,5±3,2	47 4,3±2,7	15 5,5±3,5
Gesamtscore CNPI	32 1,0±1,4	100 1,3±1,9	47 1,8±2,0	15 2,8±2,0
Schmerzbeobachtung (Pflegende) - Aktivitätssituation				
Gesamtscore BESD (mit Trost)	34 3,3±2,9	99 4,1±3,3	47 6,6±2,7	15 6,7±3,3
Gesamtscore CNPI	34 1,2±1,6	99 1,7±2,0	47 3,0±1,9	15 3,7±2,9

Datenbasis: HILDE2 2006; N=214 (Diagnostik), N=199 (Pflegeinterview).

Die Pflegenden schätzten in allen Demenzgruppen deutlich mehr Bewohner als zur Zeit schmerzbelastet ein als sich zuvor zum Zeitpunkt des diagnostischen Gespräches selbst als unter Schmerzen leidend beschrieben hatten (McNemar's S=7,6, df=1, $p<.006$). Insgesamt betrug die Rate übereinstimmender Urteile zum Vorliegen von Schmerzen knapp 62 Prozent. In der Tendenz ist diese Übereinstimmung für die Gruppe der leicht demenzkranken Bewohner etwas reduziert, während für die Gruppe SD-S mit 70 Prozent die höchste Übereinstimmung zwischen Selbst- und Fremdauskunft erreicht wird. Wie zuvor für die Selbstauskunft berichtet, sind auch die fremdeingeschätzten Schmerzintensitäten zwischen den Gruppen nicht substanziell verschieden stark ausgeprägt (F=2,1, df=3/76, $p<.111$).

Wenngleich die Demenzgruppe der körperlich und kognitiv stark beeinträchtigten Bewohner (SD-S) am ehesten als multimorbid angenommen werden kann, ist die Rate bekannter chronischer Schmerzbelastung in der Gruppe der leicht Demenzkranken mit 61 Prozent am höchsten und nimmt mit steigendem Beeinträchtigungsgrad ab. Die gefundenen Unterschiede lassen sich jedoch nicht statistisch absichern (χ^2=4,6, df=3, $p<.206$).

In der Schmerzbeobachtung werden in beiden Untersuchungsbedingungen sowohl mit dem BESD- wie auch dem CNPI-Inventar jeweils deutlich unterschiedliche Schmerzbelastungen für Bewohner der unterschiedenen Kompetenzgruppen festgestellt. Dabei werden für die Gruppen LD und MD jeweils vergleichbare Schmerzbelastungen geschätzt und innerhalb der schwer demenzkranken Bewohner unterscheiden sich die mittleren Beobachtungsscores ebenfalls jeweils nicht. Deutliche Unterschiede können allerdings in beiden

Situationen mit beiden Instrumenten zwischen den stark kognitiv Beeinträchtigten (SD-S und SD-P) und den weniger stark Betroffenen (MD und LD) belegt werden (komplexer Kontrast SD-S/P vs. L/MD). Die Verhaltensbeobachtung zeichnet damit ein sowohl von der Selbst- als auch von der Fremdbeurteilung deutlich abweichendes Bild vom Ausmaß der Schmerzbelastung in unterschiedlichen Demenzschweregraden.

Auf eine Darstellung der Schmerzbelastung in Abhängigkeit von der (wahrscheinlichen) Ätiologie bzw. Form der Demenz soll im Rahmen dieser Arbeit aufgrund des in dieser Hinsicht anfallenden Samplings, der nur eingeschränkten Möglichkeiten einer Differenzialdiagnostik (insbesondere mit Blick auf die notwendigen bildgebenden Verfahren) und der für alle Demenzursachen mit Ausnahme der Alzheimer Demenz vergleichsweise geringen Häufigkeiten in der Stichprobe verzichtet werden. Insofern, als der durch die Untersuchung berücksichtigte Mix verschiedener Demenzschweregrade und Formen weitestgehend der im Pflegeheim üblicherweise anzutreffenden Verteilung entspricht, können die Ergebnisse an dessen statt dazu genutzt werden, die generelle Schmerzbelastung demenzkranker Heimbewohner im Allgemeinen abzuschätzen.

6.9 Validität der Schmerzerfassung

Durch die Konzentration der bisher dargestellten psychometrischen Analysen auf Aspekte der Reliabilität der Messinstrumente blieb die Frage nach der Validität der beiden untersuchten Schmerzassessments weitestgehend unbeantwortet.

6.9.1 Zusammenhang der BESD- und CNPI-Schmerzwerte

Da mit der BESD und CNPI-Skalen zwei konkurrierende Assessmentverfahren mit gleichem theoretischen Geltungsanspruch in der Studie Verwendung fanden, können die Übereinstimmungen in den durch beide Skalen erfassten Gesamtscores als erster Hinweis auf die konvergente Validität beider Instrumente gelten. Die in der Stichprobe ermittelten Pearson-Korrelationskoeffizienten von r=.63 (ohne Trost) bzw. r=.64 (mit Trost) in Ruhe und r=.74 (ohne/mit Trost) in der Aktivitätssituation können entsprechend als Hinweise darauf gewertet werden, dass beide Instrumente tatsächlich zu einem großen Teil ein gemeinsames Merkmal Schmerz messen.

Weiterführende Analysen identifizierten innerhalb beider Instrumente einzelne wenig diskriminative Indikatoren, und ermöglichten eine messfehlerbereinigte Abbildung des latenten Schmerzmerkmals. Da die Korrelation zweier mit einem Messfehler behafteter Messungen die wahren Merkmalszusammenhänge stets unterschätzt (vgl. Gl. 58), gelten Strukturgleichungsmodelle mit expliziter Differenzierung von Messstruktur und interessierendem Zusammenhang zwischen den wahren Merkmalswerten als die beste Methode zur Überprüfung der Konstruktvalidität (vgl. Skrondal & Rabe-Hesketh, 2004 p. 7). Auch beim Vergleich von BESD- und CNPI-Scores kann demnach von einer in Wirklichkeit eher stärkeren Evidenz für die Validität der Instrumente ausgegangen werden. Tatsächlich wird die messfehlerbereinigte Korrelation der wahren BESD- und CNPI-

Schmerzmessungen in Ruhe auf r=.94 und in Aktivität sogar auf r=1.0 geschätzt. Offensichtlich messen beide Instrumente dasselbe latente Merkmal oder Merkmalskomposit. Trotz dieser bestechenden Evidenz darf gefragt werden, ob nicht vielleicht beide Schmerzinventare etwas Anderes erfassen als Schmerz.

6.9.2 Zusammenhang mit Selbst- und Fremdauskunft

Weitere Hinweise auf den durch die Verhaltensbeobachtung abgebildeten Merkmalsbereich können durch den in Tabelle 24 dargestellten Vergleich mit der Selbst- und Fremdauskunft zur aktuellen Schmerzbelastung durch die Bewohner und Bezugspflegenden gewonnen werden.

Tabelle 24: Vergleich der BESD- und CNPI-Scores mit den Ratings zur Schmerzbelastung

	Ruhe		Aktivität	
	BESD	CNPI	BESD	CNPI
Selbstauskunft akute Schmerzbelastung[1]				
– gegenwärtig kein Schmerz	116 3,6±3,0	116 1,4±1,9	116 4,5±3,0	116 1,9±2,1
– akuter Schmerz	50 3,7±2,8	50 1,3±1,6	51 4,3±3,5	51 1,8±2,0
– keine Auskunft	27 4,0±3,2	27 2,0±1,9	27 6,1±4,0	27 2,9±2,1
Selbstauskunft Schmerzintensität[2]	45 −.13	46 −.01	45 −.04	46 −.01
Fremdurteil akute Schmerzbelastung[1]				
– gegenwärtig kein Schmerz	110 3,7±3,0	110 1,5±2,0	110 4,5±3,2	110 2,0±2,1
– akuter Schmerz	82 3,9±3,0	82 1,6±1,8	83 5,0±3,4	83 2,3±2,2
Fremdurteil Schmerzintensität[2]	82 .35	83 .25	82 .20	83 .12

[1] Mittlere Skalenscores: N M±SD; [2] Spearman-Korrelationen (N r).
Datenbasis: HILDE2 2006; N=195 (Ruhe) N=196 (Aktivität).

Vergleicht man die mittleren Scores der BESD und CNPI-Skala mit der Einschätzung des Vorliegens von Schmerzen, so fällt auf, dass auch bei solchen Bewohnern, die im diagnostischen Interview nicht über Schmerzen klagten oder von den Pflegekräften als gegenwärtig nicht schmerzbelastet eingeschätzt wurden, eine beträchtliche Anzahl potenzieller Schmerzindikatoren beobachtet wurden. Unterscheidet man nur zwischen als schmerzfrei und schmerzbelastet ausgewiesenen Bewohnern, lassen sich keine signifikanten Mittelwertsunterschiede in den beobachteten BESD- oder CNPI-Scores für diese Gruppen finden. Allerdings werden für solche Probanden, deren Auskunftsfähigkeit im diagnostischen Gespräch als zu stark beeinträchtigt eingeschätzt wurde, als dass valide Selbstauskünfte über ihr Schmerzerleben erhoben werden könnten, zumindest in der Aktivitätssituation deutlich – und im komplexen Kontrast mit den auskunftsfähigen Bewohnern statistisch auf dem 5-Prozent-Niveau signifikant – mehr möglicherweise schmerzbezogene Ausdrucksweisen dokumentiert.

Auch wenn man die Fremdeinschätzung aktueller Schmerzbelastung durch die Pflegenden betrachtet, lassen sich für die als gegenwärtig unter Schmerzen leidend beurteilten Bewohner in beiden Beobachtungssituationen zwar erwartungsgemäß höhere, jedoch keine substanziell gesteigerten mittleren BESD- oder CNPI-Scores beobachten als für die als zur Zeit im Allgemeinen schmerzfrei eingeschätzten Bewohner.

Für diejenigen Bewohner, die sich selbst als zur Zeit des diagnostischen Gesprächs als akut unter Schmerzen leidend beschrieben hatten, besteht kein Zusammenhang zwischen der angegebenen Schmerzintensität und dem im Rahmen der zeitlich späteren Beobachtung durch die Pflegenden dokumentierten schmerzbezogenen Ausdrucksverhalten.

In der Gruppe der durch die Pflegenden als zur Zeit schmerzbelastet eingeschätzten Bewohner korreliert die fremdeingeschätzte Schmerzintensität dagegen zumindest in der Ruhesituation substanziell (BESD: $r=.35$, $p<.002$; CNPI: $r=.25$, $p<.026$) mit der Anzahl konkret beobachteter schmerzbezogener Verhaltensweisen bzw. den BESD- und CNPI-Scores. Die für die Aktivitätssituation gefundenen geringeren Zusammenhänge lassen sich nicht mehr statistisch sichern.

Dabei ist anzumerken, dass sich die Selbstauskunft zu akut vorliegenden Schmerzen und deren Intensität auf einen anderen Zeitpunkt bezieht als die Fremdbeurteilung und die schmerzbezogene Verhaltensbeobachtung, womit der Beitrag eines solchen Vergleiches für die Abschätzung der Validität der Beobachtungsskalen prinzipiell eingeschränkt erscheint. Vor diesem Hintergrund können die leicht gesteigerten Skalenscores für die per Fremdurteil als schmerzbelastet eingeschätzten Bewohner und die doch deutlich gesteigerten Zusammenhänge dieser mit der fremdeingeschätzten Schmerzintensität auch aufgrund der größeren zeitlichen Nähe zwischen Fremdurteil und Beobachtung schon eher als ein Hinweis auf die Validität der Schmerzbeobachtung gewertet werden. Sicherlich kann erwartet werden, dass eine Schmerzbeobachtung sich auch an einem vorangegangenen allgemeinen Schmerzurteil orientiert.

6.9.3 Zusammenhang mit Kompetenzbeeinträchtigungen

Die zuvor berichteten Ergebnisse zur Vorhersagbarkeit von Ruheschmerz und Schmerzveränderung bei Aktivität durch mobilitätsbezogene Beeinträchtigungen können auch als Hinweis auf die Validität der Schmerzskalen (oder zumindest des letztlich berücksichtigten Subsets schmerzbezogener Verhaltensindikatoren) betrachtet werden. Die Bedeutung der gefundenen systematischen Bezüge zum kognitiven Status und vor allem der nichtkognitiven Demenzsymptome für die Validitätsbeurteilung verhaltensbezogener Schmerzmessung jedoch ist vor dem Hintergrund des gegenwärtig unvollständigen Kenntnisstandes und der mitunter widersprüchlichen empirischen Befunde zum Zusammenhang der demenziellen Symptomatik und dem Schmerzerleben bzw. -ausdruck schwer abzuschätzen. Insoweit, wie beispielsweise das Neuropsychiatrische Inventar Verhaltensmerkmale misst (z.B. Wahn, Halluzination), die theoretisch nicht mit Schmerzen verbunden sind, müsste der empirisch nachgewiesene Zusammenhang als Hinweis auf eine unzureichende *diskriminante* Validität des Schmerzassessments gewertet werden. Insoweit aber, wie theore-

tisch sinnvolle Bezüge zwischen beiden Merkmalsbereichen hergestellt werden können (z.B. zwischen Schmerz und Aggression oder Depression), könnten die gefundenen Zusammenhänge als Hinweise auf die *nomologische* Validität des Schmerzassessments gelten.

7 Diskussion

Die vorliegende Arbeit hatte sich zum Ziel gesetzt, die gegenwärtige Praxis der verhaltensgestützten Schmerzmessung bei demenzkranken Menschen mit Blick auf die Möglichkeiten und Grenzen der zu diesem Zweck vorgeschlagenen Verhaltensinventare kritisch aufzuarbeiten.

Damit wird an einige jüngst veröffentlichte Übersichtsarbeiten angeschlossen, die ebenfalls eine Bewertung der bislang vorgeschlagenen Verfahren zur Schmerzerfassung bei Menschen mit einer demenziellen Erkrankung und entsprechend eingeschränkten Möglichkeiten zur schmerzbezogenen Selbstauskunft zu leisten und Empfehlungen hinsichtlich der Wahl eines spezifischen Verfahrens zu geben suchen. Dabei gelangen alle Arbeiten übereinstimmend zu dem Schluss, dass die psychometrische Güte einzelner Instrumente aufgrund der jeweils wenigen empirischen Arbeiten und der Heterogenität der Verfahren gegenwärtig nicht abschließend bewertet werden könne. Angesichts der dünnen empirischen Befundlage wird die Forderung stark gemacht, die mittlerweile verfügbaren Instrumente einer gründlicheren empirischen Testung zu unterziehen, anstatt stets neue Verfahren vorzuschlagen. Mit dem Einsatz der beiden international recht weit verbreiteten Verhaltensinventare CNPI und PAINAD an einer Stichprobe von nahezu zweihundert demenzkranken Menschen in stationären Einrichtungen der Altenpflege kommt die vorliegende Studie dieser Forderung nach und stellt die Beurteilung der psychometrischen Güte dieser Verfahren prinzipiell auf eine breitere Basis.

Die vorgelegte Arbeit geht jedoch deutlich über diesen *additiven* Beitrag hinaus, indem sie die Hindernisse einer inhaltlichen Bewertung der bislang vorgeschlagenen Assessmentverfahren herausarbeitet und methodisch neue Wege einer angemesseneren Beurteilung ihrer psychometrischen Eigenschaften aufzeigt.

Als Herausforderungen für eine vergleichende Diskussion konkurrierender Verfahren wurden insbesondere Unterschiede in den erfassten Verhaltensbereichen, im Skalenaufbau und Itemscoring oder der Beobachtungsdauer herausgestellt. Die Schmerzeinschätzung demenzkranker Menschen erscheint darüber hinaus in der Praxis durch eine Vielzahl nur schwer zu kontrollierender Kontextfaktoren bestimmt, welche die Interpretierbarkeit der beobachteten Verhaltensweisen als Schmerzausdruck in Frage stellen. Besondere Aufmerksamkeit wurde diesbezüglich in der vorliegenden Arbeit dem Grad der körperlichen Aktivierung der Bewohner gewidmet. Schwierigkeiten für eine Synopse des Geltungsbereiches verschiedener vorgeschlagener Verfahren ergeben sich auch aus der Heterogenität des Demenzsyndromes selbst. Gegenwärtig scheinen die zur Schmerzbeobachtung bei demenzkranken Menschen vorgeschlagenen Verfahren eher darum zu ringen, ihre prinzipi-

elle Anwendbarkeit auch in dieser schwierigen Population nachzuweisen, als dass sie ihre spezifische Eignung für diese – hinsichtlich ihres Schmerzerlebens und -ausdrucks vielschichtige und hochgradig dynamische – Personengruppe belegen.

Aus der Kritik an der bestehenden Diskrepanz zwischen dem heterogenen Anwendungsfeld der Schmerzerfassung bei demenzkranken Menschen und den restriktiven Konzepten der sowohl die gegenwärtige Praxis der Skalenentwicklung als auch -bewertung dominierenden klassischen Testtheorie werden spezifische Anforderungen an eine stärker inhalts- und kontextbezogene Testtheorie für den Forschungsbereich der Schmerzmessung bei Demenz abgeleitet. Eine angemessene methodische Konzeption begreift danach den Schmerz als eine unbeobachtete latente Merkmalsdimension (Latent Variable Modelle), die durch konkrete Verhaltensindikatoren mit einem (hinsichtlich Reliabilität und indiziertem Merkmalsniveau) jeweils unterschiedlichen Bezug zum Schmerzerleben angezeigt wird (Item Response Theorie), und die ihren Ausdruck in verschiedenen Erfassungskontexten (Messinvarianz) oder über die Zeit hinweg (längsschnittliche Modellierung) gegebenenfalls auf unterschiedliche Weise finden kann.

Mit der Zusammenführung der beiden Traditionen der Item-Response-Theorie und Latent Variable Modelle nimmt diese Arbeit auch eine nicht geringe methodische Herausforderung an. Insbesondere die Übertragung der im wesentlichen an Leistungstests entwickelten Logik probabilistischer Verfahren auf die schmerzbezogene Verhaltensbeobachtung, die anwendungsorientierte Integration der doch sehr verschiedenen Anschauungen zur Modelltestung und -optimierung, und nicht zuletzt die formale Darstellung wesentlicher Modelle aus beiden Bereichen anhand einer einheitlichen Notation können dabei als ein eigenständiger Beitrag dieser Arbeit für die auch methodische Weiterentwicklung des Forschungsfeldes begriffen werden.

Am Beispiel der beiden in der zweiten Feldphase des HILDE-Projektes berücksichtigten deutschen Adaptionen der Schmerzinventare PAINAD (BESD) und CNPI wurden die erweiterten Möglichkeiten entsprechender statistischer Verfahren für die Diskussion der psychometrischen Eigenschaften beider Skalen veranschaulicht. Um der Forderung nach einer stärkeren Orientierung an den durch die jeweiligen Inventare vorgeschlagenen potenziellen Schmerzindikatoren nachzukommen, wurden dabei alle in den Instrumenten aufgeführten konkreten Verhaltensweisen zur Beobachtung vorgegeben und analysiert. Die der ursprünglichen Skalenkonstruktion zugrundeliegenden Annahmen zur Indikationsgüte der Einzelitems und die durch die Items jeweils angezeigte Schmerzintensität wurden dabei sowohl auf der Grundlage der herkömmlichen Itemanalyse, als auch der durch ein IRT-Modell geschätzten Itemparameter diskutiert.

Da für die meisten der in den Inventaren enthaltenen Ausdrucksbereiche mehrere einzelne Verhaltens- bzw. Ausdrucksweisen beschrieben sind, wurden mithilfe der probabilistischen Verfahren zunächst die Zusammenhänge zwischen den Indikatoren innerhalb der Ausdrucksbereiche analysiert, bevor die Itemparameter schließlich auf der Grundlage des implizierten Second-Order-Faktorenmodelles geschätzt wurden, um entsprechende bereichsspezifische und -übergreifende Anteile des Schmerzausdrucks differenzieren zu können.

Die in Kapitel 3.4.5 referierten Arbeiten, die mehrere konkurrierende Beobachtungsverfahren zur Schmerzmessung bei Demenz parallel an derselben Stichprobe einsetzten, haben deutlich gemacht, dass die Möglichkeiten, die für mehrere Schmerzinventare separat geschätzten Item- und Skalenkennwerte direkt miteinander zu vergleichen sehr (nämlich auf den Vergleich von Beobachtungsraten und den korrelativen Vergleich der Gesamtskalenwerte) beschränkt bleiben. Um in der vorliegenden Arbeit dagegen eine direkte Gegenüberstellung der beiden Instrumente bzw. ihrer Indikatoren möglich zu machen, wurden in einem weiteren Analyseschritt darum alle Einzelindikatoren auf dem gemeinsamen Merkmalsfaktor Schmerz skaliert. Da item-response-analytische Verfahren auf der Grundlage von Latent Variable Modellen es erlauben, Indikatoren mit unterschiedlicher Kategorienanzahl in einem gemeinsamen Modell zu berücksichtigen, konnte der Ausdrucksbereich Trost der BESD seiner inneren Logik folgend durch einen dreistufigen Indikator repräsentiert werden. Während die Reliabilität, und folglich auch der Informationsgehalt der Schmerzassessments im Kontext der vorgestellten konventionellen Skalenanalyse als über den gesamten latenten Merkmalsraum hinweg konstant angenommen werden muss, aggregieren sich die durch die verwendete IRT-Modellierung geschätzten individuellen Beiträge der Einzelitems zu einer über verschiedene Schmerzausprägungen hinweg variierenden Informationsfunktion für das Gesamtinventar, und entsprechend konnten BESD und CNPI hinsichtlich ihrer Abbildungsbereiche, der Präzision der Schmerzmessung, und ihrer relativen Effizienz über diese Bereiche hinweg direkt kontrastiert werden.

Aus der großen Menge der verschiedenen in Kapitel 3.4.6 identifizierten Merkmale des Praxisfeldes Schmerzmanagement in der (stationären) Versorgung demenzkranker Menschen, die auf die verhaltensbezogene Schmerzerfassung Einfluss nehmen können, wurde im Rahmen der vorliegenden Studie der Aspekt der Aktivität herausgegriffen und analysiert. Die Gültigkeit der Interpretation der in durch Bewegung und Aktivität charakterisierten Beobachtungssituationen – wie auch in dieser Untersuchung – häufig gefundenen höheren Scores für verschiedene Verhaltensinventare als gesteigerte Schmerzbelastung ist allerdings nur dann gesichert, wenn die eingesetzten Verhaltensinventare unter der Bedingung geringer und hoher Aktivierung bzw. Bewegung identisch funktionieren. Ein systematischer Nachweis einer solchen Invarianz des Messinstrumentes über verschiedene Situationen hinweg wird in den bisherigen Veröffentlichungen zu den vorgeschlagenen Assessments nicht geführt. In der vorliegenden Arbeit werden dagegen Unterschiede in der Struktur des Schmerzkonzeptes als ganzem und im Referenzrahmen für die Beobachtung einzelner Verhaltensmerkmale, die sich zwischen einer Ruhe- und einer Aktivitätssituation ergeben, systematisch überprüft um zuverlässige Aussagen zur Schmerzveränderung der Bewohner über beide Beobachtungsbedingungen hinweg machen zu können. Die geschätzte individuelle aktivitätsinduzierte Veränderung im Schmerzerleben wurde dabei theoretisch als latenter Vulnerabilitätsfaktor demenzkranker Menschen herausgestellt und nachfolgend mit verschiedenen kognitiven, nicht-kognitiven und alltagspraktischen Kompetenzen (d.h. also mit den hier unterschiedenen Kernmerkmalen des Demenzsyndroms) der Bewohner in Bezug gesetzt.

Weder auf der Grundlage der klinisch experimentellen noch der neuropathologischen

Befundlage oder der durch die Pflegenden selbst berichteten Schmerzkennzeichen wurde bislang ein kohärentes Bild der demenzbedingten Veränderungen im Schmerzerleben und Schmerzausdruck entwickelt, an dem sich die Entwicklung eines tatsächlich demenzspezifischen Assessmentinstrumentes hätte orientieren können. Mit wenigen Ausnahmen scheinen verhaltensbezogene Schmerzassessments für Menschen mit Demenz ihre Angemessenheit für dieses spezifische Klientel ausschließlich aus dem Verzicht auf die Selbstauskunft durch die Betroffenen abzuleiten. Einige Verhaltensinventare – darunter auch die hier betrachtete BESD und CNPI – erfassen jedoch mehr oder minder direkt auch verbale schmerzbezogene Äußerungen der demenzkranken Menschen. Einen vergleichsweise eng auf die nicht-kognitive Symptomatik der Demenz ausgerichteten Zugang wählen Mahoney und Peters (2008) mit ihrem Versuch, Schmerzerleben und nicht-schmerzbezogene Agitation voneinander zu differenzieren. Letztlich bleibt zu konstatieren, dass sich demenzkranke Menschen in ihrem Erleben und Verhalten durch einen Mix von natürlichen Alters- und Krankheitsprozessen auszeichnen, weswegen sich ein demenzspezifisches Schmerzassessment vielleicht überhaupt nicht ausschließlich auf die Kernsymptome der Demenzerkrankung beschränken *sollte*. Sicherlich aber greift eine Beschränkung auf die Denk- und Gedächtnisfähigkeit im Sinne eines MMST-Wertes für die Binnendifferenzierung der Gruppe demenzkranker Menschen zu kurz. Im Rahmen der vorliegenden Arbeit wurde eine alternative Kategorisierung von Prägnanztypen des Demenzsyndroms auf der Grundlage des Musters erhaltener Kompetenzen in den Bereichen Kognition, Verhaltensauffälligkeit und alltagspraktische Fähigkeiten aufgegriffen. Dabei wurden die unterschiedlichen Möglichkeiten zur verbalen und non-verbalen (Schmerz-)Kommunikation von vier Kompetenzgruppen demenzkranker Bewohner mit unterschiedlichem Grad der kognitiven Beeinträchtigungen – und im Falle schwerer intellektueller Einbußen mit unterschiedlicher somatischer oder psycho-pathologischer Lagerung – untersucht.

Den Abschluss der Untersuchungen zur verhaltensgestützten Schmerzeinschätzung auf der Grundlage der HILDE-Daten bildet die Zusammenschau der verfügbaren empirischen Hinweise auf die Validität der beiden Beobachtungsinstrumente.

7.1 Identifizierte Bedarfe und Potenziale

Aus der Zusammenschau der bisherigen Befunde zur Bewertung von Verfahren zur verhaltensgestützten Schmerzmessung bei Menschen mit Demenz konnten Entwicklungsbedarfe auf drei Dimensionen identifiziert werden. Dabei sind selbstverständlich die Einzelaspekte einer gegenwärtig noch nicht hinreichenden Datenbasis, einer nur ungenügend ausdifferenzierten, oder zumindest in den vorgeschlagenen Instrumenten nicht angemessen umgesetzten konzeptionellen Entwicklung sowie die nur zögerliche Übernahme neuer testtheoretischer und statistischer Ansätze in den stark klinisch geprägten Forschungsbereich Schmerzassessment bei Demenz nicht vollkommen unabhängig voneinander diskutierbar.

Insbesondere die Diskussion der zuvor in einem vergleichsweise eigenständigen Teil der Arbeit abgehandelten spezifischen Potenziale aktueller Methoden für die Schmerzbeobachtung bei Demenz soll darum gemeinsam mit den entsprechenden Bedarfen erfolgen.

7.1.1 Datenbasis

Ein erstes offensichtliches Problem für die Abschätzung der Güte der meisten vorgeschlagenen Verhaltensinventare besteht in der gegenwärtig noch dünnen empirischen Datenbasis. Paradoxerweise liegen für den Großteil der vorgeschlagenen Instrumente mittlerweile mehr Reviews vor als Originalarbeiten. Die in den empirischen Arbeiten zur Entwicklung oder Evaluation der Verfahren realisierten Stichproben sind darüber hinaus in aller Regel klein. Insoweit, wie dieser geringen Stichprobengröße eine inhaltlich detailreiche Exploration des Forschungsfeldes auf der Grundlage qualitativer Ansätze entgegengestellt wird, muss diese Dominanz kleinerer Studien als Indikator für eine noch nicht hinreichend entwickelte Methodologie zur Beantwortung der in der Praxis aufscheinenden Fragen gewertet werden. Auch die offenbar präferierte Neukombination von Erfassungsinhalten aus bereits bestehenden Verfahren und die Anpassung bzw. Ergänzung dieser Instrumente zeigen an, dass die praktischen Bedarfe und situativen Umstände durch das Konvolut verfügbarer Verfahren offensichtlich noch nicht hinreichend gut adressiert sind. Entsprechend sollte diesen Kontextmerkmalen bei der Entwicklung und Überprüfung von Verfahren zukünftig eine größere Aufmerksamkeit zukommen. Ein Teil des Problems der geringen empirischen Basis resultiert auch aus dem Umstand, dass die gegenwärtigen Verfahren zur psychometrischen Beurteilung der Skalen als ihren vorrangigen Gegenstand das Gesamtinstrument en bloc definieren. Die ausschließliche Orientierung am Gesamttest macht es aber schwer, die in verschiedenen Instrumenten enthaltenen gemeinsamen Anteile herauszuarbeiten und zu bewerten. Geht man davon aus, dass die Selektion von Verhaltensweisen zur Schmerzbeobachtung nicht völlig willkürlich geschieht, sondern sich an den Vorerfahrungen früherer empirischer Arbeiten und einer besonders guten (augenfälligen) Passung zum Forschungsfeld und -zweck orientiert, so sollten insbesondere diese gewissermaßen kondensierten Kernanteile aus verschiedenen Verfahren stärker in den Blick genommen werden. Neue Chancen für eine bessere Integration der Erkenntnisse aus verschiedenen Studien mit nur teilweise identischen Skalen in einen gemeinsamen Wissenskorpus bieten auch die in dieser Arbeit nur knapp angerissenen probabilistischen Ansätze des *Test Equatings* (vgl. Kap. 4.3.2.2).

7.1.2 Konzeptionelle Entwicklung

Die gegenwärtig verfügbaren Instrumente zur beobachtungsgestützten Schmerzmessung bei Demenz bleiben zur Zeit deutlich hinter dem aktuellen Kenntnisstand zum Schmerzerleben bei Demenz und zur Versorgungssituation demenzkranker Menschen zurück. Charakteristische Merkmale des Alters aber auch der Demenz selbst werden bei der Konzeption der Schmerzerfassung nicht berücksichtigt. Auch hier scheinen die bereits zuvor beschriebenen Aspekte der mangelnden Orientierung am eigentlichen Erfassungsinhalt und den die Schmerzmessung bestimmenden Kontextfaktoren, sowie die nur zögerliche Nutzung der Potenziale neuerer methodischer Ansätze für die Beurteilung von Assessmentinstrumenten und ihrer theoretisch angenommenen Eigenschaften auf.

7.1.2.1 Strukturierung schmerzbezogener Verhaltensweisen

Was in einzelnen Verfahren zum Schmerzassessment durch Verhaltensbeobachtung als ein Item definiert wird, variiert mitunter beträchtlich. Neben molekularen Bestandteilen schmerzbezogenen Verhaltensausdrucks (z.B. Kontraktionen einzelner Gesichtsmuskeln) werden auch komplexere Gesten (z.B. sich entziehen) oder gar langfristige Verhaltensänderungen (z.B. in Sozialkontakten) zur Beobachtung vorgegeben. Einige neuere Arbeiten nehmen auf die durch das AGS Panel for Persistent Pain in Older Adults vorgeschlagenen übergeordneten Kategorien schmerzbezogenen Ausdrucksverhaltens (s. Tab. 3) Bezug. Im Einzelnen erscheinen sich jedoch die verwendeten Definitionen von Verhaltensweisen und Ausdruckskategorien mitunter auf deutlich unterschiedlichen Abstraktionsniveaus zu bewegen, was eine klare Abgrenzung nahezu unmöglich macht.

Die drei durch das AGS Panel vorgeschlagenen Kategorien Gesichtsausdruck, Vokalisation und Körperbewegung erscheinen besonders für eine Abbildung akuten Schmerzerlebens geeignet. Das Verhältnis von Schmerzerleben und Verhaltensausdruck muss dabei jedoch nicht immer der implizierten simplen Sequenz folgen. Ängstliches oder vermeidendes Verhalten aufgrund einer Antizipation von Schmerzreizen, selbst wenn diese durch eine demenzbedingte Beeinträchtigung der evaluativ-kognitiven Schmerzverarbeitung herabgesetzt sein mag, mögen wohl Schmerz anzeigen, allerdings nicht unbedingt eine aktuelle Schmerzbelastung in der Beobachtungssituation. Einzelne Verfahren greifen zur Abbildung akuter Schmerzbelastung daneben auch auf physiologische Schmerzmarker zurück. Die Befunde zur Adaption der vegetativen Reaktion bei chronischen Schmerzen, und einer bei Demenzkranken reduzierten vegetativen Schmerzreaktion sprechen eher gegen eine Berücksichtigung dieser Verhaltenskategorie. Für den durch die BESD-Skala eingebrachten stark auch vegetativ bestimmten Ausdrucksbereich Atmung wurde in der vorliegenden Arbeit tatsächlich eine insgesamt vergleichsweise geringe Indikationsgüte zur Abbildung von Schmerzen nachgewiesen.

In einigen der referierten Übersichtsarbeiten wurde das Argument stark gemacht, man sollte den Blick zukünftig nicht allein auf unmittelbare schmerzbezogene Verhaltensreaktionen richten, sondern darüber hinaus auch längerfristige subtile Veränderungen in den Gewohnheiten, Vorlieben und im psychischen Befinden demenzkranker Menschen als potenziell schmerzbedingte Verhaltensweisen berücksichtigen. Da weder die CNPI noch die BESD explizit[13] entsprechende Verhaltensindikatoren zu längerfristigen Verhaltensänderungen enthalten, können auf der Grundlage der hier geleisteten empirischen Bearbeitung keine Aussagen zum Stellenwert dieser subtileren Schmerzindikatoren gemacht werden. Bei der Darstellung der Schwierigkeiten, die sich aus einer nur mangelnden Berücksichtigung zeitbezogener Kontextmerkmale der Schmerzmessung und der Eigendynamik des Schmerzerlebens selbst für die Bestimmung der Reliabilität und Validität der vorgeschlagenen Skalen ergeben, konnten zumindest auf theoretischer Ebene einige Voraussetzungen für eine sinnvolle Ergänzung der herkömmlichen Verfahren um diese veränderungsbezo-

[13] Natürlich könnte man sich fragen, ob das BESD-Item trauriger Gesichtsausdruck nicht implizit als langfristige affektive Folge (chronischer) Schmerzbelastung angenommen werden muss.

genen Schmerzmarker herausgearbeitet werden. Nicht zufriedenstellend gelöst ist dabei gegenwärtig vor allem die Frage nach einem angemessenen Referenzzeitraum für die Bestimmung von Gewohnheiten und üblichem Verhalten, und damit verbunden die Frage nach der zu fordernden Stabilität der beobachteten Verhaltensänderung – insbesondere vor dem Hintergrund einer chronisch progredienten Erkrankung mit teilweise diskontinuierlichem Verlauf, aber auch möglicher Adaptionsprozesse – selbst. Es kann davon ausgegangen werden, dass eine Abschätzung solcher Veränderungen praktisch wohl eher über den Einbezug retrospektiver Urteile durch Pflegende oder Angehörige erfolgen wird als durch die systematische Wiederholung konkreter Verhaltensbeobachtung, womit dem unterschiedlichen Erfahrungshintergrund dieser Informantengruppen auf jeden Fall Beachtung geschenkt werden müsste.

Ein besonderes Problem für die Weiterentwicklung von Schmerzassessments ergibt sich auch daraus, dass nicht nur konkrete Verhaltensweisen zur Beobachtung vorgegeben werden, sondern häufig konkret beobachtbares Verhalten lediglich als *beispielhafter* Schmerzausdruck für eine abstraktere übergeordnete Verhaltenskategorie beschrieben ist. Im Einzelfall kann dann jedoch nicht mehr nachvollzogen werden, auf welcher konkreten Grundlage dieses Bereichsscoring beruht. Durch die Vorgabe von Beispielitems soll sichergestellt werden, dass die Beobachter ihrem Rating ein vorgesehenes Verständnis der übergeordneten Verhaltenskategorie zugrundelegen, und nicht ausschließlich ihrem subjektiven Verständnis folgen, das aus der Sicht der Skalenentwickler gegebenenfalls zu eng oder zu breit wäre. Inwieweit sich die Einschätzenden tatsächlich an diesen Vorgaben orientieren bleibt unklar. Es bleibt letztlich die Aufgabe der Forschenden, reliable und valide konkrete Verhaltensindikatoren mit möglichst geringem Interpretationsbedarf zu identifizieren. Das wird aber nur möglich sein, wenn auch die Wahrnehmungsgewohnheiten der Pflegenden berücksichtigt werden, und es gelingt, potenziellen Schmerzausdruck auf einer alltagsrelevanten Auflösungsebene zu beschreiben. Sowohl fragmentierte molekulare, als auch abstrakt-konzeptionelle Verhaltensindikatoren erscheinen dafür wenig geeignet, weswegen die Zukunft der verhaltensbezogenen Schmerzmessung bei demenzkranken Menschen meines Erachtens weder durch FACS, noch durch die drei unkommentierten Einzelitems Mimik, Körperhaltung und Lautäußerung bestimmt sein wird (vgl. dagegen Prkachin, 2007).

Eine interessante Komplikation auch für die Suche nach dem besten Auflösungsgrad für die Einzelindikatoren eines verhaltensgestützten Schmerzassessments bringen Verfahren ein, die strukturell eine Beobachtung unter mehreren Bedingungen vorgeben und die Ergebnisse dieser Messungen gewissermaßen als Einzelitems miteinander verrechnen. Die Originalskala CNPI sieht eine solche Summation der in Ruhe und bei Bewegung beobachteten Skalenwerte vor, und auch ECPA, EPCA-2, BISAD und in besonderem Maße MOBID kombinieren Situationen als Einzelitems zu einem Gesamtscore. Die Diskussion der Eigenschaften einzelner konkret zu beobachtender schmerzbezogener Verhaltensweisen wird damit noch einmal schwieriger. Tiefgreifend sind auch die Implikationen für die Bestimmung des Messgegenstandes Schmerz selbst. Mag die Kombination der bei verschiedenen standardisierten Bewegungen beobachteten Schmerzwerte vielleicht noch als

multilokale Bewegungsschmerz-Gesamtbelastung verstanden werden, fallen entsprechende Beschreibungen für die Kombination von Ruhe- und Bewegungs- oder Pflegesituationen schon schwerer.

7.1.2.2 Angezeigte Schmerzintensität

Ein substanzieller Erkenntnisbedarf besteht gegenwärtig auch mit Blick auf den Zusammenhang zwischen der erlebten Schmerzintensität und dem schmerzbezogenen Ausdrucksverhalten. Das ist bemerkenswert auch darum, weil schließlich alle in dieser Arbeit referierten Verhaltensinventare von der Grundannahme ausgehen, dass mit der Intensität des Schmerzes auch die Wahrscheinlichkeit für die Beobachtung der beschriebenen schmerzbezogenen Verhaltensweisen ansteigt, und darum hohe Skalenwerte mehr Schmerz anzeigen, und geringe Skalenwerte weniger Schmerz. Diese Annahme ist sicherlich nicht trivial; wird sie jedoch in Zweifel gezogen, verlieren de facto alle bisherigen Bemühungen um eine Entwicklung und Bewertung entsprechender Verfahren ihre konzeptionelle Grundlage.

Selbst Messzugänge, die qualitative Übergänge im Schmerzerleben beschreiben wie beispielsweise Ansätze zur Bestimmung von Schmerz- und Toleranzschwelle, aber auch Verhaltensinventare, die vermeintlich zurückhaltend einen Cut-off-Wert vorsehen, ab dem höchst wahrscheinlich Schmerzen vorliegen, können sich nicht klar vom Prinzip eines latenten, auf der Dimension der Empfindungsintensität gesteigerten Schmerzkontinuums lösen.

In die Berechnung des Gesamtskalenwertes fließen die beschriebenen konkreten Verhaltensweisen bei den meisten Verfahren gleichgewichtet ein, es werden also keine bestimmten Aussagen zu Unterschieden in der durch die Einzelitems jeweils angezeigten Schmerzintensität gemacht. Die BESD trifft dagegen explizite Annahmen zur relativen Schmerzintensität, die durch einen bestimmten (beispielhaften) Verhaltensausdruck jeweils angezeigt wird, indem bestimmte Verhaltensweisen stärker gewichtet werden als andere.

Die Möglichkeiten zur Bestimmung der Schwierigkeit eines Verhaltensindikators auf der Grundlage der üblichen Skalenanalyse sind durch die Stichprobenabhängigkeit dieses Kennwertes stark eingeschränkt. Vereinzelt wird eine stärkere Berücksichtigung der Auftretenshäufigkeiten einzelner potenzieller Schmerzindikatoren gefordert. Allerdings geschieht dies weniger dazu, um die implizierte Schmerzintensität abzuschätzen, sondern vielmehr um das Instrumentarium um seltene Verhaltensindikatoren zu ‚erleichtern'.

Insbesondere für Verfahren zur Schmerzmessung, die den Anwender bei der Entscheidungsfindung, beispielsweise hinsichtlich einer Schmerztherapie oder weiterer diagnostischer Schritte unterstützen sollen, ist zu fordern, dass der abgebildete Merkmalsbereich hinreichend gut beschrieben ist und in kritischen Schmerzabschnitten eine besonders gute Diskrimination geleistet wird.

7.1.2.3 Kontextfaktoren

Die wahrscheinlich größte Einschränkung erfahren die zur Zeit verfügbaren Verhaltensinventare dadurch, dass sie die Umstände der Schmerzmessung weitestgehend außer Acht lassen. Selbst wenn situative Bedingungen, wie beispielsweise durch die wiederholte Erfassung in Ruhe und Aktivität berücksichtigt werden, bleibt in der Regel unklar wie die erfassten Merkmalswerte zueinander in Bezug gesetzt werden können.

Der Begriff des Kontextes wurde durch die vorliegende Arbeit verhältnismäßig breit definiert. Herausgearbeitet wurden als die Schmerzmessung beeinflussende Faktoren dabei zum ersten Charakteristika der demenzkranken Menschen selbst (z.B. Alter, demenzbedingte Beeinträchtigungen, Aktivierung), zum zweiten Merkmale der beobachtenden Personen (z.B. Expertise), und schließlich auch Aspekte der zeitlichen Dimension der Schmerzerfassung (z.B. Beobachtungsdauer, Eigendynamik des Schmerzerlebens).

Mit der Berücksichtigung von Unterschieden zwischen Bewohnern mit verschiedenen Graden oder Mustern ihrer demenziellen Beeinträchtigungen verbinden sich vorrangig Fragen nach der Demenzspezifität der verhaltensgestützten Schmerzmessung. Unterschiede in der Aktiviertheit bzw. dem Ausmaß der Mobilisation der Bewohner werden dagegen häufig auch als systematische Bedingungsvariation bewusst hergestellt, und dienen aufgrund der bekannten schmerzrelevanten muskulo-skelettalen Degeneration und der so bei Bewegung erwartbaren höheren Schmerzbelastung nicht selten als Nachweis der Validität des Schmerzassessments. Unterschiede in der Schmerzeinschätzung, die auf Merkmale der Beobachter zurückgeführt werden können, werden dabei als Beeinträchtigung der Objektivität und Reliabilität der Verfahren diskutiert. Ein hoch bedeutsames Kontextmerkmal, das in seiner Bedeutung für die Schmerzbeobachtung in der Regel nicht hinreichend diskutiert wird, ist die Dauer der Beobachtung. Während es bei nicht-teilnehmenden Beobachtungen in der Regel problemlos möglich ist, ein Beobachtungsintervall vorzugeben, finden einige besonders relevante Schmerzbeobachtungen in Situationen statt, in die der Beobachter selbst eingebunden ist (v.a. Pflegesituationen), und in der keine einheitliche Definition eines Beobachtungszeitraumes möglich ist. Insbesondere für die wiederholte Schmerzmessung fehlt es gegenwärtig noch an einer konzisen Beschreibung der erwartbaren Veränderungen bei einzelnen schmerzbezogenen Verhaltensindikatoren, aber auch des zu messenden Schmerzes an sich. Gütekriterien mit einem Zeitbezug wie die prädiktive Validität oder die Test-Retest-Reliabilität erscheinen solange wenig aussagekräftig, wie diese zeitliche Dynamik nicht hinreichend theoretisch beschrieben ist.

Tatsächlich sind damit eine beträchtliche Menge verschiedener, die beobachtungsbasierte verhaltensbezogene Schmerzmessung potenziell konfundierender Faktoren beschrieben, und noch eine ganze Reihe weiterer impliziert. Um die aufgezeigten Interpretationsprobleme zu überwinden, wurden in dieser Arbeit die erweiterten Möglichkeiten neuer methodischer Ansätze herausgearbeitet, die Vergleichbarkeit der Schmerzmessung unter verschiedenen Erfassungsbedingungen, in verschiedenen Gruppen demenzkranker Heimbewohner, oder über mehrere Zeitpunkte hinweg systematisch zu überprüfen. Als zwei zentrale Aspekte der Invarianz einer Messung werden dabei die Bedeutungsgleich-

heit des Schmerzkonstruktes, und die Äquivalenz der für die (Nicht-)Beobachtung einzelner Indikatoren herangezogenen Schwellenwerte unter mehreren Erfassungsbedingungen identifiziert. Hierbei zeigt sich in besonderer Weise auch der Mehrgewinn einer kombinierten Betrachtung von faktoranalytischen und probabilistischen Ansätzen, da diese Facetten der Messinvarianz in beiden Konzeptionen ein deutlich unterschiedliches Gewicht besitzen, für die Analyse der Äquivalenz einer beobachtungsgestützten Schmerzmessung jedoch beide von Bedeutung sind.

7.2 Diskussion zentraler Befunde aus dem HILDE-Projekt

Mit dieser Arbeit wurden eine Reihe von Möglichkeiten aufgezeigt, wie die bislang häufig impliziten Annahmen zum Verhältnis von Schmerzerleben und schmerzbezogenem Verhaltensausdruck, die eine beobachtungsgestützte Schmerzerfassung bestimmen, expliziert und einer empirischen Überprüfung zugeführt werden können.

Vorrangig zum Zwecke der weiterführenden Konstruktvalidierung der BESD-Skala, für die im Rahmen der ersten empirischen Untersuchungsphase des Projektes HILDE bereits erste psychometrische Befunde erarbeitet wurden, wurde für die zweite empirische Untersuchungsphase des HILDE-Projektes zusätzlich die CNPI-Skala eingesetzt.

Nachfolgend soll der konzeptuelle und empirische Mehrgewinn einer Analyse dieser beiden für das beobachtungsgestützte Schmerzassessment bei Demenz recht repräsentativen Verhaltensinventare auf der Folie der aktuellen statistischen Verfahren der Latent Variable- und Item-Response-Modelle diskutiert werden.

7.2.1 Analyse der einzelnen Verhaltensinventare

Sowohl in der Ruhe- als auch in der Aktivitätssituation wurden alle der insgesamt 39 beschriebenen Einzelindikatoren beobachtet, was als ein erster grober Hinweis darauf gewertet werden kann, dass beide Verhaltensinventare in der Gruppe demenzkranker Menschen prinzipiell relevante Verhaltensweisen umfassen. Allerdings wurden einzelne Verhaltensweisen dabei sehr unterschiedlich häufig, und insgesamt vor allem in der Ruhebedingung nur wenige Schmerzindikatoren beobachtet. So betrug die mittlere Auftretenshäufigkeit für ein beliebiges BESD-Item nur knapp 15 Prozent in Ruhe und knapp 19 Prozent in Aktivität, für die CNPI-Items dagegen 10 Prozent in Ruhe und 14 Prozent in Aktivität. Entsprechend konnten durch beide Instrumente in der Aktivitätssituation signifikant höhere Gesamtskalenwerte erreicht werden als in der Ruhesituation. Die berücksichtigten Bewohner scheinen also bei Bewegung und Aktivität mehr bzw. intensivere Schmerzen zu verspüren.

Die wechselseitigen Zusammenhänge zwischen den einzelnen potenziell schmerzbezogenen Verhaltensweisen müssen mit Werten für die interne Konsistenz (KR-20) zwischen .68 und .71 für beide Inventare und Erfassungsbedingungen als moderat gelten. Offensichtlich zeigen einige der BESD- und CNPI-Indikatoren nur zu einem kleinen Anteil ein allen beobachteten Verhaltensweisen gemeinsames latentes Merkmal Schmerz

an. Dieser Befund bestätigt ein Stück weit auch die Vorerfahrungen zu den CNPI- und PAINAD-Originalskalen im anglo-amerikanischen Raum und die eigenen Ergebnisse zur BESD aus der ersten HILDE-Untersuchungsphase. Der direkte Vergleich mit den bisherigen psychometrischen Befunden wird jedoch dadurch erschwert, dass in der vorliegenden Untersuchung die in den Inventaren beschriebenen Verhaltensweisen jeweils als konkret zu beobachtendes Verhalten (dichotom) vorgegeben wurde. Dem Umstand, dass zwischen den Einzelindikatoren aus ein- und demselben Ausdrucksbereich (und für die BESD zusätzlich innerhalb der unterschiedenen Stufen verschieden starken Schmerzes) höhere Zusammenhänge erwartet werden können als zwischen Indikatoren aus unterschiedlichen Bereichen, wurde durch eine hierarchische Modellierung mit entsprechenden Ausdrucksbereichs-Faktoren, und einem übergeordneten Generalfaktor Rechnung getragen. Die Differenzierung dieser Ebenen schmerzbezogenen Verhaltensausdrucks erlaubt auch eine präzisere Diskussion darüber, welche Art von Schmerz durch die Inventare überhaupt erfasst werden bzw. inwieweit ein spezifisches Schmerzerleben (z.B. Gelenkschmerzen in der Hüfte vs. Kopfschmerz) in bestimmten Kommunikationskanälen seinen Ausdruck finden kann. Der gegenwärtige Stand der Forschung kann diese Diskussion jedoch kaum informieren. Es bleibt darum eine Aufgabe zukünftiger Schmerzforschung Verhaltensindikatoren zu entwickeln, die nicht wie gegenwärtig nach den beteiligten Körperregionen (Körperhaltung, Mimik, Lautäußerungen) strukturiert werden, sondern sich stärker an beispielsweise den emotional-motivationalen, kogntiv-evaluativen oder sensorisch-diskriminatorischen Qualitäten des Schmerzerlebens ausrichten. Insoweit wie dies gelingt, wird auch die Diskussion des Verhältnisses zwischen verschiedenen Ausdrucksbereichen bzw. -qualitäten und dem anzunehmenden Schmerzkonzept als Ganzem möglich. Trotz des im Vergleich dazu gegenwärtig noch unklaren Konzeptes von bereichsspezifischem und bereichsübergreifendem Schmerzausdruck bestätigen die vorgelegten Analysen die Vermutung früherer Arbeiten, dass die Ausdrucksbereiche Atmung und Trost der BESD zwar jeweils substanziell, aber vergleichsweise wenig zur Abbildung des generellen Schmerzverhaltens demenzkranker Menschen beitragen. Beispielsweise die Tröstbarkeit zeigt damit entweder nur zu einem geringen Anteil Schmerzen an, oder ist zumindest ein Indikator für eine besondere Art des Schmerzes, die in anderen Verhaltensweisen als der Beruhigung bzw. Ablenkung durch Andere eben einfach schlechter zum Ausdruck kommen kann. Es ist naheliegend, hierbei auch an seelischen Schmerz zu denken.

Neben der Überprüfung der implizierten strukturellen Beziehungen innerhalb des Indikatorpools beider Instrumente ergeben sich durch die in dieser Arbeit gewählte dichotome Erfassungsform und die Schätzung der Itemparameter auf der Folie der probabilistischen Testtheorie erweiterte Chancen für eine Überprüfung der getroffenen Annahmen zum Verhältnis von Schmerzausdruck und Schmerzintensität. Die vorgelegten detaillierten IRT-Analysen der Itemschwierigkeit der CNPI und BESD-Indikatoren haben dabei deutlich gezeigt, dass das für beide Originalskalen vorgesehene Itemscoring die jeweils tatsächlichen Unterschiede in der angezeigten Schmerzintensität nur unzureichend repräsentiert. So können beispielsweise einzelne CNPI-Indikatoren aus demsel-

ben Ausdrucksbereich bei sehr unterschiedlichen Schmerzbelastungen erwartet werden, während einzelne BESD-Items, die hohen Schmerz anzeigen sollten, bereits bei geringeren Schmerzintensitäten erwartet werden können und umgekehrt.

7.2.2 Vergleich der Verhaltensinventare

Bei der separaten Analyse der BESD- und CNPI-Inventare wurden zunächst nur die grundsätzlichen Vorteile einer probabilistisch orientierten Itemanalyse veranschaulicht. Aus den für die Einzelindikatoren geschätzten Wahrscheinlichkeitsfunktionen für die Beobachtung dieses Verhaltens über unterschiedliche Ausprägungsgrade des zugrundeliegenden Schmerzerlebens hinweg lassen sich jedoch weitere Kennwerte ableiten, die insbesondere eine Gegenüberstellung von Itemsubsets (z.B. Ausdrucksbereichen) erleichtern. Diese Anwendungsmöglichkeiten wurden in der vorliegenden Arbeit am direkten Vergleich der beiden berücksichtigten Instrumente BESD und CNPI bzw. ihrer Verhaltensindikatoren veranschaulicht.

Der Gesamtumfang der durch die beschriebenen BESD-Indikatoren schmerzbezogenen Information liegt über den gesamten Bereich latenten schmerzerlebens höher als die durch die CNPI-Indikatoren zusammengetragene Information zum Schmerzerleben, was nicht verwundert, da insgesamt 24 BESD-Items, aber nur 15 CNPI-Verhaltensweisen berücksichtigt wurden. Relativiert man den Informationsgehalt am unterschiedlichen Skalenumfang, so wird deutlich dass die CNPI über weite Bereiche mittleren und hohen Schmerzes hinweg eine effizientere Erfassung des Schmerzerlebens erlaubt. Beide Inventare scheinen jedoch erst in einem Bereich gesteigerten Schmerzes maximal informativ, während der relative Informationsgehalt beider Skalen für Bewohner mit einer (in dieser Stichprobe und in Ruhe) durchschnittlichen Schmerzbelastung vergleichsweise gering erscheint. Für beide Inventare ist damit zu erwarten, dass in der Aktivitätssituation – für die ein höheres Schmerzniveau erwartet werden kann – nicht nur höhere Schmerzwerte gemessen werden, sondern dass diese auch präzisere Abbildungen der wahren Schmerzbelastung darstellen. Dabei wird unmittelbar einsichtig, welche enormen Vorteile sich aus dem gewählten probabilistischen Analyseansatz und der sowohl test- als auch stichprobenunabhängigen Schätzung der Itemparameter für eine auf die jeweilige Fragestellung optimal angepasste Selektion und Zusammenstellung schmerzbezogener Verhaltensindikatoren ergeben. Sowohl die Breite des abgebildeten Merkmalsbereiches, als auch die über dieses Intervall erreichte Präzision der Schmerzmessung können bewusst gesteuert werden. Von besonderer Bedeutung ist daneben auch die Möglichkeit einer Voraussage des schmerzbezogenen Ausdrucksverhaltens in anderen als der hier betrachteten Stichprobe demenzkranker Heimbewohner. Die durch die vorgelegten Analysen gewonnenen psychometrischen Befunde können damit aber einen Grad an Generalisierbarkeit für sich beanspruchen, der durch keine der bisher veröffentlichten Arbeiten zur verhaltensgestützten Schmerzmessung bei demenzkranken Menschen erreicht wurde.

7.2.3 Schmerzänderung bei Aktivierung

Die Interpretation der berichteten Unterschiede in den BESD- und CNPI-Skalenscores für Ruhe- und Aktivitätssituationen als unterschiedliche Schmerzbelastung setzt voraus, dass die Verfahren das Schmerzerleben in Ruhe und Aktivität vergleichbar gut abbilden. In insgesamt drei weiteren Analyseschritten wurde diese Annahme direkt überprüft.

Während das Ziel der vorangegangenen Auswertungen darin bestand, für jeden einzelnen durch BESD und CNPI vorgeschlagenen Indikator möglichst reichhaltige psychometrische Information zu erhalten, liegt die Zielsetzung für diesen Analyseschritt nun etwas anders. Solche Verhaltensweisen, die sich in den vorangegangenen Analysen als vergleichsweise schlechte Repräsentanten des Schmerzes erwiesen hatten, sollten für die Abbildung der wahren Merkmalsveränderung nun nicht mehr weiter mitgeführt werden. Aus der Überprüfung der Messinvarianz wurden darum insgesamt ein Drittel aller ursprünglich 39 Verhaltensindikatoren ausgeschlossen.

Für die verbleibenden Items wurde die Messstruktur des kombinierten Itempools für beide Situationen parallel geschätzt. Die Itemdiskriminationen wurden für alle 24 verbliebenen Items unter beiden Bedingungen als gut miteinander vergleichbar geschätzt. Es kann also davon ausgegangen werden, dass der Messung sowohl in Ruhe als auch Aktivität das gleiche strukturelle Verständnis von Schmerz zugrunde liegt.

Für insgesamt acht Einzelindikatoren wurden jedoch in beiden Situationen unterschiedliche Itemschwierigkeiten festgestellt. Da in Aktivität insgesamt mehr schmerzbezogene Verhaltensweisen beobachtet worden sind, hätte man auf der Grundlage der KTT erwartet, dass die Indikatoren in Aktivität eher geringere Itemschwierigkeit besitzen. Tatsächlich sind nur die beiden Indikatoren angespannte Körperhaltung und Worte des Unbehagens in Aktivität etwas leichter zu beobachten. Für eine aussagekräftige Analyse der wahren Merkmalsunterschiede in beiden Situationen verbleiben damit noch 17 in Ruhe hinreichend reliable und trans-situativ invariante Schmerzindikatoren.

Tatsächlich liegt die geschätzte wahre Schmerzbelastung der demenzkranken Bewohner in der Aktivitätssituation ungefähr um eine halbe Standardabweichung höher als in der Ruhebedingung. Anstatt jedoch die wahren Merkmalsverteilungen in beiden Erfassungsbedingungen miteinander zu vergleichen, wurde der Unterschied in der wahren Schmerzbelastung zwischen Ruhe und Aktivität als latente Differenzkomponente konzeptualisiert. Inhaltlich werden die wahren Merkmalsveränderungen dabei als *individuelle Vulnerabilität, bei Aktivität eine Schmerzsteigerung zu erfahren* interpretiert. Insgesamt scheinen die Bewohner dabei vergleichsweise einheitlich mit einer gewissen Schmerzsteigerung auf Aktivierung zu reagieren, während die Schmerzbelastung in Ruhe deutlich unterschiedlicher ausfällt. Schmerz und aktivitätsbezogene Schmerzänderung wurden als moderat negativ miteinander verknüpft gefunden, so dass für in Ruhe eher schmerzfreie Bewohner eine deutlichere Schmerzsteigerung in Aktivität erwartet werden kann als für bereits in Ruhe stark schmerzbelastete Bewohner. Der lineare Zusammenhang repräsentiert das gefundene Muster von Ruheschmerz und Schmerzsteigerung jedoch nur unzureichend. So reagieren die Bewohner offensichtlich mit steigender ‚Grundbelastung' bis zu einem Punkt mäßig

gesteigerten Schmerzes zunehmend uneinheitlicher auf Aktivierung, während Bewohner, die bereits in der Ruhesituation stark schmerzbelastet sind, in der Aktivitätssituation nur selten noch eine deutliche weitere Schmerzsteigerung erfahren. Nur für wenige Bewohner ist die Aktivitätssituation jedoch mit geringeren Schmerzen verbunden als die Ruhesituation.

Zur Erklärung der Unterschiede in der Schmerzbelastung der Bewohner in Ruhe und ihrer aktivitätsbezogenen Schmerzvulnerabilität wurden kognitive, nicht-kognitive und alltagspraktische Kompetenzen der Bewohner herangezogen. Kognitive Beeinträchtigung erscheint dabei tendenziell mit höherem Schmerzerleben in Ruhe verbunden. Erhaltene mobilitätsbezogene Alltagskompetenzen waren dagegen erwartungsgemäß, allerdings ebenfalls nur tendenziell, mit geringeren Schmerzzuwächsen bei Aktivität verbunden. Besonders bemerkenswert ist der Befund, dass durch die Belastung mit nicht-kognitiven Demenzsymptomen ein höheres Schmerzniveau in Ruhe vorhergesagt werden kann, aber ein geringeres Ausmaß an Schmerzzuwachs bei Aktivität. Anstatt zu folgern, dass verhaltensauffällige Bewohner stärker, und über verschiedene Erfassungsbedingungen hinweg konstant schmerzbelastet sind, könnte vermutet werden, dass sich die demenzbedingten Verhaltensauffälligkeiten ohne wesentlichen realen Schmerzbezug in den Verhaltensindikatoren zur Schmerzmessung niederschlagen.

7.3 Implikationen für die Praxis

Aus der in dieser Arbeit geleisteten Identifikation zentraler Forschungsfragen, methodischer Herausforderungen und Möglichkeiten ihrer Bearbeitung, und nicht zuletzt auch aus den empirischen Analysen zur Beobachtung der BESD- und CNPI-Indikatoren leiten sich in erster Linie Empfehlungen für diejenigen Personengruppen ab, die im Rahmen ihrer (pflege-)wissenschaftlichen Forschungsarbeit Verfahren zur Schmerzmessung bei Demenzkranken Menschen beurteilen oder entwickeln.

In der direkten Gegenüberstellung mit der BESD erwiesen sich die CNPI-Items über nahezu den gesamten abgebildeten Schmerzbereich als der insgesamt in sich stimmigere und effizientere Indikatorpool zur Schmerzbeobachtung in dieser Population. Beide Skalen teilen jedoch die Schwäche, dass sie ihren maximalen Informationsgehalt erst bei vergleichsweise starker Schmerzbelastung erreichen, die zuverlässigsten Aussagen damit aber in einem für die pflegerische Entscheidungsfindung (i.S. v. Interventionsbedarf) weniger zentralen Ausschnitt des latenten Schmerzkontinuums getroffen werden können. Die BESD kann im Bereich geringen Schmerzes besser zu einer Früherkennung von Schmerzen beitragen und damit rechtzeitiger einen Interventionsbedarf anzeigen. Die vorgestellten veränderungsbezogenen Analysen haben deutlich gemacht, dass sowohl die BESD- als auch die CNPI-Skala lediglich Subsets von wenigen Items umfassen, die eine verlässliche Abbildung aktivitätsinduzierter Schmerzen erlauben. In Anbetracht des Umstandes, dass das pflegerische Handeln häufig mit einer Mobilisierung und Aktivierung des demenzkranken Menschens einhergeht, erscheinen sowohl BESD als auch CNPI en bloc in diesen zentralen Momenten des Pflegealltags weniger gut geeignet.

Diese Arbeit hat deutlich gemacht, dass evidenzbasiertes pflegerisches Schmerzmanagement mehr bedeutet, als nur Informationen mit standardisierten Verfahren zu sammeln. Ein valides Schmerzmanagement ist ganz wesentlich auch davon abhängig, dass dem Endnutzer die impliziten Annahmen und der Aussagebereich eines spezifischen Instrumentes zur Schmerzerfassung bewusst sind. Das gilt umso mehr, als die Schmerzmessung im pflegerischen Kontext verschiedene Ziele verfolgt (z.B. kontinuierliches Schmerzmonitoring, Planung von Interventionen, Optimierung des Pflegehandelns, Überprüfung von Behandlungserfolgen) für die ein- und dasselbe Instrument unterschiedlich gut geeignet sein mag. Im empirischen Teil dieser Arbeit wurde überprüft, inwieweit mit den vergleichsweise gut etablierten Verfahren BESD und CNPI einige der zentralen Herausforderungen der beobachtungsgestützten Schmerzerfassung bei Demenz bearbeitet bzw. überwunden werden können. Zusammenfassend können weder die BESD- noch die CNPI-Skala *en bloc* uneingeschränkt empfohlen werden. Beide Verfahren schlagen jedoch Schmerzindikatoren vor, die sich zur Beantwortung spezifischer Probleme der Schmerzmessung, beispielsweise der Frage nach dem Zusammenhang von Aktivität und Schmerz, in besonderem Maße eignen.

Neben dem erreichten hohen Auflösungsgrad der Analysen zu einzelnen in Frage stehenden Schmerzindikatoren kommt der vorliegenden Arbeit auch darum eine besondere praktische Bedeutung zu, da die berichteten psychometrischen Eigenschaften der BESD- und CNPI-Indikatoren erstmals test- und stichprobenunabhängig geschätzt werden konnten, und somit auch auf andere Instrumentenzusammensetzungen und Stichproben übertragen werden können. Damit wird es möglich abzuschätzen, welche Folgen sich für die Präzision und den Messbereich der BESD- bzw. CNPI-Skalen ergeben, wenn beispielsweise ganze Ausdrucksbereiche – wie die Körpersprache bei vollständig immobilen Demenzkranken – nicht mehr zur Schmerzkommunikation genutzt werden können. In ähnlicher Weise wird somit auch eine Voraussage der Funktionsweise der Instrumente in solchen Stichproben demenzkranker Heimbewohner erleichtert, die beispielsweise nach der Rückkehr aus der Behandlung in einem Akutkrankenhaus eines besonderen Schmerzmonitorings bedürfen. Wie durch die vorangegangene Diskussion deutlich wurde, kann eine konkrete Zusammenstellung von Verhaltensindikatoren zu einer Schmerzskala nur mit Blick auf den spezifischen Zweck und die begleitenden Umstände der Erfassung bewertet werden. Die auf der Grundlage der IRT-Modellierung test- und stichprobenunabhängig geschätzten Itemparameter in Tabelle 18 auf Seite 230 und die für das trans-situativ invariante Subset von Verhaltensindikatoren geschätzten Itemparameter (Tabelle 19 auf Seite 246) können über die hier bearbeiteten inhaltlichen Fragestellungen hinaus im Sinne einer *Item Bank* genutzt werden. Damit wird der Forderung nach einer “item bank specifically for the target group” (Zwakhalen et al., 2006) unmittelbar nachgekommen. Darüber hinaus verbindet sich damit die Hoffnung, dass auf der Grundlage der vorgelegten Analysen – je nach spezifischer Fragestellung und dem leitenden Forschungsinteresse der Leser – mehrere verschiedene Zusammenstellungen von Einzelindikatoren gewählt und in der Praxis erprobt werden.

7.4 Limitationen

Arbeiten, die mit einem deutlich methodischen Fokus auf vorliegende Daten aus größeren Forschungsprojekten zugreifen stehen nicht selten im Verdacht, dass der Konzeption und Erfassung der betrachteten Inhalte, hier der schmerzbezogenen Informationen der zweiten HILDE-Feldphase, nicht dieselbe Aufmerksamkeit und Sorgfalt gewidmet wurde wie für spezifische Untersuchungen zu enger umrissenen Fragestellungen erwartbar. Auch die vorliegende Arbeit greift mit der Schmerzmessung bei Demenz einen Teilaspekt aus einem größeren Forschungszusammenhang heraus. Nicht zuletzt aufgrund der Eingaben durch und den kontinuierlichen Austausch mit dem Arbeitskreis Schmerz im Alter der DGSS wurde dabei jedoch sichergestellt, dass auf der Schmerzerfassung von Beginn der Studie (erste HILDE-Projektphase) an stets ein besonderes Augenmerk lag.

Durch die Anbindung an das HILDE-Projekt standen dem Autor für die vorgelegten empirischen Analysen sicherlich für die betrachtete Population demenzkranker Heimbewohner bundesweit herausragend umfangreiche und differenzierte Informationen zur Verfügung. Darüber hinaus konnte aufgrund der Kontinuität des Projektes auch auf detaillierte schmerzbezogene empirische Befunde aus früheren Projektphasen zurückgegriffen werden. Dementsprechend konnten für die vorliegende Arbeit komplexere Modellierungen durchgeführt und ein ungewöhnlich hoher Auflösungsgrad der Analysen erreicht werden. Was zunächst als Stärke dieser Arbeit gelten kann impliziert jedoch auch eine prinzipielle Beschränkung der praktischen Reichweite der hier herausgearbeiteten Potenziale neuerer statistischer Verfahren zur Entwicklung und Bewertung verhaltensbezogener Schmerzassessments bei Demenz. Manche der bislang vorgelegten Arbeiten zur Erprobung oder Validierung vorgeschlagener Verhaltensinventare erfüllen mit Blick auf die realisierte Stichprobengröße (bzw. den Anteil kognitiv beeinträchtigter Menschen) gerademal die Mindestanforderungen für grundlegende statistische Analysen. In dem Maße jedoch, wie sich der Augenmerk der Schmerzforschung – den Forderungen einiger Experten nachkommend – von der immer neuen Kombination bekannter und neu eingebrachter Komponenten von Verhaltensinventaren weg-, und auf einen Praxiseinsatz bestehender Instrumente bei möglichst vielen demenzkranken Menschen zubewegt, werden sich auch vermehrt Möglichkeiten für den Einsatz komplexerer Analyseverfahren ergeben.

Schließlich muss hier eingeräumt werden, dass selbst die für die vorliegende Arbeit verfügbare Stichprobengröße nicht zuletzt auch wegen des geringen Skalenniveaus der Beobachtungsdaten mitunter hinter den Forderungen mancher Statistiker zurückbleibt. Bei der Abfassung dieser Arbeit wurde darum darauf geachtet, durch die Darstellung möglicher Potenziale zwar zu einem innovativen und kreativen Umgang mit den Potenzialen neuerer Verfahren aufzufordern, andererseits jedoch auch die Anforderungen an sowohl die theoretische Modellbildung (z.B. bei der notwendigen Explikation impliziter Annahmen zum Aufbau und der Funktionsweise einer Skala) als auch die verfügbare Datenbasis (v.a. hinsichtlich Stichprobengröße und Itemanzahl) deutlich zu machen.

Eine Beschränkung hinsichtlich der Vergleichbarkeit der vorliegenden Arbeit mit dem bisherigen Forschungsstand zu den Verhaltensinventaren BESD (bzw. PAINAD) und CN-

PI ergibt sich sicherlich dadurch, dass in der vorliegenden Arbeit alle in den beiden Originalskalen angeführten Verhaltensweisen als konkret zu beobachtende Schmerzindikatoren vorgegeben wurden. Der Verzicht auf Verhaltensweisen mit *Beispielcharakter* ist m.E. eine notwendige Voraussetzung für die Überwindung der dominierenden Testorientierung zugunsten einer detaillierten Kenntnis der psychometrischen Eigenschaften einzelner konkret beobachtbarer Verhaltensweisen. Der mit der gewählten Erfassungsform gewährleisteten Interpretationssicherheit stehen aber auch Nachteile entgegen. Zum ersten müssen die Pflegenden nun insgesamt 24+15=39 Items bearbeiten statt 5+6=11 Items. Bereits in Kapitel 3.4.2.2 wurde jedoch kritisch hinterfragt, welchen Stellenwert Beispielindikatoren bei der Schmerzbeobachtung tatsächlich einnehmen. Es bleibt unklar, wie eng sich Pflegende an den beschriebenen Verhaltensweisen orientieren bzw. wie häufig und wie weitgehend sie von der gebotenen Möglichkeit zur Abstraktion bzw. Interpretation tatsächlich Gebrauch machen. Ein weiteres potenzielles Problem bei der gemeinsamen Vorgabe von Verhaltensweisen aus mehreren Inventaren sind unerwünschte Effekte der Reihenfolge. Auch wenn ein Vergleich der Beobachtungsraten für verschiedene Instrumentenabschnitte keine Hinweise auf eine geringere Berücksichtigung der später aufgeführten Verhaltensweisen erbrachte, können entsprechende Effekte zulasten der CNPI aufgrund der nichtrandomisierten Vorgabe nicht ausgeschlossen werden.

Selbstverständlich wäre ein „Nachbau“ der bereichsspezifischen Kennwerte und des Gesamtscores im Sinne des ursprünglichen Formates sowohl für die BESD als auch für die CNPI möglich, beispielsweise um eine optimale Anbindung der vorliegenden Forschungsarbeit an die bisher für beide Skalen berichteten Befunde zu ermöglichen. Auf eine entsprechend (gewichtete) Aggregation der beobachteten Verhaltensindikatoren wurde jedoch bewusst verzichtet, da zum ersten konkret beobachtete Verhaltensweisen gegebenenfalls anders zu bewerten sind als Beispielindikatoren (s. letzter Abschnitt), zweitens eine explizite Vorgabe der durch die einzelnen Verhaltensweisen implizierten Schmerzintensität (1 vs. 2 Punkte im BESD) diese Interpretation beeinflussen sollte, und drittens die implizierte Struktur der Skala (beispielsweise mit Blick auf die Schwierigkeit und Diskriminationskraft der Einzelindikatoren) durch die vorgestellten differenzierten Analysen ja gerade überprüft werden sollte.

Bei der Interpretation der geschätzten individuellen Vulnerabilität, bei Aktivierung Schmerzen zu erfahren sollte berücksichtigt werden, dass der Grad der Aktiviertheit der Bewohner weder für die Ruhesituation, noch für die Aktivitätssituation einheitlich vorgegeben wurde. Zwar wurden für beide Beobachtungsbedingungen konkrete Verhaltensweisen beschrieben (z.B. Aktivität: beim Aufstehen), dennoch variieren die tatsächlich beobachteten Situationen – auch nach Beeinträchtigungsgrad – deutlich. Entsprechend ist auch der Vergleich beider Beobachtungssituationen für die Bewohner vermutlich durch unterschiedlich starke Aktivitätssteigerungen bestimmt. Ruhe und Aktivität wurden im Rahmen dieser Untersuchung vergleichsweise breit gefasst, insbesondere aber nicht ausschließlich auf körperliche Bewegung eingegrenzt. Insbesondere für schwer körperlich beeinträchtigte Menschen in weit fortgeschrittenen Phasen der Demenzerkrankung erscheint eine solche auch subtilere Zustände von Aufmerksamkeit und Engagement berücksichti-

gende Konzeption von Aktivität sinnvoll. Dennoch bleibt festzuhalten, dass die zugestandenen Freiheiten bei der Definition ruhiger und aktivierter Situationen die Möglichkeiten einer schlüssigen monokausalen Interpretation des modellierten Vulnerabilitätsfaktors als Ausdruck muskulo-skelettaler Degeneration oder Erkrankung erschweren. Eine letzte Untiefe der veränderungsbezogenen Analyse ergibt sich aus der mangelhaften Kontrolle des zeitlichen Abstandes zwischen beiden Beobachtungen. Selbst wenn die Merkmalsveränderungen hier auf den systematisch variierten Faktor Aktiviertheit bezogen werden, ist natürlich nicht auszuschließen, dass Veränderungen zwischen beiden Situationen auch durch die Eigendynamik des Schmerzerlebens selbst zustandekommen.

Im Zuge der Kooperation mit der Sektion Gerontopsychiatrie der Universitätsklinik Heidelberg erfolgte zu Studienbeginn eine ausführliche Untersuchung der Bewohner durch einen gerontopsychiatrischen Facharzt im Rahmen eines diagnostischen Interviews. Obwohl damit eine gute Abbildung der erhaltenen Bewohnerkompetenzen und der Belastung mit nicht-kognitiven Demenzsymptomen geleistet werden konnte, blieben die Möglichkeiten einer Differenzialdiagnose verschiedener Demenzätiologien in Ermangelung von Befunden bildgebender Verfahren häufig (auf die klinische Urteilsbildung des Facharztes) begrenzt. Trotz der Hinweise neuerer Forschungsarbeiten auf möglicherweise unterschiedliche Schädigungen schmerzrelevanter Hirnstrukturen bei unterschiedlichen Demenzformen (Kap. 2.2.4.2) wurden darum auch bei den Analysen zur Demenzspezifität der Schmerzbeobachtung (Kap. 6.8) nicht einzelne Ätiologien unterschieden, sondern eine Orientierung am Gesamtmuster erhaltener Kompetenzen im Sinne von Prägnanztypen des Demenzsyndroms gewählt (Kap. 6.8.6). Obwohl im Methodenteil der Arbeit beschrieben (Kap. 4.3.4), wurde bei der Differenzierung dieser Kompetenzgruppen aufgrund des begrenzten Stichprobenumfanges auf eine Überprüfung der Invarianz der Messstruktur anhand eines Multi-Group-LVMs verzichtet.

Der allergrößte Teil der dargestellten Analysen wurde ausschließlich mit den unter zwei Beobachtungsbedingungen erfassten 39 potenziell schmerzbezogenen Verhaltensindikatoren aus der BESD und CNPI bestritten. Während die wechselseitigen Beziehungen dieser Schmerzindikatoren sicherlich in einem Auflösungsgrad analysiert werden konnten, der durch keine der bislang verfügbaren empirischen Studien erreicht wurde, musste der Grad der Einbettung dieser Schmerzbeobachtung in den Gesamtkontext *Lebensqualität bei Demenz* der HILDE-Untersuchung dagegen deutlich zurücktreten. Auf eine Berücksichtigung der weiterführenden Analysen, die sowohl das Schmerzerleben der Bewohner in Ruhe, als auch deren geschätzte Vulnerabilität, bei Aktivierung eine Schmerzsteigerung zu erfahren mit verschiedenen Indikatoren der individuell erlebten Lebensqualität in Bezug setzen, wurde aufgrund des für die vorliegende Arbeit gewählten psychometrischen Schwerpunktes an dieser Stelle bewusst verzichtet.

Eng verbunden mit der Frage der Einbindung der betrachteten Schmerzerfassung in die übergeordneten Kontexte des pflegerischen Schmerzmanagements (beispielsweise mit Blick auf deren Integrierbarkeit oder Implikationen für Interventionen, usw.) oder der Forschung zu Lebensqualität bei Demenz ist auch diejenige nach der Validität und Praktikabilität der Beobachtungsverfahren. Ausgehend von der Annahme, dass die durch BESD

und CNPI beschriebenen Verhaltensindikatoren Schmerz anzeigen, war der überwiegende Anteil der vorgestellten Analysen darauf ausgerichtet, reliabilitätsbezogene Aspekte der verhaltensgestützten Schmerzeinschätzung zu prüfen und eine maximale Präzision und trans-situationale Konstanz dieser Abbildung sicherzustellen. Zur Überprüfung der Validität der Schmerzbeobachtung wurden konstruktgleiche und diskriminante Maße herangezogen. Dabei konnte auch die im Rahmen des gerontopsychiatrischen Interviews erfragte schmerzbezogene Selbstauskunft durch die Bewohner als Goldstandard auf ihre Übereinstimmung mit den BESD- und CNPI-Skalenscores geprüft werden. Wie zu erwarten – und Anlass für die Beschäftigung mit der alternativen Methodik der Verhaltensbeobachtung – war, konnten für einen substanziellen Anteil insbesondere schwer beeinträchtigter Bewohner jedoch keine Selbstauskünfte eingeholt werden. Zusätzlich konnte nicht sichergestellt werden, dass die initiale Begutachtung durch den Gerontopsychiater des Projektteams und die Durchführung der Verhaltensbeobachtung durch die Pflegenden tatsächlich maximal zeitnah, nach Möglichkeit noch am selben Tag erfolgte, wodurch die Selbstauskunft aktueller Schmerzen nur eingeschränkt als ultimativer Vergleichsmaßstab gelten kann. Dass im Rahmen der Schmerzeinschätzung durch die Pflegenden auf eine direkte Befragung des Bewohners verzichtet wurde, stellt somit einen gewissen Schwachpunkt des gewählten Studiendesigns dar. Wie bereits in Kapitel 3.4.2.3 angemerkt, bleibt selbstverständlich auch der Vergleich der konkurrierenden Verfahren BESD und CNPI zur Schmerzbeobachtung (s. Kap. 6.9.1) aufgrund der bislang für noch keines der Instrumente endgültig nachgewiesenen (und nachweisbaren) Validität, nur eingeschränkt informativ.

Über die empirische Untersuchung der schmerzbezogenen HILDE-Inhalte hinaus verfolgte die vorliegende Arbeit auch das pädagogische Ziel, dem Leser die Potenziale bislang in diesem Forschungszweig noch weitestgehend ungenutzter neuer statistischer Verfahren näherzubringen. Entsprechend wurde der Darstellung verschiedener aufeinander aufbauender testtheoretischer Konzepte und deren Bedeutsamkeit für spezifische Herausforderungen der Schmerzforschung besondere Aufmerksamkeit geschenkt bzw. Raum gegeben. Da die Familie der Item-Response-Modelle dabei als Spezialfall eines allgemeinen Latent-Variable-Model-Ansatzes verstanden wurde, orientierte sich die Auswahl einer für alle Entwicklungsebenen statistischer Verfahren einheitlichen Notation für diese Arbeit an den Vorgaben von Muthén und Asparouhov (2004) sowie Skrondal und Rabe-Hesketh (2004). Lesern, die mit der üblichen Notation von Item-Response-Modellen vertraut sind, wird damit unter Umständen eine gewisse Schwierigkeit auferlegt.

Während wesentliche Fragestellungen adressiert und am Beispiel konkreter Daten exemplarisch bearbeitet werden konnten, war für andere zumindest kurz angerissene Optionen neuer Analyseansätze, beispielsweise für das Adaptive Testen, oder auch für die Abschätzung der Demenzspezifität der postulierten Messstruktur der Beobachtungsverfahren auf der Grundlage eines entsprechenden Multi-Group-LVM aufgrund der Struktur und Anzahl der verfügbaren Daten keine praktische Veranschaulichung möglich.

7.5 Ausblick

Die beschriebenen Chancen alternativer statistischer Verfahren wie der Item-Response-Theorie auch für die Optimierung von Verhaltensinventaren zur Schmerzmessung werden gegenwärtig selbstverständlich nicht nur vom Autor der vorliegenden Arbeit erkannt. Ebenfalls im Rahmen einer Dissertation beschäftigt sich zum Zeitpunkt der Fertigstellung dieser Arbeit eine niederländische Schmerzforscherin aus dem Umfeld von Zwakhalen mit den Möglichkeiten, die psychometrischen Eigenschaften einer niederländischen Version der PACSLAC auf der Basis faktoranalytischer und probabilistischer Ansätze zu explorieren (van Nispen, 2008).

Um mit der vorgelegten Arbeit einen maximalen Wirkungsgrad zu erzielen, wurde den inhaltlichen Analysen der beiden Schmerzassessments BESD und CNPI eine ausführliche Diskussion der den Forschungsbereich dominierenden Herausforderungen und geeigneter methodischer Ansätze für deren Bearbeitung vorangestellt. Die dabei angesprochenen z.T. komplexen Modelle und ihre formale Darstellung, sowie der durch diese doppelte Zielsetzung nahezu unvermeidbar höhere Seitenumfang stellen sicherlich ein Hindernis für die breite Rezeption durch die an einer handhabbaren alltagspraktischen Schmerzmessung für demenzkranke Menschen interessierten Forschenden dar.

Angesichts der diskutierten Vielseitigkeit der vorgeschlagenen Methoden und des anschaulich nachgewiesenen Mehrwertes dieser Ansätze im Vergleich zu herkömmlichen Verfahren der Analyse und Bewertung verhaltensgestützter Schmerzassessments darf man dennoch zuversichtlich sein, dass die vorgelegte Arbeit einen erkennbaren Beitrag zur Weiterentwicklung des Forschungsfeldes leistet und mittelbar auch denjenigen demenzkranken Menschen zugute kommt, die heute noch unter unerkannten Schmerzen leiden.

A Notation

In der folgenden Übersicht ist die in dieser Arbeit verwendete Notation zusammengestellt. Um deutlich zu machen, dass sich Aspekte aus der probabilistischen Testtheorie in den übergeordneten Rahmen des Generalized Latent Variable Modeling integrieren lassen, wurde dabei durchgängig eine SEM-orientierte Notation gewählt.

Symbol		Beschreibung
Indizes		
$j = 1, 2, 3 \ldots N$		Merkmalsträger (i.d.R. Personen)
$i = 1, 2, 3 \ldots k$		Einzelitems, Subtests oder Tests
$s = 1, 2, 3 \ldots m$		Erhebungsbedingungen (Situationen, Zeitpunkte)
$d = 1, 2, 3 \ldots l$		Dimensionen eines Merkmalsraumes
$c = 0, 1, 2 \ldots C - 1$		Kategorien eines mehrfach gestuften Items
Beobachtete Variablen		
y_{ij}		bei Merkmalsträger j beobachteter Wert einer Zufallsvariablen Y_i
x_j		bei Merkmalsträger j beobachteter Wert der Kovariaten X
Latente Variablen		
η_j	[eta]	wahrer Merkmalswert des Merkmalsträgers j
ζ_j	[zeta]	latentes Residuum einer endogenen latenten Variable η_j
ϵ_{ij}	[epsilon]	Messfehler des Tests i bei Merkmalsträger j
y^*_{ij}		latente kontinuierliche Responsevariable für nicht-metrische (z.B. dichotome) beobachtete Zufallsvariablen Y_{ij}
(Ko-)Varianzen und Korrelationen		
σ^2_i	[sigma]	Varianz der Werte y_{ij} der Zufallsvariablen Y_i
σ^{*2}_i		Varianz der latenten Responsevariablen y^*_{ij}
ψ	[psi]	Varianz der wahren Merkmalswerte η_j
θ_i	[theta]	Varianz der Messfehler ϵ_{ij} des Tests i
$\psi_{ii'}$		Kovarianz zweier Zufallsvariablen i und i'
$\rho_{ii'}$	[rho]	Korrelation zweier Zufallsvariablen i und i'
Itemanalyse CTT/IRT		
p_i		Itemschwierigkeit in üblicher CTT-Notation
r_{it}		Diskriminationsfähigkeit eines Items in üblicher CTT-Notation
a_i		Itemschwierigkeit in üblicher IRT-Notation
b_i		Diskriminationsfähigkeit eines Items in üblicher IRT-Notation
c_i		Zufalls- bzw. Ratewahrscheinlichkeit eines Items in üblicher IRT-Notation
D		Skalierungsfaktor zur Angleichung der logistischen an die Normalverteilungsfunktion (D=1,7)
SEM und LRV		
λ_i	[lambda]	Pfadkoeffizient für lineare Regressionsbeziehungen zwischen Variablen
κ	[kappa]	Pfadkoeffizient der Kovariate x_j

Symbol		Beschreibung
Mittelwertsstrukturen		
$E(\eta_j)$	([eta])	Erwartungswert einer Zufallsvariable, d.h. der über potenzielle Replikationen hinweg erwartbare mittlere Wert
α	[alpha]	Erwartungswert bzw. Mittelwert der wahren Merkmalswerte η_j
μ_i^*	[mu]	Erwartungs- bzw. Mittelwert der latenten Responsevariablen y_{ij}^*
ν_i	[nu]	Intercept-Parameter zur Bestimmung des Mittelwertes μ_i^* der latenten Responsevariablen y_{ij}^*
τ_c	[tau]	Threshold-Parameter, der den Wert der latenten Responsevariablen y_{ij}^* anzeigt, ab dem die beobachteten Werte y_{ij} in die Kategorie c fallen (s.a. Indizes)
Generalisierungstheorie		
$\sigma^2(\delta)$	([delta])	relative Fehlervarianz bei normorientierter Interpretation von Testwerten
ρ^2	[rho]	Generalisierungs-(G-)Koeffizient als Reliabilitätskoeffizient bei normorientierter Interpretation von Testwerten
Wahrscheinlichkeiten		
$P(y_{ij} = 1\|\eta_j)$		bedingte Wahrscheinlichkeit, dass die beobachtete Variable y_{ij} unter dem gegebenem Ausprägungsgrad des latenten Merkmalswertes η_j den Wert 1 annimmt
z_α		Cut-off-Wert der Standardnormalverteilung zur Konstruktion von 95%- bzw. 99%-Konfidenzintervallen (α=.05 bzw. .01)
d_{ij}		Differenz zwischen Itemwert $y_{ij} \in (0, 1)$ und geschätzter Antwortwahrscheinlichkeit $P(Y_{ij}\|\eta_j)$

Literatur

Abbey, J., Piller, N. & De Bellis, A. (2004). The Abbey Pain Scale: a 1-minute numerical indicator for people with end-stage dementia. *Int J Palliative Nurs*, 10, 6 – 13.

Ackerman, Terry A. (1987). Robustness of logist and bilog IRT estimation programs to violation of local independence. Research Report 87-14: Iowa City, IA, American College Testing (ACT).

Adams, Raymond J., Wilson, Mark & Wang, Wen-chung (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1 – 23.

AGS Panel on Persistent Pain in Older Persons (2002). The management of persistent pain in older persons. *J Am Geriatr Soc*, 50, 205–224.

Alagumalai, S. & Keeves, J.P. (1996). Disattenuated Correlation and Unidimensionality. Australian Association for Research on Education (AARE) Conference, Singapore.

Allen, S. J. & Hubbard, R. (1986). Regression Equations for the Latent Roots of Random Data Correlation Matrices with Unities on the Diagonal. *Multivariate Behavioral Research*, 21, 393–398.

Anand, K. J. S. & Craig, Kenneth D. (1996). New perspectives on the definition of pain. *Pain*, 67(1), 3 – 6.

Andersen, Erling B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38(1), 123 – 140.

Andersen, Erling B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, 50(1), 3 – 16.

Andersen, R. (1995). Revisiting the Behavioral Model and Access to Medical Care: Does it matter? *Journal of Health and Social Behavior*, 36(1), 1 – 10.

Andrich, David (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 2(4), 581 – 594.

Andrich, David (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561 – 573.

Arbuckle, J. L. & Wothke, W. (1999). *Amos 4.0 User's Guide*. Chicago: SmallWaters.

Baiardi, J., Parzuchowski, J., Kosik, C., Aimes, T., Courtney, N. & Locklear, J. (2002). Examination of the reliability of the FLACC pain assessment tool with cognitively impaired elderly. Annual National Conference of Gerontological Nurse Practitioners, Chicago, IL, 2002.

Baker, A., Bowring, L., Brignell, A. & Kafford, D. (1996). Chronic pain management in cognitively impaired patients: A preliminary research project. *Perspectives*, 20(2), 4 – 8.

Barton, Sue, Findlay, David & Blake, Roger A. (2005). The management of inappropriate vocalisation in dementia: A hierarchical approach. *International Journal of Geriatric Psychiatry*, 20(12), 1180 – 1186.

Basler, H.D., Bloem, R., Casser, H.R., Gerbershagen, H.U., Griessinger, N., Hankemeier, U., Hesselbarth, S., Lautenbacher, S., Nikolaus, T., Richter, W., Schroter, C. & Weis, L. (2001). Ein strukturiertes Schmerzinterview für geriatrische Patienten. *Schmerz*, 15, 164 – 171.

Basler, H.D., Hüger, D., Kunz, R., Luckmann, J., Lukas, A., Nikolaus, T. & Schuler, M.S. (2006). Beurteilung von Schmerz bei Demenz (BESD) Untersuchung zur Validität eines Verfahrens zur Beobachtung des Schmerzverhaltens. *Schmerz*, 20, 519 – 526.

Bathgate, D., Snowden, J. S., Varma, A., Blackshaw, A. & Neary, D. (2001). Behaviour in frontotemporal dementia, Alzheimer's disease and vascular dementia. *Acta Neurologica Scandinavica*, 103(6), 367 – 378.

Becker, Janine (2004). *Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT).* Dissertationsschrift, Freie Universität Berlin, Fachbereich Erziehungswissenschaft und Psychologie.

Becker, S., Kaspar, R. & Kruse, A. (2006). Die Bedeutung unterschiedlicher Referenzgruppen für die Beurteilung der Lebensqualität demenzkranker Menschen. *Zeitschrift für Gerontologie und Geriatrie*, 39(5), 350–357.

Becker, S., Kruse, A., Schröder, J. & Seidl, U. (2005). Das Heidelberger Instrument zur Erfassung von Lebensqualität bei Demenz (H.I.L.DE.). *Zeitschrift für Gerontologie und Geriatrie*, 38, 1–14.

Beckerman, H., Roebroeck, M.E., Lankhorst, G.J., Becher, J.G., Bezemer, P.D. & Verbeek, A.L.M. (2001). Smallest real difference, a link between reproducibility and responsiveness. *Quality of Life Research*, 10, 571 – 578.

Benedetti, Fabrizio, Arduino, Claudia, Vighetti, Sergio, Asteggiano, Giovanni, Tarenzi, Luisella & Rainero, Innocenzo (2004). Pain reactivity in Alzheimer patients with different degrees of cognitive impairment and brain electrical activity deterioration. *Pain*, 111(1), 22 – 29.

Benedetti, Fabrizio, Vighetti, Sergio, Ricco, Claudia, Lagna, Elisabetta, Bergamasco, Bruno, Pinessi, Lorenzo & Rainero, Innocenzo (1999). Pain threshold and tolerance in Alzheimer's disease. *Pain*, 80(1), 377 – 382.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.

Bentler, P. M. & Bonett, Douglas G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588 – 606.

Bergh, I., Jakobsson, E. & Sjöström, B. (2008). Worst experiences of pain and conceptions of worst pain imaginable among nursing students. *Journal of Advanced Nursing*, 61(5), 484 – 491.

Bergh, I. & Sjöström, B. (1999). A comparative study of nurses' and elderly patients' ratings of pain and pain tolerance. *J Gerontol Nurs*, 25(5), 30–36.

Bickel, A. (1996). Pflegebedürftigkeit im Alter: Ergebnisse einer populationsbezogenen retrospektiven Längsschnittstudie. *Das Gesundheitswesen*, 58(Sonderheft 1), 56 – 62.

Bickel, H. (2000). Demenzsyndrom und Alzheimer-Krankheit: Eine Schätzung des Krankheitsbestandes und der jährlichen Neuerkrankungen in Deutschland. *Das Gesundheitswesen*, 62, 211 – 218.

Bickel, H. (2001). Demenzen im höheren Lebensalter: Schätzungen des Vorkommens und der Versorgungskosten. *Zeitschrift für Gerontologie und Geriatrie*, 34, 108 – 115.

Bickel, H., Bürger, K., Hampel, H., Schreiber, Y., Sonntag, A., Wiegele, B., Förstl, H. & Kurz, A. (2006). Präsenile Demenzen in Gedächtnisambulanzen: Konsultationsinzidenz und Krankheitscharakteristika. *Nervenarzt*, 75, 1079 – 1085.

Bickel, H. & Cooper, B. (1994). Incidence and relative risk of dementia in an urban elderly population: findings of a prospective field study. *Psyhol Med*, 24(1), 179 – 192.

Bock, R. Darrell & Aitkin, Murray (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443 – 459.

Bock, R. Darrell & Lieberman, Marcus (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2), 179 – 197.

Bollen, K. (1989). *Structural Equations with latent variables.* New York: Wiley.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In: Jöreskog, K.G. & Wold, H. (Hrsg.), *Systems under indirect observation: Causality, structure, prediction.*, S. 143 – 173: Amsterdam: North-Holland.

Braak, Heiko & Braak, Eva (1994). Morphological criteria for the recognition of Alzheimer's disease and the distribution pattern of cortical changes related to this disorder. *Neurobiology of Aging*, 15(3), 355 – 356.

Brennan, Robert L. (2001). *Generizability Theory.* Statistics for social science and public policy: New York: Springer.

Brieri, D., Reeve, R.A., Champion, G.D. & et al. (1990). The face pain scale for the self assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for the ratio scale properties. *Pain*, 41, 139 – 150.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K. A. & Long, J. S. (Hrsg.), *Testing structural equation models*, S. 445–455: Newbury Park, CA: Sage.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.

Bruder, Jens (2009). Alten- und Pflegeheime. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis.* (2. Aufl.), S. 431 – 443: Heidelberg: Springer Verlag.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: basic concepts, applications, and programming.* Multivariate Application Series: Mahwah, New Jersey: Lawrence Erlbaum Associates.

Carr, Eloise C.J. & Mann, Eileen M. (2002). *Schmerz und Schmerzmnagement: Praxishandbuch für Pflegeberufe*. Bern: Verlag Hans Huber.

Chen, Wen-Hung & Thissen, David (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265 – 289.

Chibnall, John T. & Tait, Raymond C. (2001). Pain assessment in cognitively impaired and unimpaired older adults: A comparison of four scales. *Pain*, 92(1), 173 – 186.

Chou, Chih-ping, Bentler, P. M. & Satorra, Albert (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, 44(2), 347 – 357.

Closs, José, Cash, Keith, Barr, Bridget & Briggs, Michelle (2005). Cues for the identification of pain in nursing home residents. *International Journal of Nursing Studies*, 42, 3 – 12.

Closs, S. José, Barr, Bridget & Briggs, Michelle (2004). Cognitive status and analgesic provision in nursing home residents. *Br J Gen Pract*, 54, 919 – 921.

Closs, S. José, Barr, Bridget, Briggs, Michelle, Cash, Keith & Seers, Kate (2004). A Comparison of Five Pain Assessment Scales for Nursing Home Residents with Varying

Degrees of Cognitive Impairment. *Journal of Pain and Symptom Management*, 27(3), 196 – 205.

Coehlo, Deborah Padgett, Hooker, Karen & Bowman, Sally (2007). Institutional placement of persons with dementia: What predicts occurrence and timing? *Journal of Family Nursing*, 13(2), 253 – 277.

Cohen-Mansfield, Jiska (2008). The relationship between different pain assessments in dementia. *Alzheimer Disease & Associated Disorders*, 22(1), 86 – 93.

Cohen-Mansfield, Jiska & Creedon, Michael (2002). Nursing staff members' perceptions of pain indicators in persons with severe dementia. *Clinical Journal of Pain*, 18(1), 64 – 73.

Cohen-Mansfield, Jiska & Lipson, Steven (2008). The utility of pain assessment for analgesic use in persons with dementia. *Pain*, 134(1), 16 – 23.

Collins, L. G. & Stone, L. A. (1966). Family structure and pain reactivity. *Journal of Clinical Psychology*, 22(1), 33.

Collins, L. Glenn & Stone, Leroy A. (1966). Pain sensitivity, age and activity level in chronic schizophrenics and in normals. *British Journal of Psychiatry*, 112(482), 33 – 35.

Cook, Ailsa K. R., Niven, Catherine A. & Downs, Murna G. (1999). Assessing the pain of people with cognitive impairment. *International Journal of Geriatric Psychiatry*, 14(6), 421 – 425.

Cook, L. L., Eignor, D. R. & Taft, H. (1985). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates. Research Report 85-38: Princeton, NJ: Educational Testing Service.

Cornu, Fr. (1975). Disturbances of the perception of pain among persons with degenerative dementia. *Journal de Psychologie Normale et Pathologique*, 72(4), 461 – 472.

Costardi, D., Rozzini, L., Costanzi, C., Ghianda, D., Franzoni, S., Padovani, A. & Trabucchi, M. (2007). The Italian version of the Pain Assessment IN Advanced Dementia (PAINAD) scale. *Arch Gerontol Geriatr*, 44, 175–180.

Craig, A. D. (Bud) (2003). A new view of pain as a homeostatic emotion. *Trends in Neurosciences*, 26(6), 303 – 307.

Craig, Kenneth D., Hill, Marilyn L. & McMurtry, Bruce W. (1999). Detecting deception and malingering. In: Block, Andrew R., Kremer, Edwin F. & Fernandez, Ephrem (Hrsg.), *Handbook of pain syndromes: Biopsychosocial perspectives.*, S. 41 – 58: Lawrence Erlbaum Associates Publishers.

Cummings, Jeffrey L., Mega, M., Gray, K. & Rosenberg-Thompson, S. (1994). The Neuropsychiatric Inventory: Comprehensive assessment of psychopathology in dementia. *Neurology*, 44(12), 2308 – 2314.

Danek, Adrian, Wekerle, Gabi & Neumann, Manuela (2009). Pick-Komplex und andere fokale Hirnatrophien. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis.* (2. Aufl.), S. 123 – 140: Heidelberg: Springer Verlag.

Davies, E., Male, M., Reimer, V. & Turner, M. (2004). Pain assessment and cognitive impairment: part 2. *Nursing Standard*, 19(13), 33 – 40.

Davies, E., Male, M., Reimer, V., Turner, M. & Wylie, K. (2004). Pain assessment and cognitive impairment: part 1. *Nursing Standard*, 19(12), 39 – 42.

Dawadi, Bhaskar Raj (1998). *Robustness of the polytomous irt model to violations of the unidimensionality assumption.* Mastersthesis, ProQuest Information & Learning.

Decker, Sheila A. & Perry, Anne G. (2003). The Development and Testing of the PATCOA to Assess Pain in Confused Older Adults. *Pain Management Nursing*, 4(2), 77 – 86.

Deutsches Netzwerk für Qualitätsentwicklung in der Pflege DNQP (2004). *Sonderdruck Expertenstandard Schmerzmanagement in der Pflege*. Osnabrück: DNQP.

Dilling, H., Mombour, W., Schmidt, M.H. & Schulte-Markwort, E. (1997). *Internationale Klassifikation psychischer Störungen: ICD-10 Kapitel V (F) Forschungskriterien* (1 Aufl.). Weltgesundheitsorganisation WHO, Bern: Verlag Hans Huber.

Donoghue, John R. & Isham, Steven P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22(1), 33 – 51.

Edwards, Robert R., Ness, Timothy J. & Fillingim, Roger B. (2004). Endogenous Opioids, Blood Pressure, and Diffuse Noxious Inhibitory Controls: A Preliminary Study. *Perceptual and Motor Skills*, 99(2), 679 – 687.

Edwards, R.R. & Fillingim, R.B. (2001). Effects of age on temporal summation and habituation of thermal pain: clinical relevance in healthy older and younger adults. *Journal of Pain*, 2, 307 – 317.

Ekman, P.E. & Friesen, W.V. (1978). *Facial Action Coding System.* Palo Alto, CA: Consulting Psychologists Press.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, Susan E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3), 495 – 515.

Epps, Cynthia D. (2001). Recognizing pain in the institutionalized elder with dementia. *Geriatric Nursing*, 22(2), quiz 78 – 79.

Feldt, Karen S. (2000). The Checklist of Nonverbal Pain Indicators (CNPI). *Pain Management Nursing*, 1(1), 13 – 21.

Fennessy, Lynda M. (1995). *The impact of local dependencies on various IRT outcomes.* Masterthesis, ProQuest Information & Learning.

Fischer, Gerhard (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Verlag Hans Huber.

Fischer, Thomas (2005). Schmerzmanagement bei alten Menschen, Teil 2 – Schmerzen richtig erkennen und einschätzen. *Pflegezeitschrift*, 58, 355 – 358.

Fischer, Thomas (2007). Instrumente für die Schmerzeinschätzung bei Personen mit schwerer Demenz: Hilfsmittel für die Beobachtung, aber kein Ersatz der Fachlichkeit. *Pflegezeitschrift*, 60, 308 – 3111.

Fischer, Thomas (2008). Schmerzeinschätzung bei Menschen mit schwerer Demenz. Pflegekongress, November 2008, Wien, Österreich.

Fischer, Thomas (2009). *Schmerzeinschätzung bei Menschen mit schwerer Demenz.* Bern: Verlag Hans Huber (vorangekündigt).

Fischer, Thomas, Spahn, Claudia & Kovach, Christine (2007). Gezielter Umgangmit herausforderndem Verhalten bei Menschen mit Demenz: Die „Serial Trial Intervention“ (STI). *Pflegezeitschrift*, 7, 370 – 373.

Fisher, S. E., Burgio, L. D., Thorn, B. E. & Hardin, J. M. (2006). Obtaining self-report data from cognitively impaired elders: Methodological issues and clinical implications for nursing home pain assessment. *Gerontologist*, 46(1), 81–88.

Fisher, Susan E., Burgio, Louis D., Thorn, Beverly E., Allen-Burge, Rebecca, Gerstle, John, Roth, David L. & Allen, Scott J. (2002). Pain assessment and management in cognitively impaired nursing home residents: Association of certified nursing assistant pain report, Minimum Data Set pain report, and analgesic medication use. *Journal of the American Geriatrics Society*, 50(1), 152 – 156.

Flor, Herta (2001). Psychophysiological Assessment of the Patient with Chronic Pain. In: Turk, Dennis C. & Melzack, Ronald (Hrsg.), *Handbook of pain assessment (2nd ed.).*, S. 76 – 96: Guilford Press.

Folstein, M. F., Folstein, S. E. & McHugh, P. R. (1975). Mini-Mental-State. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12, 189–198.

Förstl, Hans (2009). *Demenzen in Theorie und Praxis.* (2. Aufl.). Heidelberg: Springer Verlag.

Förstl, Hans, Kurz, Alexander & Hartmann, Tobias (2009). Alzheimer-Demenz. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis.* (2. Aufl.), S. 43 – 64: Heidelberg: Springer Verlag.

Fuchs-Lacelle, Shannon & Hadjistavropoulos, Thomas (2004). Development and Preliminary Validation of the Pain Assessment Checklist for Seniors With Limited Ability to Communicate (PACSLAC). *Pain Management Nursing*, 5(1), 37 – 49.

Gagliese, Lucy & Melzack, Ronald (1997). Chronic pain in elderly people. *Pain*, 70(1), 3 – 14.

Galicia-Castillo, M.C. & McElhaney, J. (2003). Persistent pain in the elderly. *Compr Ther*, 29, 43 – 46.

Galloway, S. & Turner, L. (1999). Pain assessment in older adults who are cognitively impaired. *J Gerontol Nurs*, 25, 34 – 39.

Gauthier, Serge (2006). For debate: Is mild cognitive impairment a clinically useful concept? *International Psychogeriatrics*, 18(3), 393 – 414.

Gauvain-Piquard, A. & Pichard-Leandri, E. (1989). *La douleur chez l' enfant.* Paris: Medsi/McGraw Hill.

Giacobini, E. (2000). Present and future of Alzheimer therapy. *J Neural Transm Suppl*, 59, 231 – 242.

Gibson, Stephen J. & Chambers, Christine T. (2004). Pain Over the Life Span: A Developmental Perspective. In: Hadjistavropoulos, Thomas & Craig, Kenneth D. (Hrsg.), *Pain: Psychological perspectives.*, S. 113 – 154: Lawrence Erlbaum Associates Publishers.

Gibson, Stephen J., Voukelatos, Xenophon, Ames, David, Flicker, Leon & Helme, Robert D. (2001). An examination of pain perception and cerebral event-related potentials following carbon dioxide laser stimulation in patients with Alzheimer's disease and age-matched control volunteers. *Pain Research & Management*, 6(3), 126 – 132.

Gnass, Irmela & Sirsch, Erika (2007). Schmerzeinschätzung bei kognitiv beeinträchtigten Menschen. MScN Qualifikationsarbeit, Universität Witten/Herdecke, Institut für Pflegewissenschaft.

Golembiewski, Robert T., Billingsley, Keith & Yeager, Samuel (1976). Measuring Change and Persistence in Human Affairs: Types of Change Generated by OD Designs. *Journal of Applied Behavioral Science*, 12, 133 – 157.

Graham, J.E., Rockwood, K., Beattie, B.L., Eastwood, R., Gauthier, S., Tuokko, H. & McDowell, I. (1997). Prevalence and severity of cognitive impairment with and without dementia in an elderly population. *Lancet*, 349, 1793 – 1796.

Greig, Nigel H., Lahiri, Debomoy K. & Giacobini, Ezio (2005). Editorial: Advances in Alzheimer therapy: Something old, something new, something borrowed, something blue. *Current Alzheimer Research*, 2(3), 275 – 279.

Haberl, Roman L. & Schreiber, Angela K. (2009). Morbus Binswanger und andere vaskuläre Demenzen. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis.* (2. Aufl.), S. 65 – 84: Heidelberg: Springer Verlag.

Haberman, S. (1975). Maximum likelihood estimates in exponential response models. *Annals of Statistics*, 5, 815 – 841.

Hadjistavropoulos, Heather D., Craig, Kenneth D., Grunau, Ruth Eckstein & Whitfield, Michael F. (1997). Judging pain in infants: Behavioural, contextual and developmental determinants. *Pain*, 73(3), 319 – 324.

Hadjistavropoulos, T. & Craig, K. D. (2002). A theoretical framework for understanding self-report and observational measures of pain: A communications model. *Behaviour Research and Therapy*, 40(5), 551 – 570.

Hadjistavropoulos, Thomas, Baeyer, Carl von & Craig, Kenneth D. (2001). Pain assessment in persons with limited ability to communicate. In: Turk, Dennis C. & Melzack, Ronald (Hrsg.), *Handbook of pain assessment (2nd ed.).*, S. 134 – 149: Guilford Press.

Hadjistavropoulos, Thomas & Craig, Kenneth D. (2004). *Pain: Psychological perspectives.* Lawrence Erlbaum Associates Publishers.

Hadjistavropoulos, Thomas, Craig, Kenneth D. & Fuchs-Lacelle, Shannon (2004). Social Influences and the Communication of Pain. In: Hadjistavropoulos, Thomas & Craig, Kenneth D. (Hrsg.), *Pain: Psychological perspectives.*, S. 87 – 112: Lawrence Erlbaum Associates Publishers.

Hadjistavropoulos, Thomas, Herr, Keela, Turk, Dennis C., Fine, Perry G., Dworkin, Robert H., Helme, Robert, Jackson, Kenneth, Parmlee, Patricia A., Rudy, Thomas E., Beattie, B. Lynn, Chibnall, John T., Craig, Kenneth D., Ferrell, Betty, Ferrell, Bruce, Fillingim, Roger B., Gagliese, Lucia, Gallagher, Romayne, Gibson, Stephen J., Harrison, Elizabeth L., Katz, Benny, Keefe, Francis J., Lieber, Susan J., Lussier, David, Schmader, Kenneth E., Tait, Raymond C., Weiner, Debra K. & Williams, Jaime (2007). An Interdisciplinary Expert Consensus Statement on Assessment of Pain in Older Persons. *Clinical Journal of Pain*, 23(1), S1 – S43.

Hair, J.F., Wiliam, C, B., Babin, B.J., Anderson, R.E. & Tatham, R.L. (2006). *Multivariate Data Analysis.* Pearson University Press.

Hambleton, Ronald K. (1989). Principles and selected applications of item response theory. In: Linn, Robert L. (Hrsg.), *Educational measurement (3rd ed.).*, S. 147 – 200: Macmillan Publishing Co, Inc.

Harrison, David A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, 11(2), 91 – 115.

Hattie, John A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139 – 164.

Helme, R.D. & Gibson, S.J. (2001). The epidemiology of pain in elderly people. *Clin Geriatr Med*, 17, 417 – 431.

Herr, K.A. & Garand, L. (2001). Assessment and measurement of pain in older adults. *Clinics in geriatric medicine*, 17(3), 457 – 478.

Herr, Keela, Bjoro, Karen & Decker, Sheila (2006). Tools for Assessment of Pain in Nonverbal Older Adults with Dementia: A State-of-the-Science Review. *Journal of Pain and Symptom Management*, 31(2), 170 – 192.

Herr, Keela, Titler, Marita G., Schilling, Margo L., Marsh, J. Lawrence, Xie, Xianjin, Ardery, Gail, Clarke, William R. & Everett, Linda Q. (2004). Evidence-based assessment of acute pain in older adults: Current nursing practices and perceived barriers. *Clinical Journal of Pain*, 20(5), 331 – 340.

Hirsch, R.D. (2001). Sozio- und Psychotherapie bei Alzheimerkranken. *Zeitschrift für Gerontologie und Geriatrie*, 34, 92 – 100.

Holland, Paul W. (1981). When are item response models consistent with observed data? *Psychometrika*, 46(1), 79 – 92.

Horgas, Ann L. & Tsai, Pao-Feng (1998). Analgesic drug prescription and use in cognitively impaired nursing home residents. *Nursing Research*, 47(4), 235 – 242.

Horn, J. L. (1965). A rationale and test for the factors in factor analysis. *Psychometrika*, 30, 179–186.

Hoyle, R. H. & Panter, A. T. (1995). Writing about structural equation models. In: Hoyle, R. (Hrsg.), *Structural equation modeling: concepts, issues, and applications*, Kap. 9, S. 158–176: Thousand Oaks: Sage Publications, Inc.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6 (1), 1–55.

Hu, Li-tze & Bentler, Peter M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424 – 453.

Huffman, Jeff C. & Kunik, Mark E. (2000). Assessment and understanding of pain in patients with dementia. *The Gerontologist*, 40(5), 574 – 581.

Hulin, Charles L., Lissak, Robin I. & Drasgow, Fritz (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, 6(3), 249 – 260.

Hurley, Ann C., Volicer, Beverly J., Hanrahan, Patricia A. & Houde, Susan (1992). Assessment of discomfort in advanced Alzheimer patients. *Research in Nursing & Health*, 15(5), 369 – 377.

Husebo, Bettina S., Strand, Liv I., Moe-Nilssen, Rolf, Husebo, Stein B. & Ljunggren, Anne E. (2009). Pain behaviour and pain intensity in older persons with severe dementia: Reliability of the MOBID Pain Scale by video uptake. *Scandinavian Journal of Caring Sciences*, 23(1), 180 – 189.

Husebo, Bettina Sandgathe, Strand, Liv Inger, Moe-Nilssen, Rolf, Husebo, Stein Borge, Snow, Andrea Lynn & Ljunggren, Anne Elisabeth (2007). Mobilization-Observation-Behavior-Intensity-Dementia Pain Scale (MOBID): Development and validation of a nurse-administered pain assessment tool for use in dementia. *Journal of Pain and Symptom Management*, 34(1), 67 – 80.

Huynh, H., Michaels, H. & Ferrara, S. (1995). Statistical procedures to identify clusters of items with local dependency. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Janig, W. (1995). The sympathetic nervous system in pain. *Eur J Anaesthesiol Suppl*, 10, 53 – 60.

Jeitziner, Marie-Madlen (2008). Analgosedation auf der Intensivstation – und die Aufgabe der Pflege. NAPF-Symposium am Luzerner KantonsSpital, Wolhusen, Schweiz.

Jensen, Mark P. (2003). Questionnaire validation: A brief guide for readers of the research literature. *Clinical Journal of Pain*, 19(6), 345 – 352.

Jensen, M.P., Turner, J.A. & Romano, J.M. (1994). What is the maximum number of levels needed in pain intensity measurement? *Pain*, 58(3), 387 – 392.

Jöreskog, Karl G. & Sörbom, Dag (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Scientific Software International.

Jöreskog, K.G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology*, 23, 121–145.

Kaasalainen, Sharon & Crook, Joan (2003). A Comparison of Pain-Assessment Tools for Use with Elderly Long-Term-Care Residents. *CJNR: Canadian Journal of Nursing Research*, 35(4), 58 – 71.

Kaasalainen, S.J., Robinson, L.K., Hartley, T., Middleton, J., Knezacek, S. & Ife, C. (1998). The assessment of pain in the cognitively impaired elderly: a literature review. *Perspectives*, 22, 2 – 8.

Kale, B.K. (1962). On the solution of likelihood equation by iteration processes: the multiparametric case. *Biometrika*, 49, 479 – 486.

Kaspar, R., Becker, S. & Kruse, A. (2007). Abschlussbericht der Ergebnisse der H.I.L.DE.-Erfassung im Kooperationsprojekt MeDiA in Cura. Projektbericht: Institut für Gerontologie, Heidelberg.

Keefe, Francis J. & Block, Andrew R. (1982). Development of an observation method for assessing pain behavior in chronic low back pain patients. *Behavior Therapy*, 13(4), 363 – 375.

Keefe, Francis J., Williams, David A. & Smith, Suzanne J. (2001). Assessment of Pain Behaviors. In: Turk, Dennis C. & Melzack, Ronald (Hrsg.), *Handbook of pain assessment (2nd ed.).*, S. 170 – 190: Guilford Press.

Kovach, C., Griffie, J., Muchka, S., Noonan, P.E. & Weissman, D.E. (2000). Nurses' perceptions of pain assessment and treatment in the cognitively impired elderly. It's not a guessing game. *Clinical Nurse Specialist*, 14(5), 215 – 220.

Kovach, Christine R., Cashin, Jeffrey R. & Sauer, Linda (2006). Deconstruction of a complex tailored intervention to assess and treat discomfort of people with advanced dementia. *Journal of Advanced Nursing*, 55(6), 678 – 688.

Kovach, Christine R., Noonan, Patricia E., Griffie, Julie, Muchka, Sandy & Weissman, David E. (2002). The Assessment of Discomfort in Dementia Protocol. *Pain Management Nursing*, 3(1), 16 – 27.

Kovach, Christine R., Weissman, David E., Griffie, Julie, Matson, Sandy & Muchka, Sandy (1999). Assessment and treatment of discomfort for people with late-stage dementia. *Journal of Pain and Symptom Management*, 18(6), 412 – 419.

Kral, V.A. (1962). Senescent forgetfulness: Benign and malignant. *Canadian Medical Association Journal*, 86, 257 – 260.

Kremer, E., Atkinson, J.H. & Ignelzi, R.J. (1981). Measurement of pain: patient preference does not confound pain measurement. *Pain*, 10(2), 241 – 248.

Kretzschmar, Hans A. & Förstl, Hans (2009). Creutzfeldt-Jakob-Erkrankung und andere Prionkrankheiten. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis*. (2. Aufl.), S. 115 – 122: Heidelberg: Springer Verlag.

Krulewitch, Harry, London, Marla R., Skakel, Victoria J., Lundstedt, Glenda J., Thomason, Heidi & Brummel-Smith, Kenneth (2000). Assessment of pain in cognitively impaired older adults: A comparison of pain assessment tools and their use by nonprofessional caregivers. *Journal of the American Geriatrics Society*, 48(12), 1607 – 1611.

Kruse, A. & Schmitt, E. (2008). Kompetenzgruppen von an Demenz erkrankten Menschen – Den Blick auf die Ressourcen lenken. *Pflegezeitschrift*, 61(2), 77 – 81.

Kunz, Miriam (2006). *Veränderungen in der Schmerzverarbeitung bei Demenzpatienten: subjektive, mimische, motorische und vegetative Indikatoren*. Dissertationschrift, Otto-Friedrich-Universität Bamberg.

Kunz, Miriam & Lautenbacher, Stefan (2005). Veränderung des Schmerzerlebens bei Alzheimer-Patienten. *Zeitschrift für Neuropsychologie*, 16(4), 201 – 209.

Kunz, Miriam, Scharmann, Siegfried, Hemmeter, Uli, Schepelmann, Karsten & Lautenbacher, Stefan (2007). The facial expression of pain in patients with dementia. *Pain*, 133(1), 221 – 228.

Kurz, A. & Greschniok, P. (1994). Überlebenswahrscheinlichkeit bei Alzheimer-Krankheit. *Versicherungsmedizin*, 270, 59 – 62.

Larivière, Marianne, Goffaux, Philippe, Marchand, Serge & Julien, Nancy (2007). Changes in pain perception and descending inhibitory controls start at middle age in healthy adults. *Clinical Journal of Pain*, 23(6), 506 – 510.

Lautenbacher, S. (2004). Schmerzmessung. In: Basler, H.D., Franz, C., Kröner-Herwig, B. & Rehfisch, H.P. (Hrsg.), *Psychologische Schmerztherapie.*, S. 271 – 288: Heidelberg: Springer Verlag.

Lautenbacher, Stefan, Kunz, Miriam, Strate, Peter, Nielsen, Jesper & Arendt-Nielsen, Lars (2005). Age effects on pain thresholds, temporal summation and spatial summation of heat and pressure pain. *Pain*, 115(3), 410 – 418.

Lawton, M. P. (1991). A Multidimensional View of Quality of Life in Frail Elders. In: Birren, J. E., Lubben, J. E., Rowe, J. C. & Deutchman, D. E. (Hrsg.), *The Concepts and Measurement of Quality of Life in the Frail Elderly*, Kap. 1, S. 3–27: San Diego: Academic Press.

le Quintrec, J.L., Maga, M. & Baulon, A. (1995). L'échelle comportementale simplifiée (E.C.S.). *La Revue de Gériatrie*, 20(6), 363 – 368.

Lefebvre-Chapiro, S. & the DOLOPLUS group (2001). The DOLOPLUS2 scale – Evaluating pain in the elderly. *European Journal of Palliative Care*, 8, 191 – 194.

Leong, Ian Yi-Onn, Chong, Mei Sian & Gibson, Stephen J. (2006). The use of a self-reported pain measure, a nurse-reported pain measure and the PAINAD in nursing home residents with moderate and severe dementia: a validation study. *Age and Ageing*, 35, 252 – 256.

Leventhal, Elaine A., Leventhal, Howard, Shacham, Saya & Easterling, Douglas V. (1989). Active coping reduces reports of pain from childbirth. *Journal of Consulting and Clinical Psychology*, 57(3), 365 – 371.

Lewis, Thomas (1942). *Pain.* Macmillan.

Lord, F.M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M., Novick, M.R. & Birnbaum, Allan (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Lord, Frederic M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1(1), 95 – 100.

Loyd, B.H. (1988). Implications of item response theory for the measurement practitioner. *Applied Measurement in Education*, 1, 135 – 143.

Loyd, Brenda H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17(3), 179 – 193.

Luck, Tobias, Luppa, Melanie, Weber, Stephan, Matschinger, Herbert, Glaesmer, Heide, König, Hans-Helmut, Angermeyer, Matthias C. & Riedel-Heller, Steffi G. (2008). Time until institutionalization in incident dementia cases–Results of the Leipzig Longitudinal Study of the Aged (LEILA 75+). *Neuroepidemiology*, 31(2), 100 – 108.

Lueken, U., Seidl, U., Schwarz, M., Völker, L., Naumann, D., Mattes, K., Schröder, J. & Schweiger, E. (2006). Die Apathy Evaluation Scale: Erste Ergebnisse zu den psychometrischen Eigenschaften einer deutschsprachigen Übersetzung der Skala. *Fortschritte der Neurologie, Psychiatrie*, 74(12), p714 – 722.

Lueken, Ulrike, Seidl, Ulrich, Völker, Lena, Schweiger, Elisabeth, Kruse, Andreas & Schröder, Johannes (2007). Development of a short version of the Apathy Evaluation Scale specifically adapted for demented nursing home residents. *American Journal of Geriatric Psychiatry*, 15(5), p376 – 385.

Luppa, Melanie, Luck, Tobias, Brähler, Elmar, König, Hans-Helmut & Riedel-Heller, Steffi G. (2008). Prediction of institutionalisation in dementia: A systematic review. *Dementia and Geriatric Cognitive Disorders*, 26(1), 65 – 78.

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.

Mahoney, Alison E. J. & Peters, Lorna (2008). The Mahoney Pain Scale: Examining pain and agitation in advanced dementia. *American Journal of Alzheimer's Disease and Other Dementias*, 23(3), 250 – 261.

Mahoney, F. I. & Barthel, D. W. (1965). Functional evaluation: The Barthel Index. *Md Med*, 14, 56–61.

Marsh, H.W., Balla, J.R. & Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In: Marcoulides, G.A. & Schumacker, R.E. (Hrsg.), *Advanced statistical modeling*, S. 315–353. Mahwah: Lawrence Erlbaum.

Marzinski, Lynn R. (1991). The tragedy of dementia: Clinically assessing pain in the confused, nonverbal elderly. *Journal of Gerontological Nursing*, 17(6), 25 – 28.

Masters, Geoff N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149 – 174.

McArdle, J. J. & Boker, S. M. (1990). *RAMpath: Path diagram software.* Denver, CO: Data Transforms.

McArdle, J.J. & Aber, M.S. (1990). Patterns of change within latent variable structural equation models. In: von Eye, A. (Hrsg.), *Statistical models in longitudinal research*, Bd. 1, S. 151–224. San Diego, London: Academic Press.

McCaffery, M. & Beebe, A. (1994). *Pain, clinical manual for nursing practice.* London: Mosby.

McDonald, Roderick P. & Marsh, Herbert W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, 107(2), 247 – 255.

McDonald, Roderick P. & Mok, Magdalena M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, 30(1), 23 – 40.

McGrath, P.A., de Veber, L.L. & Hearn, M.T. (1985). Multidimensional pain assessment in children. In: Fields, H.L., Dubner, R. & Cervero, F. (Hrsg.), *Advances in pain research and therapy.*, S. 387 – 393: New York, NY: Raven Press.

McGrath, Patricia A., Seifert, Cheryl E., Speechley, Kathy N. & Booth, John C. (1996). A new analogue scale for assessing children's pain: An initial validation study. *Pain*, 64(3), 435 – 443.

McKinley, R.L. & Reckase, M.D. (1982). The use of the general Rasch model with multidimensional response data. Research Report ONR 82-1: Iowa City, IA, American College Testing (ACT).

Meade, Adam W., Lautenschlager, Gary J. & Hecht, Janet E. (2005). Establishing Measurement Equivalence and Invariance in Longitudinal Data With Item Response Theory. *International Journal of Testing*, 5(3), 279 – 300.

Melzack, R. (2001). Pain and the neuromatrix in the brain. *Journal of Dental Education*, 65, 1378 – 1382.

Melzack, R. & Wall, P.D. (1965). Pain mechanisms: a new theory. *Science*, 150, 971 – 979.

Melzack, Ronald & Katz, Joel (2004). The Gate Control Theory: Reaching for the Brain. In: Hadjistavropoulos, Thomas & Craig, Kenneth D. (Hrsg.), *Pain: Psychological perspectives.*, S. 13 – 34: Lawrence Erlbaum Associates Publishers.

Mense, S. (1993). Nociception from skeletal muscle in relation to clinical muscle pain. *Pain*, 54(3), 241 – 289.

Mense, Siegfried (1993). Peripheral mechanisms of muscle nociception and local muscle pain. *Journal of Musculoskeletal Pain*, 1(1), 133 – 170.

Merkel, S.I., Voepel-Lewis, T., Shayevitz, J.R. & Malviya, S. (1997). Practice applications of research. The FLACC: a behavioral scale for scoring postoperative pain in young children. *Pediatr Nurs*, 23, 293 – 297.

Merskey, H. & Bogduck, N. (Hrsg.) (1994). *Classification of Chronic Pain.* (2 Aufl.). Seattle, WA: IASP Press.

Möltner, A., Hölzl, R. & Strian, Friedrich (1990). Heart rate changes as an autonomic component of the pain response. *Pain*, 43(1), 81 – 89.

Mäntyselkä, P., Hartikainen, S., Louhivuori-Laako, K. & Sulkava, R. (2004). Effect of dementia on perceived daily pain in home-dwelling elderly people: a population-based study. *Age Ageing*, 33, 496 – 499.

Morello, R., Jean, A., Alix, M. & Groupe Regates (1998). L'ÉCPA: une échelle comportementale de la douleur pour personnes âgées non communicantes. *InfoKara*, 51(3), 22 – 29.

Morello, Remy, Jean, Alain, Alix, Michel, Sellin-Peres, Dominique & Fermanian, Jacques (2007). A scale to measure pain in non-verbally communicating older patients: The EPCA-2 Study of its psychometric properties. *Pain*, 133(1), 87 – 98.

Morris, J. C., Heyman, A., Mohs, R. C. & Hughes, J. P. (1989). The consortium to establish a registry for Alzheimer's disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology*, 39(9), 1159 – 1165.

Muraki, Eiji (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159 – 176.

Muthén, B. O. & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modelling in Mplus. *Mplus Web Notes: No. 4, Version 5*.

Muthén, B. O. & Muthén, L. K. (2006). *IRT in Mplus. Mplus Technical Appendix*.

Muthén, Bengt (1993). Goodness of fit test with categorical and other nonnormal variables. In: Bollen, K.A. & Long, J.S. (Hrsg.), *Testing structural equation models.*, S. 205 – 234: Newbury Park: Sage.

Muthén, Bengt & Hofacker, Charles (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika*, 53(4), 563 – 577.

Muthén, Bengt O. (1992). Latent variable modeling in epidemiology. *Alcohol Health & Research World*, 16(4), 286 – 292.

Muthén, L. K. & Muthén, B. O. (1998-2006). *Mplus User's Guide*. Los Angeles, CA : Muthén & Muthén.

Nevitt, Jonathan & Hancock, Gregory R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, 8(3), 353 – 377.

Nunnaly, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Nygaard, Harald A. & Jarland, Marit (2006). The Checklist of Nonverbal Pain Indicators (CNPI): testing of reliability and validity in Norwegian nursing homes. *Age and Ageing*, doi:10.1093/ageing/afj008, 79 – 81.

O'Bryant, Sid E., Humphreys, Joy D., Smith, Glenn E., Ivnik, Robert J., Graff-Radford, Neill R., Petersen, Ronald C. & Lucas, John A. (2008). Detecting Dementia With the Mini-Mental State Examination in Highly Educated Individuals. *Archives of Neurology*, 65(7), 963–967.

Olsson, Ulf Henning, Foss, Tron, Troye, Sigurd V. & Howell, Roy D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7(4), 557 – 595.

Parmelee, Patricia A., Mostofsky, David I., Lomranz, Jacob, Mostofsky, David I. & Lomranz, Jacob (1997). *Pain and psychological function in late life.*, S. 207 – 226. Plenum Press.

Pautex, S., Michon, A., Guedira, M., Emond, H., Le Lous, P., Samaras, D., Michel, J. P., Herrmann, F., Giannakopoulos, P. & Gold, G. (2006). Pain in severe dementia: Self-assessment or observational scales? *J Am Geriatr Soc*, 54, 1040–1045.

Ponocny, Ivo (2002). On the applicability of some IRT models for repeated measurement designs: Conditions, consequences, and goodness-of-fit tests. *Methods of Psychological Research*, 7(1), 21 – 40.

Porter, Fran Lang, Malhotra, Kristen M., Wolf, Cynthia M. & Morris, John C. (1996). Dementia and response to pain in the elderly. *Pain*, 68(2), 413 – 421.

Price, Donald D., McGrath, Patricia A., Rafii, Amir & Buckingham, Barbara (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, 17(1), 45 – 56.

Price, Donald D., Riley, Joseph L. & Wade, James B. (2001). Psychophysical Approaches to Measurement of the Dimensions and Stages of Pain. In: Turk, Dennis C. & Melzack, Ronald (Hrsg.), *Handbook of pain assessment (2nd ed.).*, S. 53 – 75: Guilford Press.

Prkachin, Kenneth M. (2007). The coming of age of pain expression. *Pain*, 133(1), 3 – 4.

Prkachin, Kenneth M. & Craig, Kenneth D. (1995). Expressing pain: The communication and interpretation of facial pain signals. *Journal of Nonverbal Behavior*, 19(4), 191 – 205.

Proctor, Wendey R. & Hirdes, John P. (2001). Pain and cognitive status among nursing home residents in Canada. *Pain Research & Management*, 6(3), 119 – 125.

Radbruch, Lukas, Sabatowski, Rainer, Loick, Georg, Jonen-Thielemann, Ingeborg, Kasper, Mario, Gondek, Barbara, Lehmann, Klaus A. & Thielemann, Ingeborg (2000). Cognitive impairment and its influence on pain and symptom assessment in a palliative care unit: Development of a minimal documentation system. *Palliative Medicine*, 14(4), 266 – 276.

Rainero, Innocenzo, Vighetti, Sergio, Bergamasco, Bruno, Pinessi, Lorenzo & Benedetti, Fabrizio (2000). Autonomic responses and pain perception in Alzheimer's disease. *European Journal of Pain*, 4(3), 267 – 274.

Rasch, Georg (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Raway, B. (1994). Pain behaviors and confusion in elderly patients with hip fracture. *Dissertation Abstracts International, 55*, 55(02B).

Raykov, Tenko (1997). Scale reliability, Cronbach's Coefficient Alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32(4), 329 – 353.

Raykov, Tenko (2007). Reliability if deleted, not 'alpha if deleted': Evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, 60(2), 201 – 216.

Re, Susanna (2006). *Erleben und Ausdruck von Emotionen bei schwerer Demenz*. Hamburg: Verlag Dr. Kovač.

Reckase, Mark D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25 – 36.

Reese, Lynda M.. (April 1995). The impact of local dependencies on some LSAT outcomes. Statistical Report 95-02: Law School Admission Council.

Richards, J. Scott, Nepomuceno, Cecilio, Riles, Maxine & Suer, Zehra (1982). Assessing pain behavior: The UAB Pain Behavior Scale. *Pain*, 14(4), 393 – 398.

Roman, G.C. (2005). Cholinergic dysfunction in vascular dementia. *Curr Psychiatry Rep*, 7, 18 – 26.

Rosenbaum, Paul R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49(3), 425 – 435.

Rosier, Elisa M., Iadarola, Michael J. & Coghill, Robert C. (2002). Reproducibility of pain measurement and pain perception. *Pain*, 98(1), 205 – 216.

Rost, Jürgen (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Verlag Hans Huber.

Rost, Jürgen & Langeheine, Rolf (Hrsg.) (1997). *Applications of latent trait and latent class models in the social sciences.* Waxmann Publishing Co.

Rudinger, G. & Rietz, C. (2001). Structural equation modeling in longitudinal research on aging. In: Birren, J.E. & Schaie, K.W. (Hrsg.), *Handbook of the psychology of aging* (5 Aufl.), S. 29–52. San Diego: Academic Press.

Samejima, Fumiko (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4), 100.

Samejima, Fumiko (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika*, 38(2), 221 – 233.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, 22, 249 – 278.

Satorra, A. & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In: von Eye, A. & Clogg, C.C. (Hrsg.), *Latent variables analysis: Applications for developmental research*, S. 399 – 419: Thousand Oaks, CA: Sage.

Satzger, W., Hampel, H., Padberg, F., Bürger, K., Nolde, Th., Ingrassia, G. & Engel, R. R. (2001). Zur praktischen Anwendung der CERAD-Testbatterie als neuropsychologisches Demenzscreening. *Nervenarzt*, 72(3), 196 – 203.

Scherder, E., Oosterman, J., Swaab, D., Herr, K., Ooms, M., Ribbe, M., Sergeant, J., Pickering, G. & Benedetti, F. (2005). Recent developments in pain and dementia. *BMJ*, 330, 461–464.

Scherder, E.J., Sergeant, J.A. & Swaab, D.F. (2003). Pain processing in dementia and its relation to neuropathology. *Lancet Neurol*, 2, 677 – 686.

Scherder, Erik & van Manen, Femke (2005). Pain in Alzheimer's disease: Nursing assistants' and patients' evaluations. *Journal of Advanced Nursing*, 52(2), 151 – 158.

Scherder, Erik J. A. (2000). Low use of analgesics in Alzheimer's disease: Possible mechanisms. *Psychiatry: Interpersonal and Biological Processes*, 63(1), 1 – 12.

Scherder, Erik J. A. & Bouma, Anke (1997). Is decreased use of analgesics in Alzheimer disease due to a change in the affective component of pain? *Alzheimer Disease & Associated Disorders*, 11(3), 171 – 174.

Scherder, Erik J. A. & Bouma, Anke (2000). Visual analogue scales for pain assessment in Alzheimer's disease. *Gerontology*, 46(1), 47 – 53.

Scherder, Erik J. A., Slaets, Joris, Deijen, Jan-Berend, Gorter, Yvonne, Ooms, Marcel E., Ribbe, Miel, Vuijk, Pieter-Jelle, Feldt, Karen, van de Valk, Marinus, Bouma, Anke & Sergeant, Joseph A. (2003). Pain Assessment in Patients with Possible Vascular Dementia. *Psychiatry: Interpersonal and Biological Processes*, 66(2), 133 – 145.

Schilling, O. (2004). *Längsschnittliche Analysen zur Entwicklung der Zufriedenheit im höheren Lebensalter*. Dissertationsschrift, Ruprecht-Karls-Universität Heidelberg, URL: http://www.ub.uni-heidelberg.de/archiv/4578.

Schofield, P., Clarke, A., Faulkner, M., Ryan, T., Dunham, M. & Howarth, A. (Apr 2005). Assessment of pain in adults with cognitive impairment: a review of the tools. unpublished manuscript, http://auraserv.abdn.ac.uk:9080/aura/handle/2164/167.

Schofield, P. & Reid, D. (2006). The assessment and management of pain in older people: A systematic review of the literature. *Int J Dis Human Dev*, 5(1), 9 – 15.

Scholderer, J. & Balderjahn, I. (2005). PLS versus LISREL - Ein Methodenvergleich. In: Bliemel, F., Eggert, A., Fassott, G. & Henseler, J. (Hrsg.), *Handbuch PLS-Pfadmodellierung: Methode, Anwendung, Praxisbeispiele.*, S. 87 – 98: Stuttgart: Schäffer-Poeschel.

Schröder, Stefan G. (2006). *Psychopathologie der Demenz*. Stuttgart: Schattauer.

Schuler, M. S., S., Becker, Kaspar, R., Nikolaus, T., Kruse, A. & Basler, H. D. (2007). Psychometric properties of a scale for the behavioural assessment of pain in nursing home residents with advanced dementia (PAINAD-G). *J Am Med Dir Assoc*, 8 (6), 388 – 395.

Schuler, Matthias (2008). *Schmerzerfassung bei Demenz – Evaluation einer Beobachtungsskala*. Habilitationschrift, Ruprecht-Karls-Universität Heidelberg.

Seidl, U., Lueken, U., Völker, L., Re, S., Becker, S., Kruse, A. & Schröder, J. (2007). Nicht-kognitive symptome und psychopharmakologische behandlung bei demenzkranken Heimbewohnern. *Fortschritte der Neurologie, Psychiatrie*, 75(12), p720 – 727.

Shega, Joseph W., Hougham, Gavin W., Stocking, Carol B., Cox-Hayley, Deon & Sachs, Greg A. (2004). Pain in Community-Dwelling Persons with Dementia: Frequency, Intensity, and Congruence Between Patient and Caregiver Report. *Journal of Pain and Symptom Management*, 28(6), 585 – 592.

Sign, B. & Orrell, M. (2003). The development, validity and reliability of a new scale for rating pain in dementia (RaPID). unpublished manuscript.

Simons, W. & Malabar, R. (1995). Assessing pain in elderly patients who cannot respond verbally. *J Adv Nurs*, 22, 663 – 669.

Sirsch, Erika (2009). Schmerzerfassung als Fremdeinschätzung. Welche Schmerzeinschätzung ist die Richtige? Dementia Fair Congress, Februar 2009, Hamburg.

Sivo, Stephen A., Fan, Xitao, Witta, E. Lea & Willse, John T. (2006). The search for 'optimal' cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education*, 74(3), 267 – 288.

Sjöström, Björn (1995). *Assessing acute postoperative pain: Assessment strategies and quality in relation to clinical experience and professional role*. PhD thesis, University of Göteborg.

Skrondal, Anders & Rabe-Hesketh, Sophia (2004). *Generalized latent variable Modeling.* Interdisciplinary Statistics: Washington, D.C.: Chapman & Hall/CRC.

Slinde, Jeffrey A. & Linn, Robert L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, 15(1), 23 – 35.

Smith, Marianne (2005). Pain Assessment in Nonverbal Older Adults with Advanced Dementia. *Perspectives in Psychiatric Care*, 41(3), 99 – 111.

Snow, A. Lynn, O'Malley, Kimberly J., Cody, Marisue, Kunik, Mark E., Ashton, Carol M., Beck, Cornelia, Bruera, Eduardo & Novy, Diane (2004). A Conceptual Model of Pain Assessment for Noncommunicative Persons With Dementia. *The Gerontologist*, 44(6), 807 – 817.

Steyer, R., Eid, M. & Schwenkmezger, P. (1997). Modeling true intraindivdual change: True change as a latent variable. *Methods of Psychological Research Online*, 2, 21–33. Internet: http://www.pabst-publishers.de/mpr/ (20.02.2009).

Steyer, R., Partchev, I. & Shanahan, M.J. (2000). Modeling true intraindividual change in structural equation models: The case of poverty and children's psychosocial adjustment. In: Little, T.D., Schnabel, K.U. & Baumert, J. (Hrsg.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*, S. 109–126. Mahwah: Lawrence Erlbaum.

Stolee, Paul, Hillier, Loretta M., Esbaugh, Jacquelin, Bol, Nancy, McKellar, Laurie & Gauthier, Nicole (2005). Instruments for the Assessment of Pain in Older Persons with Cognitive Impairment. *Journal of the American Geriatrics Society*, 53(2), 319 – 326.

Stout, William (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589 – 617.

Stout, William F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55(2), 293 – 325.

Suen, Hoi K. (1990). *Principles of test theories.* Lawrence Erlbaum Associates, Inc.

Swaminathan, H. & Gifford, J.A. (1983). Estimation of parameters in the three parameter latent trait model. In: Weiss, D. (Hrsg.), *New horizons in testing*, S. 13 – 30: New York: Academic Press.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In: Bollen, K.A. & Long, J.S. (Hrsg.), *Testing structural equation models*, S. 10–39. Newbury Park, London, New Dehli: Sage.

te Marvelde, Janneke M., Glas, Cees A. W., Van Landeghem, Georges & Van Damme, Jan (2006). Application of Multidimensional Item Response Theory Models to Longitudinal Data. *Educational and Psychological Measurement*, 66(1), 5 – 34.

Teske, Karen, Daut, Randall L. & Cleeland, Charles S. (1983). Relationships between nurses' observations and patients' self-reports of pain. *Pain*, 16(3), 289 – 296.

Thalmann, B. & Monsch, A.U. (1997). *CERAD – Neuropsychologische Testbatterie: Vorläufige Normen.* Memory Clinic Basel, Hebelstrasse 10, CH-4031 Basel.

Thalmann, B., Monsch, A.U., Schneitter, M., Ermini-Fünfschilling, D., Spiegel, R. & Stähelin, H.B. (1998). *Die CERAD Neuropsychologische Testbatterie. Ein gemeinsames minimales Instrumentarium zur Demenzabklärung*. Memory Clinic Basel, Hebelstrasse 10, CH-4031 Basel.

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning.* Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Tisak, J. & Meredith, W. (1990). Descriptive and associative developmental models. In: von Eye, A. (Hrsg.), *Statistical methods in longitudinal research*, Bd. II, S. 387–406. San Diego: Academic Press.

United Nations Department of Economic and Social Affairs/Population Disvision (Hrsg.) (2006). *World population prospects: The 2004 Revision, Volume III: Analytical Report.*

van Baalen, B., Odding, E., van Woensel, M. P. C, van Kessel, M. A., Roebroeck, M. E. & Starn, H. J. (2006). Reliability and sensitivity to change of measurement instruments used in a traumatic brain injury population. *Clinical Rehabilitation*, 20(8), 686 – 700.

van Herk, Rhodee, van Dijk, Monique, Baar, Frans P. M., Tibboel, Dick & de Wit, Rianne (2007). Observation Scales for Pain Assessment in Older Adults with Cognitive Impairments or Communication Difficulties. *Nursing Research*, 56(1), 34 – 43.

van Nispen, Stephanie (2008). A new approach to the item reduction of PACSLAC. 9th European Doctoral Conference in Nursing Science, September 2008, Maastricht, The Netherlands.

Velicer, Wayne F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321 – 327.

Villanueva, M.R., Smith, T.L., Erickson, J.S., Lee, A.C. & Singer, C.M. (2003). Pain Assessment for the Dementing Elderly (PADE): reliability and validity of a new measure. *J Am Med Dir Assoc*, 4, 1 – 8.

Volicer, L., Fabiszewski, K.J., Rheaume, Y.L. & Lasch, K.E. (1988). *Clinical management of Alzheimer's disease.* Rockville, MD: Aspen Publishers.

von Davier, Alina A. & Wilson, Christine (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement*, 67(6), 940 – 957.

von Davier, Matthias (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research*, 2(2), 29 – 48.

Walsh, N., Schoenfeld, L., Ramamurth, S. & Hoffman, J. (1989). Normative model for the cold pressor test. *Am J Phys Med Rehabil*, 68, 6 – 11.

Ward, W.C. (1986). Measurement research that will change test design for the future. In: Freeman, E.E. (Hrsg.), *The redesign of testing for the 21st century, Proceedings of the 1985 ETS Invitational Conference.*, S. 25 – 34: Princeton, NJ: Educational Testing Service.

Warden, V., Hurley, A. & Volicer, L. (2003). Development and psychometric evaluation of the Pain Assessment IN Advanced Dementia (PAINAD) Scale. *J Am Med Dir Assoc*, 4, 9–15.

Ware, E. Jr. (November 2001). Using Item Response Theory To Construct QOL Measures For Children: The Dynamic Health Assessment (DynHA) Approach. 5th EC Framework Programme, Plenary Session Quality of Life and Management of Living Resources Leiden, The Netherlands.

Ware, J. E. Jr., Bjorner, J. B. & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing. A brief summary of ongoing studies of widely used headache impact scales. *Med Care*, 38(9, Suppl. 2), 73–82.

Wary, B. & Doloplus Consortium (1999). Doloplus-2, a scale for pain measurement. *Soins Gerontol.*, 19, 25 – 27.

Webb, Eugene J., Campbell, Donald T., Schwartz, Richard D. & Sechrest, Lee (1966). *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago, IL: Rand McNally.

Weindl, Adolf (2009). Parkinson Plus, Lewy-Körperchen-Demenz, Chorea Huntington und andere Demenzen bei Basalganglienerkrankungen. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis.* (2. Aufl.), S. 85 – 114: Heidelberg: Springer Verlag.

Weiner, Debra, Peterson, Bercedis & Keefe, Francis (1999). Chronic pain-associated behaviors in the nursing home: Resident versus caregiver perceptions. *Pain*, 80(3), 577 – 588.

Weyerer, Siegfried, Schäufele, M. & Hendlmeier, I. (2005). Besondere und traditionelle stationäre Betreuung demenzkranker Menschen im Vergleich. *Zeitschrift für Gerontologie und Geriatrie*, 38(2), 85 – 94.

WHO, World Health Organisation (2009). Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme 10. Revision German Modification. Online-Ressource: http://www.dimdi.de: Deutsches Institut für Medizinische Dokumentation und Information DIMDI.

Williamson, A. & Hoggart, B. (2005). Pain: a review of three commonly used pain rating scales. *J Clin Nurs*, 14, 798 – 804.

Wimo, A., Winblad, B., Aguero-Torres, H. & Von Strauss, E. (2003). The magnitude of dementia occurrence in the world. *Alzheimer Dis Assoc Disord*, 17(2), 63–67.

Wolf-Klein, G.P., Siverstone, F.A., Brob, M.S., Levy, A. Folcy, C.J., Termotto, V. & Breuer, J. (1988). Are Alzheimer patients healthier? *J Am Geriatr Soc*, 36, 219–224.

Woodrow, Kenneth M., Friedman, Gary D., Siegelaub, A. B. & Collen, Morris F. (1972). Pain tolerance: Differences according to age, sex and race. *Psychosomatic Medicine*, 34(6), 548 – 556.

Wright, B.D. & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

Yen, Wendy M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125 – 145.

Yen, Wendy M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187 – 213.

Zaudig, Michael (2009). Leichte Kognitive Beeinträchtigung im Alter. In: Förstl, Hans (Hrsg.), *Demenzen in Theorie und Praxis.* (2. Aufl.), S. 23 – 42: Heidelberg: Springer Verlag.

Zenisky, April L., Hambleton, Ronald K. & Robin, Frédéric (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, 63(1), 51 – 64.

Zenisky, April L., Hambleton, Ronald K. & Robin, Frédéric (2003). DIF Detection and Interpretation in Large-Scale Science Assessments: Informing Item Writing Practices. *Educational Assessment*, 9(1), 61 – 78.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223 – 233.

Zwakhalen, S., Hamers, J., Abu-Saad, H. H. & Berger, M. (2006). Pain in elderly people with severe dementia: A systematic review of behavioural pain assessment tools. *BMC Geriatr*, 6(3), doi: 10.1186/1471–2318–6–3.

Zwakhalen, S., Hamers, J. & Berger, M. (2006). The psychometric quality and clinical usefulness of three pain assessment tools for elderly people with dementia. *Pain*, 126, 210–220.