Multimodales Assessment und Beschwerdenvalidierung mittels MMPI-2, MMPI-2-RF und BHI-2

Analyse einer Stichprobe deutscher Patienten mit chronischen Schmerzen

Inauguraldissertation

zur Erlangung der Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

Vorgelegt von

Wolfgang Richter

aus Bünde

Bonn 2016

Gedruckt mit Genehmigung der Philosophischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Vorsitzender:	Prof. Dr. Rainer Banse
Betreuer und erster Gutachter:	PD Dr. Ralf Dohrenbusch
Zweiter Gutachter:	Prof. Dr. André Beauducel
Weiteres prüfungsberechtigtes Mitglied:	Prof. Dr. Ulrich Ettinger

Tag der mündlichen Prüfung: 31.08.2016

Zusammenfassung

Einleitung: In der vorliegenden Studie wurden 400 chronische Schmerzpatienten zu Beginn einer zweiwöchigen, stationären verhaltensmedizinischen Therapie hinsichtlich Symptom-Überhöhungen untersucht, wie sie bei sozialmedizinischen Konfliktkonstellationen gehäuft vorkommen. Dabei sollte die Nützlichkeit von zwei im deutschen Sprachraum neuen Diagnostika (MMPI-2-RF, BHI-2) zur Aufdeckung der sog. Malingered Pain-Related Disability (MPRD) überprüft werden. Etablierte MMPI-2-Validitätsskalen wurden mit ihren Äquivalenz-Skalen der kürzeren Revisionsform MMPI-2-RF verglichen sowie mit Skalen der deutschen Battery of Health Improvement (BHI-2).

Methode: Entsprechend dem Ansatz von Bianchini et al. (2005) wurden die Patienten vor der Behandlung mit sechs externen Beschwerden-Validierungsverfahren (BV) zum Assessment behavioral-somatischer, psychopathologischer und kognitiver Symptome vier Gruppen zugeordnet: (1.) Patienten mit authentischen Beschwerdeschilderungen, (2.) Patienten mit externalen Motiven für Overreporting, aber ohne Auffälligkeiten in den externen BV, (3.) Patienten mit externen Aggravationsmotiven und möglicher sowie (4.) sicherer Aggravation.

Ergebnisse: Multivariate Varianz- und ROC-/AUC-Analysen belegten eine signifikante, vergleichbare Diskriminanzgüte aller Validitätsskalen des MMPI-2-RF mit dem MMPI-2 (mit Ausnahme der L-r-Skala), wie auch beider BHI-2-Validitätsskalen, aber auch diverser weiterer BHI-2-Basisskalen. Trennschärfste Prädiktoren des MMPI-2/-RF waren die Seltene-Symptome-Skala F-r, die Response-Bias-Skala (RBS-r) sowie der integrative Meyers Validity Index. Zusätzlich konnte mittels acht, für den MMPI-2-RF adaptierter Validitätsskalen eine Identifikation von Probanden mit sicherer und möglicher Antwortverzerrung dokumentiert werden. Ein empirisch entwickelter, gewichteter Index der sechs trennschärfsten MMPI-2-RF-Validitätsskalen (Revised Overreporting Index ROI) zeigte eine exzellente Diskriminanz (Sensitivität 96 %, Spezifität 90 %, Cohen's d = 3,39 bzw. 1,21), erwies sich jedoch gegenüber den Standard-Validitätsskalen des MMPI-2-RF (integriert im MI-r-Index) als nicht signifikant trennschärfer. Patienten mit negativen Antwortverzerrungen zeigten nach der Therapie ein signifikant geringeres Therapie-Outcome.

Schlussfolgerung: MMPI-2-RF und BHI-2 stellen somit trennscharfe und valide Diagnostika zur Identifizierung nicht-authentischer Beschwerden-Darstellungen bereit, die insbesondere bei der Begutachtung schmerzassoziierter Beschwerden eingesetzt werden können.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Wenn wir die Arbeitsfähigkeit eines Menschen einschätzen, ist in diesem Urteil immer und unvermeidlich auch eine Einschätzung seiner Willenskraft enthalten. Demnach ist die Arbeitsfähigkeit überhaupt nicht objektiv messbar, ihre Einschätzung ist geradezu ein Akt des eigenen Willens des Urteilenden.

Viktor von Weizsäcker (1931)

1 Einleitung

Im Einleitungskapitel werden die Grundlagen multimodaler Erhebungen und Methoden der Beschwerdenvalidierung bei der Begutachtung von Beeinträchtigungen, die Patienten mit chronischen Schmerzen beklagen, hergeleitet und diskutiert.

Dazu dient zunächst ein kurzer Überblick über die Phänomenologie, die diagnostische Klassifikation und das Auftreten chronischer Schmerzen in der bundesdeutschen Allgemeinbevölkerung sowie über Einflussfaktoren auf den Chronifizierungsverlauf, die bei der gutachterlichen Einschätzung schmerzbedingter Beeinträchtigungen bedeutsam sind. Neben familiären, sozialen und arbeitsplatzbezogenen Aspekten wird dabei Faktoren der medikamentösen Behandlung chronischer Schmerzen besondere Beachtung geschenkt. Im Anschluss werden die Folgen einer chronischen Schmerzsymptomatik im Hinblick auf die häufig drohende Erwerbsminderung und damit die finanziell-ökonomischen Probleme des Patienten mit chronischen Schmerzen diskutiert.

Im Anschluss werden historische und moderne Ansätze der Begutachtung chronischer Schmerzen in den letzten Jahrzehnten dargelegt, die in jüngster Zeit entsprechend eines multimodalen Assessment-Modells von Bianchini et al. (2005) [31] zum Einbezug spezieller Methoden der Validierung der von Patienten präsentierten Einschränkungen und Beschwerden in drei Bewertungsdomänen führten: der Ebene kognitiv-neuropsychologischer Leistungsdefizite, der Ebene psychopathologischer Symptome sowie der Ebene somatischer und verhaltensbezogener Einschränkungen.

Zu allen Bewertungsebenen des vorgeschlagenen Begutachtungsprozesses werden detaillierte methodische Möglichkeiten der Diagnostik sowie deren Vorteile und Begrenzungen diskutiert.

Insbesondere werden dabei die beiden in der durchgeführten Studie untersuchten Fragebogenverfahren des MMPI-2, der bislang nur in den USA validierten Revisionsfassung MMPI-2-RF sowie der in Europa noch weniger bekannte BHI-2 hinsichtlich ihres Vorgehens, ihrer Möglichkeiten und Chancen kritisch besprochen. Die Diskussion der Studienergebnisse zu den in beiden Fragebögen integrierten diversen Subskalen sowie neueste Untersuchungen zur Revisionsfassung des MMPI-2 führen schließlich zur theoretischen Begründung und Festlegung der fünf Haupt-Hypothesen der hier dargelegten wissenschaftlichen Studie.

1.1 Akutschmerz, Chronische Schmerzen, Schmerzkrankheit

Schmerzen sind, entsprechend der Definition der Welt-Gesundheits-Organisation (WHO), ein unangenehmes Sinnes- und Gefühlserleben in Verbindung mit einer (tatsächlichen oder drohenden) Gewebeschädigung (Nilges & Nagel 2007 [215]). Die Ursachen solcher Beschwerden sind meist akut (Entzündung, Verletzung, Druck, Ischämie o.ä.), so dass die Beschwerden nach einer Ausheilung meist abklingen. Schmerzen haben bei solchen Beschwerden eine lebens-, manchmal überlebenswichtige Warn-Funktion, die den Organismus vor weiteren Schädigungen schützen soll, zum Beispiel durch Ruhigstellung der betroffenen Körperregion und Schonung. Ziel der Behandlung von Akutschmerzen ist somit in der Regel die Beseitigung der Schmerzursache mit dem Ergebnis einer Schmerzfreiheit (s. Abb. 1, S. 3).

Schmerzen bezeichnen wir als chronisch, wenn sie länger als die zur Ausheilung übliche Zeit anhalten oder periodisch wiederkehren (z.B. bei der Migräne). Bonica (1953) [36] benannte als üblichen Zeitraum zunächst einen ca. vierwöchigen Heilverlauf. Insofern trifft auf die meisten Schmerzen der heute meist genannte Chronifizierungsbeginn nach einem halben Jahr nicht auf alle Schmerzarten zu (z.B. Chronifizierung der postherpetischen Neuralgie auf spinaler ganglionärer Ebene mit Maximum der Degeneration bereits binnen zwei Wochen; vgl. Hempel 1993 [130]).

Chronische Schmerzen lösen sich rasch als organismisches Warnsignal von ihrer eigentlichen Ursache ab, verselbstständigen sich, breiten sich auf Rezeptorebene exponentiell aus, werden als Schmerzgedächtnis gespeichert und reproduziert. Sie überdauern ihren ursprünglichen Anlass und können so zur Krankheit selbst werden. Insofern sind ihre Ursachen oft

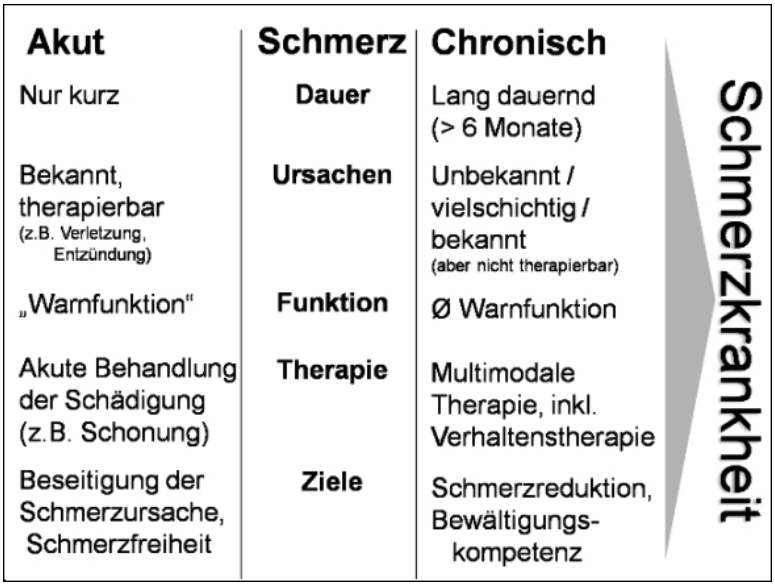


Abb. 1. Unterschiede akuter und chronischer Schmerzen (nach: Kröner-Herwig B (2008). Der Schmerz. Hamburg: TK-Schriftenreihe)

vielschichtig oder unbekannt. Eine einfache diagnostische Festlegung oder auch einfache ätiologische Zuordnung ist oft nicht möglich (Zimmermann 1996 [328]).

Oder aber die Ursachen der Beschwerden sind bekannt (z.B. eine chronische Osteoporose oder eine chronische Encephalitis dissiminata), diese sind aber nicht heilbar. Damit verändert sich das Therapieziel chronischer Schmerzen: Gegenüber dem Maximalziel einer Beschwerdefreiheit gewinnt die Schmerzreduktion auf nach der Behandlung verbleibende, aushaltbare Restschmerzen durch Erlernen einer besseren Bewältigungskompetenz durch die betroffenen Patienten Priorität. Eine solche Therapie kann konsequenterweise nicht nur eindimensional die organisch-somatische Komponente der Schmerzen zum Behandlungsfokus haben, sondern muss einem biopsychosozialen Modell durch Einbezug psychischer und sozialer Beschwerdeanteile (z.B. durch Einbezug psychologischer Verhaltenstherapie) Rechnung tragen (vgl. Kröner-Herwig et al. 2011 [160]).

Wenn Schmerzen chronisch werden, verändern sie meist den gesamten Menschen. Anhaltende Schmerzen beanspruchen oft die gesamte Aufmerksamkeit der Betroffenen und rauben die notwendige Energie für das alltägliche Leben (Richter et al. 2000 [231]). Oft werden Ein- und Durchschlafstörungen durch Schmerzen und grübelnde Gedanken ausgelöst und verstärkt. Die Stimmung wird gereizter und es kann zu interpersonellen Spannungen kommen. Die alltägliche Schaffenskraft der Patienten nimmt ab, die Stimmung wird mit an-

haltenden Beschwerden hoffnungsloser und resignierter (Richter 2012 [229], 2016 [230]). Wenn alles Denken, Fühlen und Wollen fast nur noch um Schmerzen kreist, nennt man das Schmerzkrankheit. Schmerzkranke Patienten bedürfen einer spezialisierten fachärztlichen und fachpsychotherapeutischen Behandlung.

Merckelbach & Merten 2012 [191]) berichten, dass in manchen Studien 30 Prozent der Patienten mit somatoformen Schmerzstörungen in mindestens einem Verfahren zur Beschwerdenvalidierung Auffälligkeiten zeigten und 11 Prozent dieser Probanden in mindestens zwei SVTs (Symptom-Validierungs-Verfahren) auffällig waren. Hierbei waren Probanden mit somatoformer Schmerzstörung nicht von instruierten Simulanten unterscheidbar.

Schmerzen als subjektives Erleben sind nicht objektiv messbar, sondern nur indirekt über verbale Aussagen und / oder non-verbale Verhaltensmuster zu erschließen. Sie werden von psychischen Alterationen überlagert und im Fall somatoformer Störungen ausschließlich durch psychische Mechanismen ausgelöst und aufrechterhalten. Dies erschwert insbesondere eine objektive Begutachtung der durch Schmerzen bedingten Funktionseinschränkungen und Behinderungen.

1.2 Epidemiologie chronischer Schmerzen

Schätzungen zufolge sind in der Bundesrepublik Deutschland 17 % der Bevölkerung bzw. 15 Millionen Menschen von chronischen, länger andauernden oder wiederkehrenden Schmerzen betroffen. Fünf bis sechs Millionen Menschen leiden unter anhaltenden und sehr starken Schmerzen; 900.000 Patienten sind entsprechend der vorangehenden Erläuterungen schmerzkrank. Am häufigsten sind die Betroffenen von Schmerzen des Bewegungsapparates beeinträchtigt. Chronische Rückenschmerzen kommen bei 22 % der Frauen und 15 % der Männer vor, muskuloskelettale Schmerzen treten bei 16 % der Allgemeinbevölkerung auf. Chronische Kopfschmerzen sind mit 15 % betroffener Frauen und 8 % beeinträchtigter Männer etwas seltener. Selbst Neuropathien kommen als vergleichsweise seltene Syndrome bei 8 % der Allgemein-Bevölkerung vor (Wenig et al. 2009 [312]).

Die Hälfte der betroffenen Menschen befürchtet aufgrund von Arbeitsausfallzeiten durch die Schmerzen negative Folgen für ihre Berufs-Tätigkeit. 18 % der Patienten haben anhaltend eine schmerzbedingte Arbeitsunfähigkeit. 60 Prozent der Patienten mit Schmerzen werden bereits bei einer Krankschreibung von mehr als 6 Monaten frühberentet und bei Arbeitsunfähigkeitszeiten von mehr als 12 Monaten werden 80 % der Patienten vorzeitig erwerbsunfähig. Durch Arbeitsunfähigkeitszeiten entstehen ca. 49 Millionen Euro Kosten jährlich in Europa, was 2,2 % des Bruttoinlandsproduktes der Bundesrepublik Deutschland entspricht (Wenig et al. 2009 [312]).

Krankheiten der Wirbelsäule sind die häufigste Ursache für Arbeitsunfähigkeitszeiten, deren resultierende Produktivitätsausfälle volkswirtschaftlich auf 16 bis 22 Milliarden Euro geschätzt werden. Allein chronische Rückenschmerzen verursachen jährlich in der Bundesrepublik neun Millionen Euro Produktionsausfälle. Alleinlebende Männer höheren Lebensalters beispielsweise mit geringer Schulbildung, die unter höherem Grad chronifizierter Rückenschmerzen leiden und arbeitslos werden, hatten entsprechend multivariater Varianzanalysen den signifikant größten Anteil an diesen Krankheitskosten. Mit der Dauer der Arbeitslosigkeit erhöhen sich zudem signifikant die psychopathologischen Komorbiditäten. Hierzu zählen neben Angst- und Depressions-Erkrankungen insbesondere auch Sucht- und Abhängigkeits-Erkrankungen (Schubert et al. 2013, IAB Forschungsbericht [265]).

1.3 Klassifikation und Nosologie chronischer Schmerzsyndrome

Chronische Schmerzen können mit Erweiterung der 9. Version der Internationalen Krankheits-Klassifikation ICD-10 mit der Ziffer F45.41 als biopsychosoziale Störung umschrieben werden. Die Diagnose F45.41 bezieht sich auf Schmerzen, die durch eine somatische Störung ausgelöst und wahrscheinlich auch aufrechterhalten werden. Solche Faktoren sind beispielsweise muskuloskelettale Funktionsstörungen von Muskeln, Bändern oder Sehnen, aber auch pathophysiologische knöcherne Prozesse (Bandscheibenvorfall, Osteoporose), wie auch entzündliche Ursachen (Arthrose, Sakroilitis o.ä.), nervale Schäden (z.B. Neuralgien), Folgen maligner Erkrankungen, Folgen zerebraler Störungen (z.B. nach einem Apoplektischen Insult) oder Durchblutungsminderungen (z.B. ischämisch bedingter, peripherer Verschlusskrankheit). Hinzukommen im Chronifizierungsverlauf solcher Schmerzen psychische Alterationen, in Verbindung mit Stress und Belastungen, aber auch depressive Stimmungsveränderungen, Rückzugsverhalten, Angstaffekte oder auch Schon- und Vermeidungs-

Verhalten. Meist werden die Schmerzen auf der psychischen Ebene fehlerhaft verarbeitet, führen zu dysfunktionalen Kognitionen (z.B. Schmerzzentriertheit, Katastrophisierung und Dramatisierung), wie auch zu familiären und sozialen Konfliktkonstellationen (Geissner & Jungnitsch 1992 [97]).

Differentialdiagnostisch sind chronische Schmerzen abzugrenzen von akuten Schmerzen aufgrund einer umschriebenen Gewebsverletzung sowie von rein körperlich verursachten Schäden ohne psychische Begleitkomponente, die dann entsprechend ihrer körperlichen Zuordnung in den ICD-10-Kapiteln außerhalb der F-Ziffern zu kodieren sind. Die Kombinations-Diagnose F45.41 ist aber auch abzugrenzen gegenüber primär psychischen Erkrankungen (z.B. spezifischen Angst-Symptomatiken F40.0 - 48.9 oder Depressions-Erkrankungen F30.0 - F39). Die Diagnose F45.41 beschreibt nicht die primär anhaltende somatoforme Schmerzstörung (F45.40), bei der ein beschreibbarer psychischer Auslöser der Schmerzerkrankung vorliegt, der hauptverantwortlich für Beginn, Schweregrad, Exazerbation oder Aufrechterhaltung der Schmerzen ist. Beide Schmerz-Diagnosen (F45.40 und F45.41) sind differenzialdiagnostisch abzugrenzen von undifferenzierten Somatisierungsstörungen entsprechend F45.1, die sich auf unspezifische Befindlichkeitsstörungen beziehen, die keiner konkreten Erkrankung zuzuordnen sind.

Mit der Ziffer F45.41 wurde erstmals eine angemessene Beschreibungsentität für die biopsychosoziale Diagnostik chronifizierter Schmerzzustände gefunden, die bis 2009 nur durch Hilfsdiagnosen (z.B. Andauernde Persönlichkeitsänderung bei chronischem Schmerzsyndrom F62.80) in den Klassifikationskatalogen ICD-9 und DSM-IV realisiert wurde.

Nilges & Rief (2010) [216] weisen jedoch zurecht auf einige auch weiterhin optimierbare Mängel der aktuellen Klassifikations-Möglichkeiten hin, die auch in den aktuell gültigen Katalogen nicht behoben sind (vgl. http://www.g-drg.de/cms, S. 133pp.). So findet sich der chronische Schmerz im Kapitel 18 unter den sog. „Symptomen und abnormen klinischen (Befunden) und Laborbefunden, die nicht anders klassifizierbar sind“. Auch sei der Kode für die Schmerzlokalisation als Hauptdiagnose anzugeben, wenn ein Patient speziell zur Schmerzbehandlung stationär aufgenommen werde. Insofern hat die Diagnose F45.41 trotz ihres Eingangs in die Diagnoseklassifikation noch immer keine wirkliche Repräsentanz im Krankheitenkatalog gefunden. Fehlt eine Haupt-Lokalisation der Schmerzen sind die Beschwerden nach den derzeitigen Verschlüsselungsregeln als „Chronisch-unbeeinflussbarer Schmerz“ (R52.1) oder als „Sonstiger chronischer Schmerz (R52.2)“ zu umschreiben. Die

Autoren weisen zurecht auf hier notwendige Veränderungen der Kodierrichtlinien in künftigen Diagnosekatalogen hin.

Häuser (2002a) [126] unterscheidet chronische Schmerzen adäquaten Ausmaßes mit adäquater Reaktion von solchen Schmerzen inadäquaten Ausmaßes sowie Schmerzen mit auch zusätzlich inadäquater Verarbeitung. Häuser schlägt ferner vor, chronische Schmerzen adäquaten Ausmaßes, bei denen es zu psychischen Reaktionen käme, die jedoch als noch nicht „inadäquat“ einzuschätzen wären, mit den Zusatzdiagnosen Anpassungsstörung (F43.2) oder anhaltende affektive Störung (F34.8) zu versehen. Zu unterscheiden seien die ersten drei Schmerz-Klassifikationen von chronischen Schmerzen bei primärer psychischer Krankheit. Die mit diesen Festlegungen verbundene hohe Subjektivität mit erheblichem Ermessensspielraum auf Seiten des Beurteilenden verdeutlicht ein Grundproblem der Klassifikation bei chronischem Schmerz.

Einen alternativen Klassifikations-Ansatz, der auch die Schmerzlokalisation einbezieht, ohne ätiologische Aussagen zu treffen, stellte Mersky (1986 [193], in deutscher Adaptation nach Maier et al. 2000 [175]), mit der Multiaxialen Schmerzklassifikation vor. Die Psychologische Dimension MASK-P (Klinger et al. 2000, 2011 [152], [151]) stellt eine phänomenologische Beschreibung der Schmerzen auf fünf psychosozialen Ebenen dar (motorisch-verhaltensbezogen, emotional, kognitiv, Stressoren und Personenmerkmale).

Die MASK-S-Klassifikation ist das Ergebnis einer mehrjährigen intensiven Diskussion einer bundesdeutschen, multizentrischen Arbeitsgruppe der Deutschen Schmerzgesellschaft, die sich zwischen 1993 und 1999 regelmäßig zu Konsens-Diskussionen traf. Im Schmerzdiagnose-System MASK-S werden die Beschwerden des Patienten in acht Hauptgruppen detailliert hinsichtlich der Lokalisation der Schmerzen gegliedert. Folgende MASK-Hauptgruppen wurden auf Expertenebene übereinstimmend definiert: Gruppe 1 Kopfschmerz-Syndrome, Gruppe 2 Gesichtsschmerzen, Gruppe 3 Schmerzen des Nervensystems, Gruppe 4 Schmerzen bei Störungen des Durchblutungssystems, Gruppe 5 Chronische Rückenschmerzen, Gruppe 6 Muskuloskelettale Schmerzsyndrome, Gruppe 7 Viszeral-Schmerzen, Gruppe 8 Psychiatrische und Somatoforme Schmerzsyndrome. Vorteil dieses Klassifikations-Verfahrens ist es, eine sehr genaue Beschreibung der chronischen Schmerzsymptomatik hinsichtlich ihrer Lokalisation unter bewusstem Verzicht jeglicher ätiologischer Annahmen zu ermöglichen, wie sie der ICD-10-Schematisierung zugrunde liegen. Damit werden mit der MASK-Beschreibung noch keine Entscheidungen über die Ursachen, Auslöser oder

Aufrechterhaltungs-Faktoren der berichteten Schmerzen getroffen, die sowohl auf somatischer, psychischer oder sozialer Ebene liegen können oder auf einer Kombination solcher Faktoren beruhen.

Eine an modernen Regeln orientierte Beurteilung der Behinderungen durch chronische Schmerzen erfordert die wertneutrale Identifikation somatischer, psychischer und sozialer Aspekte der Störung, unabhängig von ätiologischen Erklärungen. Wenn chronische Schmerzen somit multidimensional bedingt sind und aufrechterhalten werden, muss auch die Einschätzung aus der Erkrankung resultierender Leistungsdefizite mehrdimensional erfolgen.

1.4 Biopsychosoziale Aspekte chronischer Schmerzen

Mit Entwicklung der Verhaltensmedizin ermöglichte der Paradigmenwechsel in der Erforschung psychophysiologischer Störungen (Schwartz & Weiss 1978 [266]) die heute moderne Erkenntnis, dass Schmerzen nicht nur eine nozizeptive Sinneswahrnehmung sind, sondern mehrdimensionale Verhaltensreaktionen umfassen (Fordyce 1976 [93]); neben physiologisch-biomedizinischen Anteilen sind Schmerzsyndrome nicht nur auf motivational-emotionaler Ebene beschreibbar, sondern auch als kognitiv-evaluative und auch als behaviorale Reaktionen.

Die Bedeutung psychischer Störungen ist dabei in ihrer Kausalität nicht festgelegt: Psychische Störungen können einer chronischen Schmerz-Erkrankung prämorbid vorangehen und diese ursächlich auslösen. Sie sind insbesondere aber auch die häufigste reaktive Folge einer chronifizierten Schmerzsymptomatik.

So berichteten bei Polatin et al. (1993) [224] 57 % der befragten Patienten mit chronischen Rückenschmerzen bereits vor ihrer Erkrankung von Ängsten und Depressionen; jedoch gaben 43 % der Patienten solche Störungen auch als Folge ihrer Schmerz-Symptomatik an. Nach Currie & Wang (2004) [61] wiesen Patienten mit Rückenschmerzen in der Allgemeinbevölkerung ein mehr als dreimal höheres Risiko für die Entwicklung einer klinisch relevanten Depression auf als nicht unter Schmerzen Leidende. Einige wenige prospektive Studien,

die den noch ungeklärten Ursache-Wirkungs-Zusammenhang zwischen Schmerzen und der Belastung durch psychische Störungen untersuchten, vermerkten mehrheitlich psychische Störungen häufiger als Folge des Leidens an den Schmerzen als ihre Bedingung durch die prämorbiden Belastungen der Patienten (z.B. Linton 2000 [173], Fishbain et al. 1997 [88]). Psychische Störungen scheinen bei chronisch schmerzkranken Patienten durch spezifische dysfunktionale gesundheitsbezogene Erwartungen, Attributionen, Einstellungen und Überzeugungen negativ begünstigt zu werden sowie durch familiäre und arbeitsbezogene Faktoren verstärkt zu werden (Ruoß 1997 [255]).

Im Folgenden soll kurz auf einige wichtige psychosoziale Folgen chronischer Schmerzen eingegangen werden, die bei jeder Diagnostik und Begutachtung betroffener Patienten berücksichtigt werden müssen: die Rolle der Familie und des sozialen Umfelds der Patienten, unter den Behandlungsfaktoren insbesondere Einflüsse der medikamentösen Therapie auf den Erkrankungsverlauf sowie den Einfluss von Arbeitsplatz und Berufsstatus des Patienten.

1.4.1 Familiäre Einflüsse und Schmerzchronifizierung

Fordyce (1976) [93] lenkte mit der Formulierung eines operanten Bedingungsmodells für die Aufrechterhaltung chronischer Schmerzen den Focus auf die verhaltenssteuernden Einflüsse der sozialen Umwelt des Patienten im Sinne positiver Verstärkung von Krankheitsverhalten sowie negativer Verstärkung des Patienten durch Entlastung, Schonung und Rückzug. Während letztere Verhaltensweisen durchaus bei akuten Schmerzen die Ausheilung von Erkrankungen erleichtern können, tragen sie bei chronischen Schmerzzuständen eher zu deren Fortschreiten bei (Kröner-Herwig 2011 [160]).

Familiäre Bedingungen und Einflüsse der sozialem Umwelt des Patienten können vermutlich ebenso protektiv wie auch dysfunktional auf die Entwicklung einer chronischen Schmerzsymptomatik einwirken.

Klinische Erfahrungsberichte, aber auch explorative Studien stützen die Annahme, dass Patienten mit chronischen Schmerzen, die allein oder ohne soziale Unterstützung leben, eine schlechtere Therapieerfolgschance haben, als Patienten, die mit zumindest einer weiteren Bezugsperson zusammenleben (Sternbach 1974 [283], Roy 1982 [253]). Gleichzeitig belegen eine Reihe retrospektiver Studien, dass Patienten mit chronischen Schmerzen häufig aus schwierigen Familienverhältnissen stammen, in denen teils kollusive, übermäßig enge oder

tabuisierende Zusammenhänge dominieren. Zudem leidet häufig bei Patienten mit chronischen Schmerzen zumindest ein Elternteil ebenfalls unter chronischen Schmerzen (Turk et al. 1987 [301]). Hierbei dürften Modelllerneffekte der Kinder von ihren Eltern mit Übernahme von Krankheitsverhalten eine chronifizierende Rolle spielen, worauf auch Übereinstimmungen der Schmerzen der Patienten mit dem Schmerzsyndrom meist gleicher Lokalisation bei ihren Eltern hinweisen. Auch Payne & Norfleet (1986) [219] bemerkten in einem Review über eine Vielzahl vergleichbarer Studien, dass Patienten mit chronischen Schmerzen häufig aus Familien mit hoher allgemeiner Erkrankungsrate und hoher Rate von Schmerz-Krankheiten stammten. Ziesat (1978) [327] (s. auch Payne & Norfleet 1986 [219]) stellte zudem fest, dass Patienten mit chronischen Schmerzen signifikant häufiger die jüngsten Kinder einer Familie oder Einzelkinder waren, verbunden mit der Hypothese besonderer familiärer Aufmerksamkeit dieser Kinder.

Aber auch die Schmerzkrankheit hat oft erhebliche Auswirkungen auf das familiäre Zusammenleben. Freizeit-Einschränkungen, die Abnahme außerfamiliärer Kontakte. Aber auch emotionale Schwierigkeiten zwischen den Familienmitgliedern entwickeln sich ebenso häufig als Folge anhaltender Schmerzen, wie auch Einschränkungen beruflicher und finanzieller Art, mit dann negativen Folgen für das eheliche und familiäre Zusammenleben (Turk et al. 1983 [302]). So berichteten Rowat & Knafl (1985) [252] (s. auch Flor et al. 1989 [92]) bei vier von fünf Partnern von Patienten mit chronischen Schmerzen emotionale Störungen und soziale Belastungen, aber auch eigene physische Beschwerden. Dabei waren insbesondere die Frauen männlicher Patienten hoch belastet.

Block (1981) [34] berichtete, dass vor allem Patienten mit einer hohen emotionalen Belastung bei Schmerzexpressionen der Patienten höhere physiologische Erregungen zeigten. Ähnlich berichteten Maruta et al. (1981) [179] eine hohe eheliche und sexuelle Unzufriedenheit bei mehr als 70 Prozent der Patienten mit chronischen Schmerzen, häufiger waren eheliche Unzufriedenheiten bei den Partnern der Patienten festzustellen.

In einer Studie von Romano et al. (1989) [249] schätzten ebenfalls die Lebenspartner der Patienten ihre eheliche Unzufriedenheit und Depression höher ein als die Patienten. Auch Flor et al. (1987) [91] bestätigten, dass die angegebene Schmerzintensität der Patienten und ein niedriges Aktivitätsniveau hoch mit der Erwartung eines eher unterstützenden Partnerverhaltens korrelierte. Swanson & Maruta (1980) [288] bestätigten zudem, dass eine hohe Übereinstimmung zwischen Patienten und ihren Partnern hinsichtlich bestimmter Schmerz-

merkmale einen signifikanten Prädiktor für einen späteren Therapie-Misserfolg darstellte. Flor et al. (1989) [92] sahen diesen Zusammenhang zwischen unterstützendem Partnerverhalten und hoher subjektiver Schmerzbelastung der Patienten in gleicher Weise, und zwar insbesondere bei den männlichen Patienten.

Eine Beobachtungs-Studie von Block et al. (1981) [34] erbrachte zusätzlich das Ergebnis, dass die untersuchten Patienten mit chronischen Schmerzen intensivere Belastungsangaben in Anwesenheit eines Arztes machten als bei einer Befragung durch eine als nicht helfend eingeschätzte Person (z.B. Stationshilfe). Dies stützt die Annahme einer auch in Diagnose-, Behandlungs- und Begutachtungs-Kontexten erhöhten Beschwerdendarstellung von Patienten (Verdeutlichungstendenz).

So kann es in Behandlungs-, wie auch in familiären Kontexten (Kröner-Herwig 2011) [160] zu einer therapeutisch ungünstigen, „komplementären (Behandlungs-)Konstellation“ kommen, bei der die „Schwäche des Patienten zur Stärke des Gegenüber (des Therapeuten)“ wird, wobei der Patient einer eigenen Verantwortungsübernahme für den Behandlungsverlauf enthoben wird, und damit auch des Effizienz-Erlebens eigener Bewältigungsstrategien, einhergehend mit einer Zunahme der Depression des Patienten.

Interaktionsmuster und die ihnen zugrundeliegenden Verstärkungsmechanismen zwischen Patienten mit chronischen Schmerzen und ihren Bezugspersonen (insbesondere Familienangehörigen, aber auch Therapeuten, Gutachtern) können die Beschwerdepräsentation verstärken oder reduzieren. Das Ausmaß an Beschwerdenaggravation sollte deshalb in Begutachtungs-Situationen stärker ausgeprägt und häufiger sein als unter klinischen Bedingungen. Zu prüfen wäre demnach die Hypothese, ob vermehrte *familiäre Interaktionsprobleme* ebenso wie *operante Aspekte der Schmerzaufrechterhaltung* eindeutig mit nachweislichen Tendenzen der Beschwerdeüberhöhung einhergehen, z.B. durch Einsatz des BHI-2.

1.4.2 Medikamentöse Einflussfaktoren und Suchtgefahren

Während noch in den 1990er Jahren eine „schmerztherapeutische Unterversorgung von Patienten mit chronischen Schmerzen mit Opioid-Analgetika“ diskutiert wurde, stieg mit der Verordnungs-Möglichkeit retardierter Opioide exponentiell deren Einsatz in der Schmerztherapie an. Zenz et al. (1990) [326] sahen bei mehr als der Hälfte von 70 Patienten mit chronischen Schmerzen mit nicht-malignen chronischen Schmerzen unter Opioid-Behandlung eine signifikante Schmerzreduktion. Teilweise hätten Erkrankungen mit jahrzehntelangen Anamnesen beendet werden können. Dieser Ansatz wurde seinerzeit teilweise sogar bei Patienten mit chronischen Schmerzen mit Suchtanamnese als mögliches Behandlungsprozedere diskutiert (Sipos et al. 2000 [276]). Portenoy (1996) [225] konstatierte sechs Jahre später, dass nach kritischem Überblick erster Erfahrungen zumindest bei einer Subgruppe von Patienten mit chronischen Schmerzen unter Berücksichtigung der Effizienz, Sicherheit und der Suchtgefahren der Einsatz retardierter Opioide Vorteile für die Therapie bieten kann. Hingegen machten bereits in derselben Zeit andere Autoren (z.B. Mindach 2000 [206]) darauf aufmerksam, dass das Fehlen von Abhängigkeits-Entwicklungen durch Opioide bei Patienten mit chronischen nicht-malignen Schmerzen bisher nicht hinreichend durch Studien belegt ist. Weitere acht Jahre später wird der Einsatz zentral wirksamer Analgetika bei chronischen Schmerzen sehr viel kritischer bewertet (Maier 2008 [174]), insbesondere, wenn schnell anflutende, kurzwirksame Opioide in der Therapie Einsatz finden. In der Bundesrepublik stiegen die Therapietage mit hochpotenten Opioiden seit Einführung retardierter Opioide in 15 Jahren auf das 13-fache an, hingegen fanden sich nur geringere Verbesserungen der Lebensqualität, der Arbeitsfähigkeit und der Schmerzstärke im selben Zeitraum (Kalso et al. 2004 [147], Martell et al. 2007 [178]).

Insofern sprechen heute einige Schmerztherapeuten von einer Fehl-Entwicklung. Iatrogene Einflüsse auf den Chronifizierungsprozess mancher Patienten mit chronischen Schmerzen werden daher heute kritisch diskutiert. Maier (2008) [174] plädiert dafür, Patienten mit chronischen Schmerzen mit speziellen Risikofaktoren zur Entwicklung einer Sucht frühzeitig zu *identifizieren*. Zudem sollten nach Ansicht dieser renommierten Schmerztherapeuten entsprechende Leitlinien überprüft werden, die bisher selbst somatoforme Störungen nur als „relative“ Kontraindikation für eine Opioid-Therapie benennen. Es ist offenbar, dass auch das Vorliegen einer sozialmedizinischen Konfliktsituation mit Overreporting der Beschwer-

desituation zu diesen Risikofaktoren einer Fehlindikation von Opioid-Behandlungen gezählt werden muss.

Aus moderner Sicht werden Opioid-Analgetika in der Schmerztherapie zu häufig und zu hoch dosiert eingesetzt, obwohl sie häufig bei sekundären Motiven der Krankheitsaufrechterhaltung zu keiner Symptomreduktion beitragen. In Begutachtungen wird hingegen häufiger die Opioid-Verabreichung im falschen Umkehrschluss als Argument für stärkere Schmerzen und größere Behinderung proklamiert. Wenn dies so ist, sollten Patienten mit negativen Antwortverzerrungen häufiger eine Einnahme von Opioid-Analgetika zeigen, als Patienten mit authentischen Beschwerdeschilderungen.

1.4.3 Arbeitsplatzbezogene Einflussfaktoren

Gralow (2000) [109] berichtet, dass anhaltende Schwerstarbeit, aber auch eine nicht ergonomische Gestaltung des Arbeitsplatzes, monotone Arbeiten, auch in Zwangshaltungen, die Entstehung und Chronifizierung von Schmerzen negativ begünstigen kann. Aber auch ungünstige persönliche Merkmale, wie ein niedriges Ausbildungsniveau, sowie prädisponierende Erkrankungen chronischer Schmerzen (z.B. jugendlicher M. Scheuermann), eine geringe Berufsqualifikation, Unzufriedenheit mit der Arbeitssituation und Konflikte mit Vorgesetzten oder Kollegen können die Chronifizierung von Schmerzen befördern. Mit der Dauer der Krankschreibung wächst zudem subjektiv die Ursachenattribution der Beschwerden auf den Arbeitsplatz und damit die Unzufriedenheit, wie Burton (1997) [46] berichtet. Lange Zeiten der Krankschreibung korrelierten zudem in weiteren Studien mit einem niedrigen Bildungsstand und der Ausübung von Arbeiten ohne besondere Qualifikation (Junge et al. 1996 [145]).

In einer bevölkerungsrepräsentativen Stichprobe in New South Wales (Australien) stellten Blyth et al. (2003) [35] bei 484 telefonisch befragten Probanden fest, dass sie im Durchschnitt an 16,4 Tagen im letzten halben Jahr wegen Schmerzen arbeitsunfähig (4,5 Tage) waren oder ihre Arbeit mit Schmerzen am Arbeitsplatz verrichteten (83,8 Tage). Die Autoren folgerten, dass in den meisten Fällen eine Arbeitstätigkeit trotz Schmerzen möglich

war, somit das Therapieziel einer Schmerzbehandlung nicht notwendig Schmerzfreiheit sein müsse.

Insofern könnten Veränderungen in Bezug auf die psychische oder körperliche Belastungssituation am Arbeitsplatz möglicherweise, bei erhaltener Motivation zur Wiedereingliederung in den Beruf, eine Erleichterung des Berufseinstiegs fördern, auch bei nach erfolgreicher Behandlung verbleibenden Schmerzen. Ferner könnten Veränderungen der zeitlichen Struktur der Arbeitsabläufe arbeitsmedizinische Verbesserungs-Möglichkeiten darstellen, aber auch Veränderungen der räumlichen Gestaltung des Arbeitsplatzes, wie Veränderungen zu Aspekten zur Förderung höherer Leistungsmotivation und der Leistungsfähigkeit am Arbeitsplatz.

Im Umkehrschluss sind in der Anamnese deutliche Diskrepanzen und Inkonsistenzen zwischen den angegebenen Einschränkungen in der beruflichen Tätigkeit gegenüber sonst eher unbelastet ausführbaren Aktivitäten im Freizeitverhalten sehr klare Hinweise auf negative Angabeverzerrungen (Widder et al. 2008 [314]).

> Wenn physikalische und psychosoziale Mehr-Belastungen am Arbeitsplatz eine bedeutsame Rolle bei der Entwicklung von Schmerzbeschwerden spielen, sollten diese Aspekte bei Patienten mit chronischen Schmerzen mit Beschwerde-Überhöhungen ebenfalls häufig als Belastungsmomente genannt werden.

Da aus chronischen Schmerzsymptomen häufig Arbeitsausfallzeiten und dauernde Erwerbsminderungen resultieren, wird auf diesen Aspekt schmerzbedingter Arbeitsunfähigkeit im Folgenden näher eingegangen.

1.5 Chronische Schmerzen und Erwerbsminderung

Frühberentungen aufgrund chronischer Schmerzstörungen, beispielsweise Rückenschmerzen oder Störungen des muskulo-skelettalen Systems, stellten bis zum Jahr 2000 die häufigsten Gründe für eine Frühberentung dar. Seit dem Jahr 2000 waren zunehmend mehr Frühberentungen aufgrund psychischer Störungen und Verhaltensstörungen zu verzeichnen,

beispielsweise wegen eines Burnout oder somatoformer Schmerzstörungen. Solche psychischen Gründe stellen in der Bundesrepublik Deutschland seit 2002 bei Männern und Frauen einen kontinuierlich steigenden Anteil der Rentenneuzugänge wegen Erwerbsminderung dar (2009: 44 % der Frauen und 32 % der Männer; 2012: 41 % bei allen Frühberentungen, s. Dt. RV Bund 2011 [65], s.a. Schneider 2007 [261]).

Ebenfalls steigt seit 20 Jahren stetig die Inanspruchnahme von Berufsunfähigkeits-Versicherungen wegen psychischer Erkrankungen (Dohrenbusch 2011 [70]). Hierbei hat sich bei der Risikobewertung einer Inanspruchnahme von Leistungen das Kriterium einer Psychotherapie in der Vor-Anamnese als prognostisch nicht hinreichend valide erwiesen. Vielmehr zeigten sich u.a. folgende Faktoren als prognostisch relevant für eine spätere Frühinvalidität: eine psychische Erkrankung im frühen Erwachsenenalter, geringe Arbeitszufriedenheit, Arbeitslosigkeit nach Schulabschluss, geringes Ausbildungsniveau, geringe kognitive Fähigkeiten, soziale Isolation, Gebrauch von Alltagsdrogen, auch negatives Allgemeinbefinden.

Auch in der Allgemeinbevölkerung hat die Prävalenz psychischer Störungen generell zugenommen; nach einer Einjahresprävalenz leiden 27 % der Erwachsenen in Europa an einer psychischen Störung (Wittchen & Jacobi 2005 [319]). Hierbei handelte es sich vor allem um eine Zunahme der Diagnosen aus dem psychoneurotischen und psychosomatischen Formenkreis. Bei schweren psychiatrischen Diagnosen als Begründungen der Erwerbsunfähigkeit war hingegen keine epidemiologische Zunahme zu vermerken.

Frühpensionierungen von Beamten aufgrund psychischer Störungen, insbesondere die Zahl frühpensionierter, verbeamteter Lehrer, nahmen in den letzten Jahren deutlich zu, so dass mehr als die Hälfte dieser Arbeitnehmer im Durchschnitt 10 Jahre vor Erreichen des Altersruhestands in die Frührente überwechselten (Weber et al. 2004 [310]).

Nach einer jüngsten Erhebung der Bundespsychotherapeutenkammer (2014) [44] zeigte sich nach Auswertung der Daten der größten sechs deutschen Krankenkassen (AOK, Barmer, GEK, BKK, DAK und TK) nahezu eine Verdoppelung der Arbeitsunfähigkeitstage wegen psychischer Erkrankungen von 2000 bis 2012 (Zunahme um 96 %). Ebenso verlängerte sich die Zahl der Krankschreibungen wegen psychischer Störungen um 31 % in diesem Zeitraum.

In diesen jüngsten Erhebungen spiegelte sich vor allem die exponentiell weiter bis in das Jahr 2012 steigende Anzahl von Frührenten aufgrund psychischer Störungen wider (s. Abb. 2, S. 16). Bei den psychischen Diagnosen war dieser Anstieg vornehmlich auf eine Zunahme

der depressiven Störungen (ICD-10-F30.X) zurückzuführen (s. Abb. 3, S. 17). Die Diagnose einer psychischen Störung führte in 2012 dazu, dass die betroffenen Patienten bereits mit durchschnittlich 49 Jahren frühberentet wurden, hingegen Frühberentungen wegen körperlicher Erkrankungen erst im Durchschnitt mit 50 bis 55 Jahren erfolgten.

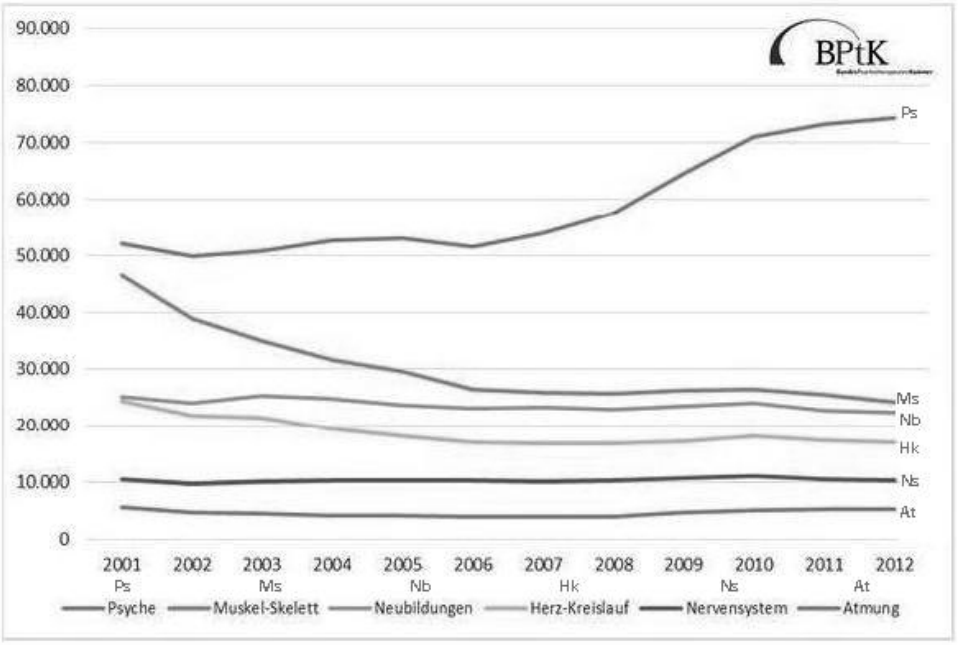


Abb. 2. Rentenneuzugänge wegen verminderter Erwerbsfähigkeit pro Jahr aufgrund der sechs wichtigsten Krankheitsarten; adaptiert nach Quelle: Bundespsychotherapeutenkammer (2014) [44]

Nach den Statistiken der Bundespsychotherapeutenkammer waren im Jahr 2012 20,9 % der mit (psychopathologisch basierend) F-Diagnosen begründeten Frührenten auf ICD-10 F4-Diagnosen zurückzuführen (s. Abb. 3, S. 17). Damit lag der Gesamtanteil der durch *chronische Schmerzsyndrome* (nach ICD-10 F45.41 und F45.40) begründeten Frührenten bei 20,9 % von 42,1 % (dem Anteil von Frührenten aufgrund psychischer Gründe) und somit bei maximal 8,8 % aller vorzeitigen Renten wegen Erwerbsunfähigkeit.

Von dieser Zahl (9 % aller Frührenten wegen chronischer Schmerzen) sind zusätzlich die Frührenten wegen anderer F4-Diagnosen abzuziehen, z.B. Posttraumatischer Belastungsstörungen (F43.1), Anpassungsstörungen (F43.2) oder Somatisierungsstörungen (F45.1).

Angesichts dieser Zahlen stellt sich die Frage, ob diese Verschiebung Abbild einer tatsächlichen epidemiologischen Zunahme psychischer Störungen ist, oder aber Abbild einer seit 1999 zunehmenden Beteiligung psychologischer Psychotherapeuten am Gesundheitswesen, einer geringeren sozialen Stigmatisierung psychischer Erkrankungen, einer verbesserten me-

dizinischen Versorgung und Behandlung somatischer Erkrankungen, der Verbesserung der Klassifikation von Erkrankungen oder aber auch mangelnder Trennschärfe des psychodiagnostischen Instrumentariums.

Dabei ist gerade die Differenzierung tatsächlicher körperlicher oder psychischer Funktionsdefizite von anderen, in Untersuchungs- und Behandlungs-Situationen teilweise häufigen Tendenzen der Symptom-Verdeutlichung (Overreporting, Exageration), der Aggravation oder sogar der Simulation von besonderer Bedeutung.[1] Auch stellt die Diskrimination unbewusster und intentionaler Mechanismen der Symptomverfälschung (z.B. durch hysterische Konversion / Somatisierung vs. Malingering / Übertreibung / Simulation) eine der zentralen psychodiagnostischen Fragen dar (Bianchini et al. 2005 [31]).

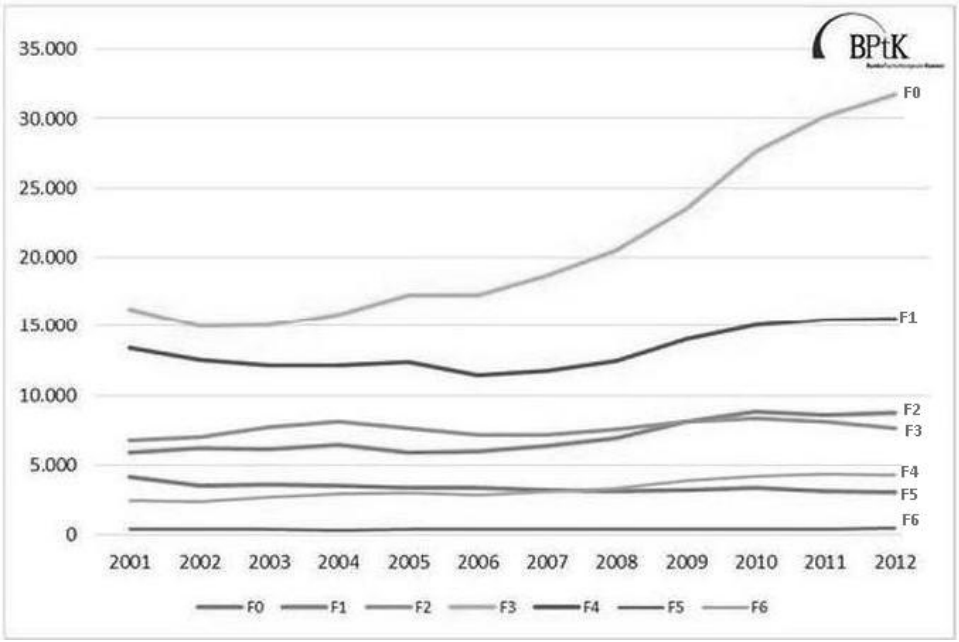


Abb. 3. Zahl der Rentenneuzugänge wegen verminderter Erwerbsfähigkeit pro Jahr aufgrund verschiedener psychischer Störungsgruppen; adaptiert nach Quelle: Bundespsychotherapeutenkammer (2014) [44]

So wurden in einer großen Studie mit mehr als 33.000 Gutachtenpatienten in den USA bei 31 Prozent der Patienten mit chronischen Schmerzen Hinweise auf Simulationsverhalten festgestellt (Mittenberg et al. 2002 [208]). Auch verstärkt nach Studien von Rohling et al. (1995) [248] die Aussicht auf finanzielle Kompensation die Angaben von Intensitätswerten bei Patienten mit chronischen Schmerzen erheblich und mindert den Behandlungserfolg. Die Aussicht auf finanzielle Kompensation hat zudem reziproke Effekte auf die angegebene

[1] Detaillierte Begriffs-Erläuterungen finden sich im folgenden Kapitel 1.6, ab S. 19

Beeinträchtigung durch Schmerzen und durch Depression (Rainville et al. 1997 [227]). Die Möglichkeit eines langen Krankengeldbezuges verlängert und intensiviert anderen Untersuchungen zufolge das Leiden an Rückenschmerzen (Cassidy et al. 2000 [53], McNaughton et al. 2000 [186]). Nach Abschluss einer Berentung wurden zudem in einer Studie abnehmende Angaben der Belastung durch Schmerzen beobachtet (Swartzman et al. 1996 [289]). Verschiedene Studien zeigen, dass Patienten im Streit um Entschädigungszahlungen ihre Beschwerden tendenziell stärker hervorheben als Patienten innerhalb einer Therapie (Birke et al. 2001 [32], Blyth et al. 2003 [35]).

Selbstberichte über Schmerzen ebenso wie Angaben zu Funktions- und Leistungsbehinderungen beinhalten Risiken der Interpretation, die erst durch Einsatz psychodiagnostischer Tests kontrolliert und überprüft werden können (Dohrenbusch 2009 [69]). Bei der Begutachtung von Patienten mit chronischen Schmerzen wird deshalb zunehmend neben der Erhebung subjektiver Symptomcharakteristika (wie Schmerzintensität, schmerzassoziierter Psychopathologie wie Depression und Angst, Behinderungs- und Lebensqualitätserleben) Wert auf den kreuzvalidierten Abgleich unterschiedlicher Datenquellen inklusive wenig- oder nicht-durchschaubarer Untersuchungsmethoden gelegt (vgl. Wagner, Richter et al. 2003 [309], wie es auch die aktuellen Begutachtungsleitlinien für chronische Schmerzen der Fachgesellschaften nahelegen, Dohrenbusch et al. 2008 [72]).

Frühberentungen aufgrund psychischer Störungen erfolgten in den letzten Jahren zunehmend früher im Lebensalter und zunehmend häufiger. Dadurch entstehen öffentlichen Trägern anhaltende hohe, erhebliche Krankheitsfolgekosten. Patienten mit Symptom-Verdeutlichungen sind zudem im Vergleich zu glaubwürdig psychisch Beeinträchtigten die jüngeren Patienten mit noch ausgeprägterer Beschwerdepräsentation. Lange Arbeitsunfähigkeit und zu erwartende finanzielle Kompensationen begünstigen nach bisherigen Studien ein solches Aggravationsverhalten. Deshalb sind Entwicklung und Überprüfung sensibler diagnostischer Instrumentarien zur Identifikation von Aggravationsverhalten und Overreporting besonders wichtig.

1.6 Begutachtung chronischer Schmerzsyndrome

Die Begutachtung von Funktionsbeeinträchtigungen und Erwerbsminderungen stellt eine besondere Herausforderung für den medizinischen wie psychologischen Gutachter dar. Prozesstechnisch steht dabei der Antragsteller, gerichtlich der Kläger, gegenüber dem Gericht in der Beweispflicht seiner Beschwerden. Der Gutachter leistet seine Einschätzungen hingegen als sachverständiger, unabhängiger Zeuge zur besseren Information für das beurteilende Sozialgericht. Der Gutachter ist damit keiner der Prozessparteien verpflichtet (auf Patientenseite: dem Kläger, seinen Rechtsvertretern und Vorbehandlern; auf Seite der Beklagten: der beklagten Versicherung, dem medizinischen Dienst der Krankenkasse, der Berufsgenossenschaft etc.). Der Gutachter ist einer möglichst objektiven, dem Patienten gerecht wertenden Einschätzung verpflichtet.

Diese Einschätzung stellt bei Patienten mit chronischen Schmerzen aufgrund der Komplexität, Vielschichtigkeit von somatischen und psychischen Befunden, aber auch oft langer Anamnese mit entsprechend vielen Vorbehandlungen, eine besondere Herausforderung dar. Trotz dieser Komplexität werden zumeist nur Einzelfachgutachten auf einem der medizinischen Fachgebiete erstattet. Seltener, bei algesiologischen Gutachten beispielsweise, werden allerdings komplexe und kombinierte, medizinisch-psychologische Zusammenhangs-Gutachten gefordert und von den Auftraggebern beantragt. Bei Wagner et al. 2003 [309] gaben beispielsweise 86 % der angefragten **schmerztherapeutischen Gutachter** an, kombinierte, medizinisch-psychologische Gutachten zu erstatten.

Der Gutachter sollte unter Würdigung sämtlicher Vorberichte und Vorgutachten in einer umfassenden Exploration möglichst alle relevanten, die Erkrankung beeinflussenden körperlichen und psychosozialen Faktoren klären, so z.B. neben Merkmalen der Schmerzsymptomatik, -modulation und des -verhaltens, bisherige Behandlungserfahrungen des Patienten, seine Attributions- und kausalen Erklärungsmodelle für seine Symptomatik, ferner Perspektiven und Veränderungsmotivationen, Informationen zum alltäglichen Verhalten sowie zu Reaktionen wichtiger Bezugspersonen. Ebenso wichtig ist die Berücksichtigung emotionaler Reaktionen im Krankheitsverlauf, beispielsweise Depressivität und Angst, aber auch Bewältigungsansätze, Ressourcen und persönliche Kompetenzen, aber auch ausführliche biographische Informationen unter Würdigung der Persönlichkeit des Untersuchten, die die Krankheits-Entstehung und -Chronifizierung erklären könnten.

Hierbei kommt der Berücksichtigung kritischer Lebensereignisse besondere Bedeutung zu. Hierzu zählen insbesondere Belastungen in der Ursprungsfamilie (z.B. durch Süchte der Eltern, chronische Krankheiten, Konflikte der Eltern, Scheidung, Gewalterfahrungen, Übergriffe, materielle Not, Größe der Familie, Probleme in der Kindheit, Adoption, Heimaufenthalte etc.), aber auch sozialer Verlust-Ereignisse (z.B. durch den Tod eines nahen Verwandten, Eheprobleme, Scheidung, Wohnungswechsel, Verlassen des Elternhauses durch die Kinder, persönliche Enttäuschungen). Neue berufliche Anforderungen (berufliche Veränderungen wie Berufswechsel, Beförderung, Arbeitslosigkeit, Ruhestand) oder auch berufsbezogene Stressoren (wie Ärger mit dem Chef oder den Kollegen, Beginn einer Arbeit des Ehepartners, Beendigung des Berufs, Veränderung der Arbeitsstätte, Ausbildungsende, Prüfungen, Abbruch der Ausbildung) können krankheitsrelevant sein.

Materielle Nöte (z.B. durch Schulden, finanzielle Probleme, Verlust von Eigentum, juristische Probleme) können belastend und krankheitsaufrechterhaltend sein. Andere Stressoren können in zusätzlichen Gesundheitsproblemen bestehen (durch Verletzungen, Unfälle, Operationen, Substanzabhängigkeiten oder psychische Krankheiten), aber auch durch neue, nicht erprobte Verpflichtungen (durch Heirat, Sorge um oder Pflege eines Familienmitglieds, Geburt eines Kindes) hervorgerufen werden.

Nur eine umfassende, bio-psychosoziale Schmerzanamnese berücksichtigt in der Regel diese Vielzahl von Einflussfaktoren auf die chronische Schmerzentwicklung (vgl. Kröner-Herwig 2011 [160], Egle & Hoffmann 1993 [83]).

Die **Beurteilung etwaiger tendenziöser Haltungen** sollte nach der AWMF-Leitlinie Nr. 051 (Arbeitsgemeinschaft der Wissenschaftlichen Fachgesellschaften 2011 [11]) zwischen dem Grad der Bewusstheit der dargestellten Beeinträchtigung und unbewussten, sich der willentlichen Beeinflussung entziehenden Reaktionsanteilen unterschieden werden. Begrifflich sind dabei vor allem Auffälligkeiten der *Verdeutlichung*, der *Aggravation* und der *Simulation* zu unterscheiden (s. Abb. 4, S. 22).

- *Verdeutlichung* beschreibt die unwillkürlich-übertreibende Beschwerdedarstellung verbunden mit dem Motiv, den Untersucher oder die Behandler von der Symptomstärke zu überzeugen. Verdeutlichung ist in Behandlungs- und Begutachtungssituationen in einem gewissen Ausmaß normal, hierin kann sich der erhöhte Therapiewunsch aus-

drücken und die Absicht, den Behandler von der Dringlichkeit einer Intervention überzeugen zu wollen;

- *Aggravation* bezeichnet eine verschlimmernde Präsentation tatsächlich vorhandener Beschwerden;
- *Overreporting* wird im engl. häufig äquivalent verwendet und bezeichnet die unbewusste oder auch bewusste, übertriebene Schilderung vorhandener Beschwerden;
- *Simulation* (im engl. häufig *Malingering*, *Feigning*, oder als Verhaltensweise *Deception* (Täuschung) oder auch *fake bad* genannt) bezeichnet dann das Vortäuschen nicht vorhandener Beschwerden. In diesem Zusammenhang wird auch zuweilen von *Negative Impression Management* gesprochen, was eine Art der Selbstbeschreibung kennzeichnet, sich selbst möglichst negativ beeinträchtigt darzustellen.

In diversen anderen Diagnostik- und Behandlungs-Kontexten kann ebenso vorkommen, dass Patienten ihre Beschwerden, Einschränkungen oder Behinderungen eher bewusst gering darstellen wollen. Motive hierzu sind vielfältig und hängen meist mit befürchteten Konsequenzen der Betroffenen bei einer Feststellung von Defiziten zusammen. Solche Motive sind z.B. die Vermeidung eines Medikamentenentzugs, die Vermeidung des Führerscheinverlustes, das Umgehen einer psychiatrischen oder psychopathologischen Diagnose mit entsprechend indizierter Behandlung. Bei älteren Menschen soll oft mit dem Versuch, Beschwerden zu bagatellisieren, die Diagnose einer neurokognitiven Störung (z.B. Demenz) vermieden werden, aus Angst vor Verlust der Selbstbestimmung und vor befürchteten Folgen (z.B. Heim-Unterbringung, Vormundschaft etc.).

Zur Kennzeichnung solcher Antwortverzerrungen in überhöht positiver Weise werden ebenfalls unterschiedliche Begriffe verwendet, z.B.

- *Dissimulation* bezeichnet die willkürliche und absichtlich untertreibende und positiv verfälschende Beschwerdedarstellung; sie wird im englischen Sprachraum auch als *fake good*, manchmal als *Positive Impression Management* oder als *Defensiveness* (dt. *Beschwerde-Abwehr*) bezeichnet.

Die Trennung dieser Begriffe wird nicht immer einheitlich gehandhabt (Bianchini et al. 2005 [31]). So wird in manchen Gutachten artifiziell zwischen einer „bewussten“ und einer „unbewussten“ Aggravation unterschieden, insbesondere wenn der Gutachter eine Negativ-

Stigmatisierung der untersuchten Person vermeiden will oder u.U. negative Reaktionen der Probanden bei Kenntnisnahme solcher bewusster Verfälschungstendenzen befürchtet.

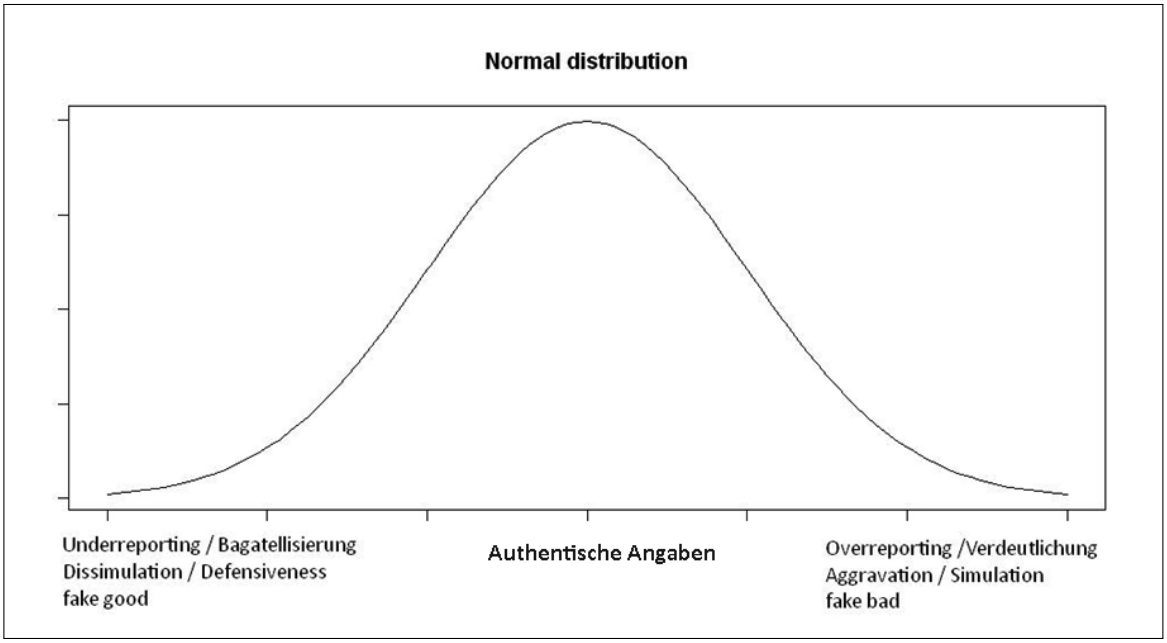


Abb. 4. Kontinuum der Authentizität von Aussagen; mod. nach Green (2008, S. 163) [115]

Als Kriterium für sog. „Aggravation“ werden häufig zwei Merkmale genannt, die jedoch nicht hinreichend operationalisiert sind, worauf Dohrenbusch (2009) [69] aufmerksam macht: das Merkmal der Bewusstseinsnähe und auch der dann fraglichen Situationsangemessenheit des Verhaltens in einer Untersuchungssituation. Mangels klarer Operationalisierung unterliegt die Beurteilung dieser Kriterien oft den Erfahrungswerten des jeweiligen Untersuchers und ist deshalb nicht frei von Gegenübertragungsreaktionen. Dohrenbusch & Pielsticker (2011, 355pp.) [73] betonen zudem, dass in sozialgerichtlichen Gutachten oft abgefragte Einschätzungen zur „zumutbaren Willensanspannung“ des Untersuchten (zur Überwindung seiner Beschwerden) oder „Bewusstseinsnähe von Motiven“ ebenfalls wegen unzureichender Operationalisierung Gutachtern erhebliche Entscheidungs-Freiheiten gewähren und künftig klarerer Definitionen bedürfen. Der in Gutachten häufiger zu findende Verweis auf die Erfahrung des Gutachters dürfte für die Validität seiner Einschätzungen hier nicht hinreichend sein.

Fishbain et al. (1999) [89] konnten in einer Metaanalyse nachweisen, dass Inkonsistenzen zwischen subjektiven Klagen und fehlendem pathologischen, anatomischen oder physiologischem Korrelat nicht zur Beschwerdenvalidierung ausreichen. Die ausführliche Exploration und Untersuchung des Patienten dient letztlich der Abwägung somatischer und physiologi-

scher Befunde, aber auch biographischer und aktueller Belastungssituationen, die zur Entstehung und Aufrechterhaltung der Beschwerden beigetragen haben oder aktuell beitragen, bei gleichzeitiger Berücksichtigung protektiver Faktoren einer adaptiven und gelungenen Stressor- und Krankheitsbewältigung. Insofern ist bei der gutachterlichen Einschätzung das *positive* und das *negative Leistungsbild* des Probanden von besonderer Bedeutung.

Bei der **Frage der Krankheitsprognose** wurden verschiedene Kriterien für einen ungünstigen Heilverlauf von unterschiedlichen Autoren genannt (s. Häuser 2002b [127]): chronisch-kontinuierlicher Verlauf der Erkrankung trotz regelmäßiger Therapie, vielfache stationäre Therapie- oder Kur-Maßnahmen, Veränderungsangst, resignative Erwartungshaltung sowie primärer, sekundärer oder tertiärer Krankheitsgewinn. Diese Faktoren sollten bei Beurteilung der Krankheitsschwere besonders beachtet werden.

Bei einer kritischen Befundwürdigung sollten ferner entsprechend der AWMF-Leitlinie Nr. 030/102 [10] der Deutschen Gesellschaft für Neurologie (DGN), der Deutschen Gesellschaft für Orthopädie und Orthopädische Chirurgie (DGOOC), der Deutschen Gesellschaft für Psychosomatische Medizin und Psychotherapie (DGPM) sowie des Deutschen Kollegiums für Psychosomatische Medizin (DKPM) und der Deutschen Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN) folgende **Diskrepanzen und Inkonsistenzen der Befunderhebung** berücksichtigt werden:

- Diskrepanz zwischen Beschwerdeschilderung (einschließlich Selbsteinschätzungen in Fragebögen) und körperlicher und/oder psychischer Beeinträchtigung bzw. beobachtbarem Verhalten in der Untersuchungssituation
- Wechselhafte und unpräzise-ausweichende Schilderung der Beschwerden und des Krankheitsverlaufs
- Diskrepanzen zwischen eigenen Angaben und fremdanamnestischen Informationen (einschließlich Aktenlage)
- Fehlende Modulierbarkeit der beklagten Schmerzen
- Diskrepanz zwischen geschilderten Funktionsbeeinträchtigungen und zu eruierenden Aktivitäten des täglichen Lebens
- Fehlen angemessener Therapiemaßnahmen und/oder Eigenaktivitäten zur Schmerzlinderung trotz umfänglicher Beschwerden

- Fehlende Möglichkeit einer sachlichen Diskussion möglicher Verweistätigkeiten bei Begutachtungen zur beruflichen Leistungsfähigkeit.

Meyer & Deitsch (1996) [200] fassten dazu bereits acht Jahre vorher zusammen, dass Simulation (Vortäuschung nicht vorhandener Beschwerden) in den folgenden vier Bedingungs-Gefügen erwogen und dann näher untersucht werden sollte:

1. Juristische Auseinandersetzungen, bei denen die untersuchte Person durch einen Rechtsanwalt vertreten wird;
2. Bedeutsame Inkonsistenzen zwischen dem angegebenen Ausmaß an Stress oder Beeinträchtigung und den objektiven Befunden;
3. Mängel in der Kooperation während der Diagnostik oder Incompliance hinsichtlich der verschriebenen Medikation;
4. Das Vorliegen einer Antisozialen Persönlichkeitsstörung.

Die beiden zuletzt genannten Kriterien sind bei näherer Betrachtung durchaus kritisch zu diskutieren. Zunehmend häufiger wird in sozialgerichtlichen Auseinandersetzungen um die Zuerkennung einer Minderung der Erwerbsfähigkeit von Patienten-/Klägerseite aus argumentiert, die Schmerzbeschwerden seien so stark, „dass jetzt sogar eine Opiat-Medikation erforderlich sei“, die aber auch nur wenig helfe. Hierzu ist festzustellen, dass zwar medikamentöse Incompliance ein mögliches Indiz für Beschwerdeüberhöhung ist. Umgekehrt jedoch stellt Medikamentencompliance kein sicheres Indiz für authentische Beschwerdendarstellungen dar.

Zum zweiten wird die Diagnose einer dissozialen Persönlichkeits-Störung (ICD-10 F60.2) im klinischen Kontext chronischer Schmerzen, insbesondere bei bislang in Arbeitsverhältnissen Beschäftigten, die die Zuerkennung einer Rente wegen Erwerbsminderung beantragen, sehr selten gestellt. Dissoziale Störungen sind eher in Komorbidität mit Sucht- und -Abhängigkeits- Erkrankungen oder anderen psychiatrischen Störungen zu finden (z.B. Borderline-Persönlichkeits-Störungen). Die bei Meyer & Deitsch (1996) [200] genannten Kriterien sind damit weder vollständig noch hinreichend typisch.

Dohrenbusch (2009) [69] ergänzt weitere **Indizien für nicht-authentisches Verhalten**:

- Die Klagen des Untersuchten wirken demonstrativ und übertrieben verzerrt;
- Inhaltliche Festlegungen werden vom Untersuchten gemieden;

- Leugnung jeglichen bisherigen Behandlungserfolges;
- Hinweise auf Behandlungs-Incompliance.

Hierbei wird deutlich, wie schwierig teilweise die Operationalisierung einzelner Kriterien ist (z.B. das Vermeiden inhaltlicher Festlegungen). Möglicherweise sind deshalb einige Prüfstrategien bislang nur in Interviews (z.B. Structured Interview of Reported Symptoms (SIRS), Rogers et al. 1992 [237], s. folgendes Kap. 1.7.3, S. 44) realisiert.

Der Abgleich von Angaben zum körperlichen Funktionsniveau gegenüber spezifischen Körpermerkmalen, z.B. Beschwielung der Hände trotz Angaben weitgehenden Rückzugs von körperlichen Aktivitäten, oder dem Zustand der Muskulatur (atrophisch oder trainiert) in der ärztlichen Untersuchung kann ebenfalls Hinweise auf Inkonsistenzen ergeben. Zur Validitätsprüfung gehört ebenfalls die Überprüfung therapeutischer Routinetätigkeiten, also zum Beispiel die Überprüfung der Durchführung angegebener regelmäßiger krankengymnastischer Übungen. Dohrenbusch (2009) [69] empfiehlt zudem in der Praxis, zunächst differenzierte schriftliche Beschreibungen des körperlichen Funktionsniveaus abzufragen, um sie später in der Exploration mit Merkmalen des positiven und negativen Leistungsbildes erneut zu explorieren und dadurch intraindividuelle Abweichungen überprüfen zu können.

Ein Problem bei der Beurteilung von **Inkonsistenzen unterschiedlicher Datenquellen** (z.B. zwischen Fremd- und Eigen-Anamnesen) ist natürlich, ob unterschiedliche Einschätzungen auf gleicher Wissensbasis bestehen (beispielsweise muss der Ehepartner nicht korrekt über alle Freizeit-Aktivitäten des Patienten informiert sein). Deshalb sind Inkonsistenzen dann als höher valide zu beurteilen, wenn der Proband sich selbst inkonsistent zu seinen Aktivitäten äußert oder inkonsistent verhält (Dohrenbusch 2009 [69]). Interessant wäre auch ein Inkonsistenzabgleich unterschiedlicher Fragebögen, beispielsweise zum körperlichen Funktionsniveau, so z.B. des Pain Disability Index, der eine bekannte und reliable Korrelation von 0,75 zum Funktionsfragebogen Hannover (FFBH-R) aufweist. Dabei könnten - unter Einbezug der Reliabiliätswerte der Fragebögen - über das Ausmaß zufälliger interindividueller Abweichungen hinausgehende Angaben Hinweise auf überhöhte Beschwerdeangaben zeigen.

Auch nach der jüngsten AWMF-Leitlinie (AWMF 2011, Leitlinien-Reg. Nr. 051/029 [11]) stellt die „Weichheit“ oder „Unschärfe“ der zu bewertenden Variablen und Befunde bei der Begutachtung chronischer Schmerzen eine besondere Anforderung für den Gutachter dar. In

dieser Verfahrensleitlinie wird vorgeschlagen, die maximal erreichbare Leistungsfähigkeit einer Person entsprechend der **Kriterien der International Classification of Functioning (ICF)** im Arbeitsplatz-Kontext, also dem beruflichen Anforderungsprofil, zu beurteilen oder aber im Fall von Beurteilungen zur Erwerbsminderung die mögliche Leistungsfähigkeit auf dem Allgemeinen Arbeitsmarkt einzuschätzen.

Grundsätzlich sollte entsprechend der Leitlinie die Validierung der beklagten Beschwerden auf eine „möglichst breite methodische Basis“ gestellt werden. So sollten neben der Exploration auch Verhaltensbeobachtungen zur Beurteilung herangezogen werden, aber auch standardisierte Fragebögen, Fragebogenkontrollskalen oder Fragebögen zu Antworttendenzen, körperliche Funktions- und Leistungstests, psychologische Funktions- und Leistungstests, Symptomvalidierungstests, Labortests, wie z.B. die Kontrolle des Blutserumspiegels zur Überprüfung der medikamentösen Compliance oder anderer Funktionsauffälligkeiten (z.B. Leber- oder Rheuma-Indices).

Als **Beurteilungsbereiche der psychischen Funktionstüchtigkeit** werden die folgenden Bereiche vorgeschlagen: Somatisierung, Emotionalität, Antriebs- bzw. Psychomotorik, Kognition, Psychotisches Erleben, Bewusstseins- und Orientierungsstörungen, Verhaltensauffälligkeiten, zwischenmenschliche Interaktion und körperliche Funktionsbeeinträchtigungen, sowie die Gesamtbeeinträchtigung aus allen vorgenannten Bereichen. Hinsichtlich der meisten dieser Leistungsbereiche sind Teil-Leistungskompetenzen definierbar, die aus dem Mini-ICF-APP (Linden et. al. 2009 [172]) abgeleitet wurden. Zum Beispiel umfasst der Bereich Aktivität auch Teilleistungen wie Anpassung an Regeln und Routinen, Flexibilität und Umstellungsfähigkeit, Entscheidungs- und Urteilsfähigkeit oder Selbstbehauptungs- und Durchsetzungsfähigkeit oder Gruppenfähigkeit, die bei der Beurteilung der beruflichen Rehabilitations-Fähigkeit von entscheidender Bedeutung sein können.

Adaptation von Methoden der Aussagepsychologie:
Zur Erfassung von Overreporting, aber auch von Underreporting können möglicherweise ähnliche Konstrukte angewandt werden, wie bei **Glaubwürdigkeits-Beurteilungen von Zeugen** (meist gleichzeitig Opfern) von Straftatbeständen.

In diesen Begutachtungen sind beispielsweise die Glaubwürdigkeit von Aussagen unterstützende Aussage-Merkmale: (a) geschilderte Schuld- oder Scham-Empfindungen seitens der Opfer von Gewalttaten, oder auch (b) Selbstanklagen, die die eigene Tatbeteiligung bein-

halten, wie sie beispielsweise bei depressiven Patienten charakteristisch sind, oder aber (c) Entschuldigungen hinsichtlich der Verhaltensweisen und Motive des mutmaßlichen Täters (im Sinne des psychoanalytischen Konstruktes der „Identifikation mit dem Aggressor“).

Bagatellisiert somit ein Patient mit chronischen Schmerzen angesprochene psychische Reaktionen und Mitbeteiligungen eher, dann entspricht dies der üblicherweise zu erwartenden Reaktion, da solche psychischen Faktoren in der Regel als stigmatisierend aufgefasst werden und die Glaubwürdigkeit der somatischen Aspekte der Schmerzen in Zweifel ziehen. Sog. Haltungen der *Defensiveness* und *Underreporting* entsprechen hinsichtlich psychischer und psychiatrischer Symptome also eher der Normalität. Damit erhöhen diese Antwortmuster eher die Glaubwürdigkeit tatsächlich feststellbarer Psychopathologie (vgl. Green 2008 [115]).

Aus der Aussagepsychologie können aber auch andere typische Glaubwürdigkeits-Merkmale für die Begutachtung chronischer Schmerzen und assoziierter Symptome adaptiert werden, wie beispielsweise Detailtreue, Widerspruchs-Freiheit und Erlebnisnähe. Diese Glaubwürdigkeitskriterien bieten auch für testpsychologische Befragungen übertragbare Maßstäbe für die Konsistenz von Probandenangaben.

Die Anwendung allerdings der sehr aufwendigen, aus dem Strafrecht stammenden, aussagepsychologischen Methoden mit Transkription wortwörtlicher Aussagen ist jedoch nicht generell auf sozialrechtliche Fragestellungen und Begutachtungen übertragbar, da weder Beweismaß noch rechtliche Konsequenzen der unterschiedlichen Rechtsverfahren vergleichbar sind.

Einsatz psychodiagnostischer Testverfahren und Einbezug psychologischer Gutachter:

Leider wurde in früheren Leitlinien zur Begutachtung von Schmerzen (vgl. Widder et al. 2008 [314]) pauschal der Verwendung von psychologischen Fragebögen und Selbsteinschätzungsskalen eine nur geringe Validität zugeordnet. Diese Beurteilung von Testverfahren blieb als Allgemeinplatz begründeter Maßen nicht unwidersprochen (Dohrenbusch et al. 2008 [72]). Kritisiert wurde zudem die in o.g. Formulierungen postulierte „Vorrangstellung ärztlicher Gutachter“ als eher berufsständische, denn als hinreichend begründete Aussage. Insbesondere fachpsychologische, natürlich auch erfahrene psychiatrische und psychosomatische Fachgutachter weisen hinsichtlich testdiagnostischer Gütebeurteilungen, aber auch in

der Erhebung und Einschätzung biographisch relevanter Fakten eine besondere Qualifikation auf. Diese ist natürlich nicht an die Qualifikation als Arzt gebunden. Die bei Widder et al. (2008) [314] proklamierte Betonung der Bedeutung des ärztlichen Gutachters wird von Dohrenbusch et al. (2008) [72] als nicht zeitgemäße Beurteilung eingeordnet, aus der zudem keine bindende Aussage für die Auftraggeber von Begutachtungsleistungen abgeleitet werden sollten.

Dohrenbusch (2009) [69] verweist ferner darauf, dass besondere Vorteile einer Objektivierung von Begutachtungs-Befunden darin bestehen können, **testtheoretische Gütekriterien** an die Untersuchung zur Beschwerdenvalidierung anzulegen; entsprechende Verfahren sollten normiert, objektiv, reliabel (zuverlässig) und valide sein, aber auch nützlich, dem Probanden zumutbar mit dem Anspruch der Unverfälschbarkeit. Sie sollten zudem praktikabel und ökonomisch durchführbar sein, also dem gängigen Maßstab jener strengen Gütekriterien genügen, die insbesondere psychodiagnostische Testverfahren aufweisen müssen. Zusätzlich sollte jede valide und reliable Exploration von Beschwerden auch entsprechend diesem Reliabilitätsprinzip Kontrollfragen und Wiederholungen beinhalten, was in den meisten Gutachten eher nicht systematisch realisiert wird (vgl. Wagner, Richter et al. 2003 [309]).

Bei der Normierung von Testverfahren ist wichtig, dass sie den Bezug zu Normen der Allgemeinbevölkerung und damit eine Einschätzung ermöglichen, wie die Leistungsfähigkeit der zu begutachtenden Person im Vergleich zu nicht-kranken, normal-leistungsfähigen Personen einzuschätzen ist. Dahingehend unterscheiden sich die Validitätsanforderungen von Testverfahren in einem Therapiekontext nicht von jenen einer Begutachtungssituation (s. Kommentar zur Leitlinie nach Widder et al. 2008 [314] von Dohrenbusch et al. 2008 [72]). Dohrenbusch (2009) [69] macht des Weiteren darauf aufmerksam, bei der Beurteilung somatischer und psychischer Symptome einen Abgleich der individuellen Daten des untersuchten Probanden in Bezug zu Daten seiner entsprechenden Altersnorm vorzunehmen, so dass Störungen nicht nur die Selbstwahrnehmung des Probanden widerspiegeln, sondern diese zur alterstypischen Ausprägung des untersuchten Merkmals in Beziehung gesetzt werden können. Eine solche Altersnormierung ermöglichen neben körperlichen Funktionseinschätzungen wiederum insbesondere testpsychologische Untersuchungen.

Dohrenbusch (2009) [69] schlägt ferner für fraglich gültige, auf einem konkreten Ausgangsverdacht beruhende Beschwerdeangaben ein gestuftes Vorgehen vor, um die Maßnahmen zur Beschwerdenvalidierung zu begrenzen. Dabei sollten zunächst einfachere Screening-

Methoden Anfangsverdachtsmomente einer Beschwerdeüberhöhung überprüfen, die bei positivem Hinweis durch nachfolgende, aufwendigere und differenzierte Verfahren kreuzvalidiert werden sollten.

Eine umfassende, vorurteilsfreie biopsychosoziale Schmerzanamnese unter Einbezug positiver wie negativer Krankheitsfolgen sind notwendige, nicht hinreichende Bedingungen einer angemessenen Begutachtung. Alleinige Inkonsistenzen zwischen subjektiven Klagen von Patienten und pathologischen anatomischen oder physiologischen Befunden sind nicht zur Beschwerdenvalidierung geeignet.
Zusätzlich muss die Glaubhaftigkeit der Beschwerden an operationalisierbaren Kriterien hinsichtlich ihrer Konsistenz inter- und intraindividuell, altersnormiert und entsprechend der gängigen Testgütekriterien (Objektivität, Validität, Reliabilität) überprüft werden. Modulierbarkeit der Beschwerden aus Sicht des Untersuchten sowie seiner der sozialen Anpassung (insbesondere der beruflichen Reintegration) ist dabei besondere Aufmerksamkeit zu widmen.

Moderne Leitlinien empfehlen eine multimodale Methodik, insbesondere auch einen Einbezug geeigneter testdiagnostischer Verfahren bei der Begutachtung, die im Folgenden ausführlich dargestellt werden.

1.7 Moderne Ansätze der Beschwerdenvalidierung

Bianchini et al. (2005) [31] formulierten erstmals eine multidimensionale Klassifikationsmethode, um die Genauigkeit von Validitätsindikatoren zu überprüfen. Hierzu sollten zwei oder mehr externale Methoden hoher Güte eingesetzt werden, um anhand inkonsistenter Befunde auf der Verhaltensebene, fragwürdiger kognitiver Leistungen und unglaubwürdiger Angaben psychischer und somatischer Symptome Overreporting zu identifizieren. Die Autoren plädieren dabei für die Unterscheidung möglicher (*possible*), wahrscheinlicher (*probable*) und sicherer (*definite*) Belege für eine sog. *MPRD (Malingered Pain Related Disabilities)*.

Um eine MPRD zu konstatieren, muss das Ausmaß der präsentierten Einschränkungen über die in jeder klinischen Untersuchung typische leichte Verdeutlichung der sog. *Pain Related Disabilities (PRD)* hinausgehen und darf nicht allein durch psychiatrische, neurologische oder psychische Bedingungen, wie unbewussten psychischen Reaktionen auf die chronische Symptomatik (z.B. erhöhte Depressivität, Angst oder Demotivation), erklärbar sein. Insbesondere bei Patienten mit anhaltender somatoformer Schmerzstörung ist dieser charakteristische Mechanismus einer Konversion oder Somatisierung unbewusster psychischer Konflikte in Körpersymptome zu berücksichtigen und darf nicht als Malingering-Indikator missinterpretiert werden (vgl. Bianchini et al. 2008 [30]).

Zur Konstatierung einer MPRD müssen nach Bianchini et al. (2005) [31] zumindest teilweise bewusste, externale Motive des Krankheitsgewinns für eine Beschwerdeüberhöhung bei den untersuchten Patienten vorhanden sein. Dies können finanziell-existenzielle Vorteile durch Erlangen eines Krankengeldes, einer Erwerbsminderungsrente oder einer fortgesetzten Krankschreibung sein oder im Fall medikamentöser Abhängigkeiten die weitere Verschreibung entsprechender Substanzen (Opiod-Analgetika, Benzodiazepine, zentralwirksame Relaxantien).

Im Fall wesentlicher externaler Motive kann ein Symptom-Overreporting auf folgende Weise operationalisiert werden:

- als intraindividuelle Inkonsistenzen innerhalb einer und zwischen verschiedenen Validierungsmethoden,
- als interindividuelle Diskrepanzen zwischen den erhobenen Befunden und den bei verschiedenen Erkrankungen erwartbaren Einschränkungen,
- als inkonsistente Befunde auf der Verhaltensebene (z.B. bei offen durchschaubaren und verdeckten Verhaltensbeobachtungen).

Die Wahrscheinlichkeit für Overreporting und die Klassifikation einer MPRD erhöhen sich dabei mit der Anzahl nicht-authentischer, inkonsistenter Befunde.

Greve & Bianchini (2004) [117] konstatieren, dass externe BV-Verfahren zur Klassifikation von Malingering speziellen Gütekriterien genügen sollten: (a) die Verfahren sollten hinreichend theoretisch begründet und hinreichend erprobt sein, (b) sie sollten empirisch überprüft sein, (c) es sollte eine potentielle Fehlerrate bekannt sein und (d) die generelle Methodik des jeweiligen Verfahrens sollte wissenschaftlich akzeptiert sein. Sensitivität bezeichnet dabei

die Trefferrate der mit einem speziellen BV-Verfahren richtig-positiv identifizierten Probanden (Anzahl der positiv getesteten Probanden dieses Merkmals geteilt durch alle Probanden mit diesem Merkmal). Spezifität ist hingegen die Richtig-Negative-Trefferrate (Anzahl der negativ getesteten Probanden ohne dieses Merkmal, geteilt durch alle Probanden ohne Merkmal). Geringe Sensitivität tritt somit auf, wenn ein BV-Verfahren viele tatsächliche Malingerer nicht identifiziert, also eine hohe Anzahl Falsch-Negativ-Klassifizierter aufweist. Geringe Spezifität tritt auf, wenn ein BV-Verfahren viele tatsächliche Non-Malingerer falsch positiv klassifiziert.

Als Positive Prädiktive Power (+PP), also die tatsächliche Vorhersagekraft eines Diagnostik-Verfahrens, bezeichnen Greve und Bianchini (2004) [117] die Anzahl tatsächlich positiver Probanden, geteilt durch die Summe der tatsächlichen und der falsch positiv Getesteten. Als Negative Prädiktive Power (-PP) wird hingegen die Anzahl der tatsächlich negativen Probanden bezeichnet, geteilt durch die Summe der tatsächlichen Negativen und der falsch positiv Getesteten. Die Positive Prädiktive Power (+PP) bezieht sich damit auf die Spezifität eines BV-Verfahrens: das heißt, wenn Probanden ohne das Merkmal keinerlei BV-Auffälligkeit aufweisen, kann man sicher sein, das ein positives Testresultat das tatsächliche Vorhandensein des Merkmals widerspiegelt. Die Negative Prädiktive Power (-PP) hingegen bezieht sich auf die Sensitivität und zeigt an, in welchem Maß ein Proband ohne das Merkmal sicher ein negatives Testergebnis aufweist. Zusammenfassend sind Sensitivität und Spezifität Indikatoren für die Validität eines BV-Verfahrens und die Prädiktive-Power eine Art Sicherheits-Index, der angibt, in wieweit die Klassifikation eines Probanden korrekt erfolgte. Die Sensitivität eines BV-Verfahrens kann nach Greve und Bianchini (2004) [117] durch multiple Detektionstechniken, die immun gegen unterschiedliche Strategien des Malingering sein sollten, erhöht werden (vgl. beim Aggravations-Simulations-Test AST, Eberl und Wilhelm 2007 [79] mit Integration zweier unterschiedlicher Detektionsindikatoren).

Malingering-Indikatoren sollten hinreichend sensitiv für die Aufdeckung von Overreporting sein (mindestens 50 % der Fälle identifizieren), jedoch ist eine möglichst hohe bis maximale Spezifität ($\geq$ 90 %) und eine hohe Positive Prädiktive Power (+PP $\geq$ 50 %) wichtig, um falsch positive Klassifikationen zu vermeiden.

Kool et al. (2008) [156] weisen zudem auf die Gefahr der „Prosecutor's fallacy“ („Täuschung des Klägers“) hin, die zum einen im Risiko einer Falsch-Beurteilung durch Nicht-Beachten einer geringen Prävalenz von Malingering in einer Subgruppe besteht (sog. niedri-

ge „Basisrate“), sowie der Problematik des Einsatzes multipler Tests, die nach Möglichkeit nicht eng miteinander korreliert sein sollten. In beiden Fällen können BV-Verfahren trotz hinreichend hoher ermittelter Gütekriterien zu einer zu hoch eingeschätzten Zahl falsch positiv klassifizierter Personen führen.

Rosenfeld et al. (2000) [250] machen in ihrer theoretischen Analyse darauf aufmerksam, dass auch die Häufigkeit von Malingering in einer Stichprobe (die sog. „Basisraten“) die Diskriminationsgüte von Validitätsskalen beeinflussen können. So stellten Mittenberg et al. (1993) [207] in ihrer Studie mittels der Wechsler-Memory-Revised-Scales in einem forensischen Patientensample Basisraten von Malingering kognitiver Funktionabilität von 50 bis 30 Prozent fest. Bei einem Cutoff von mehr als 34 Punkten des Basis-Differenz-Scores der Wechsler-Scales (General Memory Index [GMI] minus Attention-Concentration Index [ACI]) führte diese Basisrate zu hohen Werten der Sensitivität (,77), der Spezifität (,90), aber auch der PPP (,88) und der NPP (,88). Wäre die Basisrate geringer, fielen diese Werte anders aus.

Rogers et al. (1994, 1998, 2003) [247] [241] [242] betonen, dass in klinischen Patientengruppen das Vorkommen von Malingering nur in 15-17 Prozent der Fälle vorkommt. Rosenfeld et al. (2000) [250] zeigten, dass sich schon bei einer solchen Basisrate von Malingering in einem Sample von 1000 Patienten trotz hoher Sensitivitäts- und Spezifitätswerte die Positive Predictive Power auf 0,57 reduziert, bei noch niedrigerer Basisrate von 10 % der Fälle auf 0,46 sinkt und bei einer nur einprozentigen Basisrate die Positive Predictive Power nur noch 0,07 beträgt. Das würde bereits bei einer 15-prozentigen Basisrate bedeuten, dass 43 Prozent der Patienten fälschlicherweise als Non-Malingering klassifiziert würden, die jedoch kein authentisches Antwortverhalten zeigten.

Eine solche hohe Rate falsch negativ klassifizierter Probanden ist für einen Diskriminanz-Score inakzeptabel. Sellbom & Bagby (2008, 185pp. [268]) weisen darauf hin, dass eine niedrige Basisrate in Stichproben mit authentischen Patienten in klinischen Stichproben sehr häufig vorkommen (z.B. 10,7 % bei Blanchard et al. 2003 [33]). In manchen Simulationsstudien seien ebenfalls solche niedrigen Basisraten beobachtet worden.

Insofern ist es notwendig, in Untersuchungen zur Diskriminationsgüte von Validitätsscores und -indizes, nicht nur Resultate der Sensitivität und Spezifität zu berücksichtigen, son-

dern insbesondere die Basisrate des untersuchten Merkmals, als auch die Positive Predictive Power in die Beurteilung einzubeziehen.

Grundproblem der genannten Gütekriterien für BV-Verfahren ist, dass das Bemühen um größtmögliche Sensitivität stets die Anzahl falsch-positiv klassifizierter Probanden erhöht und damit die Spezifität des Verfahrens reduziert. Bei einer hohen Anzahl falsch-positiv klassifizierter Personen belastet dies vor allem den individuellen Probanden, da Verfahren der Beschwerdenvalidierung bei niedrigem Cutoff hoch sensitiv, aber wenig spezifisch sind. Bei einer hohen Anzahl falsch-negativ Klassifizierter hingegen (bei relativ hohem Cutoff des Verfahrens, hoch spezifisch, gering sensitiv) würden die gesamtgesellschaftlichen Ressourcen verstärkt belastet. Zur Festlegung valider Cutoffs sollte deshalb zunächst der Score-Range von Non-Malingering-Probanden erhoben werden (Greve & Bianchini 2004 [117]).

Resümierend plädieren Greve und Bianchini (2004) [117] in einer Pro-Individuums-Position dafür, bei jedem BV-Verfahren primär eine maximale, optimale Spezifität (minimale Rate falsch-positiv Klassifizierter) anzustreben. Dies erfordert die Festlegung eines relativ hohen Cutoffs eines BV-Verfahrens, um größtmögliche Validität zu erreichen. Die Positive Prädiktive Power (+PP) sollte anhand einer Basisrate von ca. 0,30 bestimmt werden, wie sie der Häufigkeit von Malingering in Gutachten-Kontexten entspricht. Auf diese Weise kann die Detektions-Akkuranz eines BV-Verfahrens adäquat beurteilt werden.

Diverse PVTs überprüfen die Glaubwürdigkeit kognitiver Leistungen mittels forced-choice-Testverfahren. Da Patienten mit chronischen Schmerzen häufig über Aufmerksamkeits-, Konzentrations- und Gedächtnisdefizite klagen, kann deren Glaubwürdigkeit mittels Performance-Validierungstests geprüft werden.

Im Folgenden werden Validierungs-Verfahren entsprechend dem multidimensionalen Klassifikationsmodell auf 3 Ebenen vorgestellt: 1. der kognitiven Symptom-Ebene, 2. der Ebene verhaltens-bezogener somatischer Symptome und 3. psychopathologischer Symptome.

1.7.1 Validierung kognitiver Leistungsdefizite

Methoden der Beschwerdenvalidierung (kurz: BV oder im anglo-amerikanischen Sprachraum als Leistungstests sog. *PVTs, Performance-Validation-Tests*) wurden in jüngster Zeit aus dem Bereich der Neuropsychologie auch zur Validierung klinischer Psychopathologie erprobt und genutzt (Bianchini et al. 2005 [31], Dohrenbusch 2009 [69]). Ein Malingering kognitiver Leistungen wird dabei nach dem Konzept der submaximalen Anstrengung (*submaximal effort*) oder vermindert präsentierter Leistungskapazität erfasst. Wenn ein Patient in einer Zweifach-Wahlaufgabe (*forced two-choice recognition task*), in der die Abfragen in schneller Abfolge erfolgen (um wenig bewussten Entscheidungs-Spielraum zu ermöglichen) eine Wiedererkennungs-Leistung unter Zufalls-Wahrscheinlichkeit ($\leq$ 50 %) zeigt, obwohl der Test sehr einfach zu lösen ist und einen hohen Decken-Effekt (*ceiling effect*) aufweist, kann dies als Hinweis für eine bewusste Aggravation gewertet werden.

Beispiele solcher Methoden der Beschwerdenvalidierung sind der Amsterdamer Kurzzeitgedächtnistest AKGT (Schmand & Lindeboom 2005 [257]), der Test of Memory Malingering TOMM (Tombaugh 1996 [298]) oder - als am besten evaluiertes Verfahren - der Word Memory Test WMT (Green 2003 [111]) oder der (als verbale und non-verbale Version publizierte) Medical Symptom Validity Test (MSVT, Green 2004 [112]). Eine dem WMT ähnliche, nicht kommerzielle Variante stellten Barbeau et al. (2004) [23] mit dem DMS48 (Description of the visual recognition Memory Task) vor. Eine andere Variante eines *Performance-Validation-Tests* stellt der *Aufmerksamkeits- und Belastungstest d2* (Brickenkamp 1962 [39]) dar.

Mit dem Amsterdamer Kurzzeitgedächtnistest AKGT (Schmand & Lindeboom 2005 [257]) werden dem Probanden in zwei Durchgängen 30 zu memorierende Sets von je 5 Wörtern der gleichen Kategorie präsentiert, von denen drei nach kurzer Ablenkung aus einem weiteren Set mit 5 Wörtern wiedererkannt werden sollen. Der Proband kann maximal 90 Punkte (3 x 30) erlangen, wobei selbst Patienten mit kognitiven Störungen nach Schädelhirntrauma und Patienten mit Amnestischem Syndrom im Durchschnitt $\geq$ 87 Punkte erreichen. Simulierende Probanden zeigten hingegen meist weniger als 85 Punkte. Der AKGT zeigte in jüngeren Studien überzeugende Trennschärfe-Merkmale (z.B. Merten et al. 2005 [199]). Bei hirnorganisch schwer eingeschränkten Patienten (z.B. mit Alzheimer Demenz oder Korsakoff-Syndrom) wurden hingegen fast die Hälfte (46 %) falsch positiv eingeschätzt

(Merten et al. 2007a [196]). In einer Gruppe vorab gecoachter, mnestische Defizte simulierender Probanden identifizierte der AKGT nur ca. 70 % der Malingerer (Jelicic et al. 2007 [143]), während 90 % der nicht instruierten Personen detektierbar waren.

Mit dem Word Memory Test WMT (Green et al. 1999 [114]) wird ebenfalls in einer Lern- und Wiederkennungsphase von 15 Wörtern die Gedächtnisleistung getestet, im Anschluss erfolgt nach einer Lernphase die Erinnerung frei. Absichtlich verschlechterte Leistungen zeigen sich gewöhnlich in beiden Testteilen; der WMT ist zudem nach Brockhaus & Merten (2004) [41] wenig durch Vorinstruktionen beeinflussbar.

Auch mittels des Test of Memory Malingering TOMM (Tombaugh 1996 [298]), der auch als Kurzform (sog. Medical Symptom Validity Test MSVT, Green 2004 [112]) durchgeführt werden kann, wird die Memorierung von 50 Bildkarten nach einer Distraktionsaufgabe untersucht. In Validitätsstudien erwies sich die Testleistung von simulierenden Personen gegenüber nicht sehr schwer neuropsychologisch eingeschränkten Patienten als auffällig verringert.

Eine spezielle Differenzierung mnestischer Defizite von nur angegebenen Erinnerungs-Einbußen ermöglicht der Word Completition Memory Test (WCMT, Hilsabeck et al. 2001 [136]), bei dem die ersten drei Buchstaben der zu erlernenden Zielwörter in einer Wiederkennungsphase vorgegeben werden.

Seitens der Testautoren der meisten bekannteren Verfahren zur Beschwerdenvalidierung wird besonders darauf geachtet, deren Weiterverbreitung durch Medien, insbesondere das Internet, einzuschränken, um die Bekanntheit gering und ihre Validität aufrecht zu erhalten (Bauer & McCaffrey 2005 [24]). Allen und Green (2001) [7] vermerkten bereits 1997 über einen Sechsjahreszeitraum eine abnehmende Sensitivität eines speziellen BV-Verfahrens (Computerized Assessment of Response Bias CARB von Allen et al. (1997 [6]) und diskutierten, inwieweit Coaching von Probanden durch Anwälte in sozialmedizinischen Verfahren hierfür ursächlich sein könnte.

Als differenziertes und computerisiert auswertbares kognitives Symptom-Validierungs-Verfahren entwickelten Eberl und Wilhelm (2007) [79] im deutschen Sprachraum den Aggravations-Simulationstest (AST). Übereinstimmendes Merkmal dieses und der meisten BV-Verfahren, die in der Neuropsychologie zur Prüfung simulierter anamnestischer Symptome verwendet werden, ist die Testung mittels sog. Alternativwahlaufgaben (forced choice-

Paradigma oder Zwangswahl-Prinzip), bei denen gelernte Einzelstimuli in einer zweiten Testphase mit hohem Testdeckeneffekt und nur vorgetäuschter Schwierigkeit aus zwei Antwortalternativen wiedererkannt werden sollen. Liegt die Trefferquote unter Zufallsniveau bzw. in einem für Patienten mit schweren neurokognitiven Defiziten bekannten Ergebnisbereich kann mit großer Sicherheit auf eine negative Antwortverzerrung geschlossen werden. Patienten mit schweren neurokognitiven Störungen zeigen meist sowohl in Leistungstests, als auch in zur Beschwerdenvalidierung eingesetzten Verfahren erhebliche Defizite, unabhängig von einem äußeren Kompensationsmotiv oder -anlass (Merten et al. 2007a [196]). Insofern ist diese Patientengruppe zwar als Vergleichsgruppe interessant, eine valide Untersuchung von Malingering ist jedoch nur mit speziellen BV-Verfahren möglich.

Auch im Testverlauf inkonsistente oder untypische Leistungsprofile bestätigen bei Verfahren der Beschwerdenvalidierung eine Tendenz zu sog. negativen Antwort- oder *Effort-Bias* (reduzierte Testmotivation bzw. Anstrengung). Eberl et al. (2008) [78] konnten 62 instruierte Simulanten mit vorgetäuschter Amnesie mittels des AST bei einem Cutoff von ≥ 90 % korrekt wiedererkannter Items maximal (100 %) sensitiv und spezifisch von 39 hirnorganisch beeinträchtigten Patienten, von 40 depressiven Patienten und 52 Normprobanden unterscheiden. Die ermittelten, sehr hohen Gütemerkmale sind jedoch nicht unabhängig von der Vergleichsgruppe experimenteller Simulanten und dem Verzicht auf externe Validierungs-Indikatoren in den Patientengruppen zu sehen.

Differenzierend erwiesen sich speziell die Prozentzahl richtiger Antworten (≤ 50 % eindeutiges negatives Antwortverhalten, ≤ 90 % wahrscheinlich negative Antworttendenz) sowie der Median der Reaktionszeiten (Simulanten reagieren deutlich langsamer als gesunde Probanden und beide Patientengruppen). Da insbesondere die Reaktionsgeschwindigkeit (leichte Verzögerungen der Einzelantwort) vom Probanden kaum kontrollierbar ist, aber indirekt während der Testdurchführung computerisiert erfasst werden kann, stellt der AST ein besonders sensibles Instrument zur Aufdeckung negativer Antwortverzerrungen dar. Ähnliche Resultate zu den Reaktionszeiten berichteten auch Dunn et al. (2001) [76].

Vorgeschlagen wurden zur Überprüfung von angegebenen Einschränkungen der Aufmerksamkeit und Konzentration auch psychologische Testverfahren wie der *Aufmerksamkeits- und Belastungstest d2* (Brickenkamp 1962 [39]) oder vergleichbare kognitive Tests für die Aufmerksamkeit (Frankfurter Aufmerksamkeitsinventar FAIR [217]).

Der Test d2 oder seine Revisions-Version d2-R (Brickenkamp et al. 2010 [40]) mit gekürzter mündlicher Instruktion (auch in türkischer Sprache) und einer zusätzlichen Übungsaufgabe misst dabei sowohl die Konzentrationsleistung, das Bearbeitungstempo als auch die Genauigkeit von Probanden bei Auswahlaufgaben.

Dies ermöglicht, sowohl neben der Einordnung der individuellen Konzentrationsleistung in Bezug zu einer altersspezifischen Normgruppe die Erfassung auffälliger Verlangsamungen, aber auch überhöht schneller, flüchtiger Testbearbeitungen. Ferner ist ein Assessment von Übungseffekten oder abnehmender Genauigkeit im Testverlauf (durch Erhebung einer dreistufigen Fehlerverteilung) beurteilbar, aber auch die Einschätzung der Schwankungsbreite der Leistung (des *range*). Auch ermöglicht der Vergleich einer Wiederholungstestung beispielsweise an zwei Untersuchungstagen eine Aussage über die Zuverlässigkeit (Reliabilität) der erhobenen Leistung (s. Empfehlungen von Dohrenbusch 2009 [69]).

Mittels des d2-Tests können folgende Defizite hinsichtlich Aufmerksamkeit und Konzentrationsfähigkeit erfasst werden: Flüchtigkeit oder Verlangsamung, Schwankungen der Aufmerksamkeit, Übungs- und / oder Ermüdungseffekte, Fehler unterschiedlicher Art (sog. falscher Alarm oder sog. Verpasser) sowie die Fehlerrate.

Problem dabei ist, dass Probanden das Ergebnis wie bei jedem Leistungstest durch gezielte Testvorbereitung (sog. Coaching mit Kenntnis der Testmethode) möglicherweise verbessern können, aber auch insbesondere durch bewusst langsames Arbeiten oder absichtliche Fehler verfälschen können.

Widersprüchlich werden deshalb die Ergebnisse des d2 zur Erhebung von Aggravation bzw. sogar Simulation und Dissimulation diskutiert (vgl. Merten et al. 2007c [195] und Schmidt-Atzert et al. 2004 [259]). Letztere Autoren teilten 94 Studierende in drei Gruppen, die instruiert wurden, die Testbearbeitung entweder im Sinne eines *fake good* verbessernd zu verfälschen, im Sinne eines *fake bad* eine schlechte Leistungsfähigkeit zu simulieren oder ihre authentische Leistung zu zeigen.

Probanden der *fake bad* - Gruppe bearbeiteten den Test gegenüber den Kontrollpersonen deutlich langsamer und mit signifikant höherer Fehlerzahl. Die *fake good* - Gruppe unterschied sich hingegen kaum in der Testleistung von der Kontrollgruppe, d.h. eine bessere als die tatsächliche Testleistung war kaum zu simulieren.

In der *fake bad* - Gruppe fielen insbesondere sog. Auslassfehler (Verpasser) auf, Verwechslungsfehler (z.B. Ankreuzen falscher Zielbuchstaben, z.B. d's mit einem Strich) waren seltener. Insbesondere sogenannte Buchstabenfehler (z.B. Ablenkungs-Ziele wie p's mit zwei Strichen) traten in der *fake bad* - Gruppe sehr häufig auf.

Zusammenfassend folgerten die Autoren, dass es selbst ausgewählt intelligenten Probanden schwer fallen sollte, im Test d2 simulierend eine bessere als die tatsächliche Testleistung zu simulieren. Hingegen offenbarte sich die Strategie, sehr genau, aber bewusst langsam zu arbeiten, als leicht detektierbar. „Maßvolles" Überspringen von Zeichen war die am schwierigsten nachzuweisende Testverfälschung, wurde aber von nur wenigen Probanden erkannt und angewandt. Das Vorliegen von bereits einem Doppelfehler (z.B. p's mit einem oder mit drei Strichen - somit falscher Buchstabe und falsche Strichzahl gegenüber dem Ziel-Item) war in dieser Studie ebenfalls ein relativ sicherer Hinweis auf Simulation.

Merten et al. (2007c) [195] analysierten d2-Ergebnisse aus Archivdaten von fünf Probanden-Gruppen: 74 neurologische Patienten ohne Verdacht auf Antwortverzerrungen, je 30 Gutachtenprobanden, die im Word Memory Test unauffällig und die auffällig abgeschnitten hatten, 12 instruierte Simulanten und 12 gesunde Kontrollpersonen. Dabei zeigte sich, dass sog. Doppelfehler (z.B. Ablenkungsziele p's mit einem oder drei Strichen) selten auftraten und keine Häufung in einer Probandengruppe aufwiesen. Dasselbe Ergebnis berichteten auch Schmidt-Atzert et al. (2004) [259]. Sogenannte Buchstabenfehler (z.B. Ablenkungs-Ziele wie p's mit zwei Strichen) traten häufiger auf, jedoch zeigte der vorgeschlagene Cutoff ≥ 2 keine hinreichende Güte der Diskriminanz von Overreporting. Merten (2011) [194] urteilt folglich zur Validität des d2 als Symptom-Validierungsverfahren, dass eine systematische Untersuchung für den deutschen Sprachraum bislang noch nicht hinreichend durchgeführt worden sei.

Eine sehr gute Detektierbarkeit simulierter Gedächtnisdefekte konnten Barbeau et al. (2004) [23] mittels des Tests *DMS48 (Description of the visual recognition Memory Task)* mit hohem Deckeneffekt nachweisen, in dem selbst Patienten mit moderater Alzheimer-

Demenz mehr als 50 % Wiedererkennungsleistung erbrachten, Patienten mit mildem Amnestischem Impairment ca. 60 Prozent Testleistung und Patienten mit Parkinson-Erkrankung mit 95 % Leistung fast ebenso gute Ergebnisse wie gesunde Probanden erzielten (s.a. Guedj et al. 2006 [120]). Insofern lassen sich auch mit dem DMS48 simulierte Gedächtnisdefizite relativ sicher im Antwortverhalten differenzieren (s. Abb. 5).

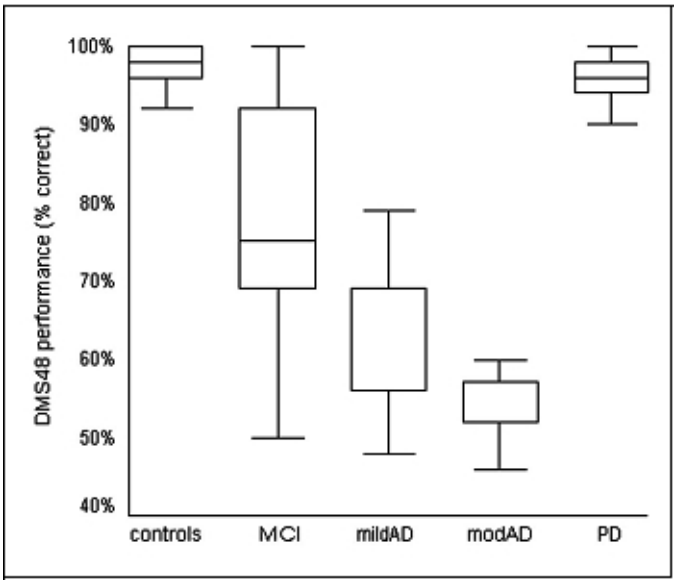


Abb. 5. Durchschnittliche Testleistung im DMS48 bei gesunden Probanden (controls), Patienten mit milder kognitiver Störung (MCI), Patienten mit milder Alzheimer Demenz (mildAD), Patienten mit moderater Alzheimer Demenz (modAD) und Patienten mit M. Parkinson (PD); nach Barbeau et al. (2004, 1319pp.) [23]

Als weiteres trennscharfes Verfahren mnestischer Simulation wird von Merten und Dohrenbusch (2010) [197] der *Rey-15-Item-Test* diskutiert, bei dem bei einem Cutoff der Wiedererkennungsleistung unter 9 Items von einer sicheren Aggravation ausgegangen werden muss. Merten (2011) [194] wertet den Fifteen Item Test (FIT) oder auch Rey Memory Test (RMT) als traditionelles Verfahren zur Beschwerden-Validierung, das als sog. Stand-Alone-Testverfahren keine hinreichende Validität aufweise und nur ergänzend im Rahmen einer umfassenderen Prüfung eingesetzt werden sollte.

Andererseits konnten Boone et al. (2002) [37] überzeugend nachweisen, dass die Validität des Rey15 (Spezifität über 90 % bei hinreichender Sensitivität von mehr als 70 %) bereits durch Ergänzung einer kurzen Wiedererkennungsaufgabe unter Verwendung einer gleichen Anzahl falscher Antwort-Möglichkeiten erheblich verbessert werden kann. So zeigten Boone, Salazar et al. (2002) [37], dass die Anwendung eines Kombinations-Scores als Summe der korrekt unmittelbar wiedergegebenen Items und einem Korrekturfaktor (richtig

wiedererkannte Items abzüglich der Anzahl falsch positiv wiederkannter Items) bei einem Cutoff-Score ≤ 21 ein Overreporting neurokognitiver Defizite sehr spezifisch trennscharf ermöglicht.

Larrabee (2007, S. 32) [166], der in einer Replikation das Testverfahren mit 105 stationären Patienten gleicher Einschlusskriterien und 90 als nicht-authentisch-antwortend beurteilte Probanden untersuchte, wertet den Rey15-Recognition-Trial-Test ebenfalls bei demselben Cutoff-Wert als nützliches und ökonomisch durchführbares Screening. Seine Untersuchung ergab eine leicht geringe Spezifität von 80 % bei ebenfalls etwas geringerer Sensitivität von 61,1 %; auch bei einer zugrundegelegten Basisrate von 30 % wies der Test eine hinreichende Positive Prädiktive Power von 57,1 % auf, wie auch eine sehr gute Negative Vorhersage-Kraft von 82,6 %. Der Autor wies darauf hin, dass der Rey15-Recognition-Test bei einem unauffälligen Score wenig Hinweise auf die kognitive Leistung des Untersuchten bietet, jedoch bei auffälligen Scores relativ sichere Hinweise auf Overreporting liefern kann.

Denselben Cutoff (≤ 21) im Rey-15-Wiedererkennungs-Test bestätigten auch Morse et al. (2013) [212] in einer jüngeren Studie, die 29 Patienten mit laufenden gerichtlichen Anerkennungsverfahren für Behinderung und mit sicheren Auffälligkeiten in PVTs mit 63 vergleichbaren Patienten mit Gerichtsverfahren ohne PVT-Auffälligkeiten in Bezug setzten, die wiederum mit 36 lernbehinderten Patienten und 54 neuropsychologischen Patienten ohne Gerichtsverfahren in Bezug gesetzt wurden. Der Test wies bei einer 92,8-prozentigen Spezifität eine hinreichend gute 70-prozentige Sensitivität auf.

Dohrenbusch (2009) [69] weist darauf hin, dass vorgetäuschte kognitive Störungen bei Patienten mit chronischen Schmerzen außerhalb des Begutachtungsettings vermutlich eher selten vorkommen (s. auch Iverson 2007 [142]). Insofern ist die Frage, in wieweit ihnen bei der Detektion von Overreporting gegenüber weiteren Bereichen möglicher Antwortverzerrungen (z.B. Überhöhung psychopathologischer Symptome oder Verhaltensauffälligkeiten) dieselbe Bedeutung zuzuordnen ist.

Merten und Dohrenbusch (2010) [197] betonen, dass ohne die Prüfung negativer Antwortverzerrungen mit Beschwerdenvalidierungsverfahren subjektive Beschwerdeangaben oder Ergebnisse von Persönlichkeitstests nur begrenzte Validität haben. So kann beispielsweise mittels des Freiburger Persönlichkeitsinventars FPI-R (Fahrenberg et al. 2001 [87]) oft nur eine faking-good-Antworttendenz (Dissimulation, Underreporting) erfasst werden (s. auch

Merten et al. 2007b [198]). Ähnlich beschreibt auch Schneider (2007) [261], dass monoforme, extreme oder inkonsistente Testergebnisse, aber auch kritische Werte in Offenheitsskalen (z. B. FPI-R), Hinweise auf „sozial erwünschtes“ oder tendenziöses Antwortverhalten geben (vgl. Wagner, Richter et al. 2003 [309]).

Bisher besteht jedoch kein wissenschaftlicher Konsens oder verbindliche Leitlinien, welche Verfahren mit welchen Cutoff-Kriterien welchen Stellenwert bei der Beschwerdenvalidierung haben. Dies mag auch darin begründet liegen, dass die meisten Belege guter Sensitivität, Spezifität und Vorhersagewerte von BV-Verfahren aus Vergleichs-Studien mit instruierten, ge-coachten gesunden Probanden stammen. Youngjohn et al. (1999) [325] zeigten jedoch bereits, dass vor-instruierte Probanden weniger überhöhte und mehr (scheinbar) glaubhafte Angaben in Beschwerden-Validierungsverfahren machen als nicht-instruierte Probanden.

1.7.2 Beschwerdenvalidierung auf Verhaltens- und somatischer Ebene

Antwortverzerrungen auf somatisch-verhaltensbezogener Ebene können als Diskrepanzen zwischen dem Patientenverhalten während der Untersuchung (berichteten oder präsentierten Einschränkungen oder verminderten Leistungskapazitäten) im Abgleich mit anderen Situationen erfasst werden (z.B. Schilderungen zum Alltagsverhalten, entsprechenden Fremdberichten, sog. nicht-organischen Untersuchungsbefunden mit Feststellung submaximaler Anstrengung oder Verdeutlichung).

Kool et al. (2008) [156] weisen nach ihren Erhebungen bei Experten und Gutachtern im Schweizer Staatsgebiet darauf hin, dass die Prüfung von Beschwerdeüberhöhungen auf der Verhaltensebene in der Regel mit nur unzureichend validierten Verfahren erfolgt, die die Verhaltens-Konsistenz überprüfen.

Mögliche Testverfahren sind funktionelle Leistungs- oder Performancetests, z.B. die Evaluation der funktionellen Leistungsfähigkeit EFL nach Susan Isernhagen (1988) [141] (vgl. Kaiser et al. 2000 [146]) oder der Performance Assessment Capacity Test PACT nach Matheson & Matheson 1989 [181] (vgl. Kool et al. 2005 [157]), die sog. Waddell-Zeichen (Waddell et al. 1980 [308]), der JAMAR Handkraft Test (Bechtol 1954 [26]) oder das Screening für somatoforme Schmerzstörung SOMS, Version 2 und 7 (nach Kool et al. 2008 [156]).

Meistenteils wurden auch diese Verfahren *nur in Simulations-Studien* auf ihre Aussagekraft untersucht.

Beim EFL-Test wird die Belastbarkeit für realitätsnahe, physische arbeitsbezogene Funktionen und Arbeitsfähigkeit anhand 29 standardisierter Leistungstests (z.B. Heben, Tragen, Überkopf-Arbeit, auf Leitern steigen, Handkoordination usw.) untersucht. Die gesamte Testbatterie erfordert jedoch eine sechsstündige Untersuchung an meist zwei aufeinanderfolgenden Tagen.

Bei der PACT-Erhebung beurteilt die befragte Person ihre Beeinträchtigung in 50 alltagsnahen Aktivitäten in stehender oder sitzender Körperposition auf einer fünfstufigen Skala zwischen „nicht durchführbar“ bis „unbeeinträchtigt ausführbar“. Eine ähnliche, aber verkürzte Strategie zur Erhebung von Funktions-Beeinträchtigungen, insbesondere bei Patienten mit Rückenleiden und rheumatoiden Schmerzen, bildet der Funktionsfragebogen Hannover (FFBH-R, nach Kohlmann & Raspe 1996 [154]). In dieser Abfrage ist bei 12 alltagsnahen Tätigkeiten (Heben und Tragen von Lasten, längeres Sitzen oder Stehen, schnelles Laufen, sich strecken, sich bücken, Oberkörper vorbeugen, Aus-der-Rückenlage-Aufsetzen) anzugeben, inwieweit der Proband diese „sehr gut“, „mit Mühe“ oder „überhaupt nicht“ ausführen kann. Die Berechnung des Summenwerts erfolgt durch Addition der Items mit anschließender Normierung auf den Bereich 0 % bis 100 %. Auch hier ist ein Konsistenz-Abgleich dieser subjektiven Einschätzungen eines Untersuchten mit dem klinischen Befund und Verhaltensbeobachtungen möglich.

Hinsichtlich der Schmerzstärke-Angaben von Patienten verglichen Dirks et al. (1993) [68] diese Selbst-Einschätzungen mit parallelen Ratings des behandelnden Pflegepersonals. Ihre Vergleiche zeigten, dass Patienten mit extrem hohen Schmerzangaben sehr viel höhere Diskrepanzen zu den Fremdratings (64,6 %) aufwiesen als Patienten mit moderateren Schmerzangaben (14,2 %). Ähnliche Ergebnisse sind auch hinsichtlich subjektiver Patientenangaben zu Fremdeinschätzungen von Funktions-Behinderungen denkbar, wie sie im klinischen Schmerz-Assessment die Regel sind.

Die sog. *Waddell-Signs* sind Indizien der körperlichen Untersuchung ursprünglich orthopädischer Patienten, die auf nicht-organisch erklärbare Ursachen der Beschwerden hindeuten (z.B. Druck und Bewegungs-Schmerzempfindlichkeit, die somatisch erklärbare Grenzen überschreitet; plötzliche unausgewogene Schwäche als Pseudozahnradphänomen, unruhige

Bewegungen, bei sonstiger Testung normaler Muskelkraft; Angabe von Schmerzen bereits bei Betasten der Haut). Die Kenntnis dieser Tests sollte zumindest in jede Untersuchung und insbesondere in die Begutachtung chronischer Schmerzen einfließen, wenngleich es sich nur um Indizien, keine sicheren Malingering-Kriterien handelt.

Der JAMAR Hydraulic-Hand-Dynamometer ist ein Standardgerät, das schon seit über 35 Jahren zur Messung der Handkraft im Gebrauch ist. Viele Länder benutzen den JAMAR als Standard-Testgerät zur Bestimmung von Schadensersatzfällen bei Handtraumata und ähnlichen Krankheitsbildern. Zur Erstellung von Normwerten untersuchten Schmidt & Toews (1970) [258] Bewerber einer Stahlfabrik und stellten deren Daten dar, Mathiowetz et al. (1985) [182] untersuchten Grob- und Feingriffe mit einer Normierung der Probanden in 12 Altersklassen. Zur Aktualisierung dieser älteren Erhebungen erfassten Ewald & Kohler (1991) [86] im Schweizer Großraum Zürich Normdaten an 1000 Personen der Normalbevölkerung.

Das sog. Screening für somatoforme Schmerzstörung (SOMS) erfasst als Selbsteinschätzung Beschwerden der letzten zwei Jahre und der letzten sieben Tage (als Itemliste von 52 Symptomen bei Frauen bzw. 48 Beschwerden bei Männern mit fünfstufiger Möglichkeit der Beantwortung). Ab einer Anzahl von ≥ 35 Beschwerden sollen Angaben als Indiz für eine Somatoforme Störung gelten; Inkonsistenzen zwischen den Beschwerden in beiden Zeiträumen sollen zudem nach Kool et al. (2008) [156] auf nicht-authentische Angaben hindeuten. Es fehlen jedoch genaue Angaben über die Diskriminanzgüte dieses Validierungsprocedere (Sensitivität, Spezifität, Prädiktive Power).

Der Einsatz der eingangs genannten, aufwendigeren und vermutlich valideren Verfahren scheitert jedoch in der klinischen Praxis oft an ihrem meist hohen Durchführungsaufwand. Insofern werden klinische Tests wie die EFL für verhaltensbezogene Beschwerdeüberhöhungen eher selten angewandt, z.B. im Patientenauswahl-Screening für das Göttinger Rücken-Intensiv-Training GRIP (Hildebrandt et al. 1996 [135]). Andere Verfahren, wie die *Waddell-Signs*, fließen zwar aufgrund ihrer Bekanntheit häufiger in schmerzmedizinische Untersuchungen ein, jedoch fehlen auch hier hinreichende Untersuchungen zu ihrer Validität.

Die Aussagekraft dieser Verfahren wird zudem durch ihre jeweilige Spezifität auf bestimmte Schmerz-Syndromgruppen eingeschränkt; so beziehen sich die meisten dieser Prüfverfahren auf Funktions-Behinderungen bei Rücken- und muskulo-skelettalen Schmerz-

syndromen (z.B. Waddell-Signs oder EFL), sind aber nur bedingt auf andere Schmerz-Diagnosen (Ischämieschmerzen, nervale Schmerzsyndrome, Kopf- und Gesichtsschmerzen) übertragbar.

Bianchini et al. 2005 [31] schlagen deshalb in ihrem multidimensionalen Konzept zur Detektion von MPRD (Malingered Pain Related Disabilities) vor, die aktuell verfügbaren Beschwerdenvalidierungsverfahren der behavioralen Domäne nur ergänzend zu Verfahren aus der kognitiven und psychopathologischer Domäne einzusetzen.

Die Prüfung der Glaubwürdigkeit verhaltensbezogener Behinderungen durch Schmerzen ist bislang am wenigsten validiert und bedarf meist aufwendiger Prüfsysteme. Eine Aggravations-Klassifikation auf dieser Funktions-Ebene ist somit am wenigsten sicher.

1.7.3 Beschwerdenvalidierung psychopathologischer Symptome

Weit besser untersuchte BV-Verfahren sind zur Prüfung von Overreporting psychopathologischer Symptome verfügbar, z.B. der SFSS (Strukturierter Fragebogen simulierter Symptome) nach Smith und Burger (1997) [281], das im englischen Sprachraum als SIMS (Structured Interview of Malingered Symptomatology) veröffentlicht wurde, sowie das in den USA verfügbare SIRS (Structured Interview of Reported Symptoms) nach Rogers et al. (1992) [237]. Ein weiteres Verfahren wie das PAI (Personality Assessment Inventory) nach Morey (1991, 1996; [211], [210]) ist in jüngster Zeit auch deutschsprachig verfügbar (Engel 2013 [85]). Als Gold-Standard der Beschwerden-Validierung werden insbesondere diverse Validitätsskalen des MMPI-2 (Minnesota Multiphasic Personality Inventory) nach Keller & Butcher (1991) [149] eingesetzt.

1.7.3.a Structured Interview of Reported Symptoms (SIRS)

Mittels des 172-Items-umfassenden SIRS kann sog. „feigning“ (dt. Simulation) psychischer Erkrankungen anhand von acht Primär-Skalen und fünf Ergänzungsskalen erhoben

werden. Das als strukturiertes Interview durchgeführte SIRS wurde seit seiner Entwicklung im Jahr 1985 mehrfach revidiert, die aktuelle Fassung ist das sog. SIRS-2.

Das Verfahren nutzt die Strategien zur Erfassung von Overreporting durch Abfrage seltener, absurder und teilweise offensichtlicher Symptome, aber auch von Underreporting-Antwortmustern, wie Bagatellisierung (Defensiveness) und Selbstzuschreibungen von Glaubwürdigkeit. Die Subscores der Basisskalen ermöglichen in vierfacher Abstufung (glaubwürdig, unklar überhöht, wahrscheinlich und definitiv überhöht) eine Klassifikation nicht-authentischer Antwortgebung. Zusätzlich lässt sich mit dem SIRS durch Integration von 32 Wiederholungs-Fragen problematisches Antwortverhalten, wie Inkonsistenzen der Antwortgebung, identifizieren.

Validitäts-Untersuchungen zur Klassifikation von Overreporting psychischer Störungen bestätigten mittels der SIRS-Indikatoren bei einer Basisrate von 31,8 % und einer Spezifität von 97,5 % eine relativ geringe Falsch-Negativ-Rate von 20 %. Für dieses Verfahren wurden zudem positiv und negativ prädiktive Werte von über 90 % berichtet (Rubenzer 2010 [254]). Das SIRS gilt deshalb im anglo- amerikanischen Sprachraum als Standard zur Erhebung von Malingering und wird häufiger auch zur Kreuzvalidierung anderer Testinstrumente verwendet. Die Entwicklung und Validierung einer deutschen Version des SIRS an einer deutschen und schweizerischen Stichprobe ist aktuell in der Vorbereitung (Lanquillon & Schmidt 2013, s. Plohmann & Merten 2013 [223]).

1.7.3.b Personality Assessment Inventory (PAI)

Das 344-Items-umfassende Personality Assessment Inventory (PAI, dt. Fassung VEI) umfasst 22 nicht-überlappende Subskalen, die psychopathologische Symptome Erwachsener umfassend erheben sollen. Einschätzungen der Probanden erfolgen auf einer vierstufigen Skala. Vier Arten von Skalen liefern Informationen über (1) die Aussage-Validität, (2) die klinisch-psychiatrische Symptomatik der untersuchten Person, (3) behandlungsrelevante Informationen (z.B. zur Suizidalität, Aggression, Stress, fehlender sozialer Unterstützung, fehlender Behandlungs-Motivation) und (4) interpersonelle Informationen, z.B. über die psychosozialen Fertigkeiten. Die klinischen Skalen erheben Aspekte Somatischer Beschwerden, Angst, Angst-assoziierte Symptome, Depression, Manische, Paranoide und Schizophre-

ne Symptome, Borderline-Charakteristika, Antisoziales Verhalten, Alkohol- und Drogen-Probleme.

Die PAI-Validitäts-Skalen erfassen vier bereits beschriebene Antwortstile: Inkonsistenz der Antworten (*Inconsistency*), Überhöhung bizarrer und seltener Symptome (*Infrequency*), positive und negative Selbst-Darstellungen (*PIM-, NIM-Scales - Impression Management*). Drei weitere PAI-Indizes wurden ergänzend publiziert: Cashel's Discriminant Function (CDF, Cashel et al. 1995 [52]) differenziert durch die Zusammenfassung von sechs PAI-Validitätsskalen authentische Probanden von Probanden, die ein Underreporting- oder Defensiveness-Verhalten zeigen. Dabei zeigten die CDF-Scores eine höhere Korrelation zum psychischen Zustand der Probanden (gesund vs. psychisch krank) als die DEF-Skala. Der Malingering Index (MAL, Morey 1996 [210]), konstruiert aus Items von 11 PAI-Subskalen, erwies sich in Kombination mit der NIM-Skala als trennscharfes Diskriminanzinstrument für Overreporting. Schließlich fasst Rogers Discriminant Function (RDF, Rogers et al. 1996 [245]) Items aus 20 PAI-Subskalen indizierend zusammen und nutzt die Strategie ungewöhnlicher Symptome zur Detektion von Overreporting. Er wurde von Rogers et al. empirisch anhand weiterer externer Klassifikations-Strategien ermittelt.

Die diversen Studien zur Klassifikationsgüte der PAI-Validitätsskalen und -indices zur Detektion von Overreporting werden differenziert von Sellbom & Bagby (2008) [268] diskutiert. In verschiedenen Simulationsstudien zeigte der CDF-Score gegenüber der DEF- und der PIM-Skala eine schlechtere Diskriminanz von Defensiveness. Sellbom & Bagby (2008) folgern, dass der CDF-Score zur Identifizierung von Probanden mit Underreporting geeigneter ist; dem gegenüber differenzieren die beiden anderen Scores mehr Patienten mit nicht-defensivem Antwortverhalten. Alle drei Skalen erwiesen sich als durch vorheriges Coaching der Probanden verfälschbar. Zur Detektion von Underreporting werden im Fall vermuteten Coaching's von Untersuchungspersonen niedrigere Cutoff-Werte (für PIM T50, DEF Rohwerte ab Score 4) empfohlen.

Zur Detektion von Overreporting erwiesen sich in klinischen und forensischen Stichproben die PAI-Skalen NIM und die Indizes MAL und RDF als am meisten trennscharf. Die RDF-Skala mit Integration von 20 Subskalen zeigte in Meta-Analysen die größte mittlere Effektstärke (d = 1,87, vgl. Sellbom & Bagby (2008) [268]). Hingegen zeigten Untersuchungen, die in einem Known-Groups-Design Probanden hinsichtlich Malingering vorab mittels

des SIRS (des häufig als Gold-Standard für Simulation psychischer Störungen benannten Verfahrens) klassifizierten, eine genau umgekehrte Detektionsgüte der drei Validitätsscores.

Die Detektions-Genauigkeit der NIM-Skala (Negative Impression Management) erwies sich in Studien mit artifiziellem Simulations- und mit Known-Groups-Design bei T-Werten zwischen 77 und 88 als hoch spezifisch in der Detektion von Overreporting, jedoch aus Sicht der meisten Malingering-Experten als nicht hinreichend sensitiv, mit falsch negativen Resultaten über 20 %. Zudem zeigte sich eine Anfälligkeit dieses Scores für erfahrene, über eine Woche trainierte simulierende Studenten (Rogers et al. 1996 [245]).

Der MAL-Index zeigte bei Cutoff-Scores ≥ 3 eine eher niedrige Sensitivität (ca. 0,60) bei Spezifitätswerten um 0,80. Dies empfiehlt aus Sicht von Sellbom & Bagby (2008) [268] den Index bestenfalls als Vorab-Screening. Höhere Spezifitätswerte bei Cutoff-Scores ≥ 5 gingen mit nur sehr geringer Sensitivität einher (Rogers et al. 1998 [244]).

Aufgrund der niedrigeren Detektionsgüte des RDF-Index in Known-Group-Designs gegenüber seiner höheren Validität in Simulationsstudien wurde dieser Index von Sellbom & Bagby (2008) [268] nicht in klinischen oder forensischen Untersuchungen zur Aufdeckung von Malingering empfohlen.

Zusammenfassend erweisen sich die unterschiedlichen Validitäts-Scores und -Indizes des PAI zwar als durchaus trennscharf und zur Detektion von Malingering geeignet, das Verfahren selbst ist jedoch eher wenig ökonomisch in der Durchführung. Der NIM-Score erwies sich unter den Validitätsmaßen im Vergleich als der trennschärfste Score.

Die deutsche Version des PAI (Verhaltens- und Erlebensinventar VEI) eignet sich ebenso wie seine Vorgänger-Fassung zur Identifikation von Antwortverzerrungen. Nach einer Untersuchung von Groves (2009) [119] zur Überprüfung der Konstruktvalidität des VEI an einer psychiatrischen Stichprobe zeigten sich hohe Korrelationen mit inhaltsähnlichen Skalen des MMPI-2 und des AMDP-Diagnostik-Systems (Baumann & Stieglitz 1983 [25]). Auch bestätigte sich die voruntersuchte Faktorenstruktur des VEI.

Tscheuschner (2011 [299]) bestätigte in einer Studie mit 101 Patienten des Medizinischen Gutachteninstituts Tübingen eine Überlegenheit der NIM-Validitätsskala des VEI gegenüber den Skalen RDF und MAL auch in der deutschen PAI-Fassung. In dieser Studie zeigte die NIM-Skala zudem eine hochsignifikante Korrelation zu den neurokognitiven Verfahren

SFSS (Strukturierter Fragebogen simulierter Symptome) und zum WMT (Word Memory Test).

1.7.3.c Structured Interview of Malingered Symptomatology (SIMS)

Das 75 Items umfassende SIMS ist ein im deutschen Sprachraum verfügbares, sehr ökonomisches und erprobtes Selbsteinschätzungs-Instrument zur Erfassung von Simulation durch Abfrage einer Vielzahl simulierter Symptome (Cima et al. 2003a [54]). Die Itemauswahl erfolgte anhand diverser Quellen wie des MMPI-2 und des SIRS. Ergänzende Items spiegeln Strategien von simulierenden Probanden vorangegangener Studien wider (Seamons et al. 1981 [267], Resnick 1984 [228], Rogers 1984 [233]).

Der Gesamt-SIMS-Wert erwies sich als der effizienteste Indikator für Simulation. Er identifizierte 95,6 % der Simulanten (Sensitivität) und 87,9 % der ehrlich antwortenden Teilnehmer (Spezifität; Smith und Burger 1992 [280], 1997 [281], Alwes 2006 [8]). Cima et al. (2003a) [54] bestätigten, dass die Werte der F-Skala im MMPI-2 von Patienten mit einem Gesamt-SFSS-Wert von 17 oder mehr signifikant höher waren als die der Patienten mit einem Gesamt- SFSS-Wert unterhalb dieses Cutoffs. Eine niedrigere Spezifität von 40 % berichtete Edens et al. (2007) mit einem tieferen Cutoff von 14 unter psychiatrisch-forensischen Patienten im Vergleich mit einer Kontrollgruppe Gesunder und simulierender Gesunder.

Das SIMS enthält fünf Skalen: Psychotische Symptome (P), Affektive Symptome (Af), Neurologische Defizite (N), Amnestische Symptome (Am) und Intelligenzmängel (LI). Die Skalen P, Af, N und Am erfassen untypische Störungsmerkmale, während die LI-Skala Items umfasst, die bei der Verdeutlichung intellektueller Defizite zu erwarten wären (Heinze & Purisch 2001 [128]).

Das SIMS weist eine hohe negative *Predictive Power* auf und damit auch eine hohe Sensitivität. Insofern sollte das SIMS mit einem Verfahren hoher Spezifität kombiniert werden (Alwes 2006 [8]), d.h. einem BV-Verfahren, das möglichst wenige tatsächliche Non-Malingerer falsch positiv klassifiziert, also mit hohem Cutoff. Alwes (2006) schlägt deshalb vor, zunächst das SIMS zur Vor-Klassifikation zu verwenden, um bei Auffälligkeiten ein weiteres BV-Verfahren (z.B. das länger dauernde SIRS) zur genaueren Klassifikation einzusetzen.

Das SIMS wurde in drei Studien evaluiert (Smith 1992 [280]); Smith & Burger 1997 [281]); Edens et al. 1999 [80]). Zwei Studien untersuchten forensische Probanden (Heinze & Purisch 2001 [128]); Lewis et al. 2002 [170]) und eine Studie jugendliche Straftätern (Rogers et al. 1996 [240]).

Smith (1992) [280] forderten in ihrer Simulationsstudie sechs Probandengruppen auf, in einer der SIMS-Skalen bewusst Symptome zu simulieren, eine Gruppe sollte zudem in dem gesamten Fragebogen überhöhte Angaben machen und eine Probandengruppe möglichst ehrlich antworten. Der SIMS-Gesamtscore ermöglichte mit einer signifikanten Trefferrate bei einem Cutoff-Score von mehr als 16 eine korrekte Klassifikation.

Rogers et al. (1996) [240] untersuchten 53 jugendliche Straftäter in einer Analogstudie, wobei die Probanden den Test zweimal ausfüllten: im zweiten Studienteil wurden sie zur bewussten Überhöhung ihrer Antworten aufgefordert. Für die Testbearbeitung und zusätzlich für die erfolgreiche Verfälschung wurden die Probanden finanziell belohnt. Bei einem Cutoff über 16 wurde eine Positive Prädiktive Power (PPP) von 0,87 erreicht, eine Negative Prädiktive Power (NPP) von 0,62. Der Cutoff von ≥ 40 zeigte eine optimale Detektionsrate mit einer PPP von 0,49 und einer NPP von 0,94.

Smith und Burger (1997) [281] untersuchten in einem Simulationsdesign 476 College-Studenten mittels des SIMS in zwei gleich großen, randomisierten Gruppen, wobei eine Gruppe zur Symptom-Simulation, die andere zu authentischen Beantwortung aufgefordert wurde. Mit Hilfe des SIMS-Gesamtscores gelang es bei einem Cutoff-Score von über 14 beide Gruppen mit einer Sensitivität von 0,95 und einer Spezifität von 0,88 zu differenzieren.

In der Studie von Edens et al. (1999) [80] füllten 196 Collegestudenten das SIMS und die Symptom-Checkliste SCL-90-R (Derogatis 1977 [64]) zweimal aus, einmal in authentischer Weise und einmal mit der Instruktion eines Overreportings, aber gleichzeitig mit der Instruktion, möglichst nicht beim Malingering entdeckt zu werden. Zur Klassifikation wurde der Cutoff-Score von ≥ 14 nach Smith & Burger (1997) [281]) verwendet. Dieser Cutoff ermöglichte, Simulation mit einer Sensitivität von 0,96 und einer Spezifität von 0,91 zu differenzieren, wobei die SIMS-Subskalen keine hinreichende Detektionsgüte aufwiesen, da 22 % der Probanden mit hohen SCL-90-R-Scores und authentischer Psychopathologie mittels des Cutoff 14 falsch positiv klassifiziert wurden. Wurde zusätzlich der Global-Severity-Index-Score der SCL-90-R von mindestens 45 zur Klassifikation herangezogen, erhöhte sich die

Detektionsgüte hinsichtlich der Sensitivität bis 1,00 bei einer Spezifität von 0,78 bis 0,91. Die Autoren plädierten für einen höheren SIMS-Cutoff-Score als 14.

Edens et al. (2007) [81] verglichen die Detektionsgüte des SIRS, des PAI und des SIMS in einem Viergruppendesign miteinander. Alle drei Instrumente zeigten bei einer Kontrollgruppe der Normalbevölkerung und bei Häftlingen, die instruiert wurden, psychische Symptome zu simulieren, eine hinreichende Detektions-Genauigkeit. Die Diskriminanzgüte war eher gering, wenn psychiatrische Patienten und Patienten, die vom psychiatrischen Pflegepersonal als simulierend eingestuft wurden, miteinander verglichen wurden. Diese Differenzierung war nur mit zwei Indikatoren des PAI möglich. Die PAI-Validitätsskala NIM (Negative Impression Management) wies zur SIMS-Erhebung die höchste Korrelation (r = 0,84) auf, aber auch die PAI-Indizes MAL (Malingering Index) und RDF (Rogers Discriminant Function) zeigten relativ ähnliche Resultate (r = 0,68 und 0,45).

Um die schwierige Frage der Differenzierungs-Güte des SIMS bei psychisch- oder psychiatrisch kranken Patienten gegenüber Patienten mit Overreporting zu klären, erhoben Merckelbach und Smith (2003) [192] in einer niederländischen Studie zur differentiellen Prävalenz SIMS-Werte von Patienten mit hohen Depressions- und Angst-Symptomen, um diese Daten in einer zweiten Studie in Bezug zu instruierten simulierenden Probanden zu setzen, die bewusst keinen externen finanziellen Anreiz erhielten. Die Reliabilität des SIMS erwies sich mit 0,72 bei den authentisch antwortenden Probanden als relativ hoch. Wie erwartet, waren bei den simulierenden Probanden höhere SIMS-Scores als bei den Patienten mit hoher Psychopathologie festzustellen, deren Scores wiederum die der authentischen Probanden signifikant überschritten. Umgekehrt war der Anteil von Probanden mit SIMS-Scores über 16 mit hoher Psychopathologie (Probanden mit einem Score im obersten Zehntel des Gesamtsamples) sehr gering (Beck Depressionsscore $\geq$ 13 mit unter 2 Prozent; Scores des State--Trait- Anxiety-Inventory über 52 in 1,1 % der Fälle). Im Durchschnitt lagen die Depressions- und Angst-Werte der Probanden mit deutlichem Overreporting im Mittel der Gesamtpopulation. In der Gruppe mit hohen Depressionswerten erreichten nur 3 von 19 Probanden einen SIMS-Score über dem Cutoff von 16 Punkten. Insofern misst das SIMS durchaus mehr den Aspekt der Simulation als reale Psychopathologie.

Schließlich analysierten Merckelbach und Smith (2003) [192] in einem Known-Groups-Ansatz die SIMS-Daten aller 298 Probanden der Studie hinsichtlich der Klassifikationsgüte. 93 % der simulierenden Teilnehmer wurden ebenso korrekt identifiziert wie 98 % der authen-

tisch antwortenden Probanden. Der SIMS-Gesamtscore erwies sich somit als hoch valides Screeninginstrument zur Klassifikation von Malingering, wenngleich die fünf Subscores keine hinreichende Klassifikationsgüte zeigten.

Heinze und Purisch (2001) [128] befragen 57 männliche Haftinsassen, deren Fähigkeit geprüft werden sollte, an einer Gerichtsverhandlung teilzunehmen und bei denen ein Malingering anzunehmen war. Die Detektionsgüte von Verfälschungstendenzen mittels des SIMS lag bei einer Sensitivität bis 0,87 bei einem Cutoff für Malingering ≥ 14 Punkten. Leider wurden keine Angaben zur Güte der Klassifikation (Predictive Power, Wahl des Cutoff) gemacht, bei großer Stärke einer wenn auch relativ kleinen Gruppe von ´Real-Life-Malingerers´.

Lewis et al. (2002) [170] untersuchten 55 männliche Haftinsassen unter ähnlichen Voraussetzungen, deren Testverhalten allerdings vorab durch Untersuchung mittels des SIRS (Structured Interview of Reported Symptoms, Rogers et al. 1992 [237]) authentisch oder testverfälschend eingeschätzt wurde. Probanden mit Klassifikation einer Testverfälschung füllten den SIMS-Fragebogen bei einem Cutoff ≥ 17 signifikant auffälliger aus als authentisch Antwortende. Die Klassifikation gelang mit einer Negative Predictive Power von 1,00 bei einer PPP von 0,54, die Sensitivität lag bei 1,00 bei einer Spezifität von 0,61.

Alwes (2006) [8] konstatiert in einem Studien-Überblick über die SIMS-Untersuchungen, dass aufgrund einiger methodischer Schwächen (z.B. Schwächen der Cutoff-Kriterien bei Heinze und Purisch (2001) [128] bzw. drei Studien mit nicht-klinisch zuzuordnenden Untersuchungspersonen sowie kleinen Gruppengrößen, teilweise fehlender Angabe von Detektionsgütekriterien) die Festlegung der SIMS-Cutoff-Scores für Malingering unklar bleibt. Die berichteten Scores zwischen 13 und 40 erzielten Trefferraten zwischen 0,73 und 0,96, bei Sensitivitätswerten von 0,87 bis 1,00 und Spezifitätswerten zwischen 0,61 bis 0,91. Auch die Rate der Positive Predictive Power rangierte von 0,49 bis 0,87, die der Negativen Predictive Power zwischen 0,62 bis 1,00. Allerdings erreichten nach Alwes (2006) [8] Cutoff-Scores zwischen 14 und 16 eine maximale NPP, so dass bei diesen Scores die Rate falsch negativ klassifizierter Personen am geringsten zu sein scheint und am wenigsten echte Verfälschung unentdeckt bleiben sollte. Deshalb bieten sich beide Cutoff-Scores auch in der hiesigen Studie als Grenzwerte für Overreporting an.

In einer eigenen Studie untersuchte Alwes (2006) [8] 308 Patienten einer privaten neurologisch-forensischen Praxis in Lexington/USA mit Kopfverletzungen und durch die Krank-

heit reklamierter Arbeitsunfähigkeit. Neben dem SIMS füllten die begutachteten Probanden den MMPI-2 aus, ferner das SIRS nach Rogers et al. (1992 [237]), den Victoria Symptom Validity Test (VSVT, Slick et al. 1995 [278]), den Test of Memory Malingering (TOMM, Tombaugh 1996 [298]) sowie den Letter Memory Test (LMT, Inman et al. 1998 [140]). Die Ergebnisse zeigten, dass der Cutoff von $\geq$ 14 mit relativ hoher Sensitivität (von 0,96 bis 1,00) Malingering auf psychischer Ebene identifizieren konnte, die Detektion kognitiver Beschwerden oder von generellem Malingering jedoch nicht hinreichend valide gelang (Sensitivität von ca. 0,80). Diese Detektionsraten lagen erst bei einem Cutoff von $\geq$ 17 im optimalen Bereich für beide Symptomkomplexe (zwischen 0,96 und 1,00). Auch wenn keines der verwendeten BV-Instrumente eine perfekte Detektionsgüte zeigte, konstatierte Alwes aufgrund ihrer Ergebnisse das SIMS als ein sehr valides BV-Verfahren zur Aufdeckung von Overreporting (21 der 24 Effektstärkemaße der Untersuchung lagen über einem Cohen's d > 0,80).

Auch neuere Literatur-Reviews (Van Impelen et al. 2014 [303]) bestätigen eine SIMS-Spezifität unter Standard (Spezifität $\geq$ 95 %), die jedoch durch höhere Cutoff-Scores $\geq$ 17 und die Kombination mit anderen SVTs verbessert werden kann.

Da dieses Verfahren mit wesentlich kürzerer Durchführung im klinischen Kontext sehr ökonomisch anwendbar ist und im Vergleich zu aufwendigeren Verfahren (z.B. PAI) von Patienten leichter akzeptiert und bearbeitet werden kann, besitzt es angesichts der dokumentierten niedrigen Rate von falsch negativ klassifizierten Probanden mit Overreporting (12 % sog. „Verpasser“) gegenüber den anderen Verfahren Vorteile.

1.7.3.d Symptom-Checkliste SCL-90-R

Ein ebenfalls häufiges im Kontext der Begutachtungs-Literatur genanntes Inventar ist die zum Screening allgemeiner Psychopathologie verwendbare Symptom-Checkliste SCL-90-R (Derogatis 1992 [64]), die bei Beschwerdeüberhöhungen in fast allen neun Subskalen ($\geq$ T70) als geeignetes Screening für Overreporting genannt wurde (McGuire und Shores 2001 [185]).

McGuire und Shores (2001) stellten in einem Vergleich von 50 Patienten mit chronischen Schmerzen mit 20 chronische Schmerzbeschwerden simulierenden Studenten fest, dass zwar Patienten mit chronischen Schmerzen generell hohe Scores in den Bereichen Depression, Zwanghaftigkeit und Somatisierung der SCL-90-R zeigten. Die Simulanten überschätzten

jedoch die Beeinträchtigung von Patienten mit chronischen Schmerzen und zeigten in allen SCL-90-R-Skalen Überhöhungen, meist weit über T-Werten von 70.

Der als „faking-bad-Score“ proklamierte Score Positive Symptom Total (PST) zeigte hingegen eine nicht hinreichende Spezifität; bessere Ergebnisse berichteten hingegen in jüngerer Zeit Sullivan und King (2008) [286].

Die Autoren stellten unter Nutzung der im Testmanual angegebenen Cutoff-Scores für den PST > 50 für Männer und > 60 für Frauen eine hinreichend hohe Detektions-Sensitivität von 0,68 in einer Simulationsstudie mit 41 College-Studenten bei einer Spezifität derselben Höhe fest. Die Studie wurde in einem randomisierten Zweigruppendesign mit extern (finanziell) motivierten Probanden durchgeführt, die in einer Gruppe authentisch und in der anderen Gruppe nach Vorinstruktion mit Täuschungsabsicht antworten sollten.

Auch Schneider et al. (2009) [262] fanden bei Gutachtenprobanden im Vergleich zu stationären Psychotherapiepatienten signifikant höhere Werte in den SCL-90-R-Subskalen Somatisierung und Phobische Angst sowie im Global-Severity-Index (s. Abb. 6, S. 54). Die Höhe der Einschätzungen psychosomatischer Symptome war zudem hoch positiv korreliert mit den Angaben der Probanden zur Beeinträchtigung ihrer beruflichen Leistungsfähigkeit sowie ihrer Belastungen am Arbeitsplatz. Diese Ergebnisse unterstützen die Nützlichkeit der SCL-90-R zur Beschwerdenvalidierung in Begutachtungskontexten.

Hardt et al. (2000) [123] weisen zudem darauf hin, dass individuelle Scores von Patienten mit chronischen Schmerzen in der SCL-90-R zu den Angaben von Normprobanden und zu psychiatrischen bzw. zu stationären psychosomatischen Patienten abzugleichen sind, da Patienten mit chronischen Schmerzen in Bezug auf diese beiden Patientengruppen (mit Ausnahme des für Patienten mit chronischen Schmerzen spezifischen, hoch ausgeprägten Bereichs Somatisierung) genau eine Zwischenstellung einnehmen. Overreporting läßt sich somit auch mit der SCL-90-R durch starke Überhöhung der Beschwerde-Angaben im Vergleich zu gesunden Personen oder Probanden mit vergleichbaren Störungsbildern identifizieren (diskriminante und konvergente Validität).

Gillard (2010) [103] untersuchte 289 Studenten in einem 2 (Scenario) x 3 (Incentive) x 2 (Coaching) faktoriellen Studiendesign mit zwölf möglichen Instruktions-Bedingungen, hinsichtlich der Detektionsgüte von Overreporting durch das SIMS und die Symptom-Checkliste SCL-90-R. In der Anreiz-*incentive*-Bedingung wurden Probanden positiv, nega-

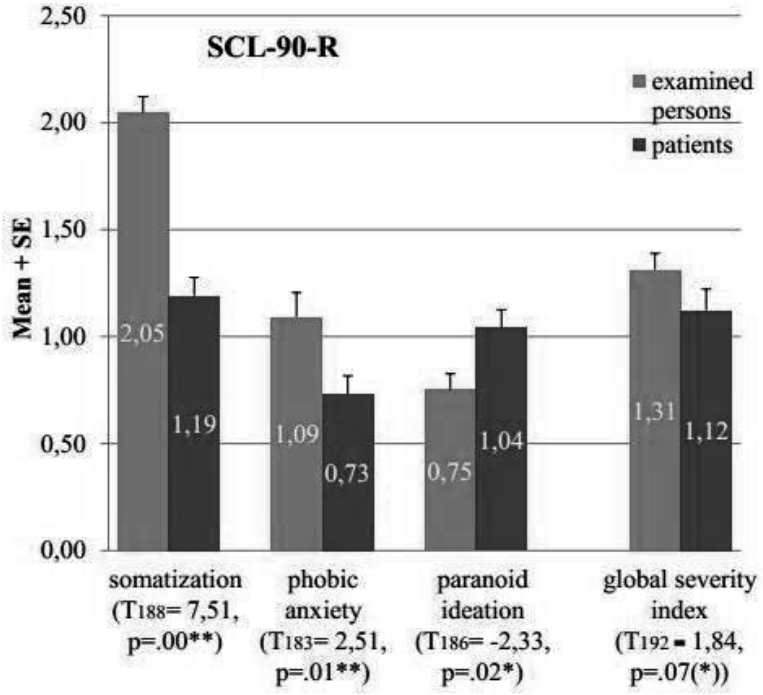


Abb. 6. Vergleiche von Gutachtenprobanden und stationären Psychotherapiepatienten mittels SCL-90-R (nach: Schneider et al. (2009) [262])

tiv und kombiniert für den Erfolg / Misserfolg bei der erfolgreichen Vermeidung verstärkt, bei der Simulation entdeckt zu werden. In der sog. *Coaching*-Bedingung wurden die Probanden unterschiedlich über die Möglichkeiten informiert, das SIMS zu verfälschen. In der Szenario-Bedingung sollten die Probanden sich unterschiedliche bekannte oder eher unbekannte Konsequenzen einer Entdeckung von Simulation vorzustellen (bekannt: Nicht-Vergabe eines Seminarscheins, unbekannt: forensische Situation).

Bereits die Erhöhung mindestens eines der neun SCL-90-Scores genügte, um die Probanden mittels Cutoff-Scores ≥ 16 in der SIMS mit hoher Spezifität von 0,96 und ebenso hoher Sensitivität von 0,82 korrekt zu identifizieren. Die Bekanntheit des Szenarios hatte einen Einfluss auf die gewählte Antwortstrategie; Probanden mit mehr Vertrautheit mit der Untersuchungs-Situation aggravierten weniger im SIMS als mit der Situation wenig vertraute Probanden. Wie erwartet, zeigten auch die vorab instruierten Probanden weniger Beschwerde-Überhöhungen in beiden Fragebögen gegenüber nicht-instruierten Personen.

Interessanterweise hatte die Art des Anreizes für Nicht-Entdeckung der Simulation von Symptomen keinen so großen Einfluss auf die Beschwerden-Überhöhung im SIMS wie in

der SCL-90-R. Tendenziell motivierten negative Konsequenzen mehr, der Entdeckung zu entgehen, als die beiden anderweitigen Verstärkungs-Bedingungen.

In der SCL-90-R zeigte die Psychopathie-Skala die am häufigsten überhöhten Scores, an zweiter Stelle die Skala Depression und Aggressivität. Simulierende Probanden mit der höchsten Anzahl überhöhter Skalen der SCL-90-R waren auch durch die SIMS-Testung am häufigsten identifizierbar. Probanden mit fünf und mehr Subskalen-T-Werten über 70 wurden zu 92 % auch mit der SIMS als simulierend klassifiziert. Zusammenfassend unterstützte die Studie somit Detektions-Möglichkeiten von Malingering auch mittels der SCL-90-R.

Zusammenfassend ist zum Assessment von Overreporting psychischer Symptome festzustellen:
Das Strukturierte Interview für vorgetäuschte Symptome SIRS besitzt eine sehr hohe Detektionsakkuranz, wurde aber bislang nicht in die Deutsche Sprache übertragen und an einer entsprechenden Population normiert. Die NIM-Skala (Negative Impression Management) aus dem PAI (Personality Assessment Inventory) zeigte in vorangehenden Studien ebenfalls gute Detektions-Eigenschaften, die jedoch nicht die Testgüte des SIRS erreichten. Das SIMS (Structured Interview of Malingered Symptomatology) weist zur NIM-PAI-Skala die höchste Korrelation auf. Ein erhöhter Cutoff-Score über einem SIMS-Summenwert 16 erwies sich in früheren Studien als am meisten resistent gegen Messfehler. Das SIMS ist nach allen verfügbaren Studien gegenüber dem SIRS und dem PAI das ökonomischste Verfahren zur Erhebung von verbalem Overreporting im deutschen Sprachraum. Obwohl ursprünglich nicht zur Beschwerdenvalidierung entwickelt, ist Overreporting auch mittels der Symptom-Checkliste SCL-90-R als ökonomisches Screening-Verfahren zu erfassen.

1.8 Grundsätzliche Konzepte der Beschwerdenvalidierung

Als einen initialen Konzeptansatz der Beschwerden-Validierung entwickelten Costa (1985, 2000) [58] und Tearnan & Lewandowski (1992, 1997) [294] erstmals Beispiele für selte-

ne somatische Symptome (s. Green 2008, S. 31 [115]). Besonders seltene und untypische Symptom-Angaben gelten dabei als Indiz für Beschwerden-Überhöhungen.

Eine andere Detektionsstrategie bezieht sich vor allem auf die von Probanden mit Malingering angegebene Stärke der Symptome. Lanyon (2003) [164] entwickelte die Health Problem Overstatement Scale (HPO) als Spezialskala des PSI (Psychological Screening Inventory-2, Lanyon 1978 [163]) um zu messen, wie Probanden ihre allgemeinen Gesundheitsprobleme übermäßig darstellen. Die Mehrzahl der Items umfasste dabei physische Symptome, wie Müdigkeit und einen allgemein schlechten Gesundheitszustand. Simulierende Probanden machten in der HPO-Skala wesentlich höhere Angaben als stationär behandelte Patienten (Cohen′s d = 2,10). In gleicher Weise zeigte die Schwere von somatischen und autonomen Symptomen im Fragebogen Modified Somatic Perception Questionaire (MSPQ; Main 1983 [176]) mit großen Effektstärken Unterschiede zwischen authentischen Patienten und Probanden mit Malingering.

Ein weiteres frühes Verfahren zur Erfassung von Dauer und Schwere von Symptomen ist das Neuropsychological Symptom Inventory (NSI, Dean 1982 [62]), das ebenfalls mit der o.g. Strategie Overreporting von Symptomen erfassen soll. In diesem Verfahren werden Paare von miteinander assoziierten Symptomen, die als sehr häufig angegeben wurden, in ihrer Antwort-Übereinstimmung erhoben (z.B. Ohrklingeln und verwaschene Sprache oder Sehstörungen und Lese-Schwierigkeiten). Damit wurde implizit auf das Konzept der Darstellungs-Inkonsistenz zurückgegriffen. Diskrepanzen in der Antwortgebung sollen hier auf Malingering hinweisen (vergleichbar der VRIN-Skala im MMPI-2).

Ergänzend konnte Larrabee (2003) [165]) demonstrieren, dass die Schmerzstärke allein nicht zwischen authentischen Patienten mit chronischen Schmerzen und Patienten mit gesicherter oder mutmaßlich überhöhter Angabe neurokognitiver Defizite unterscheiden kann. Bei hoher Spezifität von 90 % wurden mittels des McGill Pain Questionnaires (MPQ, Melzack & Torgerson 1971 [190]) nur 21 % der Probanden korrekt identifiziert, durch Überhöhungen im Pain Disability Index (PDI, Tait et al. 1987 [292]), einem Fragebogen zur Erhebung von Lebensqualitätsminderungen, wurden hingegen 59 % der Probanden mit nicht-authentischer Beschwerden-Darstellung erkannt. Allerdings erbrachte nur die Detektionsgüte des Modified Somatic Perception Questionnaire (MSPQ, Main et al. 1992 [177]) mit Erhebung diverser Somatisierungsstörungen eine Sensitivität von 90 % bei gleicher Spezifi-

tät. Die Studie unterstützte die Notwendigkeit, ungewöhnliche Symptome zur Aufdeckung nicht-authentischer Beschwerdeangaben zu untersuchen.

Ein anderer Detektionsansatz verwendet ungewöhnliche Erhöhungen der qualitativen sensorischen oder affektiven Komponente von Schmerzen als Hinweis auf eine übertriebene Beschwerdedarstellung. So konnte Windemuth (1997) [318] mit einer Genauigkeit von 86 % Beschreibungen der Schmerzqualität (Schmerzleiden, Angst, Schmerzschärfe und -Rhythmik) richtig den als glaubwürdig-antwortend eingeschätzten Patienten und den als nicht-authentisch antwortend eingeschätzten Patienten zuordnen. Auch Dohrenbusch (2009) [69] berichtet, dass eine gleichförmige Erhöhung „aller sensorischen und affektiv-emotionalen Schmerzqualitäten im Vergleich zu störungsbezogenen Normwerten“ als Hinweis auf demonstratives Herausstellen der Beschwerden gewertet werden kann. Zur Validierung können damit entsprechend differenzierte und hinsichtlich testpsychologischer Gütekriterien geprüfte Verfahren dienen, wie beispielsweise die Schmerzempfindungsskala SES nach Geissner & Schulte (1996) [98]. Dohrenbusch (2009) ergänzt, dass auch erhebliche Abweichungen ausgewählter sensorischer Merkmale der Selbsteinschätzung von untersuchten Patienten von der für das Krankheitsbild üblichen Norm (Profilabweichung) gegen die Erlebnisqualität der Schmerzen sprechen können.

Rogers (2008) [235] erläutert acht moderne Methoden der Detektion von Overreporting, die mittels diverser Fragebögen und Interviews zur Beschwerdenvalidierung durchgeführt werden können. Vor- und Nachteile dieser Methoden sowie praktische Beispiele sind in der Übersichtstabelle Tab. 1 (s. S. 58) zusammengestellt.

Hierzu zählen: (1.) Die häufige Nennung *seltener Symptome*, die sehr selten (in unter 5 % der Fälle) in klinischen Stichproben angegeben werden; (2.) *quasi-seltener Symptome*, die sehr selten von Norm-Personen angegeben werden; (3.) *bizarrer, absurder Symptome*, die extrem selten genannt werden; (4.) *nicht zueinander passender Symptom-Kombinationen*, (5.) *extremer Subskalen-Profile*, (6.) undifferenzierter Symptom-Überhöhungen, (7.) augenscheinvalider, offensichtlicher Symptome sowie (8.) fehlerhafter, nicht-stimmiger Symptom-Stereotypien.

Tab. 1. Acht Strategien zur Aufdeckung von Overreporting (nach Rogers 2008)

Aufdeckungs-Methode	Erläuterung	Beispiele	Vorteile	Grenzen
Seltene Symptome	Symptome, die sehr selten (unter 5 %) in klinischen Stichproben angegeben werden; Probanden mit Overreporting geben diese seltenen Symptome gehäuft (zu häufig) an	MMPI-2-RF Fp-r PAI-NIM	Integration in diverse Testverfahren	Patienten mit fehlender psychischer Attribution (z.B. Schmerzpatienten) vermeiden die Angabe möglichweise
Quasi-seltene Symptome	Symptome, die sehr selten (von unter 5 %) der Norm-Personen angegeben werden	MMPI-2-RF F-r	Große Effektstärken	Kann authentische Störungen ebenso wie vorgetäuschte Störungen zeigen
Bizarre, absurde, fantastische Symptome	Extreme Variante der seltenen Symptome	SIRS IA (improbable and absurde symptoms) SIMS-A	Erlaubt kaum andere Interpretation denn als Täuschung	Anfälligkeit für Verfälschung durch intelligente, strategisch vorgehende nicht-authentische Probanden
Unpassende Symptom-Kombinationen	Symptome, die in klinischen Stichproben vorkommen, aber kaum gleichzeitig; in Testteilen nicht übereinstimmend	SIRS - SC (symptom combinations) MMPI-2-RF Fb-r	Höher resistent gegen Coaching	Zur Zeit nur in strukturierten Interviews integriert
Fragwürdige Subskalen-Profile	Für Malingering charakteristische Skalen - Konfigurationen	PAI (Rogers Discriminanz Function)	Schwer zu trainieren oder zu verfälschen	Gefahr der Überinterpretation Aufgrund der Komplexität bedürfen sie besonderer Validierung
Undifferenzierte Symptom-Überhöhung	Probanden mit Overreporting überhöhen meist eine Vielzahl von Symptomen in der Intensität	MMPI-2-RF LW-r SIRS SEV (severity)	Einfache Handhabung	Führt möglicherweise schnell zu Falsch-Positiv-Klassifikationen; Mögliche Verzerrung durch echte Psychopathologie
Inkonsistenzen: Offensichtliche vs. subtile Symptome	Probanden mit Overreporting postulieren insbesondere eindrückliche, nicht-alltägliche Symptome, weil diese ihre Betroffenheit besonders zu unterstützen scheinen	MMPI-2-RF O-S-r	nach Rogers: große Effektstärken	Anwendung standardisierter Scores sinnvoll, da es sich um fünf unterschiedliche Subskalen handelt
Fehlerhafte Stereotypien	Probanden mit Overreporting haben häufig falsche Annahmen, welche Merkmale Patienten kennzeichnen	MMPI-2-RF Ds-r Dissimulation Scale PSI - Skala EPS	Resistent gegen Coaching	Nur in beiden Tests untersucht

Diese Detektions-Methoden wurden in unterschiedlichen Testverfahren als sog. Validitäts-Skalen integriert; da die meisten Methoden im MMPI-2 realisiert wurden, werden sie im Folgenden anhand der Subskalen dieses Tests zur Validierung näher erläutert.

1.9 Beschwerdenvalidierung mittels MMPI-2

Mittels des MMPI-2, des im US-amerikanischen Sprachraum am häufigsten und in seiner Vorgängerversion seit den 1940er Jahren bei Patienten mit chronischen Schmerzen am häufigsten eingesetzten Persönlichkeitstest, konnten durch die Untersuchungen von Keller und Butcher (1991) [149] vier bzw. drei für Patienten mit chronischen Schmerzen typische Test-Profile differenziert werden: neben einem Profil ohne Normabweichungen fanden die Autoren ein sog. Neurotic-Triad-Pattern (mit Auffälligkeiten in den Subskalen Hysterie Hy, Hypochondrie Hd und Depression D), das Conversion-V-Pattern (mit Auffälligkeiten in Hy und Hd), ein General-Elevated-Pattern (mit einer Vielzahl von Subskalen-Erhöhungen).

Vergleichbare Profil-Cluster wurden bereits von Cohen (1987, nach Keller & Butcher 1991 [149]) und Sternbach (1974) [283] beschrieben, wobei insbesondere das allgemeinerhöhte Profil (mit Subskalen-Auffälligkeiten im Bereich Psychopathologie Pp) im Verdacht stand, Hinweise auf testmanipulative Verfälschungstendenzen zu zeigen. Da Täuschungstendenzen auch diagnostisch zur Symptomatologie einer psychopathischen Persönlichkeitsstörung gehören (Hare und Neumann 2009 [124]), überraschen entsprechende Auffälligkeiten im MMPI-2 nicht.

Frühzeitig wurden diese speziellen Profilmuster der MMPI-Basisskalen (die sogenannte „Neurotic Triad“ und das „Conversion-V“) als typische Profilmuster von Patienten mit chronischen Schmerzen benannt, aber auch kritisiert. Franz et al. (1986) [94] beurteilten unter Einbezug der testdiagnostischen Gütekriterien die Nützlichkeit dieser Skalen wegen teilweise geringer Validität und Reliabilität hinsichtlich der Beurteilung psychosomatischer Beschwerden als gering. In ihrer Studie mit 45 Patienten mit chronischen Schmerzen mit chronischen Kopfschmerz- und 45 Rückenschmerz-Patienten zeigte sich im Vergleich mit einer Kontrollgruppe von 33 Probanden in beiden Gruppen von Patienten mit chronischen Schmerzen eine ähnliche Anzahl von Auffälligkeiten hinsichtlich somatischer Beschwerden und Angstsymptome. Beide Gruppen von Patienten mit chronischen Schmerzen schienen gleichermaßen Gefühle von Ärger und Aggression abzuwehren. Auffälligkeiten im MMPI-2

schienen somit nicht zwischen verschiedenen Patienten mit chronischen Schmerzen zu differenzieren. Messtheoretisch führen dies die Autoren auf vielfache Doppelkodierungen von Items verschiedener Subskalen zurück, ohne orthogonale Trennschärfe der Skalen, mit der Folge geringer interner Konsistenz und Validität (Franz et al. 1986 [94]).

Trotz umfangreicher Untersuchungen und Erhebungen bei Patienten mit chronischen Schmerzen (Butcher & Keller 1991 [149]) blieb deshalb die Frage nicht vollständig geklärt, welche Differenzierungs-Möglichkeiten der MMPI-2 zur Unterscheidung der Merkmale einzelner Gruppen von Patienten mit chronischen Schmerzen leisten kann.

1.9.1 Charakteristika der gebräuchlichsten MMPI-2-Validitätsskalen

Besondere Stärken der MMPI-2-Konstruktion wurden insbesondere bei der Beschwerdenvalidierung von Symptomen durch Integration einer Vielzahl testinterner Validitätsskalen gesehen.

Die Korrekturskalen L (Lie) und K (Correction) ermöglichen zunächst gemeinsam mit dem Dissimulations-Index F-K, der zwei Subskalen für Over- und Underreporting integriert, die Erfassung von sozial-erwünschten Antwortverfälschungen (faking-good) bzw. deren Abgleich mit der Häufigkeit der Angabe seltener und unüblicher Beschwerden.

Die Subskalen **VRIN** (Variable Response Inconsistency) und **TRIN** (True Response Inconsistency) legen hingegen Inkonsistenzen in der Itembeantwortung aufgrund unmotivierter, zufälliger Beantwortung oder bewusster Antwortverfälschungen offen. Die MMPI-2-VRIN-Skala überprüft bei 67 inhaltsähnlichen Items die Übereinstimmung der Antworten, die TRIN-Skala überprüft die Konsistenz der Beantwortung von 23 Item-Paaren mit quasi-umgekehrtem Inhalt.

Sozial-erwünschte Antworten können entweder bewusst als sog. Impression Management (Green 2000, 2011 [110] [113]) eingesetzt werden oder aber als Selbsttäuschung (Thies 2012 [296]). Bagby und Marshall (2004) [17] zeigten, dass sich verschiedene Validitätsskalen im MMPI-2 zur Unterscheidung beider Arten von faking-good-Tendenzen eignen (vgl. Thies 2012 [296]). Während die Lie-Skala (L), die Other-Deceptions-Scale (ODS oder Odecp, Nichols & Green 1991 [214]) und die Wiggins Social Desirability Scale (Sd, Wiggins 1959 [317]) eher *bewusste Verfälschungen* erfassen, werden mit den Validitätsskalen K (Korrek-

tur), der Superlative Scale (Butcher & Han 1995 [51]) und der Edwards Social Desirability Scale (So, Edwards 1957 [82]) *eher unbewusste Selbsttäuschungen* fehlender Pathologie identifiziert.

Die F- (Seltenheits-) Skala umfasst 60 der 64 Items der Vorgängerversion des MMPI und soll seltene somatische und psychische Symptome erfassen, die von Patienten mit Overreporting gehäuft angegeben werden. Mit einer Vielzahl von Symptomen weist die F-Skala Überlappungen mit drei klinischen Skalen (Schizophrenie, Paranoia und Psychopathie) sowie mit der Validitätsskala Fp (14 Items) auf. Die F-Skala wurde deshalb wegen ihrer mangelnden Differenzierung genuiner psychiatrischer Symptome von simulierten Beschwerdeüberhöhungen kritisiert. So fanden auch Ben-Porath et al. (2009) [28] hohe Überlappungen von Auffälligkeiten in der F-Skala mit den Basisskalen Schizophrenie und Psychasthenie.

Für die F-Skala wurden von Rogers et al. (2003) [238] Grenzwerte für sicheres Malingering bei Männern von 22 und mehr Punkten (T-Wert über 76) und bei Frauen ab Werten von 20 (T-Wert über 75) vorgeschlagen. Dies entspricht nach Butcher et al. (1989) [48] dem 95. Perzentil $\geq$ 19 Punkten für Patienten mit psychischen Störungen, bei gesunden Personen kommen solche hohen F-Werte kaum vor (99. Perzentil $\geq$ 14 Punkten).

Thies (2012) [296] ermittelte in ihrer Studie mit 240 Begutachtungs-Probanden mit Kompensationsmotiv bei einer Spezifität $\geq$ 90 % in der F-Skala bei einem Cutoff $\geq$ 15 (T64 für Männer, T66 für Frauen) eine eher geringe Sensitivität von 33,7 %. Entsprechende Daten (nach Butcher et al. 1989 [48], s. Green 2008, 174pp. [115]) belegen das 90 %-Intervall für Patienten mit psychischen Störungen in der F-Skala ab 15 Rohwertpunkten (80. Perzentil $\geq$ 10 Punkten, 99. Perzentil $\geq$ 29 Punkten).

In der Meta-Analyse von Rogers et al. (2003) [242] erreichte die F-Skala des MMPI-2 - entgegen der Kritikpunkte - die höchste Effektstärke unter den Validitätsskalen (mit d = 2,21; vgl. auch Aamodt et al. 1996 [1]). Ebenso zeigten Bagby et al. (1995) [15] im Vergleich der Angaben instruierter Studenten gegenüber den Einschätzungen psychiatrischer und forensischer stationärer Patienten eine Überlegenheit der F-Skala zur Aufdeckung von fake-bad, während die L-, die O-S- und die Mp-Skala (Positiv Malingering) fake-good-Tendenzen signifikant detektieren konnte. In einer Vergleichsstudie von simulierenden Häftlingen mit stationär behandelten psychiatrischen Patienten erwiesen sich die F-Skala und die Fp-Skala als am besten differenzierende Validitätsskalen des MMPI-2 zur Aufdeckung von Overrepor-

ting (Steffan et al. 2003 [282]). Hingegen ermittelte Thies (2012) [296] für die F-Skala eine eher mäßige Effektstärke (d = 0,73).

Meyers et al. (2002) [202] zeigen zudem in ihrer Studienübersicht bei hoher Spezifität (100 %) und Sensitivitätswerten zwischen 69 und 81 Prozent, dass T-Werte über 65 bis 70 in der F-Skala des MMPI-2 eine relativ trennscharfe Detektionssicherheit von Malingering bieten.

Arbisi und Ben-Porath (1995) [12] berichteten jedoch (vgl. auch Ben-Porath et al. 2009 [28]), dass psychiatrische Patienten in hohem Maß ebenfalls überhöhte Angaben in der MMPI-2-Skala F machen, so dass die Abgrenzung zwischen Malingering und echter Pathologie mittels dieser Skala kritisch ist. Meyers et al. (2002) [202] legten hier einen Cutoff für wahrscheinliches Malingering von 75 bis 89 T-Werten und für sicheres Malingering einem Wert von mehr als 90 fest.

Die F-Skala (zum Assessment ungewöhnlicher, seltener Symptome) erwies sich in den meisten Studien als die trennschärfste Validitätsskala des MMPI, mit eingeschränkter Sensitivität bei genuin psychiatrisch-auffälligen Patienten.

Die Fp-Skala seltener psychischer Symptome beinhaltet 27 MMPI-2-Items so ungewöhnlicher psychischer Symptome, dass maximal 20 Prozent psychiatrischer Patienten diese als zutreffende Selbstbeschreibung wählen. 15 der 60 F-Skalen-Items sind der Fp-Skala zugeordnet. Fp soll damit die Simulation schwerer psychologischer Symptome aufdecken, ihre Validität zur Aufdeckung subtilerer Verfälschungs-Strategien wird jedoch kritisch diskutiert.

Grenzwerte der Fp-Skala können als lineare T-Werte für Männer und Frauen berechnet werden; beispielsweise nennen Arbisi & Ben-Porath 1995 [12] für Männer einen Mittelwert = 1,2 (SD = 1,4) und für Frauen einen Mittelwert = 1,1 (SD = 1,25). Meyers et al. (2002) [202] legten entsprechend einen Cutoff für wahrscheinliches Malingering ab 75 T-Wertpunkten (mehr als 3 Rohscores) und für sicheres Malingering ab 89 Punkten (Rohwerte für Männer ab 8 Punkten, für Frauen ab 9 Punkten) fest. Entsprechend bestätigten Millis et al. (1995) [204] eine mäßig sichere Detektionsgenauigkeit der Skala bei einem Cutoff

von 93 T-Wertpunkten, die bei 90 % - Spezifität eine 60-prozentige Sensitivität zeigte. In einer anderen Studie (Arbisi & Ben-Porath 1998 [13]) zeigte sich ab 90 T-Wertpunkten bei 90-prozentiger Spezifität eine höhere 97-prozentige Sensitivität.

Thies (2012) [296] berechnete bei einer Spezifität $\geq$ 90 % in der Fp-Skala einen Cutoff $\geq$ 7 mit erneut sehr geringer Sensitivität von 14 %. Dieser Cutoff entspricht nach Butcher et al. 1989 [48] dem 98 - Perzentil $\geq$ 7 Punkten für Patienten mit psychischen Störungen, bei gesunden Personen kommen solche hohen Fp-Werte kaum vor (99 - Perzentil $\geq$ 5 Punkten).

Arbisi & Ben-Porath (1995) [12] sprachen sich trotz der bei ihnen gefundenen Resultate nicht für die generelle Ersetzung der F-Skala durch Fp aus, sondern sahen diese Skala eher als ergänzendes Hilfsmittel („adjunct to the interpretation of f“, Arbisi & Ben-Porath 1995, 430pp.). Die Autoren argumentierten, dass Überhöhungen in beiden F-Skalen ein relativ sicherer Hinweis für Simulation sei; wenn hingegen nur die F-Skala erhöht sei, sei dies möglicherweise eher auf tatsächliche psychische Störungen zurückzuführen.

Arbisi & Ben-Porath schätzten die Sensitivität bei einem Cutoff $\geq$ 8 auf 96 % bei üblicher Spezifität von 90 %. In einer anderen Studie (Strong et al. 2000 [285]) lag bei einem niedrigeren Cut-Punkt $\geq$ 6 die Sensitivität bei nur 0,62 bei höherer Spezifität (0,99). Dies würde allerdings bedeuten, dass in einer Begutachtung 40 % der tatsächlich Symptome simulierenden Patienten nicht richtig klassifiziert würden. Die Trennschärfe wurde in der Metaanalyse von Rogers et al. (2003) [242] mit Cohen's d von 1,90 als effektiv bewertet. Thies (2012) [296] ermittelte hingegen für die Fp-Skala eine vergleichsweise sehr niedrige Effektstärke (d = 0,35).

Die Fp-Skala (zum Assessment seltener psychopathologischer Symptome) bietet als Ergänzung der F-Skala eine gewisse Absicherung zur Diagnose von Beschwerdeüberhöhungen, die nicht nur auf einer für psychisch kranke Patienten „normalen“ oder „typischen“ Aggravation beruhen. Patienten, bei denen eine Verfälschungstendenz angenommen wird, sollten in beiden Skalen auffällig sein.

Die Fake Bad Scale (FBS) wurde entworfen, um unglaubwürdige somatische und kognitive Symptome bei Patienten zu identifizieren, die ihre Berichte künstlich und geschickt als authentisch verfälschen, um plausibel auf körperlicher Ebene behindert zu scheinen, ohne Hinweise auf eine zugrundeliegende Psychopathologie.

Nach Lees-Haley et al. (1991) [169] zeigen Rohwerte oberhalb 20 ein Overreporting bei der Reklamation von Unfallschäden an. Nach Lees-Haley et al. (1992) [168] liegen die auffälligen Werte bei Männern über einem Rohwert von 23, bei Frauen ab einem Summenwert von 26.

In der Original-Studie wurden 96 Prozent als simulierend klassifizierter Probanden richtig identifiziert (Sensitivität) und 90 % der als glaubwürdig eingestuften Patienten korrekt erkannt (Spezifität). Es ergaben sich niedrigere Durchschnittswerte bei gesunden Probanden (Rohwerte 12-14) und ebenso bei tatsächlich beeinträchtigten Patienten (Rohwerte 19-17) im Vergleich zu den als simulierend eingestuften Personen (Rohwerte über 25). Thies (2012) [296] ermittelte bei einer Spezifität $\geq$ 90 % in der FBS-Skala ab einem Cutoff $\geq$ 27 in ihrer Studie eine niedrige Sensitivität von 36 %.

Eine neuere Metaanalyse über 32 Studien zur MMPI-2-Validitätsskala FBS aus fast zwei Jahrzehnten bestätigte eine höhere Detektionsgüte dieser Skala (Spezifität 92,5 %, Sensitivität 66,7 % bis 100 %) zur Aufdeckung von Malingering im Bereich psychischer Störungen (Nelson et al. 2010) [213]. Andererseits zeigten Butcher et al. (2003) [47], dass mittels der FBS-Skala insbesondere bei psychiatrischen Patienten eine Vielzahl falsch positiver Klassifikationen von Malingering ermittelt wird, so dass die Autoren diskutierten, ob die FBSPunktenSkala vielleicht eher allgemeines Missempfinden und somatische Beschwerden erfasse als Antwortverfälschungen.

Meyers et al. (2002) [202] legten für die Fake-Bad-Skala einen Schwellenwert für Probably-Malingering von 25 bis 29 Rohwert-Wertpunkten fest und für sicheres Malingering einen Wert oberhalb 29 (entsprechend auch Ben-Porath et al. 2009 [28]). Rogers et al. (2003) [242] ermittelten eine relativ niedrige Effektstärke (Cohen's d = 0,32) für die FBS, bei Thies (2012) [296] fiel diese höher aus (Cohen's d = 0,92).

Trotz Kritik an dieser Validitätsskala (heterogene Konstruktion, geringe interne Konsistenz, hohe Angaben psychiatrischer Patienten in dieser Skala, s.a. Thies 2012 [296]), scheint

sie für die Detektion von Overreporting somatischer und auch psychischer Symptome bei moderat beeinträchtigten Personen durchaus geeignet.

Die FBS-Skala, die zur Aufdeckung besonders geschickter Verfälschungstendenzen von Patienten mit Unfallschäden entwickelt wurde, steht in demselben Verdacht wie die F-Skala, möglicherweise zu viele, für Patienten mit realer Psychopathologie typische Verfälschungen zu erfassen (Gefahr von Fehl-Klassifikationen). Bei moderat beeinträchtigten Patienten scheint die FBS dennoch Vorteile zu bieten.

Die Response Bias Scale (RBS) nach Gervais et al. 2007 [99]) wurde als 28 Items umfassende Validitätsskala des MMPI-2 vor allem dazu konzipiert, Malingering auf der kognitiven Ebene aufzudecken. Dazu wurden jene Items identifiziert, die in einer Probandengruppe mit Kompensationsmotiv trennscharf Probanden mit und ohne Verfälschungstendenz in einer Anzahl von BV-Verfahren differenzieren konnten. Gervais et al. 2010 [101] belegten an 1287 Probanden einer neurologisch-psychiatrischen Praxis, die auch den WMT (Word Memory Test, s. S. 35) als BV-Verfahren ausgefüllt hatten, dass sich die Validitätsskalen des MMPI-2-RF zur Vorhersage der Performance der Untersuchten im MCI (Memory Complaints Inventory) besser als die MMPI-2-Validitätsskalen F, Fp und FBS eigneten. Eine weitere Analyse bestätigte, dass die RBS sich in 12 % bis 32 % der Fälle besser zur Prädiktion von Klagen von Gedächtnisstörungen eignet als die MMPI-2-RF-Skalen F-r, Fp-r, Fs and FBS-r. Ähnlich bestätigten Whitney et al. (2008) [313], dass sich die RBS-Skala zur Vorhersage von Auffälligkeiten im TOMM (Test of Memory Malingering, s. S. 35) eignete.

Für die RBS des MMPI-2 wurde von Gervais et al. (2007) [99] ein Cutoff von ≥ 17 ermittelt, der bei einer Spezifität von 95 % eine relativ niedrige Sensitivität von 25 % aufwies. In der MMPI-2-Normierungsstichprobe lagen Mittelwerte für psychiatrische weibliche Patienten bei 19,1 und bei 16,9 für Männer. Rogers et al. (2003) [242] ermittelten eine ebenfalls eher geringe Effektstärke (Cohen's d = 0,32), bei Thies (2012) [296] wurde eine höhere Trennschärfe (Cohen's d = 0,92) ermittelt. Thies (2012) [296] berechnete bei einer Spezifität

$\geq$ 90 % in der RBS-Skala bei einem Cutoff $\geq$ 16 eine ebenfalls nicht sehr hohe Sensitivität von 27,9 %.

Zur Erhebung kognitiven Malingerings (Gedächtnis-, Aufmerksamkeits- und Konzentrationsstörungen) scheint sich insbesondere die RBS-Skala zu eignen, wie die Analysen zu Auffälligkeiten in externen PVTs (z.B. TOMM, Whitney et al. 2008 [313]) zeigten.

Der Henry-Heilbronner-Index (HHI) wurde als weitere Validitätsskala entwickelt (Henry et al. 2006) [132], der im eigentlichen Sinn keine Subskalen wie ein echter Index subsummiert, sondern eine 15 MMPI-2-Items umfassende Subskala ist. Die HHI-Items, die nach den Autoren sog. „pseudosomatische Symptome“ erfassen sollen, wurden aus der Fake Bad Scale und der Pseudoneurological Scale (Shaw & Matthews 1965 [274]) entwickelt. Der HHI identifizierte in der Originalstudie Probanden, die nach einem Personenschaden auf Entschädigung klagten und Auffälligkeiten in mindestens einem von vier bewährten BV-Verfahren aufwiesen im Gegensatz zu Probanden ohne Auffälligkeiten im Leistungsbemühen. Mittels des HHI konnte 85,6 Prozent der Untersuchten mit einer Spezifität von 89 % und einer Sensitivität von 80 % richtig klassifiziert werden.

In der Studie von Thies (2012) [296] mit 240 Begutachtungs-Probanden mit Kompensationsmotiv erwies sich der HHI bei einem Cutoff-Wert von $\geq$ 12 mit einer Effektstärke von 1,02 als die trennschärfste MMPI-2-Validitätsskala. Aus Gründen der Test-Sicherheit wurde die Itemliste zudem nicht von den Autoren veröffentlicht, mit dem Nachteil limitierter Anwendung. Die Effektstärke des HHI wurde von Thies (2012) [296] mit 1,02 (Cohen's d) ermittelt, womit sich dieser Validitäts-Score in dieser Studie als bester Indikator für Overreporting erwies.

Tshushima et al. (2011) [300] beschrieben bei einer Gruppe von 163 Rentenantragstellern bei einem Cutoff von mehr als 11 bei hoher Spezifität (90,2 %) eine eher niedrige Sensitivität (27,1 %). Unter Anwendung des in der Literatur teilweise genannten Cutoff $\geq$ 9 ergab der HHI-Score eine Spezifität von 77 % und eine Sensitivität von 52 %, die jedoch höher lag

als für die RBS (Spezifität von 90,8 %, Sensitivität von 39,0 %) und auch für die FBS (bei Cutoff ≥ 23 mit einer Spezifität von 91,4 % und einer Sensitivität von nur 29,7 %).

Bei Forderung einer hohen Spezifität (von mehr als 90 %) zeigte der ebenfalls kognitives Malingering erfassende Henry-Heilbronner-Index vergleichbare Detektionsergebnisse wie die RBS (ca. 30 % Sensitivität). Da die HHI-Skala mit der RSB-Skala keine gemeinsamen Items teilt, bieten sich beide Skalen zur Kreuzvalidierung an.

Die Skalen Kritischer Items nach Lachar und Wrobel (1979) [162] mit 107 Items und von Koss & Butcher (1973) [159] mit 82 Items sind Sammlungen skalenübergreifender Items, die augenschein-valide psychische Symptome erfassen. Ihr Rational zur Erhebung von Antwortverzerrungen beruht ähnlich wie bei den Obvious-Subtle-Scales (O-S, Wiener (1948) [315], S. 71) darauf, dass Patienten mit Tendenz zur Überhöhung von Symptomen diese Items besonders häufig befürworten.

Green (2008, 172pp.) [115] berichtet, dass gesunde Probanden in durchschnittlich 15 der LW-Items und 10 der KB-Items überhöhte Angaben machen. Dies zeigte nach Ansicht des Autors, dass möglicherweise nicht alle der Items trennscharf Malingering erfassen. Nach Daten von Butcher et al. 1989 [48] (s. Green 2008, 174pp. [115]) beginnt bei gesunden Probanden das 75 %-Intervall für die LW-Skala bei 23 Rohwertpunkten (80 % bei 25 Punkten, 90 % bei 31 Punkten). Bei psychisch kranken Patienten beginnen auffällige Werte der LW-Skala bei 46 Punkten (80. Perzentil-Intervall, 56 ab 90. Perzentil-Intervall). Dies ermöglicht eine gewisse Abschätzung der Grenzwerte für Overreporting.

Thies (2012) [296] berechnete bei einer Spezifität ≥ 90 % für die Lachar-Wrobel-Items einen Cutoff ≥ 49 mit eher geringer Sensitivität von 34,9 % (entsprechend dem 99. Perzentil bei Butcher et al. 1989 [48] für gesunde Probanden und dem 85. Perzentil-Intervall für psychisch Kranke).

Entsprechende Daten (nach Butcher et al. 1989 [48], s. Green 2008, 174pp. [115]) belegen das 75 %-Intervall bei gesunden Probanden für die KB-Skala ab 16 Rohwertpunkten (80 %

$\geq$ 18 Punkten, 90 % $\geq$ 23 Punkten, 99. Perzentil $\geq$ 38 Punkten). Psychisch Kranke zeigten deutlich höhere Grenzwerte (80 % $\geq$ 35 Punkten , 90 % $\geq$ 44 Punkten, 99. Perzentil $\geq$ 61 Punkten)

Da die Effektstärke der sog. Kritischen Items nach der Meta-Analyse von Rogers et al. (2003) [238] eine höhere Effektstärke der Detektionsgüte für Malingering (Cohen's d = 1,27) erst ab einem Summenscore über 79 aufwies, wird die Validität dieser Listungen zur Detektion von Overreporting als unsicher diskutiert (vgl. Thies 2012 [296]). Rogers et al. (2003) [242] ermittelten eine Effektstärke von 1,27 für die Sensitivität der LW-Items, bei Thies (2012) [296] wurde eine relativ hohe Trennschärfe mit Cohen's d von 0,92 ermittelt.

Aufgrund häufiger Auffälligkeiten auch psychisch-kranker Patienten in den sog. kritischen Item-Listen wird die Detektionsgüte dieser Skalen (Lachar-Wrobel, Koss-Butcher) nur bei hinreichend hohen Grenzwerten als aussagekräftig zur Aufdeckung von Overreporting bewertet.

Der F-K-Index oder Dissimulations-Index nutzt eine elaborierte Detektionsstrategie zur Aufdeckung von Overreporting. In diesem Score wird das Ausmaß überhöhter Beschwerden mit dem Grad bagatellisierter Symptome (Underreporting) summatorisch verrechnet. Dahinter steht die Überlegung, dass Patienten mit nicht authentisch geschilderten Beschwerden durch besonders hohe F-Werte und besonders niedrige Korrektur-Werte auffallen werden.

Der F-K-Index zeigt nach Gough (1950) [105] bei gesunden Probanden oft Mittelwerte unter 0; Index-Mittelwerte von simulierenden Probanden zeigten Werte über 0. Für Malingering definierte Gough einen F-K-Index $\geq$ 2 als Cutoff, wobei andere Autoren auch weit höhere Cutoff-Werte als Detektionsmarke angeben, zwischen Null (Lees-Haley 1991) [167] und 27 Punktwerten (Graham et al. 1991) [108]. Meyers et al. (2002) [202] legten studienübergreifend einen Cutoff für wahrscheinliches Malingering von 1 bis 9 Punktwerten und für sicheres Malingering einen Punktwert über 9 fest.

Rogers et al. (2003) [242] nannten eine sehr hohe Effektstärke (Cohen's d = 1,98) für die F-K-Skala, bei Thies (2012) [296] wurde eine ebenfalls relativ hohe Trennschärfe (Cohen's

d = 0,81) ermittelt. Hingegen ermittelte Thies (2012) [296] im F-K-Index bei einem Cutoff ≥ 5 bei einer Spezifität ≥ 90 % eine relativ geringe Sensitivität von 25,6 % (nach Butcher et al. 1989 [48] entsprechend dem 98. Perzentil mit Rohwerten ≥ 4 bei gesunden Probanden, für psychisch Kranke entsprach dieser Grenzwert dem 90. Perzentil-Intervall).

Der F-K-Index könnte konzeptuell besondere Qualitäten in der Aufdeckung von nicht-authentischen Schilderungen aufweisen, da in ihm zwei Detektionsansätze (Over- und Underreporting) synergistisch verwendet wurden.

Die Validitätsskala Dissimulating Scale (Ds) nutzt den Ansatz sog. fehlerhafter Stereotypien, die simulierende Laien eher fälschlicherweise benennen als Patienten, die diese Symptome nicht haben (z.B. "Man zeigt mir kein Verständnis", obwohl häufig Schmerzverhalten operant durch Aufmerksamkeit aufrechterhalten wird).

Nach Gough (1954) [106] soll die Ds-Skala aufgrund ihrer für Neurotizismus charakteristischen Items zwischen Simulanten und Patienten differenzieren. Diese Skala ist im MMPI-2 mit 58 Items enthalten, aber auch als verkürzte Ds-r-Skala mit 32 (der von Gough 1957 [107] benannten 40) Items einsetzbar.

Eine Mehrgruppen-Erhebung bei Patienten fiel erwartungsgemäß geringer aus als bei High-School-Absolventen, die neurotische Symptome simulieren sollten (vgl. Thies 2012 [296]). Thies (2012) [296] ermittelte bei einer Spezifität ≥ 90 % in der Ds-Skala bei einem Cutoff ≥ 24 eine wiederum relativ geringe Sensitivität von 29,1 % (entsprechend dem 98. Perzentil mit Rohwerten ≥ 24 bei gesunden Probanden und dem 85. Perzentil bei psychisch Kranken, nach Butcher 1989 [48]).

Nach Rogers et al. (2003) [238] lag die Effektstärke der Ds-Skala höher als die der verkürzten Skala, während Nelson et al. (2010) [213] eine fast doppelt so hohe Effektstärke der Ds-r-Skala berichteten (Cohen's d = 1,62), bei Thies (2012) [296] wurde eine in dieser Studie ebenfalls höhere Trennschärfe (Cohen's d = 0,80) ermittelt.

Nach der Literatur-Meta-Analyse von Meyers et al. (2002) [202] wurden bereits T-Werte über 70/72 (Rogers et al. 1993 [236], Bagby et al. 1994 [20]) mit Spezifitätswerten zwischen

86 und 100 als Grenzwerte für Malingering mit einer über 80-prozentigen Sensitivität berichtet. Meyers et al. (2002) [202] legten in konservativer Definition für die Dissimulating Scale Ds einen relativ hohen Cutoff für wahrscheinliches Malingering von 75 bis 89 T-Werten und für sicheres Overreporting oberhalb dieser Grenze fest.

Der in der Ds-Skala genutzte Detektionsansatz (bei Simulanten gehäuftes Vorkommen fehlerhafter Stereotypien) erwies sich gegenüber anderen Verfahren als besonders resistent gegen Coaching.

Die Infrequency Back Skala (Fb) analysiert die Angaben von Befragten anhand 40 Items seltener psychopathologischer Symptome, die in der zweiten Testhälfte des MMPI-2 abgefragt werden. Durch diese Skalen-Konstruktion sollte die Fb-Skala insbesondere zur Erhebung von Antwort-Inkonsistenzen dienen, die jedoch auch möglicherweise auf Unaufmerksamkeit über den Gesamttest hinweg bedingt sein könnte.

Rogers et al. (2003) [238] nennen Cutoff-Werte von 18 bzw. 19 Punkten als Indikatoren für Simulation, wobei fraglich ist, ob diese Skala auch Testaufmerksamkeit als Malingering erfasst oder andere Konstrukte (z.B. wiederum genuine psychische Störungen). Ben-Porath et al. (2009) [28] bemerkten beispielsweise hohe Fb-Scores insbesondere bei psychisch auffälligen Patienten.

Thies (2012) [296] berechnete in der Fb-Skala einen Cutoff ≥ 10 mit eher geringer Sensitivität von 29,1 % bei einer Spezifität $\geq$ 90,5 %. Rogers et al. (2003) [242] ermittelten eine relativ hohe Effektstärke (Cohen's d = 1,62) für die Fb-Skala. Bei Thies (2012) [296] wurde sogar eine Diskriminanzgüte ermittelt, die diskret über der Trennschärfe der F-Skala (Cohen's d = 0,75) einzuordnen ist.

In einer Studie von Bagby et al. (2000) [22] (s. S. 73) erwies sich die Fb-Skala gegenüber der F-Skala und der Dissimulating Scale (Ds) ebenfalls als die eindeutig trennschärfste Skala, um Depressions-Symptome authentisch antwortender Patienten von Experten, die eine Depression simulierten, zu unterscheiden. Fb-Werte $\geq$ T89 erreichten bei einer Spezifität von 85 % eine falsch negative Rate von nur 9 % (Sensitivität 91 %). Eine noch höhere Spezi-

fität (90 %) bei gleicher Sensitivität erreichte nur die Kombination aus den Skalen F und Fb als Summenscore.

Insofern könnte sich die *Fb-Skala als ein besonders trennscharfes Detektionsinstrument für Malingering* in weiteren Studien erweisen.

Die O-S (Obvious-Subtle-Scales) nach Wiener (1948) [315] ermöglichen mit 110 augenscheinlichen Items und 146 Items versteckter, subtiler Symptomabfragen durch deren Abgleich eine Aufdeckung von Over- und Underreporting. Diese Skalen folgen dem Rational, dass Probanden, die Beschwerden simulierend überhöhen, durch offensichtliche Symptome verstärkt motiviert werden, diesen häufiger zuzustimmen (vgl. Rogers et al. 2003a [242]).

Nach Green (2000) [110] gilt als Cutoff-Regel für ein Overreporting in den O-S-Skalen: Wenn die T-Werte für die offensichtlichen Items über 70 liegen, während die T-Werte der subtilen Subskalen-Items unter 50 liegen, ist eine *fake-bad*-Tendenz anzunehmen; bei umgekehrten Verhältnissen liegt eine sog. *faking-good*-Tendenz vor.

Problematisch ist, dass bei den Original-Analysen von Wiener (1948) [315] subtilere Symptome in früher ebenfalls zur Diagnostik angewandten Basis-Skalen von signifikant mehr höher gebildeten Probanden korrekt identifiziert wurden, weniger gebildete Probanden erkannten subtile Items nicht ausreichend. Insofern könnte die Abfrage *bildungsabhängig* zu Lasten höher gebildeter Probanden verfälschbar sein.

Die Schwierigkeit, geeignete Cutoff-Werte zur Beurteilung der Diskriminationsgüte eines Validitäts-Scores zu finden, sei hier ergänzend am Beispiel der Obvious-Subtle-Scales erläutert.

Während Meyers et al. (2002) [202] in ihrem MMPI-2-Validity-Gesamt-Index (vgl. Kap. 1.9.4, S. 78) einen Cutoff für Malingering bei Patienten mit chronischen Schmerzen bei T-Wert-Summenscores der fünf O-S-Skalen über 100 bis 149 T-Werten annehmen, konstatierten Rogers et al. (1994) [246] O-S-Differenzen über 80 bereits als Overreporting. Sivec et al. (1994) [277] berichten, dass Probanden, die eine Somatoforme Schmerzstörung simulieren, meist sogar O-S-Differenzen von mehr als 150 angeben (bei Sivec et al. exakt O-S-Differenzen ca. 172,9, SD = 83,1).

Auch Meyers et al. 2002 [202] berichteten, dass Patienten mit externen Kompensationsmotiven häufig T-Wert-Differenzen der O-S-Skalen von über 90 (SD = 89,0) zeigten, wäh-

rend Patienten mit primär neurologischer Symptomatik und damit assoziierten Schmerzen ohne Rentenantrag im Durchschnitt T-Wert-Differenzen unter 30 (SD = 61,8) aufwiesen. Dush et al. (1994) [77] war es mit Hilfe der T-Wert-Differenzen der O-S-Skalen möglich, 43 Patienten mit chronischen Schmerzen mit Rentenantrag und 45 Nicht-Rentenantragsteller signifikant korrekt zu differenzieren. Hierbei sei angemerkt, dass allein die Tatsache eines Berentungswunsches als Anlass für Overreporting angenommen wurde (dies kann nach jüngeren Studien nicht pauschal ohne Hinzuziehung weiterer Außenkriterien als hinreichendes Indiz für Overreporting konstatiert werden, Bianchini et al. 2008 [30]).

Hollrah et al. (1995) [137] berichteten, dass bei instruierten Simulanten psychopathologischer Symptome die O-S-Skalen Malingering in der erwarteten Richtung der Verfälschung anzeigen: Bei Probanden, die erhöhte Psychopathologie simulieren sollten, waren die Obvious-Items erhöht und die Subtle-Items deutlich verringert; dem gegenüber war bei Probanden, die instruiert wurden, eine psychische Störung eher zu bagatellisieren, das Verhältnis der O-S-Skalen-Werte reziprok.

Thies (2012) [296] ermittelte bei einer Spezifität $\geq$ 90 % in der O-S-Skala bei einem Cutoff $\geq$ 104 eine Sensitivität von 32,6 %. Dies entspricht nach Butcher et al. (1989) [48] (s. Green 2008, 174pp. [115]) dem 75. Perzentil $\geq$ 97 Punkten für Patienten mit psychischen Störungen, bei gesunden Personen kommen solche hohen Werte weit seltener vor (95. Perzentil $\geq$ 105 Punkten).

In einer großen Metaanalyse (Rogers et al. 2003 [242]) wurde eine hohe Effektstärke (Cohen's d = 1,51) für die O-S-Skalen festgestellt, bei Thies (2012) [296] fiel die Trennschärfe ähnlich der FBS und des HHI mit Cohen's d von 0,92 auf. Andere Autoren hingegen, einschließlich der deutschen Normierungsgruppe für den MMPI-2 (s. Butcher et al. 2000 [49]), werten die Testgüte der Obvious-Subtle-Scales aktuell noch als zu unsicher und empfehlen deren Verwendung als noch experimentelles Hilfsmittel nur unter Vorbehalt.

Inkonsistenzen und Widersprüche in der Antwortgebung werden mittels der Fb-Skala durch Abgleich ähnlicher Abfragen zu unterschiedlichen Zeiten der Testung (Testhälften) überprüft. Damit ermöglicht diese Skala, vergleichbar

den Obvious-Subtle-Scales, einen verdeckten Abgleich miteinander zusammenhängender Beschwerden.

Die Infrequent Somatic Complaint Skala, Fs-Skala wurde von Wygant, Ben-Porath und Arbisi (2006) [321] auf empirischer Basis mittels Sichtung von mehr als 55.000 MMPI-2-Protokollen von Patienten mit chronischen Schmerzen aus ungewöhnlichen somatischen Symptomen konstruiert. Trotz ihrer relativ großzügigen Auswahl (Vorkommen bei weniger als 25 Prozent der Probanden) lies sich mit dieser Skala eine große Detektions-Effektstärke belegen (vgl. Green 2008 [115]). Ihre Weiterentwicklung erfolgte parallel zu Fortentwicklungen des MMPI-2.

1.9.2 Weniger gebräuchliche und spezielle Validitätsskalen

Als Spezial-Validitätsskala zur Aufdeckung von *Overreporting depressiver Symptome* wurde die **Malingered Depression Scale (Md)** von Steffan et al. (2003) [282] entwickelt. In der Entwicklungsstudie konnten 32 MMPI-2-Items identifiziert werden, die zwischen Studenten mit leichter Depression und Studenten, die Depressionssymptome simulierten, trennscharf differenzieren (Cohen's d = 1,04 für nicht instruierte Simulanten, d = 0,80 für Simulanten mit Coaching). Die Diskriminanzgüte der Md-Skala war der Detektionsgenauigkeit der Validitätsskalen Fb und F vergleichbar (Bagby et al. 2005 [16]). Zu kritisieren ist jedoch die Validierung an Studenten mit eher leichter Symptomatik (Thies 2012 [296]). Thies (2012) [296] ermittelte bei einer Spezifität $\geq$ 90 % in der Md-Skala bei einem Cutoff $\geq$ 20 eine Sensitivität von 30,2 %. Die Effektstärke der Md-Skala wurde bei Thies (2012) [296] mit Cohen's d von 0,75 ermittelt, was einer mittleren Trennschärfe in dieser Studie entsprach.

In diesem Zusammenhang untersuchten Bagby et al. (2000) [22], ob es 23 ausgewählten Experten möglich sei, den MMPI-2 so zu verfälschen, als ob sie eine Depression hätten, und diese Verfälschung nicht von tatsächlich Depressionskranken zu unterscheiden sei. Die Ergebnisse zeigten, dass die Skala F, die Fb-Skala und die Dissimulating Scale (Ds) signifikant zwischen authentischen und der von Experten angegebenen Symptomatik differenzieren konnten. Dabei erwies sich die Fb-Skala gegenüber den anderen beiden Validitätsskalen

als am trennschärfsten. Die Studie belegte zudem, dass selbst Experten nicht in der Lage waren, einer Detektion von Overreporting durch die Validitätsskalen zu entgehen.

Als weitere spezielle Validitätsskala des MMPI-2 ist die 15 Items umfassende **Malingering Mood Disorder Scale** zu nennen (Henry et al. 2008) [133], in der die Autoren des Henry-Heilbronner-Index mit einer in ihrer Studie verwandten Methodik (Henry et al. 2006 [132]) besonders trennscharfe MMPI-2-Items identifizierten, die von als Simulanten eingeschätzten Probanden distinkt höher als von als authentisch beeinträchtigten Probanden beantwortet wurden. Der von den Autoren angegebene Cutoff (≥ 8) wurde mit einer Spezifität von 100 % und einer Sensitivität von 46,8 % eingeschätzt, was allerdings gleichzeitig bedeuten würde, dass jeder zweite, Depressionssymptome simulierende Proband mit der Skala nicht identifiziert würde. Thies (2012) [296] ermittelte bei einer Spezifität ≥ 90 % in der MMDS bei einem Cutoff ≥ 11 eine Sensitivität von 34,9 %.

Die Trennschärfe der von Elhai et al. (2002) [84] entwickelten Validitätsskala **Infrequency Posttraumatic Stress Disorder Scale (Fptsd)** mit 32 MMPI-2-Items, die von nur 20 Prozent der untersuchten Patienten mit posttraumatischer Belastungsstörung (PTBS) angegeben wurden, beurteilte Thies (2012) [296] als begrenzt, zumal nur männliche Soldaten mit und ohne finanzielle Kompensation in die Validierung einbezogen wurden. Zudem zeigte die Fptsd-Skala im Vergleich zu der ebenfalls ausgewerteten Fp-Skala eine geringere Diskriminanzgüte, um zwischen als authentisch eingestuften Probanden und den Probanden mit diagnostiziertem Overreporting zu unterscheiden. In der Entwicklungsstudie zur Fptsd-Skala fehlte zudem eine externe Beschwerden-Validierung.

Giger et al. (2010) [102] zeigten, dass 40 von 60 experimentellen Simulanten, die ein Gewaltdelikt inszenierten, nach entsprechender Vor-Instruktion eine tatbezogene Amnesie erfolgreich vortäuschen konnten. In ihrer Untersuchung setzen Giger et al. zwei spezielle Fragebögen zur Abfrage dissoziativer Symptome bei posttraumatischen Belastungsstörungen ein (Fragebogen zu Dissoziativen Symptomen FDS und Peritraumatic Dissociative Experiences Questionnaire PDEQ). Die simulierenden Probanden demonstrierten zwar ohne Coaching erkennbar höhere Symptome als authentisch antwortende Probanden. Wurden die experimentellen Simulanten jedoch vorab instruiert, Symptome nicht zu stark zu übertreiben, konnten sie ihr Antwortverhalten, insbesondere im FDS, deutlich weniger auffällig gestalten. Die Autoren halten deshalb Selbstbeurteilungsfragebögen ohne externe Beschwerdenvalidierung für ungeeignet, um sicher tatbezogene Amnesien zu identifizieren.

Als Spezial-Skalen zur Einschätzung überhöhter Depressions-Angaben sowie Angabe von Symptomen einer Posttraumatischen Belastungsstörung wurden die Prüfskalen Md, MMDS sowie Fptsd entwickelt. Alle Skalen wiesen bei einer Spezifität von 90 % Sensitivitätswerte von unter 50 % auf, was in Gutachtenkontexten eher gehäufter Falsch-Negativ-Klassifikationen entspräche.
Hier wäre zu prüfen, ob diese Skalen in modifizierter Form im MMPI-2-RF bessere Detektions-Qualitäten aufweisen.

1.9.3 MMPI-2-Validitätsskalen zur Erfassung von Underreporting

Underreporting von Symptomen scheint eher mit erfolgreicherer Bewältigung von Schmerzbeschwerden assoziiert zu sein. So beobachteten beispielsweise Komarahadi et al. (2003) [155] bei Patienten mit sozial erwünschtem Antwortverhalten ein intensiveres Bewältigungsverhalten und retrospektiv ein geringeres Schmerzintensitätserleben.

Zur Erhebung von Underreporting wurde seit der MMPI-Urfassung die auch in den Nachfolge-Versionen verwendete *L-(Lügen-)Skala* konzipiert, mit der sozial erwünschte Antworttendenzen erfasst werden sollen. Die L-Skala beinhaltet 15-Items, die eine positive Selbstdarstellung (Positive Impression Management, vgl. PAI - PIM-Skala, s. S. 45) und symptom-abwehrendes Antwortverhalten (Defensiveness) erfassen.

Thies (2012) [296] kritisiert, dass diese Subskala durch die generelle Antworthaltung des Befragten beeinflusst sein kann, da im MMPI-2 nur Items einer Polierung summiert werden. Nach Green (2000) [110] wird die Skala zudem von intelligenten Probanden leicht durchschaut und ist anfällig für Coaching. Bei Grenzwerten von 7 (männliche Normgruppe), 6 (weibliche Normgruppe) und 9 (psychiatrische Patienten) ergab eine Validitätsanalyse von Bear & Miller (2002) [14] eine mäßige Effektstärke (Cohen's d = 1,19), bei Grenzwerten $\geq$ 9 (T64) konnten die Autoren Testverfälschungen bei instruierten Probanden feststellen.

In der MMPI-2 L(Lie)- und der K(Correction)-Skala stellten Rogers et al. (2003a) [242] eine eher geringe Detektionsgüte für Defensiveness und Underreporting fest und mutmaßten, dass der Grund dafür in ähnlichen Verteilungsfunktionen der Angaben in der Lie-Skala bei

normalen und bei psychiatrischen Patienten zu suchen ist. Green (2008, 176pp.) [115] sah deshalb mäßige Identifikations-Möglichkeiten der L-Skala für Underreporting.

Die *Wiggins Social Desirability Scale* (Sd; Wiggins 1959) [317] wird ebenfalls als Testskala bewusster positiver Antwortverzerrungen verwendet (vgl. Thies 2012 [296]). Im MMPI-2 sind noch 33 der ursprünglich 40 mittels Befragung instruierter Studenten gefilterten Items enthalten. Die Sd-Skala weist bei hoher bis maximaler Spezifität (88 bis 100 %) bei Grenzwerten ab 17 bzw. 21 eine hohe Sensitivität (68 % bis 74 %) zum Assessment von Underreporting auf.

Bewusste Verfälschungen erfasst auch die *Other-Deceptions-Scale* (ODS oder Odecp; nach Nichols und Green 1991 [214]; vgl. Thies 2012 [296]), die in ihrer Vorläufer-Version als *Positiv Malingering Scale Mp* (Cofer et al. 1949 [56]) eine optimierte Underreporting-Detektion ermöglichen sollte. Die Auswahl der 33 Items erfolgte anhand auffälliger Fragen bei entsprechend instruierten Probanden und soll nach den Autoren ab einem Score ≥ 20 auffällige Verfälschungs-Tendenzen indizieren. Da die Mp-Skala 72 % der Items mit der Sd-Skala teilt, weisen sie ähnliche Diskriminanz-Qualitäten auf.

Die 30 Items umfassende *Validitätsskala K*(Correction) wurde mit dem Ziel konstruiert, falsch negativ klassifizierte Probanden zu identifizieren, die ebenfalls ein Underreporting im MMPI-2 zeigen. Als Items wählten die Autoren Fragen, die speziell von stationären psychiatrischen Patienten angegeben wurden, die ansonsten ein MMPI-2-Normprofil aufwiesen. Insofern misst die K-Skala nach Thies (2012) [296] eher den Aspekt der Selbsttäuschung und weniger eine bewusste Bagatellisierung.

Hohe K(Correction)-Scores $\geq$ T60 (Rohwerte ab 20 bei Männern, 21 bei Frauen) sollen geringe psychische Fähigkeiten der Selbstreflektion vorhersagen können und damit geringe psychotherapeutische Erfolge. Nach Thies [296] erhebt die K-Skala damit mehr den Aspekt „unbewusster Selbsttäuschung“ und ist besser als die L-Skala zum Assessment von Defensiveness (als Abwehrmechanismus) geeignet. Nach Bear & Miller (2002) [14] erreichte die K-Skala eine noch geringere Effektstärke (Cohen's d = 1,13) als die L-Skala, verzeichnete jedoch eine geringere Anfälligkeit für Vorinstruktionen; nach Thies (2012) [296] geben jedoch gebildete Probanden generell höhere K-Werte an als weniger gut ausgebildete Personen. Zwischen gesunden und psychisch kranken Patienten differenziert die Skala nach Green (2008, 176pp.) [115] wiederum kaum. Dennoch ist sie regulär von den MMPI-2-Testautoren zur

Berechnung jedes MMPI-2-Basisprofils mit Korrektur von 5 der 10 klinischen Basisskalen vorgesehen.

Selbsttäuschungsaspekte erfassen ebenfalls die sog. *Superlative Scale* und die *Edwards Social Desirability Scale (So)*. Die *Superlative Scale* (Butcher & Hahn 1995 [51]), zur Personalauswahl von Piloten entwickelt, umfasst 55 MMPI-2-Items, die nach Thies (2012) [296] mit hoher Effektstärke (Cohen's d = 1,51) Underreporting erfassen. Nach Green (2008) [115] weist die S-Skala eine hohe Korrelation (0,82 bis 0,88) mit der K-Skala auf. Bei dem publizierten Cutoff $\geq$ 35 unterscheidet die S-Skala jedoch nur mäßig zwischen gesunden und psychisch kranken Personen.

Dieselbe eher mäßige Diskriminanzgüte wird von Green (2008) [115] hinsichtlich *Edwards' Social Desirability Scale* (So; Edwards 1957 [82]) berichtet, die als Skala zur Detektion sozial erwünschter Antworten im MMPI-2 mit 37 von ursprünglich 79 noch im MMPI-2 auffindbaren Items konstruiert wurde. Die Effektstärke der So-Skala wird von Thies (2012) [296] geringfügig höher als die der MMPI-2-K- und der -L-Skala angegeben (Cohen's d = 1,22 bei Cutoff-Score von mehr als 30).

Bagby et al. (1997) [21] belegten eine Möglichkeit, Underreporting mittels der Other-Deception-Scale und der Superlative-Scale bei vorab instruierten Studenten nachzuweisen, die ihr Antwortverhalten im Sinne eines *faking-good* verfälschen sollten. Um solches positiv verfälschtes Antwortverhalten gegenüber psychotisch-kranken, psychiatrischen Patienten nachzuweisen, waren jedoch die Edwards-Social-Desirability-Scale und die L-Skala zur Aufdeckung von Underreporting geeigneter.

Cima et al. (2003b) [55] schlagen mit sog. *Supernormality-Scale* eine neuere Alternative vor, Tendenzen von Probanden zu erfassen, sich so normal wie möglich darstellen zu wollen. In einer Untersuchungsgruppe forensischer Patienten waren hohe Angaben in dieser Skala mit einem Verdrängungs- und bagatellisierenden Antwortverhalten korreliert. Jedoch erfordert der Supernormalitäts-Fragebogen bei einer Sensitivität von nur ca. 28 Prozent bei einer mehr als 90-prozentigen Spezifität eine weitere Optimierung.

Underreporting-Tendenzen sind am trennschärfsten mittels der Wiggins Social Desirability Scale zu erfassen, während die traditionell zu demselben Zweck konzipierte L-Skala des MMPI-2 bislang keine sichere Aufdeckungsqualität bewies. Den Aspekt der „Selbst-Täuschung“ (geschönte Selbst-Darstellung) misst vermutlich exakter die K-Correction-Skala und erwies sich als wenig verfälschbar durch das Vorhandensein einer psychiatrischen Erkrankung. Möglicherweise erfasst die K-Correction-Skala unter den vielfältigen MMPI-2-Skalen und den MMPI-2-unabhängigen Skalen den Aspekt *Underreporting* am sichersten.

1.9.4 MMPI-2-Validitätsindices

Meyers et al. (2002) [202] bündelten in ihrer Studie erstmals die einzelne Aussagekraft von sieben MMPI-2-Validitätstests in einem Gesamtscore (**Meyers Validity Index**), der aufgrund aus der Literatur abgeleiteter Cutoff-Werte eine dreistufige Gesamtbeurteilung von Overreporting (als Sicher-, Wahrscheinlich- und Nicht-überhöht) ermöglichen sollte. In einer Untersuchung mit 100 Rentenantragstellern, 100 Patienten ohne Kompensationsverfahren und 30 Simulanten gelang es, bei einem Cutoff von 5 (von maximal 14 Punkten) alle Patienten ohne Kompensationsverfahren und 86 Prozent der Simulanten richtig zu klassifizieren, 33 Prozent der Rentenantragsteller zeigten ebenfalls Hinweise auf Simulationsverhalten. Thies (2012) [296] ermittelte bei einer Spezifität $\geq$ 90 % für den MVI ab einem Cutoff $\geq$ 4 eine Sensitivität von 27,9 % bei einer in ihrer Studie mit Gutachten-Patienten relativ hohen Trennschärfe (Cohen's d = 0,70).

Greve et al. (2006) [118] berichteten eine vergleichbare Diskriminanzgüte des MVI wie die jeweiligen Testgüten der Validitätsskalen Fb, FBS und Ds-r, wobei die Effektstärken nach den zitierten Studien nur bei der Ds-r-Subskala vergleichsweise hoch ausfiel (Cohen's d = 1,5 bis 1,6). Thies (2012) [296] kritisierte, dass bei der Validierung des Meyers-Index keine externen BV-Verfahren zur Sicherung der korrekten Klassifikation des Malingering verwendet wurden.

Der Meyers-Index zeigte dennoch einen vielversprechenden Ansatz, Overreporting psychischer Symptome mittels Kombination verschiedener Validitätsskalen aufzudecken. Ins-

besondere das gewichtete Verfahren mit Integration von sieben bereits in anderen Studien vielfach validierten Detektions-Skalen sollte das Risiko falscher Entscheidungen minimieren.

In Ergänzung dieses Ansatzes zeigten Aguerrevere et al. (2008) [3], dass mittels eines um zwei Subskalen (ohne O-S- und Ds-r-Skala) reduzierten Index (**Abbrevated Meyers Validity Index**) eine ebenso hohe Detektionsquote erreicht werden konnte wie mit dem Original-MVI. Ein Cutoff ≥ 3 deutete bei Patienten mit chronischen Schmerzen sicher auf Antwortverzerrungen hin. Allerdings erscheint die Auswahl der aus dem Index eliminierten Validitätsskalen fragwürdig: die beiden Skalen wurden mit der Begründung entfernt, dass sie nicht zur computerisierten Standard-Auswertung des lizensierten Testverlages gehörten; eine empirische Begründung wurde nicht angeführt.

Eine Kombination diverser Validitäts-Subskalenwerte in einem integrativen Index könnte einen prinzipiellen Ansatz zur Erhöhung der Detektionsgüte darstellen. Indem ein Index durch Synergie möglichst homogener Einzelskalen (mit hoher Reliabilität) das Gesamtkonstrukt „Overreporting“ subsummiert, kann möglicherweise die Validität der Erhebungen optimiert werden.

1.9.5 Vertiefende Studien zu den MMPI-2-Validitätsskalen

Zur Untersuchung der Gültigkeit und Konstruktvalidität von BV-Verfahren werden neben Simulationsstudien auch Studien mit sogenanntem Known-Groups-Design eingesetzt (Bianchini et al. 2008 [30]). Diese Studien sollen die Beurteilung ermöglichen, ob ein Test Subgruppen eines Probandenpools hinreichend differenziert, die sich vorab anhand definierter und als valide bekannter Außenkriterien (sog. *Gold-Standard*) des interessierenden Konstrukts unterscheiden.

Bei Known-Group-Studien entscheidet die Genauigkeit der Klassifikation über die Validität der Untersuchung. Dabei setzt der Verzicht auf eine experimentelle Manipulation der unabhängigen Variablen eine klare Beziehung zwischen Malingering-Verhalten und den zur Klassifikation verwendeten Testverfahren voraus.

Andere Studienansätze zur Untersuchung von Malingering sind Simulations- oder Analog-Designs und Studien zur Differentiellen Prävalenz (Studien, die Probanden aus hinsichtlich des Verfälschungsverhaltens unterschiedlichen Populationen rekrutieren). Bei letzterem Untersuchungsansatz kann jedoch die Klassifikations-Genauigkeit nicht empirisch überprüft werden, während sich beim ersten Untersuchungsansatz die künstliche Manipulation der unabhängigen Untersuchungsvariablen als nachteilig erweist. Simulations-Studien haben somit bei hoher interner Validität eine fragwürdige externe Gültigkeit, die bei Known-group-Studien am höchsten ist. Die Vergleichbarkeit von Studien mit sog. ge-coachten Simulanten ist zudem von der Art des Coachings abhängig (z.B. Coaching über Beschwerdedarstellung vs. Coaching über Detektions-Methoden).

In einer Vielzahl von Einzelstudien konnten diverse Stärken und Schwächen der beschriebenen Validitätsskalen zur Aufdeckung negativer Antwortverzerrungen mit entsprechenden Cutoff-T-Werten bei psychiatrischen Patienten, bei Patienten in forensischen und sozialmedizinischen Kompensationsverfahren, bei Patienten mit traumatischer Hirnverletzung sowie bei experimentell-instruierten Probanden bestätigt werden.

Thies (2012) [296] untersuchte mittels eines Known-Groups-Designs 288 deutschsprachige Probanden eines neurologisch-psychiatrischen Gutachteninstituts, indem sie die Untersuchten mittels Ergebnissen des Word-Memory-Tests WMT und entsprechend der Kriterien nach Slick et al. (1999 [279]) hinsichtlich Diskrepanzen und Inkonsistenzen auf der somatischen und der psychischen Ebene in drei Gruppen klassifizierte. Alle Probanden füllten im Rahmen ihrer Begutachtung auch den MMPI-2 aus. Nach Berücksichtigung der Ausschlusskriterien (Fehlende Antworten, TRIN- und VRIN-Scores über Cutoffs der Auswertbarkeit) verblieben 240 Probanden in der Studie, von denen weitere 14 wegen nicht voll erfüllter Kriterien für eine MND (Malingered Neurocognitive Dysfunction) von der Studie ausgeschlossen wurden. 216 Probanden wiesen in der Studie ein Kompensationsmotiv für Symptom-Overreporting auf; 116 Probanden wurden als authentisch/glaubwürdig klassifiziert, 24 und weitere 30 Personen wiesen Anzeichen einer wahrscheinlichen MND (Malingered Neurocognitive Dysfunction) nach Slick-Kriterien oder nach DSM-IV-Merkmalen auf, 32 Probanden entsprechend beider Kriterien. Die vier Untersuchungsgruppen (Anreiz für Underreporting, Authentische Probanden, mögliche und sichere Simulation) unterschieden sich nicht hinsichtlich Alter, Schulbildung und Geschlecht.

Absolut entsprechend der Erwartung unterschieden sich die Untersuchungsgruppen (Anreiz für Underreporting, authentische Probanden, möglich und sicher simulierende Personen) in den non-parametrischen Gruppenvergleichen in fast allen 15 untersuchten MMPI-2-Validitätsskalen signifikant, mit Ausnahme der Skala Fptsd (Infrequency Posttraumatic Stress Disorder Scale); in der Skala Fp wiesen die Probanden mit Anreiz für Underreporting ebenfalls entgegen der Erwartung nicht die niedrigsten Werte auf.

Dies könnte einerseits auf Schwächen der genannten beiden Skalen zur Erfassung von negativen bzw. positiven Antwortverzerrungen hindeuten. Andererseits könnte der fehlende Nachweis auch auf eine höhere Anzahl von Probanden mit genuiner, ´echter´ Psychopathologie im Gesamtpool der Studie und damit fälschlich als aggravierend eingestufte Probanden hindeuten.

Einschränkend ist anzumerken, dass in der Untersuchung trotz vielfacher Hypothesentestungen keine Alpha-Adjustierung der Signifikanzgrenzen vorgenommen wurde, was die Gefahr von Zufallssignifikanzen erhöht.

Thies' Untersuchung stellte dennoch den ersten und einzigen Studienansatz dar, differenziert nach Aussagen-Kriterien die Testgüte der MMPI-2-Validitätsskalen kritisch zu vergleichen. Zur Beurteilung der Detektions-Akkuranz verwendete die Autorin (Thies 2012 [296]) eine standardisierte Spezifität von 90 % zur korrekten Identifizierung von Probanden mit Overreporting; eine skalenspezifische, optimale Cutoff-Bestimmung (z.B. der von Youden 1950 [324] beschriebene Algorithmus) wurde nicht ermittelt.

Somit wird eine 10-prozentige Rate falsch positiv und damit als simulierend klassifizierter Probanden in Kauf genommen, was nach Überlegungen von Bianchini et al. (2008) [30] in Begutachtungs-Kontexten zugunsten der Beurteilung einzelner Personen schwierig sein kann. Letztere Autoren fordern hier bei einer zugrundegelegten Basisrate von 30 % möglichst hohe Spezifitätswerte, um einzelne Patienten nicht falsch zu stigmatisieren.

Legt man eine solche Maximal-Forderung zugrunde, leistet keiner der MMPI-2-Validitätsscores eine korrekte Klassifikation, da in vielen Studien die Sensitivität der meisten Scores unter einer Zufalls-Zuweisung liegt (vgl. Thies (2012) [296]).

Bianchini et al. (2008) [30] untersuchten erstmals die Diskriminationsgüte der MMPI-2-Validitätsskalen zur Aufdeckung von Malingering mittels eines *kombinierten Known-Groups- und Simulations-Designs* bei (a) 23 Patienten mit chronischen Schmerzen ohne

finanziellem Kompensationsmotiv, (b) 34 Patienten mit chronischen Schmerzen mit finanziellem Kompensationsmotiv, (c) 32 Patienten mit chronischen Schmerzen, die aufgrund BV-Verfahren als auffällig antwortend eingeschätzt wurden, und (d) 26 College-Studenten, die eine schmerzassoziierte Disability simulieren sollten. Die MMPI-2-Validitätsskalen differenzierten mit hoher Genauigkeit Malingering von Non-Malingering. Aber auch sehr hohe Werte (> 90 T-Werte) konnten in den als Indikatoren für Somatisierung bekannten Basisskalen Hysterie (Hy) und Hypochondrie (Hd; engl. Hs) Malingering sicher identifizieren.

Unabhängig von den MMPI-2-Validitätsskalen, wiesen die Patienten mit chronischen Schmerzen mit finanziellem Kompensationsmotiv mit 3-4-mal höherer Wahrscheinlichkeit als die Patienten mit chronischen Schmerzen ohne finanzielles Kompensationsmotiv T-Werte von mehr als 80 in den Skalen Hy und Hd auf. Das bedeutet, dass Patienten mit chronischen Schmerzen mit finanziellem Kompensationsmotiv selbst ohne Aggravation in den BV-Verfahren im Durchschnitt um 10 T-Wert-Punkte höhere Angaben machten als Patienten ohne Kompensationsmotiv.

Die sicher als MPRD-Patienten identifizierten Probanden und die experimentellen Simulanten wiederum machten im Durchschnitt um 10 T-Wert-Punkte höhere Angaben als die Patienten mit chronischen Schmerzen mit finanziellem Kompensationsmotiv ohne Aggravation in den BV-Verfahren. Die Subskalen-Werte der beiden ersten Gruppen lagen mit 4-mal höherer Wahrscheinlichkeit im Bereich von T-Werten über 90 im Vergleich zu Patienten mit chronischen Schmerzen mit finanziellem Kompensationsmotiv ohne Aggravation in den BV-Verfahren. Zudem wiesen in der Studie die sicher als MPRD-Patienten definierten Probanden (mit hohen Auffälligkeiten in den externen BV-Verfahren unter Zufalls-Wahrscheinlichkeit) noch extremere Scores auf als die experimentellen Simulanten.

Die Anzahl der sicher als MPRD-Patienten identifizierten Probanden war in der Studie (erwartungsgemäß) relativ gering und die Festlegung von MPRD wurde bei weniger als 10 % der Patienten mit chronischen Schmerzen mit finanziellem Kompensationsmotiv getroffen. Auch Sellbom & Bagby (2008, 194pp.) [268] gehen in klinischen Settings von einer üblicherweise sehr niedrigen Basisrate von Malingering aus. So berichten beispielsweise Blanchard et al. (2003) [33] in ihrer Vergleichsstudie zur Detektionsgüte des PAI und des MMPI-2 eine Grundrate von 10,7 % für Simulation (s.u. Kap. 1.9.7, S. 84).

Insgesamt bestätigte die Untersuchung von Bianchini et al. (2008) [30] die Diskriminations-Güte der MMPI-2-Validitätsskalen Meyers-Index, inklusive der F-assoziierten Validitätsskalen (F-Family: F, Fb, Fp, FBS) sowie der MMPI-2-Basisskalen Hy und Hd in der Detektion von schmerzassoziiertem Malingering. Allerdings war die untersuchte Stichprobe relativ klein; Malingering wurde zudem ausschließlich anhand der Auffälligkeiten in kognitiven *Performance-Validierungstests* definiert.

1.9.6 Manipulierbarkeit der MMPI-2-Validitätsskalen

Mehrere Studien belegten die Anfälligkeit für Verfälschbarkeit einzelner Validitätsskalen bei entweder höher gebildeten oder vorab instruierten Probanden, insbesondere hinsichtlich der Detektion von Underreporting.

Allein die Fb-failed-back-Skala des MMPI-2 war selbst durch in der Depressions-Diagnostik und -Therapie ausgebildete Experten nicht erfolgreich gegenüber psychisch kranken Patienten zu verfälschen (Bagby et al. 2000 [22], s. S. 73).

Pelfrey & Aamodt (1996) [221] zeigten jedoch, dass intelligente gesunde Probanden mit Basis-Wissen über den MMPI-2 erfolgreich in der Lage sind, in diesem Test ein Overreporting simulierend zu präsentieren. So stellten die Autoren fest, dass intelligente Probanden mit Kenntnissen über den MMPI-2 in der Lage waren, die Werte der klinischen Skalen zu erhöhen, gleichzeitig die Validitätsskalen, wie die F-(Seltenheits)-Skala, eher im unauffälligen, moderaten Bereich zu beantworten.

Aamodt et al. (1996) [1] wiesen zudem darauf hin, dass in Experimenten zur Simulation aufgeforderte Probanden die MMPI-2-F-Skala bis zu vier Standard-Abweichungen höher ankreuzen als authentisch antwortende Probanden der Normalbevölkerung, die nicht entsprechend instruiert wurden. Hingegen füllten Personen, die durch andere, externe Standard-Verfahren zur Beschwerdenüberhöhung als simulierend klassifiziert wurden, die Items der F-Skala nur um maximal 1,5 Standard-Abweichungen höher aus als Norm-Probanden.

Insofern benutzen Personen mit einem Overreporting- oder Malingering-Antwortverhalten durchaus moderatere, weniger leicht detektierbare (sog. *sophisticated*) Strategien der Fragebogen-Beantwortung, insbesondere je intelligenter sie sind und je mehr Vorinformationen sie über die Art der Befragung besitzen.

Hierzu untersuchte Pelfrey (2004) [220] ergänzend die Validitäts-Skalen F und F-K; höhere Scores waren in dieser Studie in den beiden MMPI-2-Skalen ebenfalls hoch korreliert mit Scores geringer allgemeiner Intelligenz und geringem Wissen über den MMPI-2. Umgekehrt waren niedrige F- und F-K-Scores mit hohen Werten allgemeiner Intelligenz und umfangreicher Kenntnis des MMPI-2 verbunden. Ähnliche Befunde der Verfälschbarkeit einzelner MMPI-2-Skalen berichteten Rogers et al. (1993) [236]) zur Simulation psychiatrischer Symptome einer schizophrenen Erkrankung.

Insofern stellen selbst an einer Vielzahl von Probanden gesicherte Cutoff-Werte nur eine relative, keine absolute Wahrscheinlichkeit dar, Overreporting identifizieren zu können, wenn sie nicht die Informiertheit und den Bildungsgrad der Probanden berücksichtigen.

Höherer Bildungsstand, höhere Aufmerksamkeit / Konzentration während der Testdurchführung, Informationsstand über die Testverfahren und tatsächliche Psychopathologie können fälschliche Klassifikationen verursachen.

1.9.7 MMPI-2 und PAI

Das Personality Assessment Inventory (PAI) nach Morey (1991) [211] wurde ursprünglich als gegenüber dem MMPI-2 verbesserte Version konstruiert. Es sollte logischer aufgebaut sein, mit besserem Design, mit modernerer Terminologie, mit vereinfachter Bearbeitung sowie Auswertung und zukunftsweisenden Interpretations-Möglichkeiten (nach White 1993, zit. nach Tscheuschner 2011 [299]).

Blanchard et al. (2003) [33] verglichen die Diskriminationsgüte beider Validitäts-Instrumente PAI und MMPI-2 in einer Dreigruppen-Studie mit 52 Studenten, die das Antwortverhalten forensischer Patienten bzw. in einer anderen Gruppe die Antworten psychiatrischer Patienten mit einer *fake-bad*-Intention simulieren sollten, sowie einer Vergleichsgruppe von

432 Psychiatrie-Patienten. Der F-K-Index und die Fp-Skala des MMPI-2 erwiesen sich als die besten Prädiktoren für Overreporting, die in der Detektionsgüte durch das PAI nicht übertroffen wurden.

Vergleichende Studien zur Klassifikationsgüte der Validitätsskalen und -indices des MMPI-2 und des PAI zur Detektion von Under- und Overreporting bestätigten eher moderate Korrelationen zwischen den Defensiveness-Skalen MMPI-2-L und -K und der PAI-Skala PIM (r = 0,41 und 0,44, vgl. Sellbom & Bagby 2008 [268]).

Studien zur Untersuchung der Identifizierung von Overreporting mittels MMPI-2 und PAI fanden dagegen hohe Korrelationen zwischen den MMPI-2-Skalen F, Fp und Fb und den PAI-Validitätsskalen. Die PAI-Skala NIM (Negative Impression Management) zeigte dabei die höchsten Korrelationen (Bagby et al. 2002 [18]), während die Indizes MAL und RDF niedrigere Zusammenhänge mit den MMPI-2-Skalen aufwiesen.

Ähnlichkeiten zwischen PAI und MMPI-2 bestätigten sich vor allem hinsichtlich der sog. F-Skalen (F, Fp, Fb), die negative Antwortverzerungen (Negative Impression Management) erfassen. Dabei erwiesen sich die MMPI-2-Skalen gegenüber den Skalen des PAI zumeist als trennschärfer.

1.10 MMPI-2-RF: Aktuellste Methoden der Beschwerdenvalidierung

Trotz dieser erfolgreichen Bemühungen haftete dem MMPI-2 lange Jahre die Kritik an, seiner Profilkonstruktion lägen nicht mehr moderne, psychoanalytisch-fundierte, psychiatrisch-historische Konzepte zugrunde (Cronbach 1990 [60], Helmes & Reddon 1993 [129]). Diese Kritik betraf insbesondere die Diagnostik bei chronischen Schmerzpatienten, als eines bei Patienten mit chronischen Schmerzen im englischen Sprachraum am häufigsten eingesetzten Psychodiagnostikums. Fernerhin ermöglichten vielfältige Doppelkodierungen von Items in Bezug zur Subskalen-Zugehörigkeit keine orthogonale Trennschärfe der Subskalen mit der Folge geringer interner Konsistenz (Franz et al. 1986 [94]).

Zum anderen wurde die Testlänge von 567 Items mit erheblicher Bearbeitungsdauer als unökonomisch und klinisch wenig praktikabel kritisiert. Damit wurde auch die klinische

Nützlichkeit als Validitätsaspekt in Frage gestellt. Schließlich wurde in Zweifel gezogen, ob der MMPI-2 tatsächlich authentische psychiatrische Störungen von Verfälschungen, Übertreibungen oder Simulation von Beschwerden unterscheiden kann (Pincus et al. 1986 [222]).

Wenngleich dieses Argument zum Teil durch Identifizierung von MMPI-2-Items, die selbst bei psychiatrischen Vergleichsprobanden außergewöhnlich selten vorkommen, widerlegt werden konnte (vgl. Konstruktion der Fp-Skala durch Strong et al. 2006 [284]), wird im MMPI-2-Testmanual bei auffälligen Werten der Validitätsskalen auf die Notwendigkeit hingewiesen, diese im Einzelfall auch als Hinweis auf eine tatsächliche Psychopathologie zu interpretieren.

Aufgrund dieser Kritikpunkte am MMPI-2 legten Ben-Porath und Tellegen [29] in 2008 mit dem MMPI-2-RF eine bisher nur in den USA eingesetzte, neuartige Rekonstruktion des MMPI-2 vor. Diese Testversion weist eine gegenüber der 567 Items umfassenden zweiten Version auf 338 Items verkürzte Fassung auf. Zusätzlich wurden zehn Basisskalen und eine ebenso große Anzahl von Validitäts-Subskalen neu extrahiert und hinsichtlich ihrer Validität überprüft.

Welche Klassifikations-Möglichkeiten diese Test-Neukonstruktion bei der klinischen Einschätzung von Patienten mit chronischen Schmerzen bietet, kann derzeit mangels Studien noch nicht vollständig eingeschätzt werden.

Rogers et al. (2006) [243] betonten in einem ersten Überblicksartikel, dass die MMPI-2-Restructured Form weit mehr darstellt, als nur ein „Retrofitting“ eines überholten, veralteten Diagnostikums, sondern vielmehr einen Paradigmen-Wechsel der Skalen-Entwicklung einleiten kann.

Gegenüber der tradierten psychiatrisch-psychoanalytischen Klassifikation wendet sich die Neudefinition der MMPI-2-RF-Basisskalen einer modernen, verhaltensorientierten Diagnostik zu. Erfasst werden folgende Skalen: Demoralisation (RCd), Somatic Complaints (RC1), Low Positive Emotions (RC2), Cynismn (RC3), Antisocial Behavior (RC4), Ideas of Persecution (RC6), Dysfunctional Negative Emotions (RC7), Aberrant Experiences (RC8), Hypomanic Activation (RC9), die mit den traditionellen MMPI-/MMPI-2-Skalen verglichen werden können ([1] Hd-Hypochondrie-Skala, [2] D-Depressions-Skala, [3] Hy-Hysterie-Skala / Konversion, [4] Pp-Psycho-Soziopathie-Skala, [5] Pa- Paranoia-Skala, [6] Pt- Psychasthenie-

Skala, [7] Sc- Schizophrenie-Skala, [8] Ma- Hypomanie-Skala, [9] Si- Soziale Introversions-Skala).

Ferner wurden sechs klassische, im MMPI-2-RF nicht überlappende Validitätsskalen integriert (die Infrequent Responses F-r, die Infrequent Psychopathology Fp-r, die Symptom-Validity-Scale FBS-r, die Response Bias Scale RBS sowie die revidierten L- und K-Skalen).

Zusätzlich konnte die zur Erfassung seltener somatischer Symptome konzipierte, sogenannte Fs-Skala (Infrequent Somatic Complaints, vgl. Kap. 1.9.1) hinzugefügt werden. Tellegen et al. (2006) [295] benutzen bei der Entwicklung der Basis- und der Validitätsskalen des MMPI-2-RF eine spezielle Konstrukt-Validierungs-Technik (Jackson's Test-Entwicklungs-Methode aus dem Jahr 1970), mit der die Skalen-Homogenität optimiert wurde und Interskalen-Korrelationen minimiert wurden.

1.10.1 Die neuen Restructured Clinical-Scales (RC)

In neueren MMPI-2-RF-Studien zur Untersuchung der RC-Skalen wurden bei Patienten psychotherapeutischer Kliniken und bei Militärveteranen (Simms et al. 2005 [275], Sellbom et al. 2006 [271], Van Der Heijden et al. 2010 [305] in einer Niederländischen Patientengruppe) sehr ähnliche Ergebnisse der zwei MMPI-2-Subskalen-Testsets bestätigt. Handel et al. (2010) [121] stellten zudem eine relative Robustheit der Validitätskoeffizienten der RC-Skalen gegen moderate Abweichungen der Testvalidität in den nicht-inhaltsbasierten Kontrollskalen (TRIN, VRIN) fest.

Rouse et al. 2008 [251] fanden bei fünf der neun RC-Skalen (RC1, RC3, RC7, RC8 und RCd) hohe Korrelationen, die die Autoren als Hinweis auf eine hohe Replizierbarkeit der MMPI-2-RF-Skalen gegenüber den klassischen MMPI-2-Skalen interpretierten. Alpha- und Inter-Item-Korrelations-Analysen bestätigten hingegen keine höheren Reliabilitäten der RC-Skalen gegenüber Skalen der Vorgängerversion MMPI-2.

Deutlich überhöhte MMPI-2-RF-Profile der RC-Skalen fanden Sellbom et al. (2010) [272] bei einer Gruppe forensischer Patienten (vgl. Abb. 7, S. 88) mit Überhöhungen in fast allen RC-Skalen mit Ausnahme von RC3 (Cynismn) und RC9 (Hypomanic Activation).

Die Werte der nicht-authentischen Patienten lagen zwischen T-Werten über 80 bis zu Werten von 110 (speziell in der Skala RC6, Ideas of Persecution oder paranoide Gedanken), während die Scores der authentischen Patienten im T-Wert-Bereich T50-60 lagen.

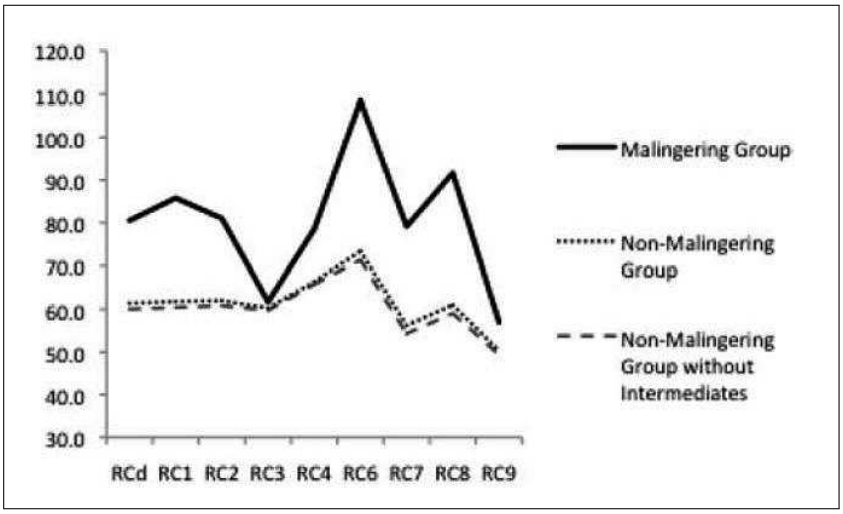


Abb. 7. Mittlere MMPI-2-RF-Clinical Scale Profile einer Malingering- und zweier Non-Malingering-Gruppen (aus: Sellbom et al. 2010 [272])

Diese Ergebnisse lassen vermuten, dass sich Malingering auch möglicherweise im MMPI-2-RF durch die RC-Basis-Skalen identifizieren lässt, wie dies bereits im MMPI-2 für die Basisskalen Hy und Hd festgestellt wurde (Bianchini et al. (2008) [30], s. Kap. 1.9.5, S. 79).

Ähnlichkeiten zwischen den Basisskalen des MMPI-2-RF (RC-Scales) und des MMPI-2 lassen vermuten, dass die neuen RC-Skalen teilweise ähnliche Detektions-Eigenschaften wie die Basisskalen des MMPI-2 haben und sich in Skalen-Überhöhungen abbilden.

Die Konsistenzprüfungsskalen TRIN und VRIN des MMPI-2/-RF: prüfen zusammen mit den nicht oder doppelt-beantworteten Items (*engl.* Cannot-Say CNS, $\geq$ 15, max. 10 % aller Items) die Antwortgüte. Mit der Skala *True Response Inconsisteny (TRIN)* wird die Konsistenz der Beantwortung von 26 Item-Paaren mit quasi-umgekehrtem Inhalt kontrolliert. Ein TRIN-Score $\geq$ T80 indiziert nicht mehr interpretierbare Aussagen. Denselben Zweck erfüllt die Skala *Variable Response Inconsisteny (VRIN)* mit 53 inhaltsähnlichen Items, die ab VRIN-Score $\geq$ T80 anzeigen, dass ein Proband den Test nicht kooperativ ausfüllte oder dazu (wegen kognitiver oder sprachlicher Defizite) nicht in der Lage war. Entsprechend auffällige Tests sollten von einer Analyse ausgeschlossen werden.

1.10.2 Charakteristika der neuen MMPI-2-RF-Validitätsskalen

Die Validitätsskala F-r - Seltener Symptome des MMPI-2-RF umfasst 32 der ursprünglich 66 Items der MMPI-2-Skala F und wurde von den Testautoren als genereller Indikator für Overreporting einer Vielzahl somatischer, psychologischer und kognitiver Symptome konstruiert. Sie gilt ab Cutoff-Werten $\geq$ T90-T99 (Rohwerte 11-12) als valides Indiz für Malingering, mit auch weit höheren möglichen Summenwerten (Ben-Porath & Tellegen 2008, [29]). Die Testautoren weisen hinsichtlich der Ergebnisse jeder der MMPI-2-RF-Validitätsskalen daraufhin, dass höhere Werte nicht nur mögliches Overreporting signalisieren, sondern auch als Hinweis auf invalide Antwortprotokolle überprüft werden müssen oder auch als Indiz für eine tatsächlich schwere Psychopathologie zu interpretieren sind.

Die F-r-Skala teilt mit der RBS-Skala drei Items (74, 106 und 120), mit ansonsten keinem übereinstimmenden Item mit weiteren Validitätsskalen. Die F-r-Skala teilt 28 Items mit verschiedenen Basisskalen, mit Ausnahme der RC3-Skala.

Die Fp-r-Skala- Seltener psychopathologischer Symptome des MMPI-2-RF beinhaltet 21 der 27 Items der MMPI-2-Skala Fp mit Indiz für überhöht dargestellte psychische oder psychiatrische Symptome bei Cutoff-Werten $\geq$ T80-T99 (Rohwerte 5-6). Die für die Skala selektierten Items sind im Gegensatz zur F-Skala weniger mit allgemeinem Distress oder schweren Störungen korreliert, da sie insbesondere von psychiatrischen Patienten selten positiv bejaht werden. Nach Ben-Porath und Tellegen (2008) [29] eignen sie sich deshalb speziell bei psychiatrisch auffälligen Patienten zur Identifikation von Overreporting.

Die Fp-r-Skala hat mit der FBS- und der RBS-Skala jeweils dasselbe Item (79) gemeinsam, sie teilt ein Item (252) mit der RBS-Skala, ferner kein Item mit weiteren Validitätsskalen und 5 Items mit der Basisskala RC6.

Die Fs-Skala - Seltener somatischer Symptome des MMPI-2-RF Die 16 Items umfassende Fs-Skala wurde im Rückgriff auf einen großen Pool archivierter MMPI-2-Daten entwickelt, als Maß für von Patienten mit chronischen Schmerzen ungewöhnliche, selten beklagte somatische Symptome (Wygant, Ben-Porath und Arbisi (2006) [321], s. S.55). Die Auswahl der Items erfolgte relativ großzügig mit Überhöhungen bei weniger als 25 % der Probanden.

Bei Überschreitungen von T-Werten von $\geq$ T80-T99 (Rohwerte 5-7) wurden Hinweise auf Overreporting konstatiert [320].

Die Fs-Skala hat mit der FBS-Skala drei Items (15, 43, 133) und mit der RBS-Skala zwei Items (137, 159) gemeinsam. Sie teilt ein Item (2) mit der Basisskala RC1, drei Items mit der Basisskala RC8. Darüber hinaus weist kein Item der Skala Übereinstimmung mit anderen Validitätsskalen oder Basisskalen auf.

Studien zu den F-Skalen des MMPI-2-RF (F-r, Fp-r und Fs) Einige wenige Studien untersuchten bislang geeignete Cutoff-Überschreitungen zur Diskriminationsgüte von Overreporting in den MMPI-2-RF-Validitätsskalen.

In einer Simulations-Studie stellten Burchess & Ben-Porath (2010) [45] signifikant höhere Werte der simulierenden Probanden psychopathologischer Symptome in den wesentlichen drei Validitätsscores (F-r, Fp-r, F-s mean $\geq$ T110) fest, im Vergleich zu Probanden, die somatische Symptome simulieren sollten ($\geq$ T80).

Wygant et al. (2009) [322] beobachteten in einer Studie mit medizinische Symptome simulierenden Probanden hohe Effektstärken der Skala FBS-r (Cohen's d = 2,31), der F-r-Skala (d = 2,01) und der F-s (d = 1,97) zur Identifizierung der auffälligen Probanden. In einer Simulationsgruppe, in der Symptome einer Schädel-Verletzung simuliert werden sollten, deckte dieses Overreporting die Fs-Skala am sichersten auf. Probanden, die in Symptom-Validierungs-Tests auffällig waren, wurden am deutlichsten durch die Skalen FBS-r und F-r identifiziert.

Ähnliche Effekte wurden in einer Known-Groups-Studie mit Einsatz kognitiver Symptom-Validierungstests zur Klassifikation des Aggravationsverhaltens der Untersuchten gesehen (Gervais et al. 2010 [101]), in der sich die neuen Skalen des MMPI-2-RF F-r, Fp-r, Fs und FBS-r tendenziell besser zur Overreporting-Detektion als ihre MMPI-2-Vorgänger erwiesen.

In einer anderen Simulations-Studie verglichen Sellbom & Bagby (2010) [269] MMPI-2-RF-Protokolle von Studenten, die psychische Symptome simulieren sollten, mit den Protokollen von psychiatrisch kranken, stationären Patienten. Die Autoren bestätigten die optimalste prädiktive Diskriminanz in diesem Vergleich bei Auswertung der Fp-r-Skala, die überhöhte psychische Symptome erfassen soll. Auch entsprachen die Testwerte den bei den Testautoren genannten Cutoff-Werten (Ben-Porath & Tellegen 2008 [29]).

In einer Drei-Gruppen-Simulation-Studie (Rogers et al. 2011 [239]) mit Probanden, die eine psychische Störung simulierten, und Probanden, die neurokognitive Störungen vortäuschten, zeigten diese Gruppen deutliche Auffälligkeiten gegenüber authentischen Patienten mit großer Effektstärke ($d \geq 2{,}0$) in den Skalen F-r und Fp-r. Cutoff-Scores in der Fp-r-Skala $\geq$ 90T führten zu nahezu keinen falsch-positiven Ergebnissen und nur zu moderaten Raten an falschen Alarmen.

Sellbom et al. (2012) [273] verglichen in einem Drei-Gruppen-Design authentische somatische Patienten ohne Kompensations-Anspruch mit einer Gruppe von Patienten mit somatoformer Schmerzstörung und einer Gruppe von Simulanten somatischer Symptome. Erneut bestätigten sich die Validitätsskalen Fs und Fp-r als beste Diskriminations-Indikatoren zwischen den Gruppen. Die Skala Fs identifizierte mit höchster Sensitivität die Vortäuschung somatischer Symptome, während die Skala Fp-r die höchste Spezifität aufwies. Die Skala FBS-r jedoch, die generell zwischen authentisch somatisch Kranken und Patienten mit Tendenz zur Verdeutlichung somatischer Symptome differenzieren sollte, führte zu eher geringer Detektionsgüte.

Noch deutlicher überhöhte MMPI-2-RF-Profile der Validitätsskalen fanden Sellbom et al. (2010) [272] bei den von ihnen untersuchten forensischen Patienten mit Skalen-Überhöhungen zwischen T-Werten über 90 bis zu Werten von 140 (speziell in der F-r-Skala, s. Abb. 8).

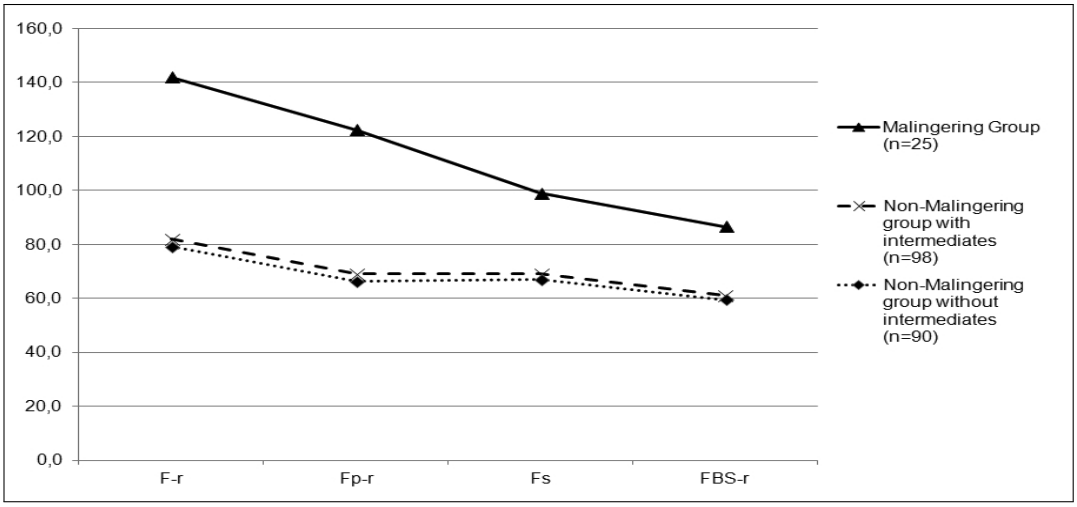


Abb. 8. Mittlere MMPI-2-RF-Validity-Scale Profile einer Malingering- und zweier Non-Malingering-Gruppen (graphisch umgesetzt, aus: Sellbom et al. (2010) [272])

Werden somit von Probanden psychiatrische oder auch teilweise somatische Symptome überbetont, zeigen sich diese Auffälligkeiten vor allem in der F-r-Skala und in der Fp-r-Skala, aber auch in geringerem Ausmaß in den Validitätsskalen Fs und FBS-r.
Bei Patienten mit Überbetonung somatischer Symptome (wie Patienten mit chronischen Schmerzen), die oft psychische Symptome eher dissimulieren, bilden sich Beschwerdeüberhöhungen möglicherweise mit anderer Pointierung ab.

Die FBS-r-Skala - Symptom-Validity-Scale des MMPI-2-RF umfasst 30 Items mit einer Überlappung von drei Items mit der F-r-Skala und mit einem Item mit der Fp-r-Skala und wurde aus dem Set von 43 Items der Vorgänger-Skala FBS aus dem MMPI-2 konstruiert. Diese Skala wurde ursprünglich im MMPI-2 entwickelt, um ungewöhnliche somatische und neurokognitive Symptome bei Antragstellern in zivilrechtlichen Prozessen zu identifizieren (s. S. 63). Auffälliges Overreporting wurde von den Testautoren bei Überschreitungen von T-Werten $\geq$ T80-T99 (Rohwerte 17-22) postuliert, Scores $\geq$ T100 werden als nicht mehr interpretierbar angegeben.

Die FBS-r-Skala weist mit anderen Skalen eine Reihe von Überlappungen auf. Sie hat mit der Fs-Skala zwei Items (43, 133) und mit der RBS-Skala zwei Items (6, 79) gemeinsam. Sie teilt drei Items (36, 99, 187) mit der K-r-Skala, ein Item mit der L-r-Skala (45) und 11 Items mit den Basisskalen RCd, RC1, RC3 und RC7. Darüber hinaus gibt es keine Übereinstimmung mit den Items weiterer Validitäts- oder Basisskalen.

Jones et al. (2012) [144] bestätigten in einer Untersuchung zur Diskriminanz der MMPI-2-RF-Validitätsskalen kongruente Überhöhungen mit externen kognitiven Symptom-Validierungstests. Große Effektstärken aller Standard-MMPI-2-RF-Validitätsskalen zeigten eine gute Detektionsgüte zur Aufdeckung suboptimalen Leistungsbemühens in den neurokognitiven Symptom-Validierungs-Tests. Dabei zeigten der Henry Heilbronner Index (HHI), die Response Bias Scale (RBS), die Fake Bad Scale (FBS) und die Symptom Validity Scale (FBS-r) eine bessere Diskriminanz als die Skalen der F-Familie.

In Übereinstimmung mit dieser Studie stellten Schroeder et al. (2012) [264] in einem Gruppendesign mit multiplen Symptom-Validierungstests neuropsychologisch auffälliger Patienten eine höhere Sensitivität der FBS-r-Skala als in anderen Untersuchungen fest. Bei der üblichen Spezifität (von 90 %) wies die MMPI-2-RF-Validitätsskala FBS-r eine Sensitivität von 48 % auf, die in dieser Studie weit höher lagen, als beispielsweise die Sensitivität der Korrektur-Skala K-r (10 %).

Gass et al. (2012) [96] beschrieben in einer Studie mit 303 nicht-forensischen neuropsychologischen Patienten eine als relativ hoch einzuschätzende, sog. „akzeptable“ Reliabilität (innere Konsistenz) der FBS-r-Skala von 0,747. Allerdings zeigten die faktoriellen Analysen bei der FBS-r zwei unterschiedliche faktorielle Dimensionen: zum einen erhebt diese Skala den Faktor „Ausmaß somatischer Symptome“, zum anderen auch den Faktor „Optimismus bzw. Abwehr sozialer Erwünschtheit / Ehrlichkeit“. Dies könnte nach Meinung der Autoren möglicherweise mehrdeutige Ergebnisse der FBSr-Scores erklären.

Die RBS-Skala - Response Bias Scale des MMPI-2-RF soll durch 28 ausgewählte Items überhöhte Angaben von Gedächtnis-Einbußen erfassen. Jedoch proklamieren die Testautoren auffällige Scores von mehr als T80-99 (Rohwerte 12-16) auch als mögliche Kennwerte für emotionale Störungen.

Scores von mehr als 100 T-Wert-Punkten indizieren nach dem Testmanual relativ sicher Overreporting. Obwohl Performance-Anstrengungstests kognitiver Leistungen zur Auswahl der Skalenitems verwendet wurden, eignet sich die RBS natürlich nicht zur Erfassung der Testanstrengung (vgl. Ben-Porath & Tellegen 2008 [29]). Sie soll vielmehr die überhöhte Angabe kognitiver Leistungsdefizite erfassen.

Die RBS-Skala weist mit anderen Skalen eine Reihe von Überlappungen auf. Sie hat mit der F-r-Skala drei Items (74, 106, 120) und mit der Fp-r-Skala zwei Items (79, 252) gemeinsam. Sie teilt zwei Items (36, 99, 187) mit der Fs-Skala (137, 159), jeweils ein Item mit der K-r-Skala (171) und der L-r-Skala (268), 11 Items mit den Basisskalen RC1, RC4, RC7, RC8 und RC9. Darüber hinaus existiert keine Übereinstimmung von RBS-Items mit weiteren Validitäts- oder Basisskalen.

In einer jüngsten Studie mit zwei unabhängigen Stichproben von Patienten mit Klageverfahren zur Anerkennung eines höheren Behinderungsgrades und einer Gruppe forensischer

Haftinsassen beobachteten Wygant et al. (2010) [323] eine akzeptable Sensitivität der RBS (von 0,38) bei Cutoff-Werten ab 90T, um negative Antwortverzerrungen neurokognitiver Symptome bei Patienten mit einer Behinderung nachzuweisen. Wurden die Fragebögen der Haftinsassen herangezogen, lagen die entsprechenden RBS-Scores erheblich höher bei T100 bis T110 (ab einer Spezifität von 89 %). Erneut zeigte sich, dass forensische Patienten erheblich deutlichere Auffälligkeiten aufweisen als Patienten mit zumindest teilweise somatischen Behinderungen. Entsprechend höher war somit die Sensitivität der RBS (Sicherheit der Detektion) bei den forensischen Probanden.

Die Response Bias Scale zeigte in dieser Untersuchung im Vergleich mit den Standard-Validitäts-Skalen des MMPI-2-RF die größte Diskriminanzgüte (Cohen′s d = 1,24 und d = 1,48) zur Aufdeckung von Verdeutlichung neurokognitiver Defizite.

Gervais et al. (2010) [101] beobachteten, dass die Validitätsskalen des MMPI-2-RF (insbesondere FBS-r, RBS) in vergleichbarem Ausmaß mit Auffälligkeiten in verschiedenen Gedächtnis-Prüf-Inventaren assoziiert waren, wie ihre entsprechenden MMPI-2-Skalen-Äquivalente. Teilweise wiesen die neueren Validitäts-Skalen des MMPI-2-RF sogar höhere Werte auf als die Skalen der Vorgängerversion (FBS Cohen's d = 0,97 gegenüber FBS-r Cohen's d = 1,11).

Tarescavage et al. (2013) [293] bestätigten zusätzlich durch eine Untersuchung von 916 nicht-neurologisch verletzen Klägern, die nach Symptom-Validierungstests (mit sog. Malingered Neurocognitive Dysfunction) als auffällig testverfälschend klassifiziert wurden, eine hohe Diskriminanzgüte der MMPI-2-RF-Skala RBS. Die Autoren stellten zudem fest, dass die Einbeziehung weiterer Patienten-Informationen die Sicherheit der Klassifikation erhöhte. Hierzu wurden die Zusatzskalen des MMPI-2-RF über Distress, Internalisierungstendenz, Gedankliche Störungen und Vermeidung von Sozialkontakten mit einbezogen.

Die Validitäts-Skalen Henry Heilbronner Index (HHI-r), Response Bias Scale (RBS), Fake Bad Scale (FBS) und die Symptom Validity Scale (FBS-r) scheinen mehr als die F-Skalen Überhöhungen kognitiver Funktionsdefizite zu erfassen (s. Studie von Jones et al. (2012) [144]). Trotz ermutigender erster Ergebnis-

se wird die Wertigkeit der FBS-r zur Erfassung von Malingering derzeit noch als unklar und mehrdeutig eingeschätzt.

1.10.3 MMPI-2-RF-Skalen zum Assessment von Underreporting

Die 14 Items umfassende K-r-Validitätsskala dient ebenso wie ihre MMPI-2-Vorgängerversion der Erfassung von Underreporting. T-Umrechungs-Scores ≥ 70 deuten gegenüber niedrigeren Scores auf eine im Vergleich zu den Normprobanden übermäßige und damit unglaubwürdige soziale Anpassung hin.

Die K-r-Skala weist mit anderen Skalen einige wenige Überlappungen auf. Sie hat mit der FBS-Skala drei Items (36, 99, 187) und mit der RBS-Skala ein Item (252) gemeinsam. Sie teilt zwei Items (36, 99, 187) mit der Fs-Skala (137, 159), jeweils ein Item mit der K-r-Skala (171) und der L-r-Skala (171) und 12 Items mit den Basisskalen RCd, RC2, RC4, RC7 und RC9. Kein sonstiges Item der K-r-Skala weist eine Übereinstimmung mit weiteren Validitäts- oder Basisskalen auf.

Die L-r-Validitätsskala dient mit ebenfalls 14 Items wie ihre MMPI-2-Vorgängerversion der Erfassung positiver Antwortverzerrungen. Die Testautoren sehen auffällige Scores von mehr als T80 (Rohwerte ≥ 9) als Marker für eine unglaubwürdig „geschönte“ Selbstdarstellung an. Die L-r-Skala hat jeweils mit der FBS-Skala ein Item (45) und mit der RBS-Skala ein Item (268) gemeinsam. Diese Skala teilt jeweils ein Item mit den Basisskalen RC2 und RC9, kein Item weist Übereinstimmungen mit weiteren Validitäts- oder Basisskalen auf.

Thies (2012) [296] kritisierte an der MMPI-2-L-Skala, sie könne durch die generelle Antworthaltung des Befragten beeinflusst werden, da nur Items einer Polierung summiert würden (s. S. 75). Diese Kritik gilt jedoch nicht für die Auswertung mittels MMPI-2-RF: hier werden in der Validitätsskala L-r Items beider Polierungen integriert. Insofern könnte die neu konzipierte Skala L-r besser zur Detektion von Underreporting geeignet sein als ihre Vorgängerfassung.

Eine Simulationsstudie von Sellbom et al. (2008) [270]) bestätigte eine valide Diskriminations-Möglichkeit der L-r- und der K-r-Skalen zwischen tatsächlich psychiatrisch kranken

Patienten und instruierten Studenten, die Beschwerden bagatellisieren sollten. Dieselben Skalen zeigten auch Unterschiede im sog. „*Underreporting*“ bei einer Gruppe von Antragstellern für das Sorgerecht ihrer Kinder im Vergleich zu entsprechend zu Underreporting instruierten Studenten.

Obwohl die L-r-Skala in Simulations-Studien und in Studien mit Probanden, die Motive für positive Antwortverzerrungen aufwiesen, Diskriminanz-Merkmale für dieses *faking - good* - Verhalten zeigte, erwies sie sich bei der Validierung negativer Beschwerde-Überhöhungen als nicht hinreichend trennscharf.

1.10.4 MMPI-2-RF: Spezielle jüngste Studien zur Beschwerdenvalidierung

Eine der wenigen Studien (Greiffenstein et al. 2012 [116]), die eine spezielle Gruppe von Patienten mit chronischen Schmerzen untersuchte, stellte bei Patienten mit Chronischem Regionalem Schmerzsyndrom (CRPS) Typ I mit einem Kompensations-Begehren zu 75 % in einem der beiden verwendeten externen Verfahren zur Beschwerdenvalidierung (Test of Memory Malingering - 23 % der Probanden, Reliable Digit Span - 50 % der Probanden) Auffälligkeiten fest; gleichzeitig wiesen die Patienten zu mehr als der Hälfte in mindestens einem MMPI-2-RF-Validitäts-Score eine Beschwerdenüberhöhung auf.

Die Autoren deuteten das Ergebnis als Hinweis auf möglicherweise häufige Beschwerdeüberhöhungen bei CRPS-Patienten mit Kompensationswunsch. Ungünstigerweise fehlt in der Studie eine Kontrollvergleichsgruppe, die ein geringeres Ausmaß an Auffälligkeiten bei einer anderen Gruppe von Patienten mit chronischen Schmerzen oder einer Patientengruppe ohne gesundheitliche Störungen belegen würde.

McCord & Drerup (2011) [184] untersuchten die Nützlichkeit und Anwendbarkeit des MMPI-2-RF bei 239 depressiven Patienten und 77 nicht-depressiven ambulanten neuropsychologischen Patienten, die alle auch chronische Schmerzen hatten. Während die depressiven Patienten in den klassischen MMPI-2-Subskalen in sieben der acht Basisskalen Auffäl-

ligkeiten aufwiesen, waren im neuen MMPI-2-RF nur zwei von acht Skalen erhöht. Insofern scheint der MMPI-2-RF eine höhere Differenzierungs-Spezifität zu besitzen.

In einer Studie zur *Verfälschbarkeit des MMPI-2-RF* zeigten Burchess & Ben-Porath (2010) [45], dass Probanden, die psychopathologische Symptome im MMPI-2-RF simulierten, auffälligere Werte in den Validitätsskalen (F-r, Fp-r und F-s im Durchschnitt $\geq$ 100) aufwiesen, als Probanden, die somatische Symptome überbetonten (F-r, Fp-r und F-s im Durchschnitt $\geq$ 80). Dies entsprach der Erwartung der Autoren, da die meisten Skalen im MMPI-2-RF sich auf psychopathologische Symptome beziehen und weniger auf die Abfrage somatischer Beschwerden.

Überraschenderweise machten die Probanden, die psychopathologische Symptome simulieren sollten, auch in der Skala F-s höhere Angaben im Vergleich zu den Probanden, die somatische Symptome überhöht angeben sollten.

Burchess & Ben-Porath (2010) [45] mutmaßten aufgrund einer Zusatzbefragung, dass die erste Gruppe die Instruktionen häufiger befolgte. Aber auch frühere Studien bestätigten bereits das höhere Overreporting von Probanden, die psychopathologische Symptome simulieren (Sivec et al. 1994, [277]).

Es ist zu vermuten, dass dieses Phänomen mit unterschiedlichen gesellschaftlichen Stereotypen über somatische und über psychische Erkrankungen zusammenhängt. Während bei somatischen Symptomen bizarre Überhöhungen leicht (augenscheinvalide und offensichtlich) als unglaubwürdig erkannt werden, erscheinen möglicherweise vielen Menschen psychische Symptome umso glaubwürdiger, je bizarrer und je ausgefallener diese sind.

Bei Patienten mit somatisch einzuordnenden Symptomen (also bei Patienten mit chronischen Schmerzen), die psychische Symptome eher bagatellisieren und somatische Symptome überbetonen, sollten sich nicht-authentische Auffälligkeiten eher in den Skalen F-r, FBS-r und RBS abbilden; möglicherweise finden sich bei diesen Patienten weniger pointierte Überhöhungen in den Skalen Fs und Fp-r.

Anderson (2011) [9] untersuchte 169 Patienten einer forensischen neuropsychiatrischen Praxis mit Kompensations-Ansprüchen, die durch drei kognitive Performance-Validierungsverfahren, drei Symptom-Validierungsverfahren und den Abgleich angegebener und beobachtbarer Verhaltenseinschränkungen in drei Gruppen von Overreporting klassifiziert wurden, hinsichtlich Identifizierbarkeit von Overreporting mittels der MMPI-2-RF-Validitätsskalen. Als externe SVTs wurde das SIMS (Structured Interview of Malingered Symptomatology) nach Smith und Burger (1997) [281], der Test M-FAST (Miller Forensic Assessment of Symptoms Test) nach Miller (2001) [203] und das SIRS (Structured Inventory of Reported Symptoms) nach Rogers et al. (1992) [237]) verwendet. Als externe PVTs wurden der Test of Memory Malingering (TOMM nach Tombaugh 1996 [298]) verwendet, der Victoria Symptom Validity Test (VSVT nach Slick et al. 1995 [278]) und der Letter Memory Test (LMT nach Slick et al. 1995 [278]).

In der Studie erlaubten alle MMPI-2-RF-Validitätsskalen eine gute Identifizierung von Overreporting, insbesondere durch die RBS-Skala (Cohen's d = 1,67), die F-r-Skala (d = 1,63) und die Skala Fs (d = 1,37). Die Autorin kam aufgrund ihrer Akkuranz-Analysen zu dem Schluss, in den vier Standard-Validitäts-Skalen F-r, Fs, FBS-r und RBS T-Wert-Erhöhungen $\geq$ T90 als Indiz einer Aggravation zu werten (s. Abb. 9, S. 99). Demgegenüber sollten in der Fp-r-Skala bereits T-Wert-Erhöhungen $\geq$ T70 sichere Auffälligkeiten zeigen, da diese Skala für die Differenzierung von Overreporting die geringste Sensitivität (0,26, bei einer Spezifität von 90 %) aufwies.

Bei kritischer Betrachtung könnte dies inhaltlich natürlich auch ein Indiz dafür sein, dass mittels der Vor-Klassifikation zu viele genuin psychisch kranke Patienten als „symptomverfälschend" beurteilt wurden, d.h. zu viele falsch-positive Klassifikationen erfolgten.

Die Autorin konnte zudem in sechs der neun Restructured-Clinical-Skalen signifikante Gruppendifferenzen beobachten, mit Ausnahme der RC-Skalen RC3, RC4 und RC9 (s. Abb. 10, S. 99).

Dieses Ergebnis überrascht zunächst, da die Restructured-Clinical-Skalen nicht zu Validierung von Beschwerde-Überhöhungen konstruiert wurden, sondern um das Ausmaß klinischer Symptome zu erfassen.

Anderseits zeigten bereits Bianchini et al. (2008) [30], dass Patienten mit MPRD (Malingered Pain Related Disabilities) auch durch Auffälligkeiten in den Basisskalen Hy und Hd

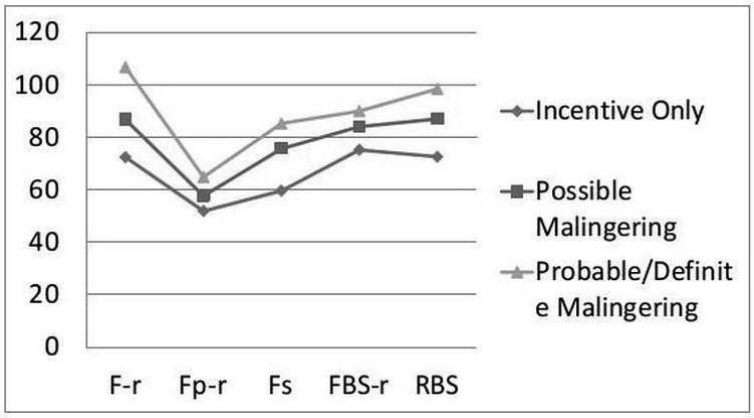


Abb. 9. Mittlere MMPI-2-RF-Overreporting Validitäts-Scores in drei Klassifikations-Gruppen (aus: Anderson (2011) [9])

des MMPI-2 identifizierbar waren - also ein im Vergleich zur Normpopulation außergewöhnliches Maß an Überhöhung hypochondrischer und hysterischer Symptome aufwiesen.

Auch die Studiengruppe von Wiggins et al. (2012) [316], die eine Probandengruppe von 2.275 Klägern untersuchten, beobachteten in der mit „Overreporting“ klassifizierten Studiengruppe signifikant höhere Scores in allen Restructured-Skalen des MMPI-2-RF gegenüber den Antworten der als nicht-aggravierend beurteilten Probanden.

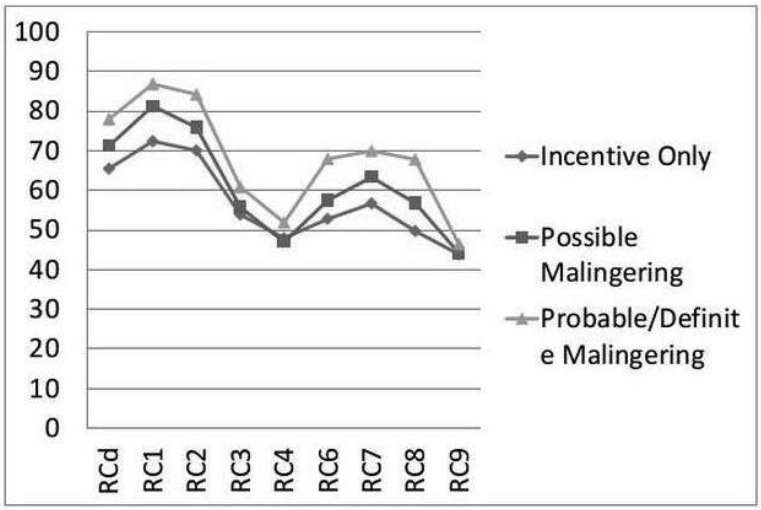


Abb. 10. Mittlere MMPI-2-RF-Clinical-Scale-Scores in drei Klassifikations-Gruppen (aus: Anderson (2011) [9])

Zusammenfassend bilden sich somit auch in den MMPI-2-Basisskalen bzw. den Restructured-Clinical-Skalen des MMPI-2-RF Beschwerde-Überhöhungen ab.

Henry et al. 2012 [131]) zeigten mittels der im MMPI-2-RF auf 11 Items reduzierten Skala HHI-r bei Cutoff-Werten ≥ 7 eine gute Sensitivität von 68,9 % zur Aufdeckung von Overreporting bei Klägern mit Kopfverletzungen im Vergleich zu Patienten mit Verletzungen ohne Kompensationsansprüche.

Somit lässt sich auch mit einer kurzen Liste besonders trennscharfer Items (z.B. HHI) eine vergleichbar gute Detektion von Overreporting leisten wie mit längeren Listen (z.B. RBS, FBS-r). Dabei bleibt unklar, ob dieselbe Detektionsgüte des auf kognitive Störungen zielenden Subtests HHI-r auch für Patienten mit chronischen Schmerzen bei nicht-neurogenen Erkrankungen gilt.

Nur wenige Studien versuchten, die MMPI-2-RF-Skalen als *Prädiktoren der Therapieprognose* von Patienten zu verwenden.

So wurden in einer Studie von Sellbom et al. (2008) [270] die RC-Skalen zusätzlich zu anamnestischen Daten genutzt, um das Rückfall-Risiko bei 483 männlichen Haftinsassen (mit juristischen Vorstrafen, Abhängigkeits-Problemen, mit psychotherapeutischen Vor-Behandlungen, Aggressions-Auffälligkeiten und Gewalt in der Partnerschaft) zu bemessen, nachdem sie an einem psychoedukativen Interventions-Programm teilgenommen hatten. Auffälligkeiten der Untersuchten in den RC-Skalen des MMPI-2-RF RC 4 (antisoziales Verhalten) und RC9 (hypomanisches Verhalten) zeigten ein relativ höheres Rückfallrisiko an.

In einem stationären psychotherapeutischen Behandlungsprogramm mit 179 Patienten lies sich mit der MMPI-2-RF-Skala RC2 (Depression) ebenso gut wie mit der MMPI-2-

Basisskala D eine geringere Therapieerfolgsquote in einem 36-Monats-Follow-up zeigen (Scholte et al. 2012 [263]).

In einer Studie mit traumatisch hirnverletzten Patienten korrelierte das Profil der MMPI-2-RF-Validitätsskalen positiv mit einer geringen Performance in den kognitiven Beschwerdenvalidierungstests und einer höheren Bereitschaft, Symptome verdeutlicht anzugeben (Thomas et al. 2009 [297]). Gleichzeitig korrelierte das Profil der MMPI-2-RF-Validitätsskalen reziprok mit dem Ausmaß an hirnorganischer Schädigung. Auch Auffälligkeiten in den klinischen RC-Skalen konnten zur Bestätigung einer Beschwerdenaggravation eingesetzt werden, mit Ausnahme der RC-Skala 3 (Cynicism). Eine konkrete Untersuchung, ob die als aggravierend identifizierten Patienten auch längerfristig die geringeren Therapieerfolge aufwiesen, erfolgte in der Studie jedoch nicht.

Nur wenige Studien untersuchten bislang Vorhersage-Möglichkeiten der MMPI-2-RF-Skalen als Prädiktoren einer Therapieprognose und verwendeten hierzu meist die Restructured-Clinical-Scales. Keine Studie untersuchte bislang, inwieweit eine durch die Validitätsskalen nachgewiesene Aggravation den Therapieerfolg bei Patienten mit chronischen Schmerzen begrenzt.

Aguerrevere (2010) [2] untersuchte mittels retrospektiv ausgewerteter MMPI-Daten 847 konsekutiv zwischen 1998 und 2008 in einer Praxis im Südosten der USA behandelte Rückenschmerz-Patienten mit einer sozialmedizinischen Problematik (Arbeitslosenkompensation oder juristisches Verfahren) im Alter von 18 bis 59 Jahren. Die Schmerzdauer betrug bei den Probanden zwischen 6 Monaten und 15 Jahren. In der Studie wurden Testergebnisse mit T-Scores von mehr als 80 in den MMPI-2-Prüfskalen TRIN oder VRIN oder Fehlende-Scores von mehr als 29 ausgeschlossen. 608 MMPI-2-Erhebungen gingen in die Studie ein, das Durchschnittsalter der Probanden betrug 42,4 Jahre (SD = 8,8), 64,6 Prozent waren Männer. Die durchschnittliche Schmerzintensität betrug entsprechend einer VAS (von 0 = „schmerzfrei“ bis 10 = „maximaler Schmerz“) 6,7 (SD = 1,9). Weniger als die Hälfte der Probanden (40,5 %) hatten eine objektivierbare somatische Schmerzursache. Voroperationen waren bei einem Viertel bis einem Drittel der Patientengruppe erfolgt. 82,6 % der Proban-

den erhielten eine Arbeitslosenkompensation, 15,3 % der Patienten strebten eine Kompensation mit einem juristischen Verfahren an. Über die Hälfte der Untersuchten war durch einen Rechtsanwalt vertreten (55,9 %).

Die Probanden wurden mittels der Kriterien nach Bianchini (2005) [31] klassifiziert. Als BV-Verfahren wurden folgende sieben, zusätzlich verfügbare Test-Daten der Neurokognitiven Domäne verwendet: (1) Test of Memory Malingering (TOMM), (2) Portland Digit Recognition Test, (3) Word Memory Test, (4) Wechsler Adult Intelligence Scale, (5) California Verbal Learning Test, (6) Millon Multiaxial Clinical Inventory-III und (7) Minnesota Multiphasic Personality Inventory-II (nur bei den als MPRD klassifizierten Probanden). Die entsprechenden Cutoff-Werte wurden aus der vorhandenen Literatur abgeleitet.

Zusätzlich war es aufgrund der ärztlichen Voruntersuchungen möglich, auch Malingering auf der körperlich-behavioralen Ebene einzuschätzen. Hierzu verwendete der Autor vier Kriterien: (a) Inkonsistenz zwischen dem Patientenverhalten während der Untersuchung und ihrem Verhalten in Situationen, in denen sie sich unbeobachtet fühlten, (b) Funktionelle, nichtorganische Befunde während der physischen Untersuchung (mit Ausnahme der Funktionellen Kapazitäts-Evaluation, s. Kriterium d), (c) Inkonsistenz zwischen dem Patientenbericht über die Symptomatik oder Vorgeschichte und der Akten-Anamnese und (d) Feststellung submaximaler Anstrengung, Symptom-Verdeutlichung oder funktioneller, nicht-organischer Befunde in der Funktionellen Kapazitäts-Untersuchung. Als zusätzliche Erhebungen lagen dem Autor Daten der Schmerz-Katastrophisierungs-Skala (Sullivan et al. 1995 [287]) sowie des Pain-Disability-Index der Probanden vor.

Zur Klassifikation der Probanden wurden diese als mit

- ”Non Malingered Pain-Related Disability”‘ beurteilt, wenn in allen BV-Verfahren keine Auffälligkeiten festgestellt wurden;
- Probanden wurden als „Personen mit sozialmedizinischer Konfliktsituation“ klassifiziert, wenn ein einzelner suspekter Befund auf der körperlich-behavioralen Ebene festgestellt wurde oder mehr als eine suspekte Inkonsistenz auf dieser qualitativen Ebene beobachtbar war, aber keine Auffälligkeit in den Performance-Validierungstests vorlag;
- Als Probanden mit „Possible Malingered Pain-Related Disability“ wurden Probanden mit Test-Auffälligkeiten beurteilt, die nicht alle Malingering-Kriterien erfüllten, z.B.

Fälle mit Auffälligkeiten in nur einem Performance-Validierungstest oder mit zwei oder mehr qualitativen Inkonsistenzen in der körperlich-behavioralen Domäne;

- Probanden wurden als „Personen mit Probably Malingered Pain-Related Disability“ identifiziert, wenn zwei oder mehr positive, Performance-Validierungstest-Befunde (PVT) für Malingering vorlagen oder der Proband zwei oder mehr qualitative Inkonsistenzen zusammen mit einem oder mehr positiven PVT-Befunden zeigte;

- Patienten wurden als „Personen mit Definite Malingered Pain-Related Disability“ klassifiziert, wenn alle entsprechenden Kriterien erfüllt waren, vor allem in einem oder mehr der kognitiven Tests ein Testperformance unter Zufallsniveau bekannt war.

Aufgrund dieser Kriterien erfüllten nach Aguerrevere (2010) [2] 37,2 % der Untersuchten die Kriterien für Non-MPRD, 29,9 % der Probanden wurden als Possible-MPRD identifiziert, mit Probably-MPRD wurden 25,3 % der Probanden and als Definitiv-MPRD 7,6 % klassifiziert.

Aguerrevere (2010) [2] untersuchte im Anschluss daran aufgrund der vorliegenden MMPI-2-Daten sowie die MMPI-2-RF-Auswertungen den Patientenpool mittels einer Zwei-Stufen-Cluster-Analyse hinsichtlich übergeordneter Profil-Pattern; die in anderen Studien verwendeten Methoden hierarchischer Clusteranalysen und K-Means-Analysen wurden bei großen Samples als weniger effizient beurteilt.

Aufgrund 50 nach einer Zufallsordnung der Probanden durchgeführten Clusteranalysen mittels MMPI-2-Basisskalen einschließlich der -Validitätsskalen ermittelte Aguerrevere (2010) [2] zwei übergeordnete Profilpattern: Patienten im ersten Cluster zeigten vor allem Auffälligkeiten entsprechend des Neurotic-Triad-Pattern (Auffälligkeiten in den drei ersten MMPI-2-Basisskalen Hypochondrie, Depression und Hysterie/Konversion), Patienten im zweiten Cluster zeigten ein General-Elevated-Profil mit Auffälligkeiten in den ersten drei Basisskalen sowie (viermal so häufig wie in Gruppe 1) Auffälligkeiten in den Skalen 4, 6, 7 und 8 (Psychopathie, Paranoide Ideen, Psychasthenie und Schizophrenie). Gruppe 1 gehörten mehr Patienten mit hohen Korrektur- und Lie-Werten an (Problem-Bagatellisierung), während Probanden der Gruppe 2 insgesamt mehr außergewöhnlich-hohe Scores aufwiesen, die sogar höher in den drei ersten Basisskalen waren als die der Patienten der Gruppe 1 (Neurotic-Triad-Pattern).

Aguerrevere (2010) [2] führte im Folgenden weitere Clusteranalysen mit den MMPI-2-RF-Basisskalen und den MMPI-2-RF-Validitätsskalen aufgrund der mittels MMPI-2 gefundenen Zwei-Cluster-Lösung durch. Hier zeigten sich nach ebenfalls 50 Analysen mit randomisierten Probanden-Sortierungen in den zwei mittels MMPI-2 gefundenen Profilclustern folgende distinkte Profilpattern mittels MMPI-2-RF-Daten:

- Patienten im ersten Cluster zeigten ausschließlich Auffälligkeiten in der MMPI-2-RF-Basisskala 1 (RC1 = Somatic Complaints).

- Patienten im zweiten Cluster wiesen folgende Auffälligkeiten auf: zwei Durchschnittswerte lagen im suspekt-auffälligen Score-Range (Werte der Validitätsskalen Fs and FBS-r), ein Score lag im möglich invaliden Bereich (Skala F-r), fünf Scores waren überhöht (RCd - Demoralisation, RC2 - Low Positive Emotions, RC6 - Ideas of Persecution, RC7 - Dysfunctional Negative Emotions, RC8 - Aberrant Experiences) und ein Score lag in einem sehr hohen Bereich (Skala RC1 - Somatic Complaints). Die auffälligsten Gruppenunterschiede zeigten sich in den Basisskalen RCd, RC6, RC7 und RC8. Hier wiesen nur 10 Prozent der Probanden der Gruppe 1 Auffälligkeiten auf, während 60 % der Probanden mit General-Psychopathological-Profil (Cluster II) überhöhte T-Werte aufwiesen.

Die mittels ANOVA und Chi-Quadrattest geprüften Unterschiede beider Gruppen zeigten, dass Probanden mit Neurotic-Triad-Pattern eine kürzere Dauer ihrer Symptomatik als die Probanden mit General-Pathological-Profil aufwiesen sowie ein überwiegend höheres Bildungsniveau zeigten. In der Gruppe mit Neurotic-Triad-Pattern waren signifikant mehr Männer als in der Gruppe 2 (General-Pathological-Profil). Es bestanden keine Gruppen-Unterschiede hinsichtlich der Verletzungsart, der Schmerzlokalisation und der Schmerzätiologie sowie bezüglich des sozialmedizinischen Status.

Hingegen unterschieden sich die Gruppen signifikant hinsichtlich der Malingering-Klassifizierungen: Entsprechend Odds-ratio-Analysen wiesen Probanden mit General-Pathological-Profil 10,5-mal häufiger als Probanden der ersten Gruppe eine MPRD-Klassifizierung auf ($p < 0{,}001$). Unter Einbezug der zusätzlichen MMPI-2-Validitätskriterien für Malingering erhöhte sich diese Wahrscheinlichkeit einer Positiv-Klassifikation auf 27,8 gegenüber der Probanden der Gruppe mit Neurotic-Triad-Pattern. Auch wiesen die Probanden mit General-Pathological-Profil signifikant höhere Schmerzintensitätswerte auf als die Gruppe-1-Patien-

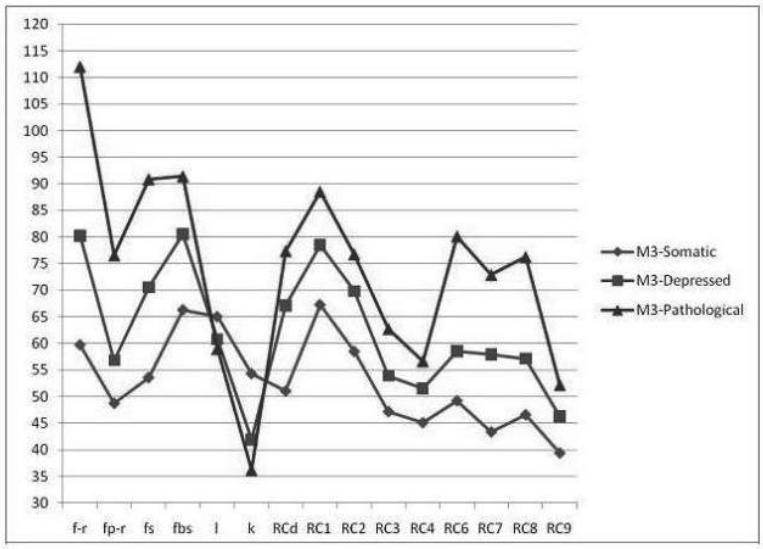


Abb. 11. Mittlere MMPI-2-RF-Validitäts-Skalen- und Clinical-Scale-Profile dreier clusteranalytisch ermittelter Overreporting-Gruppen (aus: Aguerrevere (2010) [2])

ten und hatten signifikant höhere Angaben in der Schmerz-Katastrophisierungs-Skala sowie im Pain-Disability-Index.

In der Auswertung der Basisskalen des MMPI-2-RF zeigte sich, dass die Subskalen RC3 (Cynism) und RC4 (Antisocial Behavior) am wenigsten sensitiv für Gruppenunterschiede waren (Anmerkung: zur geringen Assoziation der MMPI-2-RF-Subskala RC3 (Cynism) mit der MMPI-2-Subskala Hy siehe auch: Thomas & Youngjohn 2009 [297], Butcher et al. 2006 [50], Simms et al. 2006 [275], Rouse et al. 2008 [251], Van der Heijden et al. 2008 [304]). Thomas & Youngjohn (2009) [297] postulierten, die Skala RC3 des MMPI-2-RF weise keine Validität zur Diagnose von Somatisierung und Malingering auf. Andere Autoren hingegen berichteten eine Assoziation von RC3 (Cynism) mit der FBS-Validitätsskala des MMPI-2 (Downing et al. 2008 [74]). Ingram et al. (2011) [139] postulierten ferner, die restrukturierte Skala RC3 (Cynism) sei mit dem Motiv assoziiert, willentlich zu lügen und zu manipulieren.

Aguerrevere (2010) [2] diskutiert in seiner Arbeit, dass die gefundenen Unterschiede beider Patienten-Cluster auch eine geringere Therapieerfolgsprognose für Patienten mit General-Elevated-/ Pathological-Profil gegenüber Patienten mit Neurotic-Triad-Pattern vermuten lassen. Diese Annahme kann jedoch mit dem hier gewählten Studien-Ansatz nicht geklärt werden.

In einer weiteren Clusteranalyse (Methode 3, s. Abb. 11) analysierte Aguerrevere (2010) [2] die Klassifikation seiner Daten chronischer Rückenschmerz-Patienten, indem er die Da-

ten der MMPI-2-RF-Basisskalen und der MMPI-2-RF-Validitätsskalen direkt einer clusteranalytischen Auswertung unterzog. Durch die wiederholten, randomisiert durchgeführten Analysen kristallisierte sich hierbei eine Drei-Cluster-Lösung als die valideste Klassifikation heraus:

- Eine erste Patientengruppe wies überwiegend Auffälligkeiten in der Skala Somatic--Complaints (RC1) auf,
- Eine zweite, bekannte Subgruppe zeigte ein sogenanntes General-Pathological-Profil mit einer überhöhten Validitätsskala Fp-r (Infrequent Psychopathology), drei überhöhten T-Werten > 85 in den Validitätsskalen F-r (Infrequent Responses), Fs (Infrequent Somatic Complaints) and FBS-r (Fake Bad Scale), vier überhöhten Werten in den Basisskalen RCd (Demoralisation), RC2 (Low Positive Emotions), RC7 (Dysfunctional Negative Emotions) and RC8 (Aberrant Experiences) und zwei weiteren Skalen mit sehr starken Auffälligkeiten (RC1 - Somatic Complaints und RC6 - Ideas of Persecution).
- Eine dritte Patientengruppe wurde in den MMPI-2-RF-Analysen als sogenannte Depressed-Subgroup identifiziert, mit Skalen-Auffälligkeiten in der FBS-r-Validitätsskala, drei überhöhten Scores in den Subskalen Demoralisation (RCd), Somatic Complaints (RC1) und Low Positive Emotions (RC2). Insofern ergab die Analyse für den Patientenpool von Rückenschmerzpatienten mit Kompensationsmotiv anhand der MMPI-2-RF-Daten eine Drei-Cluster-Lösung.

Erneut zeigten die Gruppenwertvergleiche hinsichtlich der sonstigen Merkmale in der Gruppe 1 (Somatic Complaints) signifikant niedrigere Schmerzintensitätswerte, ein höheres Ausbildungsniveau und eine signifikant geringere Häufigkeit von Malingering-Klassifikationen als in der Cluster-Gruppe 3 (General-Pathological-Pattern). Alle drei Gruppen unterschieden sich signifikant hinsichtlich der Klassifikation als Malingered Pain-Related Disability MPRD ($p \leq 0{,}001$). Mehr als die Hälfte der Probanden der Gruppe 3 mit *Pathological-Pattern* (59-71 %) wurde als beschwerdeüberhöhend klassifiziert, gegenüber 37-43 % solcher Patienten mit Overreporting in der zweiten Gruppe mit *Depressed-Pattern* und nur 18-20 % der Patienten in der Gruppe 1 mit sog. *Somatic Complaints*.

Angesichts einer so extrem hohen Anzahl von MPRD-Klassifikationen ist sicherlich zu diskutieren, ob sich hier Overreporting als Ausdruck von Malingering widerspiegelt, oder

überhöhte Angaben eher ein Spiegel der Psychopathologie der Untersuchten ist. Insofern besteht auch weiterhin die Frage der Spezifität und des prognostischen Aussagewertes der MMPI-2-RF-Basisskalen wie auch der Diskriminations-Güte der jeweiligen einzelnen ValiditätsSubskalen.

In der Studie von Aguerrevere (2010) [2] wurde kein Normal-Cluster von MMPI-2-RF-Daten ohne pathologische Auffälligkeit gefunden; eine mögliche Erklärung wäre die Auswahl der Daten von ausschließlich Rückenschmerzpatienten, zusätzlich mit sozialmedizinischem Kompensationsmotiv. Es ist zu vermuten, dass es in einem Pool von Patienten mit chronischen Schmerzen mit allen möglichen Schmerzdiagnosen auch eine solche (möglicherweise große) Subgruppe ohne MMPI-2-RF-Auffälligkeiten identifiziert werden könnte.

Nach eigenen Erhebungen des Ev. Krankenhaus Bielefeld erhalten nur ca. 40 Prozent der Patienten die Hauptdiagnose Rückenschmerz. Die überwiegende Zahl der Patienten gehört anderen MASK-S-Diagnosegruppen an (Diagnose-Statistik Ev. Johannes-Krankenhaus Bielefeld, Patientendaten der Jahre 2000 - 2001, N = 1240). Auch die ausgeprägt hohe Anzahl von Patienten mit MPRD-Klassifikation erklärt sich bei Aguerrevere (2010) [2] nur aus der Probanden-Auswahl mit ausschließlich sozialmedizinischem Kompensationsmotiv.

Insofern stellen sich für konsekutiv in einem medizinisch-psychologischen Behandlungsprogramm behandelte Patienten erweiterte Fragestellungen hinsichtlich einer MMPI-2/MMPI-2-RF-Gruppen-Klassifizierung sowie hinsichtlich der Häufigkeit und Detektion von Malingering. Auch die vergleichende Diskriminanzgüte der einzelnen MMPI-2-RF-Validitätsskalen (welchen Anteil haben sie an der Sensitivität der MPRD-Identifikation) wurde mit der Studie von Aguerrevere (2010) [2] nicht geklärt.

Mittels clusteranalytischer Methoden (Aguerrevere 2010 [2]) ließen sich aus einem spezifischen Pool von Rückenschmerz-Patienten (hoher Anteil von 40 % somatoformer Störungen, hohe Arbeitslosenquote von 83 %, hoher Anteil von Rechtstreitigkeiten von 56 %) die Klassifikation von MPRD (Overreporting) mittels MMPI-2-RF beurteilen. Aguerrevere unterschied drei Profilcluster:

Somatic-Complaints-Pattern ohne Symptom-Aggravation, Depressed-Pattern mit leichter Aggravation und Pathological-Pattern mit sicherer Beschwerden-Überhöhung. Die dritte Probandengruppe fiel auf durch: geringes Bildungsniveau, hohe Schmerz-Intensitäts-Angaben, hohe subjektive Beeinträchtigung (Disability-Index) sowie einen hohen Anteil an MPRD-Klassifikationen. Die dritte Patientengruppe wies auch im MMPI-2-RF die deutlichsten Erhöhungen der Validitätsskalen wie auch der meisten Basisskalen auf.

In einer *jüngsten Studie* adaptierten Meyers et al. [201] im Jahr 2013 den Meyers--ValiditätsIndex (MVI oder MI-r) entsprechend der im MMPI-2-RF standardisiert ermittelbaren fünf Validitätsscores nach einer einfachen, gewichteten Summenformel ($T \leq 75 = 0$, $T \leq 89 = 1$, $T \geq 90 = 2$).

Die Autoren stellten bei einem Cutoff des MI-r (revidierter MVI) ≥ 5 bei 145 Patienten mit chronischen Schmerzen bei einer Spezifität von 0,93 eine Sensitivität des MI-r von 0,85 fest. Die Übereinstimmungen der Patientenklassifikationen des MVI mit dem MI-r lagen nach ihren Untersuchungen bei 93,5 %, wobei sich nach logistischen Regressions-Analysen insbesondere die Skalen F-r und Fp-r sowie die FBS-r-Skala als signifikante Prädiktoren für Malingering erwiesen.

Eigentlich erübrigt sich angesichts einer so hohen Sensitivität bei entsprechender Spezifität jede weitere Suche nach noch besser differenzierenden Verfahren. Wie Bianchini et al. (2005) [31] diskutierten, ist eine perfekte Sensitivität eines Validierungsverfahrens vermutlich nicht zu erreichen. Die Autoren halten es für nicht möglich, ein Detektions-Verfahren gegen alle Möglichkeiten der Verfälschung zu schützen. Die Schwierigkeiten liegen in der unterschiedlichen Verfälschbarkeit der Detektionsmethoden begründet, aber auch in ihrer Erkennbarkeit durch Probanden sowie der Trainierbarkeit („Coaching“) von Testverfahren, um einer Entdeckung zu entgehen.

Dennoch kommt der Summationsscore der fünf optimiertesten Validierungsverfahren des MMPI-2-RF nach Meyers et al. (2013) [201] dieser Forderung nach Optimierung der Sensitivität am nächsten.

Allerdings wurde zur externen Klassifikation für Overreporting in der Studie nach Meyers et al. (2013) [201] eine Vielzahl (insgesamt neun) Verfahren zur Beschwerden-Validierung verwendet. Hierbei handelte es sich *ausschließlich* um sog. kognitive Performance-Validationtests (PVTs). Klinische Verhaltens-Beurteilungen oder andere, externe Symptom-Validierungstests (SVTs) zur Erfassung überhöhter psychopathologischer Symptome wurden nicht verwendet.

Insofern ist Mi-r möglicherweise hochspezifisch zur Erfassung von Überhöhungen kognitiver Symptome, jedoch weniger sensibel zur Aufdeckung von Aggravationen psychopathologischer Symptome.

Möglicherweise wäre die Detektionsgüte des MI-r weniger eindeutig ausgefallen, wenn weitere unabhängige Beurteilungs-Methoden zur Klassifikation von Malingering herangezogen worden wären (z.B. tatsächlich ersichtliches Aggravationsverhalten in Beobachtungs-Situationen oder aber Beschwerdeangaben in externen Abfragen seltener psychopathologischer Symptome).

Zum zweiten wurden in der Studie nur 43 Patienten mit chronischen Schmerzen ohne Klageverfahren im Vergleich zu 102 Patienten mit chronischen Schmerzen mit Klageverfahren auf eine Berentung verglichen. Zum einen ist die Kontrollgruppe ohne Klageverfahren sehr klein und dadurch fraglich repräsentativ, zum anderen wurden in klinischen Kontexten häufig anzutreffende Probanden mit ebenfalls grenzwertigen externen Motiven für Overreporting nicht in die Studie einbezogen (z.B. Patienten mit längerer Dauer der Arbeitsunfähigkeit und drohender Aussteuerung durch die Krankenkasse, Probanden mit beabsichtigter Rentenantragstellung, Probanden mit zeitbefristeter auslaufender Rente, Probanden mit Begehren einer höheren Einstufung des Grades an Behinderung).

Zusammenfassend stellt die Studie von Meyers et al. (2013) [201] einen ersten, vielversprechenden Ansatz dar, durch eine systematische Kombination verschiedener Validierungs-Methoden die Sensitivität der Detektion von Overreporting zu erhöhen. Limitierungen dieses Ansatzes bestehen in der Beschränkung auf ausschließlich kognitive PVTs als externe Validitätskriterien sowie der kleinen Zahl untersuchter Patienten.

1.11 Beschwerdenvalidierung mittels BHI-2

Der BHI-2 (Battery of Health Improvement 2) ist ein 217 Items umfassendes psychologisches Inventar, das insbesondere zum Assessment von Patienten mit chronischen Schmerzen entworfen wurde. Es wird mit Hilfe von Papier und Bleistift durchgeführt und ist in 30-40 Minuten bearbeitbar. Der BHI 2 wurde über 18 Jahre fortentwickelt und liegt zur Zeit in der sechsten Version vor. Herausgeber ist Pearson Assessments, ebenfalls Herausgeber des MMPI. Zweck des BHI-2 ist es, die Wechselwirkung zwischen medizinischen und psychologischen Faktoren zu erfassen.

Zur Validierung wurden in den USA Daten in 106 Städten in 36 US-Staaten gesammelt. Die Validierungsstichprobe setzt sich aus zwei vergleichbaren (*ge-matchten*) normativen Stichproben von US- Volkszählungsdaten zusammen. Ein Sample setzt sich aus Rehabilitations-Patienten zusammen, das andere Sample sind Normpersonen. Das BHI-2-Studienprojekt war die größte einzelne Forschungsstudie zu psychologischen Faktoren bei Rehabilitations-Patienten. Dabei wurden u.a. folgende Erhebungen einbezogen: Assessment durch BHI-2, MMPI-2, MCMI-III (Millon Clinical Multiaxial Inventor, Millon et al. 2006 [205]), die Tennessee- Selbstkonzeptskala (Fitts 1965 [90]), der Minnesota-Satisfaction-Questionnaire (Weiss et al. 1967 [311]), die Toronto Alexithymia-Skala (Bagby et al. 1994 [19], dt. Fassung: Kuofer et al. 2001 [161]), der McGill Schmerzfragebogen (Melzack 1975 [187], Melzack 1987 [188], Melzack 2005 [189]), eine projektive Schmerzzeichnung sowie weitere medizinische und andere Patienten-Informationen.

Zudem wurde mit der BHI-2-Validierungsstudie die Schmerzratingskala in den USA nationalweit normiert und validiert. Mittels des BHI-2 können aber auch Informationen zu Schmerzbeschwerden mit den Diagnosen in Bezug gesetzt werden, so dass beispielsweise ein durchschnittliches Profil von Patienten mit chronischen Rückenschmerzen identifiziert werden kann. Aber auch Unterschiede älterer und junger Patienten mit derselben Diagnose oder Angaben von Patienten mit und ohne sozialmedizinische Kompensationsverfahren können analysiert werden. Ferner könnte das Ausmaß der Depressivität und Angststörungen bei Patienten mit chronischen Schmerzen mittels der BHI-2-Daten stichprobenspezifischer benannt werden.

So ermöglichen die ermittelten Vergleichsnormen die Einschätzung, in wieweit ein Proband mehr depressive Symptome als andere vergleichbar kranke Personen angibt. Bereits

1989 berichtete Maruta [180] weit höhere Angst- und Depressionsangaben chronisch kranker Patienten gegenüber der Allgemeinbevölkerung. Der BHI-2 ermöglicht dem Behandler zusätzlich in einer Art **Kreuzvalidierung** die Symptomatik des Patienten mit den psychologisch zu erwartenden Symptomen in derselben Patientengruppe abzugleichen. Eine der Stärken des BHI-2 sind hierbei Normdaten zu folgenden Patientenmerkmalen: Unfälle und Verletzungen, Neurogene Erkrankungen, wie z.B. Karpaltunnel-Syndrom, CRPS, Fibromyalgie-Syndrom, chronische Kopfschmerzen, chronische Rückenschmerzen und andere Schmerz-Symptomatiken, Abhängigkeits-Erkrankungen, aber auch sozialmedizinische Aspekte der Erkrankung.

Der BHI-2 wurde konstruiert, um die interdisziplinäre Behandlung medizinischer Patienten zu vereinfachen. Dazu wurden unterschiedliche Subskalen des BHI-2 konstruiert: Die Defensiveness Skala (DEF) soll die bewusste oder unbewusste Tendenz von Patienten erfassen, Informationen entweder positiv oder negativ zu beeinflussen und die Validität der Angaben einschätzen zu können. Die Skala Self Disclosure (DIS) soll die Offenheit von Probanden erfassen, Angaben über ihre Gedanken, Gefühle und Verhaltensweisen abzugeben. Hierzu wurden 100 Items des BHI-2 zu den Bereichen Depression, Angst und Feindseligkeit, aber auch zu Borderline-Symptomen, chronischer Fehlanpassung, Substanzmissbrauch und Durchhalteappellen in der DIS-Skala zusammengefasst.

Zudem wurde eine 4-Items-umfassende Validitätsskala in den BHI-2 integriert. Zu diesem Zweck wurden jene Items des Original-BHI ausgewählt, die ein zufälliges Antwortverhalten oder Hinweise auf desorientiertes Denken identifizieren könnten. Initial wurden dabei solche Items selektiert, die selten überhöht angekreuzt wurden. Von diesen Items wurden fünf besonders bizarre, inhaltlich nicht mögliche oder auf schwere Psychopathologie hinweisende Items ausgewählt („Ich bin auf das Glas in Glasbehältern allergisch“, „Ich mag das Gefühl von Schmerzen“, „Es ist schon einmal vorgekommen, dass ich von Regierungsspionen verfolgt wurde“ und „Ich bin wahrscheinlich der einzige Mensch, der je gelebt hat“).

Der BHI-2 wurde mehrfachen Überarbeitungen unterzogen. Bei der Auswahl der endgültigen Items des BHI-2 entfiel ein weiteres Item der BHI-Vorgänger-Versionen („Ich bin ungewöhnlich empfindsam für Schwerkraft“), das von Patienten mit chronischen Schmerzen zu häufig bejaht wurde. Zudem wurde das Kriterium für eine Auffälligkeit in den vier Items auf mindestens drei Items im BHI-2 erhöht (mit „Übereinstimmung“ und „Starke Übereinstimmung“ angekreuzt), um die Sicherheit des Untersuchers zu erhöhen. Ziel des resultierenden

Validitätsindex ist die Identifizierung berechtigter Zweifel an der Zuverlässigkeit und Interpretierbarkeit des BHI-2. Allerdings könnte das 3. Item („Verfolgungsgedanken“) durchaus Realitätsgehalt haben, wenn Probanden sich in einem Rechtsstreit-Kontext befinden.

Bei der Überprüfung der Skalenkonstruktion wurden von 203 Probanden 99 einer sog. Fake-Bad-Gruppe und 104 Probanden einer Fake-Good-Bedingung zugeordnet. Probanden der Simulationsgruppe wurden instruiert, im Fragebogen Symptome möglichst so zu übertreiben, als wenn sie eine finanzielle Kompensation erreichen wollten; gleichzeitig sollten die Symptome nicht zu sehr übertrieben werden, um nicht als Simulanten aufzufallen. Die Fake-Good-Probanden sollten Symptome so gering wie möglich darstellen, so dass ihre Antworten ein unrealistisch positives Selbstbild widerspiegelten, als wenn sie beispielsweise Angst hätten, vom Abschluss einer Lebensversicherung ausgeschlossen zu werden. Auch diese Probanden sollten sich jedoch nicht so unglaubwürdig darstellen, dass sie durch Dissimulation auffielen. Ein Gruppenvergleich mittels univariater Varianzanalyse zwischen der Fake-Good-Gruppe, der Fake-Bad-Gruppe, einer Patientengruppe und dem Norm-Sample ergab signifikante Unterschiede sowohl für die Simulationsskala (DIS), als auch die Dissimulations-Skala (DEF).

Die weiteren Skalen wurden unter klinischen Gesichtspunkten entwickelt: die Skala Körperliche Beschwerden (SOM) soll die Tendenz von Patienten erheben, diffuse somatische Beschwerden anzugeben, die häufig mit Stress und somatoformen Störungen verbunden sind. Die Skala Pain Complaints (PAIN) erhebt auf einer Skala von 0 bis 10 die Schmerzstärke in 10 Körperregionen. Zu diesem Zweck konnten in der BHI-2-Validierungsstudie erstmals nationale Normwerte der USA erhoben werden. Die 3. Skala Functional Complaints (FNC) soll das Erleben oder die Darstellung von Funktionsdefiziten und Disability bei Patienten mit chronischen Schmerzen erfassen. Die 4. Skala Muscular Bracing (MB) soll die Tendenz von Patienten erfassen, auf Stress, Krankheit und Verletzung mit erhöhter Muskelanspannung zu reagieren (entsprechend einer nach Schachter & Singer 1962 [256] sog. Kampf- oder Flucht-Bereitschaft). Die 5. Depression (DEP) und die 6. Skala Anxiety (ANX) sollen entsprechende Krankheitsaspekte von Verlust, Trauer und erlernter Hilflosigkeit sowie Angst, Sorge und physische Angstsymptome erheben.

Die 7. BHI-2-Skala Hostility (HOS) soll mit der Erkrankung verbundene feindselige Tendenzen von Patienten erheben, die mit Ärgergefühlen, Aggressivität und Zynismus verbunden sind. Entsprechende Zusammenhänge zur subjektiven Arbeits- und Erwerbsunfähigkeit

wurden wiederholt untersucht; so berichten mehrere Studien einen Zusammenhang zwischen Frühberentung und berufsbezogener Erschöpfung sowie erhöhter Tendenz zu zynischen Wertkonzepten (Fuchs et al. 2009 [95], Ahola et al. (2009) [4], Von Känel (2008) [148]), wie sie auch im MMPI-2-RF als Basisskala RC3 (Cynism) erfasst werden.

Die 8. Skala BOR wurde konzipiert, um Aspekte einer Borderline-Persönlichkeitsstörung bei den Befragten zu erfassen. Die 9. Skala Symptom Dependency (SYM) untersucht die Tendenz mancher Patienten, mittels der Krankheit Aufmerksamkeit und Zuwendung zu suchen (vgl. Kap. 1.4.1, S. 9). Die 10. Skala Chronic Maladjustment (CHR) soll Schwierigkeiten von Patienten erfassen, eine stabile Lebensführung aufrechtzuerhalten, hinsichtlich Ausbildung, Arbeit, finanziellen Angelegenheiten, interpersonalen Beziehungen und juristischen Konflikten. Die 11. Skala Substance Abuse (SUB) soll das Vorhandensein einer aktuellen oder durchlaufenen Abhängigkeits-Erkrankung klären (vgl. Kap. 1.4.2, S. 12).

Die 12. Skala Perseverance (PER) wurde erstellt, um die Tendenz von Patienten zu untersuchen, ihren Beschwerden mit sogenannten Durchhalteappellen dysfunktional zu begegnen. Die 13. Skala Family Dysfunction (FAM) soll untersuchen, in wieweit der Patient durch seine Familie hinsichtlich gesundheitlicher Belange positiv unterstützt wird. Die 14. Skala Survivor of Violence (SRV) erfasst traumatische Vorerfahrungen, die in der Anamnese von Patienten mit chronischen Schmerzen häufig zu finden sind. Die vorletzte 15. Skala Doctor Dissatisfaction (DOC) soll die Unzufriedenheit von Patienten mit Gesundheitsinstitutionen und ihrer bisherigen Behandlung erfassen. Die 16. Skala Job Dissatisfaction (JOB) wurde konzipiert, um Unzufriedenheit von Patienten mit ihrer beruflichen Situation (hinsichtlich Kollegen, Arbeitgeber, Chef, Kunden und Klienten oder der Tätigkeit selbst) zu erheben, die möglicherweise Ursachen für erhöhte Beschwerdeangaben sein könnten.

Die Testautoren betonen, dass die Validität der BHI-2-Erhebungen dadurch besonders erhöht sein könnte, dass das Wort *Schmerz* in nur 5 der 217 Fragebogen-Items explizit vorkommt. Bruns & Disordio (2000) [42] belegten in einer Studie mit 214 Probanden, die instruiert wurden, den BHI-2 in positiver oder negativer Weise zu verfälschen, im Vergleich mit der BHI-2-Patienten- und der Normgruppe, dass die DIS- und die DEF-Skala varianzanalytisch signifikant zwischen allen vier Gruppen unterscheiden konnte (*subtle fake good, community, patient, subtle fake bad*). Eine Ausnahme bildete der nicht signifikante Unterschied zwischen der positiv verfälschenden Gruppe und der Normgruppe.

Beide BHI-2-Validitätsskalen diskriminierten zudem signifikant zwischen Probanden, die von einem Rechtsanwalt vertreten wurden, und Probanden ohne Rechtsbeistand. Jedoch wurden für die Validitätsskalen keine abgesicherten Cutoff-Werte und keine anderen Validitäts-Merkmale wie Sensitivität und Spezifität berichtet.

Die im BHI-2-Manual berichteten Mittelwerte der Gruppen (Bruns & Disorbio 2004 [43]), zeigten in der Gruppe mit Overreporting deutlich höhere DIS-Werte (DIS-Rohwerte bei 134,96) im Vergleich zu der Patientengruppe, der Normgruppe und der fake-good--Simulanten (MW = 103,18, MW = 90,81 und MW = 76,13). Entsprechend fielen die mittleren Scores der DEF-(Defensiveness)-Skala in umgekehrter Ausprägung in den Gruppen geringer aus (MW = 7,76 vs. MW = 13,38, MW = 16,11, MW = 17,52).

In der folgenden Tab. 2 sind die auf besser vergleichbare T-Werte umgerechneten mittleren Werte in den vier Normgruppen der Testautoren in den Validitäts- und den Basisskalen des BHI-2 tabellarisch aufgelistet.

Tab. 2. BHI-2-Validitäts- und Basisskalen: Mittlere T-Werte der vier Normgruppen (umgerechnete Rohwerte nach Bruns & Disordio 2000, S. 31 [42]

	Fake Bad Group	Patient Group	Community Group	Fake Good Group
DIS	59	50	47	42
DEF	38	49	56	60
SOM	57	50	44	42
PAIN	57	50	44	42
FNC	63	49	42	42
MB	57	50	43	43
DEP	60	50	45	42
ANX	60	53	47	45
HOS	55	49	49	45
BOR	54	50	48	45
SYM	65	50	44	47
CHR	54	50	48	44
SUB	54	49	49	46
PER	45	49	51	58
SRV	48	50	48	44
FAM	55	50	48	46
JOB	54	50	57	59
DOC	64	50	39	41

Zusammenfassend unterstützen diese Resultate die Validität und klinische Nützlichkeit der neuen MMPI-2-RF-Symptom-Validierungs-Skalen ebenso wie zumindest tendenziell eine erhöhte Detektionsgüte der BHI-2-Skalen zur Aufdeckung von Antwortstilen mit Overreporting.

Trotz dieser Vorteile wurden bis dato weder der MMPI-2-RF noch der BHI-2 an einer deutsch-sprachigen Normalpopulation untersucht, noch in einem entsprechendem Sample chronischer Schmerzpatienten, so dass bislang keine Basisraten für Overreporting sowie differenzierte Gütekriterien der Validitätsskalen in einer solchen Patientengruppe vorliegen.

1.12 Zusammenhänge zwischen MMPI-2 und BHI-2

Die Entwicklung und Validierung des BHI-2 wurde sehr eng am MMPI-2 und am MCMI-III orientiert, als den in den USA am umfangreichsten untersuchten und am häufigsten eingesetzten psychologischen Testinstrumenten (Bruns & Disorbio 2004 [43]).

Wie Bruns & Disordio (2000) [42] zeigten, korreliert die DIS- und die DEF-Skala signifikant mit allen Validitätsskalen des MMPI-2.

Nach den BHI-2-Validierungsdaten korreliert die DIS-Skala hoch (r = 0,69) mit dem Dissimulations-Index F-K des MMPI-2. Die Disclosure-Skala DIS korreliert laut Testmanual zudem hoch (r = 0,58) mit der MMPI-2-F-Skala (Infrequent Responses) und negativ mit der K-(Correction)-Underreporting-Skala (r = -0,57). Die Defensiveness-Skala (DEF) des BHI-2 korreliert hoch negativ (r = -0,62) mit T-Wert-Erhöhungen in den 10 MMPI-2-Basisskalen. Symptombagatellisierung geht damit vermehrt mit einer geringen Angabe psychischer Störungen im MMPI-2 einher.

Auch zwischen den übrigen BHI-2-Subskalen und den Basisskalen des MMPI-2 stellten Bruns & Disorbio (2004) [43] inhaltsähnliche Zusammenhänge fest. So korreliert die SOM-Skala (Körperliche Beschwerden) des BHI-2 hoch mit den Skalen Hy (Hysterie), Angst- und Depression des MMPI-2 (r = 0,79; 0,64; 0,74). Die Pain-(Schmerzintensitäts)-Skala korreliert hoch mit den MMPI-2-Skalen Hy (Hysterie) und Hd (Hypochondrie) (r = 0,59; 0,58). Ähnliche Korrelationen wurden auch für die Skalen Functional Complaints (FNC) und Muscular Bracing (MB) mit den MMPI-2-Basisskalen Hysterie, Hypochondrie, Depression und Angst festgestellt (r = 0,68 bis 0,59). Die BHI-2-Skala Hostility (HOS) ist eng mit der MMPI-2-Subskala Ärger korreliert.

Die Symptom-Dependency-Skala (SYM) korreliert hoch mit den MMPI-2-Basisskalen Hysterie und Angst (r = 0,54; 0,48). Auch die BHI-2-Skala Chronic Maladjustment (CHR) zeigte eine positive Korrelation (r = 0,46) mit der MMPI-2-Basisskala Psychopathie. Die BHI-2-Skala Substanzabusus war ebenfalls hoch mit der MMPI-2-Subskala „Sucht und Abhängigkeit“ (Admission) korreliert (r = 0,55). Die Skala Durchhalteappelle (PER) korrelierte hoch mit der MMPI-2-Skala ES (Ich-Stärke, r = 0,51). Die BHI-2-Skala Family Dysfunction (FAM) korrelierte mit der MMPI-2-Skala „Family Problems“ (r = 0,70). Die Skala Doctor Dissatisfaction (DOC) zeigte einen moderaten Zusammenhang mit der MMPI-2-Skala „Negative Treatment Indicators“ (r = 0,29). Alle BHI-2-Subskalen zeigten zudem eine stabile Reliabilität; die Werte zur internen Konsistenz waren entsprechend der Studie von Bruns & Disorbio (2004) [43] in der Patientengruppe als akzeptabel (Cronbach's α 0,74), gut bis teilweise exzellent zu bezeichnen (Cronbach's α 0,96).

Die BHI-2 (Battery of Health Improvement 2) stellt eine neuartiges, im deutschen Sprachraum noch nicht validiertes Testverfahren dar, um spezifische bio-psycho-soziale Veränderungen, aber auch Overreporting bei Patienten mit chronischen Schmerzen zu identifizieren. Aus diesem Grund ist von besonderem Interesse, welche Gütekriterien die BHI-2-Validitätsskalen in der Detektion von Overreporting aufweisen und wie ihre differentielle Diskriminanz im Vergleich mit den Validitätsskalen des MMPI-2-RF einzuschätzen ist.

1.13 Zusammenfassung und Herleitung der Untersuchungs-Hypothesen

Schmerzen als subjektives Erleben sind nicht objektiv messbar, sondern nur indirekt über verbale Aussagen und / oder non-verbale Verhaltensmuster zu erschließen (Kröner-Herwig et al. 2011 [160]). Die Ausgestaltung von Schmerzen unterliegt dabei vielfältigen somatischen, psychischen und sozialen Einflussfaktoren (Zimmermann 1996 [328]). Das Schmerzerleben wird von psychischen Alterationen überlagert und im Fall somatoformer Störungen ausschließlich durch psychische Mechanismen ausgelöst und aufrechterhalten (Nilges & Rief 2010 [216]). Zusätzlich können die von Patienten präsentierten Behinderungen und Einschränkungen durch unbewusste, aber auch bewusstseinsnahe negative und (im Fall von Dis-

simulation) positive Antwortverzerrungen verfälscht sein (Dohrenbusch 2009 [69]). Dies erschwert insbesondere die Begutachtung der durch Schmerzen bedingten Funktionseinschränkungen und Behinderungen (AWMF-Leitlinie 2011 [11])).

Das Ausmaß individueller Antwortverzerrungen wird durch externe Motive der betroffenen Personen in unterschiedlichem Grad beeinflusst (Slick et al. 1999 [279]). Externale Einflussfaktoren bestehen meist in einem sekundären Krankheitsgewinn (Erlangen von Vorteilen durch eine überhöhte oder verringernde Symptom-Präsentation). Insofern müssen insbesondere bei Begutachtungen von Patienten mit chronischen Schmerzen solche sekundären Gewinnmomente erkannt werden.

Dies ist umso wichtiger, als Frühberentungen aufgrund psychischer Störungen in den letzten Jahren zunehmend früher im Lebensalter und zunehmend häufiger insgesamt erfolgten (Bundespsychotherapeutenkammer 2014 [44]). Finanzielle Kompensationen und lange Arbeitsunfähigkeitszeiten sollten deshalb das Aggravationsverhalten begünstigen (Wenig et al. 2009 [312]). Angesichts der gesellschaftlichen Folgekosten ist deshalb die Validierung trennscharfer Messinstrumente zur Detektion nicht-authentischer Symptome von vorrangigem sozioökonomischem Interesse. Begutachtungen bedürfen dabei standardisierter Methoden hoher diagnostischer Güte zur Differenzierung zwischen authentischen und überhöht präsentierten Beschwerdeangaben (Dohrenbusch 2009 [69]).

Der Einsatz psychologischer Testverfahren (besonders von Symptom- und Performance-Validierungstests) wird in der Bundesrepublik Deutschland eher kontrovers diskutiert (exemplarisch für die Contra-Position: Dressing et al. 2011 [75]; exemplarisch für die Pro-Position: Dohrenbusch et al. 2008 [72]). Angesichts dieser wichtigen Kontroversen stellt sich die Frage, mittels welches methodischen Vorgehens die Nützlichkeit dieser Verfahren zur Beschwerdenvalidierung empirisch überprüft werden kann.

Nach Bianchini et al. (2005) [31] erfordert eine an modernen Regeln orientierte Beurteilung von Behinderungen durch chronische Schmerzen den Einbezug von mindestens drei Bewertungsebenen: der Domäne kognitiver Symptome, der Domäne somatischer und psychischer Störungen und der Domäne verhaltensbezogener Leistungsdefizite. Diese Argumentation beinhaltet folgende Logik: Wenn chronische Schmerzen multidimensional bedingt sind und aufrechterhalten werden, muss auch die Beurteilung aus der Erkrankung resultierender

Leistungsdefizite mehrdimensional erfolgen (als Erweiterung der sog. Slick-Kriterien, Slick et al. 1999 [279]).

Interaktionsmuster und die ihnen zugrundeliegenden Verstärkungsmechanismen zwischen Patienten mit chronischen Schmerzen und ihren Bezugspersonen (insbesondere Familienangehörige, aber auch Therapeuten, Gutachter etc.) können die Beschwerdepräsentation verstärken oder reduzieren. Es ist deshalb anzunehmen, dass das Ausmaß an Beschwerdeaggravation in Begutachtungs-Situationen stärker ausgeprägt und häufiger ist als unter klinischen Bedingungen, was in einigen Studien Bestätigung fand (Mittenberg et al. 2002 [208]). Nachgewiesene Verfälschungstendenzen von Beschwerden (sog. *Malingering*) könnten dabei das Ausmaß von Beschwerdeüberhöhungen durch sog. *Verdeutlichung* signifikant überschreiten (AWMF-Leitlinie 051, 2011 [11]).

Da Patienten mit chronischen Schmerzen häufig über Aufmerksamkeits-, Konzentrations- und Gedächtnisdefizite klagen, sollte deren Vorhandensein stets zusätzlich zur Schmerzdiagnostik geprüft werden. Zur Erfassung nicht-authentischer kognitiver Leistungsdefizite haben sich Performance-Validierungstests (sog. forced-choice-Verfahren) als hilfreich erwiesen, die kognitive Leistungen unter Zufallsniveau erkennen helfen oder andere, von den typischen Leistungen ähnlicher Probanden abweichende Ergebnisse erfassen (z.B. Word Memory Test WMT, Green et al. 1999 [114]; Medical Symptom Validity Test MSVT, Green 2004 [112] oder DMS48, Barbeau et al. 2004 [23]).

Zur Vor-Klassifikation von Overreporting kognitiver Symptome haben sich insbesondere kürzere, ökonomische forced-choice-Testverfahren wie das DMS-48 oder der Rey15-Item-Memory-Test bewährt. Zum Screening psychischer Symptome eignet sich das SIMS oder die SCL-90-R, während zur Erfassung verhaltenspsychologischer Auffälligkeiten Selbsteinschätzungs-Verfahren zur Funktionsbeeinträchtigung oder aber auch entsprechende Fremd-Einschätzungs-Verfahren zur Beurteilung überhöhter Einschränkungen auf Verhaltensebene geeignet sind.

Negative Antwortverzerrungen auf psychischer Selbstberichts-Ebene fallen vor allem durch Häufung ungewöhnlicher, seltener, bizarr-absurder oder sich widersprechender Symptom-Kombinationen auf (Rogers 2008 [235]). Aber auch undifferenzierte Symptomüberhöhungen lassen sich unter Bezugnahme auf Angaben von Patienten mit ähnlicher klinischer Symptomatik unschwer identifizieren. Bizarre und extrem-seltene Angaben sind mittels spe-

zifischer Itemlisten entsprechender Symptome zu erheben (z.B. SIMS, Cima et al. 2003a [54]; Critical-Items nach Koss-Butcher 1973 [159]; Critical-Items nach Lachar-Wrobel 1979 [162]).

Hingegen erwiesen sich andere, ausführlichere Verfahren als sog. Routineverfahren nicht als praktikabel. So besitzt das Strukturierte Interview für vorgetäuschte Symptome SIRS hohe Detektionsakkuranz, wurde aber bislang nicht in die Deutsche Sprache übertragen und an einer entsprechenden Population normiert (Kool et al. 2008 [156]). Als Interview erfordert das SIRS aber auch einen höheren Durchführungs-Aufwand.

Die NIM-Skala (Negative Impression Management) aus dem PAI (Personality Assessment Inventory) zeigte in vorangehenden Studien ebenfalls gute Detektions-Eigenschaften, die jedoch nicht die Testgüte des SIRS erreichten. Das SIMS (Structured Interview of Malingered Symptomatology) weist im Vergleich zur NIM-PAI-Skala die höchste Korrelation auf (Edens et al. 2007 [81]). Das Verfahren wurde wegen teilweise geringer Spezifität kritisiert, die jedoch durch einen erhöhten SIMS-Summenwert über 16 Punkten in Kombination mit anderen SVTs kompensiert werden kann (Cima et al. 2003a [54], Van Impelen et al. 2014 [303]). Das SIMS ist gemäß allen verfügbaren Studien im deutschen Sprachraum gegenüber dem SIRS und dem PAI das ökonomischste Screening-Verfahren zur Erhebung von verbalem Overreporting. Obwohl ursprünglich nicht zur Beschwerdenvalidierung entwickelt, ist Overreporting auch mittels der Symptom-Checkliste SCL-90-R als ökonomisches Screening-Verfahren zu erfassen (McGuire und Shores 2001 [185], Schneider et al. 2009 [262]).

Die Prüfung der Glaubwürdigkeit verhaltensbezogener Behinderungen durch Schmerzen ist bislang am wenigsten validiert und bedarf meist aufwendiger Prüfsysteme. Eine Aggravations-Klassifikation auf dieser Funktions-Ebene ist somit am wenigsten sicher. Dennoch lassen sich insbesondere durch Inkonsistenzen gekennzeichnete Verhaltensmuster identifizieren, die für eine überhöhte oder bewusst verfälschte Symptom-Darstellung sprechen (z.B. Waddell-Zeichen, Waddell et al. 1980 [308]; EFL, Isernhagen 1988 [141]; AWMF-Leitlinie 030, 2004 [10]).

Aufgrund der Ergebnisse vorausgehender Studien wird angenommen, dass sich in einer Gruppe konsekutiv stationär schmerztherapeutisch behandelter Patienten ein spezifischer Prozentsatz von Patienten befindet, bei denen aufgrund externer Motive ein Aggravations-

verhalten besteht. Dieser Patientenanteil wird amerikanischen Studien zufolge bei 15-17 Prozent erwartet (Rogers et al. 1994, 1998, 2003 [247] [241] [242]).

Diese Patienten mit MPRD sollten sich durch einen multimodalen Algorithmus mit Einsatz von je zwei Validierungsverfahren an Auffälligkeiten hinsichtlich ihrer kognitiven Leistungskapazität in Performance-Validierungstests erkennen lassen, aber auch an spezifischen Verhaltensauffälligkeiten sowie an überhöhten Angaben psychischer Symptome (Bianchini et al. 2005 [31], Anderson 2011 [9]).

Hinsichtlich dieser Vor-Klassifikation wird angenommen, dass bei Einbezug von mindestens drei Domänen zur Erhebung negativer Angabe-Verzerrungen eine exaktere Identifikation von MPRD-Patienten erreichbar ist, als ausschließlich durch Einbezug kognitiver PVTs, wie es in den meisten us-amerikanischen Studien üblich ist.

Diese Annahme basiert auf deutlich unterschiedlichen Basisraten für Malingering in den drei Aggravations-Domänen. So wurde bei 20 % forensischer Probanden und bei bis zu 30 % klinischer Patientengruppen Overreporting psychischer Symptome berichtet (Mittenberg et al. 2002 [208], Rogers et al. 1994 [247], 1998 [241]). Bei Patienten mit chronischen Schmerzen wurde Overreporting somatischer und physischer Symptome bei ca. 30 % der Patienten in der Begutachtung beobachtet (Mittenberg et al. 2002 [208]). Bei zu begutachteten Unfallopfern mit leichten neurologischen Verletzungen waren Auffälligkeiten in den kognitiven PVTs bei 38-41 % der Probanden festzustellen. Da Patienten offenbar in Abhängigkeit von ihrer Symptomatik unterschiedliches Aggravationsverhalten zeigen und auch unterschiedliche Strategien der Antwortverzerrung anwenden (Rogers 2008 [235]), ist zur Detektion von Overreporting ebenfalls ein multimodaler Ansatz angemessener als die Beschränkung auf nur eine Aufdeckungsstrategie.

Sind auf allen drei Verhaltensebenen sichere Anzeichen für nicht-authentische Beschwerdedarstellungen ersichtlich, so kann nach Bianchini (2005) [31] von einer wahrscheinlichen oder gesicherten MPRD (Malingered Pain Related Disability) ausgegangen werden.

Ein Anliegen der vorliegenden Studie ist es, die Nützlichkeit des MMPI-2-RF zur Aufdeckung solcher nicht - authentischer Beschwerdedarstellungen bei Patienten mit extern gesicherter MPRD zu prüfen. Diese Annahme basiert darauf, dass sich in diversen Studien mit Known-Groups-Design bei Patienten ohne chronische Schmerzsymptomatiken sehr gute Detektions-Möglichkeiten von Overreporting mittels des MMPI-2-RF zeigten.

So erwies sich der MMPI-2-RF als geeignet, Overreporting bei psychiatrischen Patienten trennscharf zu identifizieren (Sellbom & Bagby 2010 [269], Sellbom et al. 2012 [273], Rogers et al. 2011 [239]), wie auch Patienten mit neurokognitiven Symptomen mit Kompensationsbegehren aufzudecken (Gass & Odland 2012 [96], Tarescavage et al. 2013 [293], Jones et al. 2012 [144]). Letztere Patientengruppe wies in einer Studie auch teilweise chronische Schmerzsyndrome auf (Gervais et al. 2010 [101]). Ferner war der MMPI-2-RF geeignet, Overreporting bei forensischen Psychiatriepatienten festzustellen (Sellbom et al. 2010 [272], Wiggins et al. 2012 [316], Wygant et al. 2010 [323]) oder auch bei instruierten und nicht instruierten Probanden aufzudecken, die Symptome simulierten (Burchess & Ben-Porath 2010 [45]).

Overreporting bei Patienten mit chronischen Schmerzen wurde nur in wenigen MMPI-2-RF-Studien untersucht, in denen meist ausschließlich kognitive Performance-Validierungstests (PVTs) zur externen Evaluation eingesetzt wurden. Meyers et al. (2013) [201] (vgl. Kap. 1.10.4) untersuchten in ihrer Studie zum Validitätsindex MI-r 43 Patienten mit chronischen Schmerzen mit und 102 Schmerzpatienten ohne Kompensationsbegehren und konnten eine hohe Detektionsgüte dieses Summations-Index der fünf Standard-Validitätsskalen des MMPI-2-RF nachweisen. Allerdings wurden zur Klassifikation von Overreporting bei den Patienten ausschließlich kognitive Performance-Validierungstests in großer Anzahl (n = 9) verwendet.

Aguerrevere (2010) [2] (vgl. Kap. 1.10.4) untersuchte die Häufigkeit definitiver MPRD (Malingered Pain Related Disability) mittels Clusteranalysen der Fragebögen (MMPI-2-RF) von 608 Rückenschmerzpatienten mit Kompensationsbegehren durch externe Validierung, ebenfalls unter Vor-Klassifikation durch sechs kognitive PVTs. Externe Validierungen in den anderen Domänen wurden nicht vorgenommen.

Anderson (2011) [9] (vgl. Kap. 1.10.4) untersuchte die Detektionsgüte der Validitätsskalen des MMPI-2-RF an 169 neuropsychiatrischen Patienten mit Rentenklageverfahren, die auch chronische Schmerzen angaben. Die Probanden wurden vorab, entsprechend dem Ansatz von Bianchini et al. (2005) [31], durch Einsatz von drei SVTs und drei PVTs sowie anhand von Diskrepanzen zwischen den angegebenen und den objektiv beobachtbaren Verhaltenseinschränkungen hinsichtlich Overreporting-Auffälligkeiten klassifiziert. Die Autorin stellte sehr gute Detektionsergebnisse der Validitätsskalen des MMPI-2-RF fest. Jedoch wurde in die Untersuchung keine Vergleichsgruppe von Patienten ohne externe Kompensa-

tionsmotive einbezogen; ferner bezogen sich Hauptdiagnose und Einschlusskriterien nicht primär auf chronische Schmerzen.

Keine Studie zur Detektionsgüte des MMPI-2-RF untersuchte bislang konsekutiv behandelte Patienten mit chronischen Schmerzen mit und ohne externe Aggravationsmotive, mit und ohne somatoforme Symptomatik, bei denen gleichzeitig das Vorhandensein einer MPRD (Malingered Pain Related Disability) entsprechend Bianchini's multimodalem Ansatz unter Einsatz verschiedener PVTs und SVTs in mehreren Aggravations-Domänen (behavioral-, kognitiv und psychopathologisch) untersucht wurde.

Eine solche mehrdimensionale Überprüfung mit konservativen Entscheidungsregeln wäre wichtig, um Patienten mit chronischen Schmerzen mit möglichen Aggravations-Motiven nicht pauschal unter den Generalverdacht von Beschwerdeüberhöhungen zu stellen und möglicherweise therapiemotivierten Patienten eine Schmerztherapie vorzuenthalten. Gleichzeitig sollten aber bei jenen Patienten, die nachgewiesen ihre Symptomatik nicht-authentisch überhöhen, sinnlose Therapieversuche wie auch Frühberentungen ohne hinreichenden Anlass vermieden werden.

Deshalb sollte die vorliegende empirische Studie die Praktikabilität, Validität und Reliabilität der verfügbaren Basis- und Validitätsskalen des MMPI-2-RF sowie des BHI-2 in neu übersetzter deutscher Fassung zur Aufdeckung von Beschwerdeüberhöhungen im Abgleich zu bisher verfügbaren Instrumenten überprüfen.

1.14 Untersuchungs-Hypothesen

Als **erste Untersuchungs-Hypothese** H_0**1** wird angenommen, dass Patienten mit extern gesicherter MPRD-Klassifikation **in signifikant deutlicher Ausprägung** Auffälligkeiten in fünf MMPI-2-RF-Validity-Scales gegenüber Patienten ohne MPRD-Klassifikation aufweisen.

Die Annahme höherer Angaben der MPRD-Patienten in den Validitätsskalen des MMPI-2-RF gründet vor allem darauf, dass die meisten Menschen dem falschen Stereotyp folgen,

nach der sich die Glaubwürdigkeit von (körperlichen oder psychischen) Symptomen insbesondere am Grad ihrer Auffälligkeit bemisst (vgl. Rogers 2008 [235]).

Wenn demnach die Einteilung der Stichprobe (z.B. nach den Slick-Kriterien) hinsichtlich dem Konstrukt „Overreporting“ gültig ist (sie also eine valide Unterscheidung unterschiedlich stark ausgeprägter Antwort-Verzerrungen leistet) und die Kennwerte der Validitätsskalen des MMPI-2-RF negative Antwortverzerrungen valide abbilden, dann sollten MPRD-Patienten **insbesondere in der F-r-Skala (Skala seltener Symptome) und der RBS-Skala (Bias-Skala kognitiver Symptome) deutlich höhere Werte** aufweisen als Patienten mit nur möglicher MPRD, diese wiederum größere Effekte als Patienten mit nur externalen Aggravations-Motiven oder ohne Anreize für Overreporting.

Dieses Ergebnis kann erwartet werden, da die F-r-Skala im Vergleich zu den übrigen Validitätsskalen des MMPI-2-RF das weiteste Spektrum psychischer und körperlicher Auffälligkeiten erfasst - und damit die größte Wahrscheinlichkeit hat, Patienten mit genereller Tendenz zur Beschwerdeüberhöhung zu identifizieren. Zudem wurde die Itemhomogenität dieser Skala durch die überarbeitete Item-Selektion in der RF-Version des MMPI-2 optimiert (vgl. Ben-Porath & Tellegen 2008 [29]).

Die RBS-Skala wurde von den Testautoren im Jahr 2011 zu den revidierten Validitätsskalen des MMPI-2-RF hinzugenommen, da sie aufgrund einer Vielzahl weiterer Studien am geeignetsten erscheint, unglaubwürdige Gedächtnisdefizite zu verifizieren (z.B. Gervais et al. 2010 [101]).

Ähnliche Effekte dieser beiden Validitätsskalen des MMPI-2-RF beobachtete Aguerrevere (2010) [2] an einer Patienten-Stichprobe mit chronischen Rückenschmerzen, die mittels PVTs drei Clustern von Patienten mit unterschiedlichem Overreporting-Verhalten zugeordnet werden konnten. Anderson (2011) [9] beobachtete vergleichbare Effekte in denselben Skalen bei 169 Patienten mit chronischen Schmerzen aus einer forensischen neuropsychiatrischen Praxis nach mehrdimensionaler Vor-Klassifikation hinsichtlich Overreporting.

Ähnliche Gruppenunterschiede wurden auch in anderen Studien mit forensischen Patienten (Gervais et al. 2010 [101]) mittels Erhebungen durch MMPI-2-RF berichtet, die sogar besser differenzierten, als die vergleichbaren Skalen des MMPI-2. Bei Haftinsassen und hinsichtlich ihrer Einschränkungen durch Behinderungen klagenden Patienten beobachteten Wygant et al. (2010) [323] insbesondere in der RBS-Skala ähnlich hohe Effekte.

Als **Nebenaspekt** wäre zu prüfen, ob sich die erwarteten Gruppenunterschiede auch in der im MMPI-2-RF auf 11 Items verkürzten Henry-Heilbronner-Index (HHI-r) zeigen, wie es verschiedene andere Studien nahelegen (z.B. Henry et al. 2012 [131], Tsushima et al. 2011 [300], Jones et al. 2012 [144]). Da die HHI-Skala mit der RSB-Skala keine gemeinsamen Items teilt, würden sich - im Fall einer Hypothesen-Bestätigung - beide Skalen zur gegenseitigen Überprüfung (Kreuzvalidierung) in Gutachten anbieten.

Gruppen-Unterschiede **in mittlerer Ausprägung** zwischen Patienten mit und ohne MPRD werden in der ebenfalls neu konzipierten Fs-Skala (Overreporting somatischer Symptome) sowie in der FBS-r-Skala erwartet. Die Fs-Skala des MMPI-2-RF wurde speziell als Auswahl eher untypischer Angaben bei Patienten mit chronischen Schmerzen konstruiert, zeigte moderate Detektion-Eigenschaften jedoch bislang ausschließlich in Simulationsstudien (Sellbom et al. 2012 [273]).

Die FBS(-r)-Skala, die speziell zur Erfassung möglicher Beschwerdeüberhöhungen somatischer und psychischer Symptome bei Patienten mit Unfallschäden konzipiert wurde, zeigte nach jüngeren Metaanalysen (Nelson et al. 2010 [213]) eine gute Detektions-Qualität zur Aufdeckung von Overreporting. Ferner bewies sie ihre Detektionsstärke in einer Studie mit Probanden, die medizinische Symptome simulierten (Wygant et al. 2009 [322]).

Die FBS-r führte jedoch in anderen Studien (Sellbom et al. 2012 [273]) mit authentischen somatischen Patienten ohne Kompensations-Ansprüche und einer Gruppe von Simulanten zu einer eher **geringen** Detektionsstärke. Die FBS-r-Skala wurde aber auch wegen ihrer heterogenen faktoriellen Struktur und deshalb mehrdeutiger Ergebnisse kritisiert (Gass et al. 2012 [96]).

Die **geringsten Unterschiede** zwischen den Untersuchungsgruppen werden in der Fp-r-Skala zur Erhebung untypischer psychopathologischer Symptome erwartet. Nach Burchess & Ben-Porath (2010) [45] neigen Simulanten somatischer Symptome dazu, weniger überhöhte Angaben psychopathologischer Symptome zu machen, als Simulanten, die bewusst Psychopathologie überhöht angeben sollten.

Deshalb ist zu vermuten, dass Patienten mit chronischen Schmerzen, die Beschwerden aggravieren, dazu neigen, mehr somatische Symptome zu überhöhen als psychische Symptome, um eine psychische Stigmatisierung zu vermeiden. Dieses Resultat lässt in der vorliegenden Studie einen höheren Differenzierungsgrad der Fs-Skala bei MPRD-Patienten vermuten als

in der Fp-r-Skala. Auch detektierte die Fp-r-Skala in den meisten der bisher genannten Studien am schlechtesten unter den fünf Standard-Validitätsskalen des MMPI-2-RF (Anderson 2011 [9]. Bei Aguerrevere (2010) [2] differenzierte die F-r-Skala siebenmal stärker als die Fs- und die Fp-r-Skala zwischen den Clustergruppen.

Als **weiteres Teilergebnis** ist zu vermuten, dass sich die erwarteten Gruppenunterschiede auch in einem gewichteten **Validitätsindex** der fünf Standard-Validitätsskalen (Meyers-Validity-Index MI-r) **in deutlicher Ausprägung** widerspiegeln. Die hohe Ausprägung dieses Detektionseffektes wird erwartet, da die Autoren dieses Index bei einer Spezifität von 0,93 eine sehr hohe Sensitivität des MI-r von 0,85 bei Patienten mit chronischen Schmerzen konstatierten (Meyers et al. 2013 [201]).

Ferner wird als **Nebenaspekt** erwartet, dass Patienten mit MPRD auch in den **Restructured Clinical Scales** des MMPI-2-RF ein **generell überhöhtes Profil** zeigen und sich gegenüber den Patienten mit chronischen Schmerzen zu erwartenden Betonungen der Basisskalen RCd (Demoralisierung), RC1 (Somatisierung) und RC2 (Depression) zusätzlich Auffälligkeiten überhöhter allgemein-psychopathologischer Symptome zeigen (z.B. in den Skalen RC6 bis RC8). Dies legen ähnliche Ergebnisse einer Reihe anderer Studien nahe (Anderson 2011 [9], Aguerrevere 2010 [2], Sellbom et al. 2010 [272]).

Als **zweite Untersuchungs-Hypothese $H_0$2** wird angenommen, dass unter der Voraussetzung einer gültigen Einteilung der Stichprobe nach den Bianchini-Kriterien Patienten mit gesicherter MPRD-Klassifikation deutlichere Auffälligkeiten **in signifikanter Ausprägung** in der Gültigkeitsskala DIS (Disclosure) der BHI-2 (Battery of Health Improvement-2) gegenüber Patienten ohne MPRD-Klassifikation aufweisen.

Diese Hypothese gründet auf den Validitätsstudien der Testautoren Bruns & Disordio (2000, 2004) [42] [43], denen zufolge die Validitätsskalen des BHI-2, insbesondere die DIS-Skala, signifikant zwischen Patienten mit chronischen Schmerzen, Patienten der us-amerikanischen Normpopulation, aber auch instruierten Probanden, die ihre Angaben in negativer

Weise oder aber in positiver Weise verfälschten, unterscheiden konnte. Deshalb sollten in der Untersuchungs-Stichprobe Patienten mit MPRD (mit nachgewiesener Beschwerdeüberhöhung) deutlich höhere DIS-Werte aufweisen als Patienten mit nur möglicher MPRD oder ohne MPRD. Vergleichbare Gruppenunterschiede berichteten auch Bruns & Disorbio (2004) [43] in ihrer Eichstichprobe.

Die Annahme einer hinreichenden Trennschärfe der Disclosure-Skala unter den Validitäts- und den Basisskalen des BHi-2 begründet sich zudem damit, dass diese Skala das Konstrukt „Overreporting“ durch eine Analyse von 100 Items des BHI-2 aus sieben BHI-2-Subskalen-Domänen subsummiert: Depression, Angst, Feindseligkeit, aber auch zu Borderline-Symptomen, chronischer Fehlanpassung, Substanzmissbrauch und Durchhalteappellen.

Da in die Validitätsskala DIS ca. die vierstufigen Wertungen der Probanden aus ca. der Hälfte aller im BHI-2 integrierten Subskalen integriert wurden, ist zu vermuten, dass - wie im MMPI-2 - auch ein erheblicher Teil der übrigen 9 Validitätsskalen „Overreporting“ erfassen kann. Diese Annahme soll als Nebenfragestellung überprüft werden.

Die Subskalen des BHI-2 würde sich damit zur Detektion von „Overreporting“ im Sinne der Strategie „Undifferenzierte Symptom-Überhöhungen“ (Rogers 2008 Rogers2008:14) eignen.

Erkennbare Gruppenunterschiede **in geringerer Ausprägung** sollten sich hingegen in der zur Erfassung von sog. „Underreporting“ konzipierten DEF-Validitätsskala (Defensiveness) zeigen, da Patienten die Tendenz zu negativer Antwortverzerrung vermutlich reziprok in geringstem Ausmaß ihrer Beschwerden bagatellisieren und positiv verzerren werden. Die Regeln der externen Klassifikation der hiesigen Studie selektieren primär Patienten mit Overreporting; problembagatellisierende oder dissimulierende Patienten wurden explizit nicht erfasst. Insofern sind eher **schwache Detektions-Effekte** in der **DEF-Skala** des BHI-2 zu erwarten.

Indirekt fließt diese Überlegung paradoxer Aufdeckungs-Merkmale der DEf-Skala auch in die BHI-2-Erhebungen zu sog. Durchhalteappellen und -Verhaltensweisen (Subskala: Perseverance) ein. Wenngleich extremes Durchhalteverhalten die Chronifizierung beispielsweise chronischer Rückenschmerzen negativ verstärken kann (Hasenbring 1993 [125]), korrelieren fehlende Durchhaltekognitionen zumeist mit weniger aktivem Schmerzbewältigungsverhalten und Tendenzen zu Rückzug und Vermeidung in sozialen, beruflichen und persönlichen

Bezügen. Fehlende Durchhaltekognitionen gehen auch mit geringerer Frustrationstoleranz bei Misserfolgen einher (vgl. Porter et al. 2003, 132pp. [226]).

Es ist deshalb anzunehmen, dass Patienten mit deutlicher Tendenz zu Beschwerdeüberhöhungen reziprok im geringsten Umfang Durchhalteverhaltensweisen angeben.

Zur **dritten Untersuchungs-Hypothese $H_0$3** ist festzustellen:

Besondere Vorteile des neuen MMPI-2-RF sind vor allem in seiner verbesserten Durchführungs-Ökonomie trotz zusätzlich geringer Item-Anzahl und geringerer Überlappung seiner Subskalen zu sehen. Hinsichtlich der Validitätsskalen ist jedoch festzustellen, dass in der US-amerikanischen Version bislang nur fünf Standard-Skalen zur Beschwerdenvalidierung einbezogen wurden. Eine Reihe im alten MMPI-2 bewährter Validitätsskalen wurden nicht auf den MMPI-2-RF übertragen und hinsichtlich ihrer Anwendbarkeit und Validität überprüft.

Beispielsweise könnte der F-K-Dissimulations-Index (Gough 1950 [105]) zur Erfassung sozial-erwünschter Antwortverfälschungen besondere Qualitäten in der Aufdeckung von nicht-authentischen Schilderungen aufweisen, da in ihm zwei Detektionsansätze (Over- und Underreporting) synergistisch Anwendung finden. Eine Integration in den MMPI-2-RF wäre möglich, da auch im neuen (MMPI-2-FR-)Fragebogen die K-(Correction-)-Skala (als K-r-Skala) einbezogen wurde.

Aufgrund häufiger Auffälligkeiten bei psychisch-kranken Patienten wird auch die Detektionsgüte der sog. kritischen Item-Listen (Lachar-Wrobel, Koss-Butcher), die vor allem augenschein-valide Items psychopathologischer Symptome abfragen, bei hinreichend hoch angepassten Grenzwerten als aussagekräftig zur Aufdeckung von Overreporting bewertet (Butcher et al. 1989 [48], Green 2008, 174pp. [115]). Ihr Rational zur Erfassung von Antwortverzerrungen beruht, ähnlich wie bei den Obvious-Subtle-Scales (O-S-Skalen, Wiener 1948 [315]), darauf, dass Patienten mit Tendenz zur Überhöhung von Symptomen diese Items besonders häufig befürworten sollten.

Zum dritten können mittels der Fb(*-failed-back*)-Skala (Butcher et al. 1989 [48]) Inkonsistenzen und Widersprüche in der Antwortgebung durch Abgleich ähnlicher Abfragen zu unterschiedlichen Zeiten der Testung (Testhälften) überprüft werden. Damit ermöglicht diese Skala einen verdeckten Abgleich von Konsistenz / Diskrepanzen miteinander zusammenhän-

gender Beschwerden. In einer Studie von Bagby et al. (2000) [22] erwies sich die Fb-Skala gegenüber der F-Skala und der Dissimulating Scale (Ds) als die trennschärfste Skala, um Depressions-Symptome authentisch antwortender Patienten von Experten, die eine Depression simulierten, zu unterscheiden.

Zu prüfen ist, ob der Algorithmus eines Konsistenz-Vergleichs von Item-Paaren auch bei einer anderen Verteilung im Test gleich gute Detektions-Möglichkeiten wie im MMPI-2 bietet, da die im MMPI-2-RF verbliebenen Items sich nicht auf zwei Testhälften verteilen. Dies könnte jedoch auch ein Vorteil dieser Erhebungsmethode sein, da andere Autoren die Aufteilung der Items im MMPI-2 kritisierten (fragliche Überlagerung mit Testaufmerksamkeit, Rogers et al. 2003 [238]).

Eine weitere, im MMPI-2-RF nicht verwendeter Validitätsskalen ist die O-S-Validitäts-Methodik (Obvious-Subtle-Scales, Wiener 1948 [315]), die Konsistenzen zwischen Angaben bei augenscheinlichen Abfragen und versteckten, subtilen Symptomabfragen überprüft. Sie könnte sich insbesondere bei Patienten mit chronischen Schmerzen als trennscharf erweisen, da sich bei Probanden, die eine Somatoforme Schmerzstörung simulierten, stark erhöhte O-S-Differenzen fanden (z.B. bei Sivec et al. 1994 [277]). Dush et al. (1994) [77] war es im MMPI-2 möglich, 43 Patienten mit chronischen Schmerzen mit Rentenantrag und 45 Nicht-Rentenantragsteller mit Hilfe der T-Wert-Differenzen der O-S-Skalen signifikant korrekt zu identifizieren.

Hier ist zu prüfen, ob diese Skalen in modifizierter Form im MMPI-2-RF ebenso gute Detektions-Qualitäten aufweisen wie im MMPI-2.

Dies gilt auch für die Md-Skala (Malingered Depression Scale, Steffan et al. 2003 [282]), die als Spezial-Skalen zur Einschätzung verfälschter (überhöhter) Depressions-Angaben entwickelt wurde. Allerdings erwies sich diese Skala in der Studie von Thies (2012) [296] bei einer Spezifität ≥ 90 % und einer Sensitivität von 30,2 % als nur im mittleren Umfang trennscharf.

Auch der in der Ds-Skala (nach Gough 1954 [106]) genutzte Detektionsansatz (bei Simulanten gehäuftes Vorkommen fehlerhafter Stereotypien) erwies sich gegenüber anderen Verfahren als besonders resistent gegen Coaching. Rogers et al. (2011) [239] adaptierten den Dissimulations-Index als erste Autoren an den MMPI-2-RF und bestätigten seine gute De-

tektionsgüte für simulierte psychische Störungen. Bei Patienten mit chronischen Schmerzen wurde die Nutzbarkeit dieses Index bislang noch nicht untersucht.

Aufgrund der Vielzahl im MMPI-2-RF untersuchbaren Validitätsskalen wurde angenommen, dass eine hinreichende Anzahl unterschiedlicher, besonders trennscharfer Skalen zu finden ist, die gleichzeitig mit hoher innerer Konsistenz unterschiedliche Aspekte von Malingering und möglichst unterschiedliche Verfälschungsstrategien erfassen können.

Meyers et al. (2013) [201] demonstrierten mittels des MI-r-Index die Nützlichkeit eines gewichteten Summations-Indikators der kaum überlappenden fünf Validitätsskalen des MMPI-2-RF zur Aufdeckung von Overreporting.

Deshalb wird in **Untersuchungs-Hypothese $H_0$3** angenommen, dass ein neu konzipierter, gewichteter Summations-Index von hoch trennscharfen MMPI-2-RF-Validitätsskalen mit möglichst hoher innerer Konsistenz gegenüber den herkömmlichen Validitätsskalen bei gleich hoher Spezifität ($\geq$ 90 %) eine maximal hohe Sensitivität ($\geq$ 95 %) aufweist, welche die Akkuranzwerte der fünf Standard-Validitätsskalen des MMPI-2-RF sowie des Kombinations-Index dieser Skalen (MI-r) signifikant überschreitet.
Dieser neue Index sollte sich aus den Standard-Validitätsskalen des MMPI-2-RF sowie den an diesen Test adaptierten *älteren* Skalen zusammensetzen.

Als **vierter Untersuchungs-Aspekt der Studie** wurde die Fragestellung aufgenommen, ob die aus dem neuen MMPI-2-RF ableitbaren Validitätsskalen trotz Verkürzung des Fragebogens eine gleich gute oder sogar bessere Detektionsgüte aufweisen als ihre Äquivalenz-Skalen im alten MMPI-2.

Diese Annahme gründet zum einen darauf, dass die Validitätsskalen des MMPI-2-RF gegenüber den traditionellen MMPI-2-Validitätsskalen eine reduzierte Anzahl von Item-Überlappungen aufweisen sollen, ferner auf einer optimierten Skalenkonstruktion basieren und damit eine präzisere Selektion trennscharfer Items zur Aufdeckung von Overreporting beinhalten (Ben-Porath & Tellegen 2008 [29]).

Zum anderen wurde in neueren MMPI-2-RF-Studien zur Nützlichkeit der RC-Skalen bei Patienten psychotherapeutischer Kliniken und bei Militärveteranen (Simms et al. 2005 [275], Sellbom et al. 2006 [271], Van Der Heijden et al. 2010 [305] in einer Niederländischen Patientengruppe) bestätigt, dass die Skalensets der zwei MMPI-Versionen sehr ähnliche klinische Einordnungen ermöglichen. Alpha- und Inter-Item-Korrelations-Analysen bestätigten hingegen nicht die erhofften höheren Reliabilitäten der RC-Skalen gegenüber Skalen der Vorgängerversion MMPI-2 (Rouse et al. 2008 [251]).

In einer Known-Groups-Studie (Gervais et al. 2010 [101]) mit Einsatz kognitiver Symptom-Validierungstests zur Klassifikation des Aggravationsverhaltens der Untersuchten erwiesen sich die neuen Validitätsskalen des MMPI-2-RF (F-r, Fp-r, Fs und FBS-r) jedoch tendenziell als besser zur Overreporting-Detektion als ihre MMPI-2-Vorgänger. Gervais et al. (2010) [101] beobachteten, dass die RBS-Validitätsskala des MMPI-2-RF in einem vergleichbaren Ausmaß Overreporting in verschiedenen Gedächtnis-Prüf-Inventaren erfasste wie ihre äquivalente MMPI2-Skala. Die FBS-r-Skala erwies sich in dieser Studie sogar gegenüber ihrer Vorgängerversion als überlegen (FBS Cohen's d = 0,97 vs. FBS-r Cohen's d = 1,11).

Die Detektionsgenauigkeit der Validitätsskalen beider Fragebogen-Versionen des MMPI sollte nach Bianchini et al. (2005) [31] anhand ihrer Spezifität (Ausschluss maximal vieler falsch positiv klassifizierter Probanden ohne Overreporting) und ihrer Sensitivität (maximale Anzahl korrekt klassifizierter Probanden mit Overreporting) beurteilt werden.

Zum Vergleich der beiden Akkuranzwerte der jeweils in beiden MMPI-Versionen äquivalenten Validitätsskalen eignen sich nach DeLong et al. (1988) [63] Vergleiche ihrer jeweiligen Receiver Operating Characteristic-Kurven bzw. der Detektionsgenauigkeit der sog. AUC-Kurven (Area under the Curve). Eine solche systematische Analyse aller verfügbaren Validitätsskalen wurde bislang in keiner anderen Studie durchgeführt, wenngleich einzelne Studien AUC-Kurven einzelner Validitätsskalen des MMPI-2-RF gegenüberstellten (z.B. Sellbom et al. 2010 [272]).

Als **vierte Untersuchungs-Hypothese** H_0**4** wird deshalb festgelegt: Wenn die Validitätsskalen des MMPI-2-RF bessere oder zumindest gleich gute

Detektions-Qualitäten gegenüber ihren Äquivalenzskalen des MMPI-2 aufweisen, dann sollten die Area-under-the-Curves der jeweiligen Skalen des MMPI-2-RF gleich oder größer als die AUC-Kurven ihrer Äquivalenzskalen sein.

Zur Festlegung der **fünften Untersuchungs-Hypothese $H_0$5** führten folgende Überlegungen:

Nur wenige Studien untersuchten bislang Vorhersage-Möglichkeiten spezifischer Patientenmerkmale auf den Erfolg von Rehabilitations-Maßnahmen. So fanden z.B. Kobelt et al. (2010) [153], dass die subjektive Prognose der Erwerbstätigkeit von Rehabilitanden durch Unterschichtszugehörigkeit verschlechtert wird. Andere Autoren (vgl. Häuser 2002b [127]) fanden, dass u.a. hohe Veränderungsangst von Patienten, eine resignative Erwartungshaltung sowie Krankheitsgewinn-Aspekte einen ungünstigen Heilverlauf von Patienten mit chronischen Schmerzen prognostizieren. Kompensationsansprüche und finanzielle Vorteile zur Zeit einer stationären Aufnahme zur Behandlung war bei Patienten mit medizinisch nicht erklärbaren motorischen Symptomen mit einem schlechtem Therapie-Outcome assoziiert (Crimlisk et al. 1998, nach Merckelbach & Merten 2012 [191]).

Einzelne Studien untersuchten die Verwendbarkeit der Restructured-Clinical-Scales des MMPI-2-RF zur Prognose spezifischer Behandlungserfolge. So ließen erhöhte Werte in den RC-Skalen des MMPI-2-RF RC 4 (antisoziales Verhalten) und RC9 (hypomanisches Verhalten) bei Haftinsassen mit Abhängigkeits- und Aggressions-Auffälligkeiten nach einem psychoedukativen Interventions-Programm ein höheres Rückfallrisiko prognostizieren (Sellbom et al. 2008 [270]). Auffälligkeiten in der Depressions-Skala RC2 des MMPI-2-RF prognostizierten nach einem stationären Psychotherapieprogramm einen geringeren Erfolg (Scholte et al. 2012 [263]).

Keine Studie untersuchte bislang die Validitätsskalen des MMPI-2-RF zur Beschwerdenvalidierung hinsichtlich ihrer prädiktiven Eignung zur Therapieerfolgsprognose. Insbesondere bei Patienten mit chronischen Schmerzen mit längerer Krankheits- und Arbeitsunfähigkeitsdauer hat die Therapieprognose jedoch besondere Relevanz. Entsprechend §9 SGB VI sowie § 4 und § 8 SGB IX sind vor einer Erwerbsunfähigkeitsberentung stets die Möglichkei-

ten einer Rehabilitation auszuschöpfen („Reha vor Rente“). Ist diese Prognose jedoch durch externale Motive, durch mangelnde Therapiemotivation oder Overreporting extrem gering, sollte dies vor Aufnahme einer Behandlung bekannt sein.

Zur Vermeidung unsinniger Behandlungen und entsprechender Kosten wäre es besonders wichtig, vor Beginn einer Rehabilitations-Maßnahme deren Erfolg abschätzen zu können. Häuser (2002b) [127]) stellten hierzu aus klinischer Erfahrung fest, dass Patienten mit chronischen Schmerzen, die ihre Beschwerden aufgrund externer Motivationen überhöhen, prognostisch einen geringeren Reha-Erfolg aufweisen.

Für die hier durchgeführte Untersuchung ist daraus zu folgern:

Wenn eine Vorab-Klassifikation Patienten mit chronischen Schmerzen unter Einbezug einer multimodalen Erhebung negativer Angabe-Verzerrungen eine sichere Aggravation anzeigt, dann ist zu erwarten, dass diese Patienten mit MPRD aufgrund mangelnder Motivation, eine therapeutische Verbesserung zu erzielen, nach einer medizinisch-psychologischen Kombinationstherapie einen geringere Schmerzreduktion aufweisen, als Patienten ohne Overreporting.

Auf dieser Basis wurde für den in der vorliegenden Studie entwickelten ROI-Index (zur Aufdeckung negativer Antwortverzerrungen) als fünfte Untersuchungs-Hypothese $H_0$5 angenommen:

Untersuchungs-Hypothese $H_0$5: Wenn sich Beschwerdeüberhöhungen bei Patienten mit chronischen Schmerzen durch den zur Identifikation von Overreporting entwickelten ROI-Index (MMPI-2-RF) trennscharf identifizieren lassen, dann ist zu erwarten, dass MPRD-Patienten, die einen sicher auffälligen ROI-Score aufweisen, aufgrund mangelnder Therapie-Motivation nach einer medizinisch-psychologischen Kombinationstherapie einen geringere Schmerzreduktion aufweisen, als Patienten, die eine unauffälligen ROI-Score aufweisen.

Wenn sich diese Hypothese nicht bestätigt, dann mag das daran liegen, dass

(1) der ROI-Index möglicherweise weniger als erhofft zur Detektion von Overreporting geeignet ist, oder

(2) die VAS-Skala zur Erhebung der Schmerzintensität nicht vollständig in der Lage ist, das Konstrukt „Therapieerfolg“ abzubilden (der sich auch an den unterschiedlichsten anderen Parametern messen ließe, z.B. Rückkehr zur Arbeit, Medikamentenkonsum, verhaltensbezogenen Tests [z.B. Wegstrecke, Finger-Boden-Abstand u.ä.], affektiven Veränderungen [z.B. in Depressions- oder Lebensqualitäts-Fragebögen]) oder

(3) Overreporting keinen unmittelbaren Einfluss auf die Therapiemotivation der Patienten hat oder möglicherweise sogar mit einer höheren Therapiemotivation einhergeht (unter der Annahme, dass Overreporting auch Ausdruck eines Hilfe-Ersuchens sei).

2 Methode

2.1 Patienten-Rekrutierung und Klassifikation

520 chronische Schmerzpatienten durchliefen in einem 18-monatigen Zeitraum von Dezember 2012 bis Juli 2014 ein zweiwöchiges multidisziplinäres stationäres verhaltensmedizinisches Behandlungsprogramm, an dem der Untersucher kontinuierlich beteiligt war. Dabei handelte es sich um Patienten mit chronischen Schmerzbeschwerden jeder Art und Lokalisation, unter Einbezug von Patienten mit einer Akut-Schmerz-Symptomatik, sofern diese als Phase eines chronischen Schmerzsyndroms eingeschätzt wurden (z.B. akute Exazerbation einer Trigeminus-Neuralgie; Akutphase einer postherpetischen Neuralgie).

Vor Therapiebeginn beantwortete jeder Patient den Deutschen Schmerzfragebogen (Korb, Pfingsten 2003 [158]), ein umfassendes Erhebungsinstrument für schmerzrelevante sowie demographische Daten. Als schmerzspezifische und -assoziierte Daten werden u.a. erhoben: Schmerzlokalisation, zeitliche Charakteristika der Schmerzen, Schmerzdauer und Schmerzintensität, schmerzassoziierte Symptome, affektive und sensorische Schmerzwahrnehmungen, Faktoren der Schmerzverstärkung und -verringerung, Information über Vorbehandlungen, schmerzrelevante Beeinträchtigungen, Angststörungen, Depressive Störungen, Komorbiditäten, soziale Faktoren, Lebensqualitäts-Einschätzungen. Ergänzend ist im Deutschen Schmerzfragebogen ein Angst-Depressions-Screening mittels Hospital Anxiety and Depression Questionaire (HADS, Herrmann-Linger et al. 1995 [134]) integriert.

Am ersten Behandlungstag erfolgte ein umfassendes vor-klinisches Assessment, mit folgenden Anteilen: Sichtung aller relevanten Vortherapie-Berichte, umfassende medizinisch-körperliche Untersuchung, umfassende psychologische Exploration, einschließlich verschiedener Funktionstests (z.B. Waddell's Zeichen nach Waddell et al. 1980 [308]), Informationen einbezogener Angehöriger, Untersuchung der sprachlichen Rezeptionsleistung (Lese-Sinnverständnis und auditives Sprachverständnis der deutschen Sprache) und in Fällen auffälliger kognitiver Störungen, ergänzende Demenz-Screening-Untersuchungen (z.B. Demtect nach Calabrese et al. 2000 [150]).

Ausschlussfaktoren: Hinweise auf schwere psychiatrische Störungen (z.B sog. *Major Depression*, schwere Angststörung oder eine andere psychotische Symptomatik), schwergradige psychosomatische Störungen (z.B schwere posttraumatische Belastungsstörung, Artefakt-

oder Konversions-Symptomatik), gerontopsychiatrische Störungen, neurologische oder andere somatische Krankheiten (z.B. Schädelhirntrauma, Altersdemenz, akute Vigilanz-Einschränkungen (z.B. durch Entzugs-Symptomatik) wurden jeweils von konsiliarisch hinzugezogenen Kollegen anderer Fachdisziplinen untersucht, insbesondere der psychiatrischen oder psychosomatischen Disziplin. Patienten mit einer entsprechenden positiven Hauptdiagnose, die gegenüber der Schmerz-Erkrankung vorrangig war, wurden von der Studie ausgeschlossen, und die Behandlung wurde in einer entsprechend spezialisierten externen Krankenhaus-Einheit fortgesetzt.

In die Studie konnten 401 Patienten eingeschlossen werden, die alle Einschlusskriterien erfüllten und alle Erhebungen durchliefen.

Diese Patienten wurden **zwei deutschen ICD-10-Diagnosen** (definiert seit 2009, ICD-Version DIMDI 2015 [66]) zugeteilt: Chronische Schmerzstörung mit somatischen und psychologischen Faktoren (F45.41) oder anhaltende somatoforme Schmerzstörung (F45.40; Schmerzstörung, die ausschließlich durch psychologische Faktoren begründet ist).

Das Assessment von Overreporting basierte auf zwei Verfahren der Beschwerdenvalidierung, die in drei Domänen unabhängig von der diagnostischen Klassifikation durchgeführt werden sollten:

1. Glaubwürdigkeit funktionaler Defizite durch Erhebung von Angaben der Patienten und der Therapeuten (**behaviorale Domäne**),
2. Glaubwürdigkeit kognitiver Defizite, die durch Performance Validierungstests (PVTs) überprüft werden sollten (**kognitive Domäne**),
3. Glaubwürdigkeit der Angabe somatischer und psychischer Symptome, die anhand fragwürdiger Befunde in Symptom-Validierungstests (SVTs) identifiziert werden sollten (**psychopathologische Domäne**).

Eine gesicherte MPRD wurde nur bei auffälligen Befunden beider Tests einer Domäne für diesen Assessment-Bereich akzeptiert.

2.2 Eingangs-Assessment

Zur Erfassung **verhaltensbezogener Auffälligkeiten** wurde zum einen eine unabhängige Bewertung der Überhöhung von Symptomen durch regelmäßige Team-Besprechungen des sta-

tionären Behandlungspersonals (Ärzte, Psychotherapeuten, Pflege, Sozialarbeiter, Seelsorger, erweitert auch Physio- und Ergotherapie) verwendet, insbesondere des Untersuchungsarztes des jeweiligen Patienten. Die Bewertungen erfolgten in den drei Stufen: 0 = keine Auffälligkeit, 1 = kontrovers diskutierte oder nur angenommene, mögliche MPRD, 2 = sicher von Arzt, Psychologen und dem gesamten Team angenommene Beschwerden-Überhöhung.

Ferner wurden Funktionsdefizite durch den Patienten und den untersuchenden Psychologen im Funktions-Fragebogen Hannover (FBBH-R, Kohlmann und Raspe 1996 [154]) zur Erhebung der Übereinstimmung zwischen Experten- und Patientensicht dokumentiert.

Der FFBH-R ist ein 12-Items-umfassender, dreistufiger Fragebogen zur Erhebung alltagsnaher funktionaler Einschränkungen durch schmerzassoziierte, inflammatorische oder muskuloskelettale Erkrankungen. Der FFBH-R wurde hier als ein der in Kap. 1. erläuterten PACT-Erhebung (nach Matheson & Matheson 1989 [181]) ähnliches Verfahren gewählt, jedoch mit höherer Durchführungs-Ökonomie. Die funktionale Kapazität von Patienten wird als Prozentzahl der Itemsumme (von 0 = „trifft gar nicht zu“ bis 2 = „trifft völlig zu“) ausgedrückt. Die bekannte Interrater-Reliabilität von Kappa-Index 0,60 wurde zur Erfassung von Symptomüberhöhungen auf der Physikalischen Symptom-Ebene entsprechend der Kriterien nach Bianchini et al. (2005) [31] genutzt. Fehlende Daten zur Detektions-Genauigkeit von Overreporting erfordern eine vorsichtige Interpretation dieses Instrumentes. Entsprechend bekannter Daten wurden Kappa-Interrater-Indizes $\leq$ 0,60 (zwischen Patient und Untersucher) als Indiz für mögliches (*possible*) Overreporting funktionaler Defizite angenommen. Kappa-Werte $\leq$ 0,30 wurden als sicheres Indiz für eine Überhöhung der berichteten Einschränkungen definiert.

Beschwerdeüberhöhung kognitiver Symptome, insbesondere fragwürdige Angaben von Gedächtsnisdefiziten, wurden mittels des aus Frankreich stammenden Performance Validierungstests DMS48 (Delayed Matching to Sample, Barbeau et al. 2004 [23]) überprüft. Hierbei handelt es sich um eine non-verbale, forced-choice Wiedererkennungsaufgabe, die aus einem Zweierset verschiedenfarbiger Zeichnungen mit 48 Objekten besteht. In einer impliziten Lernphase werden die Probanden instruiert, nur die Anzahl der Stimulusfarben zu benennen („bis zu drei Farben“ oder „mehr als drei Farben“), um in einer zweiten Testphase die beschriebenen Stimuli wiederzuerkennen, die jetzt gleichzeitig mit einem Ablenkungsreiz präsentiert werden. Die Testleistung entspricht der Prozentzahl korrekt erkannter Stimuli in den 48 Zielvergleichen.

Der hohe Deckeneffekt des DMS48 ermöglicht eine Zuordnung von wahrscheinlich nicht-authentischen Leistungsdefiziten bei Wiedererkennungsleistungen unter 90 % bei ansonsten unauffälliger neurokognitiver Funktion. Patienten mit M. Alzheimer erbringen im DMS48 im Durchschnitt mindestens eine 50-prozentige Leistung, Patienten mit M. Parkinson erreichen durchschnittlich mindestens 90 % (Mondon et al. 2007 [209]).
DMS-Summenwerte $\leq$ 90% wurden deshalb als Hinweis auf möglicherweise nicht-authentische Gedächtnisdefizite gewertet. Ein sicheres Overreporting wurde bei einer Performance unterhalb der 50-prozentigen Trefferrate angenommen.

Als zweites Assessment-Instrument wurde der Rey-15-Items-Memory-Test mit einer ergänzenden kurzen Wiedererkennungsaufgabe entsprechend Boone et al. (2002 [37], s. auch Alison et al. 2000 [5]), verwendet, mit bekannter hoher Spezifität (92 %) und ebenso hoher Sensitivität (71 %) im Vergleich zum standardisierten Rey-15- Test. Dieses Vorgehen wurde entwickelt, um den Standard-Rey-15 mit höherer Spezifität (93 %) and Sensitivität (70 %) einsetzen zu können.

Testwerte von $\leq$ 11 (80-prozentige Trefferrate) bzw. $\leq$ 9 (60-prozentige Trefferrate) im Standardtest sind als Hinweise für eine wahrscheinliche bzw. sicher nicht-authentische Angabe kognitiver Leistungsdefizite im Rey-15 zu werten, wenn sie zusätzlich durch eine unrealistisch hohe falsch positive Trefferrate im Wiedererkennungstest bestätigt wurden (vgl. Sweet et al. 2008, 225pp. [290]). Kombinierte Scores (Korrektes Abrufen + [Korrekte Wiedererkennung - Anzahl falsch positiv wiedererkannter Stimuli]) von weniger als 20 (Trefferrate < 60 %) bestätigen entsprechend verschiedener Studien (Boone et al. 2002 [37], Morse et al. 2013 [212]) sicher fragwürdige Resultate. Entsprechend wurden Trefferraten von weniger als 80 % (< 24 Punkten im Rey-15-Recognitiontest) als Indiz für zumindest fragwürdige Gedächtnisperformance angesehen.

Beschwerdeüberhöhungen in der Affekt-Domäne wurden ebenfalls zweistufig untersucht: Zum einen beantworteten die Patienten das Structured Inventory of Malingered Symptomatology SIMS (Smith & Burger 1997 [281]) mit aus der Literatur angenommenen Cutoff-Werten für *mögliches Overreporting* psychopathologischer Symptome bei Gesamtscores $\geq$ 13 Items und Annahme *sicheren Overreportings* bei Testwerten von $\geq$ 17 der 75 dichotom zu beantwortenden Items.

In früheren Studien (z.B. Merckelbach & Smith 2003 [192]) wurde eine hohe Sensitivität (93 %) und eine hohe Spezifität (98 %) zur richtigen Identifizierung instruierter Simulanten gegenüber psychiatrischen Patienten und normalen Kontrollprobanden berichtet. Jedoch zeigten Reviews (Van Impelen et al. 2014 [303]) eine Spezifität des SIMS unter gängigem Standard. Diese ließ sich jedoch durch Anhebung der Cutoff-Scores ($\geq$ 17) und Kombination des SIMS mit Einsatz eines weiteren Symptom-Validierungsverfahrens verbessern.

Zum zweiten beantworteten die Patienten die Symptom-Checkliste SCL-90-R (Derogatis 1992 [64]), um ein wahrscheinliches bzw. sicheres Overreporting psychopathologischer Symptome zu erheben. Hardt et al. (2000) [123] zeigten, dass chronische Schmerzpatienten typische Überhöhungen in fast allen SCL-90-R-Subskalen gegenüber Normpatienten machen, aber substantiell niedrige Angaben im Vergleich zu psychiatrischen und schwer belasteten psychosomatischen Patienten zu verzeichnen sind. Die durchschnittlichen Subscores dieser Patienten erreichen jedoch nicht die charakteristischen Überhöhungen instruierter Simulanten (McGuire und Shores 2001 [185]).
Entsprechend dieser Studien wurden Überhöhungen der untersuchten Schmerzpatienten in 5 oder 6 der neun Subscores als Hinweise für eine mögliche Überhöhung von Beschwerden eingeschätzt. Bei überhöhten Angaben in $\geq$ 7 der neun Subscores wurde im Vergleich mit den typischen Werten psychosomatischer Patienten eine gesicherte MPRD angenommen.

2.3 Abhängige Variablen und weitere Patienten-Selektion

Alle Patienten beantworteten in den ersten Tagen der Behandlung den MMPI-2 (Minnesota Multiphasic Personality Inventory) nach Keller & Butcher (1991) [149] und eine neue deutsche Testversion des BHI-2 (Battery of Health Improvement, Dohrenbusch & Brockhaus, 2016, in prep [71]). Die MMPI-2-RF-Daten wurden aus den MMPI-2-Erhebungen ermittelt.

Ein *Ausschluss von 33 Datensätzen* erfolgte aufgrund der im Test-Manual benannten Ausschlusskriterien für die Gültigkeit von MMPI-2-RF-Protokollen: 6 Protokolle wiesen eine zu hohe Anzahl fehlender Itemantworten auf (Can-Not-Say *CNS* $\geq$ 15), bei 18 Protokollen wies die Prüfskala (*VRIN-r*) mehr als 9 nicht-gültige Inkonsistenzen inhaltsähnlicher Items auf (T-Wert $\geq$ 80), bei weiteren 9 Protokollen wies die *TRIN-r*-Skala Inkonsistenzen der Beantwortung bei $\geq$ 14 oder $\leq$ 8 Item-Paaren (T-Wert $\geq$ 80 oder $\leq$ 20) mit quasi-umgekehrtem

Tab. 3. Regeln der Klassifikation der Patienten nach externen Kriterien

Kognitive Domäne	Gewichtung	Wertung	
Rey-15-Item-Recognition-Score $\geq$ 24	0	unauffällig	NoMRPD
Rey-15-Item-Recognition-Score < 24 und $\geq$ 20	1	möglich auffällig	P-MPRD
Rey-15-Item-Recognition-Score < 20	2	sicher auffällig	MPRD
DMS-Summenscore $\geq$ 91	0	unauffällig	NoMRPD
DMS-Summenscore $\leq$ 90 und > 50	1	möglich auffällig	P-MPRD
DMS-Summenscore $\leq$ 50	2	sicher auffällig	MPRD
Psychische Domäne			
SIMS-Gesamtscore < 14	0	unauffällig	NoMRPD
SIMS-Gesamtscore $\geq$ 14 und < 17	1	möglich auffällig	P-MPRD
SIMS-Gesamtscore $\geq$ 17	2	sicher auffällig	MPRD
SCL-90-R $\leq$ 4 Subscores $\geq$ T-Wert 70	0	unauffällig	NoMRPD
SCL-90-R 5 oder 6 Subscores $\geq$ T-Wert 70	1	möglich auffällig	P-MPRD
SCL-90-R $\geq$ 7 Subscores $\geq$ T-Wert 70	2	sicher auffällig	MPRD
Behaviorale Domäne			
FFBH-R Interrater-Kappa-Index > 0,60	0	unauffällig	NoMRPD
FFBH-R Interrater-Kappa-Index $\leq$ 0,60 und > 0,30	1	möglich auffällig	P-MPRD
FFBH-R Interrater-Kappa-Index $\leq$ 0,30	2	sicher auffällig	MPRD
Team - Beurteilung:	0	unauffällig	NoMRPD
Team - Beurteilung:	1	möglich auffällig	P-MPRD
Team - Beurteilung:	2	sicher auffällig	MPRD
Gesamt-Wertung			
Non-Malingered Pain-Related Disability	0	Max. 5 x mögl. Auffälligkeiten (1)	
Possible Malingered Pain-Related Disability	1	6 x mögl. Auffälligkeiten (1) und Max. 3 x sichere Auffälligkeiten (2); Aber: keine Domäne sicher auffällig	
Definite Malingered Pain-Related Disability	2	Vier oder mehr sichere Auffälligkeiten (2), Damit mindestens eine Domäne sicher auffällig	

Inhalt auf. Aus dieser Vorauswahl gültiger MMPI-2-RF resultierten 368 gültige Patientenprotokolle, die in der Studie berücksichtigt wurden.

Studienteilnehmer: In der Stichprobe mit vollständigen und gültigen Datensätzen (n = 368) befanden sich 233 Frauen (63,3 %) und 135 Männer (36,7 %); das mittlere Alter lag bei 55 Jahren (Median; Mittelwert 55,1 Jahre, SD = 14,7).

Die meisten Patienten wiesen eine Schulbildung mit Hauptschulabschluss auf (191 Patienten, 51,9 %), 10,6 % der Patienten hatten keinen Abschluss (n = 39), 20,9 % hatten einen Realschulabschluss (n = 77), 8,2 % eine Fachhochschulreife (n = 30) und 8,4 % das Abitur (n = 31).

Hinsichtlich der Diagnosen wiesen 5,2 % (n = 19) als Hauptdiagnose Kopf- oder Gesichtsschmerzen auf, 15,8 % (n = 58) hatten neuropathische, ischämische oder viszerale Schmerzsyndrome. Unter Rückenschmerzen oder muskuloskelettalen Schmerzen litten 69,6 % der Patienten (n = 256) und 9,5 % der Patienten (n = 35) wiesen als Hauptdiagnose eine Somatoforme Schmerzstörung auf.

2.4 Klassifikation der Probanden

Eine Overreporting-Klassifikation der Studienteilnehmer erfolgte entsprechend der bei Bianchini et al. 2005 [31] beschriebenen drei Domänen von Malingering (Kognitiver Funktionsbereich, Bereich Psychischer Beeinträchtigungen, somatisch-verhaltensbezogener Ebene).

Für sicheres Overreporting (MPRD) wurde eine **konservative Entscheidungsregel** entwickelt (s. Tab. 3, S. 139). Wir nahmen an, dass gesicherte Beweise für Overreporting in mindestens einer der drei Domänen erfasst werden müssten. In den Fällen, in denen einzelne Domänen keine sichere Bestätigung durch beide Beschwerden-Validierungsverfahren aufwiesen, sollten mindestens vier der sechs Gültigkeitsinstrumente sicher fragwürdige Resultate zeigen. Wir wählten diesen **konservativen Entscheidungs-Algorithmus** in der Abweichung von den Kriterien nach Slick et al. (1999) [279], um die Voraussetzungen für Malingering und die Klassifikations-Genauigkeit zu erhöhen.

Die Probanden wurden im Anschluss einer von vier Untersuchungsgruppen zugewiesen:

1. **No Incentive, Nicht-Malingering (NoInc):** Gruppe 1 (n = 231, 62,8 %) bestand aus Patienten mit den schmerzbezogenen Diagnosen ICD-F45.41 (Chronische Schmerzstörung mit somatischen und psychologischen Faktoren), sowie ICD-10-F45.40 (Somatoforme Schmerzstörung). Diese Patienten zeigten zudem maximal in fünf der sechs externen Verfahren zur Beschwerden-Validierung leichte und damit sog. mögliche Auffälligkeiten. Ferner wiesen diese Patienten keine finanziellen Kompensations-Konflikte oder -begehren auf, die definiert wurden bei: (1.) Patienten mit einer Arbeitsunfähigkeit von mehr als 9 Monaten, mit zunehmender Gefahr einer Aussteuerung durch ihre Krankenkasse, spätestens nach 18 Monaten; (2.) Patienten mit einem Rentenantragsverfahren; (3.) Patienten mit abgelehntem Rentenantrag, die einen Widerspruch gestellt hatten; (4.) Patienten mit abgelehntem Rentenantrag und Klageverfahren vor einem Sozialgericht (auf örtlicher, Landes- oder Bundesebene); (5.) Patienten

mit bewilligter Zeitrente, die nicht bereits in eine Dauerrente umgewandelt war oder in eine Altersrente übergegangen war; (6.) Patienten mit anderweitigen Kompensations-Konflikten (Antrag auf Erhöhung einer Schwerbehinderten-Anerkennung, Antrag auf eine Berufsunfähigkeits- oder Unfallrente; Strittige Klageverfahren nach Unfällen gegen Drittbeteiligte u.ä.).

48,1 % (111) dieser Patienten wiesen keine Medikation mit einfachen oder Opioid-Analgetika auf, 24,2 % (56 Patienten) hatten eine regelhafte Verordnung von Opioid-Analgetika und 27,7 % (64 Patienten) nahmen Benzodiazepin-Präparate oder unretardierte Opioid-Analgetika ein. Trotz dieses potentiellen Abhängigkeits-Problems (mit möglichem Aspekt eines externalen Motivs für Overreporting) wurden diese Patienten nicht als „INC"-Patienten oder mit „MPRD" klassifiziert, da sie keine sicheren Auffälligkeiten in den Beschwerdenvalidierungsverfahren aufwiesen.

2. **Incentive-only, Nicht-Malingering mit externem Aggravationsmotiv (IncOnly)**: Die Schmerzpatienten dieser Gruppe (n = 78, 21,2 %) kennzeichneten neben ihren Diagnosen ICD-F45.41 (Chronische Schmerzstörung mit somatischen und psychologischen Faktoren) sowie F45.40 (Somatoforme Schmerzstörung) nachgewiesene zusätzliche äußere Motivationen für eine Beschwerdeüberhöhung (s.o.). Jedoch zeigten auch diese Patienten keine gehäuften Auffälligkeiten in den Verfahren zur Beschwerdenvalidierung (leichte Auffälligkeiten in maximal fünf der externalen Tests, jedoch in keiner der drei Domänen sichere oder auch nur leichte Hinweise auf eine MPRD).

   42,3 % (33) dieser Patienten hatten keine Medikation mit einfachen oder Opioid-Analgetika, 32,1 % (25 Patienten) hatten eine regelhafte Verordnung von Opioid-Analgetika und 25,6 % (20 Patienten) nahmen Benzodiazepin-Präparate oder unretardierte Opioid-Analgetika ein. Auch diese Patienten wurden trotz Abhängigkeit nicht den Gruppen „P-MPRD" oder „MPRD" zugeordnet, da sie keine sicheren Auffälligkeiten in den Beschwerdenvalidierungsverfahren aufwiesen.

3. **Patienten mit möglicher Malingered Pain Related Disability (P-MPRD)**: Diese Patientengruppe (n = 34; 9,2 %) mit chronischen Schmerzstörungen und somatoformen Schmerzsyndromen, wiesen in allen sechs externalen Beschwerdenvalidierungsverfahren zumindest leichte Auffälligkeiten auf. Sie zeigten teilweise sogar in bis zu drei der Verfahren erhöhte Abweichungen (z.B. SIMS > 17, Rey15 = 9, FFBH-R Kappa = 20). Da jedoch in keiner Domäne beide Tests eine sichere Auffälligkeit bestätigten, wurde

das Antwortverhalten dieser Probanden als möglicherweise Overreporting klassifiziert. Diese Gruppe diente der indirekten Überprüfung der Klassifikationsgüte.

52,9 % (n = 18) dieser Patienten hatten keine Medikation mit einfachen oder Opioid-Analgetika, 23,5 % (8 Patienten) hatten eine regelhafte Verordnung von Opioid-Analgetika und ebenso 23,5 % (8 Patienten) nahmen Benzodiazepin-Präparate oder unretardierte Opioid-Analgetika ein. Auch diese Patienten wurden trotz Abhängigkeit nicht als „MPRD“ klassifiziert, da sie keine sicheren Auffälligkeiten in den BV aufwiesen.

4. **Patienten mit definitiver MPRD**: Diese Patienten (n = 25, 6,8 %) wiesen sichere Auffälligkeiten in mindestens vier der Verfahren zur Beschwerdenvalidierung auf, mit fragwürdigen Befunden in mindestens einer Domäne, entsprechend den Kriterien nach Bianchini et al. (2005) [31]. 21 Patienten wurden mit der ICD-10-Diagnose F45.41 (Chronische Schmerzstörung mit somatischen und psychologischen Faktoren) klassifiziert; vier Patienten wiesen eine anhaltende somatoforme Schmerzstörung (ICD-10-F45.40) auf.

   44,0 % (11) dieser Patienten hatten keine Medikation mit einfachen oder Opioid-Analgetika, 40,0 % (10 Patienten) hatten eine regelhafte Verordnung von Opioid-Analgetika und 16,0 % (4 Patienten) nahmen Benzodiazepin-Präparate oder unretardierte Opioid-Analgetika ein.

2.5 Charakteristika der Klassifikationsgruppen

Die Untersuchungsgruppen unterschieden sich nicht signifikant hinsichtlich ihres Alters ($F_{364,3}$ = 0,78, p = 0,507, vgl. Tab. 4, S. 143), das im Durchschnitt bei 55,1 (SD = 14,7) Jahren lag. Auch bezüglich der Schulbildung der Teilnehmer der vier Studiengruppen fanden sich keine relevanten Unterschiede ($F_{364,3}$ = 0,34, p = 0,796). Die Studienteilnehmer hatten durchschnittlich 10 Jahre die Schule besucht (SD = 2,7). In allen vier Untersuchungsgruppen befanden sich ebenso gleich viele Frauen und Männer (χ^2 = 0,99, p = 0,803). In der gesamten Stichprobe befanden sich durchschnittlich 63,3 % Frauen und 36,7 % Männer.

In einer kleineren Teilstichprobe (n = 275), die zur Vorab-Publikation einiger Fragestellungen verwendet wurde, fanden sich Unterschiede hinsichtlich des Alters (Patienten ohne Auffälligkeiten waren die signifikant jüngeren), auch befanden sich mehr Männer in der

MPRD-Gruppe als in den anderen Studiengruppen. Diese Unterschiede wurden durch Erhöhung der Stichprobengröße aufgehoben.

Tab. 4. Merkmale der vier Klassifikationsgruppen für Overreporting

N	231	78	34	25			
	No Incentive	Incentive Only	Possible MPRD	Definite MPRD	F	P	SN
Alter	55,9 (14,5)	53,3 (15,4)	53,6 (14,2)	56,3 (15,2)	0,78	0,507	n.s.
Schulbildung (in Jahren)	10,3 (2,6)	10,2 (2,7)	10,3 (3,1)	9,7 (2,0)	0,34	0,796	n.s.
Geschlecht					χ^2		
Anteil Frauen (%)	61,5	66,7	67,6	64,0	0,99	0,803	n.s.
Diagnosen (%)							
Kopf- und Gesichts-Schmerzen	5,2	5,1	5,9	4,0	13,19	0,154	n.s.
Rückenschmerzen / Muskuläre Syndrome	66,7	68,0	91,2	72,0			
Somatoforme Schmerzstörungen	10,8	7,7	0,0	16,0			
Andere Schmerzsyndrome	17,3	19,2	2,9	8,0			
Analgetika (%)							
Non-Opioide	48,1	42,3	52,9	44,0	0,090	0,966	n.s.
Retardierte Opioide	24,2	32,1	23,5	40,0			
Unretardierte Opioide / Benzodiazepine	27,7	25,6	23,6	16,0			
					F	P	SN
Depression (HADS-D)	10,3 (4,8)	10,1 (4,9)	10,6 (4,4)	15,4 (3,1)	9,24	0,0001	***
Angst (HADS-A)	9,0 (5,3)	9,3 (4,8)	11,2 (4,3)	13,1 (5,3)	6,18	0,0004	***

Die Untersuchungsgruppen unterschieden sich ebenfalls nicht hinsichtlich der Haupt-Schmerzdiagnosen (χ^2 = 13,2, p = 0,154). Die meisten Patienten (durchschnittlich 69,6 %) litten vorrangig an Rückenschmerzen oder muskuloskelettalen Schmerzsyndromen, 16 % der Patienten der MPRD-Gruppe wiesen eine anhaltende somatoforme Schmerzstörung (ICD-10 F45.40) auf und machten damit tendenziell den höchsten Anteil dieser Patienten gegenüber den anderen drei Studiengruppen aus. Viszerale, ischämische und neuropathische Schmerzsyndrome wurden tendenziell am häufigsten in den Studiengruppen ohne Hinweise auf Beschwerdeaggravation vordiagnostiziert.

Hinsichtlich ihrer Medikamenten-Einnahmen unterschieden sich die Studiengruppen ebenfalls nicht signifikant (χ^2 = 0,09, p = 0,966). Die Patienten der Gruppen (NoInc und Possible MPRD) wiesen tendenziell den höchsten Anteil an Verordnungen von Nicht-Opioid-Analgetika auf. In der Patientengruppe ohne Kompensationsmotive fand sich gleichzeitig tendenziell die höchste Anzahl an Einnahmen unretardierter Opioide und Benzodiazepine (in der Vorstudie waren letztere Abhängigkeiten am häufigsten bei den somatoformen Pati-

enten). Tendenziell wurden retardierte Opioide hingegen am häufigsten den Patienten mit MPRD verordnet, was auch den Erhebungen der Vorstudie entsprach.

Die Untersuchungsgruppen unterschieden sich ausschließlich signifikant hinsichtlich der in der Hospital Anxiety and Depression Scale (HADS) angegebenen Depressions- ($F_{355,3}$ = 9,24, p = 0,000) und Angstwerte ($F_{354,3}$ = 6,18, p = 0,0004). Dabei wiesen die MPRD-klassifizierten Probanden die höchsten Depressionswerte auf (MW = 15,4; SD = 4,4) ebenso wie die höchsten Angstwerte (MW = 13,1; SD = 5,3).

Post-Hoc-Analysen mittels Scheffé-Tests zeigten, dass die MPRD-Patienten signifikant höhere HADS-Depressionsscores gegenüber allen drei anderen Gruppen ohne sichere Auffälligkeiten in den BV-Verfahren angaben (MPRD vs. P-MPRD, Diff = 4,77, p = 0,002; MPRD vs. INC, Diff = 5,31, p = 0,000 und MPRD vs. NoINC, Diff = 5,03, p = 0,000).

Post-Hoc-Scheffé-Tests zeigten ferner signifikant höhere HADS-Angstscores in der Patientengruppe mit MPRD gegenüber den beiden Gruppen ohne Auffälligkeiten in den BV-Verfahren (MPRD vs. INC, Diff = 3,87, p = 0,014 und MPRD vs. NoINC, Diff = 4,12, p = 0,002). Patienten mit definitiver und mit möglicher MPRD unterschieden sich nicht hinsichtlich der Angstwerte (MPRD vs. P-MPRD, Diff = 1,94, p = 0,554). Die Patienten mit und ohne externale Kompensationsmotive unterschieden sich hingegen nicht hinsichtlich der HADS-Depressions- und Angstwerte.

Die Bedeutung dieser Auffälligkeit zwischen den Gruppen mit und ohne fragwürdige und auffällige Befunde in den Beschwerden-Validierungsverfahren wird in der Ergebnis-Diskussion ausführlich besprochen und interpretiert.

2.6 Datenanalyse

Eine Analyse der Fragestellungen erfolgte im Anschluss die Auswertung der Testverfahren. Für jeden Patienten wurden die Rohscores und amerikanischen T-Wertnormen für alle Basisskalen des MMPI-2 und alle Basisskalen des MMPI-2-RF ermittelt.

- Basisskalen (MMPI-2): 1 Hd - Hypochondrie-Skala, 2 D - Depressions-Skala, 3 Hy - Hysterie-Skala / Konversion, 4 Pp - Psycho-Soziopathie-Skala, 6 Pa - Paranoia-Skala, 7 Pt - Psychasthenie-Skala, 8 Sc - Schizophrenie-Skala, 9 Ma - Hypomanie-Skala, 10 Si - Soziale Introversions-Skala;

- Basisskalen (MMPI-2-RF): Demoralisation (RCd), Somatic Complaints (RC1), Low Positive Emotions (RC2), Cynicism (RC3), Antisocial Behavior (RC4), Ideas of Persecution (RC6), Dysfunctional Negative Emotions (RC7), Aberrant Experiences (RC8), Hypomanic Activation (RC9).

Im Anschluss wurden für jeden Patienten Rohscores und T-Werte für alle Validitäts-Skalen, ebenfalls getrennt für die Auswertung entsprechend MMPI-2 und MMPI-2-RF ermittelt.

- Validitätsskalen (MMPI-2): Skalen Lie L, Korrektur K, interne Validitätsskalen VRIN und TRIN, Seltenheitsskala F, Infrequent Psychopathology Fp, Fake Bad Scale (FBS), Response Bias Scale RBS, ES-Scales nach Keller & Butcher (1991) [149]), Validitätsindex F-K nach Gough (1950) [105], Obvoius-Subtle-Scales (O-S) nach Wiener (1948) [315], Dissimulation Scale Ds nach Gough (1954) [106], Dissimulation Scale Ds-r nach Gough 1957 [107], Malingered Depression Scale nach Steffan et al. 2003 [282], Kritische Items nach Lachar-Wrobel 1979 [162], und nach Koss-Butcher 1973 [159], Validitätsindex MIV nach Meyers et al. (2002) [202], Henry-Heilbronner-Index (HHI) nach Henry et al. (2006) [132];
- Validitätsskalen (MMPI-2-RF): VRIN-r, TRIN-r, Infrequent Responses F-r, Infrequent Psychopathology Fp-r, Infrequent Somatic F-s, Symptom-Validity FBS-r, Response Bias Scale RBS sowie die revidierten L-r- und K-r-Skalen.

Die zusätzlich im MMPI-2 auswertbaren 15 Inhalts- und die 15 Zusatzskalen sowie die 27 Inhaltskomponentenskalen wurden nicht in die Auswertung einbezogen, da sie nur bedingt für die Fragestellungen relevant sind. Ebenso wurde mit den für den MMPI-2-RF ermittelbaren Zusatz-Skalen (High-Order-Scales mit Emotional Internalizing und Externalizing, Somatic und Cognitive Scales, Internalizing und Externalizing Scales, Interpersonal Scales, Interest Scales sowie Personality Psychopathology Five PSY-5) verfahren.

Zusätzlich wurden für jedes MMPI-2-Protokoll adaptierte und verkürzte Summen-Rohscores für den MMPI-2-RF von folgenden zwölf weiteren Validitäts-Skalen und -Indizes ermittelt (Kodierungen siehe Anhang Kap. I, S. 366):

- Publizierte Validitätsskalen (MMPI-2-RF): Meyers Validitätsindex MI-r (Meyers et al. 2013 [201]) für den MMPI-2-RF, Henry-Heilbronner-Index (HHI-r) nach Henry et al. 2012 [131].

- Adaptierte Validitätsskalen (MMPI-2-RF): Fb-Infrequency-Back nach Butcher et al. 1989 [48]; Kritische Items nach Lachar-Wrobel 1979 [162] wie auch nach Koss-Butcher 1973 [159]; Wiener & Harmon's Obvious-Subtle-Scales 1948 [315]; den F-K-Index nach Gough 1950 [105]; Dissimulation Scale (Ds-rf), adaptiert nach Gough (1954) [106]; Dissimulation Scale Ds-r-r adaptiert nach Gough 1957 [107]; die adaptierte Malingered Depression Scale nach Steffan et al. 2003 [282]).

Für jedes BHI-2-Protokoll wurden die Rohwerte für die drei publizierten Validitätsskalen und 15 Basisskalen sowie eine Skala Kritischer Items berechnet sowie die für zwei Validitätsskalen und alle Basisskalen vorliegenden amerikanischen T-Wertnormen. Folgende Skalen wurden ausgewertet:

- Publizierte Validitätsskalen (BHI-2): Validitätsindex, Self Disclosure (DIS) und Defensiveness Skala (DEF),
- Basisskalen (BHI-2): Somatic Complaints (SOM), Pain Complaints (PAIN), Functional Complaints (FNC), Muscular Bracing (MB), Depression (DEP), Anxiety (ANX), Hostility (HOS), Borderline (BOR), Symptom Dependency (SYM), Chronic Maladjustment (CHR), Substance Abuse (SUB), Perseverance (PER), Family Dysfunction (FAM), Survivor of Violence (SRV), Doctor Dissatisfaction (DOC), Job Dissatisfaction (JOB) und Critical Items (BHICI).

An jeden BHI-2-Fragebogen wurden zusätzlich acht Fragen (Items 218 bis 225) angefügt, die den Aspekt der Bewusstseinsnähe näher untersuchen sollten.

Die Fragen zur Bewusstseinsnähe betrafen:

- Offenheit gegenüber der Untersuchung (Items 218 und 219),
- Interesse hinsichtlich der Befragungs-Ergebnisse (Items 220 und 222),
- Zweifel an verantwortungsvollem Umgang mit den Angaben (Item 221),
- Sorgfalt bei der Fragebogen-Beantwortung (Item 223),
- Generelle Ablehnung gegenüber der gesamten Befragung (Item 224)
- Ängste hinsichtlich persönlich negativer Auswirkungen der Befragung (Item 225)

Alle personenbezogenen Daten wurden ausschließlich anonymisiert weiterverarbeitet.

Die nachfolgenden statistischen Datenanalysen erfolgten computerisiert mittels der open-scource-Statistik Software R oder SPSS-19.

2.6.1 Demographische und Varianzanalytische Auswertungen

Die Auswertung der kategorialen, nominalen Daten, insbesondere zur Demographie der Patienten, erfolgte mittels Chiquadrat-Tests.

Zur Beantwortung der Fragestellungen $H_0 1$ und $H_0 2$ wurden Multivariate Varianzanalysen MANOVA's sowie Univariate Varianzanalysen ANOVA's mit den Faktoren Patienten-Subgruppe und den abhängigen Variablen „Validitätsskalen" durchgeführt. Entsprechend der Fragestellungen wurden jeweils mehrere Analysen mit den vorab klassifizierten Subgruppen von Patienten durchgeführt. Dies erfolgte nach einem sog. *Differential Prevalence Design* (Thies 2012 [296]).

Die Anwendungsbedingungen der Varianzanalysen (Normalverteilung der Stichprobendaten, Varianzhomogenität des *within*-Faktors) wurden jeweils vorab mit dem *Kolmogorov-Smirnov*-Test und dem *Bartlett-Box-F*-Test überprüft. Da sich Varianzanalysen als relativ robust gegenüber diesen Voraussetzungen erweisen, wurden die Analysen auch bei Abweichungen von diesen Bedingungen zur Auswertung herangezogen.

Die Prüfung der Einzeleffekte der abhängigen Variablen wurden bei signifikanten Ergebnissen der Varianzanalysen mittels Scheffé-Tests geprüft, aber zur besseren Darstellung mittels ANOVAs endgültig durchgeführt.

2.6.2 Bonferroni-Adjustierungen

Die Korrektur nach Bonferroni wird generell angewandt, um bei multiplen paarweisen Vergleichen ein Kumulieren des Alphafehlers (sog. α-Fehler-Inflation) zu minimieren (s. Bortz 1993 [38]).

Dieses eher sehr konservative Korrektur-Verfahren wurde bei Testung einiger Hypothesen dieser Untersuchung angewandt, da beispielsweise in Post-Hoc-Analysen, aber auch bei der Untersuchung multipler Validitätsskalen eines Fragebogens, durch diese Vielzahl Schein-Signifikanzen übersehen würden.

Bei jenen Analysen, bei denen solche Adjustierungen der Fehler-Signifikanz sinnvoll waren und angewandt wurden, sind je gesondert gekennzeichnet.

2.6.3 ROC- und AUC-Analysen

Die Beantwortung der Fragestellungen $H_0 4$ (Spezifität und Sensitivität der MMPI-2-RF-Validitätsskalen zur Identifizierung von sicherem Overreporting, mutmaßlichem Overreporting und unauffälligen Angaben) wurden mit Univariaten Varianzanalysen ANOVAs mit den Faktoren Overreporting-Subgruppe und den abhängigen Variablen „Validitätsskalen“ geklärt. Zusätzliche Berechnungen zur Spezifität und Sensitivität der Subskalen erfolgten mittels der ”CohensD”-function im R-package ”lsr” sowie mittels der ”prediction”- und der ”ROC”-Funktionen in den R-packages ”ROCR” und ”Epi”.

2.6.4 Analysen zur Effektstärke

Analysen der Effektstärken erfolgten mittels des auf unterschiedliche Gruppengrößen angepassten Effektmaßes *d* nach Cohen (1988) [57]. Dabei wird die Prüfgröße

$$d = \frac{\overline{X}_1 - \overline{X}_2}{SD_{Gewichtet}}$$

verwendet, wobei in die Formel die unterschiedlichen Gruppengrößen (n) in die Berechnung wie folgt integriert sind:

$$SD_{Gewichtet} = \sqrt{\frac{(n_1 - 1) * {SD_1}^2 + (n_2 - 1) * {SD_2}^2}{n_1 + n_2 - 2}}$$

Für jede untersuchte Validitätsskala wurden die Kennwerte der Sensitivität und Spezifität sowie die Positive und Negative Prädiktive Power ihrer Diskriminanzgüte ermittelt, wie in Kap. 1.7 (s. S. 29) beschrieben.

2.6.5 Analysen unter Bezug auf die Basisrate

Die Berechnung des positiven prädiktiven Werts ist nach dem Theorem nach Bayes (zitiert nach Glaros & Kline 1988 [104]) wie folgt berechenbar:

$$P(D+/T+) = \frac{p(D+) * p(T+/D+)}{[p(D+) * p(T+/D+)] + [p(D-) * p(T+/D-)]}$$

wobei *p(D+)* die Basisrate des interessierenden Merkmals (Overreporting) ist, *p(D-)* = 1 - Basisrate, *p(T+/D+)* die Sensitivität und *p(T+/D-)* die 1- Spezifität.

Umgekehrt wird die Negative Prädiktive Power als

$$P(D-/T-) = \frac{p(D-) * p(T-/D-)}{[p(D-) * p(T-/D-)] + [p(D+) * p(T-/D+)]}$$

berechnet, mit *p(T-/D-)* die Spezifität und *p(T-/D+)* die 1- Spezifität.

Nach dieser Formel kann auch die Basisrate des Merkmals in die Berechnung der PPP und NPP einbezogen und die PPP/NPP auch bei höherer oder geringerer Basisrate berechnet werden.

Berechnungen der Receiver Operating Characteristic-Kurven (ROC, Swets 1973 [291]) veranschaulichen durch eine Gegenüberstellung der Sensitivität und der Rate falsch-positiv klassifizierter Probanden (1 - Spezifität) die binäre Zuordnungs-Genauigkeit eines Merkmals, die zusätzlich durch die sog. Area-Under-the-Curve (AUC-Statistik) beurteilt werden kann. Ursprünglich zur Detektion von Radarsignalen entwickelt, wurde diese Methode frühzeitig auch bei psychologischen Fragestellungen und medizinischen (radiologischen) Verfahren angewandt. AUC-Werte über der Zufallswahrscheinlichkeit zeigen eine zumindest hinreichende Diskriminanzstärke der untersuchten Validitätsskala an, die optimalerweise bei AUC = 1,0 läge.

2.6.6 Algorithmus des Revised Overreporting Index

Auf der Suche nach einem neuen Index für den MMPI-2-RF wurde ein spezifischer Algorithmus verwendet. Der Index wurde aus verschiedenen MMPI-2-RF-Validitätsskalen zusammengefügt. Dazu wurden zunächst bisher nicht im MMPI-2-RF umgesetzte MMPI-2-Validitätsskalen für den MMPI-2-RF ermittelt (O-S-r, F-K-r, LW-r, FB-r, Ds-rf, KB-r, Ds-r-r). Dann wurde für alle MMPI-2-RF-Validitätsskalen jeweils die innere Konsistenz entsprechend einer α-Analyse nach Cronbach durchgeführt (Cronbach 1951 [59], s. Bortz 1993, 517pp. [38]), um jene Skalen mit einer exzellenten Zuverlässigkeit zu ermitteln.

Als Bewertungs-Maßstäbe der Reliabilitätsgüte gelten die in Tab. 5 (S. 150) dargestellten Werte.

Tab. 5. Bewertung von Reliabilitätsmaßen

Reliabilitätskoeffizient		Bewertung
$\alpha > 0{,}9$	=	„exzellent“
$\alpha > 0{,}8$	=	„gut“
$\alpha > 0{,}7$	=	„akzeptabel“
$\alpha > 0{,}6$	=	„fragwürdig“
$\alpha > 0{,}5$	=	„schlecht“
$\alpha \leq 0{,}5$	=	„inakzeptabel“

Während für die standardisierten Validitätsskalen des MMPI-2-RF aus der Normierungs-Studie T-Wert-Verteilungen vorliegen, existieren diese für die in dieser Studie experimentell geprüften Skalen des MMPI-2 nicht.

Um diese Skalenwerte in einem Index zu gewichten, mussten sie zunächst standardisiert werden. Die Verteilung der Skalenwerte konnte zwar bei Annahme einer normalverteilten Grundgesamtheit als normalverteilt vermutet, jedoch nicht vorausgesetzt werden. Deshalb mussten mögliche Messungenauigkeiten durch eine Flächentransformation korrigiert werden.

Zu diesem Zweck wurde eine Standardisierung nach McCall (1939) der Messwerte jeder adaptierten Validitätsskala für die gesamte Stichprobe durchgeführt (McCall 1939 [183], s. Lienert 1998 [171], Diehl & Kohr 1989 [67]).

Hierzu werden zunächst die Skalenrohwerte in Prozentränge umgerechnet. Die Prozentränge entsprechen dabei dem prozentualen Flächenanteil in der Standardnormalverteilung nach der Formel: $(Cf - f/2) * 100/N$. Die jeweilige kumulative Häufigkeitsverteilung *Cf* der Testrohpunkte schließt dabei das ganze Rohwertintervall ein, so dass ein Grenzprozentrang einem Prozentrang am Ende des Intervalls entspricht. Dabei erfolgt eine Korrektur der kumulativen Frequenzen zur Intervallmitte, indem jeweils die Hälfte der Rohwertfrequenzen *f/2* von den summierten Häufigkeiten abgezogen wird. Die so ermittelten Prozentränge werden dann anhand der Standardnormalverteilung in z-Werte transformiert. Die z-Werte können wiederum linear in entsprechende Stanine- und T-Werte transformiert werden.

Für die trennschärfsten experimentellen Validitätsskalen wurde in einem zweiten Schritt eine von 0 bis 2 gestaffelte Umrechnung in einen neuen Gesamt-Validitätsindex vorgenommen (sog. Score *ROI, Revised Overreporting Index* für den MMPI-2-RF). Hierzu wurde nach der bei Meyers et al. (2002) [202] und bei Meyers et al. (2013) [201] beschriebenen Methode die höchsten Ausprägungen des Subscores (Staninewerte 8 und 9; bei Meyers et al.

entsprechend zumeist Rohwerten $\geq$ T75 und T90) als Grenzwerte auffälliger Beschwerdeüberhöhungen festgelegt und der Revised Overreporting Index als Summenscore (zwischen 0 und 2) für jeden Patienten ermittelt.

Zur Beantwortung der Fragestellung $H_0 3$ (gewichteter Validitätsgesamtindex *ROI* für den MMPI-2-RF) wurden mittels Univariater Varianzanalyse ANOVA mit den Faktoren Malingering-Subgruppe 1 bis 3 und der abhängigen Variablen „Validitätsindex“ untersucht.

2.6.7 Vergleich der ROC- und AUC-Analysen

Zur Klärung der Fragestellung $H_0 4$ (Diskriminanz-Qualität der neuen Validitätsskalen des MMPI-2-RF gegenüber den Skalen des MMPI-2) wurden Vergleiche der Spezifitäts- und Sensitivitätswerte der jeweils äquivalenten Subskalen beider Fragebogen-Versionen, abgebildet in Receiver Operating Curves (ROC-Kurven), durchgeführt.

Diese Berechnungen wurden mittels der ”roc.test”-function im R-package ”pRoc” (Robin et al. 2011 [232]) und auch alternativ mit der ”comproc”-function im R-package ”pcvSuite” realisiert.

Generell folgen die Flächen unter ROC-Kurven (Area under the Curve) derselben Statistik wie nicht-parametrische Rang-Tests (Wilcoxon-Statistik).

Hanley & McNeil (1983) [122] (s.a. DeLong et al. 1988 [63]) erläutern ihr Vergleichs-Verfahren zweier ROC-Kurven folgendermaßen:
Angenommen eine Gruppe C_1 besteht aus *n (= N - m)* Probanden, die ein Ziel-Merkmal (z.B. *Overreporting*) zeigen, das in einer zweiten Gruppe C_2 nicht vorhanden ist, dann kann eine empirische ROC-Kurve erstellt werden, die die diagnostische Akkuranz dieses Testverfahrens erfasst.

Für eine gegebene Ausprägung des Merkmals gilt:

$$(1.) \qquad sens(z) = \frac{1}{m}\sum_{i=1}^{m} I(X_i \geq z),$$

wobei I(A) = 1, wenn A wahr ist und 0, wenn A falsch ist. Ebenso sei die Spezifität des Merkmals beschrieben mit:

$$(2.) \qquad spec(z) = \frac{1}{n}\sum_{j=1}^{m} I(Y_j < z).$$

In diesem Fall ist sens(z) die empirische Sensitivität des Tests, der die Variable in positive und negative Ergebnisse entsprechend eines Cutoff-Wertes z einteilt und spec(z) ist die entsprechende Spezifität. Wenn z alle möglichen Werte der Variablen annimmt, kann eine ROC-Kurve aus sens(z) versus $[1 - spec(z)]$ erstellt werden. Die ROC-Kurve bezeichnet eine Differenzierung, wenn sie über der 45°-Linie zwischen den Endpunkten der ROC-Grafik (0,0) und (1,1) liegt.

Wie oben erläutert, entspricht die empirische ROC-Kurve einer Wilcoxon-Verteilung bzw. einer U-Verteilung.

Der Wilcoxon-Rangsummentest baut zwar auf den Rängen einer gepoolten Stichprobe auf und nicht auf Paarvergleichen (wie der Mann-Whitney-U-Test). Er ist aber äquivalent zur Mann-Whitney-U-Statistik und kann - wie in der folgenden Gleichung (3.) gezeigt - aus ihr berechnet werden (und *vice versa*).

$$(3.) \qquad W = U + \frac{n_K(n_K+1)}{2}.$$

Die Signifikanz einer AUC gegenüber der Diagonalen einer ROC-Kurve ist deshalb auch mit einem zweiseitigen Mann-Whitney-U-Test überprüfbar. Somit ist auch die Area-Under-the-Curve aus dieser Statistik durch die folgende Formel schätzbar:

$$(4.) \qquad AUC = \frac{U}{(N1 * N2)}.$$

mit U als Wert der Teststatistik, N1 und N2 entsprechend der Probandenzahl.

Die U-Statistik schätzt die Wahrscheinlichkeit θ, dass ein zufällig aus der Gruppe C_2 ausgewählter Wert kleiner oder gleich einem zufällig aus der Gruppe C_1 ausgewählten Wert ist. Sie kann als Durchschnitt über dem Wert Ψ liegen, entsprechend

$$(5.) \qquad U = \hat{\theta} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \Psi(X_i, Y_j),$$

wobei

$$(6.) \qquad \Psi(X,Y) = \begin{cases} 1 & (Y < X) \\ 0,5 & (Y = X) \\ 0 & (Y > X) \end{cases}$$

Für die Teststatistik werden alle möglichen Paarvergleiche zwischen Probanden mit und ohne das Merkmal (z.B. *Overreporting*) untersucht, und es wird derjenige Wert erfasst, in denen auffällige Probanden einen höheren Wert haben als Probanden ohne Auffälligkeit; im Fall von Bindungen wird der Wert 0,5 festgehalten. Somit kann die Akkuranz einer Validitätsmessung mittels einer ROC-Kurve in Bezug zu ihrer Diagonale bestimmt werden.

Vergleiche von zwei oder mehr ROC-Kurven sind jedoch **komplexer** und schwieriger. Zunächst muss entschieden werden, ob zwei ROC-Kurven „gepaart“ (oder „korreliert“) sind, also aus multiplen Messungen einer Stichprobe stammen, oder ob es sich um sog. „ungepaarte“ ROC-Kurven aus verbundenen vs. nicht-verbundenen Stichproben handelt. In der hiesigen Untersuchung wurden ausschließlich gepaarte ROC-Kurven auf Unterschiedlichkeit überprüft.

Bei sog. „ungepaarten“ Daten aus zwei unterschiedlichen Gruppen von Patienten und Kontrollpersonen können Vergleiche zwischen zwei ROC-Kurven aus den AUC-Werten (Area Under the Curve) und den beiden Standard-Fehlern (*SE*) der AUC-Kurven nach folgender Formel berechnet werden:

$$(7.) \qquad z = \frac{|\, Area_1 - Area_2 \,|}{\sqrt{SE^2_{Area1} + SE^2_{Area2}}}$$

oder auch gekürzt als:

$$(8.) \qquad z = \frac{|\, Area_1 - Area_2 \,|}{SE(Area_1) - SE(Area_2)}$$

Dabei wird nach Hanley & McNeil (1982) [122] der Standard-Fehler (*SE*) einer AUC-Kurve θ nach folgender Formel bestimmt:

$$(9.) \qquad SE(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta}) + (n_A - 1)(Q_1 - \hat{\theta}^2) + (n_N - 1)(Q_2 - \hat{\theta}^2)}{n_A * n_N}}$$

Wenn $\theta \geq 0,5$ ist, ist die AUC nach Hanley & McNeil (1982) [122] nicht mehr länger nonparametrisch verteilt; der Standardfehler der AUC SE(W) hängt von zwei verteilungsspezifischen Kennwerten Q_1 und Q_2 ab, die wie folgt definiert sind:

Q_1 ist die Wahrscheinlichkeit, dass zwei zufällig ausgewählte Probanden mit Auffälligkeiten in dem Validitätsscore einen signifikant höheren Wert aufweisen als ein zufällig ausgewählter Proband ohne Auffälligkeit;

Q_2 ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Proband mit Auffälligkeiten in dem Validitätsscore einen signifikant höheren Wert aufweist als zwei zufällig ausgewählte Probanden ohne Auffälligkeit.

Wie oben geschildert, gilt die Gleichung (7.) nur, wenn die beiden ROC-Kurven sich auf Daten unterschiedlicher Patientengruppen beziehen.

Wenn die beiden zu vergleichenden **ROC-Kurven „gepaart" (oder „korreliert")** sind, also aus multiplen Messungen einer Stichprobe stammen, muss entsprechend Hanley & McNeil (1983) [122] der Nenner der Division in Gleichung (8.) erweitert werden mit:

$$(10.) \qquad SE(Area_1) - SE(Area_2) = \sqrt{SE^2(Area_1) + SE^2(Area_2) - 2rSE(Area_1)SE(Area_2)},$$

wobei r ein zusätzlicher Wert ist, der die Korrelation zwischen den zwei AUC's derselben Patientengruppe bezeichnet.

Bei Hanley & McNeil (1983) [122] findet sich ein praktisches Beispiel, wie die Korrelation zwischen den beiden AUC's berechnet werden kann. Für ordinal-skalierte Daten sollten die Kendall - τ (tau) - Korrelationen errechnet werden und für intervallskalierte Daten die Pearson-Produkt-Moment-Korrelation. Hierzu wird zunächst der Mittelwert der Korrelationen zwischen den paarigen Originaldaten der unauffälligen Patientengruppe r_N sowie der auffälligen Patientengruppe r_A berechnet.

Ferner wird ein Mittelwert beider Werte der AUC-Areale berechnet. Aus einer bei Hanley & McNeil (1983) [122] angegebenen Tabelle aller möglichen der beiden Korrelationswerte kann dann die gesuchte Gesamt-Korrelation r abgelesen werden. Diese ist entsprechend der folgenden Formel in die z-Wert-Berechnung einzusetzen.

Der kritische z-Wert zur Entscheidung über die Unterschiedlichkeit zweier ROC-Kurven wird dann folgendermaßen errechnet:

$$(11.) \qquad z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$$

Nach dem *Hanley-McNeil*-Verfahren ist somit die Abschätzung eines signifikanten Unterschiedes von ROC-Kurven über die Berechnung der Differenz der Flächen unter den Kurven (AUC) und der Division dieses Differenzbetrags durch den Standardfehler der Flächendifferenz möglich.

Problem dieses Verfahrens ist, dass die errechnenden Korrelationen (nach Kendall - τ (tau) oder als Pearson-Produkt-Moment-Korrelation) nur in bestimmten Grenzen von den Autoren angegeben wurden (vgl. Tabelle 1 in Hanley & McNeil 1983 [122]). Wenn die Korrelationen hochgradige Zusammenhänge zeigen ($r \geq 0{,}9$) oder die AUC-Flächen $\leq 0{,}7$ oder $\geq 0{,}975$ liegen, können die AUC-Kurven mittels des *Hanley-McNeil*-Verfahrens nicht miteinander verglichen werden.

Für „gepaarte“ ROC-Kurven wurden verschiedene statistische Testverfahren entwickelt, um die Unterschiedlichkeit zweier ROC-Kurven zu bestimmen (Hanley & McNeil 1983 [122], DeLong et al. 1988 [63], Venkatraman & Begg 1996 [307]). Im Falle „ungepaarter“ ROC-Kurven sind alternative Berechnungsverfahren verfügbar (Venkatraman 2000 [306]). Die Kurvenvergleiche können durch die genannten AUC-Differenz-Vergleiche durchgeführt werden (Hanley & McNeil 1983 [122], DeLong et al. 1988 [63]) oder auch anhand des sog. „ROC-Schattens“ (Venkatraman 2000 [306]). Kurvenvergleiche sind auch anhand einer festgelegten Spezifität oder mittels der Konfidenz-Intervalle der ROC-Kurven durchführbar.

Die **in der hiesigen Studie angewandte Methode** zum Vergleich zweier ROC-Kurven wurde von DeLong et al. (1988) [63] beschrieben. **Das DeLong-Verfahren** basiert zunächst auf der nonparametrischen U-Statistik und asymptotischer Normalität, ist jedoch auch auf parametrische Daten anwendbar.

Bei dem *DeLong*-Verfahren werden die Kovarianz-Matrizen der (mindestens zwei) AUC-Kurven berechnet, dann die Varianzen der Differenz zwischen den AUC-Kurven ermittelt, um über die Unterschiedlichkeit beider AUC-Kurven entscheiden zu können. Dabei wird die oben beschriebene Sphärizitäts-Bedingung geprüft, die dann erfüllt ist, wenn die Varianz der Differenzen (Kovarianz) für beliebige Stichprobenpaare gleich ist oder wenn keine zwei Stichprobenpaare mehr oder weniger abhängig voneinander sind als zwei andere Beobachtungspaare.

DeLong et al. (1988) [63] beschreiben das Verfahren folgendermaßen: Für die *r*te Statistik, $\hat{\theta}$, sind die X-Komponenten (*m* Patienten mit Auffälligkeit) und Y-Komponenten (*m* Patienten ohne Auffälligkeit) wie folgt definiert:

$$(12.) \qquad V_{10}^{r}(X_i) = \frac{1}{n}\sum_{j=1}^{n}\Psi\left(X_i^r, Y_j^r\right) \quad (i = 1,2,...,m);$$

$$(13.) \qquad V_{01}^{r}(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\Psi\left(X_i^r, Y_j^r\right) \quad (j = 1,2,...,n).$$

Ebenso sei die *k x k* - Matrix S_{10} (der auffälligen Patienten) definiert, so dass das (r,s)te Element entspricht:

$$(14.) \qquad s_{10}^{r,s} = \frac{1}{m-1}\sum_{i=1}^{m}[V_{10}^{r}(X_i) - \hat{\theta}^r][V_{10}^{s}(X_i) - \hat{\theta}^s]$$

und in gleicher Weise S_{01} (der unauffälligen Patienten) mit dem (r,s)ten Element

$$(15.) \qquad s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^{r}(Y_j) - \hat{\theta}^r][V_{01}^{s}(Y_j) - \hat{\theta}^s].$$

S_{10} und S_{01} sind somit die Kovarianz-Matrizen von V_{10} und V_{01}.

Die approximative Kovarianz-Matrix der Vektoren der Parameterschätzungen, $\hat{\theta} = (\hat{\theta}^1, \hat{\theta}^2, ..., \hat{\theta}^k)$, entspricht somit:

$$(16.) \qquad S = \frac{1}{m}S_{10} + \frac{1}{n}S_{01}.$$

Wenn *g* eine Funktion von $\hat{\theta}$ ist, die sekundären Ableitungen in der Nachbarschaft von $\hat{\theta}$ beinhaltet, und wenn $\lim_{n\to\infty}$ m/n stetig ist und $\geq$ Null, dann ist $N^{\frac{1}{2}}[g(\hat{\theta}) - g(\theta)]$ **asymptotisch normalverteilt** mit Mittelwert 0 und Varianz σ_g^2, mit

$$(17.) \qquad \sigma_g^2 = \lim_{N\to\infty} N\sum_{j=1}^{k}\sum_{i=1}^{k}\frac{\delta g}{\delta\theta^i}\frac{\delta g}{\delta\theta^j}\left(\ \tfrac{1}{m}\xi_{10}^{i,j} + \tfrac{1}{n}\xi_{01}^{i,j}\ \right).$$

Ferner ist

$$(18.) \qquad s_g^2 = N \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{\delta g}{\delta \theta^i} \frac{\delta g}{\delta \theta^j} \left(\frac{1}{m} s_{10}^{i,j} + \frac{1}{n} s_{01}^{i,j} \right)$$

eine konsistente Schätzung für Varianz σ_g^2.

Wenn g eine einfache lineare Funktion ist, reduziert sich die Berechnung erheblich, weil die partiellen Ableitungen Konstanten sind, die die Linearfunktion vorhersagen. Deshalb hat jeder Kontrast $\mathbf{L}\hat{\theta}'$ eine Standard-Normalverteilung, mit $\mathbf{L}$ als Vektor der Koeffizienten:

$$(19.) \qquad z = \frac{\mathbf{L}\hat{\theta}' - \mathbf{L}\theta'}{\left[\mathbf{L} \left(\frac{1}{m} \mathbf{S}_{10} + \frac{1}{n} \mathbf{S}_{01} \right) \mathbf{L}' \right]^{0,5}} = \frac{\mathbf{L}\hat{\theta}' - \mathbf{L}\theta'}{\sqrt{\mathbf{L} \left(\frac{1}{m} \mathbf{S}_{10} + \frac{1}{n} \mathbf{S}_{01} \right) \mathbf{L}'}}.$$

Ein Konfidenzintervall für $\mathbf{L}\theta'$ folgt daraus.

Abschließend ist festzustellen: Da die ROC-Varianz folgendermaßen von der Kovarianz abhängt

$$(22.) \qquad var(\theta_1 - \theta_2) = var(\theta_1) + var(\theta_2) - 2cov(\theta_1, \theta_2),$$

können hoch korrelierte ROC-Kurven mit ähnlichen AUC-Werten trotzdem signifikant unterschiedlich sein (vgl. Robin et al. 2011 [232]).

2.6.8 Varianzanalysen mit Messwiederholung

Bei den zur Fragestellung $H_0 5$ (Analysen zum Therapieerfolg) verwendeten Varianzanalysen mit Messwiederholung sollte zusätzlich eine Homogenität der Varianz-Kovarianz-Matrizen vorliegen. Eine Überprüfung dieser sog. *compount-symmmetry* erfolgte mit dem *Box-M*-Test und dem *Mauchly*-Sphärizitätskriterium (Huynh & Mandeville 1979 [138]). Eine Korrektur der Sphärizitäts- bzw. Zirkularitätsvoraussetzung (bei Mauchly′s W $\leq$ 1) beispielsweise durch Korrektur der Freiheitsgrade des F-Tests für die abhängigen Daten nach der approximativen $\tilde{e}$-Methode von Huynh & Feldt (1976) [138] oder nach Geisser-Greenhouse (IGA) ist bei einmaliger Messwiederholung jedoch nicht erforderlich.

Multivariate Analysen wurden nach dem *Pillai-Bartlett D-Kriterium*-Verfahren berechnet, das im Vergleich mit anderen *union-intersection*-Verfahren (Roy′s Θ, Hotellings-Lawley′s-Kriterium T_0^2/df_e, Wilk′s Lambda) eine vergleichbare Testpower aufweist.

Auch weist diese Teststatistik eine hinreichende Stabilität bezüglich Abweichungen von der Normalverteilungs- und Sphärizitätsbedingung auf (s. Olsen 1976 [218]). Die Prüfung der Einzeleffekte der abhängigen Variablen wurden bei signifikanten Ergebnissen der Varianzanalysen mittels Scheffé-Tests durchgeführt.

Der prognostische Vorhersagewert der Klassifikation für den Therapieerfolg wurde mit einer univariaten Varianzanalyse ANOVA mit den Faktoren Patienten-Subgruppe und der abhängigen Variable VAS-Endbeurteilung der Schmerzintensität untersucht.

3 Ergebnisse

3.1 These $H_0$1: Identifikation von Overreporting mittels MMPI-2-RF

3.1.1 Detektionsgüte der Validitätsskalen seltener somatischer, psychischer und kognitiver Symptome

In der ersten Untersuchungshypothese waren deutliche Effekte des bei einigen der untersuchten Patienten mit chronischen Schmerzen durch externe Messinstrumente festgestellten Overreporting-Verhaltens auf ihre Antworten in den Validitätsskalen des MMPI-2-RF erwartet worden.

Diese Effekte wurden aufgrund einer vermuteten höheren Trennschärfe der MMPI-2-RF-Validitätsskalen gegenüber ihrer Vorgänger-Version MMPI-2 erwartet.

Aber auch vorangehende Studien mit Einsatz des MMPI-2-RF, die nach einem ähnlichen externen Klassifikations-Modell als sog. *Known-Groups-Design* durchgeführt wurden, bestätigten die Nützlichkeit dieser Validitätsskalen zur Aufdeckung von Beschwerdeüberhöhungen. Ähnliche Effekte waren von Aguerrevere (2010) [2] und Anderson (2011) [9] bei Gutachten-Patienten mit chronischen Schmerzen beobachtet worden, von denen allerdings anzunehmen ist, dass sie (ähnlich wie forensische Patienten, z.B. Gervais et al. 2010 [101]) schon allein aufgrund ihres besonderen Anliegens (Kompensationswunsch) eine ausgeprägtere Beschwerdedarstellung aufweisen würden als ein Schmerzpatient ohne diese besondere Konfliktlage.

Die Ergebnisse der univariaten Varianzanalysen zeigen Folgendes:

In fünf MMPI-2-RF-Validitätsskalen und den Kontrollskalen L-r und K-r zeigten sich zwischen den vier **Klassifikationsgruppen für Overreporting hoch signifikante Unterschiede** (vgl. Tab. 6).

Tab. 6. MANOVA: Standard-MMPI-2-Validitätsskalen in den vier Studiengruppen

	Pillai's D	Exaktes F	D.F.	Fehler-D.F.	p	SN
MMPI-2-RF Validität	0,984	2316,37	7,00	358,00	0,000	***

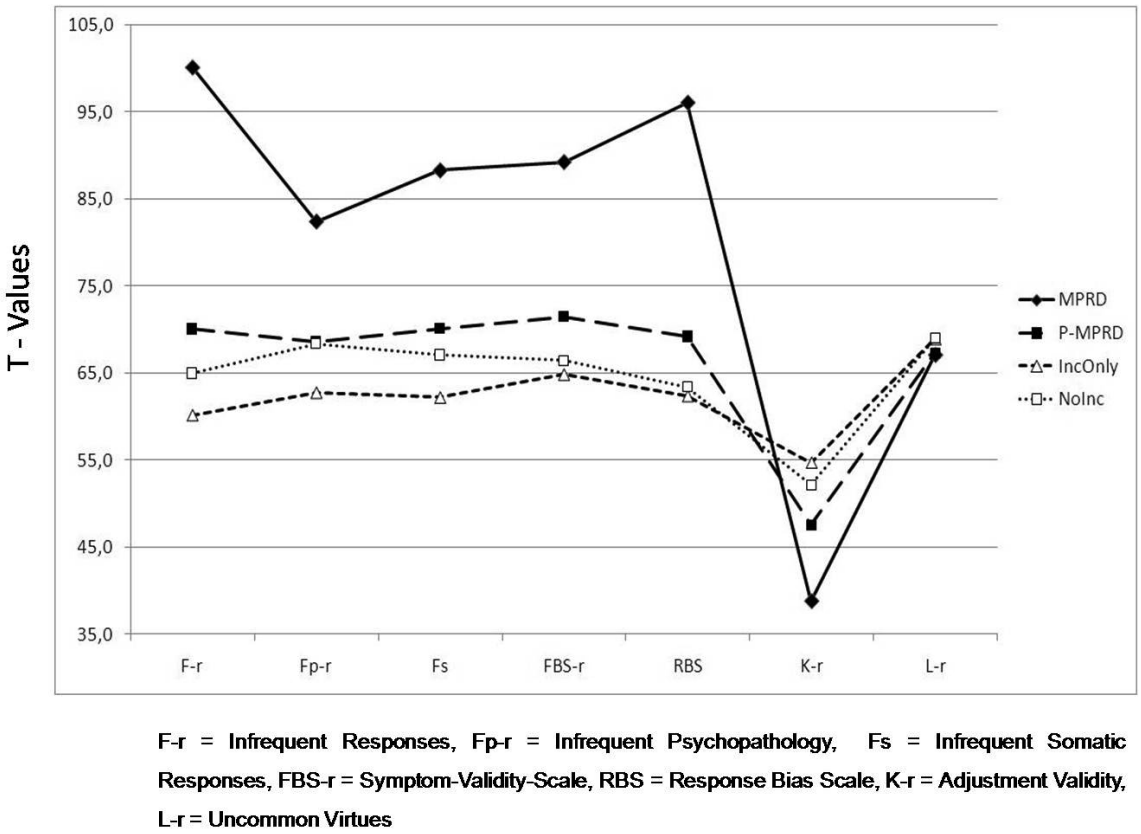


Abb. 12. Profile der vier Studiengruppen in den Standard-Validitätsskalen des MMPI-2-RF

Diese Unterschiede sind in Abb. 12 veranschaulicht. Sie zeigen je nach Overreporting-Klassifikation deutlich unterschiedliche Validitäts-Subskalenwerte der vier Patientengruppen.

Die auffälligsten Werte in allen Validitätsscores zeigte die MPRD-Gruppe, mit entsprechend der Erwartung, niedrigsten Scores in der K-r-(Korrektur)-Skala bei den Patienten mit Overreporting, da diese Skala eher Beschwerdebagatellisierung oder sog. *Underreporting* erfasst. Die Patienten mit möglicher Beschwerdeaggravation (P-MPRD) zeigten die zweithöchsten Auffälligkeiten in den Validitätsskalen, die jedoch gegenüber den MPRD-Patienten weit geringer ausfielen. Patienten ohne und mit Kompensationsmotiven (NoINC, IncOnly) zeigten eher ähnliche Validitätsprofile, wobei unerwartet jene Patienten, die allein externale Aggravationsmotive (IncOnly) aufwiesen, die niedrigsten Auffälligkeiten zeigten.

Alle Validitätsskalen, außer der Lie-r-Skala, differenzierten signifikant zwischen der stark auffälligen MPRD-Gruppe und den übrigen drei Probandengruppen.

Mittelwert-Vergleiche mittels ANOVAs (vgl. Tab. 7, S. 161) über die vier Untersuchungsgruppen hinweg zeigten, dass die F-r-Skala und die RBS-Skala am deutlichsten unter den

Standard-Validitätsskalen zwischen Patientengruppen trennen (F-r-Skala: $F_{364,3}$ = 60,7, p = 0,000; RBS-Skala: $F_{364,3}$ = 50,5, p = 0,000). Diese Ergebnisse bleiben signifikant, auch bei Berücksichtigung einer die Vielzahl durchgeführter Gruppenvergleiche einbeziehender Alpha-Adjustierung (Signifikanzniveau 0,05/9 = 0,006).

Tab. 7. ANOVAs: Standard-MMPI-2-RF-Validitätsskalen in den Studiengruppen

	1	2	3	4			
Gruppe	No Incentive	Incentive Only	Probably MPRD	MPRD			Innere Konsistenz
N Skala	231 M (SD)	78 M (SD)	34 M (SD)	25 M (SD)	F-Tests F(3, 364)	Scheffé-Tests	Cronbach's α
F-r	65,0 (14,4)	60,2 (11,0)	70,1 (11,8)	100,2 (12,3)	60,7 ***	4 > 3,2,1 3 > 2	0,745
Fp-r	68,4 (16,7)	62,8 (14,8)	68,7 (18,1)	82,4 (16,3)	9,2 ***	4 > 3,2,1	0,422
Fs	67,1 (16,1)	62,2 (13,8)	70,2 (14,6)	88,3 (20,8)	17,5 ***	4 > 3,2,1	0,537
FBS-r	66,5 (12,7)	64,9 (10,6)	71,5 (14,0)	89,3 (13,3)	28,1 ***	4 > 3,2,1	0,711
RBS	63,4 (13,2)	62,4 (11,5)	69,3 (13,4)	96,1 (14,7)	50,5 ***	4 > 3,2,1	0,677
K-r	52,1 (9,2)	54,7 (11,0)	47,6 (8,4)	38,8 (6,4)	20,7 ***	4 < 3,2,1 3 < 2	0,677
L-r	69,0 (11,7)	69,0 (11,9)	67,4 (11,2)	67,1 (9,5)	0,4 n.s.	n.s.	0,510
HHI-r	4,5 (2,0)	4,4 (1,8)	5,5 (1,8)	8,1 (1,6)	28,0 ***	4 > 3,2,1	0,569
MI-r	1,6 (2,1)	0,9 (1,4)	2,1 (2,3)	7,1 (2,2)	65,3 ***	4 > 3,2,1 3 > 2	0,789

Folgende Post-Hoc-Analysen mittels Scheffé-Tests bestätigten, dass insbesondere die Patienten mit MPRD-Klassifikation in allen Validitätsskalen außer der L-r-Skala auffälligere Werte (im Fall der K-r-Skala somit niedrige Werte) aufwiesen, als alle drei anderen Untersuchungsgruppen. Probanden mit einem möglichen Overreporting (P-MPRD) wiesen zudem gegenüber den Patienten mit alleinigem Kompensationsmotiv (IncOnly) höhere Werte in der F-r-Skala und geringere Werte in der Korrektur-Skala auf. Sie unterschieden sich jedoch in keiner anderen Skala voneinander.

Patienten ohne externale Kompensationsmotive (NoINC) wiesen in allen fünf Validitätsskalen (F-r, Fp-r, Fs, FBS-r und RBS) gegenüber den Probanden mit Kompensationsmotiven

(IncOnly) entgegen der Erwartung die höheren Auffälligkeiten auf (s. Abb. 12, S. 160). Die Feststellung eines Kompensationsmotivs rechtfertigt somit nicht grundsätzlich die Annahme einer Beschwerdeüberhöhung. Dieses Ergebnis bedarf einer ausführlichen Interpretation, auch im Zusammenhang mit ähnlichen Ergebnissen einer Studie von Bianchini et al. (2008) [30] (vgl. S. 71).

Betrachtet man zusätzlich die ebenfalls für den MMPI-2-RF konzipierte Validitätsskala Henry-Heilbronner-Index (HHI-r, Henry et al. 2008) [133]), so waren mittels dieses Kurz-Index von nur 11 Items die Patienten mit MPRD ebenso gut wie mit der FBS-r gegenüber den drei anderen Patientengruppen detektierbar ($F_{364,3}$ = 28,0, p = 0,000).

Mittels des Gesamt-Validitätsindex MI-r (Meyers et al. 2013 [201]), der gewichtet ein Symptom-Overreporting in allen fünf Standard-Validitätsskalen des MMPI-2-RF einbezieht, waren die MPRD-Patienten gegenüber allen drei anderen Untersuchungsgruppen am besten von allen Subscores identifizierbar ($F_{364,3}$ = 65,3, p = 0,000). Der MI-r zeigte zudem in den nachfolgend durchgeführten Scheffé-Tests signifikant höhere Werte der Patienten mit möglicher Beschwerde-Aggravation (P-MPRD) im Vergleich zu den Patienten mit alleinigen externen Kompensationsmotiven (IncOnly).

Die drei festgestellten **Haupteffekte** wurden in der Tabelle **mit grauer Markierung** gekennzeichnet.

Um einen Vergleich mit anderen Studien zu ermöglichen, wurden **zusätzlich** die **Effektstärken** (Cohen's d, vgl. Tab. 8, S. 164) für zwei Vergleichsgruppen von Patienten berechnet: (1.) für die Probanden ohne Auffälligkeiten in den BV-Verfahren (Gruppen NoINC und IncOnly) im Vergleich zu Patienten mit Auffälligkeiten (Possible und Definite MPRD); (2.) für den Vergleich der Patienten mit definitiver „MPRD" und den Patienten, die definitiv als „Nicht-MPRD" klassifiziert wurden.

Die ermittelten Cohen's-d-Werte zwischen den Gruppen zeigen relativ hohe Unterschiede (Werte > 1,0) an. Eine mögliche Erklärung könnte die in der hiesigen Studie zur Klassifikation einer „MPRD" gewählte, sehr konservative Entscheidungsregel sein. Die meisten **internationalen Vergleichsstudien** verwenden weniger strenge Klassifikations-Regeln (die zumeist nur auf PVT-Außenkriterien beruhen) oder aber Stichproben von homogeneren Probanden (z.B. nur Gutachten-Probanden), die sich untereinander weniger unterscheiden, als die hier untersuchten Therapie-Patienten.

Betrachtet man zum besseren Vergleich mit anderen Studien nur die Unterschiede zwischen Patienten mit möglicher / definitiver MPRD und Patienten ohne Auffälligkeit in den externen Klassifikationstests, so fallen entsprechend der Erwartung die höchsten Gruppenunterschiede in der F-r- und der RBS-Skala (d = 1,30 und 1,25) auf.

Der HHI-r zeigte erneut trotz Itembeschränkung eine hohe Detektionsgüte (d = 1,04); eine deutliche Unterscheidung zwischen den Klassifikationen „mögliche / sichere MPRD“ und „Nicht-MPRD“ zeigte sich im Meyers-Validity-Index MI-r (d = 1,25). Die Fp-r-Skala (Infrequent Psychopathology) zeigte die geringste Effektstärke (d = 0,45).

Anderson (2011) [9] berichtete in einem Sample mit Gutachten-Probanden höhere Effektstärken für die RBS-Skala (1,67), für die F-r-Skala (1,63), für die Fs-Skala (1,37), die FBS-r-Skala (1,16) sowie für die Fp-r-Skala (0,93). Dies ist vermutlich auf die in dieser Studie getroffene Auswahl trennschärferer US-amerikanischer Außenkriterien zurückzuführen.

Bei ausschließlichem Bezug auf die in der hiesigen Studie durchgeführte strengere Klassifikation definitiver „MPRD“ waren jedoch deutlich klarere Gruppenunterschiede in allen Validitätsskalen festzustellen.

Bianchini et al. (2008) [30] verglichen **Validitätsskalen des MMPI-2** von Patienten mit chronischen Schmerzen mit „definitiver MPRD“ mit einer Patientengruppe ohne überhöhte Beschwerdeangaben („No-Incentive / IncOnly“). Die Autoren fanden Effektstärken für die F-Skala von 1,8, für die FBS-Skala von 2,4, den Meyers-Validity-Index von 2,2 und für die Fp-Skala von 1,1. Im Vergleich zu der hier vorliegenden Studie zeigte die F-Skala des MMPI-2 eine geringere Detektionsstärke. Die übrigen Skalen zeigten eher vergleichbare Effekte gegenüber den Werten des MMPI-2-RF der hiesigen Studie.

Eine Einordnung dieser Resultate kann erst im Zusammenhang mit den Ergebnissen zur folgenden Hypothese („Vergleich zwischen beiden MMPI-Versionen“) erfolgen.

Vergleicht man die ermittelten Cohen’s-d-Werte mit den von Thies (2012) [296] ermittelten Effektstärken für die **Validitätsskalen des MMPI-2**, so zeigten diese zwischen den als „authentisch“ und „nicht-authentisch“ klassifizierten Probanden **geringere Werte** von 1,02 (HHI), 0,92 (FBS), 0,89 (RBS), 0,73 (F), 0,70 (MVI) und 0,35 (Fp-Skala). Dies könnte einerseits an der geringeren Trennschärfe der MMPI-2-Skalen liegen oder an der ausschließlichen Verwendung von externen PVTs.

Tab. 8 zeigt zusätzlich die **inneren Konsistenzwerte** der einzelnen Validitätsskalen (Cronbach's α) über alle 368 MMPI-2-RF-Protokolle hinweg, die sich gegenüber denselben Werten der Publikations-Stichprobe mit geringerer Probandenanzahl (n = 275) nur geringfügig veränderten (maximale Abweichung $\pm$ 0,02). Auffallend ist ein deutlicher Zusammenhang zwischen der Höhe der inneren Konsistenz und der Diskriminanzgüte der Skalen. Entsprechend internationalen Maßstäben sind die ermittelten Effektstärken als gut zu bezeichnen, zwischen „MPRD" und „Nicht-MPRD" sogar als außergewöhnlich hoch.

Tab. 8. Effektstärken der MPRD-Detektion: Standard-MMPI-2-RF-Validitätsskalen

Gruppe	NoInc & IncOnly	Possible & Def. MPRD	F-Test	d	No MPRD	Definite MPRD	F-Test	d
N	309	59			343	25		
	M	M			M	M		
Skala	(SD)	(SD)			(SD)	(SD)		
F-r	63,8	82,8	83,2	1,30	64,4	100,2	162,0	2,64
	(13,7)	(19,1)	***		(13,6)	(12,3)	***	
Fp-r	67,0	74,5	10,0	0,45	67,2	82,4	19,9	0,92
	(16,4)	(18,5)	**		(16,5)	(16,3)	***	
Fs	65,9	77,9	26,6	0,73	66,3	88,3	44,0	1,37
	(15,7)	(19,5)	***		(15,6)	(20,8)	***	
FBS-r	66,1	79,1	50,0	1,00	66,6	89,3	76,3	1,80
	(12,2)	(16,2)	***		(12,5)	(13,3)	***	
RBS	63,2	80,6	77,0	1,25	63,8	96,1	142,5	2,47
	(12,8)	(19,2)	***		(13,0)	(14,7)	***	
K-r	52,8	43,8	43,3	0,94	52,3	38,8	46,7	1,42
	(9,7)	(8,7)	***		(9,7)	(6,4)	***	
L-r	69,0	67,2	1,2	0,15	68,9	67,1	0,5	0,15
	(11,7)	(10,4)	n.s.		(11,7)	(9,5)	n,s,	
HHI-r	4,5	6,6	53,4	1,04	4,6	8,1	75,1	1,80
	(1,9)	(2,2)	***		(2,0)	(1,6)	***	
MI-r	1,4	4,2	77,2	1,25	1,5	7,1	180,6	2,78
	(1,9)	(3,3)	***		(2,0)	(2,2)	***	

d = Cohen's d: jeweils zwischen beiden Klassifikationsgruppen

Allein die Kontrollskala L-r („Lügenskala") erfüllte diese Anforderungen nicht. Ebenso wurde bestätigt, dass die Skalen F-r und RBS die höchste Detektionsgüte unter den Validitätsskalen aufweisen.

Diese Ergebnisse bleiben signifikant, auch bei Berücksichtigung einer die Vielzahl durchgeführter Gruppenvergleiche einbeziehender Alpha-Adjustierung (Signifikanzniveau 0,05 / 9 = 0,006).

Zusammenfassend wurde somit die Hypothese $H_0 1$ bestätigt, nach der chronische Schmerzpatienten mit MPRD durch überhöhte Scores in allen fünf MMPI-2-RF-Validity-Scales identifizierbar sind.

3.1.2 Aussagekraft der Restrukturierten RC-Skalen zur klinischen Symptomatik

Auch die neun Basisskalen des MMPI-2-RF wiesen entsprechend der durchgeführten Multivariaten Analyse hoch signifikante Unterschiede zwischen den vier Klassifikationsgruppen auf (vgl. Tab. 9).

Tab. 9. MANOVA: MMPI-2-RF-Basisskalen in den vier Studiengruppen

	Pillai's D	Exaktes F	D.F.	Fehler-D.F.	p	SN
MMPI-2-RF (RC)	0,988	3187,91	9,00	356,00	0,000	***

Diese Unterschiede sind in Abb. 13 (S. 166) veranschaulicht. Sie spiegeln deutlich unterschiedliche Basisskalenwerte der vier Klassifikationsgruppen, je nach Overreporting-Ausmaß, wider.

Patienten mit einer MPRD-Klassifikation zeigten in den ersten drei Basisskalen (Demoralisation, Somatic Complaints und Low Positive Emotions) des MMPI-2-RF deutliche Score-Erhöhungen, aber auch in den Basisskalen RC6 bis RC8 (Ideas of Persecution, Dysfunctional Negative Emotions und Aberrant Experiences). Sie zeigten somit ein eher generell erhöhtes Basisskalen-Profil (vgl. Aguerrevere 2010 [2]) - mit Ausnahme der Skalen RC4 (Antisocial Behavior) und RC9 (Hypomanic Activation).

Diese Ergebnisse bleiben signifikant, auch bei Berücksichtigung der Gruppenvergleiche mittels Alpha-Adjustierung (Signifikanzniveau 0,05 / 9 = 0,006).

Die als möglicherweise aggravierend klassifizierten Probanden (P-MPRD) wiesen hingegen gegenüber den Patienten ohne BV-Auffälligkeiten nur leicht erhöhte Basisskalen-Werte auf, insbesondere in der Skala RC2 (Low Positive Emotions).

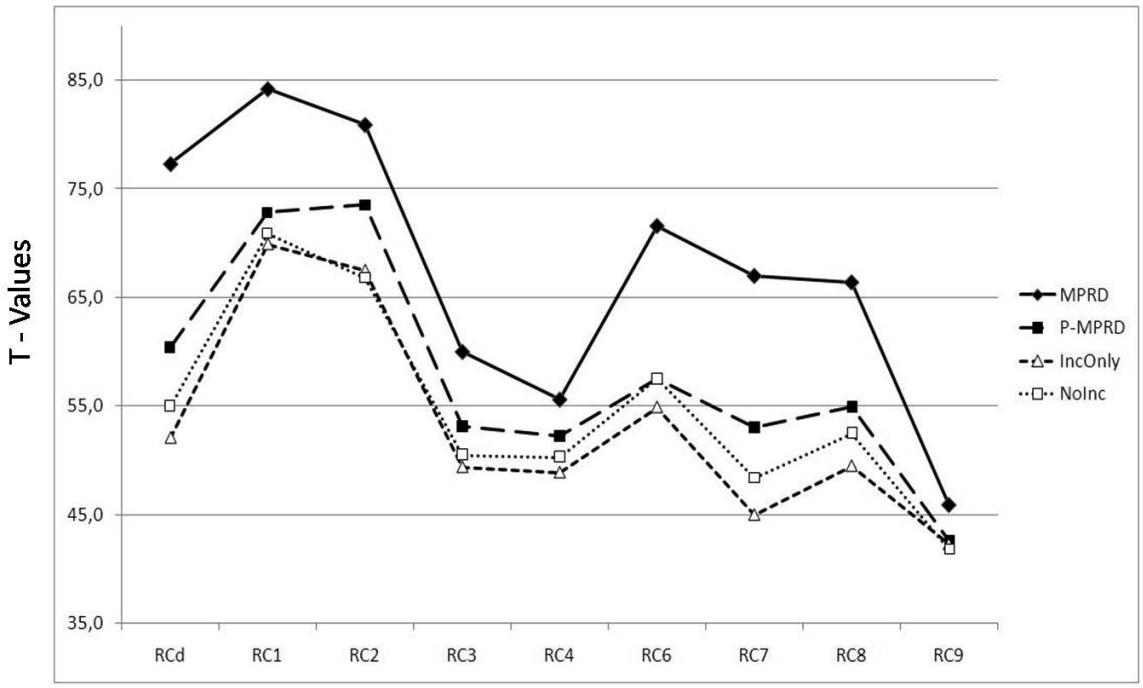


Abb. 13. Profile der vier Studiengruppen in den Basisskalen des MMPI-2-RF

Patienten mit externen Kompensationsmotiven (IncOnly) zeigten wiederum gegenüber den Patienten ohne irgendeine Auffälligkeit die niedrigsten Werte in den MMPI-2-RF-Basisskalen.

Post-Hoc-Analysen mittels Scheffé-Tests zeigten, dass sich die MPRD-Patienten in vier Basisskalen von allen drei anderen Untersuchungsgruppen unterschieden: in den Skalen RCd (Demoralisation), RC1 (Somatic Complaints), RC6 (Ideas of Persecution) und RC8 (Aberrant Experiences). Patienten mit möglicher Beschwerdeaggravation (P-MPRD) wiesen gegenüber den Patienten ohne BV-Auffälligkeiten (NoInc, IncOnly) nur in der Basisskala RCd höhere Werte auf. In den Skalen RC2, RC7 und RC8 zeigte sich dies nur gegenüber den Patienten ohne BV-Auffälligkeiten und ohne externe Kompensationsmotive (NoInc).

Im Gruppenvergleich zwischen Patienten „mit“ und „ohne MPRD“ bestätigten sich die auch graphisch deutlichen Überhöhungen der MPRD-Patienten in den genannten sechs Basisskalen (RCd, RC1, RC2, RC6 bis RC8) als signifikant (s. Tab. 10, S. 167). Insbesondere in der RCd-Skala (Demoralization) wiesen die MPRD-Patienten die deutlichsten Beschwer-

Tab. 10. ANOVAs: MMPI-2-RF-Basisskalen in den Studiengruppen

	1	2	3	4			
Gruppe	No Incentive	Incentive Only	Probably MPRD	MPRD			
N	231	78	34	25	F-Tests	Scheffé-Tests	d
Skala	M (SD)	M (SD)	M (SD)	M (SD)			
RCd	55,0 (10,6)	52,1 (8,5)	60,4 (9,4)	77,3 (6,4)	45,9 ***	4 > 3,2,1 3 > 2,1	2,23
RC1	70,9 (9,6)	69,9 (8,8)	72,8 (8,2)	84,2 (8,7)	17,0 ***	4 > 3,2,1	1,45
RC2	66,8 (13,0)	67,5 (12,2)	73,5 (14,2)	80,9 (10,6)	11,0 ***	4 > 2,1 3 > 1	1,03
RC3	50,5 (10,4)	49,4 (9,3)	53,1 (11,9)	60,0 (11,9)	7,4 ***	4 > 2,1	0,91
RC4	50,3 (8,0)	48,9 (6,9)	52,2 (6,5)	55,6 (9,3)	5,3 **	4 > 2,1	0,70
RC6	57,5 (12,2)	54,9 (12,7)	57,5 (12,1)	71,6 (12,6)	11,9 ***	4 < 3,2,1	1,19
RC7	48,4 (10,2)	45,0 (8,4)	53,0 (8,4)	67,0 (9,2)	35,6 ***	3 > 2	1,92
RC8	52,5 (7,8)	49,5 (7,0)	54,9 (6,8)	66,4 (11,6)	30,6 ***	4 < 3,2,1 3 > 2, 2 > 1	1,80
RC9	41,8 (8,0)	42,3 (8,6)	42,6 (8,8)	45,9 (7,2)	1,9 n.s.	n.s.	0,48

d = Cohen's d: zwischen Klassifikationsgruppe „MPRD" vs. „Nicht-MPRD"

denangaben auf ($F_{366,1}$ = 115,9, p = 0,000), die auch die ermittelte Effektstärke bestätigte (Cohen's d = 2,23).

Auch unter Einbezug der nur möglicherweise aggravierenden Probanden (P-MPRD und MPRD) machten diese Patienten mit Auffälligkeiten in den Verfahren zur Beschwerdenvalidierung in fast allen Basisskalen des MMPI-2-RF (außer der Skala RC9) deutlich höhere Beschwerdeangaben als die Patienten ohne BV-Auffälligkeiten (NoInc, IncOnly).

Die Basisskalen-Gruppenwerte bestätigten damit indirekt das externe Modell der Overreporting-Klassifikation. Eine ausführliche tabellarische Darstellung der Zweigruppenvergleiche zur Ermittlung der Effektstärken nach Cohen findet sich im Anhang (s. Tab. 25, S. 275).

Damit wurden Ergebnisse erster US-amerikanischer Auswertungen zum neuen MMPI-2-RF bestätigt (s. Ben-Porath 2011 [27]), aber auch Resultate jüngster amerikanischer Studien

zu diesem Fragebogen (z.B. Anderson (2011) [9], Aguerrevere (2010) [2] oder Sellbom et al. (2010) [272]).

Zusammenfassend zeigten sich somit Beschwerdeüberhöhungen im MMPI-2-RF in der hiesigen Untersuchung wie auch in vorangehenden Studien nicht nur in Auffälligkeiten der Validitätsskalen, sondern auch in einem generell überhöhtem Basisprofil (Restructured Scales).

3.2 These $H_0$2: Identifikation von Overreporting mittels des deutschen BHI-2

3.2.1 Detektionsgüte der Validitätsskalen Disclosure, Defensiveness

Als zweite Untersuchungshypothese wurden vermutet, dass sich bestimmte Personen in einer Patientengruppe mit chronischen Schmerzen insbesondere mittels der Validitätsskala „Disclosure“ des BHI-2 identifizieren lassen, deren Beschwerdedarstellung durch andere, externe Validierungsverfahren als nicht authentisch klassifiziert wurde.

Eine solche nicht-authentische Beschwerdedarstellung sollte sich reziprok in einem geringerem „Under-Statement“ / Underreporting von Symptomen in der BHI-2-Validitätsskala „Defensiveness“ bei Patienten mit MPRD-Klassifikation zeigen.

Die varianzanalytische Untersuchung dieser Hypothese zeigte nun das folgende Resultat: Die Angaben der Probanden in den drei Validitätsskalen des BHI-2 wiesen **hoch signifikante Unterschiede** zwischen den vier Klassifikationsgruppen für Overreporting auf (vgl. Tab. 11).

Tab. 11. MANOVA: BHI-2-Validitätsskalen in den vier Studiengruppen

	Pillai’s D	Exaktes F	D.F.	Fehler-D.F.	p	SN
BHI-2-Skalen	0,983	7136,65	3,00	362,00	0,000	***

Mittelwert-Vergleiche mittels ANOVAs (vgl. Tab. 12, S. 170) zwischen den vier Untersuchungsgruppen belegten, dass die Disclosure-, aber auch die Defensiveness-Skala trennscharf MPRD-Patienten von den drei übrigen Probandengruppen trennen ($F_{364,3}$ = 30,3, p = 0,000).

In der DIS-Skala wiesen MPRD-Patienten die deutlich höchsten Scores auf (MW = 60,3; SD = 9,1), wobei auch Patienten mit möglicher Aggravation (P-MPRD) mehr Symptome als die Patienten mit externen Kompensationsmotiven (INC) angaben. MPRD-Patienten wiesen erwartungskonform in der „Underreporting“-DEF-Skala gegenüber den Patientengruppen ohne Aggravations-Auffälligkeit (NoINC, IncOnly) deutlich niedrigere Werte auf (MW = 42,7 vs. 46,8 bzw. 47,3), unterschieden sich jedoch nicht signifikant von den Patienten mit nur möglicher MPRD.

Der aus nur vier Items bestehende sog. Validitätsindex des BHI-2 differenzierte entsprechend der Scheffé-Tests nicht signifikant zwischen den Gruppen, wenngleich MPRD-

Tab. 12. MANOVA: Ausprägung der BHI-2-Validitätsskalen in den vier Studiengruppen

	1	2	3	4			
Gruppe	No Incentive	Incentive Only	Probably MPRD	MPRD			Innere Konsistenz
N	231	78	34	25	F-Tests	Scheffé-	Cronbach's
	M	M	M	M	F(3, 364)	Tests	α
Skala	(SD)	(SD)	(SD)	(SD)			
DIS	44,8	41,5	49,3	60,3	30,3	4 > 3,2,1	0,929
	(9,3)	(9,1)	(8,9)	(9,1)	***	3 > 2	
DEF	46,8	47,3	42,7	36,8	10,4	4 > 2,1	0,583
	(9,1)	(10,1)	(10,3)	(9,5)	***		
Validitäts-	0,1	0,0	0,2	0,3	3,3	n.s.	0,483
Index	(0,5)	(0,1)	(0,7)	(0,5)	*		

Patienten hier die höchsten Inkonsistenzen (Befürwortung „absurder" Items) zeigten. Er empfiehlt sich damit eher als Kontroll-Instrument der Validität der Test-Bearbeitung.

In Tab. 13 (S. 171) sind ergänzend die Effektstärken zwischen jeweils zwei Probandengruppen für die Validitätsskalen des BHI-2 gegenüber gestellt. Die DIS- und DEF-Skala zeigten hinreichend hohe Effektstärken zur Erhebung von „Over-" und „Underreporting" (1,72 DIS-Skala; 0,71 DEF-Skala), mit einem höheren Differenzierungsgrad der DIS-Skala als bei drei der Standard-Validitätsskalen des MMPI-2-RF (Fp-r, Fs und K-r).

Tab. 12 zeigt zusätzlich die innere Konsistenz der Validitätsskalen (Cronbach's α) über alle MMPI-2-RF-Protokolle hinweg. Die DIS-Skala wies die höchste innere Konsistenz auf (Cronbach's $\alpha = 0{,}929$). Auffallend war ein Zusammenhang zwischen der Höhe der inneren Konsistenz und der Diskriminanzgüte der Skalen.

Die Hypothese $H_0 2$ wurde somit bestätigt, nach der chronische Schmerzpatienten mit MPRD durch überhöhte Scores in der Skala DIS (Discloure) gegenüber Patienten mit chronischen Schmerzen ohne Auffälligkeiten in externen Verfahren zur Beschwerdenvalidierung identifizierbar sind.

Tab. 13. Effektstärken der MPRD-Detektion: BHI-2-Validitätsskalen

Gruppe	NoInc & IncOnly	Possible & Def. MPRD	F-Test	d	No MPRD	Definite MPRD	F-Test	d
N	309	59			343	25		
Skala	M (SD)	M (SD)			M (SD)	M (SD)		
DIS	43,5 (9,1)	54,0 (10,5)	57,3 ***	1,08	44,5 (9,2)	60,3 (9,1)	68,9 ***	1,72
DEF	47,0 (9,3)	40,2 (10,3)	25,1 ***	0,71	46,5 (9,5)	36,8 (9,5)	24,5 ***	1,03
Val-Index	0,1 (0,4)	0,3 (0,7)	4,7 *	0,31	0,1 (0,5)	0,3 (0,5)	2,7 n.s.	0,34

d = Cohen's d: jeweils zwischen beiden Klassifikationsgruppen

3.2.2 Aussagekraft der BHI-2 Basisskalen zur Aggravations-Messung

Als Nebenfragestellung wurde in der zweiten Untersuchungshypothese vermutet, dass sich im BHI-2 ausgeprägte Beschwerdeüberhöhungen von Patienten mit MPRD-Klassifikation auch in der Mehrzahl der BHI-Subskalen abbilden sollten.

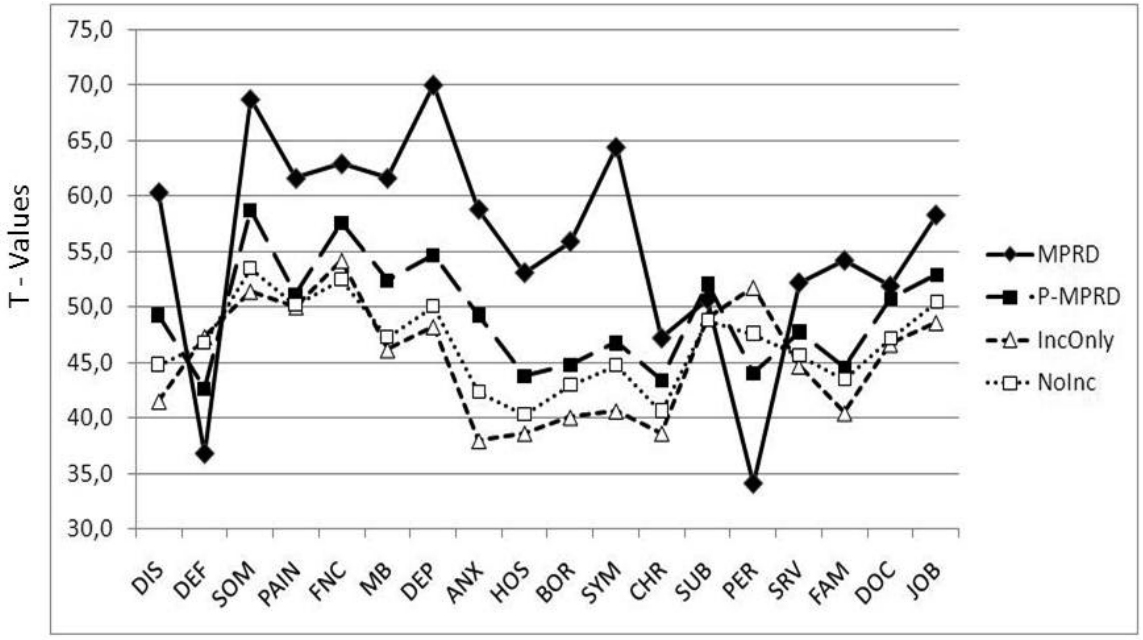
Diese Annahme beruhte darauf, dass ähnliche Effekte auch von den Testautoren bei beschwerdensimulierenden Probanden gegenüber Norm-Patienten mit chronischen Schmerzen und anderen Probanden ohne Aggravation gefunden wurden (vgl. Bruns & Disordio 2000 [42], s. S. 114).

Inhaltsabhängig erfassen die BHI-2-Basisskalen zudem diverse, für Patienten mit chronischen Schmerzen sehr spezifische Problembereiche, die in anderen Testinventaren nicht erfragt werden (z.B. FAM = Familiäre Konflikte und Probleme; Doctor - Dissatisfaction = Unzufriedenheit mit der ärztlichen Behandlung; Job - Dissatisfaction = Probleme im beruflichen Umfeld, Mobbing, Schwierigkeiten mit dem Arbeitsplatz und Arbeitgeber).

Wenn Overreporting auch emotional-motivational gesteuert wird, dann ist zu vermuten, dass Patienten mit besonders emotions-belasteter Symptomatik die Items solcher Subskalen außergewöhnlich häufig mit besonderer Intensität angeben.

Die Analyse der BHI-2-Basisskalen zeigte nun folgendes Ergebnis:

In **fast allen Basisskalen** des BHI-2 wurden Beschwerden-Überhöhungen der Patienten mit MPRD gegenüber den drei anderen Studiengruppen ersichtlich (mit Ausnahme des Subscores SUB = Substanzabusus und Doctor-Dissatisfaction). Probanden mit möglicher Be-



DIS = Disclosure, DEF = Defensiveness, SOM = Somatic Complaints, PAIN = Pain Complaints, FNC = Functional Complaints, MB = Muscular Bracing , DEP = Depression, ANX = Anxiety, HOS = Hostility, BOR = Borderline, SYM = Symptom Dependency , CHR = Chronic Maladjustment , SUB = Substance Abuse , PER = Perseverance, FAM = Family Dysfunction , SRV = Survivor of Violence , DOC = Doctor Dissatisfaction , JOB = Job Dissatisfaction , BHICI= Critical Items

Abb. 14. Profile der vier Studiengruppen in den BHI-2-Validitätsskalen

schwerdenaggravation (P-MPRD) zeigten zudem im Gruppenvergleich die zweithöchsten Scores ebenfalls über alle Basisskalen des BHI-2 hinweg. Jene Patienten, die hingegen keine Auffälligkeit in den BV-Verfahren aufwiesen (NoINC, IncOnly) zeigten in den BHI-2-Subskalen nahezu ähnliche Werte mit mittleren T-Scores (zwischen T40 bis T50).

In Abb. 14 sind die BHI-2-Validitäts- und Basisskalen-Scores summarisch als standardisierte T-Werte für die vier Studiengruppen illustriert.

Bei den nachfolgenden Mittelwert-Vergleichen mittels ANOVAs (vgl. Tab. 14, S. 173) wurde eine Alpha-Adjustierung zur Beurteilung der Signifikanz (Signifikanzniveau 0,05/17 = 0,003) angewandt.

Die Gruppenvergleiche zeigten signifikante Unterschiede in fast allen Basisskalen: nach α-Adjustierung mit Ausnahme der Skalen Substanz-Fehlgebrauch (SUB), Trauma (SRV) und Therapie-Unzufriedenheit (DOC). Diese Unterschiede beruhten entsprechend der Post-Hoc-Scheffé-Tests überwiegend auf höheren Scores der MPRD-Patienten gegenüber den drei übrigen Studiengruppen, und zwar in fast allen BHI-2-Basisskalen. Nur in vier der 16 BHI-2-Basisskalen machten Patienten mit möglichem Overreporting gleich hohe Angaben

wie die Patienten mit sicherer MPRD (SFC = Functional Complaints, ANX = Angst, CHR = Chronic Maladjustment, SRV = Survivor of Violence).

Tab. 14. ANOVAs: BHI-2-RF-Basisskalen in den Studiengruppen

	1	2	3	4			
Gruppe	No Incentive	Incentive Only	Probably MPRD	MPRD			
N	231	78	34	25	F-Tests	Scheffé-Tests	d
Skala	M (SD)	M (SD)	M (SD)	M (SD)			
SOM	53,5 (9,6)	51,4 (6,6)	58,7 (9,3)	68,7 (7,6)	28,3 ***	4 > 3,2,1 3 > 2,1	1,68
PAIN	50,2 (8,9)	50,0 (9,1)	51,1 (8,3)	61,6 (10,1)	12,5 ***	4 > 3,2,1	1,26
SFC	52,5 (8,1)	54,1 (8,1)	57,6 (8,8)	62,9 (8,4)	14,9 ***	4 > 2,1 3 > 1	1,15
MB	47,3 (12,4)	46,1 (11,9)	52,4 (11,7)	61,6 (8,0)	13,1 ***	4 > 3,2,1	1,17
DEP	50,0 (9,5)	48,2 (8,8)	54,7 (11,0)	70 (7,9)	38,6 ***	4 > 3,2,1 3 > 2	2,09
ANX	42,3 (13,4)	37,9 (11,0)	49,3 (15,3)	58,8 (10,8)	19,2 ***	4 > 2,1 3 > 2,1	1,27
HOS	40,3 (9,3)	38,6 (8,5)	43,8 (9,1)	53,1 (10,5)	17,4 ***	4 > 3,2,1	1,38
BOR	43,0 (9,1)	40,1 (6,4)	44,8 (7,5)	55,9 (11,1)	21,9 ***	4 > 3,2,1	1,54
SYM	44,7 (17,4)	40,6 (16,2)	46,8 (19,2)	64,4 (11,5)	12,6 ***	4 > 3,2,1	1,20
CHR	40,6 (9,4)	38,6 (7,6)	43,4 (8,1)	47,2 (9,6)	6,7 ***	3 > 2,1	0,75
SUB	48,8 (7,9)	49,0 (7,8)	52,1 (7,9)	50,7 (8,0)	2,1 (*)	n.s.	0,20
PER	47,6 (11,1)	51,7 (11,1)	44,1 (10,0)	34,1 (11,4)	17,1 ***	4 < 3,2,1 3 < 2 > 1	1,26
SRV	45,6 (8,9)	44,6 (8,7)	47,8 (8,2)	52,2 (10,6)	5,2 **	4 > 2,1	0,74
FAM	43,5 (9,8)	40,4 (7,2)	44,6 (10,0)	54,2 (12,4)	13,4 ***	4 > 3,2,1	1,18
JOB	50,5 (6,6)	48,6 (7,4)	52,9 (5,7)	58,3 (11,2)	13,1 ***	4 > 3,2,1 3 > 2	1,12
DOC	47,1 (11,1)	46,6 (11,2)	50,8 (10,8)	51,9 (13,2)	2,3 (*)	n.s.	0,40
BHIC	28,0 (13,7)	25,2 (11,2)	29,1 (9,2)	46,1 (23,1)	15,5 ***	4 > 3,2,1	1,36

d = Cohen's d: zwischen Klassifikationsgruppe „MPRD“ vs. „Nicht-MPRD“

Die höchsten Werte wiesen MPRD-Patienten entsprechend der Einzel-ANOVAs und der ermittelten Effektstärken nach Cohen in den Basisskalen Depression (d = 2,09), Somatisierung (d = 1,68), Borderline (d = 1,54), Hostility (Feindseligkeit, d = 1,38) und Angst (d = 1,27) auf.

MPRD-Patienten zeigten damit ein auffälliges, **generell erhöhtes psychopathologisches Profil**, vergleichbar mit der von Aguerrevere (2010) [2] gewählten Kennzeichnung von Patienten mit chronischen Rückenschmerzen mit Overreporting im MMPI-2-RF.

In sieben Basisskalen wiesen die Patienten mit möglichem Overreporting (Gruppe 3, P-MPRD) auch höhere Werte auf als Patienten ohne BV-Auffälligkeiten (NoInc, IncOnly).

Eine ausführliche tabellarische Darstellung der Zweigruppenvergleiche zur Ermittlung der Effektstärken nach Cohen findet sich im Anhang (s. Tab. 26, S. 276).

Im Gruppenvergleich zwischen den MPRD-klassifizierten Patienten mit Patienten ohne MPRD-Klassifikation bestätigten sich die auch graphisch deutlichen überhöhten Angaben der MPRD-Patienten in den oben genannten fünf Basisskalen (DEP, SOM, BOR, HOS und ANX) als signifikant, mit entsprechend hohen Effektstärken (Depressions-Skala Cohen's d = 2,09, Somatisierung d = 1,68, Borderline-Symptome d = 1,54 und Feindseligkeit d = 1,38).

Letztere beiden Kennzeichen sind nicht als typisch für andere chronische Schmerzpatienten anzusehen, sondern eher als Hinweis auf die o.g. **generelle Auffälligkeit** zu werten.

Das Antwortprofil der MPRD-Patienten in den BHI-2-Basisskalen kennzeichnete ein generell überhöhtes Profil-Muster, mit erhöht emotionalisierter Konfliktdarstellung dieser Patienten.

3.3 These $H_0$3: Entwicklung eines neuen Validitäts-Index für den MMPI-2-RF

3.3.1 Überprüfung spezifischer „älterer“ Detektionsstrategien (Adaptierte MMPI-2-RF-Skalen)

In einem dritten Analyseschritt sollte die Möglichkeit untersucht werden, einen neu konzipierten, gewichteten Summations-Index aus einer Anzahl von MMPI-2-RF-Subskalen mit möglichst optimaler (hoher) innerer Konsistenz zu entwickeln, um eine optimierte Detektions-Möglichkeit zu ermöglichen. Dieser neue Index sollte sich sowohl aus den Standard-Validitätsskalen des MMPI-2-RF als auch möglicherweise aus an diesen Test zu adaptierenden *validierten, älteren* Skalen des MMPI-2 zusammensetzen.

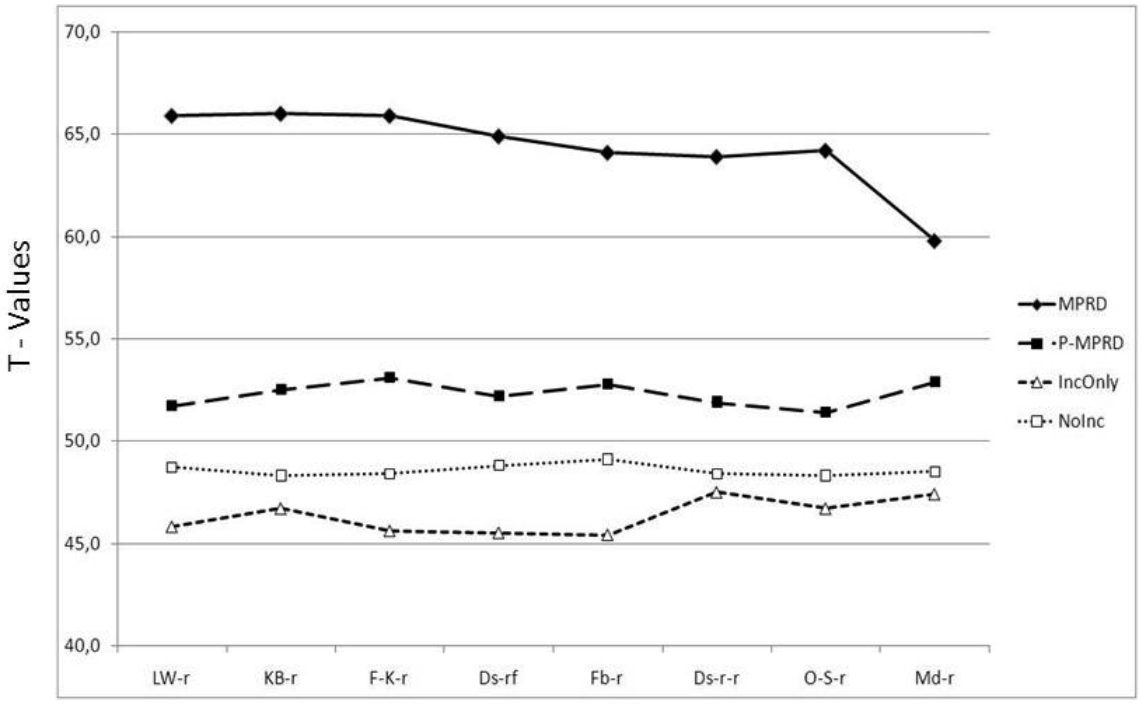
Zu diesem Zweck sollten zunächst einige ältere Skalen des MMPI-2 auf den MMPI-2-RF adaptiert werden; wie im Methodenteil erläutert, wurden die Skalen F-K-r, LW-r, KB-r, Ds-rf, Ds-r-r, Fb-r, O-S-r, Md-r aus für den MMPI-2 selektierten Items ihrer entsprechenden Äquivalenzskalen konzipiert.

Diese Skalen sollten **verschiedene Detektions-Strategien** für Overreporting überprüfen, die nicht durch die Standard-Validitätsskalen des MMPI-2-RF erfasst werden. Dabei handelte es sich um Validitätsskalen zur Erhebung von (1.) Kombinationen von Beschwerdeüberhöhung und Konfliktabwehr (F-minus-K-Konzept), (2.) Inkonsistenzen zwischen offensichtlichen und authentischen, aber subtilen Symptomen (O-S-Skalen), (3.) Undifferenzierten Symptomüberhöhungen (LW- und KB-Skalen), (4.) Simulation Neurose-spezifischer Beschwerden (Ds-Skala), (5.) Simulation Depressions-spezifischer Symptome (Md-Skala) und (6.) Inkonsistenzen zwischen inhaltsähnlichen Symptomen (Fb-Skala).

Zur Vergleichbarkeit der adaptierten Validitätsskalen wurden die Rohwerte dieser Skalen (F-K-r, LW-r, KB-r, Ds-rf, Ds-r-r, Fb-r, O-S-r, Md-r) zunächst nach dem McCall's Area-Transformation-Verfahren für die gesamte Stichprobe in z-Werte, Stanine-Scores und T-Werte transformiert (McCall 1939 [183], s. Lienert 1998 [171], Diehl & Kohr 1989 [67]).

Die ermittelten Skalen-T-Werte wurden den folgenden Berechnungen zugrunde gelegt und sind im Anhang dargestellt (s. Anhang Tab. 28, 278). An gleicher Stelle finden sich auch die Umrechnungen der Rohscores der adaptierten Validitätsskalen in standardisierte Scores für jede einzelne Skala getrennt.

Die multivariate Varianzanalyse zu den acht aus dem MMPI-2 **adaptierten Validitätsskalen** für den MMPI-2-RF zeigte das folgende Resultat:



Adaptierte MMPI-2-RF-Validitätsskalen: LW-r = Kritische Items nach Lachar und Wrobel, KB-r = Kritische Items nach Koss \& Butcher, F-K-r = Dissimulation-Index F-K nach Gough, Ds-rf = Dissimulation Scale nach Gough (1954), Fb-r = Failed-back-Skala, Ds-r-r = verkürzte Ds-Skala Ds-r nach Gough (1957), O-S-r = Obvoius-Subtle-Scales, Md-r = Malingered Depression Scale

Abb. 15. Profile der Studiengruppen in den Adaptierten Validitätsskalen des MMPI-2-RF

Die Angaben der Probanden in den adaptierten Validitätsskalen wiesen **hoch signifikante Unterschiede** zwischen den vier Klassifikationsgruppen für Overreporting auf:

Tab. 15. MANOVA: Adaptierte Validitätsskalen für den MMPI-2-RF in den vier Studiengruppen

	Pillai's D	Exaktes F	D.F.	Fehler-D.F.	p	SN
Adaptierte Skalen	0,972	1560,24	8,00	357,00	0,000	***

Diese Unterschiede in den adaptierten Validitätsskalen sind in Abb. 15 illustriert. Entsprechend der Erwartung spiegeln alle untersuchten Skalen einheitlich und deutlich auffälliges Overreporting zwischen den Klassifikationsgruppen wider - obwohl diese Itemsammlungen aus dem MMPI-2 nur in verkürzter Form im MMPI-2-RF vorhanden sind.

MPRD-Patienten wiesen in allen acht Zusatzskalen für den MMPI-2-RF die am deutlichsten überhöhten Scores auf; Patienten mit möglicher Beschwerdeüberhöhung zeigten die zweithöchsten Auffälligkeiten. Die Angaben dieser Patienten grenzten sich deutlich gegenüber beiden Probanden ohne Auffälligkeiten in den BV-Verfahren (Gruppen NoINC und IncOnly) ab.

Tab. 16. ANOVAs: Adaptierte MMPI-2-RF-Validitätsskalen in den Studiengruppen

	1	2	3	4				
Gruppe	No Incentive	Incentive Only	Probably MPRD	MPRD				
N Skala	231 M (SD)	78 M (SD)	34 M (SD)	25 M (SD)	F-Tests	Scheffé-Tests	α	d
LW-r	48,7 (9,3)	45,8 (9,4)	51,7 (6,8)	65,9 (4,2)	34,5 ***	4 > 3,2,1 3 > 2	0,873	1,96
KB-r	48,3 (9,4)	46,7 (8,3)	52,5 (7,1)	66,0 (5,5)	34,7 ***	4 > 3,2,1 3 > 2	0,886	1,98
F-K-r	48,4 (9,2)	45,6 (8,6)	53,1 (6,5)	65,9 (5,3)	38,2 ***	4 > 3,2,1 3 > 2,1	0,786	1,99
Ds-rf	48,8 (9,2)	45,5 (8,9)	52,2 (8,1)	64,9 (6,2)	32,5 ***	4 > 3,2,1 3 > 2 > 1	0,749	1,84
Fb-r	49,1 (8,9)	45,4 (7,7)	52,8 (7,9)	64,1 (1,2)	32,9 ***	4 > 3,2,1 3 > 2 > 1	0,704	1,79
Ds-r-r	48,4 (9,1)	47,5 (8,3)	51,9 (9,9)	63,9 (6,9)	24,1 ***	4 > 3,2,1	0,485	1,71
O-S-r	48,3 (9,7)	46,7 (9,0)	51,4 (8,5)	64,2 (5,4)	25,1 ***	4 > 3,2,1	0,561	1,72
Md-r	48,5 (9,4)	47,4 (9,1)	52,9 (10,7)	59,8 (8,4)	13,7 ***	4 > 2,1 3 > 2	0,712	1,17

α = Cronbach's α (Innere Konsistenz); d = Cohen's d: zwischen Klassifikationsgruppe „MPRD“ vs. „Nicht-MPRD“

ANOVA-Mittelwerts-Vergleiche (vgl. Tab. 16) über die vier Untersuchungsgruppen hinweg zeigten, dass fünf adaptierte Validitätsskalen (**grau markiert:** LW-r, KB-r, F-K-r, Ds-rf und Fb-r) eine Trennschärfe in vergleichbarer Güte aufwiesen (z.B. LW-r-Skala: $F_{364,3}$ = 34,5, p = 0,000; KB-r-Skala: $F_{364,3}$ = 34,7, p = 0,000).

Diese Gruppenvergleiche bleiben signifikant, selbst bei Berücksichtigung einer Alpha-Adjustierung an die Testanzahl (Signifikanzniveau 0,05/8 = 0,006).

Drei adaptierte Validitätsskalen (O-S, Md-r, Ds-r-r) identifizierten Overreporting mit geringerer Effektivität, wie folgende Post-Hoc-Analysen mittels Scheffé-Tests bestätigten. Die Malingered Depression Scale (Md-r) war am geringsten trennscharf, jedoch ebenfalls zur Identifikation von MPRD-Patienten und Patienten ohne Auffälligkeit in den externen BV-Verfahren geeignet ($F_{364,3} = 13{,}7$, $p = 0{,}000$).

Die ermittelten Effektstärken (Cohen's d, vgl. Tab. 16, S. 177) unterstützen erneut die hohe Trennschärfe von fünf der acht untersuchten Validitäts-Strategien.

Eine ausführliche tabellarische Gegenüberstellung der Zweigruppenvergleiche zur Ermittlung der Effektstärken nach Cohen findet sich im Anhang (s. Tab. 27, S. 277).

Auffällig wurde erneut ein Zusammenhang zwischen der inneren Konsistenz, mit der die jeweilige Validitätsskala (z.B. LW-Skala) das ihr zugeordnete Konstrukt (z.B. „Undifferenzierten Symptomüberhöhung") erfasst, und dem Grad ihrer Trennschärfe hinsichtlich des untersuchten Probanden-Verhaltens („Overreporting").

Dieser Zusammenhang bestätigte sich allerdings in der Md-r-Skala trotz hoher Konsistenz nicht. Mögliche Gründe hierfür sind in der Befunddiskussion zu erörtern (z.B. Trennschärfe des Konstruktes „Simulation depressiver Symptome"; Verkürzung der Skala im MMPI-2-RF; Spezifika der untersuchten Gruppe chronischer Schmerzpatienten).

3.3.2 Konstruktion und Evaluation eines neuen Validitäts-Index *ROI*

Zur Testung der Hypothese $H_0 3$ wurde der experimentelle Validitätsindex ROI aus fünf der adaptierten MMPI-2-RF-Validitätsskalen höchster innerer Konsistenz (LW-r, KB-r, F-K-r, Ds-rf und Fb-r) sowie unter Einbezug der F-r-Skala (als Standard-Skala höchster Trennschärfe) gewichtet zusammengefügt (vgl. Kap. 2.6.6, s. S. 149).

Zu diesem Zweck wurde ein an den Meyers Validity Index MIV (MMPI-2-) bzw. MI-r (MMPI-2-RF) angelehnter Algorithmus angewandt, nach dem mittlere (Stanine-Werte 1-7), erhöhte (Stanine-Werte 8) und maximale Angaben (Stanine-Werte 9) der Probanden in den Einzelskalen gewichtet (mit Werten von 0 bis 2) zu einem Summenwert addiert werden. Dadurch erhält ein Proband, der Maximalangaben in allen sechs Validitätsskalen macht, den Wert 12 und bei Angaben in allen Skalen unterhalb des Staninewertes 8 den Wert 0. Erreicht wird dadurch eine maximal trennscharfe, gewichtete Identifikation von Beschwerdeüberhöhungen.

Entgegen der ersten Test-Teilstichprobe (n = 275) wurden die verteilungsabhängigen Gewichte der Skalen der Verteilungsfunktion der größeren Patienten-Stichprobe (n = 368) noch einmal optimierend angepasst. Daraus resultierte in allen adaptierten Validitätsskalen eine leichte Erhöhung der Stanine-Gewichte (s. Tab. 90 im Anhang).

Die Ergebnisse der univariaten Varianzanalysen zeigten nun Folgendes:

Tab. 17. ANOVA: MMPI-2-RF-Validitätsindex ROI in den vier Studiengruppen

	1	2	3	4			
Gruppe	No Incentive	Incentive Only	Probably MPRD	MPRD			Innere Konsistenz
N	231	78	34	25	F-Tests	Scheffé-	Cronbach's
	M	M	M	M	F(3, 364)	Tests	α
Skala	(SD)	(SD)	(SD)	(SD)			
ROI	0,53	0,22	0,50	5,72	90,4	4 > 3,2,1	0,884
	(1,35)	(0,68)	(0,93)	(3,23)	***		

Mittelwert-Vergleiche mittels der ANOVAs (vgl. Tab. 17) über die vier Untersuchungsgruppen hinweg zeigten, dass der ROI-Index hoch signifikant zwischen den Patientengruppen differenzierte. Diese Differenzierung fällt entsprechend der F-Statistik am deutlichsten gegenüber allen anderen Validitätsskalen aus (ROI: $F_{364,3} = 90{,}4$, $p = 0{,}000$; zum Vergleich F-r-Skala: $F_{364,3} = 60{,}7$, $p = 0{,}000$; MI-r-Index: $F_{364,3} = 65{,}3$, $p = 0{,}000$).

Die innere Konsistenz des ROI-Index fiel vergleichbar hoch aus, wie in den adaptierten Validitätsskalen LW-r und KB-r, und ist als sehr gut zu bezeichnen (Cronbach's $\alpha = 0{,}884$).

Post-Hoc-Analysen mittels Scheffé-Tests bestätigten, dass insbesondere die Patienten mit MPRD-Klassifikation im ROI erheblich höhere Werte aufwiesen als die drei anderen Untersuchungsgruppen mit möglichem Overreporting, mit nur äußeren Anreizen für Overreporting oder ohne solche Auffälligkeiten (ROI: $F_{366,1} = 268{,}3$, $p = 0{,}000$, s. Tab. 18, S. 180).

Probanden mit einem möglichen Overreporting (P-MPRD) und MPRD-Patienten wiesen zudem gegenüber den Patienten mit alleinigem Kompensationsmotiv (IncOnly) und Patienten ohne Auffälligkeiten signifikant höhere Werte im ROI-Index auf ($F_{366,1} = 72{,}7$, $p = 0{,}000$).

Die Effektstärken (Cohen's d) für die beiden möglichen Vergleiche von Probanden mit und ohne auffällige Beschwerdeüberhöhungen zeigten entsprechend der Erwartung sehr deutliche Gruppenunterschiede im ROI-Index.

Tab. 18. Effektstärken der MPRD-Detektion: Standard-MMPI-2-RF-Validitätsskalen

Gruppe	NoInc & IncOnly	Possible & Def. MPRD	F-Test	d	No MPRD	Definite MPRD	F-Test	d
N	309	59			343	25		
Skala	M (SD)	M (SD)			M (SD)	M (SD)		
ROI	0,45 (1,40)	2,71 (3,40)	72,7 ***	1,21	0,46 (1,36)	5,72 (3,23)	268,3 ***	3,39

d = Cohen's d: jeweils zwischen beiden Klassifikationsgruppen

Im ROI-Index erreichten alle MPRD-Patienten mindestens einen Wert von 1. Diesen erreichten nur 60 der authentisch bzw. nur-wahrscheinlich auffällig Antwortenden (Sensitivität 100 %). Ein Cutoff von 12 wurde von keinem „als authentisch" klassifizierten Patienten erreicht, aber von zwei nicht-authentisch Antwortenden (Spezifität 100%). Bei einer Spezifität $\geq$ 90 % wurde ein ROI-Rohwert von 3 ermittelt. Mit diesem Score wurden 96 % der MPRD-Patienten richtig identifiziert, aber auch 6 % der als authentisch eingestuften Patienten falsch klassifiziert.

Die Wahrscheinlichkeit, dass ein Proband bei einem ROI-Cutoff 3 nicht authentisch antwortete, betrug 53 % (Positive Predictive Power). Die Wahrscheinlichkeit, dass ein als unauffällig eingestufter Proband authentisch antwortete (NPP), betrug 99 %. Die ROC-Kurve ergab eine signifikante AUC von 0,969 (s. Abb. 16). Die exakten Einzelscores zur Diskriminanzgüte des MMPI-2-RF-Index ROI sind in Tab. 89 im Anhang dargestellt.

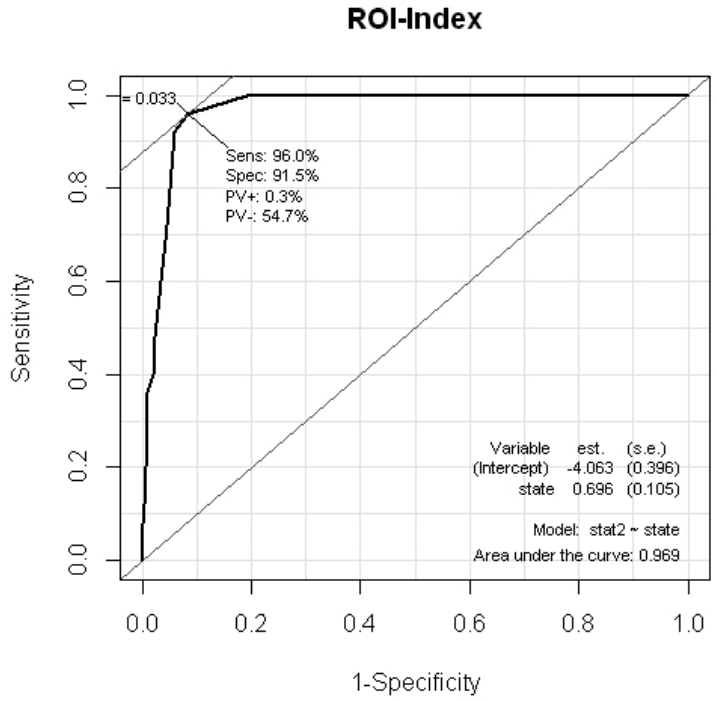


Abb. 16. ROC ROI-Index (MMPI-2-RF)

In der **Untersuchungs-Hypothese** H_0**3** wurde angenommen, dass der neu konzipierte, gewichtete Summations-Index hoch trennscharfer MMPI-2-RF-Validitätsskalen, der sog. ROI-Index, eine höhere Detektions-Akkuranz von Overreporting als die Standard-Validitätsskalen des MMPI-2-RF sowie der Kombinations-Index (MI-r) aufweist.

Zur Testung dieser Annahme wurde die ermittelte AUC-Kurve des ROI jeweils mit den AUC-Kurven der Standard-Validitätsskalen des MMPI-2-RF sowie des MI-r mittels des z-verteilten **AUC-Differenz-Verfahrens** für korrelierte ROC-Kurven **nach DeLong** et al. (1988) [63] (s. Kap. 2.6.7, S. 151) verglichen.

Wie aus Tab. 19 ersichtlich, erwiesen sich alle Validitätsskalen als trennscharf zur Aufdeckung von Overreporting (Area-under-Curve AUC $\geq$ 0,5).

Zur Vermeidung von Zufalls-Signifikanzen wurde für die sechs durchgeführten Kurvenvergleiche eine Bonferoni-α-Adjustierung (mit α = 0,05/6 Mittelwert-Vergleiche = 0,0083) vorgenommen.

Beim Vergleich der Diskriminanzgenauigkeit erwies sich der ROI-Index gegenüber den Standard-Validitätsskalen Fp-r (Infrequent Psychopathology) und Fs (Infrequent Somatic Complaints) als signifikant trennschärfer zur Aufdeckung von Overreporting.

Tab. 19. AUC-Diskriminanzgüte des ROI-Index mit 5 Standard-Validitätsskalen und dem Meyers-Validity-Index MI-r mittels DeLong-Test

ROI	MMPI-2-RF	AUC 1	AUC 2	z-Wert	p	Signifikanz
ROI	F-r	0,969	0,967	0,338	0,735	n.s.
ROI	Fp-r	0,969	0,747	5,030	0,000	**
ROI	Fs	0,969	0,796	3,185	0,001	*
ROI	FBS-r	0,969	0,880	2,275	0,023	n.s.
ROI	RBS	0,969	0,939	1,159	0,247	n.s.
ROI	**MI-r**	0,969	0,954	1,049	0,295	n.s.

Gegenüber der FBS-r-Skala, die unglaubwürdige somatische und kognitive Symptome identifizieren soll, zeigte sich eine tendenziell signifikant höhere Trennschärfe des ROI.

Die Diskriminanzgüte des ROI erwies sich hingegen gegenüber der F-r-Skala (Infrequent Responses), der RBS-r-Skala (Response Bias Scale) und im Vergleich mit der Diskriminanz des MI-r als nicht signifikant höher.

Der ROI-Validitätsindex erwies sich somit im AUC-Vergleich bei der Hälfte der untersuchten Validitätsskalen als der trennschärfste Indikator; hingegen zeigte sich beim Vergleich mit den drei am stärksten diskriminierenden Validitätsskalen keine höhere Akkuranz des ROI. Insgesamt **lies sich** somit die Untersuchungshypothese $H_0 3$ **nicht bestätigen**.

3.4 These $H_0 4$: Detektionsgüte von MMPI-2 und MMPI-2-RF

3.4.1 Akkuranz der untersuchten Detektionsstrategien und Validitätsskalen

Wie im Methodenkapitel erläutert, ist es zur Bestimmung exakter Cutoff-Werte einzelner Skalen zur Beschwerdenvalidierung notwendig, deren Akkuranz (Genauigkeit) anhand ihrer Spezifität (maximal korrekter Ausschluss authentisch antwortender Probanden) und Sensitivität (maximal korrekte Identifikation nicht-authentisch antwortender Probanden) zu bestimmen.

Zu diesem Zweck wurden die Receiver Operating Characteristic-Kurven (ROCs) äquivalenter Skalen bzw. Detektionsmethoden aus MMPI-2 und MMPI-2-RF einander gegenüber gestellt.

Unter Einbezug aller möglichen Werte der jeweils untersuchten Skala kann eine ROC-Kurve aus sens(z) versus $[1 - spec(z)]$ erstellt werden. Diese ROC-Kurve bezeichnet eine Differenzierung, wenn sie über der 45°-Linie zwischen den Endpunkten der ROC-Grafik (0,0) und (1,1) liegt.

Die exakten Cutoff-Analysen der Validitätsskalen sind ab Tab.37 (F-r-Skala, S. 292) bis Tab. 64 (Md-Skala, S. 323) im Anhang dargestellt.

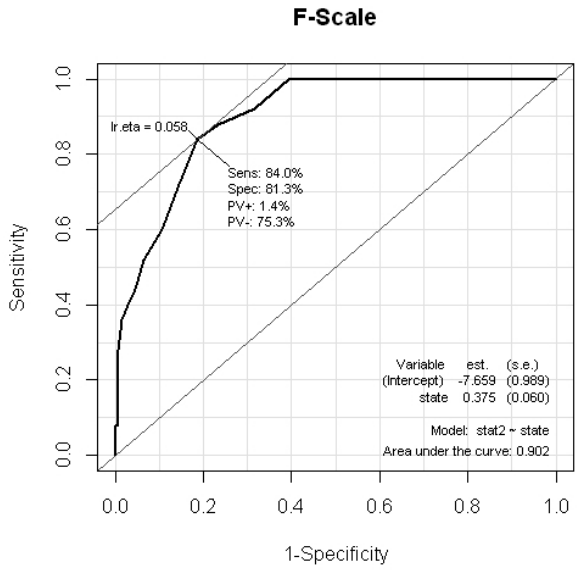


Abb. 17. ROC F-Skala (MMPI-2)

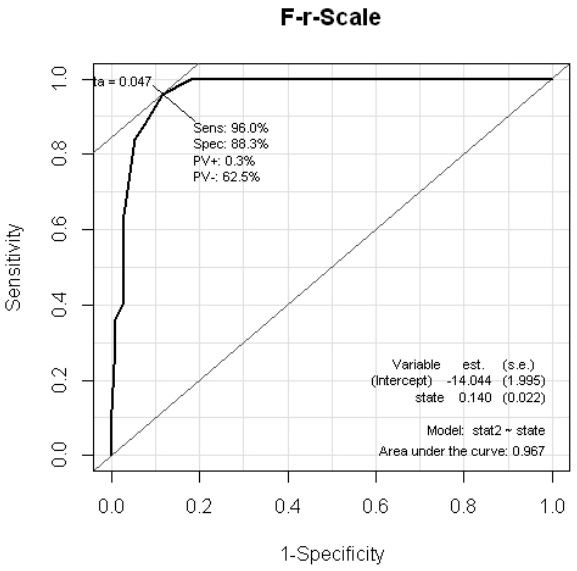


Abb. 18. ROC F-r-Skala (MMPI-2-RF)

3.4.1.a Overreporting quasi-seltener Symptome (F/-r)

Bei einer Spezifität $\geq$ 90 % wurde in der F-Skala ein Rohwert von 16 ermittelt. Hiermit wurden 60 % der nicht-authentisch antwortenden Patienten identifiziert (SN = Sensitivität), aber auch 10 % der als authentisch klassifizierten Patienten (Spezifität SP = 90). Der Grenzwert lag im Bereich anderer Studien (Rogers 2003: Männer: $\geq$ 22, Frauen $\geq$ 20; Thies 2012 [296]: 15). Die ROC-Kurve ergab eine hohe AUC von 0,902 (s. Abb. 17).

In der F-r-Skala erreichten alle MPRD-Patienten mindestens Score 8, den auch 62 (18,1 %) der authentisch bzw. nur-wahrscheinlich auffällig Antwortenden (SN = 100%) angaben. Ein Score von 17 wurde von keinem authentisch Antwortenden erreicht, aber von drei der MPRD-Klassifizierten (SP 100 %). Bei einer Spezifität $\geq$ 90 % wurde für die F-r-Skala ein Wert von 10 (T88) ermittelt, der 88 % der MPRD-Patienten richtig klassifizierte und 92 % der als authentisch identifizierten Patienten erfasste (Positive Predictive Power = 46 %, NPP = 99 %).

Die ROC-Kurve ergab eine **hoch sensitive AUC** von 0,967 (s. Abb. 18). Anderson (2011) [9] beobachtete bei Gutachten-Probanden bei ähnlicher Spezifität (87 %) eine etwas geringere F-r-Sensitivität (59 %) bei höherem Cutoff T100. Meyers et al. 2013 [201] klassifizierten als sicheres Malingering ebenfalls einen F-r-Score von $\geq$ 10.

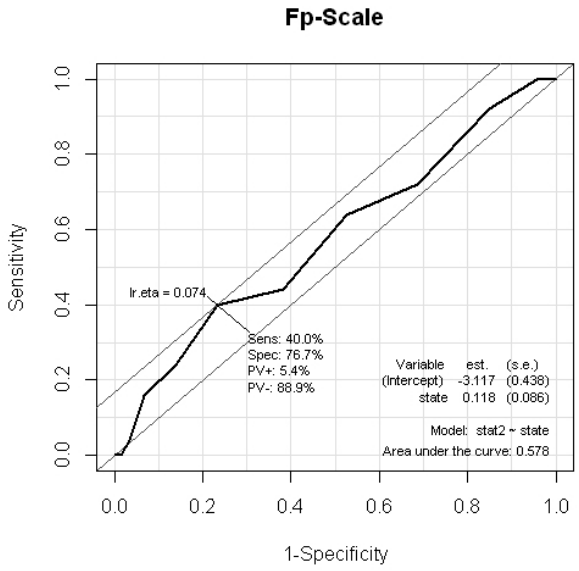


Abb. 19. ROC Fp-Skala (MMPI-2)

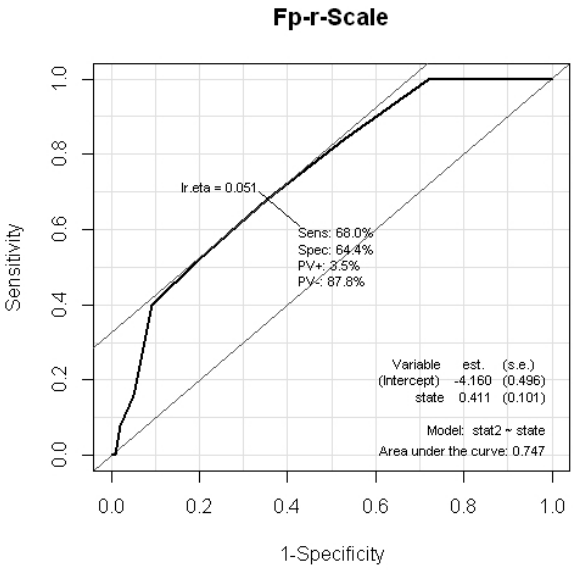


Abb. 20. ROC Fp-r-Skala (MMPI-2-RF)

3.4.1.b Overreporting seltener Symptome (Fp/-r)

Bei einer Spezifität von 93 % wurde in der Fp-Skala ein Rohwert von 8 ermittelt. Mit diesem Score wurden 16 % der nicht-authentisch antwortenden Patienten identifiziert (Sensitivität). Dieser Grenzwert entspricht dem bei Meyers et al. (2002) [202] als sicheres Malingering definierten Grenzwert ab 75 T-Wertpunkten (Thies 2012 [296]: Cutoff 7). Die ROC-Kurve ergab eine AUC von 0,578 (s. Abb. 19).

Bei einer 91 %-Spezifität wurde für die Fp-r-Skala ein Wert von 6 (T94) ermittelt, der 40 % der MPRD-Patienten richtig klassifizierte (Sensitivität).). Die Wahrscheinlichkeit, dass ein Proband bei einem Score 6 richtig klassifiziert wurde, betrug 24 % (PPP). Die Wahrscheinlichkeit, dass ein als unauffällig klassifizierter Patient richtig erkannt wurde, betrug 95 % (NPP).

Die ROC-Kurve ergab eine **mäßigere AUC** von 0,747, die signifikant von 0,5 abwich (s. Abb. 20). Anderson (2011) [9] beschrieb bei Gutachten-Probanden bei 97 % - Spezifität eine geringere Sensitivität (26 %) bei geringerem Cutoff T70 (Score 3). Die Abgrenzung von Malingering gegenüber authentischer Psychopathologie gelang damit bei Anderson (2011) [9] weniger gut. Meyers et al. 2013 [201] klassifizierten als sicheres Malingering ebenfalls einen Fp-r-Score von ≥ 6.

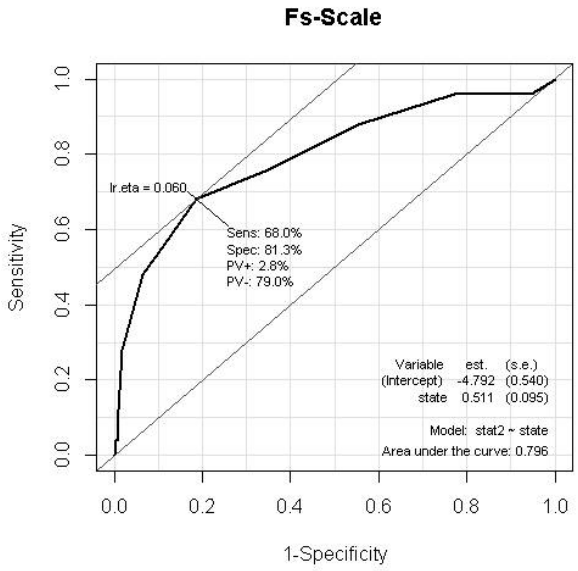


Abb. 21. ROC Fs-Skala (MMPI-2-RF)

3.4.1.c Somatische Beschwerdeüberhöhung (Fs)

In der Fs-Skala, die im MMPI-2 keine äquivalente Entsprechung hat, erreichten alle MPRD-Patienten mindestens einen Wert von 0. Diesen erreichten aber auch alle nicht mit MPRD Klassifizierten (Sensitivität 100 %. Ein Cutoff von 11 wurde von je einem als authentisch und einem MPRD-klassifizierten Probanden erreicht (Spezifität 100 %).

Bei einer 89 % - Spezifität wurde für die Fs-Skala ein Rohwert von 6 ermittelt und bei einer Spezifität $\geq$ 90 % (94 % - Spezifität) ein Cutoff von 7 (T99) erreicht. Mit diesem Score wurden 48 % der MPRD-Patienten korrekt klassifiziert (Sensitivität), aber auch 6 % der als authentisch identifizierten Patienten (PPP = 35 %, NPP = 96%). Um eine PPP von mindestens 50 % zu erreichen, wäre bei dieser Basisrate ein Score von 8 erforderlich, bei einer geringen Sensitivität von 28 %. Die ROC-Kurve ergab eine **mäßigere AUC** von 0,796, die signifikant von 0,5 abwich (s. Abb. 21).

Meyers et al. 2013 [201] klassifizierten als sicheres Malingering ebenfalls einen Fs-Score von $\geq$ 6. Anderson (2011) [9] sah bei Gutachten-Probanden bei 90 % - Spezifität eine ähnliche Sensitivität (56 %) bei geringerem Cutoff T80 (Score 5).

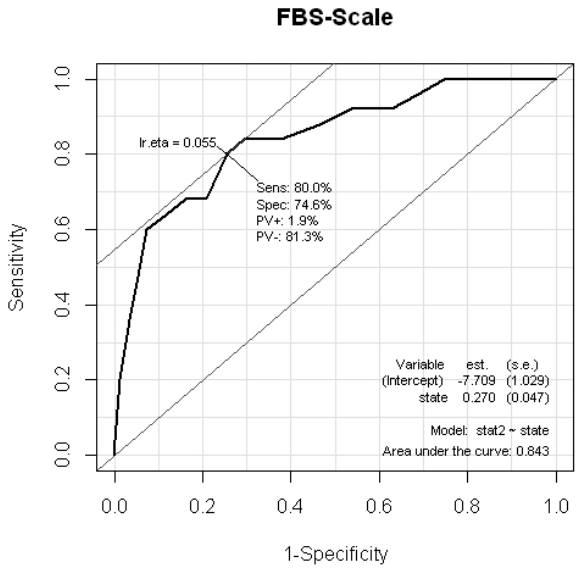


Abb. 22. ROC FBS-Skala (MMPI-2)

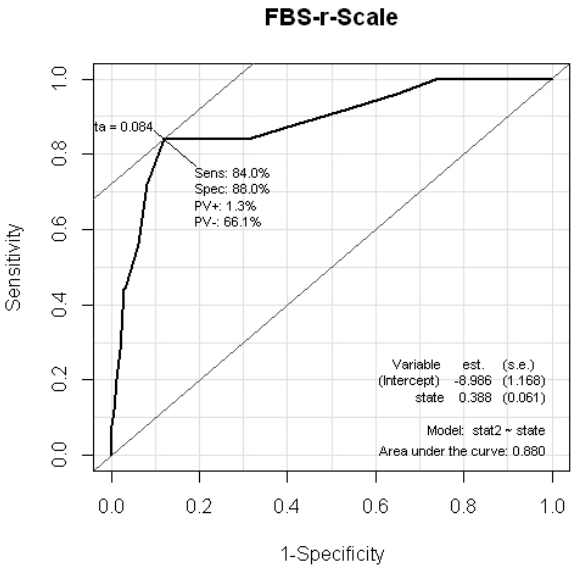


Abb. 23. ROC FBS-r-Skala (MMPI-2-RF)

3.4.1.d Overreporting unfallbezogener Beschwerden (FBS/-r)

Bei einer 93-prozentigen Spezifität wurde in der FBS-Skala ein Wert von 23 ermittelt. Mit diesem Score wurden 56 % der nicht-authentisch Antwortenden identifiziert (Sensitivität), aber auch 4 % der als authentisch Klassifizierten. Dieser Grenzwert lag im Bereich vorangehender Studien; die FBS zeigte nach Lees-Haley et al. (1992) [168] für Männer ab einem Rohwert von 24 und für Frauen ab einem Summenwert von 26 ein Overreporting bei der Reklamation von Unfallschäden (Thies 2012 [296]: Cutoff 27). Die ROC-Kurve ergab eine AUC von 0,843 (s. Abb. 22).

Bei einer 94-prozentigen Spezifität wurde für die FBS-r-Skala ein Wert von 19 (T86) ermittelt, der 56 % der MPRD-Patienten richtig identifizierte. Die Positive Predictive Power für eine richtige Klassifikation betrug 39 % (NPP = 98 %).

Meyers et al. 2013 [201] klassifizierten als sicheres Malingering einen ähnlichen FBS-r-Score von $\geq$ 21. Anderson (2011) [9] sah bei Gutachten-Probanden bei 82 % - Spezifität eine vergleichbare Sensitivität (47 %) bei einem Cutoff ab T90 (Score 21).

Die ROC-Kurve ergab eine **relativ hohe AUC** von 0,860, die signifikant von 0,5 abwich, womit sich auch die FBS-r-Skala als Detektionsstrategie empfiehlt (s. Abb. 23).

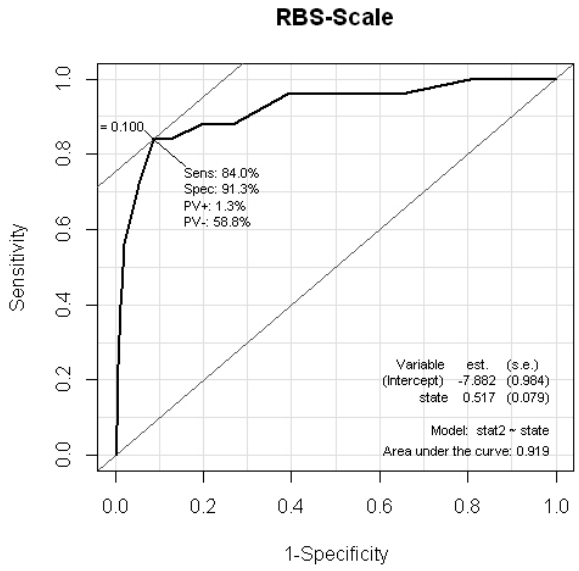


Abb. 24. ROC RBS-Skala (MMPI-2)

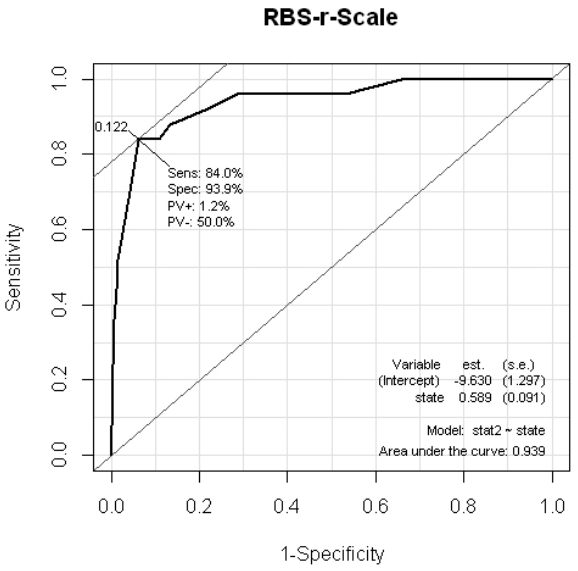


Abb. 25. ROC RBS-Skala (MMPI-2-RF)

3.4.1.e Overreporting kognitiver Symptome (RBS/ HHI/-r)

Bei einer Spezifität $\geq$ 90 % wurde in der RBS-Skala ein Rohwert von 12 ermittelt. Mit diesem Score wurden 84 % der MPRD-Patienten identifiziert (Sensitivität). Dieser Grenzwert wurde bei Tsushima et al. (2011) [300] mit $\leq$ 17 für sicheres Malingering angegeben. Der Mittelwert der Rentenantragsteller lag bei 10,7 gegenüber Nicht-Rentenantragstellern mit einem mittleren Score 6,8; bei Thies (2012) [296] lag dieser Grenzwert bei Score 16. Die ROC-Kurve ergab eine AUC von 0,919 (s. Abb. 24).

Bei einer Spezifität $\geq$ 90 % (94 %) wurde für die RSB-r-Skala ein Rohwert von 14 (T88) ermittelt. Mit diesem Score wurden 84 % der MPRD-Patienten richtig klassifiziert, aber auch 6 % der authentischen Patienten nicht erkannt. Die Wahrscheinlichkeit, dass ein Proband bei Cutoff 14 fragwürdig antwortete, betrug 50 % (Positive Predictive Power), bei einer negativen prädiktiven Power von 99 %.

Meyers et al. 2013 [201] klassifizierten als sicheres Malingering einen gleichen RBS-Score von $\geq$ 15. Anderson (2011) [9] sah bei Gutachten-Probanden bei 93-prozentigen Spezifität eine geringere Sensitivität (44 %) bei höherem Cutoff T100 (Score 17).

Die ROC-Kurve ergab eine **hoch sensitive AUC** von 0,939 (s. Abb. 25).

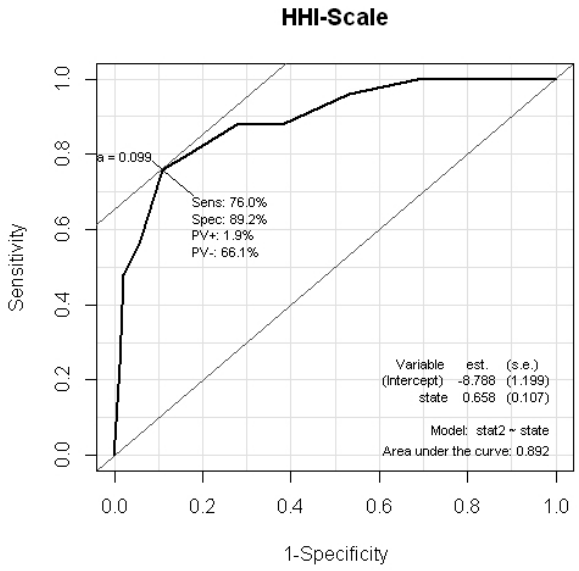


Abb. 26. ROC HHI-Skala (MMPI-2)

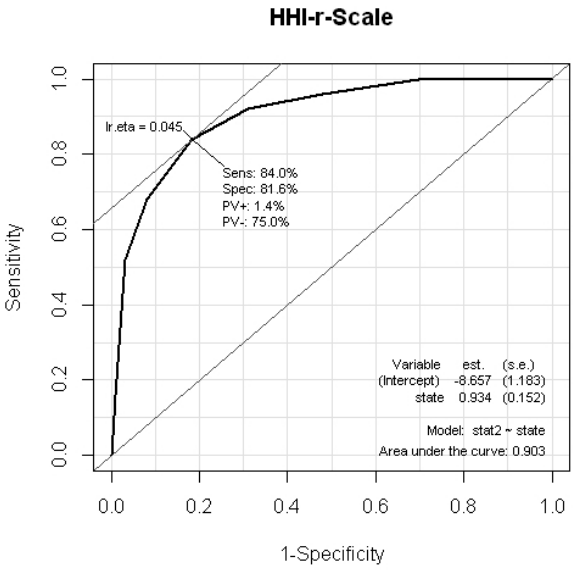


Abb. 27. ROC HHI-r-Skala (MMPI-2-RF)

Kurz-Inventar kognitiver Symptomüberhöhung (HHI/-r)

Bei einer Spezifität $\geq$ 90 % wurde in der HHI-Skala ein Rohwert von 12 ermittelt. Mit diesem Score wurden 56 % der MPRD-Patienten identifiziert (Sensitivität), aber auch 6 % der als authentisch Klassifizierten (SP = 94). Thies (2012) [296] stellte bei Gutachten-Probanden bei demselben Cutoff sicheres Malingering fest. Die ROC-Kurve ergab eine AUC von 0,892 (s. Abb. 26).

Bei einer 90-prozentigen Spezifität wurde für die HHI-r-Skala ein Rohwert von 8 ermittelt. Mit diesem Score wurden 68 % der MPRD-Patienten richtig, aber auch 8 % der als authentisch identifizierten Patienten falsch klassifiziert (PPP = 38 %, NPP = 98 %).

Henry et al. 2012 [131]) ermittelten gleichfalls mittels der im MMPI-2-RF auf 11 Items reduzierten HHI-r bei Cutoff $\geq$ 7 eine gute 69-prozentige Sensitivität zur Aufdeckung von Overreporting bei Klägern mit Kopfverletzungen im Vergleich zu den Patienten ohne Kompensationsansprüche.

Die ROC-Kurve ergab ebenfalls eine hoch sensitive AUC von 0,903 (s. Abb. 27).

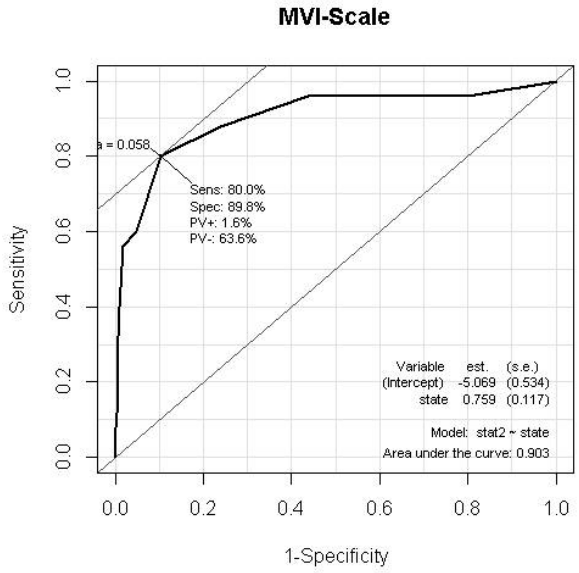


Abb. 28. ROC MVI-Skala (MMPI-2)

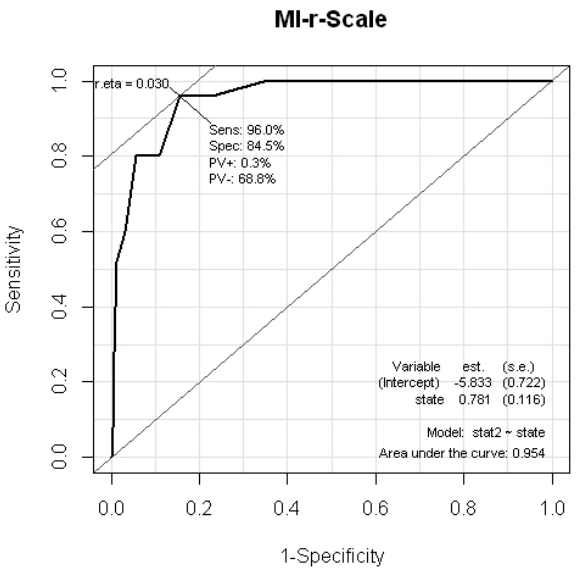


Abb. 29. ROC MI-r-Skala (MMPI-2-RF)

3.4.1.f Kombinierte Validitäts-Indikatoren (MVI / MI-r)

Bei einer 90 % - Spezifität wurde in der MVI-Skala ein Rohwert von 4 ermittelt. Mit diesem Score wurden 80 % der MPRD-Patienten sensitiv identifiziert. Dieser Grenzwert entspricht den Publikationen zur Überprüfung dieses Index (Meyers et al. 2002 [202] Cutoff 5 bei maximal 14 Punkten; Thies 2012 [296] ermittelte bei gleicher Spezifität ab MVI-Cutoff $\geq$ 4 eine Sensitivität von 27,9 %).

Die ROC-Kurve ergab eine ebenfalls **hoch sensitive AUC** von 0,903 (s. Abb. 28).

Bei 90 % - Spezifität wurde für die MI-r-Skala ein Rohwert von 5 ermittelt. Mit diesem Score wurden 80 % der MPRD-Patienten richtig klassifiziert (PPP = 35 %, NPP = 98 %).

Die Skalen-Autoren (Meyers et al. 2013 [201]) berichteten bei einem Cutoff des MI-r $\geq$ 5 bei 145 Schmerzpatienten bei 93 % - Spezifität eine 85-prozentige Sensitivität.

Die ROC-Kurve ergab eine **noch höher sensitive AUC als im MMPI-2** von 0,954 (s. Abb. 29).

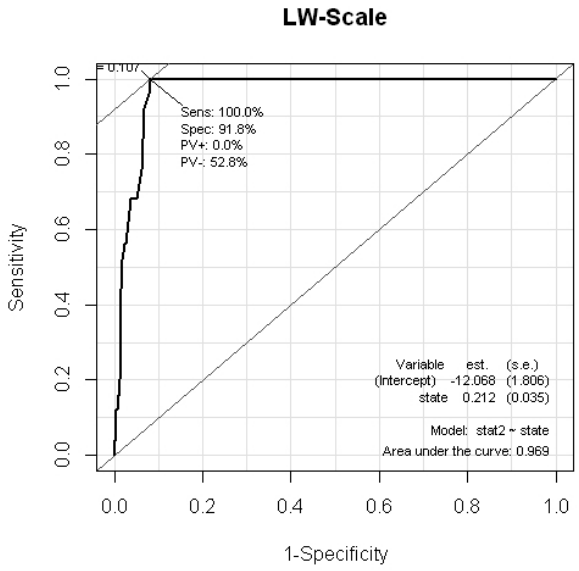


Abb. 30. ROC LW-Skala (MMPI-2)

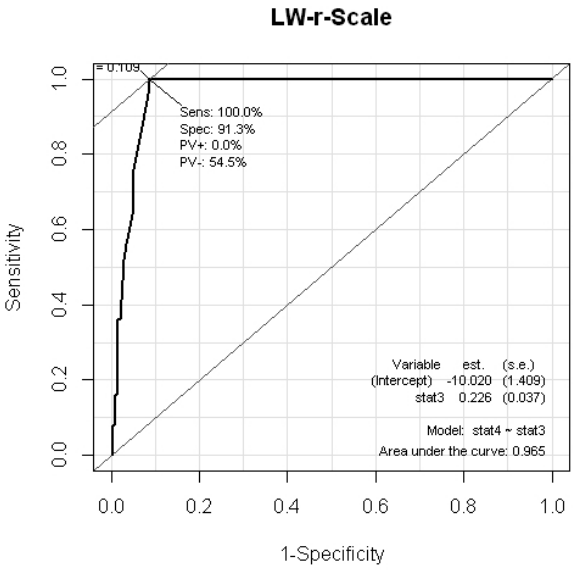


Abb. 31. ROC LW-r-Skala (MMPI-2-RF)

3.4.1.g Undifferenzierte Symptomüberhöhungen (LW/-r, KB/-r)

Bei einer 90 % - Spezifität wurde in der LW-Skala ein Wert von 47 ermittelt. Mit diesem Score wurden alle MPRD-Patienten korrekt identifiziert. Bei höherer Spezifität (95 %, 17 falsch positiv Klassifizierte) verringerte sich die Sensitivität (68%). Thies 2012 [296] berechnete bei einer Spezifität $\geq$ 90 % einen Cutoff $\geq$ 49 mit eher geringer Sensitivität von 34,9 %. Die ROC-Kurve ergab eine entsprechend hohe AUC von 0,969 (s. Abb. 30).

In der LW-r-Skala erreichten alle MPRD-Patienten mindestens einen Wert von 36. Diesen erreichten aber auch 30 der (wahrscheinlich) authentisch Antwortenden (Sensitivität 100 %). Der maximale Score von 61 wurde von keinem der MPRD-Patienten und einem authentisch Antwortenden erzielt.

Bei einer 90 % - Spezifität wurde für die LW-r-Skala ein Rohwert von 36 ermittelt. Mit diesem Score wurden alle MPRD-Patienten richtig klassifiziert, nur 9 % der als authentisch identifizierten Patienten wurde falsch eingeordnet (SN 91%). Bei einer 95%-Spezifität (Score 39) wurden 6 MPRD-Patienten nicht erkannt (SN 76%). Die Positive Predictive Power betrug 53 % bei einer 98 % - NPP.

Die LW-r-Skala erwies sich damit als **hoch detektionssensitiv** (AUC von 0,965, s. Abb. 31).

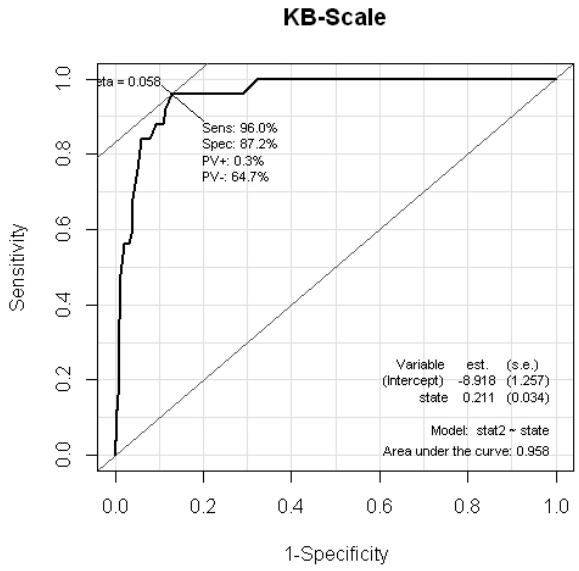


Abb. 32. ROC KB-Skala (MMPI-2)

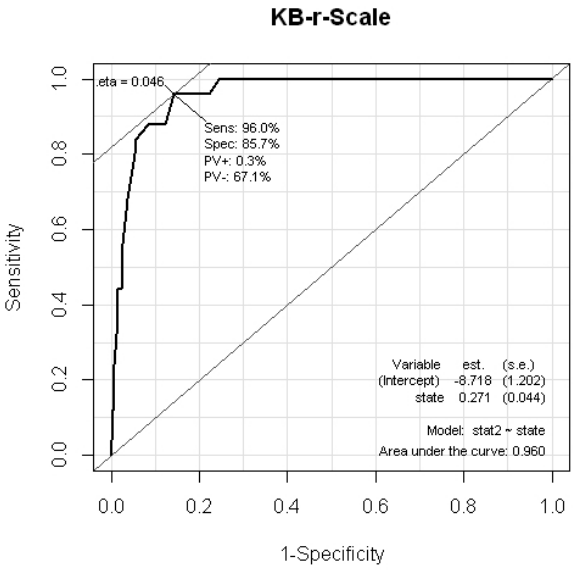


Abb. 33. ROC KB-r-Skala (MMPI-2-RF)

Diskriminanzgüte der Koss-Butcher-/KB-r-Skala

Bei einer Spezifität von $\geq$ 90 % wurde in der KB-Skala ein Rohwert von 33 ermittelt. Mit diesem Score wurden 88 % der MPRD-Patienten identifiziert (n = 22), aber auch 10 % der als authentisch Klassifizierten. Dieser Grenzwert lag im Bereich der vorangehenden Studien. Entsprechende Daten (nach Butcher et al. 1989 [48], s. Green 2008, 174pp. [115]) belegen das 90 %-Intervall bei gesunden Probanden für die KB-Skala ab 31 Rohwertpunkten. Die ROC-Kurve ergab eine AUC von 0,958 (s. Abb. 32).

In der KB-r-Skala erreichten alle MPRD-Patienten mindestens einen Wert von 19. Diesen erreichten aber auch 84 der nicht- bzw. nur-wahrscheinlich auffällig antwortenden Patienten (Sensitivität 100 %). Der Maximalscore von 46 wurde von einem MPRD-Patienten und keinem authentisch Antwortenden erreicht.

Bei einer 90 %-Spezifität wurde für die KB-r-Skala ein Rohwert von 24 ermittelt. Mit diesem Score wurden 88 % der MPRD-Patienten richtig klassifiziert (PPP = 38 % bei NPP 99 %). Die ROC-Kurve ergab die gleiche AUC wie bei der Äquivalenzskala des MMPI-2 (s. Abb. 33).

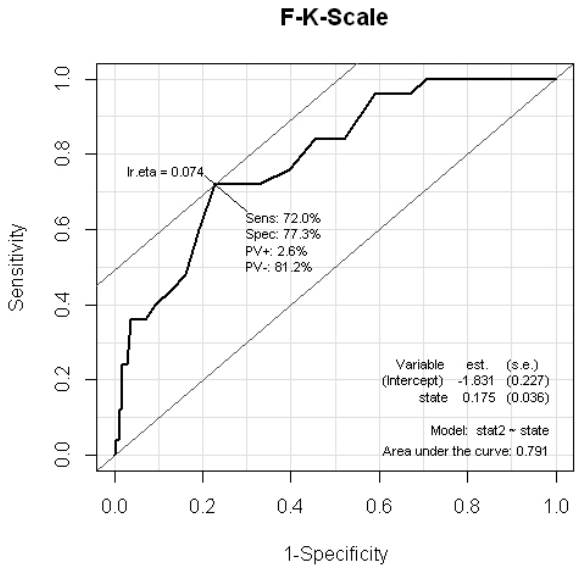


Abb. 34. ROC F-K-Skala (MMPI-2)

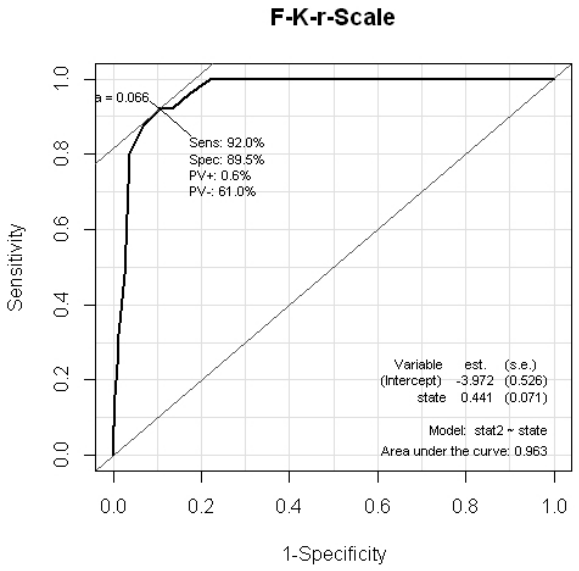


Abb. 35. ROC F-K-r-Skala (MMPI-2-RF)

3.4.1.h Kombinationen von Over- & Underreporting (F-K/-r)

Bei einer 90 % - Spezifität wurde in der F-K-Skala ein Rohwert von 1 ermittelt. Mit diesem Score wurden 40 % der MPRD-Patienten identifiziert (n = 10), aber auch 32 der als authentisch Klassifizierten falsch positiv bewertet. Dieser Grenzwert lag im Bereich anderer Studien; Gough (1950) [105] definierte als Cutoff für Malingering Werte ≥ 1, Thies (2012) [296] ermittelte bei einer gleicher Spezifität einen Cutoff ≥ 5 mit relativ geringer Sensitivität von 25,6 %. Die ROC-Kurve ergab in der hiesigen Studie eine AUC von 0,791 (s. Abb. 34).

In der adaptierten F-K-r-Skala erreichten alle MPRD-Patienten mindestens einen Wert von 1. Diesen erreichten aber auch 75 der übrigen Patienten (Sensitivität 100 %). Der maximale Score von 23 wurde von einem MPRD-Patienten und keinem authentisch Antwortenden erreicht. Bei 90 % - Spezifität wurde ein F-K-Rohwert von 4 ermittelt. Mit diesem Score wurden 92 % der MPRD-Patienten richtig klassifiziert (PPP = 39 % bei NPP 99 %).

Die ROC-Kurve ergab eine **sehr hohe AUC** von 0,963, die höher als im MMPI-2 ausfiel. Vergleichs-Studien zu diesem Score existieren bislang nicht (s. Abb. 35).

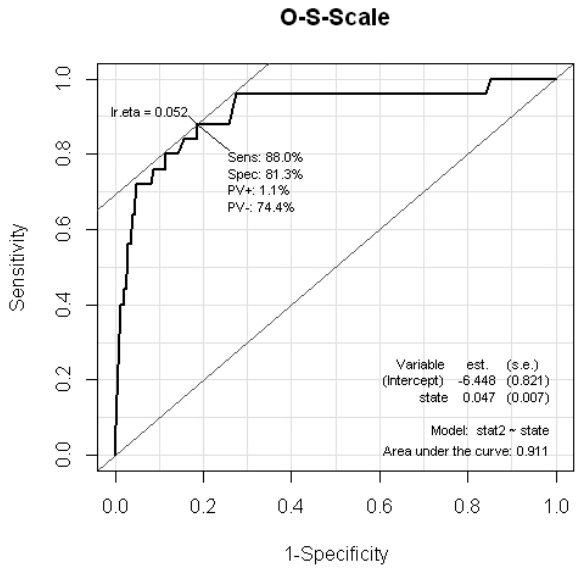


Abb. 36. ROC O-S-Skala (MMPI-2)

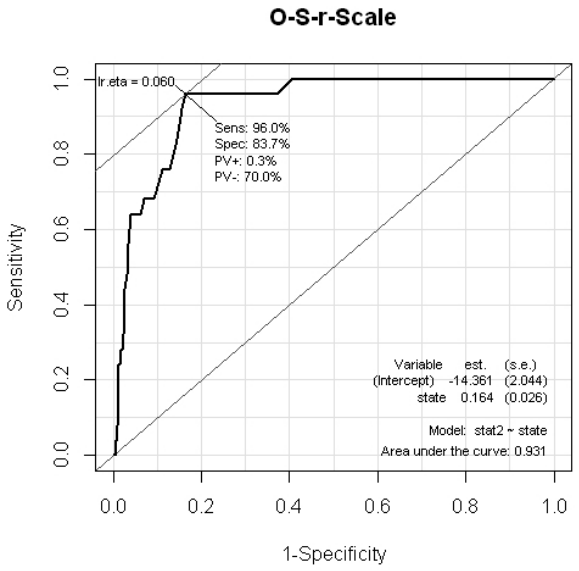


Abb. 37. ROC O-S-r-Skala (MMPI-2-RF)

3.4.1.i Inkonsistenz erkennbarer & subtiler Symptome (O-S/-r)

Bei einer 90 % - Spezifität wurde in der O-S-Skala des MMPI-2 ein Rohwert von 96 ermittelt. Mit diesem Score wurden 76 %, also 19 der 25 MPRD-Patienten identifiziert (Sensitivität). Der Grenzwert wurde bei Meyers et al. (2002) [202] als Teil des MMPI-2-Validity-Index (vgl. Kap. 1.9.4, S. 78) mit einem Cutoff für Malingering bei Schmerzpatienten bei T-Wert-Summenscores der fünf O-S-Skalen über 100 angenommen; Thies (2012) [296] ermittelte bei gleicher Spezifität bei einem Cutoff $\geq$ 104 eine Sensitivität von 32,6 %. Die ROC-Kurve der vorliegenden Studie ergab eine AUC von 0,911 (s. Abb. 36).

In der neu konzipierten O-S-r-Skala erreichten alle MPRD-Patienten mindestens einen Wert von 65. Diesen erreichten aber auch 139 der Probanden ohne MPRD (Sensitivität 100 %). Der maximale Score von 103 wurde nur von einem authentisch antwortenden Patienten erreicht. Bei einer 91-prozentigen Spezifität wurde für die O-S-r-Skala ein Rohwert von 77 ermittelt. Mit diesem Score wurden 68 % der MPRD-Patienten richtig klassifiziert (Positive Predictive Power = 35 %, NPP 97 %).

Die **hohe AUC** (0,931) ergab eine ROC-Kurve signifikant $\geq$ 0,5 (s. Abb. 37).

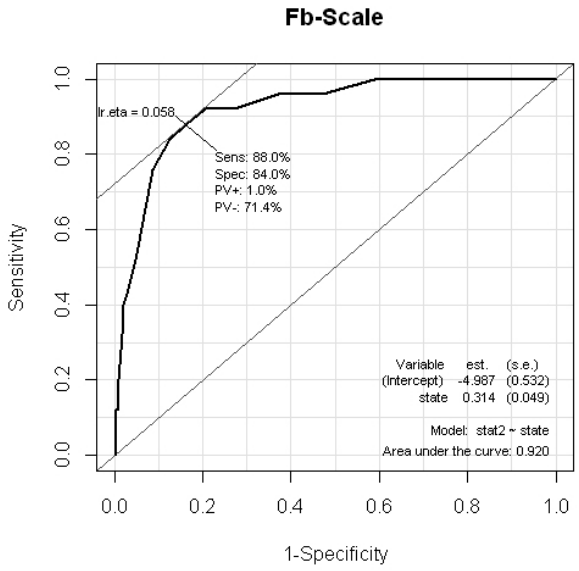


Abb. 38. ROC Fb-Skala (MMPI-2)

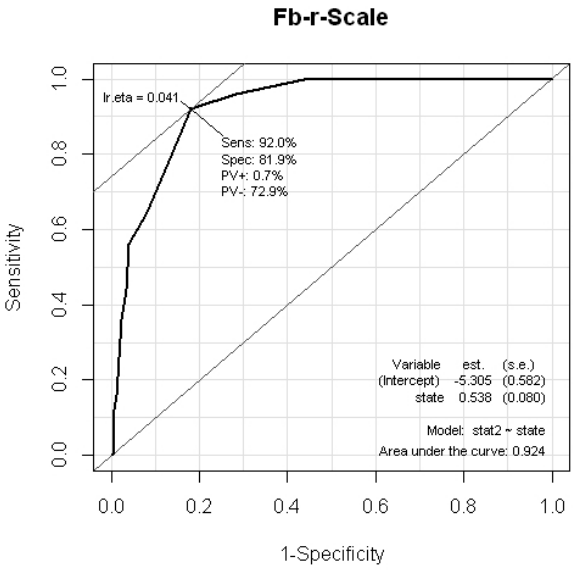


Abb. 39. ROC Fb-r-Skala (MMPI-2-RF)

3.4.1.j Inkonsistenzen inhaltsähnlicher Symptome (Fb/-r)

Bei einer 91-prozentigen Spezifität wurde in der Fb-Skala ein Rohwert von 10 ermittelt. Mit diesem Score wurden 76 % der MPRD-Patienten identifiziert (Sensitivität), aber auch 30 der als authentisch Klassifizierten falsch zugeordnet. Thies (2012) [296] berechnete bei gleicher Spezifität denselben Cutoff ≥ 10 mit einer Sensitivität von 29,1 %. Die ROC-Kurve der vorliegenden Studie ergab eine AUC von 0,920 (s. Abb. 38).

In der für den MMPI-2-RF konzipierten Fb-r-Skala erreichten alle MPRD-Patienten mindestens einen Wert von 3. Diesen erreichten aber auch 150 authentisch antwortenden Patienten (Sensitivität 100 %). Der maximale Score von 13 wurde nur von einem authentisch antwortenden Patienten erreicht. Bei einer 92 % - Spezifität wurde für die Fb-r-Skala ein Rohwert von 7 ermittelt. Mit diesem Score wurden 64 % (n = 16) der MPRD-Patienten richtig zugeordnet, aber auch 8 % der als authentisch identifizierten Patienten falsch klassifiziert (PPP = 37 % bei bei NPP = 97 %).

Die ROC-Kurve ergab eine **hoch signifikante** AUC von 0,924 (s. Abb. 39).

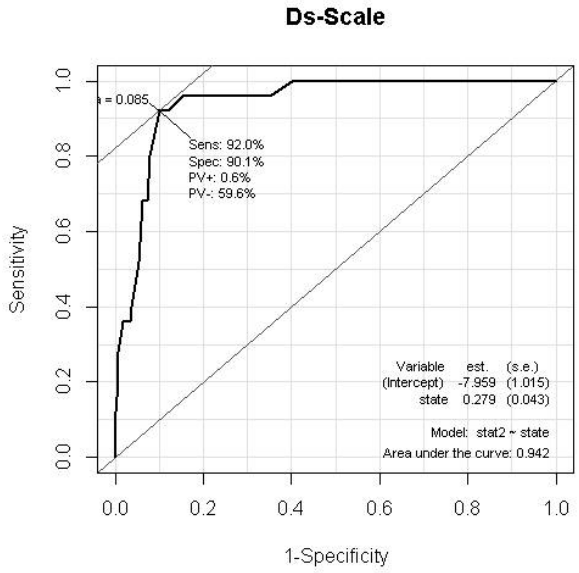


Abb. 40. ROC Ds-Skala (MMPI-2)

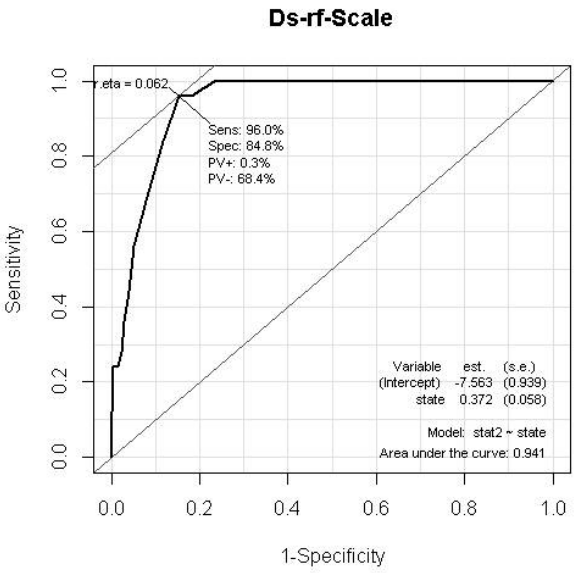


Abb. 41. ROC Ds-rf-Skala (MMPI-2-RF)

3.4.1.k Neurose-spezifisches Overreporting (Ds/-rf)

Bei einer 90 % - Spezifität wurde in der Ds-Skala ein Rohwert von 21 ermittelt. Mit diesem Score wurden 92 % der MPRD-Patienten identifiziert (Sensitivität), aber auch 21 der als authentisch Klassifizierten falsch zugeordnet. Dieser Score lag im Bereich der vorangehenden Studien; Thies 2012 [296] ermittelte bei einer 90 % - Spezifität in der Ds-Skala einen Cutoff $\geq$ 24 mit geringerer Sensitivität (29,1 %). Die ROC-Kurve ergab in der hiesigen Studie eine AUC von 0,942 (s. Abb. 40).

In der Ds-rf-Skala erreichten MPRD-Patienten mindestens einen Wert von 12. Diesen zeigten aber auch 80 der nicht auffälligen Patienten (Sensitivität 100 %). Der Maximalscore von 28 wurde nur von einem MPRD-Patienten erreicht. Bei einer 90 % - Spezifität wurde in der Ds-rf-Skala ein Cutoff-Wert von 16 ermittelt. Mit diesem Score wurden 68 % der MPRD-Patienten korrekt klassifiziert (PPP = 39 % bei NPP 98 %).

Rogers et al. (2011) [239] adaptierten die Dissimulation-Skala als erste Autoren an den MMPI-2-RF. In ihrer Studie wurde bei authentischen Patienten ein Cutoff von 11 Punkten festgestellt, bei Simulanten kognitiver Störungen ein mittlerer Score von 15 sowie bei Simulanten psychischer Symptome ein Score von 20. Die ROC-Kurve ergab eine AUC von 0,941 (s. Abb. 41), die der MMPI-2-Skala Ds entsprach.

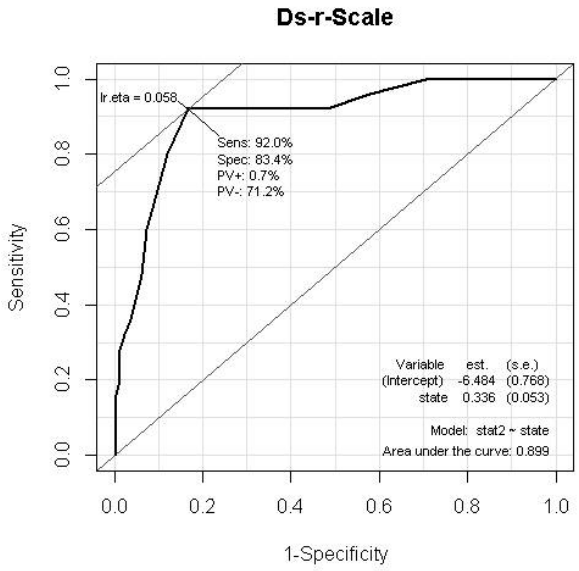


Abb. 42. ROC Ds-r-Skala (MMPI-2)

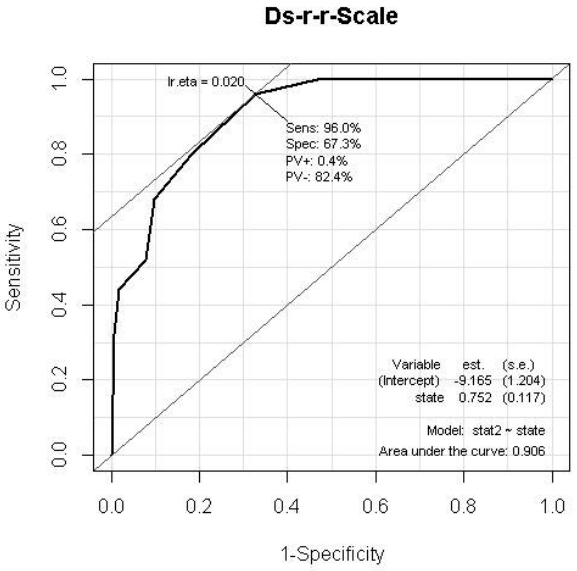


Abb. 43. ROC Ds-r-r-Skala (MMPI-2-RF)

Diskriminanzgüte der Ds-r-/Ds-r-r-Skala

In der im MMPI-2 auf 32 Items verkürzten Ds-r-Skala wurde ein Rohwert von 14 ermittelt. Mit diesem Score wurden 68 % der MPRD-Patienten identifiziert (SN), aber auch 9 % der als authentisch klassifizierten Patienten (91 % - Spezifität). Dieser Grenzwert lag im Bereich der vorangehenden Studien; Thies 2012 [296] ermittelte bei derselben Spezifität für die Ds-r-Skala einen Cutoff $\geq$ 16 mit einer Sensitivität von 29,1 %. Die ROC-Kurve ergab in der vorliegenden Studie eine AUC von 0,899 (s. Abb. 42).

In der Ds-r-r-Skala (im MMPI-2-RF integrierte Items) erreichten alle MPRD-Patienten einen Wert von 7. Diesen zeigten aber auch 160 der authentisch bzw. nur-wahrscheinlich auffällig Antwortenden (Sensitivität 100 %). Der maximale Score in dieser Patientenstichprobe von 13 wurde von acht nicht-authentisch und zwei authentisch antwortenden Patienten erreicht.

Bei einer 90 %-Spezifität wurde für die Ds-r-r-Skala ein Rohwert von 10 ermittelt. Mit diesem Score wurden 68 % der nicht-authentisch antwortenden Patienten richtig klassifiziert. Die Positive Predictive Power betrug bei diesem Cutoff 34 % bei einer NPP 97 %.

Die ROC-Kurve (s. Abb. 43) ergab ebenfalls eine **hoch signifikante** AUC (0,906).

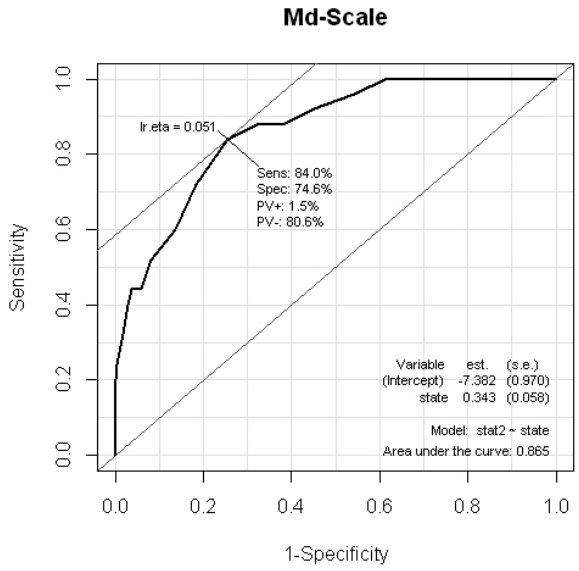


Abb. 44. ROC Md-Skala (MMPI-2)

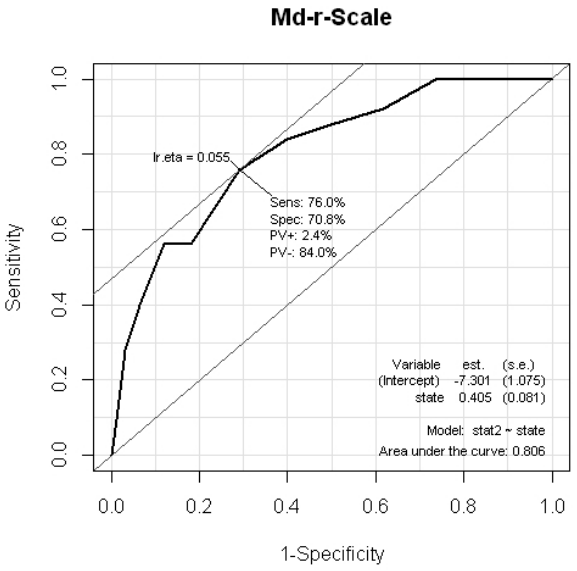


Abb. 45. ROC Md-r-Skala (MMPI-2-RF)

3.4.1.l Depressions-spezifisches Overreporting (Md/-r)

Als Cutoff-Wert der Malingered-Depression-Scale wurde bei 91-prozentigen Spezifität der Wert 14 ermittelt. Mit diesem Score wurden 68 % der MPRD-Patienten identifiziert, aber auch 32 der als authentisch Klassifizierten falsch zugeordnet. Dieser Grenzwert lag niedriger als in vorangehenden Studien; Thies (2012) [296] ermittelte bei einer 90 % - Spezifität in der Md-Skala einen Cutoff $\geq$ 20 mit einer Sensitivität von 30,2 %. Dies entsprach in der vorliegenden Studie einer Spezifität von 99 %. Die ROC-Kurve ergab eine AUC von 0,865 (s. Abb. 44).

In der für den MMPI-2-RF adaptierten Md-r-Skala hatten alle MPRD-Patienten mindestens einen Wert von 8. Diesen erreichten aber auch 252 der (wahrscheinlich) authentisch antwortenden Patienten. Der Maximalscore von 17 wurde von drei nicht-authentisch und fünf authentisch Antwortenden erreicht. Bei einer Spezifität von 94 % wurde für die Md-r-Skala ein Rohwert von 15 ermittelt. Mit diesem Score wurden 40 % der MPRD-Patienten richtig klassifiziert. Die Positive Predictive Power betrug bei diesem Cutoff 31 % bei einer NPP von 96 %.

Die ROC-Kurve ergab eine **mäßig hohe AUC** von 0,806, die signifikant von 0,5 abwich (s. Abb. 45).

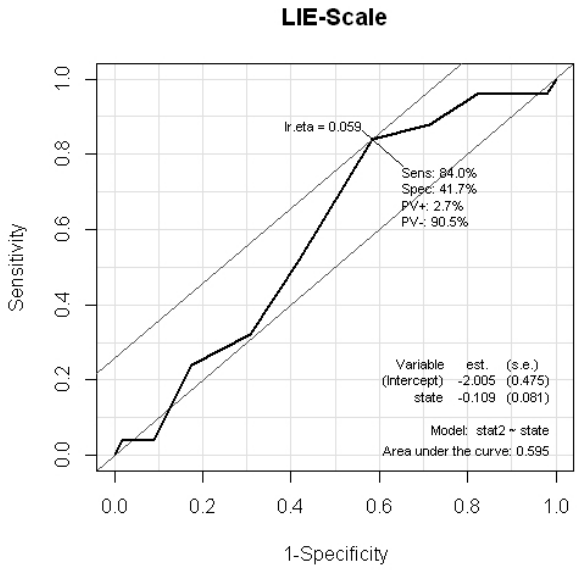


Abb. 46. ROC LIE-Skala (MMPI-2)

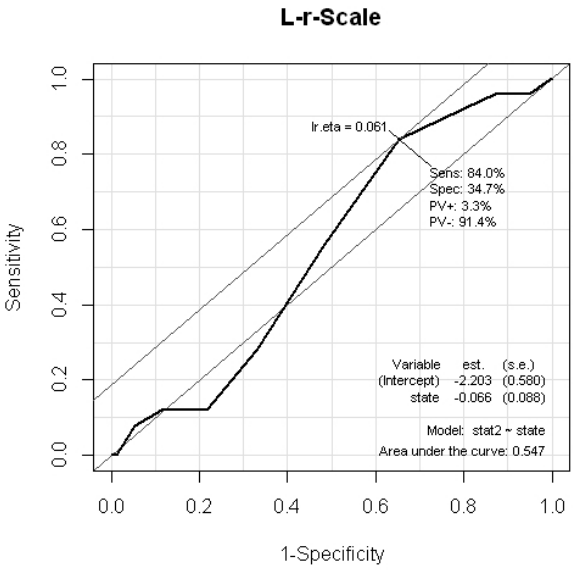


Abb. 47. ROC L-r-Skala (MMPI-2-RF)

3.4.1.m Overreporting im Sinne sozialer Erwünschtheit (K/-r, L/-r)

Sozial-erwünschte Antworten können *bewusst als sog. Impression Management* eingesetzt werden, wie in der L-Skala. In beiden L-Skalen müssen als Instrumente zur Erhebung von Underreporting die Diskriminanz-Scores reziprok errechnet werden.

Bei einer 90 % - Spezifität wurde in der LIE-Skala ein Rohwert von 3 ermittelt. Mit einem Score ≤ 3 wurden 312 der authentisch antwortenden Patienten korrekt identifiziert, aber auch 96 % der als MPRD-Patienten nicht erkannt. Bei einer höheren Sensitivität sank unmittelbar auch die Spezifität des Auswahlverfahrens. Die Detektionsgüte für die Erfassung von Overreporting war somit mäßig. Die ROC-Kurve ergab eine AUC von 0,547, die nur wenig über 0,5 lag (s. Abb. 46).

Auch in früheren MMPI-2-Studien erwies sich die L-Skala als die am wenigsten trennscharfe Subskala (Green 2008, 176pp. [115]). Rogers et al. (2003a) [242] nannten Mittelwerte für authentische Patienten von 5 ($\leq$ T48) und für Simulanten von 7 ($\leq$ T54).

In der L-r-Skala wurde bei einer 89-prozentigen Spezifität ein Rohwert von ≤ 4 ermittelt. Mit diesem Score wurden 304 (87 %) der authentisch antwortenden Patienten identifiziert (PPP = 8 %, NPP = 99 %). Die ROC-Kurve ergab eine **sehr geringe AUC** von 0,547 (s. Abb. 47).

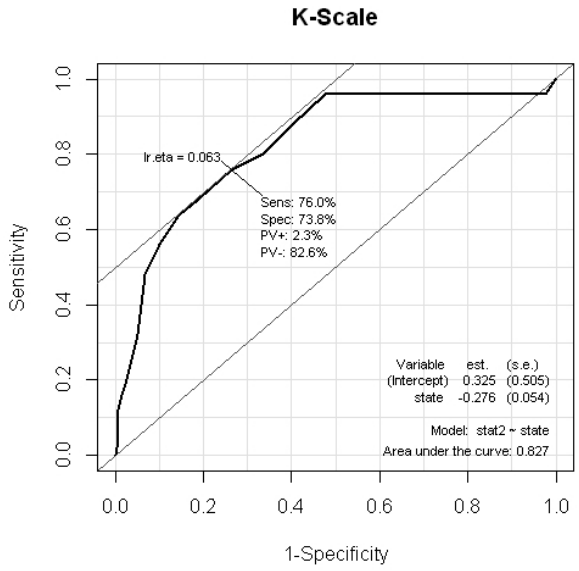


Abb. 48. ROC K-Skala (MMPI-2)

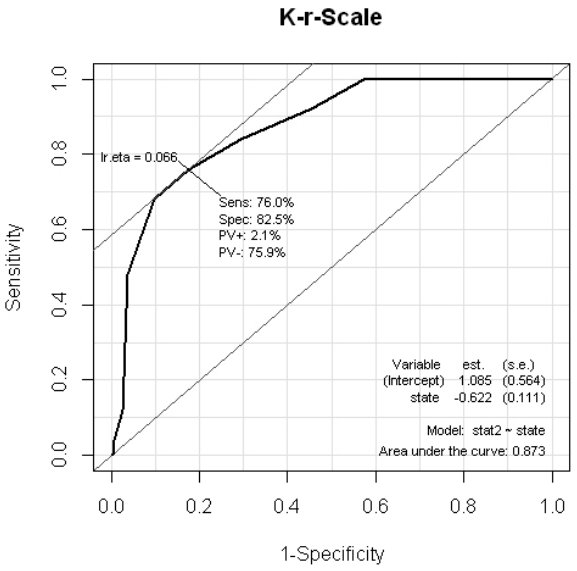


Abb. 49. ROC K-r-Skala (MMPI-2-RF)

Diskriminanzgüte der Korrektur-Skalen K-/K-r

Sozial-erwünschte Antworten können auch als *eher unbewusste Selbsttäuschungen* fehlender Pathologie mit der K-(Korrektur)Skala identifiziert werden.

Bei einer 90-prozentigen Spezifität wurde in der K-Skala ein Rohwert von 8 ermittelt. Mit diesem Score wurden 56 % der MPRD-Patienten identifiziert (Sensitivität). Der Grenzwert lag niedriger als in vorangehenden Studien; so ermittelte Thies 2012 [296] für die K-Skala in der Malingering-Gruppe einen Wert von 14,4. Rogers et al. (2003a) [242] nannten Mittelwerte für authentische Patienten von 15 ($\leq$ T48) und für Simulanten von 10 ($\leq$ T38). Die ROC-Kurve ergab eine AUC von 0,827 (s. Abb. 48).

In der K-r-Skala erreichten alle MPRD-Patienten mindestens den Wert 3. Diesen zeigten aber auch 22 von 25 MPRD-Patienten (Sensitivität 100 %). Den Maximalscore von 13 erreichten nur 6 authentische Patienten. Bei einer 90 % - Spezifität wurde ein Rohwert von 5 ermittelt. Mit diesem Score wurden 32 % der MPRD-Patienten richtig klassifiziert (PPP = 35 % bei 99 % - NPP). Die ROC-Kurve ergab eine **relativ hohe AUC** von 0,873 (s. Abb. 49).

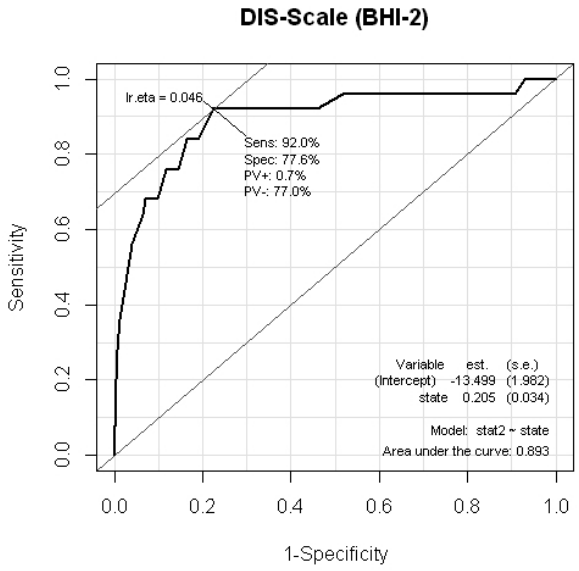


Abb. 50. ROC DIS-Skala (BHI-2)

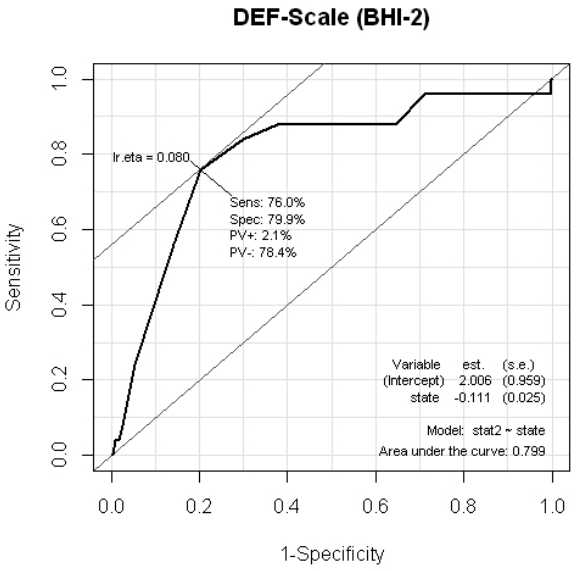


Abb. 51. ROC DEF-Skala (BHI-2)

3.4.1.n Schmerz-spezifisches Over- & Underreporting (BHI-2)

In der DIS-Skala erreichten alle MPRD-Patienten einen Wert von 31, den auch 319 (93 %) der authentischen Patienten zeigten. Den Maximalscore 74 erreichte nur ein MPRD-Patient. Bei einer 90 % - Spezifität ergab sich ein DIS-Wert von $\geq$ 58, mit dem 68 % der MPRD-Patienten richtig klassifiziert wurden (PPP = 34 %, NPP = 97 %). Die ROC-Kurve von 0,893 war **relativ hoch** (s. Abb. 50).

Fast äquivalent beschrieben Bruns & Disorbio (2004) [43] einen Mittelwert von Beschwerden simulierender Probanden von 59 und einen Mittelwert von 50 bei authentischen Patienten.

Für die Underreporting-Skala DEF wurde die Akkuranz reziprok berechnet. Die MPRD-Patienten zeigten mindestens einen DEF-Wert von 17. Der Maximalscore 74 wurde nur bei einem authentischen Patienten gesehen. Bei einer 92 % - Spezifität ergab sich bei einer **relativ hohen AUC** (0,799) ein Cutoff $\leq$ 34 (SN 68 %, s. Abb. 51).

Bruns & Disorbio (2004) [43] sahen bei ihrer fake-bad-Gruppe einen ähnlichen Mittelwert von 38 (zum Vergleich: Score 49 im Mittel bei authentischen Patienten).

3.4.2 Diskriminanz äquivalenter Validitätsskalen & -indizes

In der Untersuchungs-Hypothese $H_0 4$ wurde angenommen, dass die Area-under-the-Curves der jeweiligen Skalen des MMPI-2-RF gleich oder größer als die AUC-Kurven ihrer Äquivalenzskalen seien.

Zur Testung dieser Annahme wurden die ermittelten AUC-Kurven jeweils zwei äquivalenter Validitätsskalen mittels des z-verteilten **AUC-Differenz-Verfahrens** für korrelierte ROC-Kurven **nach DeLong** et al. (1988) [63] (s. Kap. 2.6.7, S. 151) durchgeführt.

Wie aus Tab. 20 (S. 202) ersichtlich, erwiesen sich alle Validitätsskalen als trennscharf zur Aufdeckung von Overreporting (Area-under-Curve AUC $\geq$ 0,5).

Die MMPI-2-RF-Validitätsskalen **F-r** und **Fp-r** wiesen eine **höhere Trennschärfe** gegenüber ihren beiden Äquivalenzskalen des MMPI-2 auf.

Bei einer Bonferoni-α-Adjustierung zur Vermeidung von Zufallssignifikanzen (mit α = 0,05/18 Mittelwert-Vergleiche = 0,003) erwies sich jedoch nur die Neu-Konstruktion der Fp-r-Skala als statistisch trennschärfer gegenüber ihrer Vorgängerversion.

Die FBS-r-Skala des MMPI-2-RF zeigte leicht höhere Trennschärfe-Eigenschaften als ihre Vorgänger-Version des MMPI-2. Auch der MI-r wies eine leicht höhere AUC-Diskriminanz auf als der MVI für den MMPI-2 im DeLong-Test.

Unter den adaptierten Validitätsskalen zum MMPI-2-RF erwies sich der **F-K-r-Index für den MMPI-2-RF** - auch bei einer α-Adjustierung - gegenüber seiner MMPI-2-Fassung als deutlich trennschärfer, was vermutlich vorwiegend auf die Verbesserung der F-Skala zurück zu führen ist.

Die neuerliche Verkürzung der Md-r-Skala gegenüber ihrer MMPI-2-Version (20 vs. 32 Items) führte zu einer signifikanten Verschlechterung der Detektionsgüte.

Wie aus den vorangehenden Analysen erwartbar, erwies sich die BHI-2-Subskala Disclosure gegenüber der Skala Defensiveness als tendenziell trennschärfer. Entgegen Publikationen, in denen die Validitätsskalen des BHI-2 gegenüber denen des MMPI-2 als weniger effektiv eingeschätzt wurden (vgl. Green 2008 [115]), zeigten die BHI-2-Skalen vergleichbare Detektions-Qualitäten.

Mit dem in der vorliegenden Studie **neu konzipierten Validitätsindex ROI** lies sich die höchste Detektionsgüte aller untersuchten Skalen erreichen.

Tab. 20. AUC-Diskriminanzgüte äquivalenter Validitätsskalen (DeLong-Test)

MMPI-2-RF	MMPI-2	AUC 1	AUC 2	z-Wert	p	Signifikanz
F-r	F	0,967	0,902	2,876	0,004	***
Fp-r	Fp	0,747	0,578	3,413	0,000	***
FBS-r	FBS	0,880	0,843	1,907	0,049	*
RBS-r	RBS	0,939	0,919	1,469	0,142	n.s.
L-r	Lie	0,547	0,595	1,109	0,268	n.s.
K-r	K	0,873	0,827	1,054	0,292	n.s.
HHI-r	HHI	0,903	0,892	0,502	0,616	n.s.
MI-r	MVI	0,954	0,903	1,783	0,075	(*)
LW-r	LW	0,965	0,969	1,099	0,272	n.s.
KB-r	KB	0,960	0,958	0,388	0,698	n.s.
F-K-r	F-K	0,963	0,791	3,886	0,000	***
Ds-rf	Ds	0,941	0,942	0,115	0,908	n.s.
Fb-r	Fb	0,924	0,920	0,166	0,868	n.s.
Ds-r-r	Ds-r	0,906	0,899	0,198	0,843	n.s.
O-S-r	O-S	0,931	0,911	0,570	0,568	n.s.
Md-r	Md	0,806	0,865	3,100	0,002	***
BHI-DIS	BHI-DEF	0,893	0,799	2,668	0,008	***

Der ROI-Score zeigte zudem eine ebenso hohe Detektionsgüte wie der gewichtete Meyers Validity Index (MI-r) der fünf Standard-Validitätsskalen des MMPI-2-RF.

Eine höhere Detektionsgüte der MMPI-2-RF-Validitätsskalen gegenüber ihren Äquivalenzskalen des MMPI-2 konnte jedoch nur in zwei der untersuchten 18 Skalen mittels des DeLong-Tests festgestellt werden - trotz der vermuteten Vorteile reduzierter Item-Überlappungen sowie optimierter Skalenkonstruktion und neuer Item-Selektion.

Die Hypothese $H_0 4$, nach der die Validitätsskalen des verkürzten MMPI-2-RF ebenso valide und trennscharf in der Aufdeckung von Overreporting wie die traditionellen MMPI-2-Validitätsskalen sein sollten, bestätigte sich somit. Insbesondere die Validitätsskalen F-r, MI-r sowie die neu konzipierten Validitätsskalen K-F-r und ROI-Index wiesen ausgezeichnete Diskriminations-Eigenschaften von Overreporting auf.

3.5 These $H_0$5: Einfluss von Overreporting auf den Therapie-Erfolg

In der fünften Studien-Hypothese sollte der Einfluss von Overreporting auf den in einem multimodalen Schmerztherapie-Programm erreichbaren Erfolg eingeschätzt werden.

Da das Haupt-Ziel jeder Schmerztherapie (neben diversen anderen Erfolgsparametern, z.B. Rückkehr an den Arbeitsplatz, Lebensqualitätsverbesserungen, Auftretenshäufigkeit und -dauer der Beschwerden etc.) die Reduktion dieser Beschwerden ist, wurde die Schmerzstärke vor und nach der Behandlung untersucht. Der Therapieerfolg wurde mittels Visueller Analogskala als durchschnittliche Schmerzintensität zwischen 0 (= „Schmerzfreiheit“) und 10 (= sog. „Ohnmachtsschmerz“) vor der Behandlung erhoben. Dieselbe Erhebung erfolgte als mittlere Einschätzung der Patienten an den letzten drei Behandlungstagen.

Die Ergebnisse der Erhebungen sind in den Abbildungen 52 und 53 illustriert.

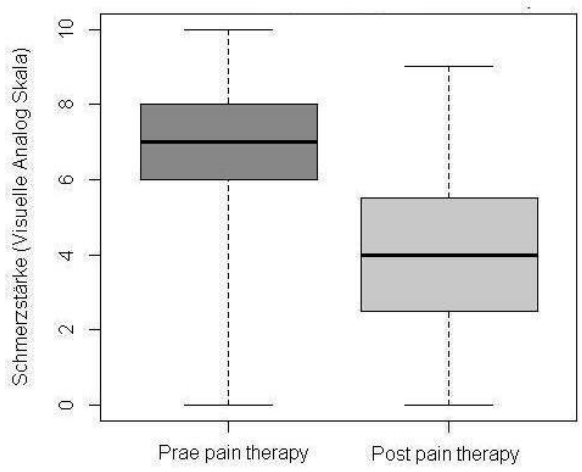


Abb. 52. Schmerzreduktion bei 325 Pbn. ohne Overreporting (ROI $\leq$ 2)

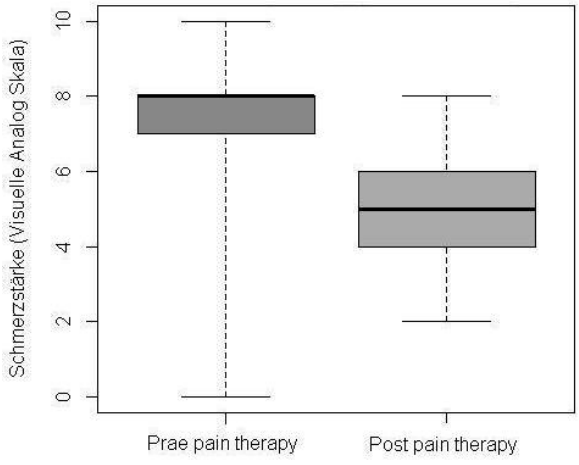


Abb. 53. Schmerzreduktion bei 43 Pbn. mit Overreporting (ROI $\geq$ 3)

Ein varianzanalytischer Vergleich zeigte bereits vor der Behandlung eine vergleichbare Schmerzbelastung in allen vier Untersuchungsgruppen ($F_{344,3}$ = 1,43, p = 0,235; s. Tab. 21, S. 204), wenngleich von den MPRD-Patienten bereits zu Behandlungsbeginn die höchste durchschnittliche Schmerzbelastung (VAS 7,7) angegeben wurde.

Auch wurde in allen vier Klassifikationsgruppen während der Schmerztherapie eine bedeutsame Schmerzlinderung erreicht (F = 370,8; p $\leq$ 0,000; s. Abb. 52 und 53). Diese Schmerzreduktion betrug bei den Patienten mit sicherem Overreporting (MPRD) ca. 30 %. Bei den Patienten ohne Hinweis auf Beschwerdeüberhöhung lag sie bei 44 % vom Ausgangswert.

Entsprechend der vergleichenden Varianzanalyse waren diese Gruppenunterschiede hoch signifikant ($F_{344,3}$ = 4,82, p = 0,003; s. Tab. 21).

Tab. 21. Schmerzreduktion nach multimodaler Schmerztherapie in den vier Klassifikationsgruppen

Gruppe	1 No Incentive	2 Incentive Only	3 Probably MPRD	4 MPRD			
N	231	78	34	25	F-Test	Scheffé-Test	ANOVA innerhalb
Skala	M (SD)	M (SD)	M (SD)	M (SD)			
VAS prae	7,0 (1,7)	7,3 (1,5)	7,4 (1,9)	7,7 (1,3)	1,4 n.s.	n.s.	370,8 0,000
VAS post	3,9 (1,8)	4,4 (1,8)	4,6 (1,9)	5,2 (1,6)	4,8 0,003	4 > 1	

Vergleicht man die 25 als definitiv auffällig klassifizierten MPRD-Patienten mit den Patienten ohne sichere Auffälligkeiten (s. Tab. 22), so machten die MPRD-Patienten auch statistisch signifikant tendenziell vor der Behandlung etwas höhere Schmerzangaben (VAS 7,7) als die Patienten ohne Auffälligkeiten (VAS 7,2; $F_{345,1}$ = 7,9, p = 0,09).

Tab. 22. Einfluss von definitivem Overreporting auf die Schmerzreduktion nach multimodaler Schmerztherapie

	No MPRD	Definite MPRD		
	343	25		
	M (SD)	M (SD)	F-Test Zwischen	ANOVA innerhalb
VAS prae	7,2 (1,6)	7,7 (1,3)	7,9 0,09	10,7 0,00
VAS post	4,0 (1,8)	4,8 (1,8)	9,0 0,00	

Bei beiden Patientengruppen lies sich die Schmerzintensität um durchschnittlich drei Punkte auf der VAS-Skala reduzieren (ANOVA mit Messwiederholung: $F_{338,1}$ = 10,7, p $\leq$ 0,001). Die angegebene **Schmerzreduktion der MPRD-Patienten** fiel jedoch **signifikant geringer** aus als das Therapieergebnis der Patienten ohne sicheres Overreporting ($F_{345,1}$ = 9,0, p $\leq$ 0,001; s. Tab. 22).

Dasselbe Ergebnis zeigte sich bei Vergleich der Probanden mit wahrscheinlichem und sicherem Overreporting (MPRD / P-MPRD) und andererseits Patienten ohne Auffälligkeiten oder mit alleiniger äußerer Motivation für Overreporting (NoInc, IncOnly).

Hier wiesen die 59 Patienten der auffälligen Gruppen bereits vor der Behandlung eine tendenziell höhere Schmerzwertung auf als die Patienten ohne Auffälligkeiten ($F_{345,1}$ = 2,9, p = 0,09). Beide Patientengruppen gaben nach der Therapie eine um 2-3 VAS-Punkte geringere Schmerzbelastung an und verbesserten sich durch die Behandlung signifikant (ANOVA mit Messwiederholung: $F_{338,1}$ = 10,0, p $\leq$ 0,000).

Jedoch gaben Patienten mit auffälligem Overreporting (MPRD / P-MPRD) nach der Therapie eine signifikant geringere Schmerzreduktion an als Patienten ohne BV-Auffälligkeiten ($F_{345,1}$ = 7,9, p $\leq$ 0,001; s. Tab. 23).

Tab. 23. Einfluss von möglichem und definitivem Overreporting auf die Schmerzreduktion nach multimodaler Schmerztherapie

Gruppe	No Incentive & IncOnly	Probably & Definite MPRD		
N	309	59		
Skala	M (SD)	M (SD)	F-Test Zwischen	ANOVA innerhalb
VAS prae	7,1	7,5	2,9	10,0
	(1,6)	(1,6)	0,09	0,00
VAS post	4,1	5,2	7,9	
	(1,8)	(1,6)	0,01	

Wurde die Vorhersage-Güte des **neu konzipierten ROI-Index** für den MMPI-2-RF (Auffälligkeit ab ROI $\geq$ 3) für den Therapieerfolg untersucht, zeigte sich in der Gruppe mit auffälligem Overreporting (n = 34) vor der Therapie eine tendenziell erhöhte Schmerzangabe, jedoch war dieser Gruppenunterschied nicht signifikant ($F_{345,1}$ = 2,2, p = 0,13).

Beide Patientengruppe zeigten therapeutische Verbesserungen (ANOVA Messwiederholung: $F_{338,1}$ = 7,7, p $\leq$ 0,01) mit einer Schmerzreduktion um drei VAS-Punkte.

Patienten mit auffälligem Overreporting (ROI $\geq$ 3) zeigten nach der Therapie eine geringere Schmerzreduktion als die Patienten ohne Auffälligkeiten ($F_{345,1}$ = 6,7, p $\leq$ 0,001, s. Tab. 24, S. 206).

Tab. 24. Prognose des Therapieerfolges nach multimodaler Schmerztherapie durch Overreporting-Klassifikation mittels ROI-Index (MMPI-2-RF)

	ROI < 3	ROI $\geq$ 3		
	325	43		
	M	M	F-Test	
	(SD)	(SD)	Zwischen	innerhalb
VAS prae	7,1	7,5	2,2	7,7
	(1,7)	(1,4)	0,13	0,01
VAS post	4,1	4,8	6,7	
	(1,9)	(1,5)	0,01	

Die Hypothese $H_0 5$, nach der Patienten, bei denen durch den Validitätsindex ROI eindeutig Beschwerdeüberhöhungen identifiziert wurden, einen schlechteren Therapieerfolg aufweisen, als Patienten ohne auffälligen ROI-Score, konnte somit bestätigt werden.

4 Diskussion

4.1 Diskussion des Studienkonzeptes: Assessment von Overreporting

Die vorliegende Untersuchung diente dem primären Ziel, die Nützlichkeit zweier neuer psychologischer Testverfahren (des MMPI-2-RF und des BHI-2) zur Aufdeckung von Overreporting bei chronischen Schmerzpatienten zu untersuchen.

Dabei sollte zum einen die Validität der auf 388 Items verkürzten revidierten MMPI-2-Version (Restructured Form) untersucht werden, aber auch der im deutschen Sprachraum nahezu unbekannte, in US-amerikanischen Reviews eher als nebenrangig eingeordnete BHI-2 (Battery of Health Improvement) mit noch kürzerer Itemzahl (217) überprüft werden (z.B. Rogers 2008, S.34 [235]).

Solche Untersuchungen erfolgten bislang weder an einer konsekutiven Patientenstichprobe chronischer Schmerzpatienten, noch wurden diese Fragebögen systematisch an einer deutschen Stichprobe untersucht.

Insbesondere im Hinblick auf Begutachtungs-Verfahren, bei denen die Glaubwürdigkeit und Glaubhaftigkeit subjektiver Symptome, wie chronischer Schmerzen und damit assoziierter Behinderungen beurteilt werden sollen, ist der Einbezug solcher Validierungsmethoden von besonderem Interesse.

Beide Fragebögen (MMPI-2-/RF, BHI-2) bieten mit der Auswertungs-Möglichkeit ihrer Basisskalen und einer Vielzahl von Validitäts-Skalen und -Indizes dem für den mit diesen Testverfahren wenig vertrauten Anwender eine schwer einschätzbare Fülle von Indikatoren (und entsprechend vielen Studien) zur Erfassung von Beschwerdeüberhöhungen. In der theoretischen Einleitung dieser Arbeit wurde ein systematischer Überblick über die meisten dieser Validierungs-Methoden integriert. Dabei wurde auch die besondere Schwierigkeit von Grenzwerten (sog. *Cutoffs*) der einzelnen Validitätsskalen diskutiert, die je nach Studiensetting unterschiedlich ausfallen und damit als Indizien, jedoch nicht als eineindeutige Beweise für Beschwerdeüberhöhungen beurteilt werden müssen.

Beide Testverfahren sind im deutschen Sprachraum bislang kaum bekannt; eine Vielzahl von Studien untersuchte jedoch die Einsetzbarkeit des MMPI-2, auch bei Patienten mit chronischen Schmerzen, dessen Validität und Trennschärfe zur Aufdeckung negativer Antwortverzerrungen überwiegend als sehr hoch eingeschätzt wird. Insbesondere in der US-amerikanischen Literatur gilt der MMPI-2 als *das Verfahren* zur Detektion von Overreporting.

Bislang ist nicht hinreichend im deutschen Sprachraum untersucht, ob die in den USA von Ben-Porath & Tellegen (2008) [29] neu publizierte, überarbeitete und verkürzte MMPI-2-Version (MMPI-2-RF) ihr erklärtes Ziel einer homogeneren Skalenkonstruktion mit vergleichbarer Qualität wie die Vorgängerversion erfüllt. Auch kann bisher nicht beurteilt werden, welche der Vielzahl von Validitäts-Skalen und -Indizes die beste Beurteilung von Beschwerde-Aggravationen gewährleisten und ob die in US-Studien publizierten Grenzwerte für diese Indikatoren auf deutsche Patienten übertragbar sind.

Mit dem BHI-2 wurde in der vorliegenden Untersuchung erstmals eine übersetzte Fassung des Fragebogens (Dohrenbusch & Brockhaus 2016, in prep. [71]) verwendet, die in nächster Zeit für Untersuchungen an deutschen Patienten mit chronischen Schmerzen verfügbar sein wird. Die Detektions-Qualität dieser deutschsprachigen Testversion mit ihrer Vielzahl an Subskalen sollte mit der vorliegenden Studie explizit an einer großen Gruppe von Patienten mit chronischen Schmerzen untersucht werden.

Eine Beurteilung der Detektionsgüte von Validierungsverfahren kann nur in Relation zu anderen, externen Assessment-Instrumenten zur Aufdeckung von Overreporting erfolgen. Wie einleitend gezeigt, wurden in den letzten Jahren zu diesem Zweck primär Prüfverfahren für kognitive Leistungen (sog. *Performance Validation Tests*, PVTs) verwendet, da diese Verfahren entsprechend ihrer Konstruktion als Zweifach-Wahlaufgaben (*forced two-choice recognition tasks*) fragwürdige Leistungsergebnisse sicher detektieren können (Testleistung unterhalb der Zufallswahrscheinlichkeit).

Die richtungsweisenden Arbeiten von Bianchini et al. (2005) [31] verdeutlichten jedoch, dass nicht allein aus der neurokognitiven Forschung entliehene PVTs zur Überprüfung von Overreporting geeignet sind bzw. zumindest weitere Assessment-Instrumente zweier ebenso wichtiger Overreporting-Dimensionen einbezogen werden sollten. In dem multimodalen, von Bianchini et al. (2005) [31] vorgeschlagenen Assessmentmodell sollten deshalb Methoden der Beschwerdenbewertung eingesetzt werden, die neben kognitiven Leistungsdefiziten auch die *psychologische Störungsebene* sowie die *Verhaltensebene* einbeziehen.

Um ein solches Modell umzusetzen, wurde in der vorliegenden Untersuchung eine umfangreiche externe Validierung durch jeweils zwei unterschiedliche Verfahren in den genannten drei Domänen für Overreporting eingesetzt.

Deutliche Auffälligkeiten in mindestens zwei der genannten Domänen führten zu einer Klassifikation von Patienten, die als Probanden mit einer sog. *Malingered Pain Related Disability (MPRD)* kategorisiert wurden. Ferner wurde eine zweite Patientengruppe unter der Gesamtgruppe der untersuchten 400 Patienten mit chronischen Schmerzen identifiziert, die hinsichtlich ihrer Beschwerde-Angaben als leicht auffällige Probanden (Probably MPRD) beurteilt wurden, wenn sie in mindestens sechs der Prüfverfahren leichte Hinweise auf negative Antwortverzerrungen zeigten.

Dieses Vorgehen entsprach einer Forschungsstrategie, die als sog. *Known-Groups-Design* Patienten mit Beschwerde-Aggravation durch externe Verfahren identifiziert, um im Anschluss die Detektions-Möglichkeiten durch die zu untersuchenden, neuen Testverfahren in einer Kreuzvalidierung zu überprüfen.

Zum dritten ließen sich Patienten aus der Gesamtgruppe vor Beginn der Behandlung identifizieren, die keine besonderen Auffälligkeiten in den externen Verfahren zur Beschwerdenvalidierung aufwiesen, aber durchaus externe Gründe für eine mögliche Beschwerdeaggravation aufwiesen. Zu diesem Zweck wurden insbesondere Schmerzpatienten mit einer befristeten Berentung, mit einem abgelehnten Rentenantrag oder einem Klageverfahren identifiziert, aber auch Patienten mit unentschiedenem Rentenantrag, die sichere finanziell-existentielle Anreize für eine Beschwerde-Überhöhung aufwiesen. Ferner wurden Patienten mit einer längeren Arbeitsunfähigkeitsdauer von mehr als neun Monaten und damit zunehmender Gefahr einer Aussteuerung durch ihre Krankenkasse als Patienten mit externalen Kompensationsmotiven einbezogen.

Letztere, in der Bundesrepublik Deutschland bestehende Besonderheit wurde in der vorliegenden Studie bewusst einbezogen, um möglichst alle Patienten mit externer finanzieller Motivierung für Aggravationen in die Studie als sog. *known group* separat zu untersuchen. In den bisher publizierten, eher wenigen US-amerikanischen Studien, die Patienten mit chronischen Schmerzen untersuchen, werden hingegen zumeist nur Patienten mit deutlichem Aggravationsmotiv eines Klage- oder Gutachtenverfahrens zur Anerkennung einer Rente untersucht.

Verschiedene Studien zeigen, dass Patienten im Streit um Entschädigungszahlungen ihre Beschwerden tendenziell stärker hervorheben als Patienten innerhalb einer Therapie (Birke et al. 2001 [32], Blyth et al. 2003 [35]). In den meisten US-amerikanischen Vergleichsstudien könnte deshalb das Ausmaß der erhobenen Aggravation von Beschwerden durch die Probandenauswahl artifiziell erhöht sein.

Diverse Studien zeigten, dass bei Probanden, die zur Simulation von Beschwerden instruiert wurden, zumeist deutlich höhere Resultate in den Validitäts-Skalen des MMPI-2-RF zur Aufdeckung von Overreporting zu beobachten sind, als bei klinischen Patienten mit zumindest teilweise authentischen Beschwerden (Wygant et al. 2009 [322], Burchess & Ben-Porath 2010 [45], Sellbom & Bagby 2010 [269], Sellbom et al. 2010 [272], Rogers et al. 2011 [239], Sellbom et al. 2012 [273]).

Insofern liefert die hiesige Studie zu diesen Patienten eine vergleichsweise in geringem Ausmaß zu negativen Antwortverzerrungen motivierte und damit schwerer zu identifizierende Gruppe dar.

Fernerhin wurden in die Studie alle übrigen Patienten einer vierten Gruppe zugeordnet, die keinerlei Kompensationsmotive und auch keine Auffälligkeiten in den Beschwerden-Validierungstests aufwiesen. Diese Patientengruppe diente als *Vergleichsnorm* ohne Aggravationsverhalten oder -Motiv. Diese Vergleichsmöglichkeit einer konsekutiv behandelten Normgruppe fehlt in fast allen bisher angewandten Studien.

Aus dieser Vorauswahl ergab sich ein Design aus vier Untersuchungsgruppen, die teils mittels objektiver Kriterien (sozialmedizinischer Statusmerkmale) und teils mittels externer Kriterien (externer Verfahren der Beschwerdenvalidierung) als sog. *known groups* klassifiziert werden konnten.

In keiner der bisher vorliegenden Untersuchungen wurde zudem die Möglichkeit exploriert, welche Auswirkungen ein Overreporting auf das Outcome einer medizinisch-psychologischen Therapie hat, obwohl dies auch anderen Autoren sinnvoll erschien (z.B. Aguereverre 2010 [2]).

Die vorliegende Studie nutzt damit einen neuen Ansatz sowohl der Beschwerdenvalidierung, als auch zweier neuartiger Verfahren der Beschwerdeüberprüfung und untersucht deren Detektionsgüte zur Aufdeckung von Overreporting. Darüber hinaus wurden in der vorliegenden Studie weitere, bisher nicht systematisch untersuchte Alternativ-Methoden der Beschwerdenvalidierung mittels des MMPI-2-RF (der sog. Revised Overreporting Index *ROI*) analysiert, die in bisherigen Studien nicht überprüft wurden.

4.2 Diskussion der externen Gruppen-Klassifikation

Der statistische Vergleich der vier Klassifikationsgruppen zeigte zunächst keine wesentlichen Unterschiede der vier Studiengruppen hinsichtlich ihres Alters, ihrer Schulbildung, des Geschlechts der Patienten, ihrer Haupt-Schmerzdiagnose sowie hinsichtlich ihres Substanzmittelgebrauchs.

Die Patientengruppen unterschieden sich ausschließlich hinsichtlich ihrer subjektiven Angaben in der offenen Abfrage von Depressivität und Angst in der Hospital Anxiety and Depression Scale (HADS, Herrmann-Linger et al. 1995 [134]). Hier zeigten insbesondere die als definitiv auffällig klassifizierten Patienten die höchsten Scores, aber auch die als wahrscheinlich auffällig eingeschätzten Probanden zeigten höhere Scores in der Angstskala als die unauffällig eingeordneten Patienten (IncOnly und NoInc). Hierbei muss berücksichtigt werden, dass in der vorliegenden Untersuchung Schmerzpatienten, die Hinweise auf eine schwere Depressions-, eine Angst-Symptomatik oder sonstige psychiatrische Erkrankungen zeigten, aus der Studie ausgeschlossen wurden.

Insofern erlauben die Auffälligkeiten in der HADS-Abfrage nicht die unmittelbare Schlussfolgerung einer tatsächlich klinisch relevanten Psychopathologie bei diesen Probanden, da es sich um offene, sog. „scheinvalide“ Abfragen handelte, die nicht den Ergebnissen der weniger durchschaubaren Validierungsmethoden (den externen *Performance und Symptom Validation Tests*) entsprachen bzw. durch diese gerade aufgedeckt werden sollten. Von erhöhten Werten in offen durchschaubaren Befragungen auf authentische Beschwerden zu folgern, würde einem naiven Beurteilungsansatz entsprechen, der zu fehlerhaften diagnostischen Einschätzungen führt (vgl. Merten und Dohrenbusch 2010 [197]).

Eine gesicherte Befundvalidierung ergibt sich erst aus der gleichzeitigen Beurteilung augenscheinvalider sowie insbesondere verdeckter für den Probanden nicht erkennbarer Beschwerdeabfragen mit unterschiedlichen Verfahren sowie deren Diskrepanzen und Inkonsistenzen. Selbstberichte über Schmerzen und assoziierte Symptome wie Depression und Angst, ebenso wie Angaben zu Funktions- und Leistungsbehinderungen beinhalten Risiken der Interpretation, die erst durch Einsatz differenzierter Kreuzvalidierungen kontrolliert und überprüft werden können (Dohrenbusch 2009 [69]).

In der hiesigen Untersuchung fiel zudem ein wichtiger korrelativer Zusammenhang auf: je höher die Angaben der Patienten in sog. augenscheinvaliden Verfahren waren (z.B. Hospital Anxiety & Depression-Scale HADS im Schmerzfragebogen), umso höher waren gleichzeitig ihre Angaben in den auf fragwürdige Befunde hindeutenden Verfahren zur Beschwerdenvalidierung (z.B. SIMS, SCL-90-R, kognitive Performance-Validierungstests etc.). Dies verstärkt Zweifel an der Validität der abgefragten Depressions- und Angst-Angaben. Gerade offen abgefragte, subjektive Angaben werden jedoch in Gutachten häufiger diskussionslos zur Beschwerdenbemessung herangezogen.

4.3 Diskussion der Detektionsgüte der Validitätsskalen (MMPI-2-RF & BHI-2)

In den Analysen zur Studienhypothese $H_0 1$ waren deutlich signifikante Unterschiede zwischen den vier Studiengruppen in allen Validitätsskalen des MMPI-2-RF, außer der sog. Lügenskala L-r, festzustellen. Diese Unterschiede fanden sich in allen Validitätsskalen, sowohl zwischen den mit definitiver MPRD klassifizierten Patienten und den als unauffällig eingestuften Probandengruppen, als auch unter Einbezug der nur möglicherweise auffälligen Patienten gegenüber den Patienten mit alleiniger externer Initiative und Probanden ohne Kompensationsmotiv (NoInc, IncOnly, s. Abb. 12, S. 160).

Ein ähnliches Ergebnis der Validitätsskalen zeigte sich auch bei Anderson (2011) [9] im Intergruppenvergleich von Gutachtenprobanden einer neuroforensischen Praxis. Dieselben Unterschiede zwischen den Validitätsskalen bestätigten auch die Studien von Sellbom et al. (2012) [273]. Aber auch Aguerrevere (2010 [2]) fand diese Unterschiede hinsichtlich aller Symptom-Validierungsskalen mit dem schwächsten Effekt in der L-r-Skala und dem deutlichsten Effekt in den Skalen F-r und Fs.

Die hiesige Studie bestätigte einen Cutoff der F-r-Skala von ≥ 10 (T88). Ähnliche Cutoff-Scores der F-r-Skala empfehlen auch die US-amerikanischen Testautoren ($\geq$ T90; Ben-Porath und Tellegen 2008 [29]).

Bei der Einzelanalyse der Standardskalen fanden sich die deutlichsten Diskriminations-Möglichkeiten in der F-r-Skala und der RSB, gefolgt von der FBS-r. Leider wurde die in der hiesigen Studie sehr trennscharfe RBS-Skala in den meisten Studien nicht einbezogen, obwohl sie im MMPI-2-RF zu den Standard-Validitätsskalen gezählt wird. Anderson (2011) [9] fand ähnliche Vorteile der F-r und RBS-Skala, mit den jeweils größten Detektions-Merkmalen.

Die Autorin führte diese Effekte auf die breite Konstruktionsbasis beider Skalen zurück. Insbesondere die RBS wurde konzipiert, um Auffälligkeiten in kognitiven PVTs zu prognostizieren (Gervais et al. 2007 [99]). Die befragten Patienten mit chronischen Schmerzen der vorliegenden Studie schienen somit gleichermaßen überhöhte neurokognitive Defizite anzugeben, wie auch seltene somatische Symptome, aber auch ungewöhnliche psychische Alterationen. Dies entspricht den in der vorliegenden Untersuchung beobachteten Auffälligkeiten der Probanden in allen drei, von Bianchini et al. (2008) [30] benannten Domänen für Overreporting.

Bei der hier untersuchten Aggravation von Patienten außerhalb von Begutachtungs-Situationen war die Überhöhung ungewöhnlicher somatischer und psychischer Symptome, wie sie durch die F-r erhoben werden, auffälliger als Symptome spezifischer Psychopathologie (Fp-r) oder der Somatik (Fs).

Die geringste Detektionsgüte zeigte die Fp-r-Skala zur Erhebung fragwürdiger psychopathologischer Beschwerden. Dies ist möglicherweise darauf zurückzuführen, dass sowohl Patienten in Rechtsstreitigkeiten, aber auch möglicherweise Patienten mit chronischen Schmerzen eher vermeiden, als psychisch krank eingestuft zu werden, um gleichzeitig aber als somatisch schwerer beeinträchtigt zu erscheinen. Diesen speziellen Aspekt soll die FBS-r-Skala erfassen, während die Fp-r-Skala eher seltene psychische Symptome bei psychiatrischen Patienten abfragt. Entsprechend bemerkten Sellbom et al. (2010) [272] bei forensischen Straftätern in den Validitätsskalen F-r und Fp-r (Infrequent Psychopathology Scale) die höchsten Trennschärfen.

Ein möglicher Grund für die eher niedrige Effektstärke der Fp-r-Skala mit höheren Cutoff-Werten (ab T90) könnte auch in der in dem untersuchten stationären Patientensample mit höherer Chronifizierung (nach Gerbershagen Stadium II/III, vgl. Schmitt & Gerbershagen 1999 [260]) und höherer Rate tatsächlicher psychischer Störungen begründet liegen, da in US-amerikanischen Studien mit Patienten mit chronischen Schmerzen häufiger niedrigere Werte dieser Skala ($\geq$ T50) beschrieben wurden (Ben-Porath 2011 [27]).

Dennoch differenzierte die Fp-r-Validitätsskala signifikant zwischen Patienten mit sicherem Overreporting und den nicht auffälligen Klassifikationsgruppen. Burchess & Ben-Porath (2010) [45] beobachteten bei instruierten Simulanten somatischer Symptome ähnliche Scores für Overreporting bei Werten $\geq$ T80 in der Fp-r-Skala. Noch stärker überhöhte Profile (insbesondere in der Fp-r-Skala mit Werten zwischen T90 und T140) wurden in der Studie von Sellbom et al. (2010) [272] bei forensischen Probanden beobachtet.

Die gegenüber den anderen Validitätsskalen geringere Detektionsgüte der Fp-r (Cohen's d = 0,92 bzw. 0,45, s. Tab. 7, S. 161) hängt zudem möglicherweise mit ihrer mehr heterogenen Konstruktion zusammen (Cronbach's α = 0,422, s. Tab. 7, S. 161, vgl. hierzu auch das Dreifaktoren-Modell zur Fp-Skala des MMPI-2, Strong et al. 2006 [284]). Die Fp-r-Skala zeigte deutliche Detektions-Vorteile vor allem in Simulations-Studien mit instruierten Probanden (z.B. Sellbom & Bagby 2010 [269], Burchess & Ben-Porath 2010 [45]), die psychische Symptome je nach Instruktion deutlich gehäufter präsentierten als tatsächlich betroffene Patienten.

Die K-r-(Correction-)Scale zeigte als Underreporting-Instrument erwartungsgemäß reziproke Effekte. Die niedrigsten Scores (MW = 38,8) wurden in der MPRD-Gruppe verzeichnet. Leicht höhere Scores zeigten sich dagegen in der Gruppe mit möglicher MPRD. Die höchsten K-r-Werte ($\geq$ 50) wiesen die Patienten ohne BV-Auffälligkeiten auf.

Die Ergebnisse stützten auch die Nützlichkeit der für den MMPI-2-RF neu konzipierten Skala Fs (Overreporting-Detektion somatischer Symptome), als auch für die kombinierte Erfassung von Overreporting kognitiver und psychischer Beschwerden durch die FBS-r. Die von manchen Autoren (Gass et al. 2012 [96], Schroeder et al. 2012 [264]) kritisierte Heterogenität der FBS-r muss sich nach den hiesigen Analysen somit nicht nachteilig auf ihre Detektions-Eigenschaften auswirken. Auch die in der hiesigen Studie ermittelte innere Konsistenz der FBS-r (Cronbach's α = 0,71) spricht für ihre hohe Trennschärfe.

Die in der hiesigen Studie beobachteten höheren Werte der Patienten mit Overreporting bei T80 in den MMPI-2-RF-Validitätsskalen Fs (Infrequent Somatic Complaints) und der Symptom-Validity-Skala FBS-r im Vergleich zu durchschnittlichen Werten um T65 bei den unauffälligen Patienten stimmen mit den aus vorangegangenen Studien bekannten US-amerikanischen Daten überein (vgl. Burchess & Ben-Porath 2010 [45]).

Letztere Autoren beobachteten, dass Personen, die psychische Merkmale negativ verzerrt darstellten, in den Standard-Validitätsskalen (F-r, Fp-r und Fs) T-Werte > 110 aufwiesen im Vergleich zu Personen, die körperliche Beschwerden verzerrten (Fs Mittelwert T84). Deshalb kann erwartet werden, dass der Vorhersagewert dieser Skalen für übertriebene Beschwerden bei Personen mit vorwiegend körperlicher Symptomatik vergleichsweise gering ausfällt.

Wie erwartet, differenzieren auch die Skalen Fs und FBS-r zwischen authentischen Angaben und Beschwerdeüberhöhungen, allerdings mit einer niedrigeren Effektstärke. Das könnte durch die Spezifität dieser Skalen für die Detektion somatischer Beschwerdenüberhöhungen bedingt sein bzw. aufgrund der Spezifität der FBS-r für die Erhebung neurokognitiver Einschränkungen bei Unfallverletzungen verursacht sein.

Bereits in einer MMPI-2-Metaanalyse berichteten Rogers et al. (2003) [242] geringere Differenzierungs-Möglichkeiten der Skala FBS (mit Cohen's d von 0,32), während die RBS-Skala des MMPI-2 am deutlichsten nicht-authentische kognitive Defizite identifizieren konnte (Gervais et al. 2008 [100]). In der Revisions-Fassung des MMPI-2-RF wurde ebenfalls eine höhere Detektionsgenauigkeit der RBS-Skala gegenüber der FBS-r berichtet (Gervais et al. 2010 [101], Tarescavage et al. 2013 [293]).

Die bei der RBS festgestellten Auffälligkeiten in der MPRD-Patientengruppe ab T90, bei den nur möglicherweise auffälligen Patienten ab T70 und bei nicht-auffälligen Patienten ab T60 bestätigte die bei Wygant et al. (2010) [323] beschriebene hohe Sensitivität der RBS.

In letzterer Studie wiesen Disability-Patienten, die keine BV-Auffälligkeiten aufwiesen, durchschnittlich RBS-Werte um T70 (exakt: 67,1) auf, während Patienten mit Beschwerden-Aggravation RBS-Scores um T90 zeigten (exakt: 88,1). Die *Response Bias Scale* zeigte ab T-Werten $\geq$ 90 bei 90-prozentiger Spezifität eine Sensitivität von 38 %, während sich in der hiesigen Studie ab Rohwert 14 (entsprechend T88) eine deutlich höhere Sensitivität von 84 % zeigte. Diese ist vermutlich auf eine höhere Selektivität unserer MPRD-Patienten durch die gewählte, konservativere Auswahl der externen BV-Kriterien zurückzuführen.

Gervais et al. (2010) [101] empfahlen für die Interpretation erhöhter RBS-Scores eine Kreuzvalidierung mit weiteren Standard-Validitätsskalen des MMPI-2-RF. Nach diesen Empfehlungen ist ab T-Werten von 80-99 in der RBS von einer deutlich erhöhten Aggravation kognitiver Symptome auszugehen. T-Werte von mehr als 99 weisen nach den Autoren auf eine sichere Aggravation hin, wenn der Proband auch in den Skalen F-r, Fp-r, F-s oder FBS-r erhöhte Scores ($\geq$ 99) aufweist und gleichzeitig Auffälligkeiten in externen BV-Verfahren zeigt.

Der speziell zum Overreporting-Assessment kognitiver Symptome konzipierte HHI-r (Henry-Heilbronner-Index) erwies sich bei einem Cutoff $\geq$ 8 als sicher sensitiv für die Auffälligkeiten von MPRD-Patienten (Sensitivität 68 % bei 90-prozentiger Spezifität), wie auch bei den leichter auffälligen P(-Probably)-MPRD-Patienten. Die innere Konsistenz dieser Kurzskala lag im Bereich der Fs-Skala (Cronbach's α = 0,57). Henry et al. (2012) [131] berichteten mittels der im MMPI-2-RF auf 11 Items reduzierten Skala HHI-r bei vergleichbaren Cutoff-Werten $\geq$ 7 eine ähnlich hohe Sensitivität von 68,9 %.

Die deutlichste Diskriminanz zwischen den Untersuchungsgruppen erreichte der sich aus fünf gewichteten Skalen konzipierte MI-r-Index, der ab einem Rohwert $\geq$ 5 eine zu 80 % korrekte Identifikation von Beschwerdeüberhöhungen ermöglichte. Meyers et al. (2013) [201] berichteten bei demselben Cutoff eine hohe 85-prozentige Sensitivität bei 93-prozentiger Spezifität. Dieses Resultat bestätigten in der vorliegenden Studie sowohl die univariaten Varianz-Analysen als auch die Berechnungen der Effektstärken. Nur bei der Gegenüberstellung von MPRD / P(-Probably)-MPRD-Patienten mit Probanden ohne BV-Auffälligkeiten erwies sich F-r gegenüber dem MI-r als trennschärfer, was dafür spricht, dass nur möglicherweise auffällige MPRD-Patienten weniger globale Beschwerde-Überhöhungen (in allen fünf Validitätsscores) zeigen als sicher auffällige Patienten.

Eine **Beschwerdenvalidierung** mittels der **BHI-2-Skalen** lies sich insbesondere durch den Einsatz der sog. Disclosure-Skala (DIS) nachweisen, die das Ausmaß an Bereitwilligkeit von Probanden erfassen soll, Informationen über sich selbst zu präsentieren.

Die DIS-Skala zeigte mit hoher innerer Konsistenz (Cronbach's α 0,93) deutliche Detektions-Möglichkeiten von Overreporting, entsprechend den Differenzierungs-Möglichkeiten der FBS-r des MMPI-2-RF. Mittels beider Validitätsskalen DIS und DEF war es möglich, Patienten mit und ohne MPRD sicher zu identifizieren (vgl. Tab. 13, S. 171). Bei 90-prozentiger Spezifität gelang es, durch die *Disclosure*-Werte 68 % der MPRD-Patienten richtig zu identifizieren.

Ähnliche Unterschiede ermittelten auch die Testautoren Bruns & Disordio (2000) [42] varianzanalytisch in einer Studie mit 214 instruierten Probanden, in denen DIS und DEF zwischen authentischen Probanden und Probanden, die ihre Beschwerden subtil verfälscht stärker oder geringer darstellten. In der vorliegenden Studie wiesen auffällige MPRD-Patienten DIS-Werte von $\geq$ T60 (Rohwert 138) auf. Die entspricht den von Bruns & Disordio (2000) [42] berichteten Scores $\geq$ 135 in der Gruppe „Fake-Bad" im Vergleich zu mittleren Werten von 103 in der Norm-Patientengruppe sowie Werten um 76 in der sog. „Fake-Good"-Gruppe.

Entsprechend fielen in der Studie der Testautoren die mittleren Werte der *Underreporting*-Skala *Defensiveness* bei den „Fake-Bad"-Patienten mit 7,7 ($\leq$ T38) weitaus niedriger aus als die Werte der „Fake-Good"-Gruppe (17,5). Die *Defensiveness*-Scores der MPRD-Patienten der vorliegenden Studie waren vergleichbar (T36,8) und bestätigten die Ergebnisse von Bruns & Disordio (2000) [42]. Der mittlere Cutoff-Wert der DEF-Skala lag in der vorliegenden Studie bei 90-prozentiger Spezifität bei T-Werten $\geq$ 34 (Rohwert 6). Die DEF-Skala erwies sich damit als die weniger sensitive der BHI-2- Validitätsskalen und zeigte auch eine größere Heterogenität (Cronbach's α) als die DIS-Skala. Die festgestellte Ähnlichkeit der Grenzwerte spricht eher für vergleichbare Normierungswerte einer us-amerikanischen und einer deutschen Patientenstichprobe.

Als wenig differenzierend mit noch geringerer innerer Konsistenz erwies sich der nur vier Items umfassende *Validitätsindex* des BHI-2, der sich vermutlich eher zur generellen Beurteilung der Auswertbarkeit des BHI-2 eignet, vergleichbar den Prüfskalen VRIN-r und TRIN-r im MMPI-2-RF.

Zusammenfassend wurden somit beide ersten Untersuchungs-Hypothesen einer akkuraten Diskriminierungs-Fähigkeit der Standard-Validitätsskalen beider untersuchten Fragebögen (MMPI-2-RF und BHI-2) bestätigt, mit einer möglichen Gewichtung der Detektionsgüte der einzelnen Validitätsskalen. In den Analysen erwiesen sich insbesondere die Skalen F-r, RBS und MI-r des MMPI-2-RF und die Skala DIS des BHI-2 als trennscharfe Indikatoren von Overreporting.

In Bezug auf die Klassifikations-Akkuranz zeigte sich, dass eher niedrig angesetzte Cutoff-Werte zu einer hohen Sensitivität bei einem Verlust an Spezifität führen; werden die Cutoff-Scores erhöht, verbessert sich die Spezifität (d.h. weniger falsch positive Klassifikationen), bei Verlust an Sensitivität des Indikators. Insofern bestätigte sich das auch bei Thies (2012) [296] gewählte, international häufige Vorgehen, als Maßstab der Akkuranz zweier Indikatoren die Sensitivität bei einer mindestens 90-prozentigen Spezifität zu vergleichen.

4.4 Diskussion der Detektionsgüte der Basisskalen (MMPI-2-RF & BHI-2)

Auch die meisten der Restructed Clinical Scales des MMPI-2-RF (mit Ausnahme der RC9) waren in der Gruppe mit definitiver MPRD wie auch bei Patienten mit möglicher MPRD deutlich erhöht, im Vergleich zu den nicht auffälligen Patienten und Probanden mit nur äußerem Anlass für Overreporting. Ähnliche Befunde stellte auch Anderson (2011) [9] in acht der neun RC-Skalen fest, mit Unterschieden zwischen den definitiv auffälligen Patienten gegenüber der Probandengruppe mit nur externalen Aggravationsmotiven.

Deutliche Subskalen-Unterschiede fanden sich vor allem in den Skalen „Demoralisierung“ (RCd) und „Somatisierung“ (RC1), was im Hinblick auf das Antwortverhalten von Patienten mit längerer schmerzbedingter Krankschreibung und zunehmend subjektiv wenig beeinflussbarer Belastungs-Situation erwartbar ist. Hypothesengemäß sollten sich insbesondere diese Patienten bei gleichzeitiger Angabe der meisten Somatisierungs-Symptome als am meisten depressiv, pessimistisch und demotiviert darstellen.

Sehr deutliche Auffälligkeiten fanden sich bei den in BV-Verfahren auffälligen Patienten jedoch auch in den Skalen RC7 (dysfunktionale negative Emotionen) und RC8 (ungewöhnliches psychopathologisches Erleben), was als Hinweise auf eine allgemein psychopathologisch erhöhte Beschwerden-Darstellung der Patienten mit Aggravationsverhalten interpretiert werden muss, wie sie auch andere Autoren berichteten (z.B. Aguerrevere 2010 [2], Anderson 2011 [9]).

Erhöhungen der MMPI-2-RF-Basisskalen $\geq$ T90 zeigten in der hier durchgeführten Studie ebenfalls Beschwerdeüberhöhungen an, im Vergleich zu signifikant niedrigen Werte $\geq$ T70 bei Patienten mit nur möglichem Overreporting, während authentisch antwortende Probanden durchschnittlich T-Wert-Erhöhungen um 60 zeigten.

Ähnliche Resultate berichtete Anderson (2011) [9] bei neurologischen Patienten mit finanziellen Kompensations-Ansprüchen, mit vergleichbaren Erhöhungen in den MMPI-2-RF-Basisskalen „Demoralization“ (RCd), „Somatic Complaints“ (RC1) und „Low Positive Emotions“ (RC2). Nur die bei Anderson (2011) [9] als gesichert nicht-authentisch antwortenden Probanden zeigten jedoch zusätzlich deutlich erhöhte Werte in den drei weiteren Basisskalen „Ideas of Persecution“ (RC6), „Dysfunctional Negative Emotions“ (RC7) und „Aberrant Experiences“ (RC8).

Vergleichbare Überhöhungen berichtete auch Aguerrevere (2010) [2] aus einer clusteranalytischen Studie unter Einbezug des MMPI-2-RF bei Rückenschmerzpatienten mit generell überhöhten MMPI-2-RF-Profilen, mit ähnlichen Überhöhungen in den Basisskalen RC6 bis RC8, die auch Sellbom et al. (2010) [272] bei einer Gruppe forensischer Patienten feststellten.

Beschwerdenaggravation wird damit offenbar weniger als „Verlust an Freude“ ausgedrückt, was auffälligen Werten in der Skala „Low Positive Emotions“ (RC2) entspräche, als durch Angaben sehr starker negativer Emotionen, verstärkter Angabe von Psychotizismus-Symptomen und depressiver Demotivation.

Ähnliche Gruppen-Unterschiede wie in der vorliegenden Studie fanden sich bei der Untersuchung von Sellbom et al. (2012) [273] im Vergleich der RC-Skalen von MMPI-2-RF-Protokollen von einerseits Simulanten somatischer Symptome, andererseits medizinischen Patienten und einer dritten Gruppe somatoform-kranker Patienten.

Während die Profile der somatoformen Patienten bei Sellbom et al. (2012) [273] im Profil am ehesten der beiden Probandengruppen ohne BV-Auffälligkeiten (NoInc und IncOnly) in der vorliegenden Studie ähnelten (mit T-Werten zwischen 50 und 60 in den meisten RC-Skalen, Erhöhungen $\geq$ T65 nur in den ersten drei RC-Skalen), wies das Profil der Probanden, die somatische Symptome künstlich überhöhten, große Ähnlichkeiten mit dem Profil der MPRD-Patienten der hiesigen Studie auf (mit hohen Abweichungen $\geq$ T65 in 6 Skalen). Das bei Sellbom et al. (2012) [273] beschriebene unauffällige Profil von Probanden mit rein medizinischen Symptomen (mit T-Werten zwischen 50 und 60 in fast allen RC-Skalen) fand in der vorliegenden Studie kein Äquivalent.

In der Hypothese $H_0$2b wurde auch für die **Basisskalen des BHI-2** eine hinreichende Unterscheidbarkeit von Schmerzpatienten mit und ohne Overreporting angenommen.

Diese Diskriminanz-Fähigkeit bestätigte sich in zwölf der 16 BHI-2-Validitätsscores mit insbesondere erhöhten Werten der „Depressivität“, „Somatisierung“, „Feindseligkeit“, „Angst“ und „Störungen der emotionalen Stabilität“ in der Gruppe der MPRD-Patienten. Signifikant höhere Scores waren aber auch in sieben BHI-2-Basisskalen bei Patienten mit nur möglicher Symptomaggravation festzustellen. Patienten mit nur äußeren Aggravationsmotiv ohne BV-Auffälligkeiten zeigten die niedrigsten Angaben in den BHI-2-Basisskalen, was bekräftigt, dass allein die Tatsache eines Aggravationsmotivs nicht zur Annahme (oder Unterstellung) einer tatsächlichen Aggravation ausreicht.

Die ermittelten varianzanalytischen Unterschiede und Effektstärken der Gruppenunterschiede (Cohen's d 2,1 bis 1,3) dokumentieren eine gute Differenzierbarkeit von Patienten mit MPRD durch die oben genannten fünf BHI-2-Basisskalen. Bisher existieren leider keine ergänzenden Studien, um diese Diskriminanz-Funktionen der BHI-2-Basisskalen vergleichend zu beurteilen. Die hier vorliegenden Analysen unterstützen jedoch die Annahme einer ähnlichen Abbildung von Symptomaggravation in den BHI-2-Basisskalen wie in den Basisskalen des MMPI-2-RF (vgl. Bianchini et al. 2008 [30]).

4.5 Detektionsgüte adaptierter Validitätsskalen im MMPI-2-RF

Zusätzlich bestätigte die vorliegende Studie Hinweise für die Nützlichkeit weiterer, traditioneller Validitätsskalen des MMPI-2, die in verkürzter Fassung auch in der Restructured Form (MMPI-2-RF) enthalten sind.

Insbesondere die Sammlungen kritischer Items nach Lachar-Wrobel (LW-r, Lachar 1979 [162]) und der Items nach Koss-Butcher (KB-r, Koss & Butcher 1973 [159]), aber auch der Dissimulation-Index F-K-r (nach Gough 1950 [105]) sowie die Dissimulation Scale (Ds, Gough 1954 [106]) scheinen wie im MMPI-2 im neuen MMPI-2-RF zusätzliche Möglichkeiten der Detektion von Overreporting zu bieten.

Zwecks Vergleichbarkeit der teils sehr heterogenen Skalen wurden T-Wert-Standardisierungen über die gesamte Patientenstichprobe erstellt, die sehr deutliche Unterschiede in allen acht untersuchten experimentellen Skalen zeigten (vgl. Abb. 15, S. 176).

Die höchste Diskriminanz zwischen den Gruppen zeigte die Skala F-K-r (Cohen's d = 1,99 zwischen MPRD-/und Non-MPRD-Patienten), was angesichts der bereits festgestellten trennscharfen Detektions-Eigenschaften der Standardskalen F-r und K-r nicht überrascht. Der F-K-Index soll nach Gough (1950) [105] sozial erwünschte Antwortmuster identifizieren, wobei simulierende Probanden mit „Faking-Bad-Verhalten" durch gegenüber der Norm deutlich erhöhte Scores erkennbar sein sollen.

Insbesondere differenzierten aber auch die für den MMPI-2-RF adaptierten Summen-Scores kritischer Items (Lachar-Wrobel LW-r und Koss-Butcher KB-r) besonders gut zwischen Probanden mit und ohne Overreporting. Ein Vorteil dieser Indikatoren könnte in der Symptom-Abfrage relativ hoch scheinvalider Items bestehen. Beide Symptomlisten wiesen entsprechend sehr hohe innere Konsistenzen auf (Cronbach's α 0,89 und 0,87).

Hohe Differenzierungs-Merkmale zur Aufdeckung von Overreporting zeigte ebenso die Dissimulation-Skala Ds-rf, der als einzige der in der vorliegenden Studie adaptierten Validitätsskalen bereits von Rogers et al. (2011) [239] auf den MMPI-2-RF adaptiert und untersucht wurde. Rogers et al. (2011) [239] beschrieben eine hohe Detektionsgüte der *Ds-rf* genannten Validitätsskala mit guter Differenzierung zwischen genuinen, authentischen Patienten einer neuropsychologischen Praxis mit Symptomen der Depression und Posttraumatischen Störungen gegenüber Probanden, bei denen mit BV-Verfahren eine überhöhte Beschwerde-Darstellung psychischer oder kognitiver Symptome festgestellt wurde.

Bei Rogers et al. (2011) [239] waren Probanden mit überhöhter Angabe psychischer Beschwerden durch Ds-rf-Werte von durchschnittlich 19,7 gekennzeichnet. Probanden mit überhöhter Angabe kognitiver Beschwerden wiesen in dieser Studie hingegen Ds-rf-Werte von durchschnittlich 14,7 auf. In der hiesigen Studie wurden für eine relativ sicherere MPRD-Klassifikation Ds-rf-Werte von 16 (T62) und mehr festgestellt, die auch einer extrapolierten T-Wertberechnung bei Rogers et al. (2011) [239] von T63 entspricht. Auch die Effektstärke bei Rogers et al. (2011) [239] zur Detektionsgüte der Ds-rf (verfälschte Angaben vs. authentische Patientenangaben) war mit 1,52 der hiesigen Studie vergleichbar (Cohen's d = 1,84 zwischen Patienten mit und ohne MPRD).

Vergleichbar gute Effektstärken berichteten Rogers et al. (2011) [239] auch für alle Standard-Validitätsskalen des MMPI-2-RF (mit besonderer Güte der F-r-Skala, der Fp-r-Skala und der RBS). Die Autoren erklären die hohe Diskriminationsfähigkeit der Validitätsskalen des MMPI-2-RF gegenüber dem MMPI-2 mit der geringeren Zahl skalenüberlappender Items, zu Lasten einer größeren Zahl gleicher Items mit den Restructed Clinical Scales.

Als den am wenigsten mit anderen Skalen überlappenden Indikator kennzeichnen die Autoren die Fp-r-Skala. Sie erklären die besondere Detektionsgüte der Fp-r-Skala in ihrer Studie mit der am besten selektierten Anzahl seltener psychopathologischer Symptome. Jedoch wurden in ihrer Studie primär Patienten mit psychischen Erkrankungen (Depression und Posttraumatische Belastungsstörung) in Begutachtungs-Prozessen untersucht, die vermutlich weit mehr entsprechender Symptome angeben als Patienten mit chronischen Schmerzen, die die Angabe solcher Symptome eher meiden, um nicht als „psychisch krank“ stigmatisiert zu werden (vgl. obige Diskussion der Fp-r, Kap. 4.3, S. 211).

Auch die adaptierte Fb-r-Skala (Failed-back-Scale) zeigte trotz Reduktion von ursprünglich 40 Items im MMPI-2 auf 27 Items im MMPI-2-RF im Intergruppenvergleich eine der Ds-rf vergleichbare Detektionsstärke von Overreporting. Rogers et al. (2003) [238] diskutierten als mögliche Einschränkungen der Aussagekraft dieser Skala, dass Auffälligkeiten auch auf dem Effekt „verminderter Testaufmerksamkeit“ beruhen könnten, da die Fb-Skala Items der ersten und der zweiten Testhälfte des MMPI-2 auf diskrepante Angaben untersucht. Letzteres Argument trifft jedoch die im MMPI-2-RF untersuchten Items dieser Skala nicht, da hier keine Testhälften untersucht werden. Vielmehr überprüfen diese Items der Fb-r-Skala im MMPI-2-RF eher die Angabe-Konsistenz inhaltsähnlicher Symptome, die offenbar ebenfalls als Validitätsindikator nützlich ist.

In anderen Studien (Bagby et al. 2000 [22], s. S. 73) erwies sich die Fb-Skala jedoch insbesondere zur Differenzierung genuiner und simulierter Depressions-Symptome als trennschärfer gegenüber anderen Validitätsskalen des MMPI-2. Aufgrund der engen Parallelität von Depressions-Erkrankungen mit der chronischen Schmerzsymptomatik (Inaktivität, sozialer Rückzug, Demotivation) und der vermehrten Angabe solcher Symptome durch aggravierende Probanden, wie sie sich in der hiesigen Studie zeigten, ist deshalb möglicherweise gerade die Fb-r-Skala besonders zur Detektion überhöhter Symptome geeignet.

Die bereits im MMPI-2 von 58 auf 39 Items verkürzte Ds-r-Skala ist im MMPI-2-RF noch mit 22 Items auswertbar. Dennoch weisen diese Items bei Zusammenfassung in einer auf den MMPI-2-RF adaptierten Validitätsskala Ds-r-r eine deutliche Diskriminationsfähigkeit von Overreporting auf, mit einer etwas geringeren Effektstärke als die Ds-rf.

Die auf den MMPI-2-RF adaptierten Obvious-Subtle-Skalen wiesen trotz einer erheblichen Item-Reduktion (von 256 Items im MMPI-2 auf 152 Items im MMPI-2-RF) eine (wenn auch verringerte) beeindruckende Detektionsgüte auf (vergleichbar dem Ds-r-r).

Auch mittels dieser adaptierten Skala war es möglich, sicher und nur möglicherweise überhöht antwortende Probanden von Schmerzpatienten ohne Auffälligkeiten in den BV-Verfahren signifikant zu unterscheiden. Die Assessment-Strategie subtil-versteckter gegenüber augenscheinvalider Items in fünf Dimensionen psychopathologischer Symptome erwies sich damit grundsätzlich als ein auch in verkürzter Itemanzahl nützlicher Detektionsansatz.

Kritisiert wurde diese von Wiener (1948) [315] entwickelte Detektionsstrategie hinsichtlich ihrer möglicherweise durch höher gebildete Probanden leichteren Durchschaubarkeit. Dieses Argument kann in der vorliegenden Studie nicht abschließend überprüft werden, da sich die vier Klassifikationsgruppen nicht hinsichtlich ihres Ausbildungsstandes unterschieden. Gegen eine solche Verfälschbarkeit der O-S-r-Skala spricht jedoch, dass gerade die MPRD-Patienten, die die auffälligsten Diskrepanzen in den O-S-r-Skalen aufwiesen, die tendenziell geringste Schulbildung aufwiesen.

Die adaptierte *Malingered Depression Scale* (Md-r) nach Steffan et al. (2003) [282] zeigte in den Intergruppenvergleichen die geringste Effektstärke. Sie ermöglichte jedoch ebenfalls eine signifikante Identifizierung der sicher und auch der nur möglicherweise überhöht antwortenden Probanden gegenüber Schmerzpatienten ohne Auffälligkeiten in den BV-Verfahren. Die Md-r umfasst in ihrer auf den MMPI-2-RF adaptierten Form 20 der ursprünglich 32 Items im MMPI-2. Kritisiert wurde ihre Validierung an Studenten mit eher leichter Symptomatik (Thies 2012 [296]). Ihre eher mittlere Trennschärfe in der Studie von Thies 2012 [296] bestätigte sich entsprechend auch in der hiesigen Untersuchung.

Zusammenfassend bestätigten die Untersuchungsbefunde Teilaspekte der Hypothese $H_0 3$, nach der frühere, „ältere“ Validitätsskalen des MMPI-2 auch in ihrer Adaptation auf den MMPI-2-RF nützliche Instrumente zur Aufdeckung von Overreporting anbieten.

Die Analysen zur inneren Konsistenz der Skalen unterstützten zudem die Annahme, dass diese Skalen hoher Homogenität die ebenfalls höchste Trennschärfe aufweisen. Dieser Aspekt wurde in bisherigen Studien zu den Validitätsskalen des MMPI-2-RF kaum berücksichtigt und könnte möglicherweise in künftigen Studien zur Entwicklung detektionsstarker Validitätsskalen weiterführend genutzt werden.

4.6 Diskussion des neu konzipierten Validitätsindex *ROI*

In der **dritten Untersuchungs-Hypothese ($H_0$3)** wurde eine Optimierung der Detektion von Overreporting durch eine gewichtete Summation möglichst homogener und trennscharfer Validitätsskalen in einem Validitätsindex untersucht. Diese Konstruktion basierte auf der Überlegung, durch Zusammenfassung möglichst konsistenter Einzelskalen, die als orthogonale Faktoren unterschiedliche Aspekte von Overreporting erfassen, die Validität eines Gesamtindex zu maximieren.

Zur Hypothesenprüfung wurde der experimentelle Validitätsindex ROI (Revised Overreporting Index) aus fünf der adaptierten MMPI-2-RF-Validitätsskalen höchster innerer Konsistenz unter Einbezug der am deutlichsten trennscharfen Standard-Skala F-r als nach Skalenverteilung gewichteter Index (Staninewerte 7 bis 9 der Validitätsskalen) konzipiert.

Der ROI-Index zeigte erwartungsgemäß eine hervorragende Diskriminanz-Möglichkeit zur Identifikation von Patienten mit Beschwerdeüberhöhungen (Possible und Definite MPRD) gegenüber den Probanden ohne Auffälligkeiten in den BV-Verfahren (Gruppen NoINC und IncOnly). Seine Differenzierungsgüte (Cohen's d = 3,4) lag höher als die mittels des Meyers Validity Index (MI-r) ermittelte Effektivität (Cohen's d = 2,8). Bei einer Spezifität von 94 % ließen sich 96 % der MPRD-Patienten mittels des ROI sicher identifizieren.

Die α-adjustierte, statistische Hypothesenprüfung erfolgte durch einen Vergleich der AUC-Kurven (Area-Under-the-Curve) mittels des *DeLong*-Verfahrens als Maß für die Detektions-Akkuranz. Dabei wurde die Detektionsgüte des neu konzipierten ROI-Index in Bezug zur Güte aller fünf Standard-Validitätsskalen des MMPI-2-RF sowie zur Güte des diese Skalen gewichtet integrierenden MI-r-Index überprüft.

Diese Prüfung ergab letztlich keine hinreichende generell höhere Detektions-Akkuranz des ROI-Index, da er sich nur gegenüber zwei der Standard-Validitätsskalen (Fp-r und Fs) als sicher trennschärfer erwies. Die hypothesengemäß erwartete Überlegenheit des ROI-Index gegenüber dem MI-r-Index, der F-r- und der RBS-Skala konnte nach den AUC-Signifikanz-Tests nicht sicher angenommen werden, obwohl er gegenüber den zum Vergleich untersuchten Validitäts-Indikatoren bei gleicher Spezifität die höchste Sensitivität aufwies.

Erklärend ist hierbei zu bedenken, dass die Konstruktion des ROI mittels einer Standardisierung an Daten der Untersuchungsstichprobe (Area-Transformation nach McCall 1939 [183]) vorgenommen wurde, die nicht notwendig eine exaktere Detektions-Akkuranz als das vergleichbare Gewichtungsverfahren des MI-r-Index ergeben muss. Die optimierte Detektionsgüte des ROI-Index beruht deshalb vermutlich tatsächlich auf der höheren inneren Konsistenz der an den MMPI-2-RF adaptierten, „älteren" Validitätsskalen und spricht für ihre weitere Verwendbarkeit.

Einschränkend ist hierbei zu berücksichtigen, dass das gewählte Standardisierungsverfahren die Gefahr einer artifiziellen, stichproben-abhängigen Optimierung beinhalten könnte. Anhand der Daten anderer Vergleichsstichproben (z.B. der ROI-Scores von Gutachten-Patienten) ist in künftigen Studien zu prüfen, ob der ROI-Index unabhängig von der vorliegenden Stichprobe seine Trennschärfe behält.

Des Weiteren war es für eine umfassende Bewertung der Nützlichkeit der in der vorliegenden Studie untersuchten Validitätsskalen, insbesondere der adaptierten Skalen einschließlich des ROI-Index erforderlich, deren Testgüte mit den Testeigenschaften der traditionellen Validitätsskalen des MMPI-2 zu vergleichen. Dieser Fragestellung wurde in der vierten Untersuchungshypothese $H_0 4$ nachgegangen.

Bevor dieser Aspekt diskutiert wird, ist es jedoch sinnvoll, zunächst die Akkuranz-Ergebnisse der untersuchten Validitätsskalen vergleichend einzuordnen.

4.7 Diskussion der Sensitivitäts-Güte der untersuchten Validitätsskalen

Die Detektionsgenauigkeit (Akkuranz) aller Validitätsskalen des MMPI-2-RF und des BHI-2 wurde mittels Receiver Operating Characteristic Curves (ROC-Kurven) detailliert überprüft. Dabei wurde insbesondere nach den sensitivsten Validitätsskalen für Overreporting bei hoher Spezifität gesucht (Ausschluss falsch positiver, authentisch antwortender Patienten).

Vergleicht man die Akkuranz der untersuchten Validitätsskalen (vgl. 3.4.1, S. 182), so wiesen drei Standardskalen des MMPI-2-RF (F-r, FBS-r und RBS) sowie der MI-r-Gesamtindex eine hohe Sensitivität ($\geq$ 70 %) zur Identifizierung von Patienten mit sicherem Overreporting auf und zeigten damit gegenüber vorangehenden Studien eine sehr hohe Diskriminationsmöglichkeit. Aber auch sechs der acht adaptierten „älteren“ Validitätsskalen wiesen eine Diskriminationsfähigkeit nahe einer 70-prozentigen Sensitivität auf.

Auch die Disclosure-Validitätsskala des neu-entwickelten deutschsprachigen BHI-2 zeigte eine vergleichbar gute Sensitivität (68 %) zur Identifikation von Overreporting. Die Konzeption der DIS-Skala (geringe Preisgabe persönlicher Informationen) scheint somit sehr geeignet zur Aufdeckung von negativen Antwortverzerrungen.

Die Sensitivität der Defensiveness-Skala des BHI-2 wies hingegen in der vorliegenden Studie eine eher mäßige Detektion von Underreporting auf (Sensitivität = 32 %), die jedoch gegenüber der wenig ermutigenden Detektionsgüte der L-r-Skala des MMPI-2-RF als weitaus besser zu beurteilen ist.

Durch die Itemselektion der L-r-Skala im revidierten MMPI-2-RF lies sich somit keine Verbesserung der bereits im MMPI-2 bekannten geringen Diskriminationsfähigkeit der LIE-Skala (z.B. Green 2008, 176pp. [115]) erreichen. Ihre eher leichte Verfälschbarkeit durch Coaching und durch einen höheren Bildungsgrad der Probanden scheint somit ein auch im MMPI-2-RF weiterhin vorhandenes Konstruktions-Problem darzustellen.

Eine endgültige Beurteilung der Detektionsgüte dieser *Underreporting*-Skalen kann jedoch erst erfolgen, wenn auch Probanden mit nachgewiesenem Anlass für **positive** Antwortverzerrungen (z.B. Dissimulation zum Zweck der Vorteilsgewinnung, z.B. zur Beibehaltung einer Opioid-Medikation oder zur Bescheinigung einer fraglichen Fahrtauglichkeit) mit Probanden ohne solche Motive verglichen werden.

In solchen Untersuchungen (beispielsweise von Thies 2012 [296]) wurden für die MMPI-2-L-Skala und alternative Validitätsskalen zur Erfassung von Underreporting höhere Diskriminations-Möglichkeiten ermittelt, als mit dem eher indirekten Ansatz der hier vorliegenden Untersuchung.

Ähnlich wie im MMPI-2-RF spiegelten auch die **Basisskalen des BHI-2** einzelne Aspekte von Beschwerdeüberhöhungen trennscharf wider, wie der Vergleich ihrer Diskriminationsfähigkeiten belegt. Insbesondere die Basisskalen zur Somatisierung, zur Schmerzintensität, zur Depression und zum Aufmerksamkeits-Gewinn (Symptom-Dependency) erwiesen sich bei mehr als 90-prozentiger Spezifität als hoch sensitiv zur Detektion von Overreporting.

Die Basisskalen „Angst“, „Funktionalität der Schmerzen“, „Borderline-Symptome“ und die „Critical-Items“ des BHI-2 zeigten eine nahezu 50-prozentige Sensitivität. Insofern eignen sich auch die BHI-2-Basisskalen zur Aufdeckung negativer Antwortverzerrungen.

4.8 Diskussion der Vergleichsanalysen beider MMPI-2-Versionen

In der **vierten Untersuchungshypothese** H_0**4** wurde die Detektions-Genauigkeit aller im MMPI-2-RF erfassten Validitätsskalen systematisch mit der Akkuranz ihrer Äquivalenz-Skalen aus der Vorgänger-Version des Fragebogens verglichen.

Zu diesem Zweck wurden die Diskriminations-Kennwerte (Receiver Operating Curves, Area-Under-the-Curves) der Validitätsskalen des MMPI-2-RF und ihrer Äquivalenz-Skalen mittels einer z-verteilten Signifikanzprüfung (nach DeLong et al. 1988 [63]) paarweise verglichen. Hierbei zeigten sich signifikant verbesserte Detektions-Eigenschaften von **drei der Standardskalen des MMPI-2-RF** gegenüber ihren Vorgängerversionen im MMPI-2.

Die neu konzipierte F-r-Skala wies eine noch höhere Diskriminanzgüte auf (Area-Under-the-Curve AUC = ,97) als die bereits sehr hohe Diskriminationsfähigkeit der F-Skala des MMPI-2 (AUC = ,90). Demgegenüber ließ die Optimierung der Fp-r-Skala im Vergleich zu ihrer Vorgängerversion mit eher schlechter Diskriminanz (AUC = ,59) Verbesserungen erkennen (AUC = ,75); die eher unzureichende Trennschärfe dieser Skala in der vorliegenden Studie spiegelte sich jedoch erwartungsgemäß auch in ihrer Akkuranzkurve wider. Als dritte Skala zeigte die FBS-r-Skala eine leicht höhere Diskriminanz gegenüber ihrem Äquivalent im MMPI-2 ($p \leq 0{,}05$), mit einer im Interskalen-Vergleich hohen Detektionsgüte (AUC = 0,88).

Unter den experimentellen, an den MMPI-2-RF adaptierten Validitätsskalen zeigte allein die F-K-r-Skala eine signifikant höhere Detektionsgüte als ihre MMPI-2-Vorgängerversion, was möglicherweise auf die Optimierung der F-Skala zurückzuführen ist. Die deutlich verbesserte Akkuranzkurve (AUC = 0,96 gegenüber AUC = 0,79) bestätigte hier auch die in den Diskriminationsanalysen festgestellte optimierte Detektionsqualität im MMPI-2-RF.

Die Verkürzung der Md-Skala im MMPI-2-RF führte hingegen gegenüber den guten Detektionseigenschaften der Malingered Depression Scale im MMPI-2 zu einer signifikant schlechteren Leistung dieser Skala, so dass Anwendern eher die Verwendung der älteren Version empfohlen werden muss.

Vergleicht man die Akkuranzkurven der beiden Validitätsskalen des BHI-2, weist die Disclosure-Scale eine signifikant höhere Detektionsgüte gegenüber der Defensiveness-Skala auf, wie sie bereits die varianzanalytischen Gruppenvergleiche zeigten.

Einschränkend muss hier jedoch – wie oben bereits erläutert – zugunsten der primär zum Assessment von Underreporting konzipierten DEF-Skala berücksichtigt werden, dass die ROC-Analysen nur auf indirekten Analysen von Patienten mit Overreporting basierten. Insofern könnten Diskriminanz-Qualitäten der DEF-Skala bei einer geeigneten Stichprobe noch höher ausfallen als in der vorliegenden Untersuchung.

Wie in der Untersuchungshypothese $H_0 3$ bestätigt, zeigte der Vergleich der Akkuranzkurven der beiden Validitätsindizes für den MMPI-2-RF eine leicht höhere Detektionsgüte des ROI gegenüber dem MI-r, die jedoch statistisch keine Signifikanz erreichte. Beide Validitätsindizes zeigten eine hohe Detektionsgüte zur Aufdeckung von Overreporting, die jedoch bei vereinfachter Auswertung auch unter Verwendung der F-r-Skala oder der adaptierten Validitätsskalen LW-r, KB-r und F-K-r erreicht werden kann.

Da bislang entsprechende internationale Studien zum Abgleich der Akkuranz der Validitätsskalen beider Fragebögen nicht vorliegen, können die Ergebnisse der vorliegenden Untersuchung nicht mit anderen Studien verglichen werden.

In Übereinstimmung mit früheren Untersuchungen (Simms et al. 2005 [275], Sellbom et al. 2006 [271], Van der Heijden et al. 2010 [305]) bestätigten die systematischen Vergleiche der Validitätsskalen des MMPI-2 und des MMPI-2-RF für beide Testformen dieselben guten Diskriminanz-Eigenschaften. Dies betraf insbesondere die sehr gute Diskriminanzgüte der Skalen F-r, der RBS-Skala und der FBS-r-Skala.

4.9 Diskussion des Einflusses von Overreporting auf den Therapieerfolg

Abschließend wurde in der **fünften Untersuchungshypothese** H_0**5** geprüft, ob sich Probanden mit externalen Kompensationsmotiven und Hinweisen auf Beschwerdeüberhöhungen auch hinsichtlich der Ergebnisse einer multimodalen, medizinisch-psychologischen Schmerzbehandlung unterscheiden. Das Fehlen dieser wichtigen Informationen, als ein bislang nicht beantworteter Forschungsaspekt, wurde bereits von Aguereverre (2012) [2] angemerkt.

Als mögliche Erfolgsparameter in der Schmerztherapie werden diverse Operationalisierungen diskutiert. Neben der Schmerzintensität werden funktionale Verbesserungen auf Verhaltensebene, aber auch Kosteneinsparungen hinsichtlich der weiteren Behandlung oder auch die Wiederherstellung der Arbeitsfähigkeit (*back-to-work*) diskutiert.

Letztlich bemisst sich der Erfolg einer Schmerzbehandlung jedoch immer am Hauptsymptom von Patienten mit chronischen Schmerzen, der subjektiven Belastung durch Schmerzen. Der Therapieerfolg wird durch Patienten und Therapeuten am häufigsten anhand der Reduktion der Schmerzstärke beurteilt. Aus diesem Grund wurde die Einschätzung des Therapieerfolgs in der vorliegenden Untersuchung an Veränderungen der Schmerzstärke vor und nach der Behandlung gemessen.

Overreporting von Beschwerden hatte offenbar einen bedeutsamen Effekt auf die berichtete Schmerzwahrnehmung. Zum einen zeigte sich, dass die Patienten mit zunehmendem Grad an externaler Motivation für Overreporting und mit Zunahme feststellbarer Auffälligkeiten in den BV-Verfahren *bereits vor der Therapie* tendenziell höhere Schmerzangaben machten. Diese Unterschiede waren jedoch vor der Therapie nicht signifikant.

Nach der Behandlung wiesen die vier Studiengruppen jedoch mit Zunahme von Aggravationshinweisen auch eine zunehmende und signifikant geringere Schmerzreduktion auf, wenngleich sich alle Patientengruppen im Therapieverlauf verbesserten.

Dieser Unterschied war sowohl bei Patienten mit möglicher und sicherer Aggravation im Vergleich zu den Patienten ohne Auffälligkeiten in den BV-Verfahren festzustellen, als auch zwischen MPRD-Patienten und Patienten ohne eine sichere MPRD.

Auch bei Patienten, die entsprechend eines höheren Cutoff-Wertes im ROI-Index (Revised Overreporting Index) als Patienten mit auffälliger Beschwerdeüberhöhung einzuordnen waren, und bei Patienten, die einen niedrigeren ROI-Wert (≤ 2) aufwiesen, war nach der Behandlung eine signifikant geringere Schmerzreduktion festzustellen. Insofern bestätigte sich die Untersuchungshypothese $H_0 5$ deutlich. Bereits vor einer multimodalen Schmerztherapie können somit jene Patienten mit auffallend negativen Antwortverzerrungen identifiziert werden, die von der Behandlung in geringerem Ausmaß profitieren werden, als die nicht entsprechend auffälligen Patienten.

Mögliche Erklärungen hierfür sind eine geringere Compliance dieser Patienten in Therapieprogrammen, die insbesondere eigentherapeutisches Bemühen und Selbstmanagement der Patienten fördern wollen. Gelingt der Aufbau solcher Eigenbewältigungs-Bemühungen nicht hinreichend, sind entsprechend schlechtere Therapie-Ergebnisse zu erwarten.

Dieses Ergebnis ist insbesondere bemerkenswert, da in der vorliegenden Studie die Behandlung unabhängig von der Klassifikation der Patienten in allen Fällen unter intensivem Einsatz vielfältiger stationärer Therapiemaßnahmen erfolgte, einschließlich teilweise invasiver Behandlungstechniken wie Nervenblockaden und in vielen Fällen auch der zusätzlichen Verordnung von Opioid-Analgetika.

An dieser Stelle sollte nur ansatzweise auf Diskussions-Impulse hingewiesen werden, wie therapeutisch mit festgestellten Beschwerdeüberhöhungen umgegangen und diese vermindert werden könnten. So berichteten Merckelbach & Merten (2012) [191], dass sich eine eher diplomatische Aufklärung der Betroffenen unter Hinweis auf moralische Maßstäbe im Sinne kognitiver Dissonanzreduktion als hilfreicher gegenüber einem konfrontativen Vorgehen erwiesen haben.

Letzteres Vorgehen führt offenbar eher zu vermehrtem Anstieg von überhöhten Beschwerdedarstellungen (quasi als Beweis der Beschwerden), was ebenfalls im Sinne der kognitiven Dissonanz-Theorie erklärbar ist.

4.10 Limitierungen der Ergebnisse und Ausblick auf künftige Untersuchungen

Unter methodischem Aspekt bestimmten in der hier durchgeführten Untersuchung die klinische Praktikabilität, die Verfügbarkeit, die Kosten, die Durchführungs-Dauer und -Ökonomie die Auswahl der eingesetzten Verfahren zur externen Beschwerdenvalidierung.

Einige möglicherweise besser erprobte *Performance Validation Tests* und *Symptom Validation Tests* existieren zurzeit noch nicht in deutscher Fassung (Structured Interview of Reported Symptoms nach Rogers et al. 1992 [237]). Manche Verfahren sind in der Durchführung zu lang, zu aufwendig und den Patienten im Rahmen ihres sonstiges Assessments nicht zumutbar (z.B. Verfahren zur Evaluation der funktionellen Leistungsfähigkeit).

Manche dieser *Golden-Standard*-Verfahren sind lizenz-gebunden, und ihre Verwendung verpflichtet den Nutzer zu wiederholten Kosten (z.B. Word Memory Test, Green 2003 [111]). Manche Verfahren sind methodisch gut konzipiert, aber optisch weniger ansprechend aufbereitet und nur mit erhöhtem Aufwand (spezieller Laptop) durchführbar (z.B. Aggravations-Simulations-Test AST, Eberl und Wilhelm 2007 [79]).

Insofern erfüllten nicht alle der in der vorliegenden Untersuchung eingesetzten Beschwerden-Validierungsverfahren zur Identifizierung von Overreporting die hohen Kriterien eines *Golden-Standards*.

Beispielsweise kann die Verwendung der Symptom-Checkliste SCL-90-R zur Aufdeckung von Beschwerdeüberhöhungen kritisch diskutiert werden, denn dieses Verfahren wurde ursprünglich zum Screening psychischer und allgemeiner psychopathologischer Symptome entwickelt. Dennoch zeigen Überhöhungen in den meisten SCL-Subskalen sensitiv ein Overreporting an (McGuire & Shores 2001 [185]), ebenso wie sich Belege für die Sensitivität von SIMS und SCL-90-R in Simulations- und Coaching-Studien finden ließen (Gillard 2010 [103], vgl. S. 52ff.).

In den vorangehenden Studien wurde jedoch wiederholt die unterschiedliche Festlegung von Cutoff-Werten für Overreporting im Fragebogen SIMS (Structured Interview of Malingered Symptomatology) mit Werten von 14, 17 oder 23 kritisch diskutiert (Smith 1992 [280], Rogers et al. 1996 [240], Edens et al. 1999 [80], 2007 [81], Heinze & Purisch 2001 [128], Alwes 2006 [8]). Denn dieses Verfahren erwies sich zur Aufdeckung eines Overreporting psychischer Symptome zwar als hoch spezifisch, aber teilweise als mäßig sensitiv (Van Impelen et al. 2014 [303]).

Diese Mängel eines *Golden-Standards*, insbesondere in der Auswahl der Verfahren zur Beschwerdenvalidierung psychischer Symptome, können kritisch diskutiert werden. Die Auswahl der externen Verfahren der Beschwerdenvalidierung wurde bereits bei der Studien-Konzeption kritisch abgewogen.

Im Studien-Design wurde den möglichen Einschränkungen mancher der eingesetzten externen Validitäts-Verfahren zum einen durch eine Erhöhung der Cutoff-Scores der SIMS Rechnung getragen, um die Möglichkeit falsch positiver Klassifikationen zu verringern, aber auch durch einen die Verfahren kombinierenden multimodalen Algorithmus.

Denkbar wäre, dass die Auswahl von zwei BV-Verfahren zur Erhebung psychopathologischer Symptome mit weniger hoher Detektionsgüte eine Ursache der in der vorliegenden Untersuchung festgestellten geringeren Sensitivität der Fp-r-Skala ist.

Dagegen spricht jedoch einerseits, dass auch der bei sicher auffälligen Probanden (MPRD-Patienten) gewählte höhere Cutoff des SIMS ($\geq$ 16) nur bedingt eine bessere Detektionsfähigkeit der Fp-r im Vergleich zu den anderen Validitätsskalen bewirkte. Ebenso spricht dagegen, dass der bei sicher auffälligen Probanden (MPRD-Patienten) gewählte Algorithmus (Auffälligkeiten in mindestens zwei der drei Domänen) eine Mehrfach-Absicherung der Klassifikation gewährleistete, so dass kein sicher identifizierter Patient *nur psychische Auffälligkeiten* in den BV-Verfahren aufwies.

Zum dritten wies die Fp-r-Skala auch in einigen vorangehenden Studien die vergleichsweise geringste Aufdeckungs-Güte unter den Validitätsskalen des MMPI-2-RF auf (z.B. Anderson 2011 [9] bei Gutachten-Patienten, Wygant et al. 2009 [322] oder Burchess & Ben-Porath 2010 [45] bei Simulanten medizinischer bzw. somatischer Symptome).

Untersuchungs-Methoden zur externen Beschwerdenvalidierung könnten in zukünftigen, noch differenzierteren Studien auf stärker an *bona-fida*-Kriterien testpsychologischer Güte orientiert werden. So wäre beispielsweise eine Verwendung des PAI (Personality Assessment Inventory) nach Morey (1991 [211], 1996 [210]) denkbar, das aktuell auch in deutschsprachiger Fassung verfügbar ist (Verhaltens- und Erlebensinventar VEI, Engel 2013 [85]). Der Aufwand für die Patienten und die damit verbundenen Schwierigkeiten der gesamten Untersuchung (z.B. geringere Mitarbeitsbereitschaft der Probanden) würde mit der Länge dieses Verfahrens jedoch erheblich steigen.

Eine Übersetzung und Verwendung des im amerikanischen Sprachraum häufig verwendeten SIRS (Structured Interview of Reported Symptoms nach Rogers et al. 1992 [237]) liegt derzeit noch nicht vor, könnte jedoch ebenfalls im klinischen Routine-Einsatz an der erhöhten Durchführungsdauer dieses Verfahrens scheitern.

Im Bereich der kognitiven *Performance Validation Tests* wäre ebenfalls der Einsatz von Instrumenten denkbar gewesen, deren höhere Trennschärfe belegt ist (z.B. Word Memory Test WMT nach Green 2003 [111]) oder Medical Symptom Validity Test MSVT nach Green 2004 [112] oder Test of Memory Malingering TOMM Tombaugh 1996 [298]).

Jedoch wurden den Schwächen der in der vorliegenden Untersuchung eingesetzten BV-Verfahren zur Erhebung kognitiven Overreportings (z.B. Rey-15-Item-Memory-Test) zum einen durch eine Modifikation Rechnung getragen, die die Validität deutlich erhöhen kann (Erweiterung zum Rey-15-Test). Zum anderen wurde dieses Verfahren durch einen zweiten *Performance Validation Test* mit einem Zwangswahl-Design (DMS-48, Description of the visual recognition Memory Task, Barbeau et al. 2004 [23]) abgesichert, so dass auch in dieser Domäne eine erhöhte Sicherheit bei der externen Validierung negativer Antwortverzerrungen bestand (Performance kognitiver Leistungen unter Zufallswahrscheinlichkeit).

Auch wäre der Einsatz aufwendigerer, computergestützter Verfahren zum Assessment neurokognitiver Beschwerden zur Validierung kognitiver Beschwerde-Aggravation denkbar gewesen. So ist es beispielsweise möglich, die Reaktionszeit der Probanden in sog. *forced-choice*-Verfahren zu erheben, wie z.B. mittels des Aggravations-Simulations-Tests (AST, Eberl & Wilhelm 2007 [79]), um testmanipulative Effekte weitgehend auszuschließen. Jedoch war der Einsatz solcher Testverfahren in der vorliegenden Untersuchung ebenfalls aus Gründen der Durchführungs-Ökonomie und Praktikabilität nicht vorgesehen.

Ferner könnte an der Durchführung der vorliegenden Studie kritisiert werden, dass die Klassifikation behavioraler Auffälligkeiten durch eine eher wenig erprobte Technik erhoben wurde (Interrater-Reliabilität der Behinderungs-Einschätzungen von Probanden und Untersucher). Jedoch sicherte diese Methodik zumindest einen höheren Grad an Objektivität als die Klassifikation der zumeist üblichen einfachen Verhaltensbeobachtungen.

Eine zweite Absicherung zur Klassifikation von Auffälligkeiten auf der Verhaltensebene erfolgte durch die Einschätzungen des untersuchenden Arztes und der Pflegekräfte in den Team-Konferenzen. Hierbei muss betont werden, dass diese Einschätzungen unabhängig von der Beurteilung des untersuchenden Psychologen zur Klassifikation möglicher und sicherer Beschwerdenaggravation abgefragt wurden. Ebenso waren die nicht-psychologischen Beurteiler nicht über die Ergebnisse der testdiagnostischen Erhebungen in den übrigen BV-Verfahren informiert. Insofern war eine Unabhängigkeit der Einschätzungen zur Beschwerdeaggravation gewährleistet.

Zusammenfassend kann somit - trotz der beschriebenen Schwächen der gewählten externen Verfahren zur Beschwerdenvalidierung - von einer validen und multimodalen Absicherung der Overreporting-Klassifikation ausgegangen werden.

Als weiterer kritischer Aspekt der durchgeführten Untersuchung könnte der in der Untersuchung gewählte **sehr konservative externe Klassifikations-Algorithmus** diskutiert werden, der zu einer eher geringen Rate sicher durch Overreporting gekennzeichneter Patienten führte. Dieses Vorgehen wurde jedoch bewusst im Sinne einer Ent-Stigmatisierung der hier untersuchten Patientengruppe gewählt, um die Anzahl falsch positiv klassifizierter Probanden möglichst gering zu halten.

Ein Grund für die Wahl dieses Algorithmus lag zum einen in den oben ausführlich bekannten Schwächen der gewählten externen Klassifikations-Indikatoren, die durch eine möglichst konservative Festlegung der Kriterien ausglichen werden sollte.

Zum anderen sind bei kritischer Betrachtung die Definitionen von Bianchini et al. (2005) [31] zur Klassifikation möglicher oder wahrscheinlicher Aggravation zur Operationalisierung nur bedingt präzise festgelegt.

Bianchini et al. (2005) [31] setzten als Grundbedingungen einer MPRD-Klassifikation folgende Kriterien fest: zum einen muss ein relevanter externer Anreiz für Aggravation vorliegen, zum zweiten dürfen die Beschwerden nicht vollständig durch psychiatrische, neurologische oder entwicklungsbedingte Faktoren/Befunde erklärbar sein. Die weiteren Kriterien für eine mögliche oder wahrscheinliche Aggravation sind jedoch weniger deutlich formuliert.

Mögliches Overreporting ist nach den Autoren festzustellen, wenn die Auffälligkeiten in den drei Domänen (kognitiv, behavioral und psychopathologisch) für die Diagnose einer wahrscheinlichen Aggravation „ungenügend“ sind. Wahrscheinliches Overreporting ist festzustellen, wenn zwei oder mehr „wahrscheinliche“ Auffälligkeiten vorliegen.

Insbesondere der Begriff einer „wahrscheinlichen“ Auffälligkeit in den Verfahren zur Beschwerdenvalidierung ist jedoch nicht klar präzisiert und von den in unterschiedlichen Studien publizierten Trennwerten (*Cutoffs*) der jeweils eingesetzten externen Klassifikations-Verfahren abhängig.

Aus diesem Grund mangelnder Präzisierung wurde in der vorliegenden Untersuchung eine annäherungsweise Operationalisierung „möglicher“ Beschwerdeüberhöhung anhand der verfügbaren Literatur zu den verwendeten BV-Verfahren definiert und zusätzlich - in Abweichung von den Kriterien nach Bianchini et al. (2005) [31] - durch eine Mehrfach-Validierung in allen sechs BV-Verfahren realisiert.

Dieses Vorgehen zur Klassifikation „möglicher Auffälligkeiten“ birgt natürlich eine gewisse Gefahr fehlerhafter Zuordnungen, die sich in den Ergebnissen möglicherweise in den eher geringen Auffälligkeiten dieser Studiengruppe (P-MPRD) gegenüber den Klassifikations-Gruppen ohne Auffälligkeiten (NoInc und IncOnly) widerspiegelt.

Möglicherweise aufgrund dieser Unschärfen der Klassifikation verwendeten Bianchini et al. (2008) [30] in ihrer eigenen Analyse zur Detektionsgüte der Validitätsskalen des MMPI-2 ausschließlich die Studiengruppen NoInc, IncOnly und MPRD.

Trotz der genannten Mängel scheinen jedoch die deutlichen Unterschiede zwischen allen vier Studiengruppen in den acht adaptierten Validitätsskalen des MMPI-2-RF (vgl. Abb. 15, S. 176), die sich durch eine besondere Trennschärfe auszeichneten, die Zuverlässigkeit des multidimensionalen Klassifikations-Algorithmus zu bestätigen.

Bemerkenswert war, dass die Patienten mit nur äußerem Anlass für Overreporting (IncOnly), die jedoch keine Auffälligkeiten in den Verfahren zur Beschwerden-Validierung aufwiesen, die *geringsten Scores* in allen Validitäts- und Basisskalen des MMPI-2-RF und des BHI-2 aufwiesen. Ihre Werte unterschritten sogar tendenziell die durchschnittlichen Werte jener Patienten, bei denen keinerlei Anzeichen für Aggravation bestand (NoInc).

Insofern stellt die reine Feststellung eines prinzipiellen Aggravationsmotivs (z.B. längere Arbeitsunfähigkeitsdauer oder Rentenantrag) keine hinreichende Bedingung dar, Beschwerdeüberhöhungen anzunehmen. Auch profitierte diese Patientengruppe - ebenso wie alle anderen untersuchten Patienten der Stichprobe - signifikant von der durchgeführten multimodalen stationären Schmerztherapie.

Bianchini et al. (2005) [31] wiesen bereits darauf hin, dass Kompensations-Ansprüche bei Arbeitstätigen oder Ansprüche von Patienten im Rahmen juristischer Auseinandersetzungen wegen Unfall- oder Verletzungsfolgen nur zwei *mögliche Motivationen* für Overreporting darstellen.

Beispielsweise ist nicht sicher, ob überhaupt, in welchem Ausmaß und zu welchem Zeitpunkt ein zeitbefristet frühberenteter Patient oder ein für eine längere Zeit krankgeschriebener Arbeitnehmer seine Beschwerden möglicherweise überhöht, um beispielsweise eine verlängerte Krankschreibung oder eine weitere Kompensations-Zahlung zu erhalten. Andererseits kann auch ein bereits frühberenteter Patient andere externe Gründe für eine Beschwerdeüberhöhung haben, z.B. um eine Höhereinstufung eines Grades der Behinderung zu erhalten oder auch um eine fortgesetzte oder erhöhte Weiterverordnung von Opioid-Analgetika zu erreichen.

Auch können andere Gründe überhöhte Beschwerdeschilderungen und erhöhten psychischen Distress bei Patienten verstärken, die dem Arzt oder Psychotherapeuten nicht bekannt sein müssen. Schließlich muss ein Patient nicht die volle Wahrheit über seine persönlichen Intentionen mitteilen, eine Therapie aufzusuchen. All diesen Gründen für Overreporting wurde in der vorliegenden Studie durch eine möglichst große Auswahl denkbarer Motive Rechnung getragen.

Der natürlichen Erwartung entsprechend zeigten sich in der vorliegenden Untersuchung bei den auffälligen Patienten *Aggravationen auf allen Verhaltens- und Untersuchungsebenen gleichermaßen*. Es darf angenommen werden, dass ein hinsichtlich des multimodalen Assessment unkundiger oder zumindest nicht vollständig informierter Patient, der bewusst oder auch unbewusst Beschwerden verstärkt präsentieren möchte, nicht zwischen den einzelnen Assessment-Ebenen differenziert.

Möglicherweise sind gerade Pointierungen spezieller Aspekte der Beschwerden (z.B. vermehrt psychische oder vermehrt körperliche Symptome) ein Hinweis auf bewusste Antwortverzerrungen. Da diese speziellen Motive jedoch in der Gruppe der nicht-authentischen Patienten vermutlich zufällig vorkommen, treten diese Besonderheiten in der Gesamtschau nicht hervor.

Frühere Studien bestätigten zudem, dass nicht alle Antragsteller sozialmedizinischer Leistungen generell überhöhte Beschwerdeangaben machen; in den meisten Untersuchungen von Probanden in Gutachtenverfahren lagen die Basisraten für Overreporting bei maximal 30 Prozent. In klinischen Stichproben sind weitaus niedrigere Basisraten eher die Regel (z.B. 10,7 % bei Blanchard et al. 2003 [33], ähnlich gering in einer psychiatrischen Stichprobe im Vergleich zu instruierten Simulanten in einer Simulationsstudie, s. auch Sellbom & Bagby 2008, 185pp. [268] zu Grundraten in klinischen Stichproben).

Auch in der hier untersuchten klinischen Stichprobe fand sich ein Bezug auf sichere Beschwerdeüberhöhung nur bei 7 Prozent der Untersuchten. Unter Einbezug der möglicherweise aggravierenden Patienten wurde eine Basisrate von 16 Prozent festgestellt. Insofern entspricht die Häufigkeit der Beschwerdeüberhöhungen weitgehend anderen Vergleichsstudien und unterstützt damit indirekt ebenfalls das verwendete multimodale Klassifikations-Modell.

Eine weitere, zu diskutierende Besonderheit könnte in der vorliegenden Untersuchung in der **Fokussierung auf die sehr homogene Stichprobe** stationärer Patienten mit chronischen Schmerzen gesehen werden, mit einer weitgehend einheitlichen Diagnose (F45.41) und einer ähnlichen medizinisch-psychologischen Kombinations-Behandlung.

Dieser Fokus der Patientenauswahl wurde aufgrund eines Mangels entsprechender Studien vorgesehen, aber auch, weil beide Fragebögen (MMPI-2-RF, BHI-2) insbesondere im Hinblick auf diese Patientengruppe konzipiert wurden. Zudem hat eine Patienten-Stichprobe aus dem klinischen Setting den Vorteil, dass vielfach diskutierte Verfälschungs-Effekte (vgl. Thies 2012 [296], z.B. Jelicic et al. 2007 [143]) durch eine „Coaching“-Instruktion von Gutachten-Probanden zum Ausfüllen der Testverfahren, praktisch ausgeschlossen werden können. Die Mehrzahl der Patienten unternimmt in einem solchen Behandlungs-Setting mangels Vorbereitung wahrscheinlich geringere Anstrengungen zur Beeinflussung von Untersuchungs-Ergebnissen als Gutachten-Probanden, die entsprechende bewertende Untersuchungen erwarten.

Dennoch könnte man argumentieren, dass die in der vorliegenden Untersuchung gewählte Patientengruppe aufgrund ihrer Vielzahl von somatischen Befunden am ehesten Patienten ähneln, die unter anderen somatischen Krankheiten leiden, und die beobachtbaren psychischen Alterationen eher „normale“ Reaktionen im Rahmen von schwierigen Erkrankungen darstellen. Aufgrund dieser Probandenauswahl wäre die **Identifikation aggravierender Probanden besonders erleichtert** und wäre weit schwieriger in einer Patientengruppe mit rein-somatoformen Störungen (F45.40) ohne eine die Beschwerden erklärende somatische Erkrankung zu erfassen (wie z.B. in der Studie von Sellbom et al. 2012 [273]). Dies könnte die Gefahr einer artifiziell zu hohen Detektions-Güte von Testverfahren aufgrund der Stichprobenauswahl erhöhen.

Auf diesen Aspekt verweist die genauere Betrachtung des Problems der Basisraten (vgl. Rosenfeld et al. 2000 [250], s. Kap. 1.7, S. 29), nach dem sich die Positive Predictive Power, d.h. die Spezifität eines Indikators, gleichermaßen mit der Reduktion der Basisrate verringert, hingegen die Sensitivität (als Negative Predictive Power) eher zunimmt.

Eine Umrechnung der ermittelten Akkuranzdaten für die Validitätsskalen des MMPI-2-RF und des BHI-2 auf eine bei Gutachtenprobanden typische Basisrate (von 30 %) findet sich im Anhang (s. Kap. H, S. 364).

Hierbei zeigte sich, dass entsprechend der bei Rosenfeld et al. 2000 [250] beschriebenen Zusammenhänge mit **höherer Basisrate** von Overreporting eine **deutliche Zunahme der *Positive Predictive Power*** der Validitätsscores einhergeht (Prozentzahl korrekt als auffällig klassifizierter Probanden). Gleichzeitig vermindert sich die *Negative Predictive Power* (Anzahl der korrekt als nicht auffällig klassifizierten Probanden) erst bei Annahme einer extrem hohen (unrealistischen) Basisrate ($\geq$ 80 %).

Der Zusammenhang zwischen möglichen Basisraten von Overreporting in einer Stichprobe und ihre (positive) Auswirkung auf die *Predictive Power* kann am Beispiel des *ROI-Index* für den MMPI-2-RF illustriert werden (s. Abb. 54, S. 234).

Wie in der Gegenüberstellung zu ersehen, erhöht sich die *Positive Predictive Power* des *ROI-Index* bei dem empfohlenen Cutoff-Wert $\geq$ 3 von 53 % in dem untersuchten klinischen Sample unter Annahme einer 30-prozentigen Basisrate in einem Sample von Gutachten-Probanden auf eine fast 90-prozentige (exakt: 87-prozentige) Detektions-Genauigkeit, und zwar bei 97-prozentiger Sicherheit, keinen unauffälligen Patienten falsch als auffällig zu klassifizieren.

Die beispielhafte Analyse unterstreicht zum einen eine besonders hohe Detektions-Sicherheit des *ROI-Index* zur Aufdeckung negativer Antwortverzerrungen bei Begutachtungen chronischer Schmerzpatienten.

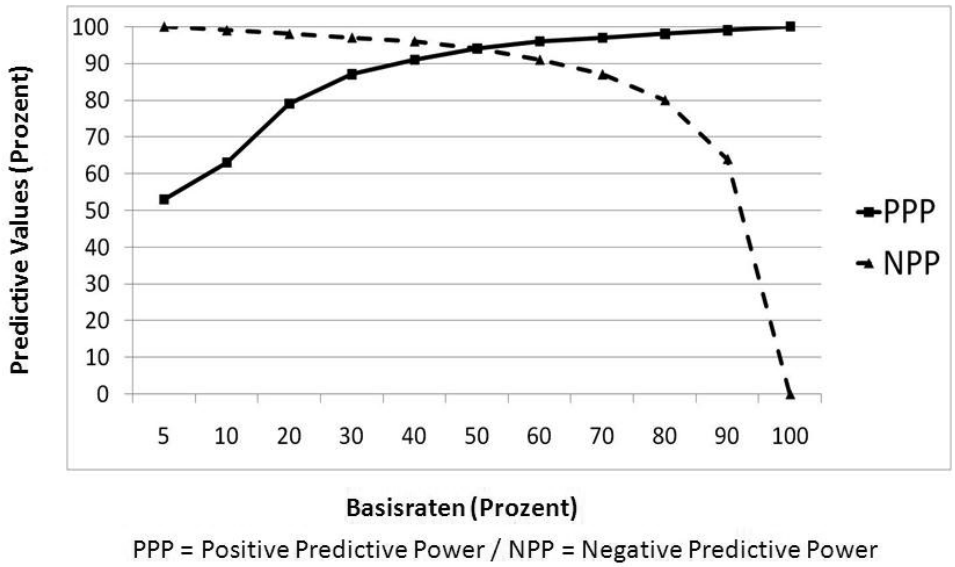


Abb. 54. Einfluss unterschiedlicher Basisraten auf die Prädiktion des **ROI-Index** (MMPI-2-RF) bei Cutoff ≥ 3

Diese kann in gleicher Weise für den MI-r-Index (PPP = 85 %, NPP = 97 %) und die F-r-Skala des MMPI-2-RF (PPP = 88 %, NPP = 98 %) angenommen werden (s. Tabellen 92 und 93 im Anhang, s. S. 364 und 365). Auch mittels der DIS-Skala des BHI-2 wären Patienten mit Overreporting in einer Stichprobe von Gutachtenpatienten mit hoher Sicherheit (PPP = 74 %, NPP = 95 %) korrekt zu identifizieren.

Zum zweiten ist somit trotz niedriger Basisrate in der klinischen Stichprobe der vorliegenden Studie von hinreichender Trennschärfe der Mehrzahl der getesteten Validitätsskalen auszugehen.

Als weitere Einschränkung einer Verallgemeinerung der Ergebnisse dieser Untersuchung muss erwähnt werden, dass alle Auswertungen auf den **US-amerikanischen T-Wert-Normen** des MMPI-2-RF und des BHI-2 beruhen. Deutsche Normwertvorgaben wurden bisher für keinen der beiden Fragebögen publiziert.

Weitere Folgestudien nach Veröffentlichung der beiden Testverfahren werden zeigen müssen, ob mittels deutscher Normdaten vergleichbar gute Detektions-Merkmale der Validitätsskalen und Basisskalen der beiden Fragebögen bestätigt werden können. Die Daten der vorliegenden Untersuchung sind gleichzeitig Teil einer großen Normierungs-Studie zum BHI-2, die in Kürze veröffentlicht wird (Dohrenbusch & Brockhaus 2016 [71], in Vorbereitung).

Erste Untersuchungen zur Akkuranz der Validitäts- und Basisskalen des MMPI-2-RF unter Verwendung der durch Herrn Prof. Engel (Mitherausgeber des MMPI-2 von Butcher et al. 1989 [48] sowie des deutsch übersetzten PAI-Fragebogens, Engel 2013 [85]) mitgeteilten deutschen Normdaten zeigten Detektionseigenschaften, die den Ergebnissen der vorliegenden Untersuchung entsprechen. Von weiterführenden Studien sind diesbezüglich detailliertere Resultate zu erwarten.

Andere methodische Einschränkungen der durchgeführten Studie betreffen die **Diskriminanzgüte der adaptierten Validitätsskalen** des MMPI-2, insbesondere ihrer Integration im Revised Overreporting Index (ROI).

Der Einsatz der Listen kritischer Items nach Lachar-Wrobel und nach Koss-Butcher basiert auf der Annahme, dass diese Items aufgrund ihrer Augenschein-Validität einen erhöhten Aufforderungs-Charakter haben, der Patienten mit Tendenz zu Beschwerdeüberhöhungen motiviert. Ähnlich wie bei den standardisierten Validitätsskalen des MMPI-2-RF können sich in überhöhten Antworten dieser Skalen psychischer Distress ebenso widerspiegeln wie negative Antwortverzerrungen, was ihre Validität zum Assessment von Overreporting einschränkt.

Zum zweiten kann unter methodischem Gesichtspunkt die Konstruktion des ROI-Index mittels des **Standardisierungsverfahrens** nach McCall (1939, [183], s. Lienert 1998 [171], Diehl & Kohr 1989 [67]) kritisch diskutiert werden. Dieses Verfahren wurde gewählt, um anhand der ermittelten Standardwerte geeignete Cutoff-Werte der Skalen zu ermitteln. Die ermittelten Werte spiegeln damit stichproben-spezifische Grenzwerte wider, die möglicherweise nur bedingt auf andere Populationen übertragbar sein könnten. So veränderten sich die standardisierten Grenzwerte im Verlauf der Untersuchung geringfügig je nach Stichprobegröße.

Andererseits hatten die mit der Stichprobengröße verbundenen Veränderungen keinen wesentlichen Einfluss auf das Gesamt-Ergebnis der vorliegenden Studie (hohe Diskriminanzgüte der Validitätsskalen des MMPI-2-RF und des BHI-2, einschließlich der Güte des ROI-Index), so dass bei der hier gewählten großen Stichprobe (n = 368) die Standardisierung als relativ stabil angesehen werden darf.

Aufgrund der Größe der untersuchten Stichprobe ist das Problem einer Stichproben-Spezifität vermutlich als marginal einzuordnen und es ist von einer weitgehenden Generalisierbarkeit der Resultate auf vergleichbare Populationen von Patienten mit chronischen Schmerzen auszugehen.

5 Theoriemodell über Genese und Assessment von Overreporting

Abschließend kann in Anlehnung an publizierte Modelle (Rogers 1984 [233], Rogers 1990 [234], Rogers et al. 1994 [247], Merckelbach & Merten 2012 [191]) ein die Ergebnisse der vorliegenden Studie integrierendes Theoriemodell zur Entstehung und Aufdeckung von Overreporting abgeleitet werden.

Dieses Modell stellt sich wie folgt dar (s. Abb. 55, S. 236):

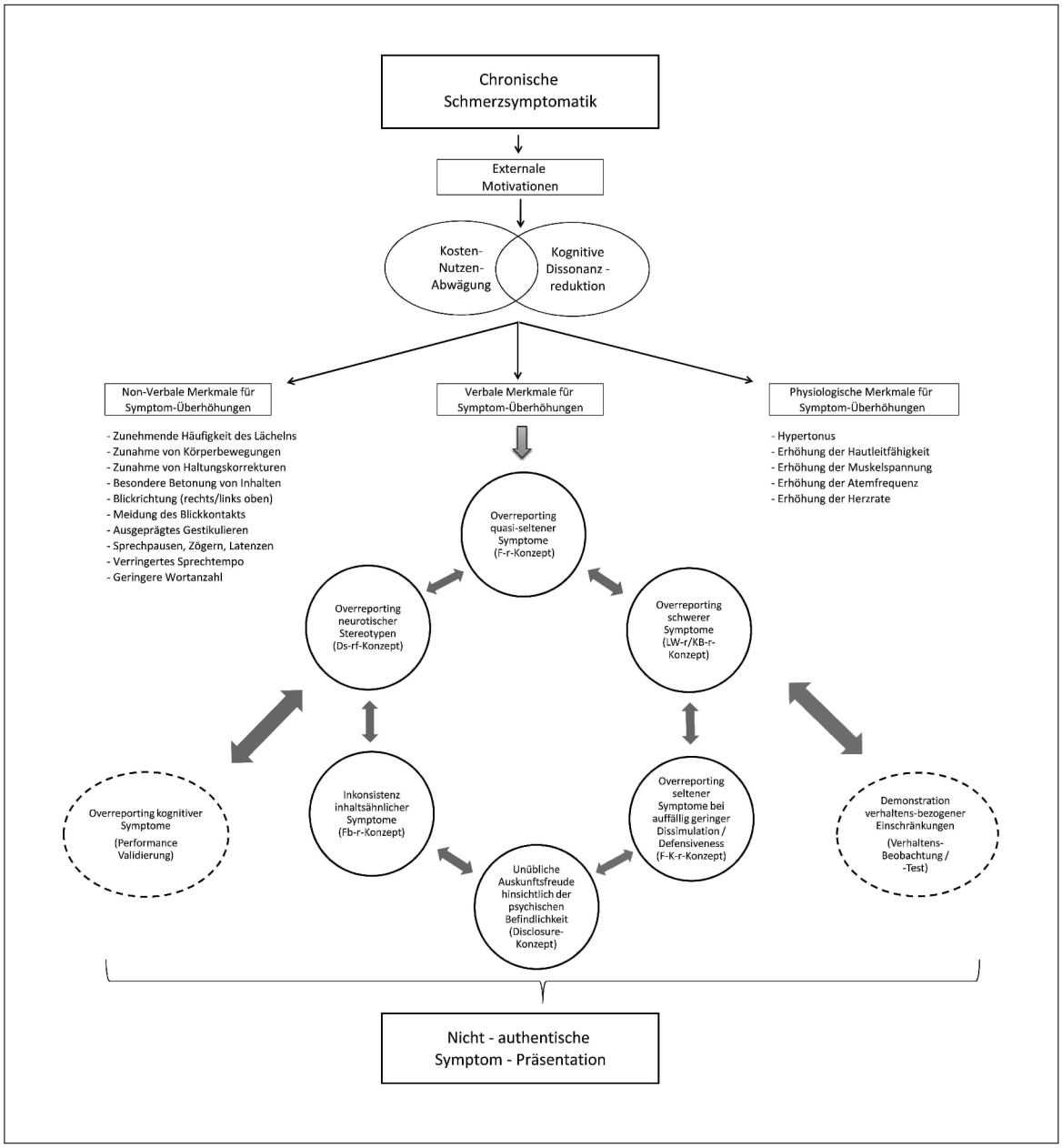


Abb. 55. Modell zu Genese und Detektion nicht-authentischer Beschwerde-Präsentation

Die Ausgestaltung einer chronischen Schmerzsymptomatik wird wie jedes Verhalten als kommunikativer Prozess zwischen Patient und anderen Menschen stets auch durch externale Motivationen beeinflusst. Diese werden verhaltenstheoretisch im Sinn positiver (z.B. Sympathie oder Unterstützung durch Behandler) und negativer Verstärkungen (Wegfall negativer Konsequenzen, z.B. durch Vermeidung unangenehmer Arbeitsbedingungen), durch Löschung (Wegfall positiver Konsequenzen) und durch negative Folgen verstärkt oder vermindert.

Wie in den Einleitungskapiteln dargelegt (z.B. Kap. 1.7, vgl. S. 29) können dabei externale Motivationen für Beschwerdeüberhöhungen in finanziell-existenziellen Kompensationen für eine schmerzbedingte Arbeits- oder Erwerbsunfähigkeit bestehen.

Die Vermeidung eines geringen Sozialhilfesatzes, beispielsweise wegen drohender Aussteuerung durch die Krankenkasse nach längerer Arbeitsunfähigkeit, Zuerkennung einer geminderten oder aufgehobenen Erwerbsfähigkeit können ebenso externale Motive für Overreporting darstellen wie der Wunsch nach Fortschreibung einer zeitbefristeten Rente vor Erreichen der Altersrente, nach einer Höherstufung eines Grades an Behinderung oder der Wunsch nach Zuerkennung eines Schmerzens- oder Verletztengeldes.

Andere externale Motive können sich auf eine individuell erwünschte Behandlung beziehen (z.B. Verschreibung von Opioid-Analgetika oder anderer abhängig machender Substanzen ohne zwingende Indikation, ein möglichst schneller Behandlungsbeginn) oder aber auf eine sozial anerkannte Ursachen-Zuschreibung der Beschwerden (als eine somatisch und nicht psychosomatisch begründete Symptomatik zur Vermeidung einer psychischen Stigmatisierung). Im Fall familiärer oder ehelicher Konflikte können Motivationen für Overreporting auch in einer gewünschten verstärkten Zuwendung oder Anteilnahme des Partners bestehen.

Im weitesten Sinn bezeichnen externale Motivationen für Beschwerdeüberhöhungen somit Faktoren eines sekundären Gewinns durch die Erkrankung. Nach Rogers et al. (Rogers 1984 [233], Rogers 1990 [234], Rogers et al. 1994 [247]) ist für diese Krankheits-Verarbeitungsprozesse ein rein pathogenes Erklärungsmodell, das die Motivation für Beschwerdeverzerrungen vornehmlich durch psychische Erkrankungen erklärt, wenig wahrscheinlich.

Nach diesem pathogenen Modell präsentiert ein Proband artifizielle Symptome zunächst, um eine Kontrolle über die tatsächlich vorhandenen psychischen Symptome zu erlangen. Wie Rogers (1990) [234] erläutert, wurde dieses sog. **pathogene Modell** zum einen weitgehend aufgegeben, weil viele Patienten eine entsprechende psychische Desorganisation nicht zeigen.

Zum zweiten veränderte sich das Krankheits-Verständnis in den letzten Dekaden von einem mehr psychiatrischen Krankheitsmodell hin zu behandelbaren, temporären Verhaltens- und Emotions-Veränderungen. Diese werden stets durch spezifische äußere Faktoren moduliert und ausgestaltet. Insofern muss ein Patient, der Beschwerdeüberhöhungen zeigt, nicht im umgangssprachlichen Sinn „verrückt“ sein, sondern kann auch ganz konkrete externale Gründe für sein Verhalten haben.

Auch das zweite, historisch präferierte **Erklärungsmodell** einer den Beschwerdeüberhöhungen zugrunde liegenden **forensischen psychopathologischen Veränderung** (z.B. als sog. anti-soziale Persönlichkeits-Störung) wurde in den letzten Jahren mangels einer Bestätigung dieses Grades an Störung in den Hintergrund gedrängt.

Die bei Patienten mit Persönlichkeits-Störungen festgeschriebene Non-Compliance (mangelndes Kooperationsverhalten) erwies sich als störungsspezifisch stark unterschiedlich (je nachdem, welche spezifischen Einbußen beklagt wurden), zum zweiten scheinen insbesondere Patienten mit Beschwerdeüberhöhungen, die besonders aufdeckungs-resistente Antwortverzerrungen nutzen, eher als besonders kooperativ.

Rogers 1990 [234] proklamiert deshalb ein wahrscheinlich zutreffenderes Modell für Beschwerdeüberhöhungen entsprechend einer individuellen Kosten-Nutzen-Analyse. Entsprechend dieser Analyse wird das Ausmaß an Beschwerdeüberhöhung insbesondere durch die Dringlichkeit der äußeren Bedingungen beeinflusst, zu denen der Betroffene keine tatsächlichen Handlungs-Alternativen sieht, sowie durch Kontext-Variablen der Behandlung oder Begutachtung, die der Betroffene als seinen Interessen entgegen gesetzt wahrnimmt, bestimmt.

Merckelbach & Merten (2012) [191] erklären mit Hilfe eines **Modells kognitiver Dissonanz**, warum manche Patienten an Schilderungen oder Darstellungen kognitiver, somatischer oder psychischer Symptome festhalten (mnestischer Defizite, Schmerz-Schilderungen oder beispielsweise depressiven Symptomen), obwohl die objektiven Befunde und Erhebungen offensichtlich keine entsprechenden Defizite belegen. Eine solche kognitive Verzerrung kann nach diesem Modell dadurch erklärt werden, dass Patienten Symptome so oft repetieren, dass sie selbst nach einer gewissen Zeit vom Vorhandensein dieser Symptome real überzeugt sind. Die Präsentation und Ausgestaltung dieser Symptome trägt dann zur intrinsischen Dissonanz-Reduktion und psychischen Stress-Bewältigung bei.

Eine zentrale Idee dieses Ansatzes ist, dass vorgetäuschte oder nicht-authentische Beschwerden durch den Prozess kognitiver Dissonanz internalisiert werden, so dass diese als real erlebt werden. Ein Phänomen dieses Prozesses besteht in einer Transition von äußerer Beschwerde-Zuschreibung zu mehr intrinsischem Beschwerdeerlebens. Dabei ist zu berücksichtigen, dass Beschwerdeüberhöhungen leicht durch Instruktionen verstärkt werden können. Bereits in früheren Simulations-Studien wurde festgestellt, dass instruierte Probanden eine klare Kenntnis über Symptome besitzen, die einfach verfälscht werden können, und auch über Symptome informiert sind, die eher schwierig vorzutäuschen und damit leicht aufzudecken sind (vgl. Kap. 1.9.6, S. 83, zur Verfälschbarkeit der MMPI-2-Skalen: z.B. Bagby et al. 2000 [22] hinsichtlich der Aufdeckung verfälschter Depressions-Symptome durch Experten mittels der Fb-Skala; Wiener 1948 [315] hinsichtlich häufigerer Erkennbarkeit versteckter Symptome in den O-S-Skalen durch gebildetere Probanden).

Neben externalen Motivationen beschreiben Merckelbach & Merten (2012) [191] als Aufrechterhaltungsfaktor der Internalisierung einer Krankenrolle in paradoxer Wirkung eine zunehmende Skepsis der Bezugspersonen (Familie, Mitarbeiter, Behandler) des Betroffenem hinsichtlich der Authentizität ihrer Beschwerden.

Andere Chronifizierungsmomente im Krankheitsverlauf bestehen in der Miss-Interpretation körperlicher Signale sowie Fehl-Informationen über körperliche Störungen durch scheinbar die Symptomatik bestätigende Befunde, die im Sinne eines Nocebo-Effektes krankheitsbezogene Überzeugungen bei den Betroffenen verstärken können. Angst könnte dabei ein weiterer modulierender Faktor im Rahmen der Ausgestaltung von Symptomen sein. Zudem verweisen Merckelbach & Merten (2012) [191] auf Studien, nach denen sich simulierende Probanden hinsichtlich der Schwierigkeit ihrer Arbeitstätigkeit als unterbezahlt empfinden und daraus möglicherweise implizit eine Berechtigung für häufigere Arbeitsausfallzeiten ableiten.

Malingering und kognitive Dissonanz führen dann vermutlich zur Überhöhung von psychischen Symptomen, die in der Allgemein-Bevölkerung als besonders selten, besonders bizarr und schwer eingestuft werden, möglicherweise weil auch bei körperlichen Beschwerden besonders schwere und auffällige Symptome als Merkmal für die Glaubwürdigkeit von Beschwerden angesehen werden. Ähnlich wie bei körperlichen Prozessen, bei denen besonders invasive Eingriffe, vielfache Diagnostika und teilweise exzessive Behandlungen meist eher zur Verschlechterung führen, wird vermutlich die besondere Aufmerksamkeit, die Patienten mit auffälligem Beschwerdeverhalten zuteil wird, zur Krankheitsaufrechterhaltung beitragen.

In ähnlicher Weise könnte in diesem Sinne auch die medikamentöse Verordnung starker Schmerzmittel (Opioid-Analgetika) paradox krankheitsaufrechterhaltend wirken, wie die aktuelle Diskussion um opioid-induzierte Schmerzen zeigt. Im Sinne eines Modells kognitiver Dissonanz könnte die Verordnung starker Schmerzmittel die Überzeugung von Patienten verstärken, an einer besonders schweren Erkrankung zu leiden. Entsprechend einem Modell kognitiver Dissonanz können auch die in der vorliegenden Studie beobachten Varianten negativer Antwortverzerrungen betrachtet werden.

Hinweise auf Beschwerdeüberhöhungen lassen sich entsprechend Studien zur Aufdeckung von Falschaussagen („Lügen“) und Täuschungen (Initiative zur Vermittlung einer falschen Information) durch spezifische non-verbale und auch spezifische physiologische Reaktionen detektieren (vgl. Rogers 1984 [233]).

Explizit erhoben oder implizit bemerkt, fließen diese Hinweise indirekt in jede Glaubwürdigkeits-Beurteilung von Aussagen ein. Die Eindeutigkeit und damit Akkuranz dieser Datenquellen wird jedoch kontrovers diskutiert. Trotz mancher Vorteile dieser Detektionsmethoden ist ihre Operationalisierbarkeit und Quantifizierung schwierig (s. Kontroverse zur Zulässigkeit des sog. Lügendetektors oder Polygraphen zur Beurteilung von Zeugenaussagen - als Beweismittel bislang nur zur Entlastung von Beschuldigten verwertbar, BGH 30.11.2010, AZ. 1 StR 509/10).

Leichter überprüfbare Merkmale für Overreporting sind aus den verbalen Beschwerdeschilderungen abzuleiten. Insbesondere präsentierte Symptome, die von Normalprobanden selten benannt werden (Infrequent-Response-F-r-Konzept), oder aber von Vergleichspersonen als zu schwergradig (LW-Konzept) oder zu offensichtlich neurotisch geprägt (Dissimulation-Ds-rf-Konzept) eingeschätzt werden, bilden Hinweise für nicht-authentische Darstellungen.

Hinsichtlich der Auffälligkeiten seltener Symptome ist einzuwenden, dass diese auch bei Patienten mit authentischen psychischen Störungen gehäuft vorkommen können, was auch für die Listen kritischer Items (nach Lachar-Wrobel / Koss-Butcher) gilt, obwohl 25 % psychiatrischer Patienten geringere kritische Angaben machen als Normalprobanden.

Dieser Prozentsatz an Underreporting unterscheidet die LW- und KB-Skalen nach Green (2008) [115] nicht von anderen Validitätsskalen. Dennoch bietet die geringe Itemüberlappung zwischen quasi-seltenen und kritischen Symptomen eine Möglichkeit der Kreuzvalidierung.

Beschwerdeüberhöhungen können sich auch in überhöht angegebenen Symptomen zeigen, die neurotische Stereotypen widerspiegeln (Ds-rf-Konzept). Dieses Konzept erwies sich in vorangehenden Studien als besonders resistent gegenüber Coaching, z.B. durch instruierte Simulanten. Rogers et al. (2011) [239] zeigten, dass Probanden mit nachgewiesenem Overreporting in externen Verfahren durch Überhöhungen solcher Stereotypen gegenüber Patienten mit genuinen psychischen Störungen identifizierbar waren.

Aber auch auffällige Inkonsistenzen zwischen inhaltlich übereinstimmenden Symptomen (Fb-r-Konzept), die meist auf fehlender Kenntnis allgemeiner Psychopathologie beruhen, oder auch Auffälligkeiten seltener Symptome bei auffallend verminderter Symptom-Bagatellisierung (F-K-r-Konzept), wie sie auch für authentische Patienten untypisch ist, bieten Hinweise auf nicht-authentische Beschwerdeschilderungen.

Weitere Hinweise sind aus dem Disclosure-(Verschlossenheits-)Konzept der BHI-2-Konzeption abzuleiten, demnach sich Beschwerdeüberhöhungen auch durch eine unübliche, überhöhte Freude und Bereitschaft von Probanden erkennen lassen, über ihre psychische Befindlichkeit bzw. das psychische Funktionieren Auskunft zu geben.

Um Schwächen einzelner Validierungsverfahren verbaler Beschwerdeüberhöhung auszugleichen, empfiehlt sich insbesondere die integrative, gewichtete Bewertung mittels der in der vorliegenden Untersuchung vorgestellten Summations-Scores (MI-r, ROI).

Gemeinsam mit Auffälligkeiten in kognitiven Leistungstests (Leistungsminderungen unter Zufallsniveau) sowie auffällig demonstrativem Beschwerdeverhalten (z.B. nicht nachvollziehbarem Schonhinken oder Finger-Boden-Abstand, insbesondere in Diskrepanz zu sonstigem Verhalten) bilden diese verbalen Merkmale von Overreporting das Zustandsbild einer nicht-authentischen Symptom-Präsentation.

6 Schlussfolgerungen und Empfehlungen für die Begutachtungs-Praxis

Zusammenfassend stellt der neu konzipierte, auf 338 Items verkürzte MMPI-2-RF gegenüber seiner Vorgängerversion eine ökonomische, trennscharfe Alternative für die Begutachtungspraxis wie auch das klinische Routine-Assessment von Schmerzpatienten mit mutmaßlichen Beschwerdeüberhöhungen dar.

Gegenüber der 567-Items umfassenden, für Patienten oft kaum zumutbaren Version des MMPI-2 besitzt der MMPI-2-RF eine deutlich homogenere Skalenstruktur mit zumindest gleichwertiger, teilweise sogar optimierter Diskriminanzstärke seiner Prüfskalen (insbesondere der Validitäts-Skalen F-r, Fp-r und FBS-r sowie des MI-r-Index).

Insbesondere die Validitätsskalen F-r (seltene somatische und psychische Symptome), RBS (zur Erfassung überhöhter Leistungsdefizite kognitiver Behinderungen) sowie der fünf Skalen integrierende MI-r bieten hoch trennscharfe Subskalen-Hinweise, negative Antwortverzerrungen zu verifizieren.

Im klinischen Alltag können mit diesen Skalen schnell Patienten mit chronischen Schmerzen mit überhöhten Beschwerdeangaben identifiziert werden und prognostische Vorhersagen über ihren wahrscheinlichen Therapieerfolg in einem medizinisch-behavioralen Behandlungs-Kontext getroffen werden.

Alternativ oder ergänzend können mit Hilfe des MMPI-2-RF auch die in der vorliegenden Studie untersuchten adaptierten Validität-Skalen sowie der fünf dieser Skalen integrierende ROI-Index zur Aufdeckung von Overreporting verwendet werden. Durch die nur marginalen Item-Überlappungen mit den Standard-Validitätsskalen bieten diese Subskalen in Begutachtungen zusätzliche Möglichkeiten einer Kreuz-Validierung der Subskalen-Auswertungen.

Die spezifische Konstruktion der in Europa kaum bekannten Battery of Health Improvement (BHI-2) im Hinblick auf die besonderen Problembereiche von Patienten mit chronischen Schmerzen empfiehlt, dieses noch kürzere Testverfahren (217 Items) als Kurz-Sreening wenn möglich im Routine-Assessment einzusetzen.

Die hier durchgeführten Untersuchungen unterstützen die Annahme, dass die im US-amerikanischen Raum häufig nur marginale Erwähnung des BHI-2 zur Overreporting-Detektion (z.B. Rogers 2008 [235]) seinen Qualitäten nicht gerecht wird.

Insbesondere die Disclosure-Validitätsskala, aber auch einige der BHI-2-Basisskalen (Depressivität, Somatisierung, Feindseligkeit, Angst und Störungen der emotionalen Stabilität) bieten dem MMPI-2-RF vergleichbare, trennscharfe Alternativen, negative Antwortverzerrungen aufzudecken.

In algesiologischen Gutachten bietet der Einsatz beider Testverfahren zusätzlich die Möglichkeit einer kreuzvalidierten Absicherung festgelegter Auffälligkeiten im Antwortverhalten einzelner Schmerzpatienten.

Dem Anwender beider Testverfahren ist zu empfehlen, zum Zweck einer maximalen Sicherheit bei der Aufdeckung negativer Antwortverzerrungen, die Ergebnisse von Validitätsskalen *möglichst geringer Item-Überlappung* miteinander zu vergleichen.

Bei Anwendung beider Fragebögen, beispielsweise in Begutachtungen, könnten insbesondere die Resultate der F-r-Skala des MMPI-2-RF mit den Ergebnissen der BHI-2-Disclosure-Skala einer unabhängigen Absicherung der Befunde dienen. Bei alleiniger Abfrage des MMPI-2-RF empfiehlt sich, insbesondere Ergebnisse der F-r-Skala mit den Listen kritischer Items LW-r und KB-r abzugleichen.

Die lange LW-r-Skala (87 Items) teilt mit der F-r-Skala (32 Items) nur drei positiv beantwortete Items (2,5 % der Items: Item 72, 232 und 311), so dass sich beide Skalen als unabhängige Indikatoren eignen.

Die kürzere KB-r-Skala (59 Items) teilt mit der F-r-Skala (32 Items) ebenfalls nur drei positiv beantwortete Items (3,3 % der Items: Item 14, 121 und 232), so dass sich beide Skalen ebenso als unabhängige Indikatoren eignen.

Bei beiden, im deutschen Sprachraum bisher kaum bekannten Verfahren (MMPI-2-RF und BHI-2) ist in der nächsten Zukunft zu erwarten, dass nach Publikation von Normierungsdaten einer deutschen Population eine Anpassung der hier vorgestellten Grenzwerte für die Basis- und Validitätsskalen erfolgen wird. Anhand dieser Daten kann in Folgestudien unter Verwendung von BV-Verfahren höherer Trennschärfe eine Verfeinerung der hier untersuchten Validierungs-Ansätze erfolgen.

Es ist zu hoffen, dass die hier vorgestellten neuen und adaptierten, multimodalen Methoden der Beschwerdenvalidierung solche Folgestudien anregen und inspirieren werden.

Abschließend muss betont werden, dass valide Klassifikationen und gutachterliche Einschätzungen zur schmerzassoziierten Beschwerdenüberhöhung nie auf einer einzigen Informationsquelle, keinem einzelnen Test oder einer einzigen Score-Prozedur basieren sollten.

Valide und reliable gutachterliche Aussagen benötigen vielmehr eine komplexe Integration multipler diagnostischer Verfahren, zu dem das klinische Assessment ebenso gehört, wie die Durchführung und Resultate unterschiedlicher Vorbehandlungen, Erhebungen zur Compliance, eine physisch-somatische, medizinische Untersuchung, Informationen wichtiger Bezugspersonen des sozialen Umfelds des Patienten (Fremdanamnese), eine umfassende psychologische Exploration, die objektive Erfassung des kognitiven und verhaltensbezogenen Leistungsbemühens und andere Datenquellen.

Insofern empfiehlt sich, insbesondere in Begutachtungs-Kontexten, in denen die Tragweite des Untersuchungsresultates für den einzelnen Patienten von weitreichender, auch finanziell-existentieller Bedeutung ist, stets ein mehrdimensionales Assessment einschließlich der Verhaltensbeobachtungen unter Einbezug der drei von Bianchini et al. (2005) [31] genannten Ebenen.

Literatur

[1] Aamodt, M., Dwight, S. & Surrette, M. (1996). Incremental validity of MMPI and MMPI-2 clinical scales in detecting malingering. *Journal of Police and Criminal Psychology*, 12,2, 42-47

[2] Aguerrevere, L. (2010). *Multivariate cluster analysis of the MMPI-2 and MMPI-2-RF scales in spine pain patients with financial compensation: Characterization and validation of chronic pain subgroups.* University of New Orleans: Dissertation

[3] Aguerrevere, L., Greve, K., Bianchini, K. & Meyers, J. (2008). Detecting malingering in traumatic brain injury and chronic pain with an abbreviated version of the Meyers Index for the MMPI-2. *Archives of Clinical Neuropsychology*, 23, 831-838

[4] Ahola, K., Gould, R., Virtanen, M., Honkonen, T., Aromaa, A. & Lönnqvist, J. (2009). Occupational burnout as a predictor of disability pension: A population based cohort study. *Occupational Environment and Medicine*, 66, 284-290

[5] Alison, L., Brauer-Boone, K., Lesser, I., Wohl, M., Wilkins, S. & Parks, C. (2000). Performance of older depressed patients on two cognitive malingering tests: false positive rates for the Rey 15-item Memorization and Dot Counting Tests. *The Clinical Neuropsychologist*, 14,3, 303-308

[6] Allen, L., Conder, R., Green, P. & Cox, D. (1997). *CARB 97. Manual for the Computerized Assessment of Response Bias*. Durham, NC: CogniSystems

[7] Allen, L. & Green, P. (2001). Declining Carb failure rates over 6 years of testing: What′s wrong with this picture. *Archives of Clinical Neuropsychology*, 16, 897-862

[8] Alwes, Y. (2006). *The utility of the Structured Inventory of Malingered Symptomatology as a screen for the feigning of neurocognitive deficit and psychopathology in a civil forensic sample.* University of Kentucky: Master's thesis

[9] Anderson, J. (2011). *A multi-method assessment approach to the detection of malingered pain: Association with the MMPI-2 Restructured Form.* Eastern Kentucky University: Master's thesis

[10] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (2004). *Leitlinie für die Begutachtung von Schmerzen.* AWMF-Leitlinien-Reg. Nr. 030/102

[11] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (2011). *Leitlinie zur Begutachtung psychischer und psychosomatischer Erkrankungen*. AWMF-Leitlinien-Reg. Nr. 051/029

[12] Arbisi, P. & Ben-Porath, YS (1995). An MMPI-2 Infrequent Response Scale for use with psychopathological populations: the Infrequency-Psychopathology Scale F(p). *Psychological Assessment*, 7, 424-431

[13] Arbisi P. & Ben-Porath, Y. (1998) The ability of Minnesota Multiphasic Personality Inventory-2 validity scales to detect fake-bad responses in psychiatric inpatients. *Psychological Assessment* 10: 221-228

[14] Baer, R. & Miller, J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment*, 14,1, 16-26

[15] Bagby, R., Buis, T. & Nicholson, R. (1995). Relative effectiveness of the standard validity scales in detecting fake-bad and fake-good responding: Replication and extension. *Psychological Assessment*, 7,1, 84-92

[16] Bagby, R., Marshall, M. & Bacchiochi, J. (2005). The validity and clinical utility of the MMPI-2 Malingering Depression Scale. *Assessment*, 10, 382-292

[17] Bagby, R. & MB (2004). Assessing underreporting response bias on the MMPI-2. *Assessment*, 11, 115-126

[18] Bagby, R., Nicholson, R., Bacchiochi, J., Ryder, A. & Bury, A. (2002). The predictive capacity of the MMPI-2 and PAI validity scales and indexes to detect coached and uncoached feigning. *Journal of Personality Assessment*, 78, 69-86

[19] Bagby, R., Parker, J. & Taylor, G. (1994a). The twenty-item Toronto Alexithymia Scale I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, 38, 1, 23-32

[20] Bagby, R., Rogers, R., Buis, T. & Kalemba, V. (1994b). Malingered and defensive response styles on the MMPI-2: An examination of validity scales. *Assessment*, 1, 31-38

[21] Bagby, R., Rogers, R., Nicholson, R. & Buis, T. (1997). effectiveness of the MMPI-2 validity indicators in the detection of defensive responding in clinical and nonclinical samples. *Psychological Assessment*, 9,4, 406-413

[22] Bagby, R.M., Nicholson, R.A., Buis, T. & Bacchiochi, J.R. (2000). Can the MMPI-2 validity scales detect depression feigned by experts? *Assessment*, 7,1, 55-62

[23] Barbeau, E., Didic, M., Tramoni, E., Felician, O., Joubert, S., Sontheimer, A., Ceccaldi, M. & Poncet, M. (2004). Evaluation of visual recognition memory in MCI patients. *Neurology*, 62, 1317-1322

[24] Bauer, L. & McCaffrey, R. (2005). Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and Word Memory Test on the internet: Is test security threatened? *Archives of Clinical Neuropsychology*, 21, 121-126

[25] Baumann, U. & Stieglitz, R. (1983). *Testmanual zum AMDP - System (Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie). Empirische Studien zur Psychopathologie*. Berlin: Springer.

[26] Bechtol, C. (1954). Grip Test: The Use of a dynanamometer with adjustable handle spacings. *The Journal of Bone and Joint Surgery*, 36A, 820-824

[27] Ben-Porath, Y. (2011). MMPI-2-RF Medical Webinar 09-2011. **http://www.pearsonassessments.com/hai/images/PDF/Webinar/MMPI-2-RF_medical_webinar_09-2011.pdf,** Stand: 02.05.2013

[28] Ben-Porath, Y., Greve, K., Bianchini, K. & Kaufmann, P. (2009). The MMPI-2 symptom validity scale (FBS) is an empirically validated measure of overreporting in personal injury ligants and claimants: Reply to Butcher et al. (2008). *Psychology, Injury and Law*, 2, 62-85

[29] Ben-Porath, Y. & Tellegen, A. (2009). *Minnesota Multiphasic Personality Inventory-2-RF (MMPI-2-RF): Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press

[30] Bianchini, K., Etherton, J.L., Greve, K., Heinly, M. & Meyers, J.E. (2008). Classification accuracy of MMPI-2-validity scales in the detection of pain-related malingering: A known-groups study. *Assessment*, 15, 435-449

[31] Bianchini, K., Greve, K. & Glynn, G. (2005). On the diagnosis of malingered pain-related disability: Lessons from cognitive malingering research. *Spine Journal*, 5, 404-417

[32] Birke, K., Schneider, W., Klauer, T. & Dobreff, U. (2001). *Wie beeinträchtigt in psychosomatisch relevanten Dimensionen sind Gutachtenpatienten wirklich? Ein Vergleich zwischen stationären Psychotherapiepatienten und Probanden in Sozialgerichtsverfahren.* In: W. Schneider, P. Henning sen & U Rüger (Eds.) *Sozialmedizinische Begutachtung in Psychosomatik und Psychotherapie*, Bern: Huber, 195-224.

[33] Blanchard, D., McGrath, R., Pogge, D. & Khadivi, A. (2003). A comparison of the PAI and MMPI-2 as predictors of faking bad in college students. *Journal of Personality Assessment*, 1980, 197-205

[34] Block, A. (1981). An investigation of the response of the spouses to chronic pain behavior. *Psychosomatic Medicine*, 4, 425-432

[35] Blyth, F., March, L., Nicholas, M. & Cousins, M. (2003). Chronic pain, work performance and litigation. *Pain*, 103, 41-47

[36] Bonica, J. (1953). *The Management of Pain*. Philadelphia: Lea & Febinger.

[37] Boone, K., Salazar, X., Lu, P., Warner-Chacon, K. & Razani, J. (2002). The Rey 15-item recognition trial: A technique to enhance sensitivity of the rey 15-item Recognition Memorization Test. *Journal of Clinical and Experimental Neuropsychology*, 24, 561-573

[38] Bortz, J. (1993). *Statistik für Sozialwissenschaftler*. Berlin: Springer, 4. Auflage

[39] Brickenkamp, R. (1962). *Aufmerksamkeits-Belastungs-Test d2*. Göttingen: Hogrefe

[40] Brickenkamp, R., Schmidt-Atzert, R. & Liepmann, D. (2010). *Aufmerksamkeits- und Konzentrationstest d2-R*. Göttingen: Hogrefe

[41] Brockhaus, R. & Merten, T. (2004). Neuropsychologische Diagnostik suboptimalen Leistungsverhaltens mit dem Word Memory Test. *Der Nervenarzt*, 75, 9, 882-887

[42] Bruns, D. & Disorbio, J. (2000). A comparison of two BHI measures of faking. In: *Paper Presentation to the American Psychological Association 2000 National Convention*.

[43] Bruns, D. & Disorbio, J. (2004). *Battery for Health Improvement (BHI-2)*. London: Pearsson

[44] Bundespsychotherapeutenkammer (2014). BPTK-Studie zur Arbeits- und Erwerbsfähigkeit. *BPTK-Newsletter*, 1/2014, 2

[45] Burchess, D. & Ben-Porath, Y. (2010). The impact of overreporting on MMPI-2-RF substantive scale score validity. *Assessment*, 17, 4, 497-516

[46] Burton, A. (1997). Spine update. Back injury and work loss: biomechanical and psycho-social influences. *Spine*, 21, 2575-2580

[47] Butcher, J., Arbisi, P., Atlis, M. & McNulty, J. (2003). *The construct validity of the Lees-Haley Fake Bad Scale. Does this scale measure somatic malingering and feigned emotional distress? Archives of Clinical Neuropsychology*, 18, 473-485

[48] Butcher, J., Dahlstrom, W., Engel, R., Graham, J., Tellegen, A. & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis: University of Minnesota Press

[49] Butcher, J., Dahlstrom, W., Engel, R., Graham, J., Tellegen, A. & Kaemmer, B. (2000). *Multiphasic Personality Inventory 2. Deutsche Fassung*. Göttingen: Hogrefe-Verlag

[50] Butcher, J., Hamilton, C., Rouse, S. & Cumella, E. (2006). The deconstruction of the HY scale of MMPI-2: Failure of RC3 in measuring somatic symptom expression. *Journal of Personality Assessment*, 87, 186-192

[51] Butcher, J. & Han, K. (1995). Development of an MMPI-2 scale to assess the presentation of self in a superlative manner: The S-scale. newblock In: J.N. Butcher & C.D. Spielberger (Eds.) *Advances in personality assessment 10*. New York: Erlbaum, 25-50

[52] Cashel, M., Rogers, R., Sewell, K. & Martin-Cannici, C. (1995). The Personality Assessment Inventory (PAI) and the detection of defensiveness. *Assessment*, 2, 333-342

[53] Cassidy, J., Carroll, L., Cote, P., Lemstra, M., Berglund, A. & Nygren, A. (2000). Effect of eliminating compensation for pain and suffering on the outcome of insurance claims for whiplash injury. *New England Journal of Medicine*, 342, 1179-1186

[54] Cima, M., Hollnack, S., Kremer, K., Knauer, K., Schellbach-Matties, R., Klein, B. & Merckelbach, H. (2003a). Strukturierter Fragebogen Simulierter Symptome. Die Deutsche Version des „Structured Inventory of Malingered Symptomatology: SIMS“. *Der Nervenarzt*, 74, 11, 977-986

[55] Cima, M., Merckelbach, H., Hollnack, S., Butt, S., Kremer, K., Schellbach-Matties, R. & Muris, P. (2003). Other side of malingering: supernormality. *The Clinical Neuropsychologist*, 17, 235-243

[56] Cofer, N. (1949). *The response of chronic pain patients to the original and the revised versions of the Minnesota Multiphasic Personality Inventory*. University of Minnesota

[57] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd ed.

[58] Costa, P. & McRae, R. (1985). Hypochondriasis, neuroticism and aging: When are somatic complaints unfounded? *American Psychologist*, 40, 19-28

[59] Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334

[60] Cronbach, L. (1990). *Essentials of psychological testing*. Harper & Row, 5th ed., 539 pp.

[61] Currie, S. & Wang, J. (2004). Chronic back pain and major depression in the general canadian population. *Pain*, 107, 54-60

[62] Dean, R. (1982). *Neuropsychological symptom inventory*. St. Louis: Washington University, School of Medicine

[63] DeLong, E., DeLong, D. & Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837-845

[64] Derogatis, L. (1992). *SCL-90-R: Administration, scoring and procedures manual II for the revised version*. Towson: Clinical Psychometric Research

[65] Deutsche Rentenversicherung Bund (2011). *Rentenversicherung in Zahlen 2011. Aktuelle Ergebnisse, Stand: 31. Mai 2011, Entwicklung der Daten bis heute*. Berlin: Heenemann

[66] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) (2015). *Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme. ICD-10. Revision, Deutsche Modifikation.* **http://www.dimdi.de/static/de/klassi/icd-10-gm/kodesuche/onlinefassungen/htmlgm2015/** Stand: 11.09.2016

[67] Diehl, J. & Kohr, H. (1989). *Deskriptive Statistik.* Magdeburg: Klotz

[68] Dirks, J., Wunder, J., Kinsman, R., McElhinny & Jones, N. (1993). A pain rating scale and a pain behavior checklist for clinical use: development, norms, and the consistency score. *Psychotherapy and Psychosomatics*, 59, 41-49

[69] Dohrenbusch, R. (2009). Symptom- und Beschwerdevalidierung chronifizierter Schmerzen in sozial-medizinischer Begutachtung. Teil I: Terminologische und methodologische Zugänge. *Der Schmerz*, 23,3, 231-240

[70] Dohrenbusch, R. (2011). Zur Risikobewertung der Berufsunfähigkeit aufgrund psychischer Erkrankung. *Versicherungsmedizin*, 63,1, 25-32

[71] Dohrenbusch, R. & Brockhaus, R. (2016). *Deutsche Version der Battery for Health Improvement (BHI-2) von D. Bruns & J.M. Disorbio*. Frankfurt: Pearson, in prep

[72] Dohrenbusch, R., Nilges, P. & Traue, H. (2008). Leitlinie für die Begutachtung von Schmerzen (Kommentar). *Psychotherapie*, 53, 63-68

[73] Dohrenbusch, R. & Pielsticker, A. (2011). Begutachtung von Personen mit chronischen Schmerzen. In: B. Kröner-Herwig, J. Frettlöh, R. Klinger & P. Nilges (Hrsg.) *Schmerzpsychotherapie*, Heidelberg: Springer, 335-356

[74] Downing, S., Denney, R., Spray, B., Houston, C. & Halfaker, D. (2008). Examining the relationship between the Restructured Scales and the Fake Bad Scale of the MMPI-2. *The Clinical Neuropsychologist*, 22, 680-688

[75] Dressing, H., Foerster, K., Widder, B., Schneider, F., & Falkai, P. (2011). Zur Anwendung von Beschwerdenvalidierungstests in der psychiatrischen Begutachtung. Stellungnahme der Deutschen Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN). Stellungnahme Nr. 03 / 28.01.2011. *Nervenarzt*, 82, 379-390. **http://www.dgppn.de/presse/stellungnahmen/detailansicht/browse/1/select/stellungnahmen-2011/article/141/zur-anwendun.html** Stand: 02.05.2016

[76] Dunn, T., Shear, P., Howe, S. & Ris, M. (2003). Detecting neuropsychological malingering: effects of coaching and information. *Archives of Clinical Neuropsychology*, 18, 121-134

[77] Dush, D., Simons, L., Platt, M., Nation, P. & Ayres, S. (1994). Psychological profiles distinguishing litigating and nonlitigating pain patients: subtle, and not so subtle. *Journal of Personality Assessment*, 62, 299-313

[78] Eberl, A. & Wilhelm, H. (2007). *Der Aggravations- und Simulationstest AST*. Lüdenscheid: Mnemo-Verlag

[79] Eberl, A., Heusler, M., & Schimrigk, S. (2008). *Detection of malingering in prepared and unprepared experimental simulators.* Vortrag und Poster auf dem 29. International Congress of Psychology, Berlin

[80] Edens, J., Otto, R. & Dwyer, T. (1999). Utility of the Structured Inventory of Malingered Symptomatology in identifying persons motivated to malinger psychopathology. *Journal of the American Academy of Psychiatry and Law*, 27,3, 387-396

[81] Edens, J., Poythress, N. & Watkins-Clay, M. (2007). Detection of malingering in psychiatric unit and general population prison inmates: a comparison of the PAI, SIMS, and SIRS. *Journal of Personality Assessment*, 88,1, 33-42

[82] Edwards, A. (1957). *The social desirability variable in personality asessment and research*. New York: Dryden

[83] Egle, U.T. & Hoffmann, S. (1993). Transkulturelle Aspekte von Schmerzerleben. In: U.T. Egle & S. Hoffmann (Hrsg.) *Der Schmerzkranke*. Stuttgart: Schattauer, 28 - 41

[84] Elhai, J., Naifeh, J., Zucker, L., Gold, S., Deitsch, S. & Frueh, B. (2004). Discriminating malingered from genuine civilian posttraumatic stress disorder. A validation of three MMPI-2 infrequency scales (F, Fp and Fptsd). *Assessment*, 11, 139-144

[85] Engel, R. (2013). *Verhaltens- und Erlebensinventar (VEI). Deutschsprachige Adaptation des Personality Assessment Inventory (PAI) von L.C. Morey*. Bern: Huber

[86] Ewald, S. & Kohler, U. (1991). Handkraft: Richtwerte für Erwachsene. *Ergotherapie*, 9, 4-11

[87] Fahrenberg, J., Hampel, R. & Selg, H. (2001). *Freiburger Persönlichkeitsinventar. Revidierte Fassung. (FPI-R).* Göttingen: Hogrefe, 8. Auflage, 2010

[88] Fishbain, D., Cutler, R., Rosomoff, H. & Rosomoff, R. (1997). Chronic pain-associated depression: antecedent or consequence of chronic pain? A review. *Clinical Journal of Pain*, 13, 116-137

[89] Fishbain, D., Cutler, R., Rosomoff, H. & Rosomoff, R. (1999). Chronic pain disability exaggeration/malingering and submaximal effort research. *Clinical Journal of Pain*, 15, 244-277

[90] Fitts, W. (1965). *Manual for the Tennessee Self Concept Scale*. Nashville, TN: Counselor Recordings and Tests

[91] Flor, H., Kerns, R. & Turk, D. (1987). The role of spouse reinforcement, perceived pain and activity levels of chronic pain patients. *Journal of Psychosomatic Research*, 31, 251-259

[92] Flor, H., Turk, D. & Rudy, T. (1989). Relationship of pain impact and significant other reinforcement of pain behavioral. The mediating role of gender, marital status and marital satisfaction. *Pain*, 38, 45-50

[93] Fordyce, W. (1976). *Behavioral method for chronic pain and illness*. St. Louis: Mosby.

[94] Franz, C., Paul, R., Bautz, M., Choroba, G. & Hildebrandt, J. (1986). Psychosomatic aspects of chronic pain: a new way of description based on MMPI item analysis. *Pain*, 24, 33-43

[95] Fuchs, S., Endler, P., Mesenholt, E., Paß, P. & Frass, M. (2009). Burnout bei niedergelassenen Ärztinnen und Ärzten für Allgemeinmedizin. *Wiener Medizinische Wochenschrift*, 159,7-8, 188-191

[96] Gass, C. & Odland, A. (2012). Minnesota Multiphasic Personality Inventory-2 Revised form symptom validity scale-revised (MMPI-2-RF FBS-r; also known as fake bad scale): Psychometric characteristics in a nonlitigation neuropsychological setting. *Journal of Clinical Experimental Neuropsychology*, 34, 6, 561-570

[97] Geissner, E. & Jungnitsch, G. (1992). *Psychologie des Schmerzes*. Weinheim: Beltz

[98] Geissner, E. & Schulte, A. (1996). *Schmerzempfindungs-Skala*. Göttingen: Hogrefe

[99] Gervais, R., Ben-Porath, Y., Wygant, D. & Green, P. (2007). Development and validation of a Response Bias Scale (RBS) for the MMPI-2. *Assessment*, 14, 196-208

[100] Gervais, R., Ben-Porath, Y., Wygant, D. & Green, P. (2008). Differential sensitivity of the Response Bias Scale (RBS) and MMPI-2 validity scales to memory complaints. *The Clinical Neuropsycholgist*, 22, 1-19

[101] Gervais, R., Ben-Porath, Y., Wygant, D. & Sellbom, M. (2010). Incremental validity of the MMPI-2-RF overreporting scales and RBS in assessing the veracity of memory complaints. *Archives of Clinical Neuropsychology*, 25, 274-284

[102] Giger, P., Merten, T., Merckelbach, H. & Oswald, M. (2010). Willentliche Testverfälschung bei Verfahren zur Erfassung von Dissoziation: Ergebnisse einer Begutachtungsstudie. *Praxis der Rechtspsychologie*, 20,1, 131-147

[103] Gillard, N. (2010). *Methodological issues in malingering research: The use of simulation designs*. University of North Texas: Master's thesis

[104] Glaros, A. & Kline, R. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specifity, and predictive value model. *Journal of Clinical Psychology*, 44,6, 1013-1023

[105] Gough, H. (1950). The F minus K dissimulation index for the MMPI. *Journal of Consulting Psychology*, 14, 408-413

[106] Gough, H. (1954). Some common misconceptions about neuroticism. *Journal of Consulting Psychology*, 18, 287-292

[107] Gough, H. (1957). *California Psychological Inventory Manual*. Palo Alto, California: Consulting Psychologists Press

[108] Graham, J., Watts, D. & Timbrook, R. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment*, 57, 264-277

[109] Gralow, I. (2000). Pychosoziale Risikofaktoren in der Chronifizierung von Rückenschmerzen. *Schmerz*, 14, 104-110

[110] Green, P. (2000). *MMPI-2: An interpretive manual*. Needham-Heights: Allyn & Bacon, 2nd Edition

[111] Green, P. (2003). *Green's Word Memory Test. User's manual.* Edmonton /Canada: Green's Publishing.

[112] Green, P. (2004). *Medical Symptom Validity Test (MSVT) for Microsoft Windows. User's manual*. Edmonton /Canada: Green's Publishing.

[113] Green, P. (2011). *MMPI-2: An interpretive manual*. Needham-Heights: Allyn & Bacon, 3rd Edition

[114] Green, P., Iverson, G. & Allen, L. (1999). Detecting malingering in head injury litigation with the Word Memory Test. *Brain Injury*, 13,10, 813-819

[115] Green, R. (2008). *Malingering and Defensiveness on the MMPI-2.* In: R. Rogers (Ed.) *Clinical assessment of malingering and perception*. New York: Guilford, 159-181

[116] Greiffenstein, M., Gervais, R., Baker, W., Artiola, L. & Smith, H. (2013). Symptom validity testing in medically unexplained pain: A chronic regional pain syndrome type I. Case series. *The Clinical Neuropsychologist*, 27, 138-147

[117] Greve, K. & Bianchini, K. (2004). Setting Empirical cut-offs on psychometric indicators of negative response bias: A methodological commentary with recommendations. *Archives of Clinical Neuropsychology*, 19, 533-541

[118] Greve, K., Bianchini, K., Love, J., Brennan, A. & Heinly, M. (2006). Sensitivity and specificity of MMPI-2 Validity scales and indicators to malingered neurocognitive dysfunction in traumatic brain injury. *The Clinical Neuropsychologist*, 20, 491-512

[119] Groves, J. (2009). *Untersuchungen zur Konstruktvalidität des Verhaltens- und Erlebensinventars (VEI) an einer klinischen Stichprobe*. Ludwig-Maximilians-Universität München: Dissertation.

[120] Guedj, E., Barbeau, E., Didic, M., Felician, O., de Laforte, C., Ceccaldi, M., Mundler, O. & Poncet, M. (2006). Identification of subgroups in amnestic mild cognitive impairment. *Neurology*, 67, 356-358

[121] Handel, R., Ben-Porath, Y., Tellegen, A. & Archer, R. (2010). Psychometric functioning of the MMPI-2-RF VRIN-r and TRIN-r scales with varying degrees of randomness, acquiescence, and counter-acquiescence. *Psychological Assessment*, 22, 1, 87-95

[122] Hanley, J.A. & McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148, 839-843

[123] Hardt, J., Gerbershagen, H. & Franke, P. (2000). The Symptom Checklist SCL-90-R: Its use and characteristics in chronic pain patients. *European Journal of Pain*, 4, 137-148

[124] Hare, R. & Neumann, C. (2009). Psychopathy: Assessment and forensic implications. *Canadian Journal of Psychiatry*, 54, 791-802

[125] Hasenbring, M. (1993). Durchhaltestrategien - Ein in der Schmerzforschung und Therapie vernachlässigtes Phänomen? *Schmerz*, 7, 4, 304-313

[126] Häuser, W. (2002a). Gibt es eine Schmerzkrankheit? Medizinische und Psychosoziale Charakteristika von Probanden mit chronischen Schmerzsyndromen in der Sozialgerichtsbarkeit. *Medizinischer Sachverständiger*, 4, 120-126

[127] Häuser, W. (2002b). Parameter zur Beurteilung der Prognose von Probanden mit chronischen Schmerzsyndromen in der sozialgerichtlichen Begutachtung. *Medizinischer Sachverständiger*, 98, 157-160

[128] Heinze, M. & Purisch, A. (2001). Beneath the mask: Use of psychological tests to detect and subtype malingering in criminal defendants. *Journal of Forensic Psychology Practice*, 1, 23-52

[129] Helmes, E. & Reddon, J. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin*, 113, 453-471

[130] Hempel, V. (1993). *Zoster und postherpetische Neuralgie.* In: M. Zenz & I. Jurna (Hrsg.). *Lehrbuch der Schmerztherapie.* Wissenschaftliche Verlagsgesellschaft Stuttgart, 453-457

[131] Henry, G., Heilbronner, R., Algina, J. & Kaya, Y. (2012). Derivation of the MMPI-2-RF Henry-Heilbronner Index-r (HHI-r) scale. *The Clinical Neuropsychologist*, 1-7

[132] Henry, G., Heilbronner, R., Mittenberg, W., Enders, C. & Roberts, D. (2006). The Henry-Heilbronner-Index: A 15-item empirical derived MMPI-2 subscale for identifying probable malingering in personal injury ligants and disability claimants. *The Clinical Neuropsychologist*, 20, 786-797

[133] Henry, G., Heilbronner, R., Mittenberg, W., Enders, C. & Roberts, D. (2008). Empirical derivation of a new MMPI-2 Scale for identifying probably malingering in personal injury ligants and disability claimants: The 15-item Malingered Mood Disorder Scale (MMDS). *The Clinical Neuropsychologist*, 22, 158-168

[134] Herrmann-Lingen, C., Buss, U. & Snaith, R. (1995). *Hospital Anxiety and Depression Scale*. Bern: Huber

[135] Hildebrandt, J., Pfingsten, M., Franz, C., Saur, P. & Seeger, D. (1996). Das Göttinger Rücken Intensiv Programm (GRIP). Ein multimodales Behandlungsprogramm für Patienten mit chronischen Rückenschmerzen, Teil 1. *Der Schmerz*, 10,4, 190-203

[136] Hilsabeck, R., LeCompte, D., Marks, A. & Grafman, J. (2001). The Word Completition Memory Test (WCMT): A new test to detect malingered memory deficits. *Archives of Clinical Neuropsychology*, 16, 7, 669 - 678

[137] Hollrah, J., Schlottmann, R. & Scott, A. (1995). Validity of the MMPI subtle items. *Journal of Personality Assessment*, 65, 278-299

[138] Huynh, H. & Mandeville, G. (1979). Validity conditions in repeated measures designs. *Psychological Bulletin*, 86,5, 964-973

[139] Ingram, P., Kelso, K. & McCord, D. (2011). Empirical correlates and expanded interpretation of the MMPI-2-RF Restructured Clinical Scale 3 (Cynicism). *Assessment*, 18, 95-101

[140] Inman, T., Vickery, C., Berry, D., Lamb, D., Edwards, C., & Smith, G. (1998). Development and initial validation of a new procedure for evaluating adequacy of effort given during neuropsychological testing: The letter memory test. *Psychological Assessment*, 10, 128-139

[141] Isernhagen, S. 1988). Functional Capacity Evaluation. In: S.J. Isernhagen (Hrsg.) *Work injury: Management and prevention*, Gaithsburg: Aspen Publishers, 139-174

[142] Iverson, G., Page, J. L. & Koehler, B. (2007). Test of Memory Malingering (TOMM) Scores are not affected by chronic pain or depression in patients with fibromyalgia. *The Clinical Neuropsychologist*, 21, 532-546

[143] Jelicic, M., Merckelbach, H., Candel, I. & Geraerts, E. (2007). Detection of feigned cognitive dysfunction using special malinger tests: A simulation study in naive and coached malingerers. *International Journal of Neuroscience*, 117, 8, 1185-1192

[144] Jones, A., Ingram, M. & Ben-Porath, Y. (2012). Scores on the MMPI-2-RF scales as a function of increasing levels of failure on cognitive symptom validity tests in a military sample. *The Clinical Neuropsychologist*, 26, 790-815

[145] Junge, A., Fröhlich, M., Ahrens, S., Hasenbring, M., Sandler, A., Grob, D., & Dvorak, J. (1996). Predictors of bad and good outcomes of lumbar spine surgery. A prospective clinical study with 2-year's-follow-up. *Spine*, 21, 1056-1065

[146] Kaiser, H., Kersting, M., Schian, H., Jacobs, A. & Kasprowski, D. (2000). Der Stellenwert des EFL-Verfahrens nach Susan Isernhagen in der medizinischen und beruflichen Rehabilitation. *Rehabilitation*, 39, 297-306

[147] Kalso, E., Edwards, J., Moore, R. & McQuay, H. (2004). Opioids in chronic non-cancer pain: Systematic review of efficacy and safety. *Pain*, 112,3, 372-380

[148] Känel, R. von (2008). Das Burnout-Syndrom: Eine medizinische Perspektive. *Praxis*, 97, 477-487

[149] Keller, L. & Butcher, J. (1991). *Assessment of chronic pain patients with the MMPI-2*. Minneapolis: University of Minnesota Press

[150] Kessler, J., Calabrese, P., Kalbe, E., & Berger, F. (2000). Demtect. Ein neues Screening-Verfahren zur Unterstützung der Demenzdiagnostik. *Psycho* 6, 343-347.

[151] Klinger, R. (2011). *Klassifikation chronischer Schmerzen: „Multiaxiale Schmerzklassifiaktion (MASK)“.* In: B. Kröner-Herwig, J. Frettlöh, R. Klinger & P. Nilges (Hrsg.) *Schmerzpsychotherapie*, Heidelberg: Springer, 319-333.

[152] Klinger, R., Hasenbring, M., Pfingsten, M., Hürter, A., Maier, C. & Hildebrandt, J. (2000). *Die multiaxiale Schmerzklassifikation - Psychosoziale Dimension (MASK-P), Band 1.* In: R. Klinger, M. Hasenbring, M. Pfingsten, A. Hürter, C. Maier & J Hildebrandt (Hrsg.) *Die multiaxiale Schmerzklassifikation (MASK)*, Hamburg: Deutscher Schmerzverlag

[153] Kobelt, A., Winkler, M., Göbber, J., Pfeiffer, W. & Petermann, F. (2010). Hängt die subjektive Prognose der Erwerbstätigkeit vom Migrationsstatus ab? *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, 58, 3, 189-197

[154] Kohlmann, T. & Raspe, H. (1996). Der Funktionsfragebogen Hannover zur alltagsnahen Diagnostik der Funktionsbeeinträchtigung durch Rückenschmerzen (FFBH-R). *Rehabilitation*, 35, 1-8

[155] Komarahadi, F., Maurischat, C., Härter, M. & Bengel, J. (2003). Zusammenhänge von Depressivität und Ängstlichkeit mit sozialer Erwünschtheit bei chronischen Schmerzpatienten. *Schmerz*, 18, 38-44

[156] Kool, J., Meichtry, A., Schaffert, R. & Rüesch, P. (2008). *Der Einsatz von Beschwerdevalidierungstests in der IV-Abklärung. Bericht im Rahmen des mehrjährigen Forschungsprogramms zu Invalidität und Behinderung (FoP-IV).* Bern: Schweizerisches Bundesamt für Sozialversicherungen

[157] Kool, J., Oesch, P., Bachmann, S., Knuesel, O., Dierkes, J., Russo, M., de Bie, R. & van den Brandt, P. (2005). Increasing days at work using function-centered rehabilitation in nonacute nonspecific low back pain: A randomized controlled trial. *Archives of Physical Medicine and Rehabilitation*, 86, 857-864

[158] Korb, J. & Pfingsten, M. (2003). Der deutsche Schmerzfragebogen - implementierte Psychometrie. *Schmerz*, 17, 47

[159] Koss, M. & Butcher, J. (1973). A Comparison of psychiatric patients' self-report with other sources of clinical information. *Journal of Research in Personality*, 7,3, 225-236

[160] Kröner-Herwig, B. (2011). *Schmerz als biopsychosoziales Phänomen.* In: B. Kröner-Herwig, J. Frettlöh, R. Klinger & P. Nilges (Hrsg.) *Schmerzpsychotherapie*, Heidelberg: Springer, 4-13

[161] Kupfer, J., Brosig, B. & Brähler, E. (2001). *TAS-26: Toronto-Alexithymie-Skala-26. Deutsche Version. Manual.* Göttingen: Hogrefe.

[162] Lachar, D. & Wrobel, T. (1979). Validating clinical hunches: Constructing of a new MMPI critical item set. *Journal of Counseling and Clinical Psychology*, 47, 277-284

[163] Lanyon, R. (1978). *Psychological Screening Inventory II. Manual*. Michigan: Port Huron

[164] Lanyon, R. (2003). Assessing the misrepresentation of health problems. *Journal of Personality Assessment*, 81, 1-10

[165] Larrabee, G. (2003). Exaggerated pain report in litigants with malingered neurocognitive dysfunction. *The Clinical Neuropsychologist*, 17,3, 395-401

[166] Larrabee, G. (2007). *Assessment of Malingered Neuropsychological Deficits.* Oxford: University Press

[167] Lees-Haley, P. (1991). MMPI-2 and F-K scores of personal injury malingerers in vocational neuropsychological and emotional distress claims. *American Journal of Forensic Psychology*, 9, 5-13

[168] Lees-Haley, P. (1992). Efficacy of MMPI-2 validity scales and MCMI-II modifier scales for detecting spurious PTSD claims: F, F-K, Fake Bad Scale, Ego Strength, Subtle-Obvious Subscales, DIS, DEB. *Journal of Clinical Psychology*, 48, 681-689

[169] Lees-Haley, P.R., English, L.T. & Glenn, W.J. (1991). A Fake Bad Scale on the MMPI-2 for Personal Injury Claimants. *Psychological Reports*, 68, 203-210

[170] Lewis, J., Simcox, A. & Berry, D. (2002). Screening for feigned psychiatric symptoms in a forensic sample by using the MMPI-2 and the Structured Inventory of Malingered Symptomatology. *Psychological Assessment*, 14, 170-176

[171] Lienert, G. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz

[172] Linden, M., Baron, S. & Muschalla, B. (2009). *Das Mini-ICF-Rating für psychische Störungen (Mini-ICF-APP)*. Bern: Huber

[173] Linton, S. (2000). A review of psychological risk factors in back and neck pain. *Spine*, 25, 1148-1156

[174] Maier, C. (2008). Auch Sucht ist eine Krankheit. *Schmerz*, 22, 639-643

[175] Maier, C., Hildebrandt, J., Klinger, R., Hasenbring, M., Pfingsten, M., Hürter, A. & Mitarbeiter der Arbeitsgruppen EDV und Qualitätssicherung der DGSS (2000). *Die multiaxiale Schmerzklassifikation - Somatische Dimension (MASK-S), Band 2*. In: R. Klinger, M. Hasenbring, M. Pfingsten, A. Hürter, C. Maier & J Hildebrandt (Hrsg.) *Die multiaxiale Schmerzklassifikation (MASK)*. Hamburg: Deutscher Schmerzverlag

[176] Main, C. (1983). The Modified Somatic Perception Questionaire. *Journal of Psychosomatic Research*, 27,6, 503-514

[177] Main, C., Wood, P., Hillis, S., Spanswick, C. & Waddell, G. (1992). The distress and risk assessment method. A simple patient classification to identify distress and evaluate the risk of poor outcome. *Spine Journal*, 17, 42-52.

[178] Martell, B.A., O'Connor P.G., Kern, R.D., Becker, W.C., Morales, K.H., Kosten, T.R. & Fiellin, D.A. (2007). Systematic review: Opioid treatment for chronic back pain: prevalence, efficacy, and association with addiction. *Annals of Internal Medicine*, 16,146, 116-127

[179] Maruta, T., Osborne, D., Swanson, D. & Halling, J. (1981). Chronic pain patients and spouses - marital and sexual adjustment. *Mayo Clinic Proceedings*, 56, 307-310

[180] Maruta, T., Vatterott, M. & McHardy, M. (1989). Pain management as an antidepressant: long-term resolution of pain-associated depression. *Pain*, 36, 335-337

[181] Matheson, L. & Matheson, M. (1989). *Assessment and Capacity Testing*. Trabuco Canyon: PACT.

[182] Mathiowetz, V., Weber, K., Volland, G., Dowe, M. & Rogers, S. (1985). Grip and pinch strength: Normative data for adults. *Archives of Physical Medicine and Rehabilitation*, 66, 69-74

[183] McCall, W.A. (1939). *Measurement*. New York: Macmillian

[184] McCord, D. & Drerup, L. (2011). Relative practical utility of the Minnesota Multiphasic Personality Inventory-2 Restructured Clinical Scales versus the Clinical Scales in a chronic pain patient sample. *Journal of Clinical and Experimental Neuropsychology*, 33, 140-146

[185] McGuire, B. & Shores, E. (2001). Simulated pain on the Symptom Checklist 90-Revised. *Journal of Clinical Psychology*, 57, 1589-1596

[186] McNaughton, H., Sims, A. & William, T. (2000). Prognosis for people with back pain under a no-fault 24-hour-cover compensation scheme. *Spine*, 25,10, 1254-1258

[187] Melzack, R. (1975). The McGill Pain Questionnaire: Major properties and scoring methods. *Pain*, 1, 277-299

[188] Melzack, R. (1987). The short-form McCill Pain Questionnaire. *Pain*, 30, 191-197

[189] Melzack, R. (2005). The McGill Pain Questionnaire. *Anesthesiology*, 103, 199-202

[190] Melzack, R. & Torgerson, W. (1971). On the language of pain. *Anaesthesiology*, 34, 50-59

[191] Merckelbach, H. & Smith, G. (2003). Diagnostic accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in detecting instructed malingering. *Archives of Clinical Neuropsychology*, 18, 145-152

[192] Merckelbach, H. & Merten, T. (2012). A Note on Cognitive Dissonance and Malingering. *The Clinical Neuropsychologist*, 26, 1217-1229

[193] Mersky, H. (1986). Classification of chronic pain. Description of chronic pain symptoms and definitions of pain terms. *Pain*, 3, 1-225

[194] Merten, T. (2011). Beschwerdenvalidierung bei der Begutachtung kognitiver und psychischer Störungen. *Fortschritte der Neurologie, Psychiatrie und ihrer Grenzgebiete*, 79, 102-116

[195] Merten, T., Blaskewitz, N. & Stevens, A. (2007). Kann suboptimale Testmotivation mit dem Aufmerksamkeits-Belasstungs-Test (Test d2) erkannt werden? *Aktuelle Neurologie*, 34,3, 134-139

[196] Merten, T., Bossink, L. & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonligant patients. *Journal of Clinical and Experimental Neuropsychology*, 29, 308-318

[197] Merten, T. & Dohrenbusch, R. (2010). Testpsychologische Ansätze der Beschwerdenvalidierung. *Psychotherapeut*, 55, 394-400

[198] Merten, T., Friedel, E., Mehren, G. & Stevens, A. (2007). Über die Validität von Persönlichkeitsprofilen in der nervenärztlichen Begutachtung. *Nervenarzt*, 78, 511-520

[199] Merten, T., Green, P., Blaskewitz, M. H. N. & Brockhaus, R. (2005). Analog validation of german-language symptom validity tests and the influence of coaching. *Archives of Clinical Neuropsychology*, 20,6, 719-726

[200] Meyer, R. & Deitsch, S. (1996). *The clinician's handbook: Integrated diagnostics, assessment, and intervention in adult and adolescent psychopathology*. Needham Heights, Massachusetts.: Allyn & Bacon, 4th Edition

[201] Meyers, J., Miller, R., Haws, N., Murphy-Tafiti, J., Curtis, T., Rupp, Z., Smart, T. & Thompson, L. (2013). An adaptation of the MMPI-2 Meyers Index for the MMPI-2-RF. *Applied Neuropsychology: Adult*, 0, 1-7

[202] Meyers, J., Millis, S. & Volkert, K. (2002). A validity index for the MMPI-2. *Archives of Clinical Neuropsychology*, 17, 157-169

[203] Miller, H. (2001). *Miller Forensic Assessment of Symptoms Test (M-FAST) and professional manual*. Lutz, FL: Psychological Assessment Resources.

[204] Millis, S., Putnam, S. & Adams, K. (1995). *Neuropsychological malingering and the MMPI-2: Old and new indicators*. Paper presented at the 30th Annual Symposium on Recent Developments in the Use of MMPI, MMPI-2, MMPI-A, pp. 1-5

[205] Millon, T., Millon, C., Davis, R. & Grossman, S. (2006). *Millon Clinical Multiaxial Inventory (MCMI-III). Manual*. Minneapolis, MN: Pearson Education, Inc, 3d ed.

[206] Mindach, M. (2000). Keine Opioidabhängigkeit bei Schmerzpatienten? Fragen eines lesenden Arztes. *Schmerz*, 14, 186-191

[207] Mittenberg, W., Azrin, R., Millsaps, C. & Heilbronner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale-Revised. *Psychological Assessment*, 5, 34-40

[208] Mittenberg, W., Patton, C., Canyock, E. & Condit, D. (2002). Base rates of malingering and symptom exeggeration. *Journal of Clinical and Experimental Neuropsychology*, 24, 1094-1102

[209] Mondon, K., Gochard, A., Marque, A., Armand, A., Beauchamp, D., Prunier, C., Jacobi, D., de Toffol, B., Autret, A., Camus, V., & Hommet, C. (2007). Visual recognition memory differentiates dementia with lewy bodies and parkinsons disease dementia. *Journal of Neurology, Neurosurgery, and Psychiatry*, 78, 738–741

[210] Morey, L. (1991). *An interpretive guide to the Personality Assessment Inventory (PAI)*. Odessa, FL: Psychological Assessment Resources

[211] Morey, L. (1991). *Personality Assessment Inventory: Professional Manual*. Odessa, FL: Psychological Assessment Resources

[212] Morse, C.L., Douglas-Newman, K., Mandel, S. & Swirsky-Sacchetti, T. (2013). Utility of the Rey-15 recognition trial to detect invalid performance in a forensic neuropsychological sample. *Clinical Neuropsychology*, 27, 1395-1407

[213] Nelson, N., Hoelzle, J., Sweet, J., Arbisi, P. & Demakis, G. (2010). Updated metaanalysis of the MMPI-2 symptom validity scale FBS: Verified utility in forensic practice. *The Clinical Neuropsychologist*, 24, 701-724

[214] Nichols, D. & Green, R. (1991). *New measures for dissimulation on the MMPI/MMPI-2*. St. Petersburg, Florida: 26th Annual Symposium on Recent Developments in the Use of the MMPI (MMPI-2/MMPI-A)

[215] Nilges, P. & Nagel, B. (2007). Was ist chronischer Schmerz? *Deutsche Medizinische Wochenzeitschrift*, 132, 2133-2138

[216] Nilges, P. & Rief, W. (2010). F45.41 Chronische Schmerzstörung mit somatischen und psychischen Faktoren. Eine Kodierhilfe. *Schmerz*, 24, 209-212

[217] Oehlschlägel, J. (1996). *Frankfurter Aufmerksamkeits-Inventar (FAIR)*. Bern: Huber

[218] Olsen, C. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83,4, 576-586

[219] Payne, B. & Norfleet, M. (1986). Chronic pain and the family. A review. *Pain*, 26, 1-22

[220] Pelfrey, W. (2004). The relationship between malingerers' intelligence and MMPI-2 knowledge and their ability to avoid detection. *International Journal of Offender Therapy and Comparative Criminology*, 48,6, 649-663

[221] Pelfrey, W. & Aamodt, M. (1996). *Relationship between malingeres' intelligence and MMPI-2 Knowledge and their ability to avoid detection as malingerer*. Paper presented at the Annual Meeting of the Academy of Criminal Justice Sciences, Las Vegas: Nevada

[222] Pincus, T., Callahan, L., Bradley, L., Vaughn, W. & Wolfe, R. (1986). Elevated MMPI scores for hypochondriasis, depression, and hysteria in patients with rheumatoid arthritis reflect disease rather than psychological status. *Arthritis Rheumatism*, 29, 1456-1466

[223] Plohmann, A. & Merten, T. (2013). The third european symposium on symptom validity assessment: Facts and controversies. *Clinica y Salud*, 14, 197-203

[224] Polatin, P., Kinney, R., Gatchel, R., Lillo, E. & Mayer, T. (1993). Psychiatric illness and chronic low back pain. The mind and the spine - which goes first? *Spine*, 18, 66-71

[225] Portenoy, R. (1996). Opioid therapy for chronic nonmalignant pain: A review of the criticla issues. *Journal of Pain and Symptom Management*, 11,4, 203-217

[226] Porter, L., Bigley, G. & Steers, R. (2003). *Motivation and work behavior*. New York: McGraw-Hill, 7th edition.

[227] Rainville, J., Sobel, J., Hartigan, C. & Wright, A. (1997). The effect of compensation involvement on the reporting of pain and disability by patients referred for rehabilitation of chronic low back pain. *Spine*, 22,17, 2016-2024

[228] Resnick, P. (1984). The detection of malingered mentall illness. *Behavioral Sciences and the Law*, 2, 20-38

[229] Richter, W. (2016). *Schmerzanamnese.* In: H.G. Nobis, R. Rolke, T. Graf-Baumann (Hrsg.) *Schmerz - Eine Herausforderung*. Berlin: Springer, in prep

[230] Richter, W. (2012). *Schmerzanamnese.* In: H.G. Nobis, R. Rolke, T. Graf-Baumann (Hrsg.) *Schmerz - Eine Herausforderung*. Berlin: Springer, 79 pp.

[231] Richter, W., Hankemeier, U. & Aulbert, E. (2000). *Psychische Grundlagen von Schmerzempfindung, Schmerzäußerung und Schmerzbehandlung.* In: U. Hankemeier, F. Krizanits & K. Schüle-Hein (Hrsg.) *Tumorschmerztherapie*, Berlin: Springer.

[232] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Müller, M. (2011). PROC: An open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77, **doi:10.1186/1471-2105-12-77**

[233] Rogers, R. (1984). Towards an empirical model of malingering and deception. *Behavioral Science and the Law*, 2, 93-112

[234] Rogers, R. (1990). Development of a new classificatory model of malingering. *Bulletin of the American Academy of Psychiatry and the Law*, 18, 323-333

[235] Rogers, R. (2008). *Detection Strategies for malingering and defensiveness.* In: R. Rogers (Ed.) *Clinical assessment of malingering and perception*. New York: Guilford, 14-35

[236] Rogers, R., Bagby, R. & Chakraborty., N. (1993). Feigning schizophrenia disorders on the MMPI-2: Detection of coached simulators. *Journal of Personality Assessment*, 60, 215-226

[237] Rogers, R., Bagby, R. & Gillis, J. (1992). Improvement in the M test as a screening measure for malingering. *Bulletin of the American Academy of Psychiatry and the Law*, 20, 101-104

[238] Rogers, R. & Bender, S. (2003). *Evaluation of malingering and deception.* In: A.M. Goldstein & I.B. Weiner (Eds.) *Handbook of Psychology, Vol. 2, Forensic Psychology*. New Jersey: Wiley, 112 pp.

[239] Rogers, R., Gillard, N., Berry, D. & Granacher, R. (2011). Effectiveness of the MMPI-2-RF validity scales for feigned mental disorders and cognitive impairment: A known-groups study. *Journal of Psychopathology and Behavioral Assessment*, 33, 355-367

[240] Rogers, R., Hinds, J. & Sewell, K. (1996). Feigning psychopathology among adolescent offenders: Validation of the SIRS, MMPI-A, and SIMS. *Assessment*, 67, 244-257

[241] Rogers, R., Salekin, R., Sewell, K., Goldstein, A. & Leonard, K. (1998). A comparison of forensic and nonforensic malingerers: A prototypical analysis of explanatory models. *Law and Human Behavior*, 22, 353-367

[242] Rogers, R., Salekin, R., Sewell, K., Martin, M. & Vitacco, M. (2003). Detecting of feigned mental disorders: A metaanalysis of the MMPI-2 and malingering. *Assessment*, 10, 168-177

[243] Rogers, R. & Sewell, K. (2006). MMPI-2 at the crossroads: Aging technology or radical retrofitting? *Journal of Personality Assessment*, 87,2, 175-178

[244] Rogers, R., Sewell, K., Cruise, K., Wang, E. & Ustead, K. (1998). The PAI and feigning: A cautionary note on its use in forensic-correctional settings. *Assessment*, 5, 399-405

[245] Rogers, R., Sewell, K., Morey, L. & Ustead, K. (1996). Detection of feigned mental disorders on the Personality Assessment Inventory: A discriminant analysis. *Journal of Personality Assessment*, 67, 629-640

[246] Rogers, R., Sewell, K. & Salekin, R. (1994). A meta-analysis of malingering on the MMPI-2. *Assessment*, 1,3, 227-237

[247] Rogers, R., Sewell, K. & Goldstein, A. (1994). Explanatory models of malingering: A prototypical analysis. *Law and Human Behavior*, 18, 543-552

[248] Rohling, M. & Binder, L. (1995). Money matters: A meta-analytic review of the association between financial compensation and the experience and treatment of chronic pain. *Health Psychology*, 14, 6, 537-547

[249] Romano, J., Turner, J. & Clancy, S. (1989). Sex differences in the relationship of pain patients dysfunction to spouse adjustment. *Pain*, 39, 289-295

[250] Rosenfeld, B., Sands, S. & Gorp, W. V. (2000). Have we forgotten the base rate problem? Methodological issues in the detection of distortion. *Archives of Clinical Neuropsychology*, 15, 349-359

[251] Rouse, S., Greene, R., Butcher, J., Nichols, D. & Williams, C. (2008). What do the MMPI-2 Restructured Clinical Scales reliably measure? Answers from multiple research settings. *Journal of Personality Assessment*, 90, 435-442

[252] Rowat, K. & Knafl, K. (1985). Living with chronic pain: The spouse's perspective. *Pain*, 23, 259-271

[253] Roy, R. (1982). Marital and family issues in patients with chronic pain. A review. *Psychotherapy and Psychosomatics*, 37, 1-12

[254] Rubenzer, S. (2010). Review of the Structured Interview of Reported Symptoms-2 (SIRS-2). *Open Access Journal of Forensic Psychology*, 2, 273-286

[255] Ruoß, M. (1997). Schmerz und Behinderung als subjektive Konstruktionen. *Schmerz*, 11, 305-313.

[256] Schachter, S. & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69, 379-399

[257] Schmand, B. & Lindeboom, J. (2005). *AKGT: Amsterdamer Kurzzeitgedächtnistest. Amsterdam Short-Term Memory Test*. In Zusammenarbeit mit T. Merten und R. Millis. Göttingen: Hogrefe

[258] Schmidt, R. & Toews, J. (1970). Grip strength as measured by the Jamar Dynamometer. *Archives of Physical Medicine and Rehabilitation*, 51, 321-327

[259] Schmidt-Atzert, L., Bühner, M., Rischen, S. & Warkentin, V. (2004). Erkennen von Simulation und Dissimulation im Test d2. *Diagnostica*, 50, 124-133

[260] Schmitt, N. & Gerbershagen, H. (1990). The Mainz Pain Staging System (MPSS) for Chronic Pain. *Pain*, 5, 484, suppl.

[261] Schneider, W., Firzlaff, M., Birke, K. & Klauer, T. (2009). Sozialmedizinische Begutachtung der beruflichen Leistungsfähigkeit. *Psychotherapeut*, 54, 37-43

[262] Schneider, W. (2007). Standards der sozialmedizinischen Leistungsbegutachtung in der Psychosomatischen Medizin und Psychotherapie. *Psychotherapeut*, 52, 447-462

[263] Scholte, W., Tiemensand, B., Verheul, R., Meerman, A., Eggers, J. & Hutschemaekers, G. (2012). The RC Scales predict psychotherapy outcomes: The predictive validity of the MMPI-2's Restructured Clinical Scales for psychotherapeutic outcomes. *Personality and Mental Health*, 6, 292-302

[264] Schroeder, R., Baade, L., Peck, C., VonDran, E., Brockman, C., Webster, B. & Heinrichs, R. (2012). Validation of the MMPI-2-RF Validity Scales in criterion group neuropsychological samples. *The Clinical Neuropsychologist*, 26, 129-146

[265] Schubert, M., Parthier, K., Kupka, P., Krüger, U., Holke, J., & Fuchs, P. (2013). Menschen mit psychischen Störungen im SGB II. Iab Forschungsbericht. Aktuelle Ergebnisse aus der Projektarbeit des Instituts für Arbeitsmarkt- und Berufsforschung 12/2013. **http://doku.iab.de/forschungsbericht/2013/fb1213.pdf** Stand: 11.09.2016

[266] Schwartz, G. & Weiss, S. (1978). Yale Conference on Behavioral Medicine: A propose definition and statement of goals. *Journal of Behaviorial Medicine*, 1, 3-12

[267] Seamons, D., Howell, R., Carlisle, A. & Roe, A. (1981). Rorschach simulation of mental illness and normality by psychotic and nonpsychotic legal offenders. *Journal of Personality Assessment*, 45, 130-135

[268] Sellbom, M. & Bagby, R. (2008). *Response styles on multiscale inventories.* In: R. Rogers (Ed.) *Clinical Assessment of Malingering and Perception*. New York: Guilford, 182-206

[269] Sellbom, M. & Bagby, R. (2010). Detection of overreported psychopathology with the MMPI-2 RF form validity scales. *Psychological Assessment*, 22,4, 757-767

[270] Sellbom, M., Ben-Porath, Y., Baum, L. & Gregory, E.E.C. (2008). Predictive validity of the MMPI-2 Restructured Clinical (RC) Scales in a batterers' intervention program. *Journal of Personality Assessment*, 90, 129-135

[271] Sellbom, M., Ben-Porath, Y., McNulty, J., Arbisi, P. & Graham, J. (2006). Elevation differences between MMPI-2 Clinical and Restructured Clinical (RC) Scales: Frequency, origins, and interpretative implications. *Assessment*, 13, 430-441

[272] Sellbom, M., Toomey, J., Wygant, D. & Kucharski, L. (2010). Utility of the MMPI-2-RF (Restructured Form) validity scales in detecting malingering in a criminal forensic setting: A known-groups design. *Psychological Assessment*, 22,1, 22-31

[273] Sellbom, M., Wygant, D. & Bagby, R. (2012). Utility of the MMPI-2-RF in detecting non-credible somatic complaints. *Psychiatry Research*, 197, 295-301

[274] Shaw, D. & Matthews, C. (1965). Differential MMPI performance of brain damaged versus pseudoneurologic groups. *Journal of Clinical Psychology*, 21, 405–408.

[275] Simms, L., Casillas, A., Clark, L., Watson, D. & Doebbeling, B. (2005). Psychometric evaluation of the Restructured Clinical Scales of the MMPI-2. *Psychological Assessment*, 17, 345-358

[276] Sipos, C., Reker, M., Heinold, F., Thier, T., Stuppe, M., Richter, W. & Hankemeier, U. (2000). Behandlung von chronischen Schmerzen bei Drogenabhängigen. *Schmerz*, 14,3, 175.-183

[277] Sivec, H., Lynn, S. & Garske, J. (1994). The effect of somatoform disorder and paranoid psychotic role-related dissimulations as a response set on the MMPI-2. *Assessment*, 1, 69-81

[278] Slick, D., Hopp, G. & Strauss, E. (1995). *The Victoria Symptom Validity Test*. Odessa, FL: Psychological Assessment Resources

[279] Slick, D., Sherman, E. & Iverson, G. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, 13, 545-561

[280] Smith, G. (1992). Detection of malingering: A validation study of the slam test (Doctoral dissertation). *Dissertation Abstracts International*, 53, 3795b

[281] Smith, G. & Burger, G. (1997). Detection of malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of the American Academy of Psychiatry and the Law*, 25, 183-189.

[282] Steffan, J., Clopton, J. & Morgan, R. (2003). An MMPI-2 scale to detect Malingered Depression (MD Scale). *Assessment*, 10, 382-292

[283] Sternbach, R. (1974). *Pain Patients. Traits and Treatment*. New York: Raven

[284] Strong, D., Glassmire, D., Frederick, R., & Green, R. (2006). Evaluating the Latent Structure of the MMPI-2-F(p)-Scale in a Forensic Sample: A Taxometric Analysis. *Psychological Assessment* 18,3: 250-261

[285] Strong, D., Greene, R. & Schinka, J. (2000). A taxometric analysis of MMPI-2 infrequency scales F and F(p) in clinical settings. *Psychological Assessment*, 12, 166-173

[286] Sullivan, K. & King, J. (2008). Detecting Faked Psychopathology: A Comparison of Two Tests to Detect Malingered Psychopathology Using a Simulation Design. *Psychiatry Research*, 176,1, 75-81

[287] Sullivan, M., Bishop, S. & Pivik, J. (1995). The Pain Catastrophizing Scale: Development and validation. *Psychological Assessment*, 7, 524-532

[288] Swanson, D. & Maruta, T. (1980). The family's viewpoint of chronic pain. *Pain*, 8, 163-166

[289] Swartzman, L., Teasell, R., Shapiro, A. & McDermid, A. (1996). The effect of litigation status on adjustment to whiplash injury. *Spine*, 21,1, 53-58

[290] Sweet, J., Condit, C. & Nelson, N. (2008). Feigned Amnesia and Memory Loss. In: R. Rogers (Ed.) *Clinical assessment of malingering and perception*. New York: Guilford, 218-236.

[291] Swets, J. (1973). The relative operating characteristic in psychology. *Science*, 182, 990-1000

[292] Tait, R., Pollard, C., Margolis, R., Duckro, P. & Krause, S. (1987). The Pain Disability Index: Psychometric and validity data. *Archives of Physical and Medical Rehabilitation*, 68, 438-441

[293] Tarescavage, A., Wygant, D., Gervais, R. & Ben-Porath, Y. (2013). Association between the MMPI-2 Restructured Form (MMPI-2-RF) and malingered neurocognitive dysfunction among non-head injury disability claimants. *The Clinical Neuropsychologist*, 27, 313-335

[294] Tearnan, B. & Lewandowski, M. (1992). The behavioral assessment of pain questionnaire: The development and validation of a comprehensive self-report instrument. *American Journal of Pain Management*, 2,4, 181-191

[295] Tellegen, A., Ben-Porath, Y., Sellbom, M., Arbisi, P., McNulty, J. & Graham, J. (2006). Further evidence on the validity of the MMPI-2 Restructured Clinical (RC) Scales: Addressing questions raised by Rogers et al. and Nichols. *Journal of Personality Assessment*, 87, 148-171

[296] Thies, E. (2012). *Der deutsche MMPI-2: Effektivität der Validitätsskalen in der Aufdeckung von Antwortverzerrung*. Marburg: Tectum

[297] Thomas, M. & Youngjohn, J. (2009). Let's not get hysterical: Comparing the MMPI-2 validity, clinical, and RC Scales in TBI litigants tested for effort. *The Clinical Neuropsychologist*, 23, 1067-1084

[298] Tombaugh, T. (1996). *Test of Memory Malingering (TOMM)*. North Tonawanda: Multi-Health Systems

[299] Tscheuschner, B. (2011). *Über die Möglichkeit nicht-authentische Beschwerden zu erkennen. Diagnostische Optimierung bei Simulation*. Tübingen: Eberhard Karls Universität, Medizinische Fakultät, Dissertation

[300] Tsushima, W., Geling, O. & Fabrigas, J. (2011). Comparison of MMPI-2 Validity scale scores of personal injury litigants and disability claimants. *The Clinical Neuropsychologist*, 25,8, 1403-1414

[301] Turk, D., Flor, H. & Rudy, T. (1987). Pain and families. I. Etilogy, maintenance and psychosocial impact. *Pain*, 30, 3-27

[302] Turk, D., Meichenbaum, D. & Genest, M. (1983). *Pain and Behavioral Medicine. A cognitive-behavioral perspective*. New York: Guilford Press

[303] Van Impelen, A., Merckelbach, H., Jelicic. M., & Merten, T. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): A Systematic Review and Meta-Analysis. *The Clinical Neuropsychologist* 28: 1336-1365

[304] Van der Heijden, P.T., Egger, J.I.M. & Derkson, J.J.L. (2008). Psychometric evaluation of the MMPI-2 Restructured Clinical Scales in two dutch samples. *Journal of Personality Assessment*, 90, 456-464

[305] Van der Heijden, P.T., Egger J.I.M. & Derksen, J.J.L. (2010). Comparability of scores on the MMPI-2-RF Scales generated with the MMPI-2 and MMPI-2-RF booklets. *Journal of Personality Assessment*, 92, 254-259

[306] Venkatraman, E. (2000). A permutation test to compare Receiver Operating Characteristic Curves. *Biometrics*, 56, 1134-1138

[307] Venkatraman, E. & Begg, C. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, 83, 835-848

[308] Von Weizsäcker, V. (1931). Über den Begriff der Arbeitsfähigkeit. *Deutsche Medizinische Wochenschrift* 39(57), 1653-1657; 1996-1998

[309] Waddell, G., McCulloch, J., Kummel, E. & Venner, R. (1980). Nonorganic physical signs in low-back pain. *Spine*, 5,2, 117-125

[310] Wagner, T., Richter, W., Rothkopf, C., Staudigel, K. & Hankemeier, U. (2003). Der Schmerztherapeut bei der Begutachtung. *Schmerz*, 17, 20-33

[311] Weber, A., Weltle, D. & Lederer, P. (2004). Frühinvalidität im Lehrerberuf: Sozial- und Arbeitsmedizinische Aspekte. *Deutsches Ärzteblatt*, 101,13, A850-859

[312] Weiss, D., England, R. D.G. & Lofquist, L. (1967). *Manual for the Minnesota Satisfaction Questionnaire*. Minneapolis: University of Minnesota, Vol. 22

[313] Wenig, C., Schmidt, C., Kohlmann, T. & Schweikert, B. (2009). Costs of back pain in Germany. *European Journal of Pain*, 13,3, 280-286

[314] Whitney, K., Davis, J., Shepard, P., Herman, S. & Roudebush, R. (2008). Utility of the Response Bias Scale (RBS) and other MMPI-2 validity scales in predicting TOMM performance. *Archives of Clinical Neuropsychology*, 23,7, 777-786

[315] Widder, B., Dertwinkel, R., Egle, U., Foerster, K. & Schiltenwolf, M. (2008). Leitlinie zur Begutachtung von Schmerzen. *Psychotherapeut*, 52, 334-346

[316] Wiener, D. (1948). Subtle and obvious keys for the MMPI. *Journal of Consulting Psychology*, 12, 164-170

[317] Wiggins, C., Wygant, D., Hoelzle, J. & Gervais, R. (2012). The more you say the less it means: Overreporting and attenuated criterion validity in a forensic disability sample. *Psychological Injury and Law*, 5, 162-173

[318] Wiggins, J. (1959). Interrelations among MMPI measures of dissimulation under standard and social desirability instructions. *Journal of Consulting and Clinical Psychology*, 3, 419-427

[319] Windemuth, D. (1997). Möglichkeiten der psychologischen Aggravationsdiagnostik fachorthopädischer Schmerzpatienten durch den Einsatz einer multidimensionalen Schmerzskala. *Verhaltenstherapie und Verhaltensmedizin*, 18, 407-417

[320] Wittchen, H. & Jacobi, F. (2005). Size and burden of mental disorders in Europe - A critical review and appraisal of 27 studies. *European Neuropsychopharmacology*, 15, 357-376

[321] Wygant, D. (2007). *Validation of the MMPI-2 Infrequent Somatic Complaints (Fs) Scale.* Kent State University: Unpublished doctoral dissertation

[322] Wygant, D., Ben-Porath, Y., Arbisi, P. & Berry, D. (2006). *An MMPI-2 validity scale designed to detect somatic overreporting in civil forensic settings*. Paper presented at the Annual Conference for the American Psychology Law Society. Florida: St. Petersburg

[323] Wygant, D., Ben-Porath, Y., Arbisi, P., Berry, D., Freeman, D. & Heilbronner, R. (2009). Examination of the MMPI-2 Restructured Form (MMPI-2-RF) validity scales in civil forensic settings: Findings from simulation and known-group samples. *Archives of Clinical Neuropsychology*, 24, 671-680

[324] Wygant, D., Sellbom, M., Gervais, R., Ben-Porath, Y., Stafford, K., Freeman, D. & Heilbronner, R. (2010). Further validation of the MMPI-2 and MMII-2-RF Response Bias Scale: Findings from disability and criminal forensic settings. *Psychological Assessment*, 22,4, 745-756

[325] Youden, W. (1950). Index for Rating Diagnostic Tests. *Cancer*, 3, 32-35.

[326] Youngjohn, J., Lees-Haley, P. & Binder, L. (1999). Comment: Warning malingerers produces more sophisticated malingering. *Archives of Clinical Neuropsychology*, 14,6, 511-515

[327] Zenz, M., Strumpf, M. & Willweber-Strumpf, A. (1990). Orale Opiattherapie bei Patienten mit „nicht-malignen“ Schmerzen. *Schmerz*, 4, 14-21

[328] Ziesat, H. (1978). Are family patterns related to the development of chronic low back pain. *Perception and Motor Skills*, 46, 1062 (suppl.)

[329] Zimmermann, M. (1996). *Schmerz - Was ist das?* In: J. Apfelbach (Hrsg.) *Schmerz*. Reinbek: Einhorn, 10-21