BEITRÄGE ZUR SOZIALEN SICHERHEIT

Bericht im Rahmen des mehrjährigen Forschungsprogramms zu Invalidität und Behinderung (FoP-IV)

Der Einsatz von Beschwerdevalidierungstests in der IV-Abklärung

Forschungsbericht Nr. 4/08

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Bundesamt für Sozialversicherungen BSV
Office fédérale des assurances sociales OFAS

Das Bundesamt für Sozialversicherungen veröffentlicht in seiner Reihe "Beiträge zur Sozialen Sicherheit" konzeptionelle Arbeiten sowie Forschungs- und Evaluationsergebnisse zu aktuellen Themen im Bereich der Sozialen Sicherheit, die damit einem breiteren Publikum zugänglich gemacht und zur Diskussion gestellt werden sollen. Die präsentierten Folgerungen und Empfehlungen geben nicht notwendigerweise die Meinung des Bundesamtes für Sozialversicherungen wieder.

Autoren/Autor/innen:

Jan Kool, André Meichtry, René Schaffert, Peter Rüesch
Zürcher Hochschule für Angewandte Wissenschaften ZHAW
Departement Gesundheit
Technikumstrasse 71
8401 Winterthur
Tel. +41 (0) 58 934 63 21 / Fax +41 (0) 52 269 63 21
E-mail: jan.kool@zhaw.ch
Internet: www.zhaw.ch/de/gesundheit.html

Auskünfte:

Martin Wicki
Abteilung Mathematik Analysen Statistik
Bundesamt für Sozialversicherungen
Effingerstrasse 20
3003 Bern
Tel. +41 (0) 31 322 90 02
E-mail: martin.wicki@bsv.admin.ch

ISBN: 3-909340-59-8

Vertrieb:

BBL, Vertrieb Publikationen, CH - 3003 Bern
http://www.bundespublikationen.admin.ch

Bestellnummer: 318.010.4/08 d

Forschungsprogramm FoP-IV

Der Einsatz von Beschwerdevalidierungstests in der IV-Abklärung

Zürcher Hochschule für Angewandte Wissenschaften ZHAW, Departement Gesundheit

Jan Kool, André Meichtry, René Schaffert, Peter Rüesch

Winterthur, 28. August 2008

Vorwort des Bundesamts für Sozialversicherungen

Wer infolge von Geburtsgebrechen, Krankheit oder Unfall invalid wird und dadurch Einkommenseinbussen erleidet, erhält von der Invalidenversicherung die Kosten für eine angemessene Deckung des Existenzbedarfs ausgeglichen (Art. 1a IVG). Die Abklärung, ob eine erkrankte Person in ihrer Erwerbsfähigkeit eingeschränkt ist und damit Anrecht auf eine Rente hat, ist oft sehr anspruchsvoll. Die Komplexität hat mit der Zunahme psychisch oder schmerzbedingter Leiden, die schwieriger abzuklären sind, zugenommen. Gerade diese gesundheitlichen Beeinträchtigungen können bei klinischen Abklärungen oftmals kaum objektiviert werden, was in den letzten Jahren in der Öffentlichkeit Anlass gegeben hat, Menschen mit diesen Behinderungen, die Renten beziehen, pauschalisierend unter den Generalverdacht des Versicherungsbetrugs zu stellen.

Aus diesen Gründen gibt es einen Bedarf nach zusätzlichen und verbesserten Abklärungsmethoden von schwierig objektivierbaren Gesundheitsschäden. Im vorliegenden Projekt wird geprüft, ob zusätzlich zu den üblichen klinischen Untersuchungen auch Symptom- bzw. Beschwerdevalidierungstests eingesetzt werden können. Die vorliegende Studie wurde in Auftrag gegeben, um einen systematischen Überblick über den Einsatz von Beschwerdevalidierungstests und erste Erfahrungen mit diesen in der Schweiz zu erhalten. Ausgehend von einer Literaturanalyse holten die Forschenden die Stimmen aus der Praxis ein. Die Befunde der Studie sind auf den ersten Blick kontrovers: in der Literatur wird etlichen Validierungstests wissenschaftliche Validität zugesprochen, bei den Expertinnen und Experten in der Praxis herrscht hingegen grosse Skepsis vor.

Ein zweiter Blick ergibt aber ein differenzierteres Bild: die Fachliteratur über die im deutschen Sprachraum eingesetzten Tests verweist auf viele Probleme bei deren wissenschaftlichen Überprüfung. So sind die „Simulanten“ bei den wissenschaftlichen Testüberprüfungen in der Regel „schauspielernde“ Studierende. Es wurden keine Vergleiche zwischen Abklärungen gemacht, die Tests einsetzen und solchen ohne Einsatz von Tests. Auch weisen einzelne Tests eine geringe Spezifität auf, das heisst sie „überführen“ begutachtete Personen fälschlicherweise dem Simulationsverdacht, was nicht hingenommen werden kann. Andererseits zeigt sich, dass viele der angefragten Expertinnen und Experten mit Tests arbeiten, jedoch häufiger mit nicht wissenschaftlich geprüften. Verbindend kann festgestellt werden, dass sowohl die wissenschaftliche Literatur als auch die Praxis als zentral hervorheben, dass die Beschwerdevalidierungstests nur komplementär zur fachlich qualifizierten klinischen Untersuchung durch erfahrene Fachspezialistinnen und -spezialisten eingesetzt werden dürfen. Mit Tests lassen sich Widersprüche aufdecken, aufgrund derer einen Fall vertiefter angesehen werden sollte. Als Bestandteil im Abklärungsprozess sollten sie deshalb nicht ausgeschlossen werden.

Der vorliegende systematische Überblick über die Literatur, den Einsatz von und die Erfahrungen mit Beschwerdevalidierungstests in der Schweiz bieten eine wertvolle Basis für eine weiterführende Diskussion unter Fachleuten zur Entwicklung und Einführung von versicherungsmedizinischen Standards. Zusätzlich sollen die angelsächsische wissenschaftliche Literatur ausgewertet und wissenschaftlich begleitete Pilotprojekte in der gutachterlichen Praxis durchgeführt werden. Wenn sich die Tests in der Praxis bewähren, sollen sie als Bestandteil der Standards breiten Fachkreisen vorgestellt und die Fachleute geschult werden. Auf diesem Wege kann die Anwendung dieser Tests an den grösser werdenden Bedarf zur Abklärung von Aggravation und Simulation in der Schweiz sukzessive angepasst werden.

Alard du Bois-Reymond

Leiter des Geschäftsfelds Invalidenversicherung

Préface de l'Office fédéral des assurances sociales

Quiconque devient invalide à la suite d'une infirmité congénitale, d'une maladie ou d'un accident et doit en supporter les conséquences en termes de revenu reçoit de l'assurance-invalidité une compensation qui lui permet de couvrir ses besoins vitaux (art. 1a LAI). Mais la démarche visant à déterminer si une personne malade est limitée dans sa capacité de gain, et donc si elle a droit à une rente, est souvent très complexe. Cette complexité a encore augmenté avec la multiplication des troubles psychiques et des syndromes douloureux, les plus délicats à évaluer, car il n'est pas rare que les examens cliniques ne permettent pas d'objectiver ces atteintes à la santé. Cela a conduit l'opinion publique, ces dernières années, à suspecter de fraude à l'assurance toutes les personnes handicapées qui touchent une rente pour des motifs de cet ordre.

Il est donc nécessaire d'améliorer les méthodes employées pour instruire les atteintes à la santé difficilement objectivables, voire d'en trouver de nouvelles. Le présent projet avait pour but de savoir si, outre les examens cliniques habituels, on pourrait faire appel à des tests de validation des symptômes. L'étude devait donc passer systématiquement en revue l'utilisation de ces tests et faire le point sur l'expérience acquise en la matière dans notre pays. Partant d'une analyse bibliographique, les chercheurs ont réuni des témoignages concrets. A première vue, les résultats de l'étude paraissent contradictoires : selon la littérature, les tests seraient validés scientifiquement, tandis qu'en pratique les experts semblent pour le moins sceptiques.

Mais, en y regardant de plus près, on découvre un tableau plus nuancé. D'un côté, en effet, la littérature spécialisée traitant des tests employés dans l'espace germanophone cite de nombreux problèmes quant à leur validation scientifique. Par exemple, les « simulateurs » auxquels fait appel l'analyse scientifique des tests sont généralement des étudiants « qui jouent la comédie ». Aucune comparaison n'est faite entre les procédures d'instruction utilisant des tests et celles qui n'en utilisent pas. Certains tests sont peu spécifiques, c'est-à-dire qu'ils font croire aux expertisés qu'il s'agit d'une simulation alors que ce n'est pas le cas, situation qui n'est pas acceptable. D'un autre côté, on s'aperçoit que, si de nombreux experts interrogés travaillent avec des tests, ceux-ci sont souvent non validés scientifiquement. On peut conclure en disant que, tant pour la littérature scientifique que pour la pratique, l'emploi de tests de validation des symptômes est à réserver à des spécialistes expérimentés, et uniquement en complément des examens cliniques réalisés par un personnel qualifié. Ils peuvent permettre de repérer des contradictions amenant à étudier le cas plus en profondeur, et ne sont pas donc à exclure en tant que composante d'une procédure d'instruction.

Cette analyse systématique de la littérature relative aux tests de validation des symptômes, de leur utilisation et de l'expérience déjà acquise en Suisse constitue

une base précieuse pour une discussion plus poussée entre professionnels sur le développement et l'introduction de standards dans le domaine de la médecine des assurances. Il serait également nécessaire de passer en revue la littérature scientifique anglo-saxonne et de réaliser des projets pilotes concrets, suivis sur le plan scientifique, dans les institutions chargées des expertises. Si les tests font leurs preuves en pratique, il faudrait les intégrer aux standards qui seront présentés aux spécialistes, et y former ces derniers. Ainsi pourra-t-on adapter progressivement l'usage de ces tests aux besoins croissants de l'instruction lorsque des phénomènes d'aggravation et de simulation sont en jeu.

Alard du Bois-Reymond

Chef du domaine Assurance-invalidité

Premessa dell’Ufficio federale delle assicurazioni sociali

L’assicurazione invalidità compensa con un’adeguata copertura del fabbisogno vitale le conseguenze economiche subite da chi, a causa d’infermità congenita, malattia o infortunio è o è diventato invalido (art. 1a LAI). Accertare l’incapacità al guadagno e quindi il diritto a una rendita è spesso molto difficile. A maggior ragione, se si considera la forte crescita del numero di persone sofferenti di disturbi – generalmente psichici, ma anche fisici – pressoché impossibili da dimostrare clinicamente. Non a caso, sono proprio i beneficiari di rendita affetti da questi disturbi che negli ultimi anni l’opinione pubblica ha preso a sospettare indistintamente di simulazione e truffa ai danni dell’assicurazione.

Per accertare danni alla salute difficilmente oggettivabili sono dunque necessari nuovi e più efficaci metodi. Commissionato per avere un quadro sistematico dell’impiego e dei risultati dei test di verifica della veridicità dei sintomi/disturbi in Svizzera, il presente studio valuta la possibilità di completare gli esami clinici classici con test di questo tipo. Allo scopo, i ricercatori hanno analizzato la letteratura specifica in lingua tedesca e sondato la prassi. A prima vista, i risultati parrebbero contraddittori: mentre la letteratura riconosce la validità scientifica dei test, tra gli esperti che li applicano nella prassi regna lo scetticismo.

Ad una più attenta lettura emerge tuttavia un quadro più differenziato: gli studi considerati, infatti, segnalano molte lacune nella verifica scientifica dei test utilizzati nei Paesi di lingua tedesca. Fanno notare, p. es., che per sperimentarne la validità si è fatto ricorso ad „attori“ (in genere studenti) che recitavano la parte del simulatore. Inoltre, non sono stati fatti confronti tra accertamenti che fanno uso di questi test e accertamenti che non ne fanno uso. Non solo: alcuni di questi test sono troppo poco specifici e rischiano di „smascherare“ simulazioni in realtà inesistenti, cosa ovviamente inaccettabile. D’altro canto, molti degli esperti interpellati, a prima vista scettici, fanno regolarmente uso di test, nella maggior parte dei casi addirittura senza alcuna garanzia scientifica. Letteratura scientifica e prassi concordano tuttavia su un punto, e cioè che i test di verifica della veridicità dei disturbi possono essere impiegati soltanto a complemento di esami clinici qualificati eseguiti da specialisti esperti. Grazie ai test si possono scoprire contraddizioni e analizzare più a fondo determinati casi. Non se ne dovrebbe quindi escludere l’impiego quale parte integrante del processo di accertamento.

Il quadro sistematico della letteratura specifica in lingua tedesca e l’analisi dei risultati finora ottenuti in Svizzera con i test di verifica della veridicità dei disturbi offerti dal presente studio costituiscono una valida base su cui fondare l’introduzione e lo sviluppo di test standard per la medicina assicurativa. L’attendibilità dei test dovrà tuttavia essere verificata nella prassi mediante progetti pilota svolti ed accompagnati secondo criteri scientifici e non si potrà mancare di studiare a fondo anche la letteratura anglosassone. I test che daranno buoni risultati dovranno essere proposti

come parte integrante degli standard agli ambienti interessati e costituire materia di formazione per gli specialisti. In questo modo se ne potrà gradualmente adeguare l’applicazione al crescente bisogno di smascherare i casi di aggravamento e simulazione.

Alard du Bois-Reymond

Capo Ambito Assicurazioni invalidità

Foreword by the Federal Social Insurance Office

The purpose of invalidity insurance (IV) is to compensate individuals suffering from a disability due to congenital disease, illness or an accident for the economic consequences of this disability by covering their basic living costs (Art. 1a IVG). It is often very difficult to determine whether the earning capacity of individuals in poor health is in fact reduced, thereby entitling them to draw an IV pension. The complexity and difficulty of this task has increased concomitantly with the rise in the number of individuals presenting with psychological or pain-related ailments, which often do not lend themselves to an objective clinical assessment. This has led in recent years to a tendency among the general public to suspect even bona fide IV pension recipients of possible benefit fraud.

In order to remedy this situation, additional and more effective methods are needed to assess health disorders which are difficult to diagnose medically. The present project examines whether the use of symptom validity tests (SVT) in conjunction with the usual clinical assessments could offer a solution. The research team was asked to provide a systematic overview of the use of SVTs and to gather feedback on their trial introduction in Switzerland. The researchers first reviewed the literature on this subject before conducting a survey among test practitioners. At first glance, the study findings appear controversial – the literature asserts the scientific validity of these tests, yet those who used these tests in practice were highly sceptical.

However, a closer look produces a more nuanced picture. The literature on the specific tests piloted in the German-speaking part of Switzerland highlights many problems with regard to their scientific validity. For example, during the development stages these tests are trialled on individuals who pretend to fake their symptoms rather than on actual suspected “malingerers”. No comparisons could be made between assessments which involved the use of such tests and those that did not. Also a number of tests lacked specificity, in other words they falsely “convict” the person taking the test of suspected malingering, which is unacceptable. At the same time, the study found that many of the practitioners surveyed were using tests which were not scientifically validated. Both the scientific literature and practitioners underline that SVTs should not replace professional clinical assessments carried out by experienced specialists but rather serve as a complement to them. These types of test are useful in that they often expose discrepancies in the accounts given by the examinees, thereby indicating that the case warrants closer inspection. In light of this finding, the inclusion of SVTs as a component in the IV assessment procedure should not be ruled out.

The present systematic overview of the literature, as well as the summary report on the use of and experiences with SVTs in Switzerland offer a valuable starting point for an exhaustive debate by qualified experts on the development and introduction of insurance medicine standards. In addition, a review of scientific research from the English-speaking world as well as pilot projects in institutions involved in such asses-

sments should be conducted. If these tests are validated during the pilot trial, they should be put forward to a wider circle of experts as new components of the IV assessment procedure and the relevant experts should be given the training needed to use them. Along the way the application of these tests could be gradually adapted in line with the rising need to identify the presence of aggravation and malingering in relation to Swiss IV pension claims.

Alard du Bois-Reymond

Head of Invalidity Insurance

Inhaltsverzeichnis

Tabellenverzeichnis V
Abbildungsverzeichnis VII
Zusammenfassung IX
Résumé XV
Riassunto XXI
Summary XXVII
Glossar XXXIII
1 Einleitung, Problemstellung 1
1.1 Ausgangslage 1
1.2 Zielsetzungen, Fragestellungen 5
1.2.1 Allgemeine Zielsetzung 5
1.2.2 Fragestellungen 5
1.3 Konzeption der Untersuchung 5
1.3.1 Übersicht 5
1.3.2 Systematische Literaturrecherche 6
1.3.3 Experteninterviews 6
1.3.4 Anwenderbefragung 7
2 Theoretische und begriffliche Grundlagen 9
2.1 Simulation und verwandte Konstrukte 9
2.1.1 Definition von Simulation (engl. Malingering) 9
2.1.2 Konzept der negativen Antwortverzerrung (response bias) 11
2.1.3 Störungsbilder mit engem Bezug zu Simulation/Aggravation 13
2.1.4 Störungen mit Krankheitswert vs. Simulation/Aggravation: Inkonsistenzkriterium 13
2.1.5 Häufigkeit von Simulation/Aggravation in Abklärungssituationen 15
2.1.6 Fazit 15
2.2 Methodische Schwierigkeiten bei der Messung und Identifikation von Simulation 16
2.2.1 Identifikation von Simulation 17
2.2.2 Known-Groups 17
2.2.3 Analogstudien 17

2.2.4 Sensitivität, Spezifität und prädiktive Werte 17
2.2.5 Prosecutor's fallacy 18
2.2.6 Kreuzvalidierung, externe Validität und ökologische Validität 20
2.2.7 Fazit 21
2.3 Verfahren zur Erfassung von Simulation, Aggravation 21
2.3.1 Hintergrund 22
2.3.2 Beschwerdevalidierungstests 24
2.3.3 Leitlinien 25
2.3.4 Leitlinie für die Begutachtung der BV bei Schmerzen mit Behinderung 26
2.3.5 Fazit 28
3 Beschwerdevalidierungstests: Systematische Literaturrecherche 31
3.1 Vorgehen 31
3.1.1 Datenbanken 31
3.1.2 Definition der Suchstrategie 31
3.1.3 Erhaltene Referenzen 32
3.1.4 Anpassung der Auswertungsstrategie 33
3.2 Steigendes Interesse an Beschwerdevalidierungstests 34
3.3 Gütekriterien für Beschwerdevalidierungstests 35
3.4 Diskussion ausgewählter Tests: Strukturierter Fragebogen Simulierter Symptome SFSS / (engl.) SIMS 35
3.4.1 Entwicklung und Aufbau des SFSS / SIMS 36
3.4.2 Zusammenfassung der Erkenntnisse aus Studien zum SFSS / SIMS 36
3.4.3 Bewertung des SFSS / SIMS 37
3.5 Diskussion ausgewählter Tests: Amsterdamer Kurzzeitgedächtnistest AKGT / (engl.) ASMT 40
3.5.1 Entwicklung und Aufbau des AKGT 40
3.5.2 Zusammenfassung der Erkenntnisse aus Studien zum AKGT 40
3.5.3 Bewertung des AKGT 41
3.6 Weitere wichtige Beschwerdevalidierungstests 42
3.6.1 Word Memory Test (WMT) 43
3.6.2 Test of Memory Malingering (TOMM) 43
3.6.3 Medical Symptom Validity Test (MSVT) 44
3.6.4 Word Completition Memory Test (WMCT) 44

3.6.5 Testbatterie zur Forensischen Neuropsychologie (TBFN) 44
3.6.6 Miller Forensic Assessment of Symptoms Test (M-FAST) 45
3.6.7 Unterskalen des Minnesota Multiphasic Personality Inventory (MMPI-2) 45
3.6.8 Weitere mögliche Test zur Beschwerdevalidierung 45
3.7 Spezielle Einsatzgebiete: chronische Schmerzen 48
3.8 Diskussion der Erkenntnisse aus der Literaturstudie 48
3.9 Fazit aus der Literaturstudie 50
4 Interviews mit Experten und Gutachtenden 51
4.1 Methoden 51
4.2 Resultate 51
4.2.1 Arbeitsfeld 52
4.2.2 Das Umfeld der Begutachtung 53
4.2.3 Diagnosen der Exploranden 53
4.2.4 Inkonsistenz, Verdeutlichung, Aggravation, Simulation 54
4.2.5 Beschwerdevalidierungstests 56
4.2.6 Barrieren 63
4.2.7 Wissenschaftliche Literatur und Forschungsbedarf 63
4.2.8 System und Strukturen 64
4.2.9 Entwicklungen und Erwartungen 65
4.3 Fazit 66
4.3.1 BVT im kognitiven Bereich 66
4.3.2 Verhaltener Gebrauch von BVT in der Praxis 66
4.3.3 Ungenügend validierte BVT vor allem im körperlichen Bereich 66
4.3.4 Konsistenzprüfung versus Beschwerdevalidierung 66
4.3.5 Quantensprung und Unbestimmtheit 67
4.3.6 Andere Entscheidungshilfen 67
4.3.7 Forschungsbedarf 67
4.3.8 Häufigkeit von Simulation 67
5 Schriftliche Befragung von MEDAS- und RAD-Gutachtenden 69
5.1 Methoden 69
5.2 Resultate 70
5.3 Fazit 74

6 Synthese und Diskussion der Befunde 77
6.1 In Kürze: Antworten auf die zentralen Fragestellungen 77
6.2 Methodische Grenzen der Studie 77
6.3 Diskussion und Synthese der Befunde 78
6.3.1 Theoretische und begriffliche Grundlagen 78
6.3.2 Prävalenz von Aggravation und Simulation 79
6.3.3 BVT und Leitlinien 79
7 Schlussfolgerungen und Empfehlungen 81
Literaturverzeichnis 83

Tabellenverzeichnis

Seite

Tabelle 1	Übersicht über nicht-zielkonforme Leistungen der IV nach Ott et al. (2007, S.40-41)	3
Tabelle 2	Mögliche Kontexte negativer Antwortverzerrungen (nach Merten, 2008, im Druck)	12
Tabelle 3	Abgrenzung Simulation/Aggravation von artifziellen, somatoformen und dissoziativen Störungen	14
Tabelle 4	Zusammenhang zwischen dichotomem Testresultat und Malingering	18
Tabelle 5	Der Strukturierte Fragebogen simulierter Symptome SFSS aus der Sicht der Gütekriterien nach Hartmann	38
Tabelle 6	Der Amsterdamer Kurzzeitgedächtnistest aus der Sicht der Gütekriterien nach Hartmann	41
Tabelle 7	Selten erwähnte Tests zur Beschwerdevalidierung (Forts. nächste Seite)	46
Tabelle 8	Themenfelder der Interviews mit Experten und Gutachtenden	52
Tabelle 9	Antworthäufigkeiten pro Sprachregion und total	71
Tabelle 10	Von den Gutachtenden genannte Tests für die Identifikation von Inkonsistenz	72
Tabelle 11	Von den Gutachtenden genannte Tests zum Bestimmen von Aggravation und Simulation	73

Abbildungsverzeichnis

Seite

Abbildung 1 Entwicklung IV-Neuberentungen nach Invaliditätsursachen 1997-2006 (indexiert, 1997=1.00; Daten: IV-Statistik 2007) 2
Abbildung 2: Prävalenzabhängigkeit des positiv prädiktiven Wertes 19

Zusammenfassung

Ausgangslage, Problemstellung

In den letzten Jahren wurde in der Schweizer Politik, in den Medien und in der Öffentlichkeit die Frage nach der Grössenordnung des ungerechtfertigten Bezugs von Leistungen der Sozialhilfe und -versicherungen aufgeworfen. Die Debatte hat einerseits zwar einen Tabubruch bewirkt, indem die Vergabe von Leistungen der IV kritisch und öffentlich hinterfragt wird. Zum anderen lief sie aber auch Gefahr, Menschen mit Behinderungen und Beeinträchtigungen, die Renten beziehen, pauschalisierend unter den Generalverdacht des Betrugs zu stellen. Eine neuere Studie des BSV aus dem Jahre 2007 versucht, das Risiko, dass Leistungen nicht adäquat vergeben wurden, einzuschätzen. Die Autoren sprechen in diesem Zusammenhang von so genannten *nicht-zielkonformen Leistungen.* Bei einem kleineren Teil dieser Fälle muss davon ausgegangen werden, dass Leistungen aufgrund falscher Angaben der Klientel − d.h. durch Aggravation oder Simulation von Beschwerden − zugesprochen wurden.

Nicht-zielkonforme Leistungen der IV sind zur Hauptsache bei schwierig objektivierbaren Gesundheitsstörungen zu erwarten; dazu zählen z.B. chronische Rückenschmerzen ohne somatisch erkennbare Ursache, andere Schmerzkrankheiten, Schleudertraumata und Depressionen. Bei diesen Störungsbildern besteht ein teilweise erheblicher Ermessensspielraum für die Einschätzung der Erwerbsunfähigkeit und des IV-Grades. Es gibt deshalb einen Bedarf nach einer verbesserten Abklärungspraxis u.a. durch die Entwicklung von Standards bei der Begutachtung von schwierig objektivierbaren Gesundheitsbeeinträchtigungen.

Zielsetzungen und Fragestellungen

Die vorliegende Studie soll Kenntnisse zum aktuellen Stand der wissenschaftlichen Entwicklung von Beschwerdevalidierungstests (BVT) aufbereiten und Anwendung sowie Umgang mit BVT in der medizinischen und neuropsychologischen berufsbezogenen Abklärungspraxis darstellen. Weiter soll eine kritische Bewertung und Dokumentation der Tests und ihrer Anwendungen vorgenommen werden. Die Befunde der Studie dienen als Grundlagen für die Diskussion über die Abklärungsverfahren in der IV und für eine (verbesserte) Implementierung von BVT in den Abklärungs- und Gutachterstellen der Sozialversicherungen. Im Einzelnen wurden die folgenden Fragestellungen untersucht:

1. Welche Beschwerdevalidierungstests sind in der wissenschaftlichen Fachliteratur validiert worden und wie ist diese Validität zu beurteilen?
2. Welche BVT werden in den verschiedenen Settings der Sozial- und Unfallversicherungen (SUVA, IV, KK, Taggeld, Haftpflicht) angewandt, und wie werden sie dort beurteilt?
3. Wie können die Erfahrungen mit der Diagnostik im neuropsychologischen Kontext in andere Fachbereiche, die in der Beurteilung versicherungsrelevanter Beschwerden tätig sind, transferiert und operationalisiert werden?

Methodisches Vorgehen

Im Rahmen der vorliegenden Studie wurden drei methodische Zugänge zum Untersuchungsgegenstand gewählt.

Zunächst wurden mit einer *systematischen Literaturrecherche* in wissenschaftlichen Literaturdatenbanken potenzielle Messinstrumente und Beobachtungsverfahren im Bereich der Beschwerdevalidierung ermittelt, beschrieben und beurteilt.

In einem zweiten Schritt wurde mit leitfadengestützten *Experteninterviews* der Einsatz von Beschwerdevalidierungstests (BVT) in der Abklärungspraxis untersucht. Interviewt wurden 13 Personen: elf Gutachtende, vorwiegend tätig in Medizinischen Abklärungsstellen (MEDAS), und 2 wissenschaftliche Experten. Von den interviewten Gutachtenden arbeiteten deren fünf in der Westschweiz und sechs in der Deutschschweiz. Es handelte sich dabei um Psychiater, Rheumatologen, Allgemeinmediziner und Psychologen.

Schliesslich erfolgte eine strukturierte Befragung der (potenziellen) Anwenderinnen und Anwender von BVT. Hauptzielgruppe waren Fachpersonen, die zu Handen der IV Gutachten erstellen (insbesondere Mitarbeitende der Regionalärztlichen Dienste RAD und der medizinischen Abklärungsstellen MEDAS).

Aggravation und Simulation: theoretische und konzeptionelle Aspekte

Aggravation ist in der Fachliteratur definiert als Übertreibung oder Ausweitung von Beschwerden; tatsächlich vorhandene Symptome werden zur Erreichung eines Ziels (z.B. Rente, Erlass einer Massnahme etc.) verstärkt. Demgegenüber meint Simulation die absichtliche, reflektierte, zweckvolle Vortäuschung von Symptomen oder fälschliche Beschwerdenschilderung.

Beide Verhaltensweisen äussern sich häufig als *Inkonsistenzen* zwischen − in der Abklärung − *beobachteten* und − aufgrund der geschilderten Beschwerden − *erwarteten* Fähigkeiten oder Leistungen. In Bezug auf die gezeigten Leistungen in einem Beschwerdevalidierungstest werden diese Inkonsistenzen auch als *negative Antwortverzerrung* bezeichnet.

Das Problem liegt nun aber darin, dass Inkonsistenzen zwischen beobachteten und erwarteten Leistungen nicht nur Ausdruck von Aggravation oder Simulation sondern auch von gesundheitlichen Störungen sein können, denen Krankheitswert zugebilligt wird. Dazu werden in der Fachliteratur besonders bestimmte psychische Erkrankungen, nämlich die somatoformen Störungen und die artifizielle Störung gezählt. Es werden zwei Kriterien angeführt, welche die Abgrenzung der Aggravation/Simulation von Krankheiten erlauben sollen: zum einen die Motivierung des Klienten oder der Klientin durch einen externen Anreiz wie der Erhalt einer Rente, der Erlass einer Strafe etc., und zum anderen die Bewusstseinsnähe des Verhaltens. Postuliert wird: je stärker das Verhalten durch externe Anreize motiviert und je bewusster es ist, desto grösser ist die Wahrscheinlichkeit für das Vorliegen von Aggravation oder Simulation. Beide Kriterien − Motivierung und Bewusstseinsnähe − sind jedoch nicht einer Überprüfung durch bestimmte Messinstrumente und der Objektivierung zugänglich. Es liegt vielmehr im Ermessensspielraum der Gutachtenden, die Motivation und den Grad der Bewusstheit eines Klienten oder einer Klientin zu bestimmen.

Aufgrund dieser konzeptionellen Probleme bedarf die Bestimmung von Simulation oder Aggravation einer fundierten differentialdiagnostischen Abklärung, die Alternativerklärungen für das Klientenverhalten mit grosser Wahrscheinlichkeit auszuschliessen vermag.

Ein weiterer Problembereich betrifft die Messung von Aggravation oder Simulation. So fehlt ein eigentlicher *'Goldstandard'* an dem Beschwerdevalidierungstests geeicht werden können: Bei der Entwicklung von Beschwerdevalidierungstests können diese i.d.R. nicht an 'echten' Simulanten oder Simulantinnen geprüft werden, vielmehr werden dazu Personen eingesetzt, die Simulanten oder Simulantinnen spielen. Ein wichtiger Aspekt betrifft auch die *diagnostische Sicherheit* der Tests. Dabei geht es um zwei Anforderungen an einen BVT: er soll aggravierendes/simulierendes Verhalten anzeigen (Empfindlichkeit oder Sensitivität des Tests), und bei nicht-aggravierendem/simulierendem Verhalten soll er einen negativen Befund liefern (Spezifität). Zu vermeiden sind insbesondere falsch-positive Befunde − wenn also der Test bei Personen anzeigen würde, die effektiv nicht simulieren. Deshalb ist bei der Entwicklung von BVT zur Vermeidung von Falsch-Positiven eine hohe Spezifität anzustreben, was dann aber notgedrungen mit einer verminderten Sensitivität einhergeht.

Wissenschaftliche Perspektive: Beschwerdevalidierungstests, Arten und Eignung

Die Überprüfung der Plausibilität der geschilderten Beschwerden eines Klienten oder einer Klientin wird als Beschwerdevalidierung bezeichnet. Das diagnostische Instrumentarium für die Beschwerdevalidierung besteht zum einen aus so genannten Beschwerdevalidierungstests und zum anderen aus Leitlinien. Letztere definieren eine Reihe von Kriterien, die im Rahmen einer Begutachtung erfüllt sein müssen, damit mit grosser Sicherheit von Aggravation oder Simulation gesprochen werden kann; anerkannt sind die Leitlinien von Slick und von Bianchini. Beschwerdevalidierungstests sind ein Bestandteil dieser Leitlinien.

Folgende Arten von Beschwerdevalidierungstests werden in der Fachliteratur beschrieben: Alternativwahlverfahren, Tests mit vorgetäuschtem Schwierigkeitsgrad, Tests zur Identifikation untypischer Leistungsprofile. Allen Tests ist gemeinsam, dass sie die Erfassung von Inkonsistenzen zwischen beobachteten und zu erwartenden Leistungen erlauben.

Die systematische Recherche wissenschaftlich publizierter Literatur ergab für den Zeitraum 1997 bis 2007 rund 1'100 Referenzen. Davon entfielen 570 Beiträge auf die letzten fünf Jahre von 2003 bis und mit 2007 wovon 340 Veröffentlichungen seit 2005 erschienen waren. Insbesondere im englischen Sprachraum liegt eine breite Fachliteratur vor. Für die vorliegende Studie wurde die Recherche eingeschränkt auf Veröffentlichungen im deutschen Sprachraum und auf Themen mit Bezug zur IV; diese Einschränkung lieferte schliesslich 30 deutschsprachige Referenzen. Es werden einzelne Beschwerdevalidierungstests (BVT), die wissenschaftliche Gütekriterien erfüllen und deren Praxistauglichkeit untermauert ist, näher beschrieben.

Folgendes Fazit der wissenschaftlichen Literatur zu BVT kann gezogen werden: Es gibt ein breites Spektrum von Tests zu unterschiedlichen Einsatzbereichen für die Abklärung von vermuteter Aggravation oder Simulation. Trotz wachsender Forschungstätigkeit bleibt aber das Problem bestehen, dass all diese Tests, auch die wissenschaftlich gut untersuchten, jeweils spezifische Einschränkungen in Bezug auf ihre Anwendungsfelder haben und alle eine nicht zu vernachlässigende Rate von falsch-positiven Ergebnissen liefern. Gleichwohl können bei einem gezielten Einsatz und bei umsichtiger Interpretation von Testergebnissen mit diesen Verfahren zusätzliche Erkenntnisse gewon-

nen werden. Bei der Begutachtung von Aggravation und Simulation muss jedoch mit multimethodalen Ansätzen gearbeitet werden. Beschwerdevalidierungstests können in diesem Kontext zusätzliche Erkenntnisse liefern, wenn sie fachgerecht angewandt und interpretiert werden.

Perspektive der Gutachtenden: Kritische Bewertung von Beschwerdevalidierungstests

Die Prävalenz von Aggravation oder Simulation wird von einem Grossteil der interviewten Gutachtenden als gering beschrieben. Die meisten Gutachtenden waren der Meinung, dass Simulation kein eigentliches Kernproblem ihrer Arbeit darstelle. Die höheren Raten, die aus wissenschaftlichen Studien berichtet werden, führen die Gutachtenden auf die spezifischen Populationen der Studien sowie auf die Abstraktion von sozialen Faktoren bei der Entstehung von chronischen Beschwerden zurück.

Eine Minderheit der interviewten Gutachtenden verwendet im Rahmen von Abklärungen wissenschaftlich überprüfte Beschwerdevalidierungstests (BVT). Genannt wurden folgende Tests: Persönlichkeitsinventare (MMPI-2 und PS-16), Symptomskalen (SFSS, und SCL 90), kognitive Tests für das Gedächtnis (Rey Kurzzeit, TOMM, WMT, Kurzzeitgedächtnistest aus dem Bremer BVT) und kognitive Tests für die Aufmerksamkeit (Frankfurter Aufmerksamkeitsinventar). Häufiger arbeiten die Gutachtenden mit nicht wissenschaftlich gesicherten Verfahren zur Überprüfung der Konsistenz des Klientenverhaltens.

Eine spezielle Ausgangslage zeigt sich für die Begutachtung körperlicher Beschwerden, insbesondere im Kontext der Neurologie und der Rheumatologie. Hier werden mangels verfügbarer BVT Verfahren zur Beschwerdevalidierung angewandt, die eigentlich für andere Zwecke entwickelt wurden. Dazu zählen z.B. die Evaluation der funktionellen Leistungsfähigkeit (EFL) kombiniert mit dem Performance Assessment Capacity Test (PACT), die Waddell-Zeichen, der JAMAR Handkraft Test, psychophysische Tests, Bluttests zum Erfassen von Malcompliance, Verfahren der Dolorimetrie sowie das Screening für somatoforme Schmerzstörung (SOMS2 und der SOMS7).

Viele Gutachtende bezweifeln, ob einzelne BVT oder Testkombinationen die Beurteilung von Aggravation oder Simulation mit ausreichender Zuverlässigkeit und Sicherheit erlauben. Nach Ansicht der Gutachtenden liefern BVT primär Momentaufnahmen. Dabei bestehe die Gefahr, die vielen zusätzlichen Faktoren auszublenden, die in der Testsituation sowie im ganzen Gutachtensprozess bestimmend sind und das Endergebnis beeinflussen.

Grundsätzlich sind die Gutachtenden offen bezüglich der weiteren Entwicklung und Forschung auf dem Gebiet der BVT. Es wird in diesem Bereich auch ein eigentlicher Forschungsbedarf anerkannt, indem besonders die externe Validität der BVT − also die Übertragbarkeit auf breitere Klientenpopulationen in der Praxis − als ungenügend beurteilt wird. Viele Gutachtende lehnen jedoch eine überstürzte Einführung von BVT ab. Sie wünschen, das diagnostische Instrumentarium im Rahmen von Abklärungen weitgehend in eigener Verantwortung bestimmen zu können.

Die Gutachtenden berichteten über einen zunehmenden Druck durch den Abklärungsprozess der IV, die Arbeitsfähigkeit des Klienten oder der Klientin möglichst positiv zu beurteilen. Sie befürchten eine verminderte Unabhängigkeit, Neutralität und Objektivität in ihrer gutachterlichen Arbeit und bewerten dies als Gefahr für die Qualität der Gutachten.

Anwendung in der Praxis

Mit einer schriftlichen Befragung wurde die Anwendung von Beschwerdevalidierungstests in der gutachterlichen Praxis umfassender untersucht. Die Befragung richtete sich an Gutachtende der 18 Medizinischen Abklärungsstellen (MEDAS) die im Auftrag der IV Gutachten durchführen, und Mitarbeitende der 10 Regionalen Ärztlichen Dienste (RAD). Die RAD überwachen die medizinischen Abklärungsverfahren, um eine gesamtschweizerisch möglichst einheitliche Beurteilung der Leistungsgesuche zu garantieren. Bei den RAD und MEDAS sind insgesamt über 300 Personen tätig. Dreissig Personen nahmen an der Umfrage Teil und schickten einen ausgefüllten Fragebogen zurück. Möglich ist, dass der Rücklauf selektiv war und einen Einfluss auf die Erfassung entweder kritischer oder positiver Meinungen zur Anwendung von BVT hatte. Die Ergebnisse der Auswertung der Befragung können deshalb nicht als repräsentativ betrachtet werden. Eine qualitative oder beschreibende Auswertung der Befragung ist aber sehr gut möglich. Die Daten sind in Übereinstimmung mit der bereits in den Interviews angedeuteten Heterogenität der Meinungen bezüglich der Zusammenhänge zwischen Inkonsistenz, Aggravation und Simulation. Sie bestätigen auch eine gewisse Skepsis unter den Gutachtenden gegenüber Beschwerdevalidierungstests.

Das Erfassen von Inkonsistenzen – mehr oder weniger strukturiert − scheint unumstritten. Alle Anwenderinnen und Anwender der Stichprobe erachten das Erfassen von Inkonsistenzen wichtig für das Bestimmen der Arbeitsfähigkeit. Eine Minderheit der Gutachtenden wendet häufig standardisierte Tests zum Erfassen von Inkonsistenzen an. In der Romandie ist die Anwendung von nicht-standardisierten Tests seltener als in der Deutschschweiz. Dieser Unterschied ist jedoch aufgrund der unscharfen Definition eines nicht-standardisierten Tests und aufgrund der nicht repräsentativen Stichprobe vorsichtig zu interpretieren, zumal aus der Romandie vermehrt einzelne Fragebögen aus derselben Organisation zurückgeschickt wurden als aus der Deutschschweiz.

Inkonsistenz wird von 74% der Gutachtenden als ein mässiger oder starker Indikator für Aggravation und von 59% als ein mässiger oder starker Indikator für Simulation gesehen. Für 89% der Gutachtenden ist ein Urteil bezüglich Aggravation eher häufig bis sehr häufig möglich. Für ein Urteil bezüglich Simulation waren 38% der Meinung, dass dies eher häufig bis sehr häufig möglich ist. Für 35% bleibt ein Urteil über Simulation sehr selten möglich. 85% der Gutachtenden waren der Meinung, dass die Identifikation von Aggravation eher oder gar sicher zur Aufgabe eines Gutachters, einer Gutachterin gehört. Bezüglich Simulation waren 65% dieser Meinung.

Genannte Tests für das Bestimmen von Inkonsistenzen waren die Waddell-Zeichen, der MMPI, der PACT und der HAMD sowie verschiedene Labor- und Medikamentenspiegeltests. Die Häufigkeit der Anwendung war gross ebenfalls für die Waddell-Zeichen (sehr häufig), den PACT (sehr häufig), den MMPI (häufig) und den HAMD (sehr häufig).

Als Hilfe für das Bestimmen von Aggravation und Simulation wurden vor allem der MMPI, der ASTM und andere – nicht näher bestimmte – neuropsychologische Verfahren genannt. Die meisten Tests werden selten angewandt. Häufiger verwendet werden der HADS (sehr häufig), der HAMD (sehr häufig), der MADRS (sehr häufig), der MMST (häufig) sowie die Waddell-Zeichen (sehr häufig).

Die zusätzlichen Kommentare der Gutachtenden auf die offene Frage entsprachen zum grossen Teil den Aspekten, die bereits in den Interviews thematisiert wurden. Dazu gehören eine als mangelhaft empfundene ökologische und externe Validität von BVT-Testbatterien, die Betonung der Wichtigkeit der klinischen Erfahrung, die Betonung der Wichtigkeit der Ausbildung bezüglich Testmethodik und Testinterpretationen mit der damit verbunden Gefährlichkeit der Tests bei unsachgemässer Anwen-

dung, die Problematik einer möglichen Verlagerung eines juristischen und nicht-medizinischen Problems in die Pseudoobjektivität, der wiederholte Hinweis auf die Inkonsistenz der Abklärungsbefunde als zwar meistens notwendige, aber nicht hinreichende Bedingung für die Postulierung von Simulation und Aggravation.

Schlussfolgerungen und Empfehlungen

Aufgrund der Befunde der vorliegenden Studie können folgende Empfehlungen formuliert werden:

1. Die aktuelle Qualität der Gutachten in Bezug auf die Beschwerdevalidierung ist unbekannt. Als Grundlage für Verbesserungen ist eine Standortbestimmung der Beschwerdevalidierung in der Gutachtenspraxis notwendig;
2. Die Beschwerdevalidierung sollte systematisch und auf der Basis von fachlich anerkannten Leitlinien erfolgen und − soweit verfügbar − auch unter Anwendung von standardisierten BVT;
3. Aus- und Weiterbildungsangebote im Bereich der Versicherungsmedizin zu Fragen der Beschwerdevalidierung und damit verbunden zur Diagnostik schwer objektivierbarer Gesundheitsstörungen sind zu fördern;
4. Die Prävalenz von Aggravation und Simulation in der Schweiz ist unbekannt. Die wissenschaftliche Ermittlung der Aggravation und Simulation ist eine wichtige Voraussetzung für die verantwortbare Anwendung der BVT;
5. Die wissenschaftliche Validierung und Entwicklung von Beschwerdevalidierungstests mit Relevanz für die Abklärungspraxis sollte gefördert werden.

Résumé

Contexte et problématique

Ces dernières années, le monde politique suisse, les médias et le public se sont beaucoup intéressés à la perception indue de prestations dans l'aide sociale et les assurances sociales et se sont interrogés sur l'ampleur du phénomène. Si, d'un côté, les débats ont brisé un tabou en posant publiquement et de manière critique la question de l'attribution des prestations AI, ils portent en eux le risque de généraliser les soupçons de fraude à toutes les personnes handicapées ou malades qui perçoivent une rente. Une étude récente de l'OFAS (2007) essaye d'estimer le risque que des prestations n'auront pas été octroyées à bon escient. Les auteurs parlent à leur propos de *prestations non conformes aux objectifs de l'assurance.* Dans une petite partie de ces cas, on peut supposer qu'elles ont été octroyées parce que les assurés avaient fait de fausses déclarations, c'est-à-dire qu'ils avaient exagéré, voire simulé, les troubles allégués.

Les prestations AI non conformes aux objectifs de l'assurance sont allouées le plus souvent, comme on peut s'y attendre, pour des atteintes à la santé difficilement objectivables, telles que douleurs du dos chroniques sans cause somatique détectable, autres troubles douloureux, coup du lapin et dépressions. Ces tableaux cliniques ont fréquemment pour particularité de laisser une importante marge d'appréciation lorsqu'il s'agit de déterminer l'incapacité de gain et le taux d'invalidité. Il est donc nécessaire d'améliorer l'instruction, notamment en développant des normes applicables aux expertises qui portent sur des atteintes à la santé difficiles à objectiver.

Objectifs et questions posées

La présente étude vise à faire le point sur le développement scientifique des tests de validation des symptômes (TVS), ainsi qu'à présenter la façon dont ils sont appliqués et utilisés en pratique dans les examens médicaux et neuropsychologiques destinés à évaluer la capacité de travail. Elle doit permettre en outre une évaluation et une documentation critiques des tests et de leur utilisation. Ses résultats serviront de base à la discussion sur l'instruction dans l'AI et amélioreront l'emploi de ces techniques dans les examens et les expertises des assurances sociales. Les questions posées étaient les suivantes :

1. Quels tests de validation des symptômes ont été validés dans la littérature scientifique spécialisée et que vaut cette validation ?
2. Quels sont les TVS utilisés dans les différentes assurances sociales et assurances-accidents (SUVA, AI, LAMal, indemnités journalières, responsabilité civile), et comment y sont-ils jugés ?
3. Comment transposer l'expérience acquise avec ces techniques diagnostiques en neuropsychologie à d'autres domaines intervenant dans l'appréciation des troubles ayant une incidence sur ces assurances ?

Méthode utilisée

Pour la présente étude, nous avons fait appel à trois approches méthodologiques.

Tout d'abord, nous avons procédé à une *recherche bibliographique systématique* dans les bases de données scientifiques afin de repérer, de décrire et d'évaluer des instruments de mesure et des procédures d'observation pouvant être utiles pour la validation des symptômes.

Nous avons ensuite étudié l'emploi de tests de validation des symptômes (TVS) dans la pratique de l'instruction par le biais d'*entretiens dirigés avec des experts*. Nous avons à cet effet interviewé 13 personnes : onze experts, actifs principalement dans des centres d'observation médicale (COMAI), et deux scientifiques. Sur les experts, cinq travaillaient en Suisse romande et six en Suisse alémanique ; il s'agissait de psychiatres, de rhumatologues, de généralistes et de psychologues.

Enfin, nous avons interrogé, au moyen d'un questionnaire structuré, les utilisateurs effectifs ou potentiels de TVS. Le principal groupe cible était composé des spécialistes effectuant des expertises pour l'AI, notamment de ceux qui travaillent dans les services médicaux régionaux (SMR) et les centres d'observation médicale (COMAI).

Exagération et simulation : aspects théoriques et conceptuels

Dans la littérature spécialisée, la notion d'« exagération » est définie comme l'accentuation ou l'extension des troubles ; le patient renforce les symptômes réellement présents afin d'atteindre un objectif (rente, mesure, etc.). La simulation est l'imitation délibérée et réfléchie des symptômes, dans un but précis, ou la description mensongère des troubles.

Les deux comportements se manifestent souvent par une *incohérence* entre les capacités ou les performances qui sont *observées* (durant l'examen) et celles qui sont *attendues* (sur la base des troubles décrits). S'agissant des performances obtenues lors d'un test de validation des symptômes, ces incohérences sont souvent appelées aussi *distorsions négatives*.

Mais le problème est que les incohérences entre les performances observées et les performances attendues peuvent exprimer non seulement des phénomènes d'exagération ou de simulation, mais aussi des atteintes à la santé réelles. Dans la littérature spécialisée, on compte parmi ces troubles certaines maladies psychiques, à savoir les troubles somatoformes et le trouble factice. Deux critères devraient permettre de faire la distinction entre l'exagération/simulation et la maladie : le fait que la personne est motivée par un intérêt extérieur (obtenir une rente, voir une peine prononcée, etc.) et la conscience qu'elle a de son comportement. On postule que plus le comportement est motivé par des intérêts extérieurs et plus il est conscient, plus grande est la probabilité d'avoir à faire à une exagération ou à une simulation. Les deux critères – motivation et conscience – ne peuvent cependant ni être mesurés instrumentalement ni objectivés. L'expert dispose en fait d'une certaine marge de manœuvre pour déterminer la motivation et le degré de conscience de la personne qu'il examine.

Etant donné ces problèmes conceptuels, il faut, pour déterminer s'il s'agit d'une exagération ou d'une simulation, un diagnostic différentiel fondé, capable d'exclure avec une grande vraisemblance d'autres explications du comportement de l'assuré.

Un autre problème est la mesure de l'exagération ou de la simulation. Il n'existe pas, en effet, d'« *étalon or* » permettant de jauger un test de validation des symptômes : ceux qui développent de tels tests ne peuvent généralement pas les essayer sur de « vrais » simulateurs ; ils doivent faire

appel à des personnes qui en quelque sorte « simulent la simulation ». Un autre problème est la *sécurité diagnostique* des tests. Pour être sûr, un test doit remplir deux conditions : mettre en évidence une exagération/simulation (sensibilité) et donner un résultat négatif en l'absence d'exagération/simulation (spécificité). Il doit surtout éviter les faux positifs, c'est-à-dire ne pas être positif alors que la personne ne simule pas. Les développeurs doivent donc, pour éviter les faux positifs, parvenir à une spécificité élevée, ce qui va nécessairement de pair avec une sensibilité moindre.

Point de vue scientifique : tests de validation des symptômes, types et adéquation

On entend par « validation des symptômes » une démarche visant à prouver la plausibilité du trouble décrit par un patient. Les instruments diagnostiques utilisés à cet effet sont, d'une part, ce qu'on appelle les tests de validation des symptômes (TVS) et, d'autre part, des critères. Ceux-ci doivent être satisfaits dans le cadre d'une expertise pour que l'on puisse parler avec une grande certitude d'exagération ou de simulation. Les critères reconnus sont ceux de Slick et ceux de Bianchini. Les TVS en font partie.

La littérature spécialisée distingue différents types de tests, tels que les tests à choix forcé, les tests avec difficultés simulées et les tests d'identification de performances atypiques. Tous ont en commun le fait qu'ils permettent de détecter des incohérences entre les performances observées et les performances attendues.

Notre recherche systématique dans la littérature scientifique a donné, pour la période allant de 1997 à 2007, environ 1100 références, dont 570 travaux publiés de 2003 à 2007 (340 depuis 2005). La littérature spécialisée est particulièrement abondante dans l'espace anglophone. Pour la présente étude, nous avons limité les recherches aux articles publiés dans l'espace germanophone et aux thèmes en rapport avec l'AI ; nous avons ainsi disposé de 30 références en allemand. Quelques tests de validation des symptômes remplissant les critères de qualité scientifique et ayant fait leurs preuves en pratique sont décrits plus en détail.

On peut tirer de la littérature scientifique relative aux TVS la conclusion suivante : il existe un vaste spectre de tests utilisés dans des domaines spécifiques et permettant de confirmer des phénomènes d'exagération ou de simulation. Bien que les recherches se multiplient, un problème demeure malgré tout : ces tests, y compris ceux qui ont été bien étudiés du point de vue scientifique, ont tous un champ d'application limité et fournissent des taux non négligeables de faux positifs. Cependant, si l'on utilise les résultats de manière ciblée et qu'on les interprète avec prudence, ils peuvent fournir des données supplémentaires. Dans le cadre d'une expertise où des questions d'exagération ou de simulation sont en jeu, plusieurs approches sont cependant nécessaires, les tests pouvant jouer un rôle complémentaire s'ils sont employés à bon escient et interprétés avec professionnalisme.

Point de vue des experts : appréciation critique des tests de validation des symptômes

La majorité des experts interrogés jugent faible la prévalence des phénomènes d'exagération et de simulation, la plupart estimant que la simulation ne constitue pas un problème majeur dans leur pratique. Les experts expliquent les taux élevés rapportés dans les études scientifiques par la spécifici-

té des populations étudiées et par la non-prise en compte des facteurs sociaux dans la genèse des troubles chroniques.

Ils sont peu nombreux à utiliser dans le cadre de leurs examens des tests de validation des symptômes scientifiquement reconnus. Parmi ceux-ci, les plus souvent cités sont : les inventaires de personnalité (MMPI-2 et PS-16), les listes de contrôle des symptômes (SFSS/SIMS et SCL 90), les tests cognitifs pour la mémoire (15 mots de Rey, TOMM, WMT, Bremer BVT) et les tests cognitifs pour l'attention (Frankfurter Aufmerksamkeitsinventar). Mais en général, les experts préfèrent évaluer la cohérence du comportement des assurés par des techniques non éprouvées scientifiquement.

Une situation particulière est celle des expertises de troubles physiques, notamment en neurologie ou en rhumatologie. Comme ces domaines ne possédaient pas de TVS spécifiques, on y a importé des procédures de validation développées à l'origine dans d'autres buts. Citons l'Evaluation des Capacités Fonctionnelles (ECF) combinée au Performance Assessment Capacity Test (PACT), les signes de Waddell, le test de force de préhension de JAMAR, les tests psychophysiques, les tests sanguins permettant de confirmer une mauvaise observance du traitement, la dolorimétrie et le Screening für somatoforme Störung (SOMS2 et SOMS7).

De nombreux experts doutent que les TVS, seuls ou associés, permettent de juger des phénomènes d'exagération ou de simulation avec une fiabilité et une certitude suffisantes. Selon eux, ils ne fournissent que des « instantanés ». Le risque est donc de négliger de nombreux autres facteurs qui jouent un rôle déterminant dans la situation de test comme dans l'ensemble du processus d'expertise et qui influent sur le résultat final.

En principe, les experts sont favorables à ce que l'on poursuive le développement des TVS et les recherches dans ce domaine. Ils estiment même que ces recherches seraient nécessaires, car ils jugent insuffisante, en particulier, leur validité externe, c'est-à-dire la possibilité de les étendre à d'autres populations. Ils sont cependant nombreux à refuser leur introduction précipitée, ordonnée par l'OFAS, et préfèrent pouvoir choisir eux-mêmes, dans la mesure du possible, leurs instruments diagnostiques.

Les experts ressentent, dans le cadre de l'instruction AI, une pression grandissante pour que la capacité de travail des assurés soit estimée au niveau le plus haut possible. Ils craignent de perdre leur indépendance et de ne plus pouvoir établir leurs expertises avec objectivité et neutralité, ce qui selon eux risque de nuire à la qualité de celles-ci.

Utilisation en pratique

Nous avons étudié de façon plus détaillée, au moyen d'un questionnaire écrit, l'utilisation des tests de validation des symptômes dans la pratique de l'expertise. Ce questionnaire s'adressait aux experts des 18 centres d'observation médicale (COMAI) qui réalisent des expertises sur mandat de l'AI et aux collaborateurs des dix centres médicaux régionaux (SMR). Les SMR contrôlent le déroulement des examens médicaux réalisés dans le cadre de l'instruction afin de garantir une appréciation uniforme des demandes de prestations. SMR et COMAI occupent au total plus de 150 personnes, mais nous n'avons reçu en retour que 30 questionnaires remplis. Ces retours pouvant être sélectifs, il est possible que les pourcentages d'avis critiques ou positifs sur l'utilisation des

TVS soient biaisés. On ne peut donc pas considérer les réponses comme représentatives. En revanche, une évaluation qualitative ou descriptive est tout à fait possible. Les données confirment l'hétérogénéité des avis, déjà manifeste lors des entretiens, sur les liens entre incohérence, exagération et simulation ; elles attestent également d'un certain scepticisme de la part des experts envers les tests de validation des symptômes.

La nécessité de repérer les incohérences – qu'elles soient plus ou moins structurées – semble faire l'unanimité. Toutes les personnes incluses dans notre échantillon jugent cette pratique importante pour déterminer la capacité de travail. Une minorité d'experts utilise fréquemment des tests standardisés ; l'utilisation de tests non standardisés est plus rare en Suisse romande qu'en Suisse alémanique. Mais, étant donné la délimitation peu claire entre tests standardisés et tests non standardisés ainsi que la non-représentativité de l'échantillon, cette différence est à interpréter avec prudence, d'autant que les questionnaires retournés provenaient plus souvent de la même organisation en Suisse romande qu'en Suisse alémanique.

Parmi les experts, 74 % considèrent que l'incohérence constitue un indicateur moyennement ou très puissant d'un phénomène d'exagération et 59 % d'un phénomène de simulation. 89 % d'entre eux sont souvent ou assez souvent capables de rendre un verdict d'exagération, contre 38 % pour la simulation ; 35 % estiment que diagnostiquer cette dernière est très rarement possible. 85 % des experts sont d'avis qu'identifier les phénomènes d'exagération fait certainement ou très certainement partie des tâches d'un expert, contre 65 % pour la simulation.

Les tests cités pour le repérage des incohérences sont les signes de Waddell, le MMPI, le PACT et le HAMD, ainsi que divers contrôles des constantes biologiques ou des dosages de médicaments dans le sang ou l'urine. Les autres tests utilisés sont les signes de Waddell (très souvent), le PACT (très souvent), le MMPI (souvent) et le HAMD (très souvent).

Pour repérer l'exagération et la simulation, les experts s'aident surtout du MMPI, de l'ASTM et d'autres techniques neuropsychologiques non précisées. La plupart de ces tests sont rarement utilisés ; le HADS (très souvent), le HAMD (très souvent), le MADRS (très souvent), le MMST (souvent) et les signes de Waddell (très souvent) le sont davantage.

Les commentaires donnés par les experts en réponse aux questions ouvertes correspondent en majorité aux points qui avaient déjà été traités durant les entretiens : manque de validité contextuelle et externe des batteries de TVS ; importance de l'expérience clinique, ainsi que de la formation à la technique des tests et à leur interprétation, avec le risque concomitant d'un emploi inapproprié ; question du traitement médical objectif d'un problème juridique et non médical ; et insistance sur le fait que l'incohérence constitue une condition nécessaire mais non suffisante pour affirmer qu'il s'agit d'un phénomène d'exagération ou de simulation.

Conclusions et recommandations

On peut tirer des résultats de la présente étude les recommandations suivantes :

1. à l'heure actuelle, rien ne permet de juger la qualité des expertises en ce qui concerne les TVS. Pour pouvoir améliorer ces tests, il est nécessaire de faire le point sur la validation des symptômes dans la pratique de l'expertise ;

2. la validation des symptômes devrait être systématique et reposer sur des directives reconnues par les professionnels ainsi que, dans la mesure où ils existent, sur des TVS standardisés ;
3. dans le domaine de la médecine des assurances, il faut promouvoir les cours de formation initiale et de perfectionnement sur la validation des symptômes et, en lien avec celle-ci, sur le diagnostic des atteintes à la santé difficilement objectivables ;
4. on ignore quelle est la prévalence des phénomènes d'exagération et de simulation en Suisse. Leur évaluation scientifique constitue donc une condition importante à l'utilisation responsable des TVS.
5. encourager la validation scientifique et le développement de tests de validation des symptômes applicables à la pratique de l'expertise.

Riassunto

Contesto e problematica

Negli ultimi anni il mondo politico svizzero, i media e l'opinione pubblica hanno sollevato la questione del volume delle prestazioni ingiustificate versate dall'aiuto sociale e dalle assicurazioni sociali. Analizzando in modo critico e pubblico la concessione delle prestazioni AI, il dibattito ha certamente permesso di infrangere un tabù, con il rischio però che le persone con disabilità o problemi di salute beneficiarie di una rendita siano sospettate in generale di truffa. Uno studio pubblicato dall'UFAS nel 2007 tenta una stima delle prestazioni AI eventualmente concesse in modo improprio. Al riguardo gli autori utilizzano il concetto di *prestazioni non conformi agli obiettivi dell'AI*. Per una parte esigua di questi casi vi è da supporre che le prestazioni siano state assegnate sulla base di false indicazioni fornite dagli assicurati (ad es. aggravamento o simulazione di disturbi).

Come prevedibile, le prestazioni non conformi agli obiettivi dell'AI sono concesse principalmente in caso di problemi di salute difficilmente oggettivabili, quali ad esempio dolori alla schiena cronici senza cause riconoscibili dal punto di vista somatico, altre sindromi dolorose, traumatismi cervicali di contraccolpo o depressioni. Per questi quadri clinici il margine di valutazione della capacità al guadagno e del grado d'invalidità è talvolta notevole. Per questo motivo è necessario migliorare la prassi d'accertamento, sviluppando tra l'altro norme standard applicabili alle perizie relative ai disturbi alla salute difficilmente oggettivabili.

Obiettivi e tematiche

L'obiettivo del presente studio è di fare il punto sullo sviluppo scientifico dei test di verifica della veridicità dei disturbi (TVVD), di descriverne l'applicazione e di presentare il modo in cui sono utilizzati nella prassi d'accertamento medica e neuropsicologica a scopo professionale. Dovrà inoltre esaminare in modo critico e documentare i test e le loro applicazioni. I risultati dello studio fungeranno da base di discussione sulla procedura di accertamento nell'AI e permetteranno di migliorare la realizzazione dei TVVD nei servizi di accertamento e nell'ambito delle perizie in tema di assicurazioni sociali. Sono state poste le domande seguenti:

1. Quali TVVD sono stati ritenuti validi nella letteratura scientifica specialistica e come deve essere valutata la loro validità?
2. Quali TVVD sono applicati nei diversi settori delle assicurazioni sociali (Suva, AI, assicurazione malattie, indennità giornaliere, responsabilità civile) e come vengono valutati da chi li utilizza?
3. In che modo le esperienze maturate con queste tecniche diagnostiche nell'ambito neuropsicologico potranno essere trasferite e applicate ad altri settori specialistici attivi nella valutazione dei disturbi rilevanti dal punto di vista dell'assicurazione?

Metodo adottato

Per lo studio sono stati scelti tre approcci metodici.

Mediante una *ricerca bibliografica sistematica* nelle banche dati della letteratura scientifica sono stati dapprima individuati, descritti e valutati potenziali strumenti di valutazione e procedure di analisi nell'ambito della verifica della veridicità dei disturbi.

In un secondo tempo sono state condotte *interviste con esperti* per analizzare l'utilizzazione dei TVVD nella prassi d'accertamento. Sono state intervistate 13 persone: 11 periti e 2 esperti scientifici. Tra i periti - psichiatri, reumatologi, medici generalisti e psicologi -, attivi prevalentemente nei servizi di accertamento medico (SAM), 5 lavoravano nella Svizzera romanda e 6 nella Svizzera tedesca.

Infine è stata eseguita un'inchiesta strutturata presso chi ha già utilizzato i TVVD o potrebbe utilizzarli. Sono stati interrogati principalmente esperti che allestiscono perizie all'attenzione dell'AI, in particolare collaboratori dei servizi medici regionali (SMR) e dei SAM.

Aggravamento e simulazione dei disturbi: aspetti teorici e concettuali

Nella letteratura specialistica la nozione di aggravamento dei disturbi viene definita come l'esagerazione o l'ampliamento degli stessi, vale a dire che i disturbi esistenti vengono amplificati al fine di raggiungere un obiettivo (rendita, provvedimento ecc.). La simulazione, dal canto suo, è la produzione intenzionale di disturbi fisici o psichici falsi o grossolanamente esagerati, motivata da incentivi esterni.

Entrambi i comportamenti sono sovente contraddistinti dall'*incoerenza* tra le capacità o prestazioni *osservate* durante l'accertamento e quelle *previste* sulla base dei disturbi descritti. Per le prestazioni ottenute in un TVVD tali incoerenze sono definite *distorsioni negative*.

Il problema consiste però nel fatto che le incoerenze tra prestazioni osservate e previste possono essere l'espressione, oltre che dell'aggravamento o della simulazione dei disturbi, anche di disturbi della salute cui viene attribuito il carattere di malattia. Ne fanno parte, come menzionato nella letteratura specialistica, determinate affezioni psichiche, quali i disturbi somatoformi e il disturbo fittizio. Due criteri dovrebbero permettere di distinguere tra aggravamento o simulazione di disturbi e malattia: la motivazione dell'assicurato determinata da un incentivo esterno come l'ottenimento di una rendita o il condono di una pena e la consapevolezza del suo comportamento. Si considera che più il comportamento è motivato da incentivi esterni ed è consapevole, maggiore è la probabilità di un aggravamento o di una simulazione. I due criteri - motivazione e consapevolezza - non possono però essere misurati con determinati strumenti né verificati con elementi obiettivi. A determinare la motivazione e il grado di consapevolezza di un assicurato è piuttosto il margine di apprezzamento di cui dispone la persona incaricata di eseguire la perizia.

Considerando questi problemi concettuali, per stabilire se vi sia simulazione o aggravamento dei disturbi è necessaria una diagnosi differenziale approfondita per poter escludere con grande probabilità altre spiegazioni al comportamento dell'assicurato.

Un altro problema è costituito dalla misurazione dell'aggravamento o della simulazione, in quanto manca un vero e proprio *gold standard* (standard qualitativo) che permetta di calibrare un TVVD. Di regola chi sviluppa questi test non può eseguirli con "veri" simulatori, ma ricorre a persone che si

comportano come tali. Riveste importanza anche la *sicurezza diagnostica* dei test. Due sono i requisiti: il test deve rilevare il comportamento di aggravamento o simulazione dei disturbi (sensibilità del test) e, se questo non è riscontrato, fornire un risultato negativo (specificità). Vanno evitati in particolare risultati falsamente positivi, ossia casi in cui il risultato è positivo ma la persona non simula. Per evitare i risultati falsamente positivi, nello sviluppo dei TVVD bisogna quindi giungere ad una specificità elevata, il che però comporta necessariamente una sensibilità ridotta.

Valutazione scientifica: test di verifica della veridicità dei disturbi - tipi e idoneità

Gli strumenti diagnostici per attestare la plausibilità dei disturbi descritti da un assicurato consistono in appositi test di verifica (TVVD) e in linee direttive. Queste ultime definiscono una serie di criteri da adempiere nell'ambito di una perizia affinché si possa parlare in modo pressoché certo di aggravamento o simulazione dei disturbi. Sono riconosciuti i criteri di Slick e di Bianchini. I TVVD ne sono parte integrante.

Nella letteratura specialistica vengono descritti i seguenti tipi di test: procedura di scelta alternativa, test con grado di difficoltà simulato e test d'identificazione di profili di prestazione atipici. Tutti permettono di rilevare le incoerenze tra prestazioni osservate e previste.

Per il periodo compreso tra il 1997 e il 2007 dalla ricerca sistematica nella letteratura scientifica sono risultate circa 1'100 referenze. 570 articoli sono stati pubblicati negli ultimi cinque anni (dal 2003 al 2007 compreso), di cui 340 dal 2005. Le pubblicazioni sono redatte prevalentemente in lingua inglese. Per il presente studio la ricerca è stata limitata alle pubblicazioni in lingua tedesca e ai temi relativi all'AI; con questa limitazione sono risultate 30 referenze. Verranno descritti più dettagliatamente alcuni TVVD che adempiono criteri di qualità scientifica e la cui idoneità pratica è stata confermata.

Dalla letteratura specialistica relativa ai TVVD emerge che per accertare un presunto aggravamento o una presunta simulazione dei disturbi esiste un'ampia gamma di test adeguati a diversi ambiti d'impiego. Tuttavia, nonostante la crescente attività di ricerca, sussiste il problema che tutti questi test, anche quelli analizzati in modo approfondito dal punto di vista scientifico, presentano talvolta limitazioni specifiche nei loro campi di applicazione e forniscono una quota di risultati falsamente positivi da non sottovalutare. Ciononostante, utilizzando in modo mirato e interpretando avvedutamente i risultati dei test, questo modo di procedere può fornire nuove informazioni. Nell'ambito di perizie per attestare l'aggravamento o la simulazione è tuttavia necessario lavorare con diversi approcci. In tal caso, se applicati e interpretati in modo competente, i TVVD possono fornire informazioni supplementari.

Valutazione critica dei periti

La maggior parte dei periti intervistati ha definito esigua la frequenza dei casi di aggravamento o simulazione dei disturbi ritenendo che la simulazione non costituiva un problema essenziale nel loro lavoro. Secondo loro, i tassi elevati menzionati negli studi scientifici sono da ricondurre alla specificità delle popolazioni di assicurati analizzate e all'astrazione fatta da fattori sociali nell'insorgenza di disturbi cronici.

Nell'ambito dei loro accertamenti una minoranza dei periti intervistati utilizza TVVD riconosciuti scientificamente. Sono stati menzionati i test seguenti: inventari di personalità (MMPI-2 e PS-6), scale di valutazione dei sintomi (SFSS e SCL 90), test cognitivi per la memoria (15 parole di Rey, TOMM, WMT, test di memoria a breve termine *Bremer BVT*) e test cognitivi per l'attenzione (*Frankfurter Aufmerksamkeitsinventar*). Tuttavia, per analizzare la coerenza del comportamento degli assicurati i periti lavorano più frequentemente con procedure non riconosciute dal punto di vista scientifico.

Si osserva una situazione particolare per le perizie relative ai disturbi fisici, in particolare nell'ambito della neurologia e della reumatologia. Non essendo disponibili TVVD specifici, per verificare la veridicità dei disturbi vengono applicate procedure sviluppate per altri scopi. Tra queste vi sono ad esempio la valutazione della capacità funzionale (VCF) combinata con il *performance assessment capacity test* (PACT), i segni di Waddell, lo *JAMAR Handkraft test*, test psicofisici, test del sangue per rilevare la scarsa osservanza della cura (*malcompliance*), la tecnica di dolorimetria e lo *screening* per disturbi somatoformi (SOMS2 e SOMS7).

Molti periti esprimono dubbi sulla possibilità di valutare con sufficiente affidabilità e sicurezza l'aggravamento o la simulazione mediante singoli TVVD o combinazioni di test. Secondo loro, questi test forniscono principalmente delle istantanee e si rischia di tralasciare molti altri fattori decisivi per il test e per l'intero processo della perizia e che possono incidere sul risultato finale.

Per principio i periti sono favorevoli al proseguimento dello sviluppo dei TVVD e della ricerca in questo ambito. Essi riconoscono pure che in questo settore è necessario un vero e proprio lavoro di ricerca, in quanto ritengono insufficiente in particolare la validità esterna di questi test, ossia la possibilità di applicarli a popolazioni di assicurati più ampie. Molti di essi sono però contrari all'introduzione affrettata di questi test, ordinata dall'UFAS, e auspicano di poter continuare a decidere quali strumenti diagnostici utilizzare nell'ambito degli accertamenti, assumendosi la responsabilità della loro scelta.

I periti segnalano che il processo d'accertamento dell'AI genera una pressione sempre maggiore affinché la capacità lavorativa dell'assicurato sia valutata il più positivamente possibile. Essi temono una limitazione della loro indipendenza, neutralità ed obiettività nell'allestimento delle perizie e ritengono che questo possa compromettere la qualità del loro lavoro.

Utilizzazione dei test

Con un'inchiesta scritta abbiamo analizzato in modo più dettagliato l'utilizzazione dei TVVD nella prassi delle perizie. Il questionario era destinato a periti dei 18 SAM incaricati di eseguire perizie all'attenzione dell'AI e a collaboratori dei 10 SMR. I SMR verificano le procedure di accertamento mediche al fine di garantire, a livello nazionale, una valutazione più uniforme possibile delle richieste di prestazioni. Complessivamente nei SMR e nei SAM lavorano oltre 150 persone. Ci sono stati rinviati 30 questionari compilati. È possibile che il rinvio dei questionari sia stato selettivo e abbia influito sul rilevamento di pareri critici o positivi concernenti l'applicazione dei TVVD. I risultati dell'inchiesta non possono quindi essere considerati rappresentativi, ma possono benissimo servire per una valutazione qualitativa o descrittiva. I dati confermano l'eterogeneità dei pareri sui nessi tra incoerenza, aggravamento e simulazione, come era già emerso dalle interviste, e ribadiscono un certo scetticismo tra i periti riguardo ai TVVD.

L'esistenza d'incoerenze, più o meno strutturate, è incontestabile. Tutte le persone figuranti nel campione che hanno utilizzato i TVVD considerano che sia importante rilevare queste incoerenze per determinare la capacità lavorativa. Una minoranza dei periti utilizza frequentemente TVVD per rilevarle. Nella Svizzera romanda i test non standardizzati vengono utilizzati più raramente rispetto alla Svizzera tedesca. Tuttavia, visto che la definizione di test non standardizzato è vaga e il campione non è rappresentativo, questa differenza va interpretata con cautela, tanto più che dalla Svizzera romanda sono stati rinviati più questionari provenienti dalla stessa organizzazione rispetto alla Svizzera tedesca.

Il 74 per cento dei periti considera che l'incoerenza costituisca un indicatore mediamente o notevolmente forte per l'aggravamento (59 per cento per quanto riguarda la simulazione). L'89 per cento di essi ritiene che spesso o molto spesso sia possibile esprimere un parere per quanto riguarda l'aggravamento dei disturbi, mentre per la simulazione tale percentuale è del 38 per cento. Per il 35 per cento, solo in rari casi ci si può pronunciare sulla simulazione. L'85 per cento dei periti ritiene che l'individuazione di casi di aggravamento dei disturbi faccia piuttosto o certamente parte dei loro compiti, mentre per la simulazione la percentuale è del 65 per cento.

I test utilizzati per determinare le incoerenze sono i segni di Waddell (molto sovente), l'MMPI (sovente), il PACT (molto sovente) e l'HAMD (molto sovente) nonché diversi test di laboratorio e con medicamenti.

Per aiutarli a determinare l'aggravamento o la simulazione i periti menzionano in particolare l'MMPI, l'ASTM e altre tecniche neuropsicologiche non meglio precisate. Nella maggior parte dei casi i test sono utilizzati raramente. Quelli maggiormente utilizzati sono l'HADS (molto sovente), l'HAMD (molto sovente), il MADRS (molto sovente), l'MMST (sovente) e i segni di Waddell (molto sovente).

Gli altri commenti dei periti riguardo alle questioni poste corrispondono in gran parte agli aspetti già esaminati nelle interviste: mancanza di validità ecologica ed esterna delle serie di test, importanza dell'esperienza clinica, della formazione per quanto riguarda la tecnica dei test e la loro interpretazione con il conseguente rischio di un'applicazione inadeguata, eventuale trasformazione di un problema giuridico e non medico in pseudo-obiettività e l'insistenza sul fatto che l'incoerenza costituisce una condizione necessaria ma non sufficiente per supporre l'aggravamento o la simulazione dei disturbi.

Conclusioni e raccomandazioni

Dai risultati del presente studio si possono formulare le raccomandazioni seguenti:

1. Il grado d'affidabilità delle perizie di verifica della veridicità dei disturbi applicate nella prassi è tuttora sconosciuto. Nella prospettiva di un loro sviluppo mirato, è quindi necessario fare il punto della situazione;
2. La verifica della veridicità dei disturbi dovrebbe essere eseguita sistematicamente, sulla base di linee guida scientificamente riconosciute e − per quanto disponibili − con l'ausilio di test standardizzati;
3. Va promossa l'offerta di formazione e perfezionamento in materia di verifica della veridicità dei disturbi, e quindi di diagnosi di disturbi difficilmente oggettivabili, nel settore della medicina assicurativa;

4. La frequenza dei casi di aggravamento e simulazione in Svizzere è sconosciuta. La certezza della loro dimostrazione scientifica è condizione imprescindibile per un impiego responsabile dei TVVD.
5. Promuovere la validità scientifica e lo sviluppo di TVVD rilevanti per la prassi d'accertamento

Summary

Background

In recent years, the issue of the unintentional or deliberate abuse of the Swiss social security and social welfare systems has been raised repeatedly by the political establishment, the media as well as the wider public. While openly criticising the allocation of invalidity insurance (IV) benefits is now no longer taboo, the debate has nevertheless increased the risk of bona fide IV benefit recipients also coming under suspicion. A FSIO study in 2007 tries to estimate the risk of unjustified benefits; the authors referred to these as *IV benefit payments which are inconsistent with invalidity insurance objectives.* In a very small number of these cases, it is likely that the award of an IV pension was based on false information provided by the claimant, i.e. as the result of exaggerating their health problems or of malingering.

It is likely that the unjustified award of IV benefits mainly concerns disorders that are difficult to diagnose medically, such as chronic back pain without a detectable somatic cause, other conditions associated with chronic pain, whiplash and depression. Given that there is greater room for discretion when evaluating the incapacity to work or the level of disability of an individual suffering from such a disorder, improvements to the current IV assessment procedure appear warranted. This will involve developing standards for the assessment of disorders which are difficult to diagnose medically.

Objectives and problems

The present study aims to contribute to existing knowledge on the scientific development of symptom validity tests (SVT) and to provide an overview of the use of these tests by medical and neuropsychological professionals during their IV assessment work. We shall also describe and critically evaluate these tests and their applications. Our findings should provide a basis not only for future discussions on the IV assessment procedure but also for the (more proficient) use of SVTs by the assessment and medical examination offices of the Swiss social security system. The study will concentrate on each of the following questions:

1. Which symptom validity tests have already been evaluated in the literature and how are these validations to be judged?
2. What types of symptom validity test do Swiss social security providers and accident insurers (SUVA, IV, daily sickness allowance, daily allowance and liability insurance) use and how do they rate them?
3. How can experiences with the methods used to carry out neuropsychological assessments be transferred to and operationalised by other professions involved in the IV assessment procedure?

Methodology

To begin with, the authors decided to adopt three different methodologies.

The first phase of the study involved a *systematic search* of scientific literature databases with the aim of identifying, describing and evaluating those rating and monitoring tools which could potentially be used to assess symptom validity.

In the subsequent part we conducted a series of *guided interviews with professionals* involved in the IV assessment procedure with regard to their use of symptom validity tests. 13 people were interviewed: 11 assessors, most of whom were employed by the medical observation centres (MEDAS), and 2 academic experts. Five of the eleven assessors were employed in Western Switzerland while the remaining six were employed in German-speaking Switzerland. This group comprised psychiatrists, rheumatologists, general physicians and psychologists.

Finally, a structured survey of (potential) users of symptom validity tests was carried out. The key target group was professionals who carry out assessments on behalf of the IV (staff from the regional medical services and the medical observation centres).

Exaggeration and malingering: theoretical and conceptual considerations

The scientific literature defines exaggeration as the overstating or amplification of symptoms for the purpose of achieving a specific outcome (e.g. pension award, receipt of other measures etc.). In contrast, malingering is the deliberate, conscious, purposive fabrication of symptoms or the reporting of a factitious history.

Both of these behaviours frequently manifest themselves as *inconsistencies* between the capacity or performance which is *observed* (during the assessment procedure) and that which is *expected* (based on the person's account of his/her symptoms). In symptom validity tests, such inconsistencies with respect to observed performance are also referred to as *negative response bias*.

The problem is that inconsistencies between observed and expected performance may indicate not only exaggeration or malingering but also a valid underlying health problem. According to the literature, such problems notably include certain types of somatoform and factitious psychological disorders. There are two criteria which should enable an assessor to identify exaggeration and malingering. The first is the client's motives for his behaviour, e.g. an external incentive such as a pension award or avoidance of prosecution etc.. The second is the degree to which the client has deliberately adopted such behaviour. There is a hypothesis which states that the more the behaviour is motivated by external incentives and the more deliberate it is, the more likely it is that the person is exaggerating or malingering. However, it is not possible to evaluate either criterion objectively by means of specific monitoring instruments. It is inevitably at the discretion of the assessors to determine whether the behaviour of the client is motivated by an external incentive or whether it is a conscious decision.

Given these conceptual problems, a thorough differential diagnostic assessment is required which would be capable of identifying exaggeration or malingering with a high degree of reliability, and therefore would be more likely to rule out alternative explanations for the person's behaviour.

A further problem area is quantifying exaggeration and malingering. No *'gold standard'* currently exists which could be applied to the symptom validity test. During the test development trials, 'real' malingerers generally cannot be used. Instead, trial participants pretend to adopt malingering behaviour. Another problematic area is the *diagnostic reliability* of the test. There are two essential requirements for any SVT: it should identify exaggerating/malingering behaviour (test sensitivity) and should produce a positive result in cases of exaggeration or malingering (specificity). It is especially important that these tests do not report false-positives, i.e. the test results wrongly indicate malinger-

ing or exaggeration. When developing an SVT, every effort therefore should be made to ensure a high degree of specificity in order to prevent false-positives. The downside, however, is that higher specificity means lower sensitivity.

Scientific perspectives: symptom validity tests, types and suitability

Symptom validity refers to the process of checking the plausibility of reported symptoms. The diagnostic instruments used to validate symptoms include "symptom validity tests" and guidelines. The latter are sets of criteria which must be met in order to determine exaggeration or malingering with a high degree of certainty; the recognised guideless commonly used are those by Slick and by Bianchini. Symptom validity tests are part of these guidelines.

The literature we examined referred to the following types of symptom validity tests: alternative selection procedure, tests with a simulated degree of difficulty, and tests on the identification of atypical performance profiles. All these tests have one feature in common: they are able to pinpoint inconsistencies between observed and expected performance.

Our systematic search of the scientific literature generated around 1,100 articles for the period 1997-2007. 570 of these are from the last five years, i.e. from 2003 up to and including 2007, of which 340 were published in 2005. There is a wealth of literature from the English-speaking world. However, for the purposes of the present study, we limited our choice to German-language publications which deal with IV-related subjects. In the end, we identified 30 such articles. Individual symptom validity tests (SVT), which meet the scientific quality criteria and have been tried and tested in practice, are described in more detail.

The following summary of the literature on SVT can be made. There is a broad spectrum of tests from various fields of application that are used to identify exaggeration and malingering. Despite intensive research on the subject, an intractable problem with these tests remains, including with test that have been subjected to rigorous checks: they have a restricted field of application and all produce a non-negligible share of false-positives. Such limitations notwithstanding, the procedure can yield useful information if applied in a targeted way and if care is taken when interpreting the results. However, for a full assessment of exaggeration/malingering several methods ought to be applied in combination. In this context, symptom validity tests can offer useful additional information if applied and interpreted properly.

Assessors' opinions: Critical evaluation of symptom validity tests

Most of the assessors we interviewed were of the opinion that exaggeration and malingering were not widespread problems. The majority also stated that malingering was not a problem they encountered frequently in their work. Their explanation for the higher share of malingering cases reported in scientific articles was the sample populations on which these studies were based as well as the exclusion of the social factors involved in the onset of chronic disorders.

A minority of the interviewed assessors use scientifically proven symptom validity tests (SVT) in their IV assessment work. They cited the following tests: personality inventories (MMPI-2 and PS-16), symptom checklists (SFSS, and SCL 90), cognitive memory tests (Rey Short-Term Memory, TOMM, WMT, short-term memory test from the Bremen SVT) and cognitive attention tests (Frankfurt atten-

tion inventory FAIR). More frequently assessors use procedures which have not received scientific backing to verify the consistency of clients' behaviour.

The background to the assessment of physical disorders, particularly neurological and rheumatological ones, is rather particular. Due to the poor availability of appropriate SVT, procedures are used which were not devised with symptom validity specifically in mind. These include, for example, the evaluation of functional performance combined with the Performance Assessment Capacity Test (PACT), the Waddell's signs test, the JAMAR grip test, psychophysical tests, blood tests to check for malcompliance, pressure pain threshold tests as well as screening for somatic pain disorders (SOMS2 and SOMS7).

Many assessors doubt whether individual SVT or combinations of tests would be able to detect exaggeration or malingering with sufficient reliability and certainty. According to the assessors, SVT merely provide a snapshot of the examinee at a specific moment in time. The inherent danger with these tests is that they may mask additional factors that should be decisive not only for the test but also for the entire assessment process and its final outcome.

The assessors agree in principle that more research and development needs to be undertaken in the SVT field. The external validity of an SVT, i.e. the transferability to broader client populations, is considered to be unsatisfactory. Many assessors, however, reject a rushed introduction of SVT as recommended by the FSIO. They would like to decide as independently as possible on the diagnostic instruments they will use in their IV assessment work.

The assessors reported that, when conducting IV assessments, they were under growing pressure to rate the examinees' fitness to work as highly as possible. They fear that this will negatively affect the independence, neutrality, objectivity and quality of their IV assessment work.

Application in practice

The use of symptom validity tests in the IV assessment process was thoroughly examined via a written survey. It was aimed at 18 medical observation centres (MEDAS) tasked with conducting IV assessments, and at the staff of the 10 IV Regional Medical Services (RAD). The latter are responsible for the supervision of medical assessment procedures with the aim of ensuring a high degree of standardisation nationwide. Over 150 people are employed in RAD and MEDAS. 30 persons responded and turned in the questionnaire. It is possible that responses are distributed non-randomly, implying that our findings may be subject to a certain selection bias. The results of the survey evaluation therefore cannot be considered representative. However, we are able to provide a qualitative or descriptive evaluation. The data concur with the heterogeneity of the opinions noted already in the interviews which dealt with the links between inconsistency, exaggeration and malingering. They also confirm a certain degree of scepticism among assessors with regard to the use of symptom validity tests.

The largely structured recording of inconsistencies appears to be uncontested. All test users in our random sample consider that this is important for the determination of a person's ability to work. A minority of assessors use standardised tests to record inconsistencies. In French-speaking Switzerland the use of non-standardised tests is rarer than in German-speaking Switzerland. This difference, however, should be interpreted with caution due to the fluid definition of what constitutes a non-standardised test, as well as the non-representative sample, especially given that a greater

number of individual questionnaires were returned from the same organisation in French-speaking Switzerland than from German-speaking Switzerland.

74% of assessors consider inconsistency a reasonable or strong indicator of exaggeration; 59% see it as a reasonable or strong indicator of malingering. 89% of assessors stated that it was possible rather or very often to ascertain whether an examinee had adopted exaggerating behaviour. In terms of malingering, only 38% agreed. A further 35% of assessors said that it was very rare to be able to reach such a conclusion. 85% of assessors considered that the identification of exaggeration could or should be one of their responsibilities. With regard to the identification of malingering, 65% agreed.

Respondents cited the use of the following tests to determine inconsistencies: the Waddel's signs test, MMPI, PACT and HAMD, as well as various laboratory and drug screening tests. The most frequently used were the Waddell's signs test (very often), PACT (very often), MMPI (often) and HAMD (very often).

Assessors also cited the use of the following tests to determine exaggeration and malingering, MMPI, ASTM as well as other non-specified neuropsychological tests. Most tests were rarely used. The most frequently used tests include HADS (very often), HAMD (very often), MADRS (very often), MMST (often) as well as the Waddell's signs test (very often).

The additional comments by the assessors on open-ended questions correspond largely to those which had already been mentioned in the interviews. These include limited ecological and external validity of SVT test batteries, the importance of clinical experience, the importance of adequate training in both administering and interpreting these tests (including awareness of the dangers inherent in their appropriate application), the danger of shifting a legal but not medical problem on to pseudo-objectivity, and repeated reference to the fact that inconsistencies are necessary but not sufficient for proving malingering or exaggerating behaviour.

Conclusions and recommendations

Based on the findings of the present study, the following recommendations can be made:

1. The quality of assessments involving symptom validity has yet to be gauged. With a view to further improvements, an analysis of the use of symptom validity in the IV assessment procedure is required;
2. Assessments involving symptom validity should be systematic and based on recognised guidelines; where possible, standardised SVTs should be used;
3. Development of basic and advanced training programmes in the field of insurance medicine on symptom validity testing and on the associated issue of the diagnosis of disorders which do not lend themselves to objective testing;
4. Little is known about the prevalence of exaggeration and malingering in Switzerland. The responsible use of SVTs in the IV assessment procedure relies on the scientific identification of such cases.
5. Encouragement of scientific validation and development of symptom validity tests for use in the IV assessment procedure.

Glossar

A posteriori Wahrscheinlichkeit: Die Wahrscheinlichkeit einer Aussage oder die Wahrscheinlichkeit der Präsenz eines Merkmals nach einer Untersuchung, zum Beispiel einen Test oder eine Begutachtung.

A priori Wahrscheinlichkeit: Die Wahrscheinlichkeit einer Aussage oder die Wahrscheinlichkeit der Präsenz eines Merkmals vor einer Untersuchung, zum Beispiel einen Test oder eine Begutachtung.

Aggravation: Übertreibung und/oder Ausweitung vorhandener Beschwerden zur Erreichung eines bestimmten Ziels.

AI: Assurance-Invalidité.

Analogstudie: Im Kontext der Validierung von BVT: eine Studie, in der normale Probanden den Auftrag erhalten den Gold-Standard, das Malingering, zu simulieren.

Artifizielle Störung: Als psychische Störung aufgefasste, zielgerichtete Vortäuschung oder Erzeugung von Symptomen oder Krankheiten mit einem primären Krankheitsgewinn.

Augenscheinvalidität: Der Test, und somit die Ergebnisse, sind rein vernünftig betrachteter gültig. Die Augenscheinvalidität beruht auf Expertenmeinung, nicht auf einer formellen wissenschaftlichen Überprüfung, und bildet die niedrigste Stufe der Validität.

Bayes-Theorem: Die a posteriori Wahrscheinlichkeit der Präsenz eines Merkmals (der prädiktive Wert) ist eine Funktion der a priori Wahrscheinlichkeit (hypothetisierte Prävalenz) und der empirischen Evidenz (Sensitivität und Spezifität).

Below-chance Verfahren: Verfahren, die Antwortverhalten unter Zufall quantifizieren und als Hinweis auf Simulation interpretieren.

Beschwerdevalidierungstest (BVT): Standardisiertes Verfahren zur Überprüfung der Glaubwürdigkeit und Plausibilität der von den Klienten und Klientinnnen geschilderten Beschwerden und Beeinträchtigungen.

BSV: Bundesamt für Sozialversicherungen.

COMAI: Centres d'Observation Médicale de l'AI (COMAI), in der Deutschschweiz: MEDAS.

DSM-IV: Diagnostic and Statistical Manual of Mental Disorders Version 4.

Falsch-Negativ-Rate: Wahrscheinlichkeit eines negativen Tests gegeben Präsenz des Merkmals, das heisst, wenn das richtige Testergebnis positiv wäre.

Falsch-Positiv-Rate: Wahrscheinlichkeit eines positiven Tests gegeben Absenz des Merkmals, das heisst, wenn das richtige Testergebnis negativ wäre.

FoP-IV: Forschungsprogramm zur IV des Bundesamtes für Sozialversicherungen.

ICD-10: Internationale Klassifikation der Krankheiten Version 10.

Inhaltsvalidität: Der Grad der Abdeckung von einzelnen Items für das zu untersuchende Konstrukt. Beispiele von ‚Konstrukte' sind Depression, Konzentration, und Persönlichkeitsstruktur.

Inkonsistenz: Widersprüchlichkeit, d.h. Diskrepanz zwischen dem erwarteten Verhalten (Leistungen, Fähigkeiten) der Klientinnen und Klienten aufgrund der von ihnen geschilderten Beschwerden und dem beobachteten Verhalten.

Item: Erhebungseinheit; bei einem Test oder Fragebogen: kleinste Einheit, Aufgabe oder Frage.

IV: Invalidenversicherung.

IVST: IV-Stelle.

Known-Groups Design: Wird verwendet in Studien zur Erhebung der Konstruktvalidität um festzustellen, ob ein Instrument bekannte, bezüglich dem interessierenden Konstrukt unterschiedliche Gruppen auch als unterschiedlich bewertet.

Konstruktvalidität: Der Grad der Erfassbarkeit eines theoretischen Konstrukts durch ein Messinstrument. Beispiele von ‚Konstrukte' sind Depression, Konzentration, und Persönlichkeitsstruktur.

Kreuzvalidierung: Wiederholung einer Validierungsstudie bei einer neuen Probandengruppe. Das aus einer ersten Studie bei einer bestimmten Stichprobe entwickelte Modell oder Testverfahren sollte das Outcome in einer neuen Teststichprobe ebenfalls hinreichend gut voraussagen. Ergebnisse einzelner Studien sind nur beschränkt allgemein anwendbar.

Kriteriumsvalidität: Der Grad der Übereinstimmung zwischen den Outcomes eines Tests und den Outcomes eines Gold-Standards, das „Kriterium".

MEDAS: Medizinische Abklärungsstellen der IV.

Negativ Prädiktiver Wert: Wahrscheinlichkeit der Absenz des Merkmals gegeben negativer Test.

Negative Antwortverzerrung: Oberbegriff für Simulation und Aggravation.

Nicht-zielkonforme Leistungen der IV: Unrechtmässige Inanspruchnahme von IV-Leistungen durch versicherte Person (z.B. fahrlässiges Verhalten) oder nicht-zustehende Leistungen, die im Wesentlichen das Ergebnis von systembedingten Fehlern (z.B. Fehleinschätzungen im Rahmen der Abklärung) sind.

Objektivität: Wertfreiheit, Unparteilichkeit, Unvoreingenommenheit, grösstmögliches Ausschalten von Gefühlen und Vorurteilen. Der Begriff Objektivität wird fälschlicherweise oft mit der Abwesenheit von Fehler in der Beurteilung assoziiert.

OFAS: Office fédéral des assurances sociales.

Positiv Prädiktiver Wert: Wahrscheinlichkeit der Präsenz des Merkmals gegeben positiver Test. Siehe Seite 18 ff.

Prävalenz: Relative Häufigkeit eines Merkmals. Siehe Seite 18 ff.

Prosecutor's fallacy: Risiko einer falschen Beurteilung von bedingten Wahrscheinlichkeiten durch Nicht-Beachten der Prävalenz und der Problematik von multiplen Tests. Siehe Seite 18.

RAD: Regionale ärztliche Dienste.

Reliabilität: formale Genauigkeit einer Untersuchung.

Sensitivität: Wahrscheinlichkeit eines positiven Tests gegeben Präsenz des Merkmals. Siehe Seite 18 ff.

Simulation (engl. Malingering): Absichtliche, reflektierte, zweckvolle Vortäuschung von nicht-vorhandenen Symptomen oder fälschliche Beschwerdenschilderung zur Erreichung eines bestimmten Ziels (z.B. Erhalten einer Rente).

Skala: Abbildung eines empirischen Relativs in ein numerisches Relativ; im Rahmen von Tests auch als Ausdruck eines, aus der Kombination von einzelnen Fragen gewonnen, Wertes.

SMR: Services médicaux régionaux, in der Deutschschweiz: RAD.

Spezifität: Wahrscheinlichkeit eines negativen Tests gegeben Absenz des Merkmals. Siehe Seite 18 ff.

Testgütekriterien: Objektivität, Reliabilität und Validität.

Validierungsstudie: Studie, die die Reliabilität und Validität eines Messinstrumentes untersucht.

Validität: (von Lat. validus: kräftig, wirksam; engl. „validity“, Gültigkeit) Bei der internen Validität betrifft die Gültigkeit von Aussagen in Studien, Tests oder Gutachten für die untersuchten Personen. Die externe Validität von Aussagen in Studien betrifft die Frage, ob sich die Resultate auch auf andere Personen, Populationen und Testsituationen verallgemeinern lassen.

1 Einleitung, Problemstellung

1.1 Ausgangslage

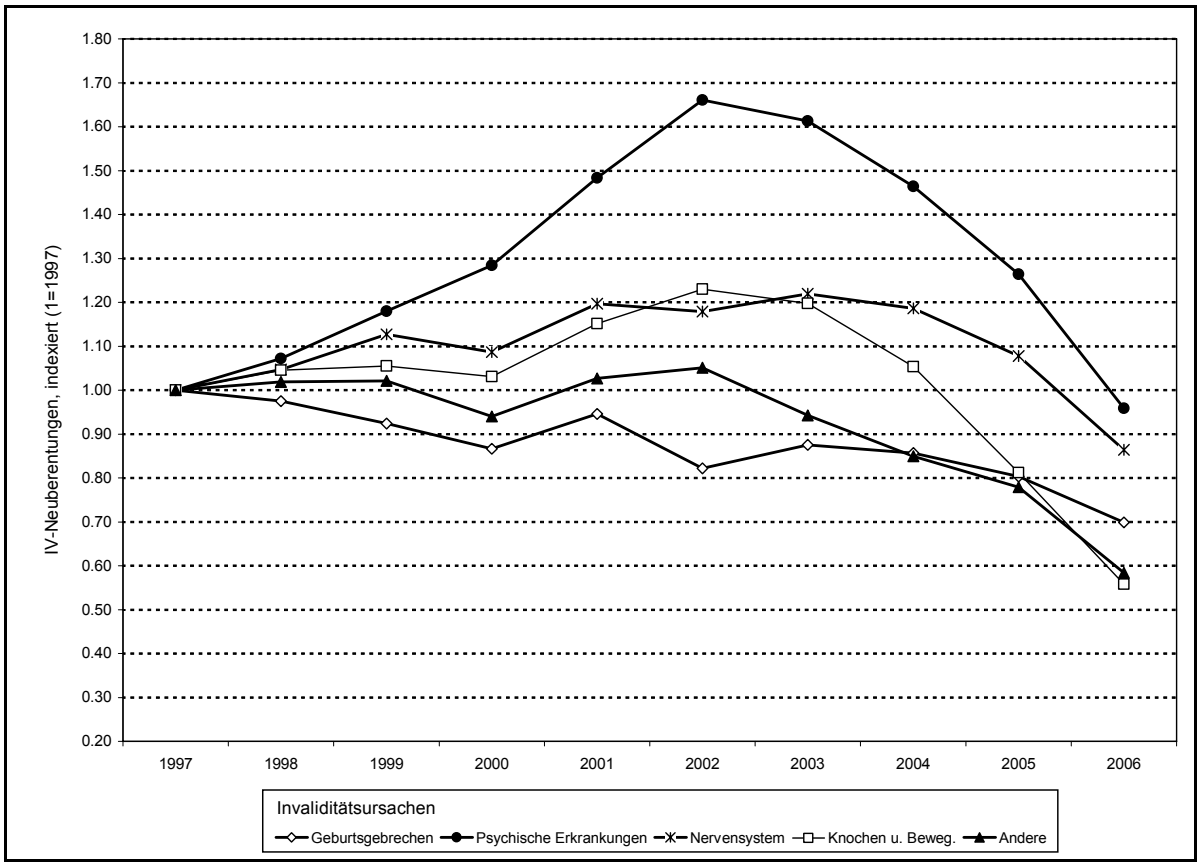
In den letzten Jahren wurde in der Schweizer Politik, in den Medien und in der Öffentlichkeit häufiger und auch vehementer die Frage nach der Grössenordnung des ungerechtfertigten Bezugs von Leistungen der Sozialhilfe und -versicherungen aufgeworfen. Hintergrund war insbesondere das jahrelange Wachstum der IV-Renten und damit verbunden die zunehmende Verschuldung der IV.

So nahmen die IV-Renten zwischen 1998 und 2007 um 42% (von 180'000 auf 253'000) zu. Seit 2003 kam es jedoch zu einer Abflachung des Wachstums und im Jahre 2007 sogar zu einem leichten Rückgang der Zahl der Rentenbezüger um 1.2% (Buri, Härter, & Sottas, 2007). Dies ist besonders auf eine deutliche Abnahme der Neuberentungen seit 2002 zurückzuführen: Während zwischen 1997 und 2002 ein durchschnittliches jährliches Wachstum von 4% bei den Neurenten zu verzeichnen war, so ergab sich zwischen 2002 und 2007 eine jährliche Abnahme von 11%. Die genauen Ursachen dieser Abnahme sind nicht klar, es können jedoch folgende Hintergründe angeführt werden: Abnahme der Erstanmeldungen für IV-Leistungsgesuche; restriktivere Praxis bei der Beurteilung von Berentungsgesuchen, insbesondere bei schwer objektivierbaren Gesundheitsschäden; die Einführung der regionalen ärztlichen Dienste (RAD). Die Abnahme der Neuberentungen ist ausserdem unabhängig von der Invaliditätsursache zu beobachten, besonders ausgeprägt ist sie jedoch bei den psychischen Erkrankungen (vgl. Abbildung 1).

Trotz dieser Abnahme hat allerdings der relative Anteil der psychischen Erkrankungen als Invaliditätsursache weiter zugenommen, indem dieser 2006 bei 41% (1997: 30%) lag, während die relative Bedeutung der Erkrankungen der Knochen und des Bewegungsapparates abgenommen hat von 27% auf 21%.

Einzelne Politiker sehen als Mittel zum Abbau des Schuldenberges der IV primär eine rigorose Kontrolle und ggf. Infragestellung von Rentenbezügen unter der Annahme einer mehr oder weniger grossen Dunkelziffer von ungerechtfertigt bezogenen Leistungen. Die politische Debatte um Scheinrenten wurde hart an der Grenze zur Verunglimpfung von Kranken und Behinderten geführt, eine sachlichere Beurteilung des Themas tut deshalb not. Die Debatte hat aber auch zu einem Tabubruch geführt, indem die Angemessenheit von Leistungsbezügen explizit diskutiert und benannt werden kann. Die Diskussion dieser Thematik − auch im wissenschaftlichen Kontext − ist im Ausland, besonders in den angelsächsischen Ländern, wesentlich weiter fortgeschritten als in der Schweiz. Das Ergebnis sind eine Reihe von Lehrbüchern und Readern zum Thema der so genannten 'Malingering'-(dt.: Simulation-) Forschung (vgl. Halligan, Bass, & Oakley, 2003b; Larrabee, 2007; Richard Rogers, 1997). Wesentliche Aspekte dieser Forschung werden in diesem Bericht im Kapitel 2 komprimiert dargestellt.

Abbildung 1 Entwicklung IV-Neuberentungen nach Invaliditätsursachen 1997-2006 (indexiert, 1997=1.00; Daten: IV-Statistik 2007)



Als erster Schritt zu einer Versachlichung der Debatte ist eine Präzisierung der Begriffe notwendig. Die im Auftrag des BSV durchgeführte Studie von Ott, Bade und Wapf (2007) schlägt als Oberbegriff die *"nicht-zielkonformen Leistungen"* der IV vor, da die Fokussierung auf den Missbrauch zu einer unfruchtbaren Verengung der Diskussion führt. Unter den nicht-zielkonformen Leistungen differenzieren Ott et al. zwischen:

- *unrechtmässiger Inanspruchnahme der IV,* d.h.: vorsätzlich missbräuchliches und fahrlässiges Verhalten der Versicherten oder von Dritten ohne allfälliges Fehlverhalten des Versicherers;
- *nicht zustehender Leistungsausrichtung der IV,* d.h.: system-/prozessbedingte nicht-zielkonforme Leistungen, zu denen die versicherte Person keinen oder nur einen marginalen Beitrag leistet (z.B. Fehleinschätzungen der IV-Stellen, der RAD und weiterer Gutachtenden).

Diesen beiden Arten nicht-zielkonformer Leistungen können eine Reihe von Unterkategorien zugeordnet werden, wie sie in Tabelle 1 zusammengefasst sind.

Tabelle 1 Übersicht über nicht-zielkonforme Leistungen der IV nach Ott et al. (2007, S.40-41)

Unrechtmässige Inanspruchnahme der IV	Nicht zustehende Leistungsausrichtung der IV
A: Vorsätzlicher Versicherungsbetrug • Simulation von Krankheiten oder Unfallfolgen. • Bewusste, massive Aggravation: Bewusstes Übertreiben oder Betonen von vorhandenen Krankheitssymptomen. • Bewusst falsche Angaben zur früheren Erwerbstätigkeit betreffend Tätigkeit, Anstellungsgrad oder Erwerbseinkommen. • Meldepflichtverletzungen: Unterlassene Meldung von Arbeitsleistungen, von veränderten Arbeitspensen und/oder von höheren Einkommen durch Rentenbeziehende. • Grobe Verletzungen der Schadenminderungspflicht und der Mitwirkungspflicht. Z.B. klar unkooperatives Verhalten, Förderung oder bewusstes Inkaufnehmen einer Verschlechterung des Gesundheitszustandes, Teilnahmeverweigerung an Massnahmen.	• Misslungene Abgrenzung der Krankheits- oder Unfallfolgen von sozio-ökonomischen und soziokulturellen Faktoren: die vorhandene Arbeits- oder Erwerbsunfähigkeit ist nicht oder nicht in ausreichendem Ausmass gesundheitsbedingt, trotzdem wird eine Leistung zugesprochen. • Unrichtige Einschätzung der Arbeitsunfähigkeit durch die RAD oder die beauftragten Gutachtenden bzw. unrichtige Einschätzung der Zumutbarkeit durch die IV-Stelle. • Der oder die Versicherte ist zwar ursprünglich integrationsbereit, aber aufgrund langwieriger Verfahren, schlechter Beratung oder Ähnlichem gelingt die Wiedereingliederung nicht. • Unrichtige Bestimmung des IV-Grades. • Mangelhafte Durchführung von Revisionen (vor allem aufgrund personeller Engpässe).
B: Fahrlässige Verstösse gegen das IVG • Bewusste, geringfügige Aggravation. • Verletzung der Schadensminderungspflicht: die versicherte Person bemüht sich nicht selbst in zumutbarem Rahmen um eine Verbesserung oder Erhaltung des bestehenden Gesundheitszustande, z.B. durch das Durchführen vom Arzt verordneter Leibesübungen bei Rückenschmerzen. • Verletzungen der Mitwirkungspflicht: die versicherte Person trägt nicht in zumutbarer Art und Weise zum Erfolg von Wiedereingliederungsmassnahmen bei.	

Aufgrund der begrenzten Datenlage können Ott et al. lediglich eine Schätzung der Grössenordnung nicht-zielkonformer Leistungen der IV vorlegen, wobei die getrennte Berechnung nach unrechtmässiger Inanspruchnahme und nicht zustehenden Leistungen nicht möglich war. Ott et al. schätzen ein

Potenzial von 8-10% des Rentenbestandes, das nicht zielkonform entrichtet wird. Mit Blick auf Neurenten kommen sie zur Schätzung einer Bandbreite von 8-18% nicht-zielkonformer Leistungen. Die im Rahmen von Experteninterviews erfolgte qualitative Bewertung der Relevanz verschiedener Arten nicht-zielkonformer Leistungen für die Rentenfallzahlen zeigt ausserdem folgendes: Bei der Ausstellung von IV-Renten wird der nicht-zustehenden Leistungsausrichtung bzw. systembedingten Faktoren eine hohe Relevanz beigemessen, von mittlerer Bedeutung sind fahrlässige Verstösse und als gering wird die Bedeutung des vorsätzlichen Versicherungsbetruges eingeschätzt. Ausserdem zeigte sich, dass das grösste Potenzial zur Reduktion nicht-zielkonformer Leistungen bei vollzugs- und systembedingten Aspekten liegt, die zu nicht zustehender Leistungsausrichtung bei der IV führen.

Aufgrund dieser Konzipierung ist der Gegenstand der vorliegenden Studie, Simulation und Aggravation als Teilmenge nicht-zielkonformer Leistungen anzusehen, wobei die Unterkategorien der fahrlässigen Verstösse gegen das IVG und des vorsätzlichen Versicherungsbetruges tangiert sind. Wichtig ist die Feststellung der Teilmenge: aufgrund der befragten Experten in der Studie von Ott et al. ist der Anteil von durch Simulation oder Aggravation erzielter nicht-zielkonformer Leistungen der IV als relativ klein zu bewerten.

Prädestiniert für die Ausrichtungen nicht-zielkonformer Leistungen sind nicht oder nur begrenzt objektivierbare Gesundheitsstörungen. Erst hier entsteht ein relevanter Ermessensspielraum für die Einschätzung der Erwerbsunfähigkeit und des IV-Grades. Zu den schwierig objektivierbaren Störungen werden insbesondere gezählt: chronische Rückenschmerzen ohne somatisch erkennbare Ursache, andere Schmerzkrankheiten, Schleudertraumata, Depressionen, Burn-out.

Für eine Minderung nicht-zielkonformer Leistungen werden u.a. die für die vorliegende Studie besonders relevanten Empfehlungen gemacht, weil sie die konkrete Abklärungspraxis betreffen:

- die Einschätzung der Arbeitsfähigkeit erfolgt durch den Versicherungsarzt, die Vertrauensärztin und nicht durch den behandelnden Arzt, die Ärztin;
- Ausbau der versicherungsmedizinischen Abklärungskapazitäten, Unterstützung der Qualitätsentwicklung versicherungsmedizinischer Abklärungen: dazu zählen neben der fachlichen Weiterentwicklung der Versicherungsmedizin in der Schweiz u.a. auch die Entwicklung von Standards bei der Begutachtung von schwierig objektivierbaren Gesundheitsbeeinträchtigungen.

Eine Reihe von Verbesserungen in der Abklärungspraxis sind durch die Einführung der Regionalen Ärztlichen Dienste (RAD) bereits erfolgt, wie die Evaluation von Wapf und Peters (2007) zeigt. Diese Verbesserungen betreffen u.a. die folgenden Aspekte: Vereinheitlichung der medizinischen Grundlagen, verbessertes versicherungsspezifisches Wissen der Ärzteteams, erhöhte Qualität der Dossierbeurteilungen. Nicht genauer untersucht bei diesen Massnahmen ist indessen das konkrete Vorgehen bei der Begutachtung schwierig objektivierbarer Gesundheitsbeeinträchtigungen und insbesondere die allfällige Anwendung von Verfahren und Tests zur Identifikation von Simulation und Aggravation.

1.2 Zielsetzungen, Fragestellungen

1.2.1 Allgemeine Zielsetzung

Die vorliegende Studie soll Kenntnisse zum aktuellen Stand der wissenschaftlich fundierten Entwicklung von anerkannten und validierten Beschwerdevalidierungstests (BVT) liefern und die Anwendung von und der Umgang mit BVT in der nationalen und internationalen neuropsychologischen und berufsbezogenen multidimensionalen Diagnostik darstellen. Die Untersuchung liefert eine kritische Bewertung und Dokumentation dieser Tests und ihrer Anwendungen. Ausserdem sollen im Rahmen dieses Berichtes Grundlagen für die Diskussion über die Abklärungsverfahren in der IV und für eine (verbesserte) Implementierung von BVT in den Abklärungs- und Gutachterstellen der Sozialversicherungen bereitgestellt werden.

1.2.2 Fragestellungen

Die folgenden spezifischen Fragestellungen wurden im Rahmen dieser Studie untersucht:

1. Welche Beschwerdevalidierungstests sind in der wissenschaftlichen Fachliteratur validiert worden und wie ist diese Validität zu beurteilen?
2. Welche BVT werden in den verschiedenen Settings der Sozial- und Unfallversicherungen (SUVA, IV, KK, Taggeld, Haftpflicht) angewandt, und wie werden sie dort beurteilt?
3. Wie können die Erfahrungen mit der Diagnostik im neuropsychologischen Kontext in andere Fachbereiche, die in der Beurteilung versicherungsrelevanter Beschwerden tätig sind, transferiert und operationalisiert werden?

1.3 Konzeption der Untersuchung

1.3.1 Übersicht

Im Rahmen der vorliegenden Studie wurden drei methodische Zugänge zum Untersuchungsgegenstand gewählt:

1) eine systematische Literaturrecherche;

2) problemzentrierte Experteninterviews;

3) schriftliche Befragung bei (potenziellen) Anwendern und Anwenderinnen von Beschwerdevalidierungstests.

Diese drei Zugänge werden im Folgenden kurz beschrieben. Die detaillierte Darstellung des jeweiligen methodischen Vorgehens erfolgt in den entsprechenden Ergebniskapiteln (3.1, 4.1 und 5.1) direkt 'vor Ort'.

1.3.2 Systematische Literaturrecherche

Mit einer systematischen Literaturrecherche sollen potenzielle Messinstrumente und Beobachtungsverfahren im Bereich der Beschwerdevalidierung ermittelt werden. Mit zuvor festgelegten Selektionskriterien wird die für den Auftrag relevante Literatur (Tests, Validierungsstudien etc.) eingegrenzt. Es soll *multidimensional*, bezüglich kognitiven-mentalen und physischen Aspekten gesucht werden (Bianchini, Greve, & Glynn, 2005). Die Recherche erfolgt in relevanten Datenbanken und Bibliotheken, nämlich in:

a) Online-Datenbanken wie Medline/PubMed, PsychInfo, Cinahl, ISI Web of Science

b) Testarchive: Psyndex (via DIMDI oder ZPID).

Anschliessend wird die erfasste Literatur auf der Grundlage eines Analyserasters quantitativ und qualitativ ausgewertet. Das Analyseraster enthält zentrale Bewertungskriterien für relevante Messinstrumente.

1.3.3 Experteninterviews

Mit leitfadengestützten Experteninterviews (Bogner, 2005; Lamnek, 2005) sollen übergeordneten Frage des Einsatzes von Beschwerdevalidierungstests (BVT) in der Abklärungspraxis geklärt werden. Dabei werden die Resultate der Literaturrecherche soweit als möglich integriert. Insbesondere kann die Operationalisierung und die Validität der angewendeten Testinstrumente in der Praxis mit der Fachliteratur verglichen werden. Um die unterschiedliche Situation in den Landesteilen der Schweiz adäquat abzubilden, werden Experten und Expertinnen sowohl in der deutschsprachigen Schweiz als auch in der Romandie befragt.

Die entsprechenden Experten und Expertinnen sollten bereits Erfahrung in der Beurteilung von BVT haben und mit dem wissenschaftlichen Formalismus der Diagnostik (Gütekriterien, Wahrscheinlichkeitsaussagen usw.) vertraut sein. Es wurden Experten und Expertinnen aus den folgenden Personengruppen berücksichtigt, wobei Interviews mit 10-15 Fachleuten vorgesehen waren:

- *‚Testentwickler'* bzw. wissenschaftliche Experten und Expertinnen (aus den Fachbereichen der Neuropsychologie, Testdiagnostik, evt. Rehabilitationsmedizin);
- *Anwender und Anwenderinnen* bzw. Gutachtende, Vertrauensärzte und –ärztinnen, u.a. Versicherungsmediziner in den Rehabilitationskliniken der Unfallversicherung, in den Evaluationszentren der IV, und bei den Haftpflicht- und Taggeldversicherern; IV-Beratende.

Im Rahmen der Interviews sollte untersucht werden, ob überhaupt und wenn ja, welche Tests in den entsprechenden Settings routinemässig eingesetzt werden. Weitere Fragen, die u.a. untersucht wurden, waren: Inwieweit sind BVT in der Praxis anwendbar? Liefern sie praxisrelevante Befunde? Wie sind die Tests operationalisiert? Wie wird mit widersprüchlichen Resultaten umgegangen? Welche Konsequenzen haben positive Testbefunde für die betroffenen Gesuchstellenden bzw. Patientinnen und Patienten. Ausserdem sind potentielle Vorbehalte (z.B. ethisch begründete Bedenken etc.) gegenüber der Anwendung von BVT zu berücksichtigen. Die Auswertung der transkribierten

Experteninterviews erfolgte nach dem Ansatz der qualitativen Inhaltsanalyse nach Mayring (Mayring, 2003; Mayring & Gläser-Zikuda, 2005)[1]. Als Software wird ATLAS TI eingesetzt.

1.3.4 Anwenderbefragung

Auf der Grundlage der Erkenntnisse der Experteninterviews sollte anschliessend eine strukturierte Befragung der (potenziellen) Anwenderinnen und Anwender (s.o.) von BVT durchgeführt werden. Hauptzielgruppe waren Fachpersonen der medizinischen Abklärungsstellen MEDAS, die zu Handen der IV Gutachten erstellen, und Mitarbeitende der Regionalärztlichen Dienste RAD, die vorwiegend versicherungsmedizinische Stellungnahmen und Berichte sowie seltener auch Gutachten erstellen.

Der Fragebogen sollte Problembereiche in der Anwendung von BVT in der Praxis beinhalten, die sich im Rahmen der Experteninterviews verdeutlicht haben. Das Instrument enthielt geschlossene und offene Fragen zu den relevanten Punkten.

Die Auswertung der strukturierten Befragung erfolgte mit einfachen deskriptiv-statistischen Methoden in SPSS.

[1] Dieses Verfahren umfasst drei zentrale Arbeitsschritte: (1) Zusammenfassung: Reduktion des Ausgangstextes auf eine überschaubare Kurzversion der wichtigsten Inhalte; (2) Explikation: Klärung unklarer Textbestandteile; (3) Strukturierung: Entwicklung und schrittweise Verfeinerung eines inhaltlichen Kategorienschemas und Zuordnung der Textinhalte.

2 Theoretische und begriffliche Grundlagen

2.1 Simulation und verwandte Konstrukte

2.1.1 Definition von Simulation (engl. Malingering)

Der Begriff Simulation ist in der psychiatrischen Diagnostik eine anerkannte Diagnose. Sie hat jedoch in den beiden etablierten diagnostischen Klassifikationssystemen, DSM-IV (Sass, Wittchen, Zaudig, & Houben, 1998) und ICD-10 (Dilling, Mombour, & Schmidt, 1993), einen unterschiedlichen Stellenwert. In der ICD ist Simulation einer Zusatzkategorie, den so genannten Z-Diagnosen (Z76.5) zugeordnet ohne weitere Angaben zur Feststellung von Simulation; die Z-Diagnosen betreffen 'Faktoren, die den Gesundheitszustand beeinflussen und zur Inanspruchnahme von Gesundheitsdiensten führen'. Im DSM-IV ist Simulation der Gruppe der 'anderen klinisch relevanten Problemen' zugeordnet (V65.2). Das DSM macht jedoch genauere Angaben, wann Simulation diagnostiziert werden kann:

1. Symptomdarbietung steht im forensischen Kontext;

2. Deutliche Diskrepanz zwischen der berichteten Belastung/Behinderung und objektiven Befunden;

3. Mangel an Kooperation bei den diagnostischen Untersuchungen und den verordneten Behandlungsmassnahmen;

4. Vorhandensein einer antisozialen Persönlichkeitsstörung.

Differentialdiagnostisch abgegrenzt wird Simulation von vorgetäuschten Störungen, Konversionsstörungen und anderen somatoformen Störungen (s.u.) Die ICD macht kaum Angaben zur Diagnosestellung; betont wird, dass Simulation bewusst erfolgt.

Die DSM-IV-Definition wurde indessen kritisiert, da sie kein Urteil über die Bewusstheit des Verhaltens und der Motivation des potenziellen Simulanten erlaubt (Halligan, Bass, & Oakley, 2003a; Larrabee, 2007). Von Rogers (1997) wurde ausserdem auf die inadäquate Assoziation von Simulation mit psychischen Erkrankungen und/oder dissozialem Verhalten hingewiesen. Rogers schlägt demgegenüber eine Konzeption von Simulation als adaptives Verhalten vor: Demnach vollzieht ein potenzieller Simulant eine Kosten-Nutzen-Analyse, wenn er mit einer Abklärungssituation konfrontiert ist, die er in Bezug auf seine Bedürfnisse als indifferent oder gar negativ bewertet. Das adaptive Erklärungsmodell sagt eine Häufung von simulierendem Verhalten voraus in folgenden Situationen: (1) die Rahmenbedingungen einer Abklärung werden vom Klienten oder von der Klientin als 'feindlich' erlebt, oder (2) die persönlichen Interessen des Klienten, der Klientin sind stark tangiert bzw. es steht viel auf dem Spiel, oder (3) keine Alternativen zu simulierendem Verhalten scheinen offen zu stehen.

Eine differenziertere Erfassung von Simulation haben Slick, Sherman und Iverson (1999) vorgelegt für den Bereich neuro-kognitiver Beeinträchtigungen. Dieses als "Slick-Leitlinen" zur Identifikation von Simulation breit anerkanntes Verfahren wurde ausserdem von Bianchini et al. (2005) erweitert, sodass auch der Bereich chronischer Schmerzen abgedeckt ist; eine genauere Darlegung des Verfahrens von Bianchini et al. folgt im Kapitel Leitlinien ab Seite 25. An dieser Stelle ist das Wesentliche der Slick- und Bianchini-Leitlinien lediglich knapp zusammengefasst, um den Unterschied ge-

genüber der DSM-IV-Definition von Simulation darzulegen. Bianchini et al. schlagen 5 Kriterien (A-E) vor, wovon 2 spezifische Kriterien (A und E) immer erfüllt sein müssen, dass mit ausreichender Wahrscheinlichkeit von Simulation gesprochen werden kann. Die verbleibenden Kriterien (B, C, D) sind relevant für die Erfassung des Schweregrades von Simulation. Die 5 Kriterien von Bianchini et al. lauten wie folgt:

- A: Nachweis relevanter externer Anreize für Simulation (z.B. Rentenabklärungen);
- B: Hinweise auf Aggravation der Behinderung im Rahmen der klinischen Untersuchung (z.B. negative Leistungsverzerrung, Inkonsistenzen);
- C: Hinweise auf Aggravation der Behinderung im Rahmen von kognitiven und neuropsychologischen Tests;
- D: Hinweise auf Aggravation der Behinderung im Rahmen der Beschreibung der Symptome und Beschwerden durch den Klienten, die Klientin;
- E: die Befunde der Kriterien A bis D sind nicht vollständig durch psychiatrische, neurologische oder entwicklungsbedingte Faktoren erklärbar.

Im Vergleich zur DSM-IV Definition fallen somit folgende Bedingungen weg: die Fokussierung auf den forensischen Kontext, die mangelnde Compliance der Klientinnen und Klienten mit den Gutachtenden [2], die Koppelung von Simulation mit dem Konstrukt der Persönlichkeitsstörungen.

In der Fachliteratur wird oft zwischen verschiedenen Ausprägungsgraden von Simulation differenziert. So unterscheidet etwa Rogers (1997) zwischen drei Schweregraden von Simulation ('mild', 'moderate', 'severe'), die sich auf das Ausmass der Symptomdarstellung bezieht (z.B. ob bestehende Beschwerden nur verstärkt oder aber gänzlich neu 'erfunden' werden). Rogers grenzt Simulation auch ab von *Unzuverlässigkeit (Unreliability)* und von *Dissimulation (Defensiveness).* Ersteres kann auf der phänomenalen Ebene nahe bei simulierendem Verhalten liegen, die Intention des Klienten, der Klientin lässt sich jedoch nicht klar identifizieren. Letzteres, Dissimulation ist das Gegenstück von Simulation, also die Leugnung, Verschleierung von Behinderungen und Beeinträchtigungen (z.B. bei Suchtpatientinnen und Suchtpatienten).

Die Begrifflichkeit im deutschen Sprachraum zum Phänomen der unglaubwürdigen Darstellung von Beschwerden im Gutachtenskontext ist differenzierter als im englischen. (Dohrenbusch, 2007) unterscheidet mit Bezug auf die Empfehlungen des Verbandes Deutscher Rentenversicherer zwischen drei Konzepten:

- Verdeutlichung,
- Aggravation,
- Simulation.

[2] Dazu bemerkt Dohrenbusch (2007), dass ein Mangel an Kooperation auch Ausdruck entmutigender Erfahrungen des Klienten mit diagnostischen und therapeutischen Bemühungen zur Klärung seines Leidens zu tun haben kann. D.h. mangelnde Kooperation muss nicht einfach, eine willentliche Verzerrung der Beschwerden sein, um einen bestimmten externen Anreiz (z.B. Rente) zu erhalten.

Verdeutlichung wird dabei als ein häufig in Abklärungssituationen anzutreffendes und als 'legitim' zu bezeichnendes Verhalten verstanden: der Klient, die Klientin verdeutlicht Beschwerden, um sicher gehen zu können, dass der Gutachter, die Gutachterin ihn ernst nimmt. Verdeutlichung könnte auch teilweise Rogers' Begriff der Unzuverlässigkeit (Rogers, 1997) zugeordnet werden.

Demgegenüber bezeichnet *Aggravation* die bewusste Verstärkung vorhandener Beschwerden zu bestimmten, klar erkennbaren Zwecken. *Simulation* bezieht sich auf das bewusste Vortäuschen nicht-vorhandener Beschwerden oder Symptome. In der englischsprachigen Fachliteratur wird synonym zu Aggravation teilweise auch von 'symptom exaggeration' und zu Simulation von 'symptom fabrication' gesprochen (vgl. Rogers, 1997).

Die Begriffe Verdeutlichung und Unreliability bringen somit zum Ausdruck, dass ein Teil des Spektrums von unglaubwürdiger Symptomdarbietung in einer Gutachtenssituation als legitimes oder akzeptables Verhalten zu betrachten ist. Indessen bleibt die Schwierigkeit bestehen, eine klar bestimmte und v.a. objektiv messbare Grenze zu definieren zwischen Verdeutlichung auf der einen Seite und Aggravation oder Simulation auf der anderen Seite.

2.1.2 Konzept der negativen Antwortverzerrung (response bias)

Merten (2008, im Druck) schlägt anstelle von Simulation den übergeordneten Begriff der 'negativen Antwortverzerrung' und auf der Ebene der Diagnostik den Begriff der 'Beschwerdevalidierung' vor. Merten ist der Ansicht, dass diese Begriffe nicht nur politisch-korrekter sondern auch der Sachlichkeit des Diskurses in diesem Themenbereich dienlich sind.

Negative Antwortverzerrung (N.A.) wird definiert als das Bemühen eines Klienten oder einer Klientin, den Gutachter, die Gutachterin durch *ungenaue oder unvollständige Antworten* oder durch die *Demonstration eingeschränkter Leistungen* in der Abklärung zu täuschen (Merten, 2008, im Druck). Und unter der Diagnostik der Beschwerdenvalidität wird die Überprüfung der Authentizität oder Glaubhaftigkeit der demonstrierten Symptome, der geschilderten Beschwerden und der Testergebnisse, die bei einem Klienten oder einer Klientin im Rahmen der Begutachtung beobachtet werden, verstanden. Negative Antwortverzerrung kann in 2 unterschiedlichen Formen (oder in deren Kombination) auftreten, nämlich als:

- unzutreffende Beschwerdenschilderung, oder
- fälschliche Symptompräsentation.

Negative Antwortverzerrung kann auf einer phänomenologischen Ebene als eine Form von Inkonsistenz zu betrachten: der Inkonsistenz zwischen den beobachteten Leistungen in einem Beschwerdevalidierungstest und den erwarteten Leistungen aufgrund der Beschwerdenschilderung des Klienten, der Klientin.

Tabelle 2 Mögliche Kontexte negativer Antwortverzerrungen (nach Merten, 2008, im Druck)

Simulation	Absichtliche, reflektierte, zweckvolle Vortäuschung von Symptomen oder fälschliche Beschwerdenschilderung zur Zielerreichung
Aggravation	Beschwerdenübertreibung und/oder -ausweitung: tatsächlich vorhandene Symptome werden zur Zielerreichung verstärkt
Somatoforme und dissoziative Störungen	Befindlichkeits- und Verhaltensstörungen, die sich in Form körperlicher Symptome oder Krankheiten präsentieren und die als psychische Störung aufgefasst werden
Artifizielle Störung / selbstmanipulierte Störung	Als psychische Störung aufgefasste zielgerichtete Vortäuschung oder Erzeugung von Symptomen oder Krankheiten mit einem primären Krankheitsgewinn[3]
Psychiatrische Erkrankungen oder psychopathologische Phänomene…	… die mit einer eingeschränkten Kooperativität verbunden sind oder in deren Rahmen Motivationsprozesse selbst betroffen sind
Persönlichkeitsstörungen	Bestimmte Merkmale der Persönlichkeitsstruktur sind in einem besonderen Mass ausgeprägt, unflexibel oder sozial unangepasst, dass sie als Störung aufgefasst werden.
Situationsbedingte Faktoren	z.B. Atmosphäre, Setting der Abklärung

Nach Merten (2008, im Druck) ist negative Antwortverzerrung als übergeordnetes Konzept zu verstehen, das in unterschiedlichen Kontexten vorkommen kann, und dann auch unterschiedlich zu interpretieren ist. Die verschieden Facetten von N.A. werden dann − in Abhängigkeit des Kontextes − mit bereits bekannten Begriffen bezeichnet − insbesondere kann negative Antwortverzerrung auch bei Störungen mit Krankheitswert vorkommen, wie Tabelle 2 zeigt.

In dieser Definition werden Simulation und Aggravation voneinander differenziert und als Varianten des übergeordneten Konzeptes der N.A. betrachtet. Weiter aber weist Tabelle 2 darauf hin, dass das Phänomen der N.A. auch bei Störungsbildern und Erkrankungen vorkommen kann bei denen es nicht als Simulation/Aggravation zu bewerten ist. Darin liegt zugleich ein Problem, weil das Konstrukt Simulation/Aggravation sich bis zu einem gewissen Grad inhaltlich überlappt mit anderen diagnostischen Konstrukten. Damit befassen wir uns im nachfolgenden Abschnitt (2.1.3).

[3] Die Begriffe primärer und sekundärer Krankheitsgewinn gehen ursprünglich auf Sigmund Freud (Freud, 1986) zurück. Der *primäre Krankheitsgewinn* bezieht sich auf direkte Anreize, die sich aus der Krankheit selbst ergeben: insbesondere Erleichterungen, Schonung, vermehrte Zuwendung aufgrund der Krankenrolle (Freud bezog den Begriff auf das Phänomen, dass durch eine neurotische Erkrankung ein psychischer Konflikt nicht bearbeitet werden muss bzw. vermieden werden kann). Der *sekundäre Krankheitsgewinn* bezieht sich demgegenüber auf soziale Anreize, die eine Folge der Krankheit sind, wie z.B. der Erhalt einer Rente etc.. In der Praxis ist es jedoch mitunter schwierig, eine klare Trennung der beiden Begriffe vorzunehmen.

2.1.3 Störungsbilder mit engem Bezug zu Simulation/Aggravation

Es werden besonders zwei diagnostische Konstrukte genannt, die in einer mehr oder weniger engen Beziehung zu Simulation/Aggravation stehen, nämlich die so genannten artifiziellen oder vorgetäuschten Störungen (factitious disorder), die somatoformen Störungen und die dissoziativen Störungen − bei beiden handelt es sich um anerkannte Störungsbilder der psychiatrischen Diagnostik[4]. Die beiden diagnostischen Kategorien werden im ICD-10 wie folgt umschrieben:

1) artifizielle, vorgetäuschte Störung: Der betroffene Patient oder die Patientin täuscht Symptome wiederholt ohne einleuchtenden Grund vor und kann sich sogar, um Symptome oder klinische Zeichen hervorzurufen, absichtlich selbst beschädigen. Die Motivation ist unklar, vermutlich besteht das Ziel, die Krankenrolle einzunehmen. Die Störung ist oft mit deutlichen Persönlichkeits- und Beziehungsstörungen kombiniert.

2) somatoforme Störungen: Das Charakteristikum ist die wiederholte Darbietung körperlicher Symptome in Verbindung mit hartnäckigen Forderungen nach medizinischen Untersuchungen trotz wiederholter negativer Ergebnisse und Versicherung der Ärzte und Ärztinnen dass die Symptome nicht körperlich begründbar sind. Wenn somatische Störungen vorhanden sind, erklären sie nicht die Art und das Ausmass der Symptome, das Leiden und die innerliche Beteiligung des Patienten, der Patientin.

3) dissoziative (Konversions-)Störungen: Das allgemeine Kennzeichen der dissoziativen oder Konversionsstörungen besteht in teilweisem oder völligem Verlust der normalen Integration der Erinnerung an die Vergangenheit, des Identitätsbewusstseins, der Wahrnehmung unmittelbarer Empfindungen sowie der Kontrolle von Körperbewegungen. Diese Störungen wurden früher als verschiedene Formen der "Konversionsneurose oder Hysterie" klassifiziert. Sie werden als ursächlich psychogen angesehen, in enger zeitlicher Verbindung mit traumatisierenden Ereignissen, unlösbaren oder unerträglichen Konflikten oder gestörten Beziehungen. Die Symptome verkörpern häufig das Konzept der betroffenen Person, wie sich eine körperliche Krankheit manifestieren müsste. Körperliche Untersuchung und Befragungen geben keinen Hinweis auf eine bekannte somatische oder neurologische Krankheit. Zusätzlich ist der Funktionsverlust offensichtlich Ausdruck emotionaler Konflikte oder Bedürfnisse. Die Symptome können sich in enger Beziehung zu psychischer Belastung entwickeln und erscheinen oft plötzlich. Nur Störungen der körperlichen Funktionen, die normalerweise unter willentlicher Kontrolle stehen, und Verlust der sinnlichen Wahrnehmung sind hier eingeschlossen. Störungen mit Schmerz und anderen komplexen körperlichen Empfindungen, die durch das vegetative Nervensystem vermittelt werden, sind unter Somatisierungsstörungen (F45.0) zu klassifizieren.

2.1.4 Störungen mit Krankheitswert vs. Simulation/Aggravation: Inkonsistenzkriterium

Es wurde versucht, eine Abgrenzung dieser drei psychiatrischen Konstrukte von Simulation/Aggravation vorzunehmen; diese Abgrenzung basiert auf den Kriterien der Bewusstseinsnähe

[4] Diagnosecodes nach ICD-10/DSM-IV: artifizielle/vorgetäuschte Störungen, F68.1/300.19; somatoforme Störungen, F45/300.7, 300.8, 300.81, 300.11; dissoziative Störungen, F44/300.6, 300.12-300.15

und der Motivation des Klientenverhaltens (Cunnien, 1997; Merten, Stevens, & Blaskewitz, 2007). Dabei gelten Abgrenzungen die in Tabelle 3 dargestellt werden.

Tabelle 3 Abgrenzung Simulation/Aggravation von artifziellen, somatoformen und dissoziativen Störungen

Diagnostisches Konstrukt	Bewusstheit der Symptomerzeugung	Motivation des Klientenverhaltens
Simulation/Aggravation	absichtlich, gesteuert ('bewusst')	reflektiert, klar ('bewusst')
Artifizielle Störung	absichtlich, gesteuert ('bewusst')	unreflektiert, unklar ('unbewusst')
Somatoforme und dissoziative Störungen	unbeabsichtigt, nicht gesteuert ('unbewusst')	unreflektiert ('unbewusst')

Allerdings suggeriert die Darstellung in Tabelle 3 eine Trennschärfe zwischen den aufgeführten diagnostischen Konstrukten, die nicht zutreffend ist und zwar aus den folgenden Gründen (vgl. Sharpe, 2003):

- Die Abgrenzung aufgrund der Kriterien Bewusstheit und Motivation ist heikel, weil diese allein auf der subjektiven Einschätzung des Gutachtenden beruht; es gibt z.Z. kein objektives Verfahren, dass eine entsprechende Einschätzung zuverlässig erlauben würde.
- Bei artifiziellen Störungen wird angenommen, dass Klientenverhalten motiviert ist durch den sog. ‚sekundären Krankheitsgewinn' bzw. durch die Zuschreibung der Patientenrolle. Demgegenüber gehe es bei Simulation/Aggravation nicht um die Patientenrolle sondern um externe Anreize wie den Erhalt einer Rente oder den Erlass einer Bestrafung. Es ist jedoch mehr als fraglich, inwieweit die beiden Konzepte auseinander gehalten werden können, da ja auch die Einnahme der Patientenrolle durch sog. externe Anreize (z.B. die Zuschreibung von Arbeitsunfähigkeit) gekennzeichnet ist. Rogers und Neumann (2003) kommen deshalb zum Schluss, dass sowohl auf theoretischer als auch auf diagnostischer Ebene eine klare Differenzierung zwischen Simulation/Aggravation und artifizieller Störung auf der Basis von Annahmen zur Motivierung und Bewusstseinsnähe des Klientenverhaltens kaum möglich ist.

Aufgrund dieser Schwierigkeiten kommt Sharpe (2003) zum Schluss, dass die einzige Möglichkeit, Simulation/Aggravation von psychischen Störungen mit Krankheitswert zu differenzieren, in der Identifikation von *Inkonsistenzen* der Symptome besteht als Indikator für eine fehlende Psychopathologie. Zu beachten sind dabei Inkonsistenzen in den folgenden Bereichen:

- innerhalb der Anamnese, Krankengeschichte;
- Krankengeschichte vs. Verhaltensbeobachtung;
- Symptome vs. anerkannte diagnostische Kriterien;
- Krankengeschichte aus Klientensicht vs. Krankengeschichte aus Sicht einer anderen Person;

- Krankengeschichte aus Klientensicht vs. Arztberichte;
- zwischen Beobachtungen im Verlauf der Abklärung.

Das Inkonsistenzkriterium lässt sich nicht nur für psychische Krankheiten sondern auch für die Abgrenzung anderer Störungen mit Krankheitswert von Simulation/Aggravation verwenden. Die Identifikation von Inkonsistenzen ist denn auch ein zentrales Mittel der Simulationsdiagnostik (vgl. Bianchini et al. 2005). Insofern kann die Bestimmung verschiedener Arten von Inkonsistenzen nach Sharpe verallgemeinert werden.

2.1.5 Häufigkeit von Simulation/Aggravation in Abklärungssituationen

Wie häufig simulierendes oder aggravierendes Verhalten in Abklärungssituationen vorkommt, ist schwierig abzuschätzen, zumal nur Befunde aus ausländischen Studien vorliegen. Ausserdem variieren die identifizierten Raten für Simulation/Aggravation stark nach dem Kontext der Begutachtung sowie nach dem jeweiligen Störungsbild, das Anlass für die Abklärung gab. Eine viel zitierte Studie von Mittenberg, Patton, Canyock und Condit (2002) zu 'Malingering' in neuropsychologischen Abklärungen in den USA berichtet über Raten von 8% bis 35% für Simulation/Aggravation. Und in einer Untersuchung von 235 Patientinnen und Patienten, welche nach Unfällen kognitive Beeinträchtigungen angaben, fanden Merten, Friedel und Stevens (2006) in Deutschland bei über 44% der begutachteten Personen Hinweise für suboptimales Leistungsverhalten.

Sowohl die Zuverlässigkeit als auch die Übertragbarkeit dieser Raten auf Schweizer Verhältnisse ist jedoch begrenzt. So weist Dohrenbusch (2007) darauf hin, dass die Datenbasis dafür zur Zeit (noch) sehr schmal ist und die Vergleichbarkeit der in Studien berichteten Prävalenzen auch durch unterschiedliche methodische Zugänge erschwert ist. Als Fazit kann jedoch festgehalten werden, dass die Auftretenswahrscheinlichkeit von Simulation/Aggravation in Abklärungssituationen am höchsten ausgeprägt sein dürfte, deren Ausgang für die begutachtete Klientel mit erheblichen externen Anreizen (z.B. materielle Vergünstigungen) verbunden ist.

2.1.6 Fazit

Negative Antwortverzerrung meint das Bemühen eines Klienten oder einer Klientin, den Gutachter, die Gutachterin durch ungenaue oder unvollständige Antworten oder durch die Demonstration eingeschränkter Leistungen in der Abklärung zu täuschen. Die Begriffe Simulation und Aggravation können als Teilaspekte negativer Antwortverzerrung definiert werden. Dabei bezieht sich

- *Simulation* auf die absichtliche, reflektierte, zweckvolle Vortäuschung von Symptomen oder fälschliche Beschwerdenschilderung zur Erreichung eines Ziels;
- *Aggravation* umfasst die Übertreibung oder Ausweitung von Beschwerden, d.h. tatsächlich vorhandene Symptome werden zur Zielerreichung verstärkt (Merten, 2008, im Druck).

Das Phänomen negativer Antwortverzerrung im Rahmen einer Abklärung ist eine notwendige, aber nicht hinreichende Bedingung für die Identifikation von simulierendem oder aggravierendem Verhalten. Negative Antwortverzerrung kann auch die Folge von Störungen mit Krankheitswert sein, wobei im psychiatrischen Kontext besonders die somatoformen und die Konversions-Störungen zu erwähnen sind. Zu beachten ist aber auch eine mögliche Komorbidität des Klienten, der Klientin im Bereich neurologischer sowie Entwicklungsstörungen.

Eine eindeutige, absolut trennscharfe Bestimmung von Simulation/Aggravation in Abgrenzung zu (insbesondere psychiatrischen) Störungen mit Krankheitswert ist kaum möglich, sondern nur mit einer gewissen Wahrscheinlichkeit zu leisten. Dies hat mit zwei anderen Voraussetzungen zur Identifikation von simulierendem oder aggravierendem Verhalten zu tun: nämlich der Bewusstheit des Verhaltens und dessen Motivierung durch externe Anreize (z.B. den Erhalt einer Rente):

- Inwieweit jedoch ein Klient oder eine Klientin bewusst und absichtlich bestimmte Symptome produziert, kann durch die Gutachtenden lediglich vermutet werden, es gibt kein objektives Verfahren zur Bestimmung der Bewusstseinsnähe.
- Problematisch ist auch die Unterscheidung von externen und internen Anreizen: Wenn nämlich eine Person primär durch den quasi 'internen' Anreiz der Patientenrolle (und der damit verbundenen Erleichterungen) motiviert ist, so wird dies nicht als Simulation/Aggravation definiert sondern der psychiatrischen Diagnose der artifiziellen Störung zugeschrieben und als Ausdruck einer psychischen Krankheit bestimmt. Die Unterscheidung zwischen Simulation/Aggravation und artifizieller Störung ist schwierig, da die Differenzierung der Klientenmotivation nach internen oder externen Anreizen nur bedingt zu leisten ist.

Dies bedeutet, dass die Bestimmung von Simulation oder Aggravation einer fundierten differentialdiagnostischen Abklärung bedarf, die Alternativerklärungen für das Klientenverhalten mit grosser Wahrscheinlichkeit auszuschliessen vermag.

2.2 Methodische Schwierigkeiten bei der Messung und Identifikation von Simulation

Simulation/Aggravation oder Malingering sind Konstrukte für deren Erfassung es keinen eigentlichen Goldstandard gibt. Ein Goldstandard für die Erfassung von Simulation wäre ein Messinstrument, das mit Sicherheit erfassen würde, dass ein Explorand simuliert. Ausserhalb einer ständigen Beobachtung der Exploranden im Alltag ist Simulation also schwer zu erfassen. Die Validierung bei der Testentwicklung von BVT ist daher mit methodischen Schwierigkeiten verbunden.

Im Folgenden werden diese Schwierigkeiten für diagnostische Tests im engeren, formalen Sinne erläutert. Die Problematik und Argumentation ist aber prinzipiell – qualitativ − auch übertragbar auf Testkombinationen und Leitlinien, da allgemein akzeptiert ist, dass ein einzelner BVT nicht Simulation oder Aggravation diagnostizieren kann.

2.2.1 Identifikation von Simulation

Die Identifikation von Simulation ist im Wesentlichen ein Validitätsproblem. Welche Validität − auf einer kumulativen Skala von Augenscheinvalidität bis hin zur einen Goldstandard benötigenden Kriteriumsvalidität − kann nun für einen BVT bei dessen Entwicklung bestimmt werden? Wie ist diese Validität zu beurteilen? Wir gehen davon aus, dass ein Messinstrument für Simulation zumindest Konstruktvalidität aufzeigen sollte (auf einer Skala Augenschein-, Inhalt-, Konstrukt- und Kriteriumsvalidität). Augenschein- und sogar Inhaltsvalidität stellen streng genommen noch keine wissenschaftliche Validität dar; letztere ist zwar notwendig, aber nicht hinreichend für Kriteriumsvalidität (Streiner & Norman, 2006). Auf der Ebene der Inhaltsvalidität ansiedeln könnte man das Erkennen von Widersprüchlichkeiten in klinischen Mustern. Inhaltsvalidität hat aber eher etwas mit der Reliabilität zu tun (z.B. interne Konsistenz als ein Mass für Inhaltsvalidität) und ist daher nicht ausreichend für die Diagnostik von Simulation.

2.2.2 Known-Groups

Auf der Ebene der Konstruktvalidität gibt es sogenannte ‚Known-Groups Studien'. Das sind Studien zur Untersuchung der Konstruktvalidität von BVT. Dabei werden die Ergebnisse der BVT von zwei oder mehr Gruppen verglichen. Die Gruppen werden dabei so konstruiert, dass das interessierende Konstrukt plausibel unterschiedliche Ausprägungen in den jeweiligen Gruppen hat. So wird z.B. davon ausgegangen, dass Hirnverletzte mehr kognitive Probleme haben sollten als Patienten und Patientinnen mit somatoformen Störungen. Zeigt nun ein entsprechender Test das Gegenteil, so kann dies als ein Hinweis auf potentielle Aggravation/Simulation aufgefasst werden.

2.2.3 Analogstudien

In Analogstudien wird schliesslich auf einer noch höheren Ebene eine Kriteriumsvalidität angestrebt, indem ein künstlicher Goldstandard durch eine „Simulation" von Malingering kreiert wird. Ein richtiger Goldstandard ist ja ausserhalb einer ständigen Alltagsobservation des Exploranden unmöglich. Die interne Validität wird dann mit den Begriffen der Sensitivität, der Spezifität sowie mit dem prädiktiven Wert beurteilt. Formal ist der positiv prädiktive Wert eines BVT die Wahrscheinlichkeit, dass ein Explorand Malingering aufweist, gegeben einen positiven BVT oder eine positive BVT-Kombination.

Für die „Diagnostik" von Simulation oder Aggravation ist aus ethischer Perspektive eine minimale Falsch-Positiv-Rate zu fordern, was einer submaximalen bis maximalen Spezifität entspricht. Die Schwelle für ein positives Testresultat ist bei der Testentwicklung in diesem Sinne zu legen. Die Grösse der Sensitivität ist zweitrangig und wird durch diese Schwelle bestimmt.

2.2.4 Sensitivität, Spezifität und prädiktive Werte

Die Sensitivität (Sn) eines BVT oder einer BVT-Kombination ist die Wahrscheinlichkeit eines positiven Tests gegeben Malingering. Die Spezifität (Sp) ist die Wahrscheinlichkeit eines negativen Tests

gegeben Absenz von Malingering. Tabelle 4 zeigt die Beziehung zwischen Testresultat und dem Merkmal Malingering. Der positiv prädiktive Wert, PPV, beschreibt die Wahrscheinlichkeit, dass bei einem positiven Testresultat tatsächlich Malingering vorliegt. Der negativ prädiktive Wert, NPV, ist die Wahrscheinlichkeit, dass bei einem negativen Testresultat tatsächlich kein Malingering vorliegt.

Tabelle 4 Zusammenhang zwischen dichotomem Testresultat und Malingering

	Malingering +	Malingering -	Posttest-Wahrscheinlichkeit
BVT+	A	B	PPV: $P(M+ \mid BVT+) = \frac{A}{A+B}$
BVT-	C	D	NPV: $P(M- \mid BVT-) = \frac{D}{C+D}$
Sn, Sp,	$Sn = P(BVT+ \mid M+ = \frac{A}{A+C})$	$Sp = P(BVT- \mid M-) = \frac{D}{B+D}$	
FR	$FNR = 1 - Sn = P(BVT- \mid M+) = \frac{C}{A+C}$	$FPR = 1 - Sp = P(BVT+ \mid M-) = \frac{B}{B+D}$	

A: Richtig positiv, B: Falsch positiv, C: Falsch negativ, D: Richtig negativ. Sn: Sensitivität, Sp: Spezifität. M-: Malingering nicht präsent, M+: Malingering präsent, BVT+: BVT positiv, BVT-: BVT negativ, PPV: Positiv prädiktiver Wert (positive Posttestwahrscheinlichkeit), NPV: Negativ prädiktiver Wert (negative Posttestwahrscheinlichkeit). FR: Falsch-Rate, FNR: Falsch-Negativ-Rate, FPR: Falsch-Positiv-Rate.

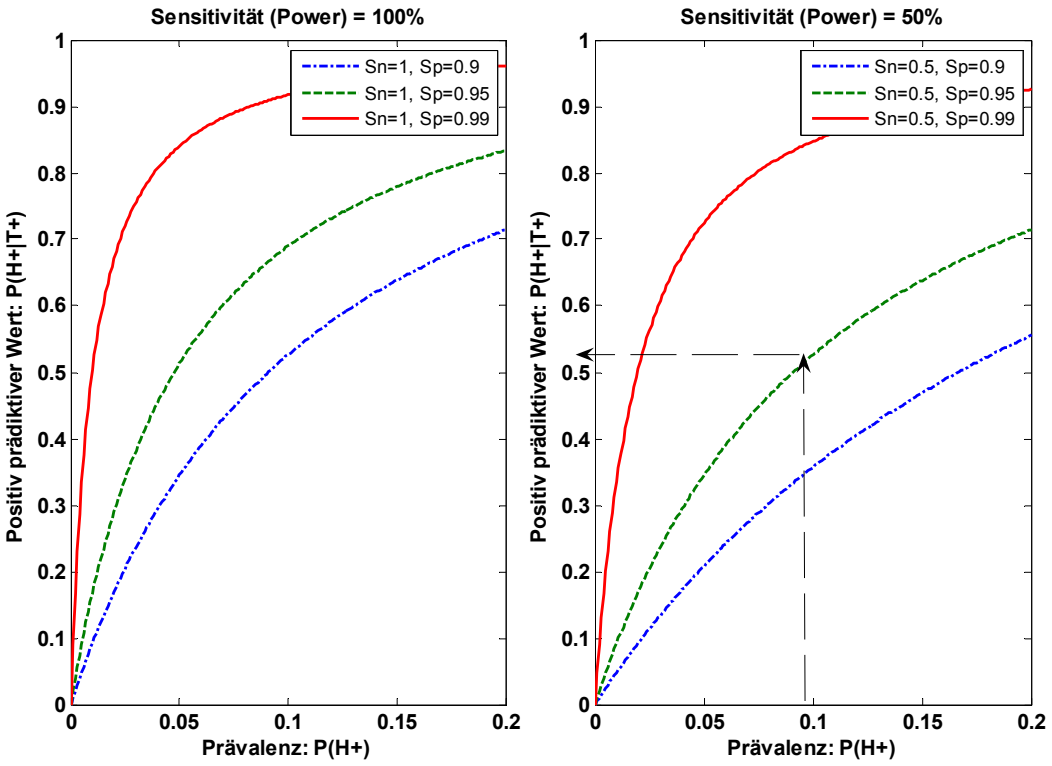
2.2.5 Prosecutor's fallacy

Testentwickler und Testanwender müssen sich der Gefahren der sogenannten ‚prosecutor's fallacy' bewusst sein. Dazu gehören zwei Aspekte: Das Risiko einer falschen Beurteilung von bedingten Wahrscheinlichkeiten durch Nicht-Beachten der Prävalenz und die Problematik von multiplen Tests.

Bedingte Wahrscheinlichkeit

Sensitivitäten, Spezifitäten und prädiktive Werte sind bedingte Wahrscheinlichkeiten. Angenommen, ein Goldstandard existiert – in Analogstudien wird der Goldstandard meistens durch falsche Exploranden „simuliert" − dann kann aus Sensitivität und Spezifität und der a priori Prävalenz der positiv prädiktive Wert eines Tests oder einer Testkombination berechnet werden. Im Gegensatz zur Sensitivität (Sn) und Spezifität (Sp) ist aber der prädiktive Wert – nach dem Bayes-Theorem − sehr stark prävalenzabhängig, wie Abbildung 2 zeigt. Grundsätzlich gilt, dass eine hohe Spezifität (kleine Falsch-Positiv-Rate) nicht gleichbedeutend ist mit einem hohen positiv prädiktiven Wert, ausser bei hohen Prävalenzen oder hoher a priori Wahrscheinlichkeit von Simulation oder Aggravation.

Abbildung 2: Prävalenzabhängigkeit des positiv prädiktiven Wertes



H+: Präsenz von Malingering, T+: BVT oder BVT-Kombination positiv. Für Sn = 0.5 und Sp = 0.95 und bei einer Prävalenz von 10% ist der positiv prädiktive Wert z.B. „nur“ 0.52.

Einfluss der Grundprävalenz

Vor allem bei tiefen Grundraten oder Prävalenzen sind Fehlinterpretationen von Testresultaten eine nicht geringe Gefahr[5]. Das gilt auch für „unter Zufall“ Resultate von so genannten „forced choice“ Tests. Bei solchen Tests wird bei Antwortverhalten unter Zufall z.T. auf Simulation geschlossen (Bianchini et al., 2005; Greve & Bianchini, 2004). Ohne Einbezug der Grundrate bleiben aber bei diesen Tests Entscheidungen delikat[6]. Beim Entwickeln von BVT ist zur Vermeidung von Falsch-

5 Diagnostik von Simulation mit BVT, Interne Validität: Es gilt allgemein für die Diagnose eines Merkmals M mit einem Test T, nach Bayes:

$$PPV = P(M+|T+) = \frac{P(T+|M+)P(M+)}{P(T+)} = \frac{P(T+|M+)P(M+)}{P(T+|M+)P(M+) + P(T+|M-)P(M-)} = \frac{Se \cdot \Pr}{Se \cdot \Pr + (1-Sp)\cdot(1-\Pr)}.$$

Dabei ist P die Wahrscheinlichkeit, M das Merkmal (Malingerer), T der BVT oder die BVT-Kombination, PPV der positiv prädiktive Wert oder die A-posteriori Wahrscheinlichkeit, Sn die Sensitivität, Sp die Spezifität des BVT und Pr die Prävalenz oder die A-priori Wahrscheinlichkeit von M in der entsprechenden Population. Es wird ersichtlich, dass nur bei einer perfekten Spezifität von 1, d.h. bei Abwesenheit von Falsch-Positiven, was aus ethischer Sicht gefordert werden müsste, der PPV 1 wird. Auch bei einer schon grossen Spezifität von 0.95 und einer Sensitivität von 0.30 bleibt der PPV bei einer angenommenen Prävalenz von 20% Malingering „nur“ bei 60%, bei einer angenommenen Prävalenz von 10% sogar „nur“ bei 40%, usw. Die Wahrscheinlichkeit, dass eine Testperson ein Malingerer ist, bleibt also auch bei einem positiven Testresultat oft zu gering.

6 Forced-choice Test: Getestet wird die Hypothese: „Explorand ist kein Simulant“. Das Testergebnis eines BVT oder einer BVT-Kombination von einem Exploranden sei r. R seien mögliche Testergebnisse. Die Wahrscheinlichkeit $P(R \le r \,|\, Sim-)$ sei z.B. unter 5%. Dann ist das – sonst übliche - Ablehnen der Hypothese gefährlich, da die Posttest-

Positiven eine hohe bis perfekte Spezifität anzustreben, was dann aber notgedrungen mit einer verminderten Sensitivität einhergeht, da eine Verminderung der Falsch-Positiven auch eine Verminderung der Richtig-Positiven nach sich zieht (Bianchini et al., 2005; Greve & Bianchini, 2004; Mossman, 2000a, 2000b, 2003). Auch eine relativ niedrige Falsch-Positiv-Rate kann bei niedriger Grundprävalenz nicht ausreichen, um hohe prädiktive Werte zu erreichen. Nur bei Tests oder Testkombinationen mit einer nahezu maximalen Spezifität sind auch bei einer niedrigen Grundprävalenz die Posttest-Wahrscheinlichkeiten bezüglich effektiver Simulation hoch.

Gefahren bei Testkombinationen

Zur Problematik von Testkombinationen: Es muss darauf geachtet werden, inwieweit die verschiedenen Tests zumindest inhaltlich voneinander unabhängig sind und effektiv einen Informationsgewinn darstellen[7]. Korrelierende Tests können die Falsch-Positiv-Rate auch bei Kombination dieser Tests nicht mehr senken und dürfen dann nicht kumulativ interpretiert werden. Dies führt in der Diagnostik oft zu Überschätzung von a posteriori oder Posttest-Wahrscheinlichkeiten (Bachmann, ter Riet, Clark, Gupta, & Khan, 2003). Es sollten daher immer unabhängige oder schwach korrelierende Tests miteinander kombiniert werden.

Schwellenwert für Posttest-Wahrscheinlichkeiten

Wie gross muss nun ein positiver prädiktiver Wert – eine Posttest-Wahrscheinlichkeit für Simulation – nach Applikation eines BVT oder einer BVT-Serie sein, damit Nicht-Simulation ausgeschlossen werden kann? Dies hängt ab vom politischen und sozialen Kontext und von der Rechtsprechung (Mossman, 2003). Diese Frage ist wissenschaftlich – messtechnisch – nicht mehr zu beantworten.

Wichtig bleibt aber, dass Wahrscheinlichkeitsaussagen (Wie wahrscheinlich ist es, das jemand Simulant ist, gegeben BVT oder BVT Kombination positiv) in Form von prädiktiven Werten gemacht werden. Die Integration von Vortestwahrscheinlichkeiten (A priori Wahrscheinlichkeit, in die Grundprävalenz von Simulation im jeweiligen Umfeld und manchmal auch andere Fakten einfliessen) ist für die Schlussfolgerungen wichtig.

2.2.6 Kreuzvalidierung, externe Validität und ökologische Validität

Kreuzvalidierung

Ist die Schwelle, ab der ein BVT als positiv gilt – unter Einbezug des gewünschten Verhältnisses von Sensitivität zu Spezifität – definiert, sollte der Test zusätzlich an anderen Stichproben kreuzvalidiert

---

Wahrscheinlichkeit, dass der Explorand ein Simulant ist, nicht 95%, sondern z.B. bei einer Prävalenz von 10% und einer Sensitivität (Power des Tests) von 70% „nur“ 60% ist! Für massiv höhere Prävalenzen und grössere Spezifitäten (kleinere Signifikanzniveaus) wird der Kontrast dann immer kleiner. Für eine Prävalenz von 40 % und einer Sensitivität von 70 % und einem p-Wert von 1% folgt z.B. eine Posttestwahrscheinlichkeit von 97%.

7 Durch Testkombinationen kann die Spezifität erhöht werden. Bei n unabhängigen Tests gilt für die Falsch-Positiven (FP) bei Positivität von T1 und T2 …und Tn: $P(T_1+,...,T_n+|M-)=P(T_1+|M-)\cdot...\cdot P(T_n+|M-)=(1-Sp_1)\cdot...(1-Sp_n)$. Theoretisch könnte also die Falsch-Positiv Rate beliebig an 0 angenähert werden, aber je mehr die verschiedenen Tests untereinander korrelieren, desto kleiner ist dieser Informationsgewinn. Bei zwei korrelierenden ten Tests z.B. gilt für diese Rate: $P(T_1+,T_2+|M-)=(1-Sp_1)\cdot(1-Sp_2)+r\sqrt{Sp_1(1-Sp_1)Sp_2(1-Sp_2)}$. Je höher die beiden Tests miteinander korrelieren (r gegen 1), umso mehr entspricht die Falsch-Positiv Rate der Testkombination derjenigen eines einzigen Tests.

worden sein, um zu testen, ob das Modell, das aus der Stichprobe entwickelt wurde, kein Zufallsbefund ist. So sollte das aus der Trainingsstichprobe entwickelte Modell (quantifizierte Sensitivität und Spezifität oder andere Validitätskoeffizienten) unter Einbezug der Grundrate das Malingering in der neuen Teststichprobe hinreichend gut voraussagen.

Externe Validität

Da der generierte Goldstandard in Analogstudien eigentlich kein solcher ist, wird das Validitätsproblem zum Teil verschoben von der internen auf die externe Validität. Die in solchen Studien beteiligten „Simulanten" repräsentieren zum Teil nicht die realen Exploranden in der gutachterlichen Praxis. So kann es sein, dass in der Testentwicklung z.B. nicht auf eine adäquate Darstellung/Simulation eines finanziellen Anreizes – eine Vorbedingung für Malingering – geachtet wurde.

Ökologische Validität

Darunter versteht man den Aspekt eines realistischen Kontextes, in dem der BVT entwickelt/validiert wurde. Hochstandardisierte Verfahren gehen natürlich immer bis zu einem gewissen Grad einher mit einem verzerrten Abbild der Alltagssituation. Diese Validität gehört neben der internen und der externen Validität aber nicht zu den eigentlichen Gütekriterien.

2.2.7 Fazit

BVT, die einzeln oder in Kombination Simulation erfassen wollen oder die Erfassung von Simulation vorgeben, müssen wie alle Tests wissenschaftliche Gütekriterien erfüllen. Inhaltsvalidität ist eng mit dem Reliabilitätbegriff verwandt. Die Konstruktvalidität von BVT kann mit Known-Groups Designs untersucht werden. Durch die Nicht-Verfügbarkeit eines richtigen Goldstandards werden schliesslich in Analogstudien Pseudo-Goldstandards generiert und mit diesen können dann die Gütekriterien eines Tests (vor allem die prädiktiven Werte) quantifiziert werden; dies aber zu Lasten der externen Validität.

Aus ethischer Sicht sollte ein BVT eine geringe bis minimale Falsch-Positiv-Rate oder submaximale bis perfekte Spezifität aufweisen. Die Schwelle für ein positives Testresultat ist in diesem Sinne bei der Testentwicklung zu definieren. Die Grösse der Sensitivität ist zweitrangig und wird durch diese Schwelle bestimmt.

Gegeben einen positiven Test, reicht submaximale Spezifität (und evtl. Sensitivität) des Tests jedoch nicht aus für die Postulierung von Simulation oder Aggravation. Der positiv prädiktive Wert – die Wahrscheinlichkeit der Simulation gegeben positiver Test – ist im niedrigprävalenten Bereich, in dem wir uns bei Simulation befinden, hochgradig von dieser Prävalenz abhängig. Bei tiefen Prävalenzen oder a priori-Wahrscheinlichkeiten und submaximaler Spezifität (und evtl. Sensitivität) bleiben Schlussfolgerungen bezüglich Simulation und Aggravation immer noch problematisch.

2.3 Verfahren zur Erfassung von Simulation, Aggravation

Die folgenden Abschnitte beleuchten zuerst den Hintergrund der Erfassung von Simulation und Aggravation. Anschliessend besprechen wir die Ergebnisse der internationalen Forschung über die

Prävalenz reduzierter Beschwerdevalidität (BV). Im letzten Abschnitt wird die Beurteilung der BV mit standardisierten Tests und Leitlinien ausführlicher besprochen. Im Fazit werden die Ergebnisse zusammengefasst.

2.3.1 Hintergrund

Die Diagnostik der BV dient der Überprüfung der Authentizität oder Glaubhaftigkeit der durch eine Person demonstrierten Symptome, der durch sie geschilderten Beschwerden und der Ergebnisse, die diese Person in einer Leistungsüberprüfung erzielt (Merten, 2008, im Druck). Da eine Begutachtung eine gültige Aussage über die Leistungen machen soll, muss im Rahmen der Abklärung die Plausibilität der geschilderten Beschwerden/Symptome und der Testergebnisse untersucht werden.

Eine wissenschaftlich abgesicherte Differenzierung zwischen Aggravation und Simulation ist nicht möglich. Die Beurteilung der BV verzichtet auf diese Differenzierung und fokussiert auf die Unterscheidung zwischen plausiblen und nicht-plausiblen Beschwerden. Die Überprüfung der Konsistenz der Befunde bildet dazu eine wichtige Grundlage (siehe Seite 14). Mit diesen Verfahren soll die Validität der geschilderten Beschwerden überprüft werden.

Die Beurteilung der BV entwickelte sich in den letzten 15 Jahren besonders im Bereich der Neuropsychologie, im Vergleich zur Medizin, sehr stark. Ein wichtiger Hintergrund dieser unterschiedlichen Entwicklung ist, dass die gutachterliche Bewertung von Funktionsfähigkeiten sich weniger auf Diagnosen als auf die Bewertung von Funktionseinschränkungen abstützt. Die Neuropsychologie ist im Vergleich zur Medizin mehr auf die Erfassung der kognitiven Leistung und weniger auf (medizinische) Diagnostik ausgerichtet. Die Neuropsychologie liefert durch die Beurteilung kognitiver Leistungen einen wichtigen Beitrag zur Beurteilung der arbeitsbezogenen Leistungsfähigkeit.

Funktionseinschränkungen bilden die Grundlage für die arbeitsbezogene Leistungsminderung der Exploranden. In den Gutachten sollen Beschwerden überprüft werden wie etwa: ich kann mich nicht mehr so gut konzentrieren; wenn ich mich anstrenge, nehmen meine Schmerzen zu; nach einem halben Arbeitstag bin ich völlig erschöpft. Für die Überprüfung dieser Beschwerden wurden, insbesondere im neuropsychologischen Bereich, eine Vielzahl standardisierter BVT entwickelt.

Aus dem Bereich der Begutachtung der arbeitsbezogenen Leistungsfähigkeit liegen erst wenige Ergebnisse zur Zuverlässigkeit des Expertenurteils bei der Erkennung ungenügender BV vor. Bei Exploranden, die nach einem fremdverursachten Unfall Schadenersatz wegen Rückenschmerzen fordern, sind die eigenen Angaben vielfach unzuverlässig und nicht valide (Carragee, 2008). Fachleute erkennen im Rahmen forensischer Begutachtung Simulationsversuche in der Regel aber kaum besser als wenn sie rein per Zufall urteilen würden (Hall & Pritchard, 1996). Ein weiterer Hinweis auf mögliche Antwortverzerrungen liefert eine Studie bei 1499 kanadischen Exploranden. Das Ausmass selbst geschilderter Gedächtnisstörungen hing in dieser Studie stärker mit der Anstrengungsbereitschaft als mit der tatsächlichen Leistung zusammen (Greene, 2005). Verschiedene Studien zeigen einen nicht erklärbaren inversen Zusammenhang zwischen der Schwere der Schädigung und den später von den Exploranden gezeigten Beschwerden. Gedächtnisprobleme traten ausgeprägter bei Exploranden mit einer leichteren traumatischen Hirnverletzung auf als bei Personen mit einem schwereren Trauma (Green, Iverson, & Allen, 1999). Mertens zitiert mehrere Studien von angrenzenden Forschungsbereichen die zeigen, „dass Menschen generell keine guten Experten in der Erkennung unaufrichtiger Kommunikation sind“. Dies schliesst die Erkennung von Simulationsversu-

chen durch Mediziner und Psychologen ein. Allein auf den klinischen Eindruck, das „Gefühl“ vertrauend, schneiden Experten in einschlägigen empirischen Untersuchungen schlecht ab und sind allzu leicht täuschbar (Ekman & O'Sullivan, 1991; Faust, 1995; Rosenhan, 1973; Vrij, 2000). Merten sagt dazu auch: „Umso bemerkenswerter ist es, wie häufig auch in der Gegenwart genau dieses individuelle, nicht ausreichend kommunizierbare, lehrbare und falsifizierbare Gespür für die Bewusstheit oder Unbewusstheit fälschlicher Symptompräsentationen beschworen und als Massstab für weit reichende gutachterliche Entscheidungen benutzt wird.“

Die BV ist in gewissen Kontexten bei einem beträchtlichen Anteil der Exploranden reduziert. Die Prävalenz einer deutlich ungenügenden BV variiert je nach Kontext stark. Sie könnte zu Beginn einer Arbeitsunfähigkeit durchaus deutlich unter 10% liegen. Auch in Relation zur gesamten Gruppe von Personen die eine Berentung erhalten dürfte die Prävalenz tief sein. In neuropsychologischen Abklärungen in den USA werden Malingering-Raten von 8% bis 35% für Simulation/Aggravation beschrieben (Mittenberg, Patton et al., 2002). Nicht nur Studien aus den USA berichten über eine ungenügende BV bei einem substantiellen Anteil der Begutachtungen. Auch in Europa ist das Problem bekannt und es gibt Hinweise darauf, dass in gewissen Kontexten eine deutlich höhere Prävalenz vorliegen kann (Merten et al., 2006; Merten, Friedel, & Stevens, 2007; Mittenberg, Patton et al., 2002; Stevens, Friedel, Mehren, & Merten, 2008). Die Verwendung von BVT gab bei Exploranden mit einem Schleudertrauma in den Niederlanden Hinweise auf eine unzureichende Leistungsbereitschaft bei 61% der Personen (Schmand et al., 1998). In einer Studie in Grossbritannien zeigten 62% der Exploranden, die über Gedächtnisprobleme klagten, eine deutlich negative Antwortverzerrung in einem Leistungsmotivationstest (Gill, Green, Flaro, & Pucci, 2007). In der Studie untersuchte ein Mediziner im Rahmen der Begutachtung die kognitiven Leistungen bei 119 Personen ohne nachweisbare Pathologie, welche Gedächtnisprobleme angaben. Die Begutachtungen fanden alle statt im Zusammenhang mit Berentungs- oder Haftpflichtfragen. Die Exploranden hatten somit möglicherweise ein Interesse daran, eine reduzierte Leistung zu zeigen. Gill verwendete einen sehr einfachen Gedächtnistest. Dabei schnitten 62% der Exploranden im Test schlechter ab als Kinder im Alter von 8-14 Jahren und Personen mit einer deutlichen Demenz. In Deutschland schlussendlich fanden Merten et al., 2006 eine fragliche BV bei 44% einer untersuchten Gruppe, 256 Patienten und Patientinnen mit fraglicher BV und kognitiven Beschwerden nach unter anderem Gehirnerschütterung, Gehirnprellung, psychische Störungen (z.B. depressive Störung, somatoforme Störung, Angststörung), Schleudertrauma der Halswirbelsäule. In der Schweiz liegen unseres Wissens keine entsprechenden Studien vor.

Der Eindruck und das ‚Gefühl’ der Expertinnen und Experten bilden also keine zuverlässige Grundlagen für die Beurteilung der BV. Die Prävalenz ungenügender BV dürfte im Kontext der Begutachtung beachtlich sein. Somit besteht ein Bedarf für spezielle Verfahren zur Erfassung der BV. Unterschieden wird dabei zwischen standardisierten BVT und Leitlinien. Beide Verfahren werden sowohl in der wissenschaftlichen Literatur beschrieben, als auch im Rahmen der Begutachtung in der Schweiz teilweise verwendet. Standardisierte BVT sind in der Regel speziell entwickelt und validiert für die Überprüfung einzelner oder mehrerer kognitiver und psychologischer Bereiche. Leitlinien beschreiben die Vorgehensweise und Kriterien für die Beurteilung der BV. Da es für die meisten Bereiche keine standardisierten BVT gibt, die den wissenschaftlichen Kriterien bei isolierter Anwendung vollumfänglich genügen, werden die BVT in der Regel im Rahmen von ähnlichen Kriterien, wie sie in den Leitlinien beschrieben werden, angewendet.

2.3.2 Beschwerdevalidierungstests

BVT beruhen auf verschiedenen konzeptionellen und praktischen Grundlagen der Testanwendung, die in Anlehnung an Merten im Folgenden beschrieben werden (Merten, 2008, im Druck).

Negative Antwortverzerrungen

Nach Bianchini et al. (2005) sind negative Antwortverzerrungen „empirisch bedeutsame Beobachtungen von Inkonsistenzen" im neurokognitiven Kontext. Bezüglich physischer Belastbarkeit brauchen Bianchini et al. den Begriff „effort bias" für die verzerrte Präsentation der körperlichen Leistungsfähigkeit als Pendant für den Begriff „response bias" der in der kognitiven Leistungserfassung verwendet wird. Der Begriff „sincerity of effort" wird von Lechner, Bradbury und Bradley (1998) definiert als „des Patienten bewusste Motivation, optimale Leistung während einer Evaluation zu zeigen". Merten definiert Antwortverzerrungen (response bias), in Anlehnung an Bush et al. (2006), als das Bemühen einer untersuchten Person, den Untersucher durch ungenaue oder unvollständige Antworten oder durch die Demonstration eingeschränkter Leistungen in entsprechenden Prüfverfahren zu täuschen. Nach Merten (2008, im Druck) können negative Antwortverzerrungen in zwei unterschiedlichen Formen oder in der Kombination beider in Erscheinung treten: a) als unzutreffende Beschwerdenschilderung und b) als fälschliche Symptompräsentation. Bei einer suboptimalen Leistungsmotivation oder eingeschränkten Anstrengungsbereitschaft des Probanden liegen die im Test gezeigten Leistungen unterhalb derer, welche die Person bei angemessener Kooperativität zu erbringen fähig wäre (Merten, 2008, im Druck).

Alternativwahlverfahren

Bei Alternativwahlverfahren (forced choice) wird ein Reiz angeboten, zum Beispiel ein Wort, eine Zahl, ein Symbol oder ein Bild. Der Explorand muss anschliessend den Reiz in einer Auswahl von angebotenen Optionen wieder erkennen. Die Zeit zwischen dem Reiz und der Wiedererkennung kann variiert werden. Auch bei stark reduziertem Gedächtnis kann, auf Grund des Zufalls, eine bestimmte minimale Trefferquote erwartet werden. Liegt die Trefferquote tiefer, so ist dies als Hinweis auf eine mögliche negative Antwortverzerrung zu interpretieren. Diese Methode wird deshalb auch als ‚below-chance Verfahren' bezeichnet. Erwartet wird bei Gedächtnistests auch ein negativer Zusammenhang zwischen der Zeitdauer zwischen dem Reiz und der Antwort, und der Trefferquote. Beispiele solcher Tests für Zahlen sind der ‚Digit Memory Test' (DMT) und der ‚Portland Digit Recognition Test'.

Testdeckeneffekt oder vorgetäuschte Schwierigkeit

Einige Tests wurden so konzipiert, dass sie eine Schwierigkeit vortäuschen, aber in Wirklichkeit so einfach sind, dass sie von den Exploranden (annähernd) fehlerfrei ausgeführt werden können. Erwartet wird also eine maximale Leistung und eine Punktezahl (score) im Bereich des Maximums (der ‚Decke'). Der Testdeckeneffekt beruht also darauf, dass die minimale Leistung für einen Maximalscore ausreicht, und auch bei einer gewissen Funktionsbeeinträchtigung erwartet werden kann. Dazu wird der Test bei Patientinnen und Patienten mit einer nachgewiesenen schweren Funktionsbeeinträchtigungen in so genannten ‚known groups Studien' durchgeführt. Angewendet werden solche Tests bei Exploranden, die keine medizinische Diagnose aufweisen, welche eine reduzierte Leistung erwarten lässt. Wenn die Leistung deutlich unter derjenigen von Patienten und Patientinnen mit nachgewiesenen schweren Funktionsbeeinträchtigungen liegt, so ist dies als Hinweis für

eine mögliche negative Antwortverzerrung zu interpretieren. Ein Beispiel eines BVT-Verfahrens, das auf dem Testdeckeneffekt oder der vorgetäuschten Schwierigkeit beruht, ist die Überprüfung von Gedächtnisleistungen mit einfachen Tests bei Exploranden ohne nachgewiesene Hirnläsionen. Positiv ist das Ergebnis, wenn die Leistungen deutlich unter der Leistung von Patienten und Patientinnen mit einer schweren Demenz liegen.

Inkonsistente oder untypische Leistungsprofile

Erwartet wird, dass die in einem Test demonstrierten Leistungen ein Profil zeigen, das konsistent ist mit den Erwartungen für die entsprechende Pathologie. Abweichende Leistungsprofile sind untypisch und deshalb nicht konsistent mit den Erwartungen. Eine deutliche Abweichung von diesen Erwartungen reduziert die BV. Die Leistungskurve kann zum Beispiel bei Konzentrationstests hinzugezogen werden, indem überprüft wird, ob mit zunehmender Dauer der Untersuchung oder bei grösserer Schwierigkeit der Testaufgaben die Leistung abnimmt.

Ein anderes Beispiel für die Überprüfung der Konsistenz ist die Erfassung besonders drastischer oder untypischer psychischer Beschwerden mit dem ‚Strukturierten Fragebogen Simulierter Symptome' (SFSS, die deutsche Version vom Structured Interview of Malingered Symptomatology oder SIMS).

2.3.3 Leitlinien

Leitlinien für die Beurteilung der BV ermöglichen durch ihre allgemeine Formulierung die Verwendung nicht standardisierter und nicht validierter Verfahren und die Berücksichtigung aller vorliegenden Befunde. Sie haben somit auch eine grosse Bedeutung im Kontext der Anwendung standardisierter BVT.

Leitlinien für die Beurteilung der BV sind aus zwei Gründen von Bedeutung: Erstens sind sowohl Gutachtende als auch Autoren und Autorinnen wissenschaftlicher Publikationen zur Beurteilung der BV der Meinung, dass ein Gutachten sich bei der Beurteilung der BV nicht nur auf BVT abstützen darf. Vielmehr sind sich alle Autoren darin einig, dass die Befunde in einem Gesamtzusammenhang von Informationen aus verschiedenen Quellen beurteilt werden müssen. Die Leitlinien stehen in Übereinstimmung mit diesen Empfehlungen. Zweitens sind in den meisten Teilbereichen der medizinischen Begutachtung kaum standardisierte BVT vorhanden. Eine Beurteilung der BV wird in diesen häufig vorkommenden Situationen vorgenommen, indem die Plausibilität und die Konsistenz der unterschiedlichen Befunde beurteilt werden. Somit nehmen die Leitlinien einen bedeutenden Platz ein.

Die Leitlinien beschreiben eine allgemeine Vorgehensweise für die Überprüfung der BV, die auch den Kontext der Begutachtung einbezieht. So berücksichtigen die Leitlinien, ob für den Exploranden ein Anreiz für eine suboptimale Leistung besteht, und setzen den Einbezug medizinischer Diagnosen bei der Interpretation der Testergebnissen voraus. In den nächsten Abschnitten werden die wichtigsten Leitlinien im Bereich der Überprüfung der BV beschrieben. Dabei werden die Leitlinien von Slick, Sherman und Iverson (1999), und von Bianchini et al. (2005) dargestellt.

Slick et al. definierten vier Kriterien zur Identifikation von Simulation (‚Malingering'): A) der Nachweis eines externen Anreizes für Aggravation; B) Hinweise für Aggravation aus der neuropsychologi-

schen Untersuchung; C) Hinweise aus der Beschreibung der Beschwerden durch den Exploranden, und D) die negativen Antwortverzerrungen können nicht durch Pathologien oder Entwicklungsstörungen erklärt werden. Blaskewitz (2005) schreibt über die Leitlinien von Slick et al.: „Im Wesentlichen haben sich bis heute die Richtlinien von Slick, Sherman und Iverson (1999) durchgesetzt, auf die in vielen Forschungsarbeiten Bezug genommen wird“. Die Autoren unterscheiden zwischen Kriterien für definitive, wahrscheinliche und mögliche Simulation. Die Voraussetzung für alle diese Diagnosen ist zunächst ein äusserer Anreiz. Die konkrete Diagnose richtet sich nach der Sicherheit der Belege für willentliche Übertreibung oder Vortäuschung (fabrication) der kognitiven Dysfunktion bei gleichzeitiger Abwesenheit von plausiblen Alternativerklärungen (psychiatrische oder neurologische Störungen oder Entwicklungsstörungen). Auffällige Ergebnisse in Beschwerdevalidierungstests oder speziell konzipierte Parameter aus neuropsychologischen Standardverfahren gelten dabei als stärkste Belege für mangelnde Leistungsbereitschaft.’

Eine Einschränkung der Kriterien von Slick, Sherman und Iverson (1999) besteht darin, dass die Kriterien nicht geeignet sind für die Beurteilung der BV im Bereich einer Behinderung mit somatischem Befund. Deshalb entwickelten Bianchini et al. (2005) die Leitlinien weiter für die Anwendung bei Patienten und Patientinnen mit Schmerzen und Behinderung (pain related disability). Für die Begutachtung von körperlichen Beschwerden und Funktionseinschränkungen erweiterten Bianchini et al. die Kriterien von Slick et al. mit einem zusätzlichen Kriterium B (siehe unten). Die übrigen Kriterien entsprechen denjenigen von Slick et al. Die Kriterien von Bianchini et al. schliessen die Leitlinien und Kriterien von Slick et al. somit ein und können deshalb nicht nur im psychologischen, sondern auch im medizinischen Kontext angewendet werden. Wir beschreiben anschliessend die Kriterien von Bianchini et al. ausführlich.

2.3.4 Leitlinie für die Begutachtung der BV bei Schmerzen mit Behinderung

Der Grundgedanke bei der Entwicklung der Leitlinien durch Bianchini at al. ist die multidimensionale Präsentation der Behinderung. Patientinnen und Patienten mit einer chronischen körperlichen Behinderung zeigen eine vielfältige Palette von Symptomen, Beschwerden und Einschränkungen der körperlichen und kognitiven Leistungsfähigkeit. Die Frage ist nicht primär, wie stark die Schmerzen sind, sondern wie stark eine Person in den Alltagsfunktionen und in der Arbeitsfähigkeit behindert ist. BVT zur Prüfung von Leistungen ausserhalb des kognitiven Bereichs sind kaum vorhanden. Mit Hilfe der Leitlinien kann die Validität der Befunde im Rahmen der Begutachtung überprüft werden. Dabei können alle verfügbaren somatischen und kognitiven Informationen berücksichtigt werden.

Bianchini et al. definierten fünf Kriterien für die Beurteilung der BV bei Patienten und Patientinnen mit einer Behinderung im Rahmen von Schmerzen. Malingering (engl. für Simulation) ist nach Bianchini et al. (2005) eine „beabsichtigte Verstärkung oder Generierung von kognitiven, emotionalen oder physischen Dysfunktionen nach Krankheit/Behinderung mit dem Ziel eines finanziellen Nutzens, Vermeiden von Arbeit oder zum Bezug von Medikamenten“. Die Bewusstseinsnähe gilt im wissenschaftlichen Kontext als nicht überprüfbar. Bianchini erkennt dieses Problem. Er geht davon aus, dass diese Frage nicht beantwortet werden muss. Wenn bei Patienten oder Patientinnen oder Exploranden Anreize für eine Antwortverzerrung oder für eine submaximale Anstrengung bestehen, ist die Möglichkeit von Antwortverzerrungen und submaximalen Leistungen im Rahmen der Begutachtung immer zu überprüfen.

Kriterium A betrifft den Nachweis relevanter externer Anreize für Aggravation oder Simulation zum Zeitpunkt der Begutachtung. Beispiele sind laufende gerichtliche Abklärungen in Zusammenhang mit Haftpflicht und Schadenersatz, Rentenabklärungen, Abklärung der Diensttauglichkeit und der Verschreibung von Medikamenten.

Kriterium B betrifft die Hinweise (‚evidence') auf Aggravation oder Simulation der Behinderung im Rahmen der klinischen Untersuchung (ärztliche körperliche Untersuchung, Untersuchung im Rahmen der Physiotherapie und Ergotherapie, Erfassung der Funktionellen Leistungsfähigkeit).

1. Wahrscheinliche negative Leistungsverzerrung (effort bias). Hinweise auf Aggravation bei der Erfassung der Leistungsfähigkeit in einer oder mehreren validierten Messungen der körperlichen Leistungsfähigkeit (zum Beispiel die Messung der Handkraft).
2. Diskrepanz zwischen berichteten Symptomen und der physiologischen Reaktion. Beispiel: fehlende Zunahme der Herzfrequenz bei einer berichteten Zunahme der Schmerzintensität.
3. Befunde im Rahmen der klinischen Untersuchung oder der Erfassung der Funktionellen Leistungsfähigkeit sind nicht vereinbar mit bekannten organischen Erklärungsmechanismen.
4. Diskrepanzen im vom Gutachter oder der Gutachterin beobachteten und unbeobachteten Verhalten.

Kriterium C beschreibt Hinweise auf Aggravation oder Simulation der Behinderung aus kognitiven und neuropsychologischen Tests.

1. Eindeutige negative Antwortverzerrungen (‚response bias') mit einer unwahrscheinlich tiefen Leistung (‚below chance') in Alternativwahlverfahren für kognitive Leistungen.
2. Wahrscheinliche negative Antwortverzerrungen im Rahmen eines oder mehrerer validierter Tests für Aggravation bei der Beschreibung der subjektiven Symptome oder kognitiven Leistungsfähigkeit.
3. Eine Diskrepanz zwischen dem kognitiven Leistungsprofil und bekannten klinischen Mustern der kognitiven Hirnleistung wird als Hinweis auf Aggravation interpretiert.
4. Diskrepanz zwischen neuropsychologischen Testergebnissen und der ausserhalb der Testsituation beobachteten kognitiven Leitungsfähigkeit.

Kriterium D betrifft Hinweise auf Aggravation aus der Beschreibung der Symptome, Beschwerden und Behinderung durch den Exploranden.

1. Eindeutige Inkonsistenzen zwischen der durch den Exploranden beschriebenen kognitiven oder körperlichen Leistung und dem unbeobachteten Verhalten.
2. Diskrepanzen zwischen den Angaben der Exploranden über die Vorgeschichte, und den Angaben in der medizinischen Dokumentation. Vorbestehende Erkrankungen können zum Beispiel bagatellisiert oder verschwiegen werden und der Gesundheitszustand vor dem Unfall wird als gut bezeichnet.
3. Die Symptombeschreibung stimmt nicht mit bekannten Mustern überein und ist physiologisch oder neurologisch nicht erklärbar, zum Beispiel wenn bei einer einseitigen Diskushernie in der Halswirbelsäule Schmerzen im ganzen Körper beschrieben werden.

4. Unterschiede zwischen der durch den Exploranden beschriebenen kognitiven oder körperlichen Leistung und dem im Rahmen der Begutachtung beobachteten Verhalten.
5. Starke Hinweise auf Aggravation körperlicher Symptome aus der psychologischen Evaluation, zum Beispiel in Subskalen des MMPI-2 für körperliche und emotionale Symptome.

Kriterium E beschreibt die wichtige Bedingung, dass die Befunde unter A bis D nicht vollständig durch psychiatrische, neurologische oder entwicklungsbedingte Faktoren erklärbar sind. Die Beobachtungen die die Kriterien erfüllen, repräsentieren wahrscheinlich bewusstseinsnahe Verhaltensweisen mit dem Ziel eines sekundäreren Gewinns. Zum Schluss wird betont, dass eine gleichzeitig objektiv beschriebene Pathologie oder Erkrankung (auch psychiatrisch) die Aggravation nicht ausschliesst.

Bianchini et al. berücksichtigen, dass es unterschiedliche Ausprägungen von Aggravation und Simulation gibt. Kriterium A und E sind notwendige Voraussetzungen. Die Kriterien B-D werden unterschiedlich stark gewichtet wobei C1 und D1, in Anbetracht der guten Forschungsgrundlage, stark gewichtet werden. Die anderen Kriterien haben eine schwächere wissenschaftliche Basis und werden nicht so stark gewichtet. Bianchini et al. beschreiben, in Anlehnung an Slick et al., drei Ausprägungsstufen der Aggravation:

- Als **Eindeutig** (definitive) gilt die Aggravation wenn ein substantieller Anreiz vorliegt (Kriterium A), ein eindeutiger Hinweis auf Absicht besteht (Kriterium C1 oder D1) und die Befunde nicht anders erklärt werden können (Kriterium E).
- **Wahrscheinlich** (probable) ist die Aggravation, wenn ein substantieller Anreiz vorliegt (Kriterium A), zwei oder mehr wahrscheinliche Hinweise vorliegen (Kriterien B, C2-5 und D2-5) und die Befunde nicht anders erklärt werden können (Kriterium E).
- **Möglich** (possible) ist eine Aggravation wenn Kriterium A und E erfüllt sind, während die Hinweise im Rahmen der Kriterien B-D ungenügend sind für die Diagnose einer wahrscheinlichen Aggravation.

2.3.5 Fazit

In der Schweiz liegen noch keine systematisch erhobenen Zahlen zur Prävalenz einer reduzierten BV bei Exploranden oder Explorandinnen mit schwer objektivierbaren Diagnosen vor. Viele Studien im gutachterlichen Kontext, sowohl in Europa als auch in Amerika, zeigen, dass die Prävalenz einer reduzierten BV in gewissen Kontexten beachtlich ist.

Für die Beurteilung der BV stehen standardisierte validierte Tests und Leitlinien zur Verfügung. Die Methoden schliessen sich nicht gegenseitig aus sondern ergänzen sich vielmehr.

Standardisierte und validierte BVT liegen für einzelne kognitive Leistungen vor. Für viele andere Bereiche sind sie nicht vorhanden. Der höchste wissenschaftliche Wert wird in der Forschung den eindeutigen negativen Antwortverzerrungen (‚response bias') beigemessen. Dabei liegen Testergebnisse vor mit einer unwahrscheinlich tiefen Leistung in Alternativwahlverfahren für kognitive Leistungen, entweder ‚below chance' oder unterhalb der minimal zu erwartenden Leistung (‚Testdeckeneffekt').

Leitlinien nehmen in der Beurteilung der BV einen zentralen Platz ein. Mit Hilfe der Leitlinien kann die Validität der Befunde im Rahmen der Begutachtung überprüft werden. Dabei können alle verfügbaren somatischen und kognitiven Informationen berücksichtigt werden. Am meisten verbreitet sind die Leitlinien von Slick et al. für Exploranden mit einer kognitiven Behinderung, und die von Bianchini et al. für Exploranden mit einer physischen und kognitiven Behinderung in Zusammenhang mit Schmerzen. Die Verwendung von Leitlinien kann aus zwei Gründen empfohlen werden. Für viele Bereiche sind keine standardisierten validierten BVT verfügbar. Hinzu kommt, dass die BV nicht mit standardisierten BVT alleine beurteilt werden kann. Wissenschaftliche Expertinnen und Experten im Bereich standardisierter BVT empfehlen nachdrücklich die Berücksichtigung entsprechender Leitlinien.

Als eindeutige Hinweise auf eine reduzierte BV interpretieren Bianchini et al. sowohl negative Antwortverzerrungen als auch Inkonsistenzen zwischen der kognitiven oder körperlichen Leistung, die der Explorand beschreibt, und seinem Verhalten.

3 Beschwerdevalidierungstests: Systematische Literaturrecherche

3.1 Vorgehen

Für die systematische Erfassung relevanter Literatur zu Beschwerdevalidierungstests wurde ein mehrstufiges Vorgehen gewählt. Auf der Grundlage von Überblicksarbeiten zur Symptomvalidierung und zur Fragestellung zu aggravierendem und simulierendem Verhalten wurden Fachbegriffe im Themenbereich gesammelt und zu einer konkreten Suchstrategie verdichtet. Um der Mehrsprachigkeit der Schweiz gerecht zu werden, wurden neben englischen Suchbegriffen auch deutsche und französische Ausdrücke berücksichtigt. Mit diesem Instrumentarium konnten dann die verschiedenen zentralen Datenbanken abgefragt und die Resultate in ein Literaturverwaltungsprogramm aufgenommen werden. Die gefundenen Referenzen wurden darauf geprüft, ob es sich um Beiträge handelt, welche sich mit Simulation (Malingering) und damit zusammenhängenden Testverfahren befassten.

Aufgrund der grossen Zahl gefundener Beiträge mussten für die weitere Analyse Einschränkungskriterien zur Auswahl der zu konsultierenden Texte gefunden werden. Auf der Basis von Beiträgen zu in deutscher Sprache erhältlichen Tests und unter Beizug einer für die Jahre 2002 bis 2005 durchgeführten Review (Blaskewitz & Merten, 2007) konnte dann eine Liste mit wichtigen Tests erstellt werden. Für zwei ausgewählte Verfahren erfolgte eine ausführlichere Analyse der veröffentlichten Studien.

3.1.1 Datenbanken

Als wichtigste Datenbanken für den Bereich der Medizin- und Gesundheitswissenschaften wurden Medline und − wegen der etwas grösseren Reichweite − auch Pubmed abgefragt. Als Portale mit spezifischem Fokus auf Psychologie und Psychiatrie kamen PsychInfo und der auf Tests fokussierte deutschsprachige Psyndex hinzu.

Um auch allfällige Artikel aus weiteren Fachbereichen der Gesundheitsversogung, wie Physiotherapie oder Pflege, finden zu können, wurde noch CINAHL in die Recherche mit einbezogen. Für die Berücksichtigung von Literatur aus einem eher allgemein sozialwissenschaftlichen Hintergrund erfolgte der Einbezug der multidisziplinären Datenbank ISI Web of Science. Zusätzlich zu diesen Abfragen wurde für die spezifische Suche nach französischsprachigen Referenzen auf die Banque de Données Santé Publique (BDSP) zugegriffen und die aus Frankreich stammende multidisziplinäre Datenbank Francis abgefragt.

3.1.2 Definition der Suchstrategie

Die Festlegung der Suchstrategie und die Auswahl zentraler Suchbegriffe erfolgten im Rahmen der Konsultation einer Auswahl von Texten zum Themenbereich. Mit diesen Begriffen wurden erste Testsuchläufe unternommen, auf deren Basis die Strategie verfeinert werden konnte. Aufgrund dessen mussten Begriffen, wie „simulation“, „response bias“ oder „deception“, welche aufgrund ihrer

allgemeinen Bedeutung teilweise zu mehr als 10'000 (in der grossen Mehrheit nicht themenrelevanten) Treffern führten, ausgeschlossen werden.

Dies führte für die definitive Suchstrategie zum Themenkomplex von Aggravation und Simulation zu folgenden englischen Begriffen:

> *malingering; symptom exaggeration; effort bias; symptom validity; sincerity of effort; false symptoms; feigning; fabricated illness; feigned illness*

Diese Terme wurden kombiniert mit einer Sammlung von Begriffen für Tests, Verfahren oder Gutachten wie:

> *measure; test; scale; instrument; psychometrics; inventory; questionnaire; assessment; expertise; opinion; guideline*

Für die verschiedenen Suchmaschinen der abgefragten Datenbanken musste diese Suchstrategie jeweils im Einzelnen angepasst werden. Für die Suche über die gemeinsam abfragbaren Datenbanken Medline, PsychInfo, Psyndex und CINAHL ergab sich folgende konkrete Syntax:

> *(((malinger* or symptom exaggerat* or effort bias or symptom validity or sincerity of effort or false symptom* or feigning or fabricated illness or feigned illness) and (measur* or test* or scale* or instrument* or psychometric* or inventory or questionnaire* or assessment* or expertise or opinion or guideline*)) in AB)or (((malinger* or symptom exaggerat* or effort bias or symptom validity or sincerity of effort or false symptom* or feigning or fabricated illness or feigned illness) and (measur* or test* or scale* or instrument* or psychometric* or inventory or questionnaire* or assessment* or expertise or opinion or guideline*)) in TI)*

3.1.3 Erhaltene Referenzen

Die Recherche über die Suchmaske von Medline inklusive der assoziierten Datenbanken führte eingeschränkt auf Jahrgänge ab 1992 bereits zu 1660 Treffern. Nach Einbezug der weiteren Datenbanken erhöhte sich die Zahl auf 2100 Artikel, wobei auch einige doppelt gezählte enthalten waren.

Eine Analyse der zeitlichen Verteilung der gefundenen Veröffentlichungen liess erkennen, dass sich die Forschungsaktivitäten in den letzten Jahren intensiviert hatten. In einer noch nicht von doppelten Nennungen (ca. 30%) bereinigten Aufstellung fanden sich, für die fünf Jahre von 1992 bis und mit 1996, rund 370 Beiträge. In der auf Erscheinungen seit 1997 reduzierten, sowie von doppelten Nennungen befreiten, Bibliothek fanden sich noch rund 1100 Referenzen. Davon entfielen 570 Beiträge auf die letzten fünf Jahre von 2003 bis und mit 2007 wovon 340 Veröffentlichungen seit 2005 erschienen waren.

Aufgrund der grossen Zahl der gefundenen Beiträge musste für eine ausführlichere Betrachtung der Beträge aufgrund von Abstracts und Titeln eine zeitliche Einschränkung auf seit 2005 erschienene Artikel vorgenommen werden. Aus diesen verbleibenden Referenzen wurden Artikel, welche sich nicht mit Testverfahren zu Malingering oder Aggravation befassten, ausgeschieden und die Namen der Tests in einer Liste erfasst. Dadurch reduzierte sich die Zahl der relevanten Referenzen auf 230 für die Jahre ab 2005 worin 56 verschiedene Verfahren zu Abklärungen von Malingering und Aggravation Erwähnung fanden.

Krankheitsbilder, welche in diesen Titeln und Abstracts genannt wurden, waren beispielsweise traumatische Gehirnverletzungen, psychische Krankheiten und post-traumatische Belastungsstö-

rungen sowie etwas seltener auch Beeinträchtigung durch Schmerzen. Häufig genannte Einsatzbereiche waren neuropsychologische Abklärungen im Zusammenhang mit rechtlichen Angelegenheiten oder bei der Regelung von Versicherungsansprüchen.

In Bezug auf die Sprache der gefundenen Artikel zeigte sich, dass mehrere Beiträge in deutscher Sprache erschienen waren. Es konnte aber nur eine einzige Publikation in französischer Sprache gefunden werden, welche sich aber mit einem Fallbeispiel aus einem forensischen Zusammenhang befasste und nicht als relevant eingestuft werden konnte. Auch eine spezifische Suche mit französischen Suchbegriffen in den auf den frankophonen Raum ausgerichteten Datenbanken ergab keine weiteren Treffer. Die bei dieser französischen Recherche identifizierten rund 200 Publikationen befassten sich alle mit nicht studienrelevanten Themen.

Eine Recherche nach deutschsprachigen Texten unter Ergänzung von deutschen Suchbegriffen in Medline (hier werden auch Originaltitel erfasst) ergab nach einer Reduktion auf themenrelevante Beiträge insgesamt 42 Publikationen. Eine Durchsicht der Titel und Abstracts führte zu einer Liste von 30 Tests oder Verfahren, die mehrheitlich auch im deutschen Sprachraum angewandt worden waren. Die genannten Tests bezogen sich überwiegend auf die Abklärung von Malingering im neuropsychologischen Bereich, wobei meist die Untersuchung der kognitiven Leistungsfähigkeit im Vordergrund stand.

3.1.4 Anpassung der Auswertungsstrategie

Aufgrund der grossen Zahl genannter Verfahren war im Rahmen des vorgesehenen Projektumfangs eine umfassende Diskussion der einzelnen Tests nicht möglich. Deshalb drängte sich eine Modifikation der ursprünglich geplanten Auswertungsstrategie auf. Die Zahl von potentiell relevanten Beiträgen musste eingeschränkt werden:

Einerseits sollten nur Tests und Verfahren weiter diskutiert werden, die in einer Schweizer Landessprache vorliegen. Da bereits in der deutschsprachigen Literatur rund 30 Tests vorlagen, musste darüber hinaus für eine eingehende Diskussion der Testgütekriterien eine weitere Beschränkung erfolgen. Deshalb stützten wir uns für die Zusammenstellung eines Sets mit möglichen Tests, für die Anwendung in der Begutachtungspraxis, auf die Resultate einer umfassenden Übersichtsstudie (Blaskewitz & Merten, 2007) und auf nach inhaltlichen Kriterien ausgewählte spezifische Zeitschriftenartikel zu einzelnen Tests.

Andererseits beschränkten wir uns für eine ausführliche Testdiskussion auf zwei einzelne Verfahren. Bei der Auswahl dieser Verfahren, die in Bezug auf eine Validitätsbeurteilung eingehender besprochen werden sollten, konzentrierten wir uns auf Tests für Krankheitsbilder mit grosser Relevanz für die IV. Dazu wurden in einer aktuellen Studie zu nicht konformen Leistungen in der IV insbesondere auch psychische Störungen, chronische Rückenschmerzen, Schleudertraumata gezählt (Ott et al., 2007).

Ausgewählt wurde deshalb der Amsterdamer Kurzzeitgedächtnistest AKGT (engl. als Amsterdamer Short-Term Memory Test ASTM bekannt), welcher von den Testautoren als geeignet bezeichnet wird für Einsatzbereiche wie Schleudertrauma, Chronisches Müdigkeitssyndrom, Multiple Sklerose und isolierte amnestische Syndrome im Zusammenhang mit Gehirntumoren und Schlaganfällen (Schagen, Schmand, deSterke, & Lindeboom, 1997). Der AKGT wurde und auch bereits zur Ab-

schätzung der Wahrscheinlichkeit von aggravierenden Verhaltensweisen im Zusammenhang mit der Diagnose Schleudertrauma eingesetzt (Schmand et al., 1998).

Andererseits wird mit dem Stukturierten Fragebogen simulierter Symptome SFSS (engl. als: Structured Interview of Malingered Symptomatology SIMS bekannt) ein Test ausführlicher diskutiert, welcher für die Untersuchung von Malingering im Rahmen von psychiatrischen Störungen eingesetzt werden kann. In einer Review zu Verfahren der Erkennung von Malingering im psychiatrischen Zusammenhang nennen Singh, Avasthi und Grover (2007) als wichtigste Tests neben dem SIMS noch das Structured Interview of Reported Symptoms SIRS, Unterskalen des MMPI-2 (Minnesota Multiphasic Personality Inventory) und des PAI (Personality Assessment Inventory). Im deutschen Sprachraum wurden von diesen Tests bislang nur das MMPI-2 und das SIMS (SFSS in der deutschen Übersetzung) angewandt. Während es sich beim MMPI-2 um ein umfassendes Instrument mit über 568 Fragen zur Beschreibung der Persönlichkeit und zur Erfassung von psychischen Beschwerden handelt, wurde das SIMS explizit als Instrument zur Erkennung von Aggravation oder Simulation entwickelt und ist aufgrund seiner Kürze (75 einfache mit Ja oder Nein zu beantwortende Fragen) gut einsetzbar und einfach in der Handhabe.

3.2 Steigendes Interesse an Beschwerdevalidierungstests

Wie die Zunahme von Studien zum Themenbereich Aggravation und Malingering zeigt, wächst das Interesse an einer wissenschaftlichen Auseinandersetzung mit Methoden und Verfahren der Erkennung von suboptimalem Leistungsverhalten und der Beschwerdevalidierung zunehmend auch im europäischen Kontext (Blaskewitz & Merten, 2007). Dieses Interesse gründet zum Einen auf der Erkenntnis, dass in Begutachtungssituationen mit einem gewissen Anteil von Personen, die Symptome aggravieren oder simulieren, gerechnet werden muss und zum Anderen auf der Einsicht, dass Simulation und Aggravation bei neuropsychologischen Tests zu einer Verzerrung der Testresultate führt.

Verschiedene Studien weisen darauf hin, dass im Kontext von Begutachtungen mit einer nicht vernachlässigbaren Zahl von Untersuchten zu rechnen ist, welche im Rahmen von Tests suboptimale Leistungen erbringen. Je nach Setting und Hintergrund der Begutachtung variieren die in verschiedenen Studien eingeschätzten Raten aber beträchtlich. In einer Untersuchung zu Malingering in neuropsychologischen Abklärungen in den USA ermittelten Mittenberg, Aguila-Puentes, Patton, et al. (2002) abhängig von untersuchten Krankheitsbildern Raten von zwischen 8% und 35% für Aggravation und mutmassliches Malingering. In einer Untersuchung bei Personen mit kognitiven Beeinträchtigungen nach Unfällen, fanden Merten et al. (2006) in Deutschland bei über 44% der Begutachteten Hinweise für suboptimales Leistungsverhalten. Auch wenn diese ermittelten Raten nicht einfach auf andere Bereiche übertragen werden können, weisen sie darauf hin, dass im Rahmen von Begutachtungen durchaus mit einer „qualifizierten Minderheit“ (Blaskewitz & Merten, 2007) von Personen mit verzerrendem Antwortverhalten gerechnet werden muss.

Neben der Erwartung der Existenz von Malingering in Begutachtungszusammenhängen wird als weiterer Grund für die Notwendigkeit der Durchführung von Beschwerdevalidierungstests auf die Folgen von negativen Antwortverzerrungen für die Zuverlässigkeit neuropsychologischer Testverfahren hingewiesen (Merten, Friedel, Mehren et al. 2007). Diesbezüglich wird in einem Positionspapier

der National Academy of Neuropsychology (Bush et al., 2005) angemerkt, dass ein Assessment der Validität der Antworten Teil einer jeden neuropsychologischen Abklärung sein sollte. Der Einsatz von Tests zur Beschwerdevalidierung dient nach dieser Argumentation auch der allgemeinen Verbesserung der Qualität in der neuropsychologischen Diagnostik.

3.3 Gütekriterien für Beschwerdevalidierungstests

Aufgrund der wachsenden Zahl von Tests zur Beschwerdevalidierung erlangt die Beurteilung dieser Verfahren eine wachsende Bedeutung. Dazu hat Hartman (2002) ein Set von acht Kriterien entwickelt, die auch in nachfolgenden Arbeiten wiederholt aufgegriffen worden sind (Blaskewitz & Merten, 2007). Nach Hartman sollten Tests zur Beurteilung von Leistungsbereitschaft und Simulation:

- Anstrengungsbereitschaft messen, aber nicht von den zu messenden kognitiven Störungen beeinflusst werden (Sensitivität und Spezifität).
- bei den Patienten und Patientinnen den Eindruck erwecken, dass es sich um eine realistische Messung der untersuchten kognitiven Funktion handelt (Augenscheinvalidität).
- Symptombereiche messen, welche von Patienten und Patientinnen häufig übertrieben werden.
- auf einer strengen wissenschaftlichen Normierung fundieren.
- auf Validierungstudien beruhen, die gesunde Probanden, Patientengruppen und Individuen umfassen, die verdächtigt oder erwiesen simulieren, in Bedingungen der Abklärung von Arbeitsunfähigkeit oder im Rahmen forensischer Begutachtungen.
- schwierig zu verfälschen oder zu coachen sein.
- einfach anzuwenden sein.
- von aktuellen Forschungsergebnissen gestützt werden.

Diese Anforderungen werden in der von Hartman aufgeführten Zusammenstellung verbreiteter Verfahren von vielen Tests nicht erfüllt – zumindest nicht in allen Punkten. Nur der Word Memory Test erfüllt alle Kriterien. Andere häufig angewandte Tests wie beispielsweise der Rey 15 Item Test oder der Dot Counting Test erfüllen keinen einzigen oder nur einen der aufgeführten Punkte.

Diese Kriterien zur Testbeurteilung sollen in dieser Literaturstudie nachfolgend auch für die Bewertung von zwei eingehender besprochenen Tests zur Anwendung kommen.

3.4 Diskussion ausgewählter Tests: Strukturierter Fragebogen Simulierter Symptome SFSS / (engl.) SIMS

Seit 2003 liegt mit dem Strukturierten Fragebogen Simulierter Symptome SFSS ein auf Deutsch übersetzter Beschwerdevalidierungstest für den Einsatz bei psychischen Krankheitsbildern vor. Der SFSS wurde nicht als strenger Test zur Unterscheidung zwischen authentischer und simulierter Symptomartikulation entwickelt, sondern als relativ einfach anwendbares Instrument, mit dessen

Hilfe ein erster Eindruck gewonnen werden kann. Dabei wird davon ausgegangen, dass diese erste Erkenntnis allenfalls in ergänzenden Tests weiter vertieft werden muss. Die in verschiedenen Studien gefundenen niedrigen Werte für die Spezifitäten und das damit einhergehende Problem von vielen falsch-positiven Ergebnissen bestätigen dieses Vorgehen in der Anwendung.

3.4.1 Entwicklung und Aufbau des SFSS / SIMS

Das Structured Inventory of Malingered Symptomatology SIMS ist Ende der 90er Jahre in den USA entwickelt worden (Smith & Burger, 1997). Es wurde konzipiert als leicht handhabbare Screeningmethode für das Erkennen einer Vielzahl von simulierten Symptomen. Zusammengesetzt aus neu gebildeten Fragen und simulationssensiblen Items bereits bestehender Messinstrumente (bspw. MMPI, SIRS WAIS-R[8]), wurde das SIMS als Selbsteinschätzungsfragebogen mit 75 Items erarbeitet. Die Fragen müssen jeweils mit Ja oder Nein beantwortet werden und sind für die Auswertung zu 5 Skalen mit jeweils 15 Fragen geordnet. Die einzelnen Skalen erfassen häufige vorgetäuschte Beschwerdeklassen wie: niedrige Intelligenz, affektive Störung, neurologische Beeinträchtigung, Psychose und amnestische Störung.

Für die Bestimmung von Hinweisen auf simulierte Beschwerden werden verschiedene Fragetypen eingesetzt. Einige Items enthalten Aussagen zu bestimmten Syndromen (bspw. Depression), nennen aber Symptome, welche aus der Sicht von Fachleuten sehr atypisch sind (bspw. „Je depressiver ich bin, um so mehr möchte ich essen“). Andere Fragen nennen bizarre Symptome (bspw. „Wenn ich Stimmen höre, fühlt es sich an, als würden meine Zähne aus dem Körper heraustreten“). weitere beziehen sich auf Tatsachen und nennen entweder fast richtige oder tatsächliche Zusammenhänge (bspw. „Gold und Silber ähneln sich, beides sind Metalle.“ / „Die Woche hat 6 Tage“). Aus den Antworten, die auf Simulation hinweisen, können für die Interpretation ein Gesamtscore sowie spezifische Werte für die einzelnen Skalen berechnet werden. Das SIMS ist auf Deutsch übersetzt worden und in dieser Form in einer Studie mit forensischen Patienten und Patientinnen und Studierenden bezüglich seiner psychometrischen Eigenschaften überprüft worden (Cima et al., 2003; Maaike Cima, Pantus, & Dams, 2007)[9].

3.4.2 Zusammenfassung der Erkenntnisse aus Studien zum SFSS / SIMS

Die verschiedenen Studien, welche sich mit dem Structured Inventory of Malingered Symptomatology oder seiner deutschen Version, dem Strukturierten Fragebogen Simulierter Symptome befassen, zeigen ein uneinheitliches Bild. Einerseits konnten in sogenannten Analogstudien mit Studierenden, die angeleitet wurden Symptome zu simulieren, durchaus gute Werte für Spezifität und Sensitivität des Tests eruiert werden (Merckelbach & Smith, 2003; Smith & Burger, 1997). Andererseits reduzierten sich diese Werte im Rahmen von Studien mit Patientinnen und Patienten aber drastisch und es ergaben sich teilweise hohe Anteile an falsch-positiven Testergebnissen (Cima et al., 2003; E-

[8] MMPI: Minnesota Multiphasic Personality Inventory-2
SIRS: Structured Inventory of Reported Symptoms
WAIS-R Wechsler Adult Intelligence Scale Revised

[9] Im Artikel von Cima et al. (2003) ist der Fragebogen mit allen Fragen abgedruckt.

dens, Poythress, & Watkins-Clay, 2007; Merckelbach & Smith, 2003; Vitacco, Rogers, Gabel, & Munizza, 2007). Diese Ergebnisse legen nahe, dass der Test, wie von mehreren Autorinnen und Autoren vorgeschlagen (Edens et al., 2007; Lewis, Simcox, & Berry, 2002), nur als Screeninginstrument zur Ermittlung eines ersten Eindruckes angewandt und gegebenenfalls mit spezifischeren Instrumenten oder Verfahren ergänzt wird. Eine detaillierte Darstellung der Erkenntnisse aus den gesichteten Veröffentlichungen zum SFSS / SIMS findet sich in einem nachfolgenden Kasten (Seite 38).

3.4.3 Bewertung des SFSS / SIMS

Eine Beurteilung des SFSS / SIMS auf der Basis der bislang zu diesem Test vorliegenden Studien ist schwierig. So ist der Forschungsstand noch nicht sehr breit und die Ergebnisse fallen je nach Studiendesign unterschiedlich aus. Dieses durchzogene Bild bestätigt sich auch beim Versuch der Bewertung des Tests mit den Kriterien von Hartmann (siehe dazu Tabelle 5). In einigen Kriterien, wie beispielsweise der einfachen Anwendbarkeit, schneidet der Test relativ gut ab und in anderen Kriterien wie etwa bei den errechneten Spezifitäten im Rahmen von Untersuchungen in Anwendungsbereichen mit Patientinnen und Patienten fällt das Urteil eher negativ aus.

Eine Empfehlung für oder gegen den Einsatz des SFSS / SIMS in der gutachterlichen Praxis kann auf Grund dieser Literaturanalyse nicht gegeben werden. Die konsultierten Veröffentlichungen behandeln entweder Testüberprüfungen in analogen Settings oder Erkenntnisse aus Studiendesigns in forensischen Zusammenhängen. Daraus kann nicht direkt abgeleitet werden, ob und in welchem Masse der Test für die Anwendung in der Begutachtung von Rentenabklärungen geeignet ist. Hier obliegt es den Fachleuten in der Praxis zu beurteilen, ob die mit dem SFSS mögliche Eruierung von simulierten Beschwerdenbereichen wie: niedrige Intelligenz, affektive Störung, neurologische Beeinträchtigung, Psychose und Amnestische Störung, von Bedeutung sind.

Wenn es im Rahmen von Begutachtungen darum geht, mit überschaubarem Aufwand einen ersten Eindruck darüber zu erhalten, ob einer der geschilderten Beschwerdenbereiche potentiell von einem Exploranden simuliert wird, und wenn dabei positive Testresultate bloss als weiter zu verfolgende Hinweise auf Aggravation oder Simulation gewertet werden, könnte der Strukturierte Fragebogen Simulierter Symptome aber durchaus einen nützlichen Beitrag leisten.

Tabelle 5 Der Strukturierte Fragebogen simulierter Symptome SFSS aus der Sicht der Gütekriterien nach Hartmann

Messung des vorgeschlagenen Symptombereiches; Insensibilität gegenüber relevanten Störungen (Sensitivität; Spezifität)	In Analogstudien widerspiegeln sich die zur Simulation instruierten Beschwerdenbereiche in den entsprechenden Skalen. Die anhand von Analogstudien ermittelten hohen Werte für Sensitivität und Spezifität bestätigen sich aber in der Anwendung bei Patientinnen und Patenten mit relevanten Störungen nicht. In solchen Settings ist mit hohen Werten von falschpositiven Resultaten zu rechnen.
Wird von Testpersonen als realistische Messung einer untersuchten Funktion wahrgenommen	Es liegen noch kaum entsprechende Studien vor. Es wird aber darauf hingewiesen, dass der Test vor allem für „nicht differenziert vorbereitete“ Personen geeignet sei (Vitacco et al., 2007).
Messung häufig übertriebener Symptombereiche	Der Test wurde als eine Kombination von Items aus bereits bewährten Verfahren entwickelt. Dadurch ist gewährleistet, dass auch häufig übertriebene Symptombereiche erfasst werden.
Strenge wissenschaftliche Normierung	Die Forschung zum SIMS / SFSS steckt noch in den Kinderschuhen, die wissenschaftliche Normierung ist noch ausbaufähig.
Validierungsstudien in unterschiedlichen Settings	Der Test ist zwar vorbildlich in Validierungsstudien mit gesunden Probanden und Patientengruppen überprüft worden und was besonders hervorzuheben ist, auch in relevanten Einsatzgebieten mit experimentell simulierenden Personen validiert. Gerade in diesen Settings, erweist sich aber die Klassifikationsgüte des Tests als problematisch.
Anfälligkeit auf Coaching und Verfälschung	Es liegen uns keine entsprechenden Studien vor.
Einfache Anwendbarkeit	Als Instrument der Selbstbeurteilung mit 75 Items einfach anzuwenden.
Aktuellem Forschungsstand entsprechend	Der aktuelle Forschungsstand ist noch eher knapp.

Ausführliche Diskussion 1 Erkenntnisse zum SFSS / SIMS

Sensitivität und Spezifität in Analogstudien und in Studien mit Patientinnen und Patienten

In Bezug auf die Bewertung der Gütekriterien des Structured Inventory of Malingered Symptomatology zeigt sich ein uneinheitliches Bild mit unterschiedlichen Studienergebnissen. In einer im Rahmen der Testkonstruktion durchgeführten Studie überprüften Smith und Burger (1997) ihren Test zusammen mit der MMPI-2 „Faking Bad – Scale“ an 476 Studierenden und fanden für das SIMS eine Sensitivität von 95.6 %, ein Wert, der die „Faking Bad-Scale“ des MMPI in den Schatten stellte.

Nachfolgende Untersuchungen in gemischten Settings, in denen neben gesunden Testpersonen, welche Symptome simulierten, jeweils auch Zielgruppen der Testanwendung einbezogen wurden, zeigten zwar weiterhin hohe Sensitivitäten und Spezifitäten bei gesunden Personen, aber deutlich niedrigere Werte für die Testgüte innerhalb der Zielgruppen. In einer Untersuchung bei 64 Männern, welche in einem staatlichen medizinischen Zentrum forensische Abklärungen zur Feststellung einer Gerichtsfähigkeit machen mussten, fanden Lewis et al. (2002) lediglich eine Spezifität von 61% für das SIMS unter Verwendung des Structured Inventory of Reported Symptoms SIRS als Gold-Standard für die Einteilung in die Gruppe von Malingerern. Auch Vitacco et al. (2007) fanden mit einem fast identischen Studien-Design zwar gute interne Konsistenzen für die Skalen des SIMS aber mit 65% Spezifität ebenfalls eine unbefriedigende Trennschärfe des Tests.

Weitere Aussagen zu Gütekriterien

In einer Studie über verschiedene Samples, in welche teilweise auch psychiatrische Patientinnen und Patienten eingeschlossen waren, fanden Merckelbach und Smith (2003) eine gute Test-Retest-Korrelation und befriedigende Konsistenzen der Skalen aber Hinweise für eine Sensibilität des SIMS für Psychopathologien. Im umfassenden Sample mit ehrlich antwortenden Studierenden, ehrlich antwortende psychiatrische Patientinnen und Patienten und simulierenden Studierenden erreichte das SIMS eine Sensitivität von 93% und eine Spezifität von 98%. In der Untergruppe von psychiatrischen Patientinnen und Patienten und Studierende mit psychischen Problemen reduzierte sich die Spezifität auf 84%. Eine bedeutend niedrigere Spezifität von 40% (allerdings mit tieferem Cut-off von 14 statt 16) unter psychiatrischen Patienten und Patientinnen, fanden hingegen Edens et al. (2007) im Rahmen einer vergleichenden Studie mit korrekt antwortenden Gesunden, simulierende Gesunde und Insassen einer psychiatrischen Station eines Gefängnisses, bei welchen 26 von 56 Insassen aufgrund der Krankenakten und anderen Untersuchungen als Malingerer eingestuft wurden.

Deutsche Übersetzung und Überprüfung der deutschen Version

Zur deutschen Version des SIMS, dem Strukturierten Fragebogen Simulierter Symptome (SFSS) liegen bislang erst zwei Studien vor. Cima et al. (2003) fanden in einer heterogenen Stichprobe mit forensischen Patientinnen und Patienten und Studierenden eine adäquate Reliabilität und interne Konsistenz sowie eine gute vorhersagende Validität in dem Sinne, dass Personen, welche bestimmte Symptome simulierten, in den entsprechenden Unterskalen auch auffällige Werte zeigten. Auch die übereinstimmende Validität zwischen dem SFSS und den F-Skalen des MMPI-2 sowie den Hinweisen auf Simulation aus den Krankenakten wurde auf der Basis von Korrelationen als befriedigend beschrieben. In Bezug auf Sensitivität und Spezifität wurden Werte von 87% bzw. 86% gefunden, welche auch in der Untergruppe der forensischen Patientinnen und Patienten nicht niedriger waren.

In einer Analyse von Akten eines neurologisch-psychiatrischen Gutachteninstitutes untersuchten Merten et al. (2007) eine Stichprobe von 378 Gutachtenprobanden für welche 339 Word Memory Test-Protokolle (WMT) und 198 Strukturierte Fragebogen Simulierter Symptome (SFSS) vorlagen. Die Autoren und Autorinnen beurteilen die internen Konsistenzen der Unterskalen des SFSS insgesamt etwas kritischer, als andere Autoren und Autorinnen. In Bezug auf die Validität des Strukturierten Fragebogens Simulierter Symptome finden sie aber eine Übereinstimmung von 77% zwischen den Klassifikationen des WMT und des SFSS. Diese Übereinstimmung bestätigte sich in der signifikanten Produkt-Moment-Korrelation von 0.57 bis 0.70 zwischen dem SFSS und verschiedenen Skalen des WMT im Rahmen der anderswo publizierten Resultate aus einem Subsample der gleichen Studie.

Fazit zum Forschungsstand

Auf internationaler Ebene ist der Strukturierte Fragebogen Simulierter Symptome oder seine englische Ursprungsversion SIMS bisher noch nicht sehr intensiv beforscht worden. Insbesondere die Forschung bezüglich seiner klinischen Nützlichkeit befindet sich noch in den Kinderschuhen (Edens et al., 2007). Gleichwohl lassen sich einige vorläufige Schlüsse in Bezug auf Vor- und Nachteile und wissenschaftliche Gütekriterien festhalten. Wie von verschiedenen Autorinnen und Autoren festgehalten wurde, ist das SIMS als Screening-Instrument konzipiert. So hält sich im Vergleich zu anderen Instrumenten, welche sich für die Erkennung von Aggravation oder Simulation psychischer Krankheiten verwenden lassen, wie zum Beispiel dem MMPI-2, der Zeitaufwand für die Beantwortung der 75 Fragen in Grenzen. Für eine Verwendung als Screening-Instrument spricht auch die Tatsache, dass mittels der fünf Unterskalen des Tests auch Aussagen dazu gemacht werden können, welche Beschwerdeklassen simuliert oder aggraviert werden. Aufgrund der gefundenen Spezifitäten, welche vor allem bei der Testanwendung in potentiellen Einsatzgebieten mit 61% bis 87% eher niedrig ausfielen, muss aber mit einer relativ hohen Rate von falsch-positiven Ergebnissen (13% bis 41%) gerechnet werden. Auch wenn der Test im Rahmen einiger weniger Studien gute Resultate im Hinblick auf die Vergleichbarkeit der Ergebnisse mit anderen Beschwerdevalidierungstests oder Testskalen zu Erkennung von Malingering zeigte (inhaltliche Validierung), kann der Strukturierte Fragebogen Simulierter Symptome aufgrund dieser hohen Rate falsch-positiver Resultate nicht als alleine verwendetes Instrument benutzt werden. Schon Lewis et al. (2002) und neulich auch Edens et al. (2007) schlugen als Fazit ihrer Studien vor, dass SIMS dazu einzusetzen, bei der Abklärung von eventuellem Malingering jene Personen auszuschliessen, bei denen ein Vortäuschen von Symptomen unwahrscheinlich erscheine und jene, welche positive Resultate im SIMS zeigten, mit aufwendigeren Instrumenten wie dem MMPI-2 oder dem SIRS weiter abzuklären.

3.5 Diskussion ausgewählter Tests: Amsterdamer Kurzzeitgedächtnistest AKGT / (engl.) ASMT

Mit dem Amsterdamer Kurzzeitgedächtnistest wurde ein spezifischer BVT entwickelt, der zur Erfassung der Leistungsmotivation bei Konzentrations- und Gedächtnisstörungen eingesetzt werden kann. Der Test kann Aufgrund der erfassten Beschwerdenbereiche bei verschiedenen Diagnosen eingesetzt werden. Beschrieben ist auch der Einsatz bei Personen mit Schleudertrauma. Bislang liegen zur Testvalidierung vor allem Analogstudien vor. In diesen Studiendesigns erreicht der Test gute Sensitivitäten und fast perfekte Spezifitäten.

3.5.1 Entwicklung und Aufbau des AKGT

Der Amsterdamer Kurzzeitgedächtnistest wurde als ein Verfahren zur Erfassung negativer Antwortverzerrungen beziehungsweise einer unzureichenden Leistungsmotivation im Rahmen neuropsychologischer Untersuchungen entwickelt. Der Test wird den Probanden gegenüber als Test zu Konzentration und Gedächtnis vorgestellt (Schmand, Lindeboom, Merten, & Millis, 2005). In zwei Übungsdurchgängen und 30 Auswertungsrunden werden den Testpersonen jeweils fünf gedruckte Worte der gleichen semantischen Kategorie (beispielsweise Ländernamen) präsentiert, welche laut vorgelesen und erinnert werden müssen. Nach einer kurzen Ablenkung muss das Individuum aus fünf Worten der gleichen Kategorie die drei mit der zuerst gezeigten Liste übereinstimmenden Begriffe identifizieren (Schmand et al., 1998). Dabei wird der Testperson jeweils die Anzahl der korrekt erkannten Worte genannt. Auf diese Weise können maximal 90 (30x3) Punkte erreicht werden. Selbst Patientinnen und Patienten mit Gedächtnisstörungen nach einem Schädel-Hirn-Trauma oder Patienten und Patientinnen mit Amnestischem Syndrom schnitten in einer Validationsstudie (Schagen et al., 1997) mit Scores von 87 bis 90 sehr gut ab. Dagegen erreichten Personen, welche instruiert wurden, Gedächtnisprobleme zu simulieren, lediglich Scores von 85 oder weniger. Gemäss den Testautoren und -autorinnen sei der Test für Patienten und Patientinnen mit Alzheimer oder dem Korsakoff's-Syndrom (eine degenerative Gehirnschädigung) nicht geeignet, jedoch dienlich für den Einsatz bei Diagnosen wie Schleudertrauma, chronisches Müdigkeitssyndrom, Multiple Sklerose und isolierte amnestische Syndrome im Zusammenhang mit Gehirntumoren und Schlaganfällen. Die Autorinnen und Autoren halten fest, dass der Test einfach in all jene Sprachen übersetzt werden könnte, für welche semantische Norm-Kategorien gebildet worden seien. Bislang liegen aber erst niederländische, englische und deutsche Versionen vor.

3.5.2 Zusammenfassung der Erkenntnisse aus Studien zum AKGT

Der Amsterdamer Kurzzeitgedächtnistest zeigt in verschiedenen Analogstudien hohe Werte für Sensitivität und Spezifität (Bolan, Foster, Schmand et al., 2002; Merten, Green, Henry, et al., 2005; Merten, Henry, & Hilsabeck, 2004; Schagen et al., 1997). Bei klinisch offensichtlichen kognitiven Beeinträchtigungen verringern sich diese Gütekriterien deutlich, es kommt zu einem nicht unbedeutenden Anteil von falsch-positiven Testresultaten (Merten, Bossink, & Schmand, 2007), dabei muss aber erwähnt werden, dass schon die Testautoren und Autorinnen vom Einsatz in diesen Beschwerdebereichen abraten. Im Moment ist die Forschungsgrundlage zum Test mit 17 Untersu-

chungen, welche 2007 im Rahmen einer Testrezension (Herzberg & Frey, 2007) erwähnt wurden, noch ausbaufähig. Wünschenswert wären diesbezüglich auch Untersuchungen mit Symptome aggravierenden oder simulierenden Patientinnen und Patienten.

3.5.3 Bewertung des AKGT

In Bezug auf die von Hartman (2002) vorgeschlagenen Gütekriterien schneidet der Amsterdamer Short-Memory Test relativ gut ab, auch wenn der Forschungsstand in den nächsten Jahren noch weiter ausgebaut werden müsste (siehe Tabelle 6).

Tabelle 6 Der Amsterdamer Kurzzeitgedächtnistest aus der Sicht der Gütekriterien nach Hartmann

Messung des vorgeschlagenen Symptombereiches; Insensibilität gegenüber relevanten Störungen (Sensitivität; Spezifität)	Es zeigen sich zwar Einflüsse von schweren kognitiven Störungen, der Einsatz des Tests kann aber auf Situationen beschränkt werden, wo solche klinisch offensichtlichen Symptome nicht vorliegen. Die in Analogstudien bestimmten Werte für Sensitivität und Spezifität sind durchwegs sehr hoch. In diesen Studiendesigns ergeben sich keine oder kaum falschpositive Testresultate.
Wird von Testpersonen als realistische Messung einer untersuchten Funktion wahrgenommen	Die vorgeschlagene Überprüfung von Konzentration und Gedächtnis wird in der Testkonstruktion umgesetzt.
Messung häufig übertriebener Symptombereiche	Mit Konzentrations- und Gedächtnisstörungen werden neuropsychologisch relevante Beschwerdenbereiche überprüft.
Strenge wissenschaftliche Normierung	Die wissenschaftliche Normierung wird mit einer Basis von 1500 Testpersonen in einer Testrezension als ausreichend bezeichnet (Herzberg & Frey, 2007)
Validierungsstudien in unterschiedlichen Settings	Bislang liegen als Validierungsstudien nur Analogstudien oder Studien mit Patientinnen und Patienten und gesunden simulierenden Personen vor. Eine Überprüfung mit aggravierenden oder simulierenden Patientinnen und Patienten fehlt noch.
Anfälligkeit auf Coaching und Verfälschung	In einer Analogstudie führt ein Coaching zu einer Verringerung der Sensitivität.
Einfache Anwendbarkeit	Der Test kann einfach angewandt werden und dauert nicht lange (10 bis 30 Minuten).
Aktuellem Forschungsstand entsprechend	Der Forschungsstand zum AKGT / ASMT müsste noch weiter ausgebaut werden.

Wenn für eine allfällige Anwendung in der Schweiz der Test, wie von Herzberg und Frey (2007) vorgeschlagen, tatsächlich nur dort eingesetzt wird, wo Testpersonen Störungen angegeben, die klinisch nicht offensichtlich sind, und wenn der Test, wie von mehreren Autorinnen und Autoren eindringlich gefordert (Thomas Merten et al., 2007; Slick et al., 1999), nur im Zusammenhang mit einer eingehenden klinischen Beurteilung angewandt wird, könnte er aufgrund seiner einfachen Anwendung und seiner im Vergleich zu anderen Tests hohen Spezifitäten und Sensitivitäten gleichwohl empfohlen werden.

Von besonderem Interesse ist dabei sein potentieller Einsatz in Situationen, wo ein äusserer Anreiz für Aggravation oder Simulation gegeben ist: bei klinisch schwierig erfassbaren kognitiven Beschwerden, beispielsweise im Zusammenhang von Diagnosen wie Schleudertrauma oder dem Chronischem Müdigkeitssyndrom.

Ausführliche Diskussion 2 Erkenntnisse zum AKGT / ASMT

Sensitivitäten und Spezifitäten

Die Forschungsgrundlagen zum Amsterdamer Short-Term Memory Test sind trotz seiner relativ einfachen Anpassung an verschiedene Sprachen noch relativ gering. So konnten im Rahmen dieser Literaturrecherche keine Studien zur Testgütekriterien gefunden werden, welche sowohl in Bezug auf die Kontrollgruppe, wie auch bei den Malingerern, in einem klinischen Zusammenhang durchgeführt wurden. In einer Rezension des Tests aus dem Jahre 2007 verweisen Herzberg und Frey (2007) auf insgesamt 17 Studien mit Total 1503 Testpersonen, in welchen durchwegs gesunde, zur Simulation instruierte Personen eingesetzt wurden.

Im Rahmen solcher Analogstudien schnitt der Test in Bezug auf Sensitivität und Spezifität meist relativ gut ab. So fanden Merten et al. (2004) in einer relativ kleinen Untersuchung eine Spezifität von 100% bei einer Sensitivität von 90%. Die Autorinnen und Autoren des Tests fanden ihrerseits eine perfekte Zuteilung in einer Studie, in welcher die Testergebnisse von Patienten und Patientinnen mit geschlossenem Schädel-Hirn-Trauma mit den Resultaten von Angehörigen verglichen wurden, welche instruiert wurden, die Symptome der Verwandten so gut wie möglich zu simulieren (Schagen et al., 1997). In einer weiteren mit 36 Teilnehmenden eher kleinen Untersuchung fanden Merten et al. (2005) mit 100% Spezifität und 100% Sensitivität ideale Testgütemasse. Sehr hohe korrekte Klassifikationen fanden auch Bolan et al. (2002) in einer mit 90 Teilnehmenden etwas grösseren Studie, in welcher für den ASTM auch vergleichbare Ergebnisse wie für den Test of Memory Malingering gefunden wurden. Niedrige Sensitivitäten fanden hingegen Jelicic, Merckelbach, Candel et al., 2007) in einer Studie zur Untersuchung des Einflusses von Coaching auf die Testresultate bei gesunden Testpersonen. Während 90% der nicht gecoachten, Symptome simulierenden Personen mit dem ASTM als Malingerer identifiziert werden konnten, wurden in der Gruppe der gecoachten Simulierenden nur noch 70% vom Test als Malingerer ausgewiesen.

Einschränkungen beim Einsatz in Zielgruppen mit schweren Störungen

Wie schon die Testautorinnen und Testautoren bei der Testveröffentlichung erwähnten (Schagen et al., 1997), gibt es für den Amsterdamer Short-Term Memory Tests gewisse Einschränkungen beim Einsatz in Zielgruppen mit schwereren Gedächtnisstörungen. In einer neuen Studie in einem klinischen Setting konnten Merten, Bossink et al. (2007) nachweisen, dass fast die Hälfte (46%) von Patienten und Patientinnen mit offensichtlichen klinischen kognitiven Beeinträchtigungen trotz guter Testmotivation (bona fide Patienten und Patientinnen) fälschlicherweise vom Amsterdamer Short-Term Memory Test als Malingerer eingestuft würden. In einer Testrezension schlagen Herzberg und Frey (2007) folgerichtig einen Einsatz nur in Untersuchungssituationen vor, in denen von den Testpersonen Gedächtnis- und/oder Konzentrationsstörungen angegeben werden, aber keine klinisch offenkundige Störungen vorliegen.

Auch wenn der Amsterdamer Kurzzeitgedächtnistest von Blaskewitz und Merten (2007) als: "originelles Testformat“, welches „solide konstruiert“ sei, beschrieben wurde und mit von seinen Klassifikationsergebnissen mit dem WMT vergleichbar sei, besteht für die weitere Validierung in konkreten Einsatzgebieten weiterhin Forschungsbedarf.

3.6 Weitere wichtige Beschwerdevalidierungstests

Aufgrund des beschränkten Projektumfangs konnten in dieser Literaturrecherche nur die bereits beschriebenen Amsterdamer Kurzzeitgedächtnistest und der Strukturierte Fragebogen Simulierter Symptome ausführlich diskutiert werden. Unter Berücksichtigung einer im Jahre 2007 veröffentlichten Review von Blaskewitz und Merten (2007) und durch Beizug weiterer spezifischer Veröffentli-

chungen liessen sich aber gleichwohl einige Angaben zu weiteren für den deutschen Sprachraum relevanten Tests zusammentragen.

Als wichtige weitere Tests sollen deshalb im folgenden noch Instrumente wie der Word Memory Test (WMT), der Test of Memory Malingering (TOMM), der Medical Symptom Validity Test (MSVT), der Word Completition Memory Test (WMCT), die Testbatterie zur Forensischen Neuropsychologie (TBFN), der Miller Forensic Assessment of Symptoms Test (M-FAST) sowie das Minnesota Multiphasic Personality Inventory (MMPI-2) kurz vorgestellt werden.

Angaben zu anderen Tests, die weniger häufig erwähnt werden oder in den Veröffentlichungen tendenziell kritisch beurteilt wurden, sind in der Tabelle 7 zusammengestellt.

3.6.1 Word Memory Test (WMT)

Der Word Memory Test ist einer der am besten untersuchten Beschwerdevalidierungstests (Blaskewitz & Merten, 2007) und aufgrund seiner geringen Anfälligkeit auf Coaching sowie der Testgüte in Bezug auf Sensitivität und Spezifität sehr beliebt (Brockhaus & Merten, 2004). Laut der Beurteilung von Hartman (2002) sei der WMT einer der besten BVT, die weltweit verfügbar sind. Der Test ist in zwölf Sprachen erhältlich (auch Französisch, Italienisch zurzeit nur in oraler Ausführung) und einfach anwendbar.

Im Testverfahren müssen 15 Substantive vorgelesen und anschliessend erinnert werden. Zuerst sollten dann so viele der genannten Worte wie möglich aus einer Liste von 30 Substantiven wieder erkannt und unterstrichen werden. Nach 10 Minuten werden erneut 15 Substantive vorgelesen, welche anschliessend frei erinnert werden müssen. Für das Testurteil wird das Abschneiden in den beiden Varianten beurteilt. Das freie Aufschreiben ist schwieriger als das Erkennen der Worte in einer Liste. Absichtlich verfälschende Personen zeigen untypischen Antwortmuster im ersten und zweiten Durchlauf (Brockhaus & Merten, 2004).

Potentielle Probleme für die Anwendung ergeben sich allerdings bei Personen mit klinisch offensichtlichen kognitiven Beeinträchtigungen. Bei dieser spezifischen Klientel fanden Merten, Bossink et al. (2007) eine Rate von nahezu 50% falsch-positiven Einstufungen.

3.6.2 Test of Memory Malingering (TOMM)

Der auch ins Deutsche übersetzte Test of Memory Malingering gilt neben dem WMT als der am besten untersuchte eigentliche Beschwerdevalidierungstest (Blaskewitz & Merten, 2007). Auf der Basis einer guten wissenschaftlichen Untersuchung eignet sich der Test für breite Einsatzbereiche wie z.B. bei Posttraumatischen Belastungsstörungen (PTSD), bei Schmerzpatienten und Schmerzpatientinnen in forensischen Abklärungen.

Im Testverfahren müssen aufgrund von 50 Karten mit Strichzeichnungen alltägliche Gegenstände memoriert werden. In wiederholten Durchgängen (nach Ablenkungen) muss dann das ursprüngliche Bild erkannt werden, wenn es zusammen mit einem Hintergrundbild präsentiert wird (Bolan et al., 2002).

Vorsicht geboten sei aber beim Einsatz bei Exploranden mit stark reduzierter Intelligenz, schweren Hirnverletzungen und Demenz (Blaskewitz, 2005) oder bei klinisch offensichtlichen kognitiven Beeinträchtigungen (Merten, Bossink et al., 2007). Die an sich in verschiedenen Validierungsstudien gefundenen hohen Sensitivitäten und Spezifitäten können sich beim Einsatz in diesen Gruppen von Patientinnen und Patienten massiv verringern. In einer Studie bei geistig behinderten Personen ergab sich eine Rate von 41% falsch-positiven Einstufungen (Hurley & Deal, 2006).

3.6.3 Medical Symptom Validity Test (MSVT)

Der Medical Symptom Validity Test ist eine neu entwickelte Kurzform des WMT. Der Test ist in 12 Sprachen erhältlich und auch in einer nonverbalen Form einsetzbar. Im deutschen Sprachraum ist der Test bislang erst in einer kleinen Studie mit gesunden Testpersonen untersucht worden. Hier zeigten sich als ausreichend beurteilte Werte für die Sensitivität und Spezifität (Blaskewitz & Merten, 2006; Merten et al., 2005).

3.6.4 Word Completition Memory Test (WMCT)

Beim Word Completition Memory Test müssen Wortanfänge (i.d.R. 3 Buchstaben) vervollständigt werden. Dies geschieht nach dem Erlernen einer Wortliste, welche Worte enthält, die ebenfalls mit den entsprechenden Buchstaben beginnen. Infolge des Priming aus den erlernten Worten der Liste fallen auch Amnestikern zuerst die vorher gelesenen Worte ein. Absichtlich verfälschende Personen fassen den Test als Gedächtnistest auf und suchen deshalb zuerst andere Worte zur Vervollständigung (Merten et al., 2004). Im deutschen Sprachraum ist der Test bislang nur im Rahmen einer Analogstudie (mit gesunden Testpersonen) überprüft worden (ebd.). Dabei ergab sich mit 100% Sensitivität und Spezifität eine sehr gute Trennschärfe. Für die weitere Validierung sind aber noch zusätzliche Studien notwendig (Blaskewitz & Merten, 2007).

3.6.5 Testbatterie zur Forensischen Neuropsychologie (TBFN)

Bei der Testbatterie zur Forensischen Neuropsychologie handelt es sich um eine deutsche Entwicklung, die aus mehreren Untertests zusammengestellt wurde (Heubrock, Eberl, & Petermann, 2002). Mittels der einbauten Tests soll untersucht werden, ob es sich bei geschilderten sensorischen oder kognitiven Störungen um authentische oder um bewusst oder unbewusst verfälschte Symptome handelt. Durch die verschiedenen eingesetzten Verfahren kann eine relativ breite Palette von Beschwerden untersucht werden. Die Gütekriterien der einzelnen angewandten Tests werden aber im Rahmen eine Rezension eher kritisch beurteilt (Merten, 2003). Bemängelt wird auch eine bislang noch fehlende Kreuzvalidierung und eine als noch unzureichend eingeschätzte Datenbasis (Blaskewitz & Merten, 2007).

3.6.6 Miller Forensic Assessment of Symptoms Test (M-FAST)

Der M-FAST ist ein Interview, welches von Fachpersonen bei der Abklärung von Simulationsverdacht bei psychischen Erkrankungen eingesetzt werden kann. Der Test ist einfach anwendbar und in wenigen Minuten durchführbar. Allerdings liegt nur die englische Version des Tests vor. In der englischsprachigen Literatur sind auch Studien für den Einsatz des M-FAST in relevanten Bereichen verfügbar (Vitacco et al., 2007). Insgesamt werden dem Test eine ausreichende Validierung und befriedigende Gütekriterien zugemessen (Blaskewitz & Merten, 2007). Aus der konsultierten Literatur lässt sich aber nicht ableiten, inwieweit der Test sich für die Anwendung in anderen Sprachen eignet. Da der Test in der Testzentrale der Schweizer Psychologen beschrieben ist und auch in der Schweiz vertrieben wird, kann angenommen werden, dass eine Verwendung durch Englisch sprechende Fachpersonen möglich ist.

3.6.7 Unterskalen des Minnesota Multiphasic Personality Inventory (MMPI-2)

Das Minnesota Multiphasic Personality Inventory ist ein in den USA weit verbreitetes Instrument zur Beschreibung der Persönlichkeit und zur Erfassung von Beschwerden. Mit seinen 568 Fragen und einer ungefähren Bearbeitungsdauer von 1 bis 1.5 Stunden handelt es sich um ein sehr umfangreiches Verfahren, in welches auch verschiedene Skalen zur Erfassung von suboptimalen Leistungsverhalten eingebaut wurden. Je nach angewandten Skalen werden hierbei unterschiedliche Klassifikationsgüten ausgewiesen, in der Regel liegen aber gute Werte vor. In einer Studie im forensischen Zusammenhang wurden beispielsweise für die Spezifität je nach Skala Werte zwischen 94% und 100% gefunden (Lewis et al., 2002).

Im deutschen Sprachraum wurde dem MMPI-2 bislang eine relativ grosse Skepsis entgegengebracht (Merten, 2002), weshalb hier auch noch ein grosser Mangel an angewandter Forschung festzustellen sei (Blaskewitz & Merten, 2007).

In den USA ist das MMPI-2 für die Abklärung von Malingering in einem breiten Spektrum von Beschwerden eingesetzt worden. So wurde unter anderem auch vorgeschlagen das MMPI-2 für die Entdeckung von Antwortverzerrungen im Zusammenhang mit chronischen Schmerzen einzusetzen (Arbisi & Butcher, 2004).

3.6.8 Weitere mögliche Test zur Beschwerdevalidierung

Aus der breiten Palette von bisher konstruierten Beschwerdevalidierungstests oder Teilskalen von umfassenderen Verfahren sind vorgängig einige häufig erwähnte oder gut untersuchte Verfahren vorgestellt worden. Die dabei vorgenommene Unterscheidung in häufiger erwähnte und seltener genannte oder unwichtigere Instrumente konnte nicht auf der Basis eines strengen Analyserasters vorgenommen werden, die konsultierten Veröffentlichungen gaben dazu keine Informationen her. Letztlich orientieren sich die Forscher und Forscherinnen bei der Auswahl von Verfahren, welche in deutschen Studien untersucht und dazu übersetzt wurden, nicht immer nur an der Bedeutung der Verfahren. So ist es erstaunlich, dass es bislang keine Übersetzung des Structured Interview of Re-

ported Symptoms SIRS gibt. Im englischen Sprachraum wird dieses Verfahren relativ häufig genutzt und ist auch schon als Gold-Standard zur Validierung von anderen Instrumenten genutzt worden.

Der Vollständigkeit halber sollen deshalb nachfolgend in Tabelle 7 noch Instrumente beschrieben werden, die bislang in deutschsprachigen Veröffentlichungen eher selten diskutiert wurden oder dabei als eher ungeeignet bewertet wurden. Auch unter diesen als weniger zentral eingestuften Tests finden sich interessante Verfahren, welche in spezifischen Settings eingesetzt werden könnten oder vielleicht auch in Zukunft mehr Bedeutung erlangen werden.

Tabelle 7 Selten erwähnte Tests zur Beschwerdevalidierung (Forts. nächste Seite)

One-in-five Test	Beim One-in-Five Test werden den Exploranden in drei Blöcken mit 12 Karten jeweils 4 Zahlen gezeigt, die sie sich einprägen sollen. Nach 5, 10 und 15 Sekunden (pro Block jeweils anders) wird eine Karte mit fünf Zahlen genannt, wobei jeweils nur eine Zahl anders ist als die vorhin gezeigten 4 Ziffern. Die Probanden müssen nun eine Zahl nennen, die auf beiden Karten vorhanden war. Mit diesem Verfahren erhöht sich die Wahrscheinlichkeit für zufällig richtige Antworten auf 80%. In einer Analogstudie wird der One-in-Five Test aber einerseits von Testteilnehmenden als Beschwerdevalidierungstest erkannt und andererseits schneidet er in Bezug auf die Klassifikationsgüte auch schlechter ab als andere Tests wie beispielsweise der Amsterdamer Kurzzeitgedächtnis Test (Blaskewitz & Merten, 2006).
Coin-in-the-Hand Test	Beim Coin-in-the-Hand Test müssen Exploranden in 10 Durchgängen angeben, auf welcher Seite der Untersucher eine Münze in der Hand versteckt. Die Münze wird zuerst gezeigt und die Hände nachher auf den Rücken geführt (ev. Hand nur schliessen), nach einer kurzen Ablenkung (Rückwärts von 10 auf 0 abzählen) werden die Exploranden befragt in welcher Hand die Münze gehalten wird. Selbst Patientinnen und Patienten mit Amnesie erreichen 10 von 10 richtige Nennungen. Patienten und Patientinnen mit suboptimalem Leistungsverhalten geben eine unterzufällige Häufigkeit von richtigen Antworten an (Blaskewitz, 2005). Der Test ist sehr anschaulich und einfach durchzuführen, wurde aber bislang noch nicht weiter wissenschaftlich erforscht.
Rey 15-Item Test (FIT)	Beim 15-Item Test von Rey handelt es sich um einen relativ alten und weit verbreiteten Test. Bei der Anwendung müssen 15 Items (Symbole, Zahlen, usw.) in kurzer Zeit gelernt werden und unmittelbar danach, sowie verzögert (10 min, 15 min), aus der Erinnerung gezeichnet werden. Der Test ist wegen einer übersichtlichen Anordnung der Symbole sehr einfach und wird von aggravierenden Personen überdurchschnittlich schlecht ausgeführt (Heubrock, 1995). Der Test erfüllt in keinem einzigen Punkt die Kriterien zur Beurteilung von Beschwerdevalidierungstests von Hartman (2002). Auch zeigen sich in verschiedenen Studien relativ schlechte Werte für die Klassifikation (Blaskewitz & Merten, 2007). Im Rahmen einer Anwendung des 15-Item Tests bei geistig behinderten Personen wurden rund 80% der untersuchten Personen fälschlicherweise als Malingerer eingestuft (Hurley & Deal, 2006).
Trail Making Test	Beim Trail Making Test müssen in zwei Testdurchläufen Kreise mit einem Stift ohne Absetzen in bestimmter Reihenfolge verbunden werden. Im ersten Durchgang werden 25 mit Zahlen von 1-25 beschriftete Kreise vorgelegt und in einem zweiten Durchgang werden 11 Kreise vorgelegt, die mit Zahlen von 1-11 und Buchstaben von A-L versehen sind, wobei hier nach Zahlen- und nach Buchstabenreihenfolge verbunden werden muss. Beurteilt wird das Verhältnis der Testzeiten und die gemachten Fehler (Blaskewitz, 2005). In einer Analogstudie im deutschen Sprachraum zeigt der Test aber mit 22% Sensitivität und 78% Spezifität eine unbefriedigende Klassifikationsgüte und weiter wird eine bislang unzureichende Kreuzvalidierung bemängelt (Blaskewitz & Merten, 2007).
Rey Dot Counting Test (DCT)	Beim Dot Counting Test werden den Exploranden in unregelmässiger Reihenfolge 6 Karten mit unsystematisch verteilten Punkten und 6 Karten mit systematisch verteilten Punkten gezeigt. Gemessen wird die Schnelligkeit, mit welcher die Punkte gezählt werden können. Abweichungen vom Normwert geben Hinweise auf Simulationsversuche, weil es schwierig ist, bewusst die Verzögerung der jeweils realen Schwierigkeit der Zählaufgabe anzupassen (Heubrock, 1995). Der Test zeigt aber unzureichende Sensitivität, von einer alleinigen Anwendung sei abzuraten (Blaskewitz & Merten, 2007). Auch erfüllt der Test nur ein Kriterium nach Hartman (2002).

Tabelle 7 Selten erwähnte Tests zur Beschwerdevalidierung (Forts. von vorangehender Seite)

Auditory Verbal Learning Test	Beim Auditory Verbal Learning Test werden zwei Listen A und B mit jeweils 15 Substantiven sowie eine Wiedererkennungsliste präsentiert. In fünf Durchgängen werden im 1-Sekundentakt die Begriffe vorgelesen um gleich anschliessend so gut wie möglich niedergeschrieben zu werden. Anschliessend wird mit der Liste B in nur einem Durchlauf analog umgegangen. Am Schluss wird die Wiedererkennungsliste (enthält die Worte aus A und B sowie 20 zusätzliche) abgegeben, mit deren Hilfe alle von Liste A noch erinnerten Worte wiedergegeben werden sollen (Schiemann, 2003). Der Test zeigt in Studien Sensitivitäten von 76% und Spezifitäten von 92% und wurde auch kreuzvalidiert (Blaskewitz & Merten, 2007), ist aber bisher im deutschen Sprachraum noch kaum angewandt worden.
Wisconsin Card Sorting Test	Beim Wisconsin Card Sorting Test wird mit einem Set von Karten gearbeitet, die nach verschiedenen Schemata oder Logiken sortiert werden könnten. Die Testpersonen müssen nach einem ihnen unbekannten Schema Karten einem Stapel zuordnen und erhalten dabei Rückmeldung ob sie diese richtig zugeteilt haben. Nach zehn Durchgängen wird das Schema geändert (Blaskewitz, 2005). Der Test zeigt aber insgesamt eine unbefriedigende Klassifikationsgüte und ist noch nicht ausreichend kreuzvalidiert (Blaskewitz & Merten, 2007).
Aufmerksamkeits-Belastungs-Test d2	Der Aufmerksamkeits-Belastungs-Test d2 ist ein im deutschen Sprachraum weit verbreiteter Leistungstest. Das Auftreten von bestimmten Fehlerarten in diesem Test wurde als Indikator für Simulation vorgeschlagen und im Rahmen einer Studie überprüft (Schmidt-Atzert, Bühner, Rischen et al., 2004). Im Rahmen einer späteren Validierungsstudie liess sich dieses erste positive Urteil aber nicht bestätigen, die Forscherinnen und Forscher raten deshalb von einer Nutzung des d2-Tests als Verfahren für die Bestimmung von suboptimalem Leistungsverhalten ab .(Merten, Blaskewitz, & Stevens, 2007).
Rey Complex Figure Test	Beim Complex Figure Test nach Rey muss von den Exploranden eine geometrischen Figur abgezeichnet und anschliessend (nach 3 und nach 30 Minuten) aus dem Gedächtnis erneut gezeichnet werden. Anschliessend muss die ursprüngliche Figur aus einer Menge von 24 Stimuli wieder erkannt werden (Blaskewitz, 2005). In einer eher kleinen Analogstudie mit gesunden Testpersonen im deutschen Sprachraum konnten als Klassifikationswerte eine Sensitivität von 75% und eine Spezifität von 92% ermittelt werden (Blaskewitz & Merten, 2006). Die internationale Forschungsgrundlage zu diesem Test ist zurzeit noch ungenügend.
Offenheitsskala des Freiburger Persönlichkeits- Inventars	Das Freiburger Persönlichkeitsinventar wird im deutschen Sprachraum relativ häufig für die Erfassung der Persönlichkeit und der Beschwerden von Personen angewandt. Erfasst werden Bereiche wie: Lebenszufriedenheit, Soziale Orientierung, Leistungsorientierung, Gehemmtheit, Erregbarkeit, Aggressivität, Beanspruchung, Körperliche Beschwerden, Gesundheitssorgen, Offenheit sowie zwei Sekundärskalen Extraversion und Emotionalität. Im Handbuch wird für das Eintreten von Werten unterhalb einer gewissen Norm in der Offenheitsskala vorgeschlagen, weitere Informationen über die Leistungsbereitschaft der Testpersonen einzuholen (Merten, Friedel et al., 2007). In einer retrospektiven Analyse von Daten eines Gutachteninstitutes konnte aber gezeigt werden, dass sich die Offenheitsskala des Freiburger Persönlichkeitsinventars nicht für die Erfassung einer unzureichenden Testmotivation eignet (ebd.).
Weitere Tests, die allenfalls in deutscher Sprache angewandt werden könnten Im Rahmen der Konsultation von Artikeln und Studien, welche mit Bezug auf den deutschen Sprachraum veröffentlicht wurden, kamen noch einige weitere Tests zur Sprache, welche im folgenden aufgezählt, aber nicht weiter beschrieben werden. Diese Tests wurden bislang nur in Einzelfällen eingesetzt. Es handelt sich dabei um den Judgment of Line Orientation Test (Blaskewitz & Merten, 2006), den Reliable Digit Span Test aus der Wechsler Adult Inteligence Scale (Merten, Bossink et al., 2007) sowie drei weitere im Rahmen der Entwicklung einer Testbatterie vorgeschlagene, aber nicht mit weiteren Studien untersuchte, Verfahren (Farbtafeln-Test, Distractions-Test, Dominiotest) (Schiemann, 2003).	

3.7 Spezielle Einsatzgebiete: chronische Schmerzen

Obwohl chronische Schmerzen epidemiologisch betrachtet und in der gutachterlichen Praxis eine grosse Bedeutung haben, kann für diesen Bereich der Einsatz von Beschwerdevalidierungstests nicht ausreichend diskutiert werden. Einerseits handelt es sich um ein, auch aus medizinischer Sicht, sehr komplexes Thema und andererseits sind auch die Forschungsaktivitäten zu Aggravation und Simulation in diesem Bereich eher punktueller Natur. In der Review von Blaskewitz und Merten (2007) werden diesbezüglich 11 Studien aufgelistet, welche von 6 unterschiedlichen Autorengruppen stammen. Als mögliche Tests werden in diesen 11 Studien einerseits die bereits beschriebenen MMPI-2 und der Test of Memory Malingering genannt und andererseits wird auf mehr als 10 weitere mögliche Verfahren hinweisen, welche von Videoanalysen über Unterskalen von Intelligenztests bis hin zum Pain Disability Index reichen.

Aufgrund der Komplexität der Situation und Zusammenhänge bei Schmerzen kommen Blaskewitz und Merten zu folgendem Schluss:

> „Zum gegenwärtigen Zeitpunkt bleibt festzustellen, dass sich eine Unterscheidung zwischen tatsächlichen und vorgetäuschten Schmerzen oft schwierig gestaltet und bislang keine Methode und kein Messinstrument angegeben werden kann, das eine sichere Aussage hierzu ermöglicht“ (Blaskewitz & Merten, 2007, S. 149).

Von besonderer Bedeutung für die Untersuchung von Aggravation und Simulation im Zusammenhang mit Schmerzen sind die Leitlinien von Bianchini et al. (Bianchini et al., 2005), welche einen ausführlichen Kriterienkatalog für die Entscheidungsfindung beim Verdacht auf Vortäuschen von Schmerzen entwickelt haben (siehe Seite 26).

3.8 Diskussion der Erkenntnisse aus der Literaturstudie

Bislang sind hier die meisten Tests zur Beschwerdevalidierung in Bezug auf ihre wissenschaftliche Untermauerung und ihre Eignung für den praktischen Einsatz eher kritisch beurteilt worden. Zusammenfassend lassen sich gleichwohl einige relativ gut erforschte Verfahren zur Beschwerdevalidierung hervorheben. International als gut bis sehr gut untersucht gelten der Word Memory Test (WMT) und der Test of Memory Malingering (TOMM) für den Einsatz bei kognitiven Beschwerden[10]. Im Einsatzbereich von psychischen und psychiatrischen Symptomen bietet sich der Strukturierte Fragebogen Simulierter Symptome (SFSS) an, wobei dieser Test aufgrund seiner hohen Zahl von falsch-positiven Ergebnissen höchstens als Screeninginstrument eingesetzt werden darf und bei positiven Ergebnissen mit weiteren spezifischeren Test ergänzt werden muss (ev. mit dem MMPI-2). Für den Einsatz bei kognitiven Beschwerden, welche sich vor allem in Gedächtnis- und Konzentrationsstörungen äussern, bietet sich für die Anwendung in Deutsch, Französisch oder Italienisch der Amsterdamer Kurzzeitgedächtnistest (AKGT) an. Hier ist einschränkend festzuhalten, dass die Datenbasis zu diesem Test noch etwas schmal ist und weitere Untersuchungen in relevanten Einsatz-

[10] Einschränkend muss gleichwohl darauf hingewiesen werden, dass der Forschungsstand zu den deutschen Versionen dieser Instrumente zur Zeit noch gering ist. Es gibt aber in diesen Studien bislang keine Hinweise dafür, dass der WMT oder der TOMM in ihren deutschen Versionen von den englischen Ausgaben abweichende Testgütekriterien hätten.

gebieten wünschbar wären. Aufgrund der Breite von erfassten Beschwerdenbereichen kommt allenfalls auch die Testbatterie zur forensischen Neuropsychologie in Betracht. Leider ist für dieses Verfahren die wissenschaftliche Basis aber noch eher schmal. Bei weiteren wichtigen im deutschen Sprachraum anwendbaren Instrumenten zeigt sich ähnlich wie beim AKGT und dem SFSS ein uneinheitliches Bild. Häufig ist die wissenschaftliche Untermauerung (gerade auch zu den deutschen Versionen der Verfahren) noch knapp.

Diese Befunde schliessen aber nicht aus, dass bei spezifischer Anwendung und einer umsichtigen Interpretation der Ergebnisse, je nach konkreter Fragestellung im Begutachtungsprozess mit Hilfe von standardisierten Tests zusätzliche Erkenntnisse gewonnen werden können. Eine umsichtige Interpretation bedeutet vor allem, dass Testergebnisse nie alleine für die Diagnose von Simulation verwendet werden dürfen. Diese Tatsache wird in verschiedenen Veröffentlichungen von unterschiedlichen Autorinnen und Autoren hervorgehoben. Exemplarisch kann dazu Thomas Merten, der für den deutschen Sprachraum führende Forscher zum Thema Beschwerdevalidierungstest, zitiert werden:

> „Allerdings muss betont werden, dass die gutachterliche Beurteilung keineswegs auf einem einzigen derartigen Testergebnis beruhen darf. Erst in der Gesamtschau aller verfügbaren Information und in besonderer Würdigung des selbst erhobenen psychischen Befundes kann eine solche Beurteilung getroffen werden. Jedes diagnostische Verfahren und damit auch jedes Instrument zur Beschwerdevalidierung weist eine Fehlerrate auf; und es muss mit falsch-negativen wie falsch-positiven Klassifikationsergebnissen gerechnet werden. Allein ein multimethodaler Ansatz zur Beurteilung der Beschwerdenvalidität ist geeignet, solche Fehlklassifizierungen zu minimieren.“ (Merten et al., 2007, S. 150)

In Anbetracht dieser Einschränkungen beim praktischen Einsatz von Beschwerdevalidierungstests stellt sich weiterhin die Frage, ob und wie in der Praxis entschieden werden kann, inwiefern es sich bei vorgetragenen Beschwerden um reale oder simulierte Symptome handelt. Hier könnten allenfalls Leitlinien, wie sie auf Seite 25-28 beschrieben werden, zum Einsatz kommen. Dohrenbusch (2007) diskutiert einige dieser Leitlinien als Kriteriologien zur Kennzeichnung von Aggravations- und Simulationstendenzen. Dabei kommt er aber zum Schluss, dass keine der vorgeschlagenen Kriteriologien empirisch durch hinreichend kontrollierte und valide Studien überprüft worden sei und es sich dabei „überwiegend um Heuristiken“ (ebd. S. 230) handeln würde. Die beschriebenen Kriterienlisten setzen aufgrund der verschiedenen fachwissenschaftlichen Hintergründe der Autorinnen und Autoren unterschiedliche methodische und inhaltliche Schwerpunkte. Dohrenbusch folgert aus den verschiedenen Ansätzen, dass es grundsätzlich wesentlich sei, Testergebnisse und Beschwerdeäusserungen immer im Untersuchungskontext und im Vergleich zu geeigneten Vergleichsnormen zu sehen. Deshalb müssen sich die Entscheidungen zur Wahrscheinlichkeit von Verfälschungstendenzen jeweils auf eine Vielzahl von geeigneten Methoden und Verfahren stützen.

Als weitere Einschränkung muss erwähnt werden, dass in den meisten Studien kaum auf potentielle Schwierigkeiten der Tests in der praktischen Anwendung eingegangen wird. Einerseits werden die Verfahren mehrheitlich in sogenannten Analogstudien mit gesunden Personen, welche zur Simulation instruiert wurden, durchgeführt und andererseits finden Untersuchungen mit Patientinnen und Patienten meist nur in einzelnen Bereichen statt; der Frage nach der Übertragbarkeit von Studienresultaten auf andere Settings wird meist nicht nachgegangen. Für den konkreten Einsatz eines Tests

muss deshalb im Einzelnen hinterfragt werden, ob der Test sich jeweils konkret für die untersuchten Beschwerdenbereiche eignet.

Trotz der wachsenden wissenschaftlichen Beschäftigung mit Instrumenten und Verfahren zur Validierung von Beschwerden zeigen die bisher entwickelten Tests weiterhin gewisse nicht zu vernachlässigende Fehlerraten und können zudem in der Regel nicht zwischen aritifizieller Störung und Simulation unterscheiden (Merten, 2002). Diese Einschränkungen bei der Interpretation von Testresultaten können bislang auch nicht durch den Einsatz von weiterführenden Kriterienlisten behoben werden.

Mertens Feststellung: *„Als einziges wirklich zuverlässiges Kriterium für eine Simulation muss das Eingeständnis des Betreffenden, in der Tat zu simulieren oder simuliert zuhaben, gelten“* (Merten, 2002), behält also weiterhin Gültigkeit.

3.9 Fazit aus der Literaturstudie

Als abschliessende Erkenntnis aus der Literaturstudie lässt sich somit festhalten, dass es mittlerweile ein sehr breites Spektrum von unterschiedlichen Tests zu unterschiedlichen Einsatzbereichen für die Abklärung von vermuteter Aggravation oder Simulation gibt. Trotz der wachsenden Forschungstätigkeit bleibt aber das Problem bestehen, dass all diese Tests, auch die wissenschaftlich gut untersuchten, jeweils spezifische Einschränkungen in Bezug auf ihre Einsatzgebiete haben und alle eine nicht zu vernachlässigende Rate von falsch-positiven Ergebnissen liefern. Gleichwohl können bei einem gezielten Einsatz und bei umsichtiger Interpretation von Testergebnissen mit diesen Verfahren zusätzliche Erkenntnisse gewonnen werden.

Wie Studien zur Einschätzung des Ausmasses an Simulation und Aggravation in verschiedenen Ländern und Kontexten zeigen (Merten et al., 2006; Mittenberg, Aguila-Puentes et al., 2002), muss im Rahmen von Begutachtungen durchaus mit einer „qualifizierten Minderheit“ (Blaskewitz & Merten, 2007) von Personen mit verzerrendem Antwortverhalten gerechnet werden. Weiter kann nicht davon ausgegangen werden, dass eine Simulation von Beschwerden von Expertinnen und Experten alleine aufgrund ihrer Erfahrung erkannt wird. Merten (2005) zitiert diesbezüglich Studien, die aufzeigen wie bis zu 50% von Personen, welche Kopfverletzungsfolgen simulieren, von Experten nicht als Simulantinnen und Simulanten erkannt wurden. Weder mittels standardisierter Tests noch alleine aufgrund eines Fachurteils kann demnach eine absolut gesicherte Feststellung darüber gemacht werden, ob in einem konkreten Fall Symptome simuliert werden.

Im Sinne der Feststellung, dass bei einem so komplexen Phänomen wie Aggravation und Simulation im Rahmen von Begutachtungen auch mit mehrschichtigen und multimethodalen Ansätzen gearbeitet werden sollte (Merten et al., 2007), können auch aus der Sicht dieser Literaturstudie standardisierte Beschwerdevalidierungstests zusätzliche Erkenntnisse liefern, wenn sie denn fachgerecht angewandt und interpretiert werden. Wie und in welchem Masse Beschwerdevalidierungstests im Alltag der gutachterlichen Tätigkeit einen Platz finden können, kann aber aufgrund der konsultierten Veröffentlichungen nicht abgeleitet werden, da diesbezüglich keine Aussagen gemacht werden.

4 Interviews mit Experten und Gutachtenden

4.1 Methoden

Es wurden leitfadengestützte Experteninterviews (Bogner, 2005) durchgeführt zur übergeordneten Frage des Einsatzes von BVT in der Abklärungspraxis. Im Rahmen der Interviews sollte untersucht werden, ob und wenn ja, welche BVT in den entsprechenden Settings routinemässig eingesetzt werden. Weitere Themen waren: Operationalisierung und Anwendbarkeit der BVT in der Begutachtung, Relevanz der Befunde, Umgang mit widersprüchlichen Resultaten, Konsequenzen positiver Testbefunde für die Exploranden. Ausserdem wurden in den Interviews potentielle Vorbehalte − wie ethisch begründete Bedenken gegenüber der Anwendung von BVT − sowie andere Barrieren zur Anwendung von BVT thematisiert.

Die Auswertung der transkribierten Experteninterviews erfolgte nach dem Ansatz der qualitativen Inhaltsanalyse nach Mayring (Mayring, 2003; Mayring & Gläser-Zikuda, 2005). Dieses Verfahren umfasst drei zentrale Arbeitsschritte: (1) Zusammenfassung: Reduktion des Ausgangstextes auf eine überschaubare Kurzversion der wichtigsten Inhalte; (2) Explikation: Klärung unklarer Textbestandteile; (3) Strukturierung: Entwicklung und schrittweise Verfeinerung eines inhaltlichen Kategorienschemas und Zuordnung der Textinhalte. Als Software wurde ATLAS TI eingesetzt.

Potentielle Teilnehmer und Teilnehmerinnen für die Interviews waren Gutachtende im Bereich der Rentenversicherung, insbesondere der MEDAS, Versicherungsmediziner der RAD, sowie Fachleute auf dem Gebiet der Aggravation-/Simulation-Forschung. Die Teilnehmenden wurden über den Zweck der Interviews informiert und unterschrieben eine schriftliche Einverständniserklärung für die Teilnahme an den Interviews.

4.2 Resultate

Interviewt wurden 13 Personen: 11 Gutachtende, vorwiegend tätig in Medizinischen Abklärungsstellen (MEDAS), und 2 Experten, Psychologen die im Bereich der Begutachtung und Überprüfung der BV mehrere Publikationen veröffentlichten. Von den interviewten Gutachtenden arbeiten 5 in der Westschweiz und 6 in der Deutschschweiz. Es handelt sich dabei um Psychiater, Rheumatologen, Allgemeinmediziner und Psychologen. Die Resultate werden in folgenden 9 Themenfeldern (Tabelle 8) beschrieben.

Tabelle 8 Themenfelder der Interviews mit Experten und Gutachtenden

1. Arbeitsfeld des Interviewpartners
2. Das Umfeld der Begutachtung
3. Diagnosen der Exploranden
4. Inkonsistenz, Simulation, Aggravation, Verdeutlichung
5. Beschwerdevalidierungstests a) Bedenken b) Konsistenzprüfung und Beschwerdevalidität c) Zustimmung d) Ablehnung von einheitlichen Vorgaben e) Nicht-validierte BVT f) Validierte kognitive BVT g) Validierte somatische BVT
6. Barrieren
7. Wissenschaftliche Literatur und Forschungsbedarf
8. System und Strukturen
9. Entwicklungen und Erwartungen

4.2.1 Arbeitsfeld

Die interviewten Gutachtenden arbeiten mehrheitlich in einer multidisziplinären Umgebung, mit Einbezug von Konsiliarärzten und -ärztinnen. Die Interviewten sind entweder nur als Gutachtende tätig, oder sie üben sowohl eine klinische als auch eine gutachterliche Tätigkeit aus. Die Interviewten haben hinsichtlich ihrer beruflichen Spezialisierung verschiedene Hintergründe: Rheumatologie, Psychiatrie, Rehabilitation und Psychologie. Die Gutachtenden sind vielfach in den Bereichen Forensik und Versicherungsmedizin tätig.

Nur selten sind Gutachten monodisziplinär ausgerichtet, meistens jedoch bi- oder polydisziplinär (ein Psychiater oder eine Psychiaterin und ein zweiter Gutachter oder eine zweite Gutachterin); eher selten − bei klaren Fällen − wurde monodisziplinär gearbeitet. Die Gutachtertätigkeit bezog sich auf Hauptgutachten und Teilgutachten. Die Gutachtenden waren bei 80-1000 Gutachten pro Jahr involviert.

Auftraggeberin für Gutachten bei den MEDAS ist mehrheitlich die IV, dann folgen private Versicherungen und Gerichte. Der Aufwand für ein Gutachten wurde auf 8-14 Stunden geschätzt, abhängig von der Komplexität des Falles. Wenn Anwälte oder Anwältinnen involviert sind, sei mit erheblichem Mehraufwand zu rechnen. Ebenfalls, wenn schon mehrere andere Gutachtende involviert seien.

Einige Interviewte erhalten ein Salär unabhängig von der Anzahl Gutachten, im Gegensatz zu der Mehrheit von vollberuflichen Gutachtenden, die keiner klinischen Tätigkeit mehr nachgingen. Nebenberuflich Gutachtende weisen darauf hin, dass die Auftraggeber Druck auf die Gutachter ausüben können. Vollberuflich Gutachtende nennen eine potentiell ungenügende Professionalität der nebenberuflichen Gutachtenden, deren Haupttätigkeit klinisch therapeutisch ist.

4.2.2 Das Umfeld der Begutachtung

Für die meisten Interviewten stellt sich die Fragestellung der Berentung bei praktisch 100% der Gutachten. Es wurde darauf hingewiesen, dass die meisten Exploranden eine Berentung erwarten und wünschen, auch wenn sie vordergründig angeben an einer Wiedereingliederung interessiert zu sein.

Die Befunde, die nach einer Begutachtung vorliegen, reichen oft nicht aus für eine fundierte Bestimmung der Arbeitsfähigkeit. Die Quantifizierung der Arbeitsfähigkeit anhand der Synthese innerhalb des Gutachtens wird deshalb als eine Art „Quantensprung“ beschrieben mit einer inhärenten Unbestimmtheit:

> On pourrait appeler ceci le «saut quantique» entre les données médicales et la décision d'incapacité de travail. Je pense qu'il y a un gap qu'on est loin d'avoir comblé. Il y a encore un grand part d'arbitraire.

Eine fundierte Beurteilung der Arbeitsfähigkeit (AF) ohne Unbestimmtheit (gewisse Unsicherheiten sind vorhanden) wird von einigen Interviewten als nur scheinbar präzise und eigentlich unseriös betrachtet. Die Unbestimmtheit in der zu definierenden AF von etwa 70% könnte mit einem Intervall als AF von 70 ± 15% dargestellt werden. Bei der Angabe von Intervallen besteht jedoch nach Aussagen einiger Befragten eine gewisse Gefahr, dass von Seiten der Versicherer zum Teil die obere Grenze "bevorzugt" wird.

Eine Grundproblematik für Gutachtende ist die notwendige Abstraktion von sozialen Faktoren. Das biopsychosoziale Modell, mit dem der klinisch tätige Arzt oder die Ärztin arbeitet, ist für die Gutachtenden bei Fragen der Berentung nicht mehr anwendbar. Gesetze und Rechtssprechung abstrahieren die sozialen Aspekte, so dass diese in der arbeitsbezogenen Begutachtung nicht behinderungsrelevant sind. Daher werden die Einschätzungen der Behinderung im Rahmen der Behandlung ganz anders sein als im Rahmen einer Begutachtung. Folgende Aussagen illustrieren diese Thematik.

> Dort [bei der Abstraktion von sozialen Faktoren in der Begutachtung] überfordert das Recht die Medizin. Wir können die psychosozialen Faktoren nicht immer ausblenden…
>
> Wenn man uns fragt: Sind es überwiegend psychosoziale Krankheitsfaktoren, dann kann man mit einer akzeptablen Treffsicherheit sagen, ja. Wenn man aber von uns fordert, wir sollen sagen, wie viele Prozent psychosoziale Faktoren es sind, dann ist dies nicht seriös zu beantworten.

Diese letzte Aussage erschwert auch die Bestimmung der Arbeitsfähigkeit, wie bereits vorher angesprochen wurde.

4.2.3 Diagnosen der Exploranden

Folgende Diagnosen werden als relevant für die Beschwerdenvalidierung (BV) und die Inkonsistenzproblematik beschrieben: chronische Schmerzen, Depressionen, somatoforme Störungen

und unspezifische Probleme am Bewegungsapparat, Weichteilrheumatismus, Fibromyalgie[11], unspezifische Rücken- und HWS Probleme, Schleudertrauma, Psychosen und Schizophrenie. Es gibt laut den Interviewten regionale Unterschiede: so scheint die Diagnose Schleudertrauma in der Deutschschweiz und Fibromyalgie in der Westschweiz häufiger vor zu kommen.

Es wird betont, dass die Probleme vieler Exploranden im Grenzbereich zwischen Psyche und Köper liegen. Bei etwa 50% der Exploranden liegt eine psychische Hauptdiagnose vor. Psychiater und Psychiaterinnen sind denn auch sehr häufig als Experten in mono- oder poly- resp. interdisziplinären Gutachten tätig. Die Häufigkeit von psychischen Diagnosen als Nebendiagnosen wurde mit 70-80% angegeben.

Verschiedene Interviewte wieien auf eine mögliche Unerfahrenheit von erstbehandelnden Ärzten und Ärztinnen und anderen Gutachtenden hin, die oft zu falschen Diagnosen, vor allem bezüglich somatoformen Störungen und HWS Schleudertrauma führe:

> Die somatoforme Störung ist nach alter Auffassung eine Konversionsstörung. Hier wird ein seelischer Schmerz, der als solcher nicht erlebt und ausgesprochen wird und bewusst da sein darf, von der Theatergruppe der Seele als Körperschmerz aufgeführt… Da muss die psychiatrische Anamnese einwandfrei sein. Dann sagen wir auch, im Gegensatz zur Rechtssprechung, natürlich gibt es somatoforme Störungen, die ausgesprochen stark und beeinträchtigend und invalidisierend sind. Die Patienten und Patientinnen können nicht erfolgreich behandelt werden vom Psychologen von der Psychologin, dafür haben sie ja den Konversionsmechanismus.
>
> Diskrepanzen haben wir vor allem bei HWS Schleudertrauma. Nach einem Unfall sagen die Orthopäden: er hat eine Blockade oder nicht, vielleicht hilft manuelle Therapie. Der Neurologe macht Messungen und wenn sie nichts finden, sagen sie: ich sehe nichts. HWS Patienten und Patientinnen sind nie psychisch krank, haben keine Depressionen und die Schmerzen sind nicht psychisch bedingt, aber sie haben hirnorganische Defizite, Konzentrationsstörungen und so weiter.

Einige Interviewte betonten hingegen, dass bezüglich der Problematik von Simulation und Aggravation nicht die „relevanten" Diagnosen für sie wichtig seien, sondern eher der kulturelle Hintergrund der Exploranden. Dieses Problem – die externe Validität von BVT – wird in späteren Kapiteln erörtert.

4.2.4 Inkonsistenz, Verdeutlichung, Aggravation, Simulation

Die Gutachtenden verwenden die Begriffe Kohärenz und Konsistenz als Synonyme. Die Kohärenz- oder die Konsistenzproblematik sei praktisch bei jedem Gutachten ein Thema:

> Il y a pratiquement toujours la problématique de l'hiatus entre l'importance des douleurs et puis les signes cliniques objectifs, donc la problématique du trouble somatoforme douloureux est toujours en filigrane dans 90% des cas.

In einem Gutachten ist der Experte, die Expertin in nicht geringem Masse beteiligt am gemessenen Zustand des Exploranden bezüglich – unten näher definierten – Verdeutlichungstendenzen, Aggravation und Simulation. In Analogie zur Messproblematik in der modernen Physik sei vollständige Objektivität eine Illusion und Aggravation zum Teil ein Resultat der ganzen „Messvorrichtung" während einem Gutachten. Ein Interviewter formulierte dies folgendermassen:

> Aggravation ist zum Teil das Artefakt einer misslungenen Begegnung in der Begutachtung.

[11] Ist laut Bundesgerichtsentscheid wie chronischer Schmerz zu behandeln

Der Begriff der englischsprachigen Literatur – Malingering – wird in der Westschweiz eher weniger angewandt. Zudem wird die angelsächsische Literatur zum Thema in der Romandie zwar als interessant, aber bezüglich den grundlegenden Konzepten und Ideen eher kritischer betrachtet als in der Deutschschweiz. Als Erklärung dafür, dass ein grosser Teil der Publikationen im Bereich der Simulationsforschung amerikanischer und europäischer Herkunft ist, gab es aber auch die Meinung, dass das Thema in der Schweiz immer noch tabu sei:

> Pendant un cours sur la simulation aux Etats-Unis, le professeur demande: Qui a déclaré plus de vol qu'il a eu pour être remboursé? Tout l'amphithéâtre dit oui. Conclusion: tout le monde est simulateur… En Suisse, même expérience. Personne ne répond par l'affirmative. Donc, c'est un tabou.

Innerhalb von Gutachten vorkommende Verdeutlichungstendenzen − also keine eigentliche Aggravation − werden allgemein als normal eingestuft. Es gibt ein normales Mass an Verdeutlichungstendenz, die situationsimmanent ist:

> Verdeutlichungen haben oft etwas mit Normalpsychologie zu tun und nicht mit Psychopathologie. Unterschiedliches Verhalten, unterschiedliche Aussagen in unterschiedlichen Untersuchungssituationen sind zuerst normal, was auch damit zu tun hat, dass die Aussage und das Verhalten vom Exploranden immer auch eine Reaktion ist auf den Gutachtenden und wie der sich verhält. Man muss also die Objektivität einschränken.

Die Interviewten gehen von einer sehr geringen a priori Wahrscheinlichkeit von Simulation aus und schätzen die Prävalenz auf weniger als 1% bis maximal 10%. Die zum Teil weitaus höheren Prävalenzen in der Malingering-Forschung werden auf eine andere Definition der Schwelle zwischen Verdeutlichungstendenzen, Aggravation und Simulation zurückgeführt, auf ein anderes Versicherungsumfeld und schlussendlich auch auf ein anderes Menschenbild:

> C'est l'abstraction des problèmes sociaux.
>
> Les situations ou le patient cherche à te tromper. Je pense que ce n'est pas ce qu'on voit en expertise.

Die Diagnostik von Simulation ist für den durchschnittlichen Gutachtenden kein relevantes Thema, eine gewisse Quantifizierung von Konsistenz jedoch schon.

Das Mass der Bewusstheit ist gering bei Verdeutlichungstendenzen, grösser bei Aggravation und sehr gross bei Simulation. Ein grosses Problem bei der Unterscheidung von Verdeutlichungstendenzen, Aggravation und Simulation stellt die Beurteilung des Bewusstseinsgrades der verstärkten Beschwerdepräsentation dar. Gutachtende finden die Frage nach der Bewusstseinsnähe relevant, wenn auch fast nicht zu beantworten, wie folgende Aussagen illustrieren:

> [Die Unterscheidung zwischen Verdeutlichung und Aggravation ist…] eine formale Unterscheidung, die üblich ist und mehr oder weniger in der Rechtsprechung mit eingegangen ist. Das Ganze steht auf ganz dünnem Eis, da dem Praktiker zuverlässige Kriterien fehlen wie er eine situationsangemessene Verdeutlichung von einer nicht mehr der Situation angemessenen Aggravation unterscheiden kann. Da ist es wirklich die Frage, ob man diese Unterscheidung braucht, wenn man sie nicht klar operationalisieren kann. […] Andererseits scheint es noch schwieriger, Begrifflichkeiten für Glaubhaftigkeit ganz streichen zu wollen. Wenn wir nur noch Beeinträchtigung und Leistungsfähigkeit nehmen, und ganz auf diese Begriffe Verdeutlichung und Aggravation verzichten, dann verliert man an Information. […] Es macht also trotzdem Sinn zu fragen wie bewusst, wie bewusstseinsnah die Beschwerden präsentiert werden.

Da die Gutachtenden das Ausmass der Bewusstheit nur mit ungenügender Sicherheit bestimmen können, werden die Formulierungen mit entsprechender Vorsicht gewählt. Und es wird nicht von „bewusst" sondern von „bewusstseinsnah" gesprochen. Ein anderer Gutachter beschreibt Verdeutlichungstendenzen als „beschwerdeorientierte Schilderungsweise", Aggravation als eine „bewusstseinsnahe Verhaltensweise" und schliesslich Simulation als ein „bewusstseinsnahes Ausdrucksverhalten". Auch die Experten sind sich der Problematik bewusst. Ein Experte schlägt vor, nicht von

„bewusst“ oder „bewusstseinsnah“ zu sprechen, sondern von „wahrscheinlichen Entlastungsmotiven“. Auch dieser Ausdruck berücksichtigt die Unsicherheit in der Bestimmung der Bewusstheit der verstärkten Beschwerdepräsentation.

Es wurde die Frage nach der nötigen Schwelle und den Kriterien erörtert, bezüglich denen Simulation definiert wird. Für nicht wenige Gutachtende ist diese Diskussion müssig, da sie davon ausgehen, dass das Problem der Simulation kein eigentliches sei. Aggravation hingegen komme viel häufiger vor als Simulation, sei aber nicht immer leicht zu unterscheiden von akzeptierten Verdeutlichungstendenzen:

> C’est une question de critère. Il y a une définition dans le DCM-4 en psychiatrie. Mais pour moi, je n’utilise pas de définition opérationnelle, c’est une question de feeling et – surtout – le consensus avec des collègues dans nos expertises pluridisciplinaires. Cela nous permet de recouper nos opinions et on va parler plutôt de la question de l’authenticité.

4.2.5 Beschwerdevalidierungstests

Bedenken gegenüber der BVT-Forschung

Die Interviewten schätzen die Prävalenz von Simulation in der Schweiz viel tiefer ein als in der – zum Teil von der Forensik beeinflussten – amerikanischen Malingering-Literatur. Ein Teil der Malingeringforschung stammt aus dem Bereich der Forensik. In der Forensik sind die Krankheitsbilder, im Vergleich zur arbeitsbezogenen Begutachtung, verschieden. Sucht, Schizophrenie, Persönlichkeitsstörungen und Psychosen kommen in der Forensik häufig vor. Chronischen Schmerzen, Fibromyalgie und somatoforme Schmerzstörungen sind dagegen fast nur in der beruflichen Begutachtung anzutreffen. Die Ausgangslage ist in der Forensik im Vergleich zur arbeitsbezogenen Begutachtung anders, weil die Tat feststeht und oft schon lange zurück liegt. Die wissenschaftliche Validität der BVT aus der Forensik für die berufsbezogene Begutachtung wird bezweifelt. Bei der Entwicklung von BVT im Bereich der Malingeringforschung wird der fehlende Goldstandard für Simulation kritisiert.

Beschwerdevalidität und Konsistenz

Viele Gutachtende unterscheiden zwischen BVT zur Erfassung von Simulation oder Aggravation und einer Inkonsistenzüberprüfung. Eine Inkonsistenzprüfung wird als Mass für die Beschwerde-Reliabilität angesehen. Die eigentliche Beschwerdevalidierung hingegen als ein Validitätsmass.

Zustimmung für nicht validierte, nicht standardisierte BVT

Breite Zustimmung finden die Beurteilung der BV bei Exploranden mit nicht validierten, individuellen Verfahren während der Expertise und der interdisziplinäre Konsens bezüglich Konsistenz der Exploranden. Wenn auch standardisierte BVT angewendet werden, betonten die Gutachtenden die Bedeutung der individuellen Bewertung der Ergebnisse durch den Kliniker. BVT werden nicht als selbständige Tests, sondern als mögliche Mosaiksteine im Gutachten betrachtet. In der Schlussbeurteilung von Gutachten werden alle Ergebnisse gesamthaft beurteilt und die Konsistenz beurteilt:

> Je ne pense pas qu’il y a un test de la validité dans ce sens. Je pense que cela tient à l’étude comparative qu’on fait à partir de tous les éléments du dossier médical, de l’observation et de ce que la personne raconte…
>
> Beaucoup de tests psychologiques sont des tests d’autoévaluation qui n’ont en expertise pas énormément de valeur. On les intègre par rapport à notre propre réflexion, jamais isolément.

Ablehnung von Vorgaben für eine einheitliche Verwendung von BVT

Die Interviewten standen standardisierten Tests oder Testbatterien, die spezifisch und mechanisch darauf abzielen, Simulation und Aggravation zu diagnostizieren, mehrheitlich ablehnend gegenüber. Die Ablehnung war zum Teil sehr stark. Viele Gutachtende meinen, dass es keine validen Tests oder Testkombinationen – im Sinne von eigentlicher Beschwerdevalidität – gibt oder geben kann.

Wenn man die Beurteilung der BV mit festgelegten BVT verlange, würde damit die Tür für andere Probleme geöffnet. Eine einheitliche Beurteilung der BV mit standardisierten BVT würde den unterschiedlichen Leistungsdefiziten und Inkonsistenzmustern der Exploranden nicht gerecht werden. Die Forschung im Bereich der BV entwickelt sich momentan schnell und Empfehlungen wären deshalb schnell veraltet. Die Vorgaben für die Beurteilung der BV müssten von Jahr zu Jahr angepasst werden.

BVT fehlen in vielen Bereichen. BVT sind vor allem für einzelne psychokognitive Bereiche verfügbar. Für die Überprüfung der körperlichen Befunde sind kaum BVT vorhanden. Ursprünglich wurden BVT für sensorische und motorische Störungen entwickelt. Die Entwicklung der BVT wurde in den letzten 20 Jahren unter anderem von Neuropsychologen vorangetrieben. Rheumatologen und andere Experten und Expertinnen im Bereich chronische Schmerzen haben sich weniger mit diesen Fragen beschäftigt.

Bezüglich der bereits oben erwähnten Messproblematik oder den Artefakten, die während einem Gutachten auftreten, wurde mehrfach betont, dass auch das Resultat eines BVT durch die eigentliche Messvorrichtung beeinflusst wird:

> Donc, dans le fond, j'ai très peu recours à ces tests spécifiques de la mémoire, parce que tous les tests – et je suis sûre que ça va être la même chose pour d'autres tests – le résultat est conditionné par l'attente du patient.

Weiterführend in Analogie zur bereits erwähnten Messproblematik in der Quantenmechanik, ist die Forderung nach einer totalen Objektivität eine Illusion. So hat das Bundesgericht in einem Urteil auf die „Heisenberg'sche Unschärferelation" Rücksicht genommen, wenn es darin entschieden hat, dass sich ein Versicherter bei der medizinischen Begutachtung nicht von seinem Rechtsvertreter begleiten lassen darf:

> Bei letzterem geht es darum, dem medizinischen Begutachtenden eine möglichst objektive Beurteilung zu ermöglichen, weshalb diejenigen Rahmenbedingungen zu schaffen sind, die sich aus wissenschaftlicher Sicht am ehesten dazu eignen, eine solche Beurteilung zu ermöglichen. Die Anwesenheit eines Rechtsvertreters, einer Rechtsvertreterin bei der medizinischen Befragung und Untersuchung führt zu einer Veränderung der Rahmenbedingungen. Zudem müsste der Gutachtende damit rechnen, dass auch die Gegenpartei ihre Rechtsvertretung an die Begutachtung schickt und die Expertise im Kampfgebiet zwischen zwei Rechtsvertretenden angefertigt werden müsste, mit unberechenbaren Folgen für das Untersuchungsergebnis.

Bezüglich der Konsistenz wurde auch die Stereotypie als wichtig erachtet. Diese zwei Dinge – Konsistenz und Stereotypie – können sich gegenseitig möglicherweise aufheben oder ausschliessen, und sie sind eng verbunden mit der Validitätsproblematik von BVT. Der Klient oder die Klientin kann sich in Untersuchungen hochgradig stereotyp verhalten und immer dasselbe sagen. Dort ist dann nur auf der Oberfläche Konsistenz gegeben und BVT können dort von Vorteil sein:

> Insofern kann ich nicht nur die Konsistenz beurteilen. Konsistenz ist ein Reliabilitätsmass. Wir haben das Dilemma zwischen Reliabilität und Validität. Wir messen etwas sehr genau aber wissen nicht mehr genau was wir da gemessen haben. […] Es gibt keine Forschung über stereotypes Verhalten im Rahmen der Begutachtung, da besteht sicher ein Forschungsbedarf.

Nicht validierte Beschwerdevalidierungstests

Nicht validierte BVT oder Verfahren sind nicht wissenschaftlich evaluiert. Sie sind aber meistens im Rahmen der Gutachten klar operationalisiert und hochgradig standardisiert. Nicht validierte BVT machen den überragenden Teil der Konsistenzprüfung und unter Umständen auch der Instrumente für die Identifikation von Verdeutlichungstendenzen, Aggravation oder Simulation aus.

Die von den Gutachtenden beschriebenen, nicht validierten Verfahren bestehen aus Vergleichen von Selbstberichten der Klienten oder Klientinnen innerhalb und zwischen Untersuchungen, Vergleichen zwischen Selbstbericht und Verhalten, Vergleichen zwischen Selbstbericht und Akten und Vergleichen zwischen Selbsteinschätzung und Fremdeinschätzung.

Vorab ein typisches Verfahren innerhalb einer Expertise, wie es sehr wahrscheinlich von einem Grossteil der Gutachtenden durchgeführt wird. Es beinhaltet eine Inkonsistenzprüfung, die nicht wissenschaftlich validiert ist:

> Les rapports sont toujours construits de la même manière. J'ai entendu les plaintes, le degré de douleur, l'intensité de la douleur, l'handicap fonctionnel subjectif, ensuite je fais mon examen de A-Z avec mes références en tête. Donc je regarde quelles sont les déficiences. Au cours de cet examen, je fais les tests de non organicité ou d'incohérence, ensuite les examens d'imageries, ensuite la batterie de tests dont l'évaluation de la capacité fonctionnelle (ECF), par exemple. Après, je recoupe ces différents éléments. Si tout concorde au niveau de l'handicap fonctionnel perçu, la douleur, inférieur à 5, et puis l'ECF, effort maximal – alors ma perception rejoint celle du patient.
>
> Sinon, je vais quand même retourner sur la santé pure et dure. Je vais regarder quelle atteinte il y a de mon point de vue. Si c'est une atteinte qui est assez spécifique, je vais me baser sur une espèce de moyenne fictive que j'ai pour fixer l'incapacité de travail. Si c'est un trouble non spécifique, c'est finalement le degré de cohérence qui va me conduire à faire une reconnaissance de quelque chose. Alors, je ne peux donner une règle de pourcent. Donc, si j'ai – par exemple – des signes objectifs de souffrances lombaires incontestables et beaucoup de signes de non organicité, une ECF avec beaucoup d'autolimitations et que le patient m'annonce un handicap fonctionnel qui correspondrait à une incapacité de travail, toute profession confondue, je vais rabaisser et m'approcher à une capacité de travail que je vais plutôt mettre en rapport avec les atteintes que j'ai pu « sentir ». Donner des chiffres, c'est impossible, parce qu'il n'y a pas de recettes.

Die interviewten Gutachtenden arbeiten also mehrheitlich mit solchen nicht validierten, im jeweiligen Kontext standardisierten Verfahren und Beobachtungen. Konsistenzprüfung ist unbestritten. Eine mehr oder weniger standardisierte Gegenüberstellung der klinischen oder anderer Zeichen mit den geäusserten Beschwerden wird praktisch immer gemacht.

Es bleibt aber unklar, inwieweit man hiermit auch die effektive *Beschwerdevalidität* untersucht. Das heisst, dass man trotz einem Aufzeigen von erheblichen Inkonsistenzen auch an der Validität des Messinstruments zweifeln kann anstatt an der Validität der Beschwerden:

> Wenn mir jemand sagt, der so aussieht wie Sie jetzt: ich habe immer eine Schmerzintensität von 10. Und ich sage ihm, 10 ist der absolut maximale Schmerz auf der Welt. Dann versteht er unter diesem 10 etwas anderes als ich.

Validierte kognitive Beschwerdevalidierungstests

Die Gutachtenden verwenden entweder selbst validierte BVT im kognitiven Bereich, die ursprünglich aus der psychologischen und der forensischen Forschung kommen, oder sie kennen BVT aus anderen Gutachten. Genannt werden Persönlichkeitsinventare (Minnesota Multiphasic Personality Inventory oder MMPI-2 und der PS-16, der 16 Persönlichkeitsfaktoren erfasst), Symptomskalen (Strukturierten Fragebogen Simulierter Symptome oder SFSS, und Symptom Checklist 90 oder SCL 90), kognitive Tests für das Gedächtnis (Rey Kurzzeitgedächtnistest, Test of Memory Malingering oder

TOMM, Word Memory Test oder WMT, Kurzzeitgedächtnistest aus dem Bremer BVT) und kognitive Tests für die Aufmerksamkeit (Frankfurter Aufmerksamkeitsinventar):

> Nous, on n'hésite pas à utiliser des tests psychologiques vraiment fait pour cette question: comme le MMPI-2. En neuropsychologie, on va aussi utiliser des tests : Le TOMM pour la mémoire, c'est un test reconnue et les figures de REY.

Als beste Methode zur Beurteilung von Simulation oder Aggravation und der Beschwerdevalidität betrachten die Experten so genannte „below chance" Tests im kognitiven Bereich. Solche Tests identifizieren Antwortverhalten unterhalb des Zufalls. Bezüglich diesen Tests sind Experten der Ansicht, dass nicht auf das Testergebnis alleine abgestützt werden darf, sondern zusätzlich die Kriterien von Slick (Slick et al., 1999) erfüllt sein müssen:

> Diese Methode ist relativ sicher wenn die Restwahrscheinlichkeiten auf Grund von Hintergrundinformationen und den Kriterien von Slick annähernd ausgeschlossen werden.
>
> Pour le TOMM pour la mémoire. Il y a un seuil normal. En dessous, cela veut vraiment dire que la personne fait des efforts pour être aussi bas sur l'échelle.

Wie bereits betont, weisen viele Gutachtende darauf hin, dass es einen Unterschied gibt zwischen der Überprüfung der Konsistenz und der Beurteilung der Beschwerdevalidität. Bei vielen kognitiven Tests wird die Konsistenz innerhalb des Tests geprüft. Auch Extremantworten auf viele Fragen die aus Studien bei Patientenpopulationen nicht bekannt sind, sind unwahrscheinlich und werden mit einer möglichen Aggravation assoziiert (fake bad). Wichtig sei es auch, meinten alle Gutachtenden und Experten, dass die Plausibilität von kognitiven Tests mit anderen Befunden beurteilt wird. Die Eignung eines Tests zur Überprüfung der Konsistenz heisst aber nicht ohne weiteres, dass der gleiche Test auch als BVT verwendet werden kann.

Bezüglich der Validität der BVT kritisierten mehrere Personen, dass gewisse Tests als BVT gebraucht werden, die nicht für diese Fragestellung entwickelt wurden. Es bleibt zu erwähnen, dass sogar bei hoher Spezifität (Rate der falsch Positiven kleiner als 5 oder 1%), ohne Einbezug der a priori Prävalenz, Schlussfolgerungen gefährlich sein können, wie oben beschrieben (Seite 18-20). Unklar ist deshalb, wie gross das Risiko von falsch-positiven Ergebnissen bei der Anwendung dieser Tests für die Beurteilung der BV ist:

> Wir brauchen [für die standardisierte Erfassung der Symptome und für die Überprüfung der BV] den Symptom Checklist 90. Es ist nicht möglich dass jemand überall hohe Werte hat. Dann hat er überall das höchste angekreuzt. Indifferente Beschreibungen werden als Hinweis auf Aggravation interpretiert. Wir vergleichen die Ergebnisse mit anderen Informationen, mit der Schmerzgeschichte. Manchmal sagt eine Person in der Anamnese auf der Frage ob er Stimmen höre ‚Nein', Während er das im SCL-90 Test angibt.
>
> Der Symptom Checklist 90 ist eine Liste, die für den klinischen Alltag entwickelt worden ist. Viele Fragebögen sind nicht für die Fragestellung der Beschwerdevalidität entwickelt worden.

Bei der Verwendung kognitiver Tests müssen die Muttersprache und das oftmals niedrige Bildungsniveau vieler Exploranden berücksichtigt werden. Die Berücksichtigung von Bildung und Sprache bei der Beurteilung der BV ist nur in einzelnen Testbereichen möglich. Sprachunabhängige Tests bieten hier offensichtliche Vorteile:

> Wir verwenden in der Neuropsychologie vor allem zwei Verfahren um die Plausibilität oder Validität von Problemen zu untersuchen. Einerseits der BVT mit visuellen Tests, der kann auch von Analphabeten ausgeführt werden. Andererseits der Kurzzeitgedächtnis Test A; dabei geht es um einen Vergleich von visuellen Informationen. Der dritte Test ist der Word Memory Test für das Gedächtnis. Der liegt vor auf Deutsch, Englische, Spanisch, Französisch, Italienisch, Spanisch, Kroatisch, Albanisch und Portugiesisch.
>
> Ich glaube nicht dass es ein MMPI in all den benötigten Sprachen gibt (Albanisch, Türkisch, Serbokroatisch). Und die Fragen sind sprachlich sehr anspruchsvoll. Viele Exploranden mit wenige Jahre Schulbildung können diese Fragebögen nicht ausfüllen.

Im Weiteren sind Beschwerden bei unspezifischen Erkrankungen, die meistens den Hintergrund der Begutachtung bilden, starken interkulturellen Unterschieden unterlegen. Viele Gutachtende bezweifeln, ob so genannt „validierte“ Tests auch interkulturell validiert wurden.

Die Gutachtenden verwenden kognitive Leistungstests nicht nur zur Überprüfung der Konsistenz, oder als BVT, sondern auch mit verschiedenen anderen Zielen. Die Beurteilung der BV steht nicht im Vordergrund. Die Ausgangslage einer Begutachtung besteht darin, dass auf der Grundlage der verbleibenden Leistungsfähigkeit anschliessend die Arbeitsfähigkeit bestimmt werden soll. Die Beurteilung der kognitiven Fähigkeit ist aus mehreren Gründen notwendig. Viele Exploranden klagen über kognitive Symptome wie Aufmerksamkeits-, Gedächtnis- und Konzentrationsstörungen. Deshalb muss die Leistungsfähigkeit in diesen kognitiven Bereichen bestimmt werden. Hinzu kommt, dass viele Exploranden die bisherigen, vielfach körperlich schweren Tätigkeiten, nicht mehr ausführen können. Sie sollen, wenn möglich, in einer anderen, leichten Tätigkeit integriert werden. Die neuropsychologische Eignung für alternative Tätigkeiten muss also abgeklärt werden.

Mehrere Gutachtende äusserten sich zur Verwendung von Persönlichkeitsfragebögen wie den MMPI und den 16-PS. Beide Instrumente sind sehr stark psychopathologisch ausgerichtet. Der MMPI ist der bekannteste Persönlichkeitsfragebogen der Welt, er liegt in vielen Sprachen vor und wird ständig weiterentwickelt. Der MMPI dürfte sich, nach der Meinung eines Gutachters, zur Validierung von anderen Symptomskalen und Listen eignen. Ein Gutachter beschrieb die gleichzeitige Anwendung des MMPI für die Leistungsdiagnostik, die Arbeitserprobung, die Beurteilung der Arbeitsfähigkeit in einer leichten Tätigkeit, und als BVT. Der MMPI dürfte ziemlich resistent sein gegen coaching. Kritisiert werden die Konstrukte die den Subskalen zu Grunde liegen:

> Der MMPI ist nicht sehr durchsichtig mit den 560 Fragen, dauert 1.5 bis 3 Stunden und ist in dem Sinne auch eine Arbeitserprobung für eine leichte Tätigkeit. Er ist problematisch hinsichtlich der Konstrukte weil diese sehr empirisch gebildet wurden und nicht theoretisch fundiert sind. Das ist ein grosser Itempool der immer wieder neu faktorisiert worden ist und da entstanden Konstrukte beziehungsweise Skalen die nicht immer genau überzeugend das abbilden was sie benennen. Die deutschsprachige Version ist jetzt glaube ich auch wieder angepasst worden an deutsche Verhältnissen mit mehr deutschen Formulierungen und so und insofern kann man den sicherlich auch hier nutzen.
>
> Es gibt ein anderes Persönlichkeitsinventar, der 16-PS mit 260 Fragen und 16 Persönlichkeitsfaktoren. Der PS-16 enthält auch Intelligenz-Items, ein paar Wissens- und Rechenaufgaben die da eingestreut sind. Da können sie die Intelligenz abschätzen und das Profil zur Intelligenz oder zur Bildungsbeeinträchtigung in Beziehung setzen.

Instrumente zur Erfassung der Beschwerden (englisch: ‚symptoms’) untersuchen unwahrscheinliche und eher unglaubwürdige Aussagen. Einige Gutachtende kannten solche standardisierte Instrumente, beispielsweise der SIRS (Structured Interview for Reported Symptoms). Für den SIRS liegt keine deutsche Übersetzung vor. Dies scheint aufgrund der Tatsache, dass dieser Test in englischer Form als Gold-Standard verwendet wird, erstaunlich. Im deutschen Sprachraum ist dafür der SFSS (Strukturierter Fragebogen Simulierter Symptome) recht bekannt, dennoch als BVT nicht ohne Berücksichtigung der weiteren Befunde zu interpretieren. Zum SFSS meinte ein Experte:

> Der SFSS ist sehr empfehlenswert, er ist das Beste was wir zu Verfügung haben. Der SFSS ist zur Hypothesengenerierung und zur Untersetzung von Befunden sicherlich geeignet. Für eine endgültige Unterscheidung ist er in vielen Grenzfällen kritisch. Es muss also wirklich ein kompetenter Untersucher sein. Wir haben leider im deutschen Sprachraum noch nicht genügend Daten um Sensitivität und Spezifität zu beurteilen. Das Instrument ist noch nicht perfekt. Das Problem ist, dass ich meine Entscheidung nicht darauf basieren darf.

Für die unterschiedlichen kognitiven Leistungen können jeweils unterschiedliche Erhebungsmethoden gewählt werden. Die Erfassung depressiver Symptome kann mittels Befragung, Persönlichkeits-

fragebogen oder Symptomliste erfolgen. Allgemein bekannt und sehr wichtig in Zusammenhang mit der Plausibilität, Konsistenz und BV ist, dass die erhaltenen Informationen stark abhängig sind von der gewählten Methode der Erfassung. Diese Tatsache zeigt einerseits, dass eine zuverlässige Beurteilung oft nur möglich ist, wenn Befunde eines Merkmals, die mit mehreren unterschiedlichen Methoden erfasst wurden, gemeinsam interpretiert werden. Andererseits ist diese Tatsache auch ein Hinweis darauf, dass gewisse Methoden empfindlicher sind für Aggravation. Die Plausibilität der Beschwerden kann durch den Vergleich der Befunde überprüft werden:

> Ich habe häufig das Phänomen, dass wenn ich psychologische Symptome abfrage, ich zu ganz unterschiedlichen Ergebnissen komme in Abhängigkeit davon ob diese Symptome eingebaut sind in einer Symptomliste oder in einem Persönlichkeitsfragebogen. Bei einem Persönlichkeitsfragebogen erwarten die Leute gar nicht, dass es um den Nachweis ihrer Beschwerden geht, sondern mehr darum dass sie beschreiben wer sie sind, wie sie sind, und es geht nicht primär ums Klagen. Wenn depressive Symptome im Beschwerdefragebogen mit 98% und im Persönlichkeitsfragebogen mit 50% angegeben werden, dann frage ich mich was bedeutet das in Hinblick auf die Ausprägung der depressiven Symptomatik.

Validierte somatische Beschwerdevalidierungstests

Es gibt keine den Interviewten bekannten wissenschaftlich klar validierten BVT, die nicht in den psychokognitiven Bereich gehören. Ausserhalb der kontinuierlichen Beobachtung des Alltags von Exploranden sind die meisten Tests streng genommen nicht validiert. Trotzdem beschreiben wir sie in diesem Kapitel, da es doch zum Teil – im Gegensatz zu individuellen Vorgehensweisen – intersubjektiv standardisierte und beschriebene Verfahren sind, die aber nicht als eigentliche BVT validiert wurden. So wird zum Beispiel das Verhalten bei der körperlichen Untersuchung mit den Waddell-Zeichen untersucht. Dieser Test wurde entwickelt um somatische von anderen Ursachen für Rückenschmerzen zu unterscheiden.

> [Les Waddell]…, je les fais pratiquement systématiquement. En principe, c'est réservé à la lombalgie et c'est réservé aux patients de notre culture.

Die Evaluation der Funktionellen Leistungsfähigkeit (EFL) nach Isernhagen ist eine standardisierte Untersuchung der arbeitsbezogenen körperlichen Leistungsfähigkeit mit über 20 Tests welche die Arbeit in unterschiedlichen Ausgangsstellungen und mit steigender Belastung mit externen Gewichten evaluiert. Das EFL wird angewendet in der zweitägigen Vollversion und in weniger bekannten – in verschiedenen Settings entwickelten – Kurzversionen. EFL werden in zertifizierten Rehabilitationskliniken von spezialisierten Therapeuten durchgeführt und beinhalten hochplausible Konsistenzkriterien, auch wenn diese nicht wissenschaftlich validiert sind. Diese Beurteilung der Inkonsistenz mit den standardisierten Kriterien weist eine Sicherheitsmarge und eine Gradierung auf. Die Konsistenz wird als gut beurteilt bei 0-1 positiven Kriterien, als mässig bei 2-4 und als schlecht bei über 4 positiven Kriterien.

Es wird aber darauf hingewiesen, dass das EFL ursprünglich vor allem als arbeitbezogene Leistungsmessung gedacht war und die integrierte Konsistenzprüfung meistens nicht im Mittelpunkt stand. In der berufsbezogenen Begutachtung für die IV nimmt die Konsistenzprüfung des EFL im Sinne einer BVT einen zentralen Platz ein:

> In der Rehabilitation können Sie eher davon ausgehen, dass jemand zeigen will, was er kann. Im gutachterlichen Sektor muss er – das ist das System – beweisen, was er nicht kann, sonst gibt es keine Versicherungsleistungen.
>
> J'ai défendu à plusieurs reprises le fait que ce que ce test nous montre, c'est que la personne collabore ou elle ne collabore pas. Et ça, un bon clinicien ou un bon rhumatologue, il le voit aussi s'il fait un examen de rhumatologue.

Anscheinend ist der Bedarf nach EFL seitens der IV nicht gross. Zusätzlich ist der Aufwand für einen klassischen EFL – 2 Tage – für viele zu hoch. Andere Anwender und Anwenderinnen haben in ihren Organisationen so genannte mini-EFL entwickelt, die vom Aufwand her beträchtlich geringer sind und die eine Mischform von Leistungs-, Belastungs-, Beschwerdevalidierungstests oder Motivationstests darstellen. Diese werden wie der klassische EFL immer in Kombination mit dem PACT (Performance Assessment Capacity Testing) durchgeführt. Über die Reliabilität und Validität vom PACT gibt es keine wissenschaftlichen Studien. Der PACT erfasst die selbst geschätzte körperliche Leistungsfähigkeit mit 50 Bildern von arbeitsbezogenen Alltagsaktivitäten. Die Angaben der Exploranden werden mit den Ergebnissen der Leistungstests und mit anderen Befunden verglichen. Mini-EFL sind in diesem Sinne nicht wissenschaftlich validiert, zeichnen sich aber aus durch ein hohes Mass an Nachvollziehbarkeit, Plausibilität und Machbarkeit.

Eine somatoforme Schmerzstörung (SFS) kann mit dem „Screening for Somatoforme Symptoms" (SOMS) erfasst werden. Eine SFS liegt vor bei 10-12 positiven Antworten auf dem SOMS. Beschwerden der letzten 2 Jahre werden mit dem SOMS2, und Beschwerden der letzten 7 Tage mit dem SOMS7 erfasst. Der Vergleich der Ergebnisse kann Hinweise auf Inkonsistenzen geben:

> Man sieht häufig, in etwa 60% der Fälle, dass Patienten und Patientinnen im zweiten Fragebogen, der die letzten 7 Tagen erfasst, Beschwerden angeben die im ersten Fragebogen für die letzten 2 Jahren nicht angegeben wurden. Das geht nicht.

Ein weiterer Zugang für somatische oder körperliche BVT ist die Dolorimetrie, die Erfassung von Schmerzen. Ein Experte unterscheidet hier wiederum zwischen Reliabilitätsprüfung und eigentlichem BVT: Bei der Fibromyalgie wird z.B. die Druckempfindlichkeit multilokal überprüft. Welche Druckpunkte das genau sind „ist eigentlich unerheblich". Es wird überprüft, wie gut Exploranden ihre eigenen Schwellenwerte reproduzieren. Wenn sie dazu nicht in der Lage sind, kann man aber nicht sagen, wie druckschmerzhaft diese Punkte sind. Das würde dann „nicht als ein BVT betrachtet werden, sondern als eine Reliabilitätsprüfung der Angaben zur eigenen Schmerzempfindlichkeit". Erwähnt wurde in diesem Zusammenhang auch die Giessener Beschwerdeskala; bei dieser geht es darum ob Schmerzen konsistent sind oder nicht.

Es gibt auf der nicht-kognitiven Ebene auch biologische Tests: Dazu gehören Bluttests. Diese verifizieren die Compliance der Exploranden bei der Medikamenteneinnahme. Aufgrund der gemessenen Blutwerte wird hier zum Teil eine hohe Häufigkeit von Exploranden beschrieben, die die verschriebenen Medikamente nicht einnehmen. Es wird aber auf der Gegenseite gerade die mangelhafte Präzision der mit diesen Tests erfassten Blutwerte kritisiert. Die Messfehler seien viel zu gross, um sichere Rückschlüsse zu erlauben:

> Dans 80%, on ne les trouve pas [les médicaments]. Par contre, quand on commence à chercher dans les urines, on trouve THC, Cocaïne, alcool…on trouve beaucoup. Personne ne parle de cette problématique.
>
> Es gibt eine Review im New England Journal, die sagt, dass Sie bei den Antidepressiva rein durch die individuelle Bioverfügbarkeit Spiegelschwankungen zwischen 1 und 20 haben. Wenn sie dort einen „ungenügenden" Spiegel feststellen, dürfen Sie nicht daraus auf Malcompliance schliessen.

Andere Tests im körperlichen Bereich die erwähnt wurden, waren der JAMAR Handkraft Test, der „Pseudo-Strength Test", und der „Step-Test". Bei der Handkraft wird eine Abhängigkeit der Kraft mit der Griffweite erwartet mit einer maximalen Kraft in einer mittleren Griffposition. Ergebnisse der Handkraft können auch mit anderen Leistungen verglichen werden, zum Beispiel mit der mit den Händen gehobenen Last. Leider liegen kaum wissenschaftliche Untersuchungen in Zusammenhang mit diesem Test vor. Der „Pseudo-Strength Test" und der „Step-Test" simulieren hohe körperliche

Anstrengungen. Sie quantifizieren Autolimitationen und wurden für die Prognostik im Rehabilitationsbereich entwickelt, z.B. für Return to work. Diese Tests sollten daher mit Vorsicht im Kontext von echten BVT verwendet werden.

4.2.6 Barrieren

Die Interviewten beschrieben verschiedene Barrieren für die Anwendung von BVT:

- Diese hängen zuerst einmal zusammen mit der grundsätzlichen Einstellung der Gutachtenden zu BVT. Wie bereits betont, sind BVT im Sinne einer Konsistenzprüfung, also als Reliabilitätsmass, praktisch unbestritten. Jedoch herrscht Dissens über BVT im Sinne von eigentlicher Beschwerde*validierung,* wenn es in Richtung von Feststellung von Aggravation oder Simulation geht.
- Die Ausbildung von Gutachtenden auf dem Gebiet der Testdiagnostik ist ebenfalls ein Thema. Als Anwender und Anwenderin müssen die Gutachtenden die Gütekriterien der BVT kennen und die interne Validität beurteilen können. Die externe Validität und auch die ökologische Validität von BVT, die zum Teil als problematisch eingestuft wird, sollte ebenfalls in solchen Ausbildungen thematisiert werden.
- Auch wird die wissenschaftliche Qualität von „validierten" BVT oder BVT-Sets angezweifelt (mit Analogstudien bestimmte Pseudo-Spezifität und Pseudo-Sensitivität oder Bestimmung der Konstruktvalidität mit Known-Groups Designs sowie mangelhafte Kreuzvalidation).
- Am meisten wird jedoch die externe Validität hinterfragt, d.h. die Anwendbarkeit von validierten BVT auf Exploranden aus anderen Kulturen, aus anderen sozialen Umfeldern und in einem anderen Versicherungshintergrund. Natürlich ist auch der ethische Konflikt, nämlich die für die gutachterliche Tätigkeit notwendige Abstraktion von möglichen sozialen krankmachenden Faktoren – also der Abstraktion der „Soziosomatik" – ein Thema.
- Ein Hindernis ist auch die Tabuisierung des Themas sowie ein steigender Druck der Auftraggeber auf die Gutachtenden. Manche Gutachtende meinen, dass man sich damit in der Schweiz sehr schnell unbeliebt macht. In Fragebögen sei z.B. der Anteil der Migranten und Migrantinnen, die ‚unter Zufall' (below chance) reagieren, „verheerend hoch". Bei vielen Gutachtenden wird es als Problem anerkannt, „aber alle scheuen sich davor es zu benennen".

Grundsätzlich sind die meisten Gutachtenden sehr interessiert an einer Diskussion bezüglich BVT, sie sind auch an spezifischen BVT interessiert. Trotzdem vertreten die Gutachtenden unterschiedliche Meinungen bezüglich der Motivation, die der Entwicklung von BVT zugrunde liegt. So befürchten einige vor allem politisch motivierte Interessen. Sie wollen in diesem Sinne eine weitere Diskussion mittragen, wollen sich aber eine individuelle Wertung von solchen Tests vorbehalten.

4.2.7 Wissenschaftliche Literatur und Forschungsbedarf

Die Interviewten haben sich mehr oder weniger mit der wissenschaftlichen Literatur auseinandergesetzt und zum Teil auch selber publiziert. Die vorwiegend angelsächsische Literatur wird mehrheit-

lich als interessant erachtet, doch die Übertragbarkeit auf die Schweiz wird eher nicht befürwortet. Deutsche Publikationen werden – vor allem in der Westschweiz – als zu „simulationslastig“ empfunden. Französische Publikationen sind international praktisch unbedeutend.

Eine Sorge unter den Gutachtenden ist die wissenschaftliche Qualität von „validierten“ BVT. In Analogstudien ist der falsche Goldstandard ein Problem vor allem für die externe Validität. Analogstudien haben daher eine eingeschränkte Validität. Eine Täuschung kann nur als sicher angegommen werden, wenn sie von Probanden im Rahmen einer Studie gemäss Auftrag ausgeführt wird. Die externe Validität dieser Forschung bei experimentellen Täuschenden ist beschnitten und daher sind die Tests nicht unbedingt anwendbar bei reellen Exploranden. Da es keinen Goldstandard gibt, kommt man über eine gewisse Konstruktvalidität nicht hinaus.

Neben den Designs von Studien wird auch die Quantifizierung der Gütekriterien selbst hinterfragt. So wird zum Beispiel die a priori Prävalenz vielfach nicht berücksichtigt, wenn man Tests mit hoher Spezifität anwendet (vgl. Seite 18-20).

Wie bereits erwähnt, gibt es laut den Fachleuten keine Forschung über stereotypes Verhalten im Rahmen der Begutachtung. Da Stereotypie an der Oberfläche Konsistenz vortäuscht, ist die Erforschung von BVT im Zusammenhang mit Stereotypie relevant.

Bei neuropsychologischen Tests wird von den Fachleuten vorgeschlagen, dass man wegkommen sollte von verbalen Tests sowie von den Testapplikationen auf Bildschirmen. Andere Modalitäten wie Sehen und Hören könnten und sollten vermehrt nonverbal geprüft werden. Auch hier besteht Forschungsbedarf.

BVT wurden vor allem im psychokognitiven Bereich entwickelt. Es besteht ein grosser Nachholbedarf für Tests im körperlichen Bereich. Im körperlichen Bereich ist die Entwicklung von validen BVT aber mit grösseren Problemen verbunden, da das Prinzip eines BVT gerade darin besteht, ganz einfache und mit Sicherheit zu leistende Aufgaben zu kreieren. Es gibt wenige BVT bezüglich motorischer Symptome. Diese versucht man meistens in Beziehung zu setzen zu neurologischen Schäden oder zu körperlichen Veränderungen, die man gefunden hat. Die Fachleute meinen aber, dass dieser Zusammenhang überschätzt wird.

4.2.8 System und Strukturen

Die Interviewten betrachten verschiedene Aspekte des Systems und der Strukturen im Umfeld einer Begutachtung als relevant für die Frage der Anwendung von BVT. Wichtig sei eine Verbesserung der Kommunikation den verschiedenen Instanzen im Umfeld der IV. Ein verbessertes Feedback ist aber für die Verbesserung der Qualität der Gutachten unabdingbar:

> Wenn man davon ausgeht, dass Systeme nur mit Feedback-Schleifen gut funktionieren können, ist die momentane Situation für die Gutachtenden unbefriedigend.

IV-Stellen richten unterschiedliche Anforderungen an die Gutachten. Teilweise bestehen auch Diskrepanzen zwischen den Anforderungen der IV-Stellen und den Leistungsverträgen mit dem BSV. Die befragten Fachpersonen weisen auch darauf hin, dass die fachlichen Grundlagen im Bereich der medizinischen Versorgung und in der Begutachtung für die IV unterschiedlich sind. In der Medizin hat das biopsychosoziale Krankheitsmodell an Bedeutung gewonnen, indem auch soziale Faktoren der Krankheitsentstehung bei der medizinischen Diagnosestellung und Behandlung einbezogen

werden. Hingegen sollen sozialen Faktoren bei der Beurteilung der Arbeitsfähigkeit für die IV nicht berücksichtigt werden.

Nach Ansicht der befragten Gutachtenden ist die IV teilweise damit konfrontiert, Probleme des politischen und sozialen Systems absorbieren zu müssen. Dies wird als inadäquat beurteilt. Die Gutachtenden befürchten, dass die politischen Entwicklungen und die restriktivere Praxis der IV gewisse soziale Probleme verschieben könnten anstatt sie zu lösen:

> Jede Kultur hatte einen gewissen Prozentsatz an Leuten, die nicht „funktionierten" so wie es der Maxime entspricht. …Was sich geändert hat im Laufe der Jahrhunderte sind die Etiketten.

Das Bestreben der IV die Kosten zu senken, könnte dazu führen, dass Gutachten vermehrt dort in Auftrag geben, wo die Arbeitsfähigkeit in der Regel höher beurteilt wird. Niedrig beurteilte Arbeitsfähigkeiten werden vom Auftraggeber viel öfter hinterfragt als hohe Arbeitsfähigkeiten. Diese Entwicklung gefährdet die Neutralität und Objektivität der Gutachtenden.

Die befragten Fachpersonen sind der Ansicht, dass die formellen Anforderungen an die Gutachten in letzter Zeit stark gestiegen sind. Die zunehmende Verwendung der Gutachten vor Gericht führt dazu, dass Gutachten nicht nur medizinisch sondern auch juristisch korrekt verfasst werden müssen. Gutachten dienen den Fallverantwortlichen der IV als Entscheidungsgrundlage. Zusätzlich sollen die Beurteilungen und Grundlagen im Falle einer gerichtlichen Verwendung eindeutig sein. Einige Gutachter betrachten die juristisch korrekte Formulierung der Gutachten nicht als ihre Aufgabe:

> Je sens parfois une pression de la part des SMR. La traduction de mon argument dans le langage juridique n'est pas mon travail.

4.2.9 Entwicklungen und Erwartungen

Die Mehrheit der Gutachtenden befürwortet eine weitere Auseinandersetzung mit BVT, besteht aber weiterhin auf einer individuellen Wertung dieser Tests durch Expertinnen und Experten. Einige Gutachtende sind wie bereits beschrieben der Meinung, dass es eigentlich sinnlos sei, nach harten BVT zu suchen, da die Frage nach der Bewusstseinsnähe z.B. bezüglich Simulation nicht zu beantworten sei.

Die Experten meinen aber, dass gerade wegen diesem Aspekt der Bewusstseinsnähe vermehrt Neuropsychologen und Psychologen in den Gutachterprozess eingebunden werden sollten, dass vermehrt Psychodiagnostik statt Medizindiagnostik gemacht werden sollte:

> BVT im kognitiven Bereich müssen von Psychologen oder Neuropsychologen beurteilt werden. Man kann sie nicht einem Arzt oder einer Assistentin übergeben. Auch in Deutschland ist die Vorstellung verbreitet, dass man zusätzlich zur körperlichen Untersuchung noch ‚schnell' ein Simulationstest macht. Das ist ganz gefährlich in beiden Richtungen.

Wieder andere wollen vermehrt BVT einsetzen und fördern, dazu bräuchte es aber eine verbesserte Ausbildung und Schulung, aber auch eine verbesserte Wertschätzung in diesem Bereich. Einige Gutachtende glauben, dass z.B. das Besuchen von Kursen über Malingering heute noch eher unüblich ist, und dass die Wertschätzung von diesbezüglichem Einsatz heute noch nicht gegeben sei.

> Je fais des cours sur le malingering. …, j'ai dés fois l'impression que celui qui fait ça se fait mal voir.

Die Interviewten der MEDAS meinen, dass sich in den letzten Jahren die Anforderung an ein Gutachten verändert hat. Der Anteil einfacher Gutachten hat abgenommen. Die seit der Inkraftsetzung der 4. IV-Revision bestehenden RAD übernehmen die bis dahin weitgehend von den Gutachtende

der MEDAS durchgeführten versicherungsmedizinischen Abklärungen. Aufwand und Umfang der Gutachten sind höher als früher. Die Komplexität der Fälle nehme zu. Immer mehr liegen viele Akten und inkonsistente Gutachten vor. Die Dicke der Dossiers nimmt zu und es liegen vermehrt Gerichtsurteile vor.

4.3 Fazit

4.3.1 BVT im kognitiven Bereich

BVT werden vor allem im kognitiven Bereich verwendet. Genannt wurden diesbezüglich von Experten und Gutachtenden Persönlichkeitsinventare (MMPI-2 und PS-16), Symptomskalen (SFSS, und SCL 90), kognitive Tests für das Gedächtnis (Rey Kurzzeit, TOMM, WMT, Kurzzeitgedächtnistest aus dem Bremer BVT) und kognitive Tests für die Aufmerksamkeit (Frankfurter Aufmerksamkeitsinventar).

4.3.2 Verhaltener Gebrauch von BVT in der Praxis

Wissenschaftlich validierte BVT werden von einer Minderheit der interviewten Gutachtenden verwendet. Eine Konsistenzprüfung innerhalb und im Umfeld von Gutachten ist jedoch unbestritten. Experten und Gutachtende sind sich einig, dass BVT, wenn für Gutachten verwendet, nur als Mosaiksteine und nie ausschliesslich in die Entscheidungen bezüglich Arbeitsfähigkeit einfliessen dürfen.

4.3.3 Ungenügend validierte BVT vor allem im körperlichen Bereich

In diesem Sinne wird in der Praxis – ausser für die genannten BVT im kognitiven Bereich – eher mit nicht validierten Verfahren gearbeitet, mit denen die Konsistenz überprüft wird. Im körperlichen Bereich, in der Neurologie und in der Rheumatologie gibt es nach unserem Wissenstand keine wissenschaftlich validierten BVT. Hingegen gibt es standardisierte Verfahren auf einer tieferen Validitätsstufe, die bei der Konsistenzprüfung helfen, aber keine eigentlichen BVT sind, weil sie zum Teil für andere Zwecke entwickelt wurden als für die Beschwerdevalidierung, oder weil eben keine genügende Validierung als BVT vorliegt. Dazu gehören z.B. die Evaluation der funktionellen Leistungsfähigkeit (EFL) in Voll- oder Kurzversionen, kombiniert mit dem Performance Assessment Capacity Test (PACT), die Waddell-Zeichen, der JAMAR Handkraft Test, psychophysische Tests, Bluttests zum Erfassen von Malcompliance, Verfahren der Dolorimetrie sowie das Screening für somatoforme Schmerzstörung (SOMS2 und der SOMS7).

4.3.4 Konsistenzprüfung versus Beschwerdevalidierung

Es hat sich herausgestellt, dass Gutachtende und zum Teil auch Experten zwischen Konsistenzprüfung und eigentlicher Beschwerde*validierung* (diagnostische Tests für die Identifikation von Simulation und Aggravation) unterscheiden. Der Anspruch von BVT ist nicht derselbe wie derjenige einer

Konsistenzprüfung. Konsistenz ist ein Mass für die Reliabilität und daher unproblematischer als ersteres, welches die schwierigere Validitätsfrage stellt. Die Frage nach der Bewusstseinsnähe bei Aggravation und Simulation muss gestellt werden, ist aber sehr schwierig zu beantworten. Hier wird von Experten ein vermehrter Einbezug von Neuropsychologen und Psychologen vorgeschlagen.

4.3.5 Quantensprung und Unbestimmtheit

Viele Gutachtende bezweifeln ob einzelne BVT oder Testkombinationen es erlauben, mit ausreichender Sicherheit Simulation zu beurteilen. Es bleiben Momentaufnahmen, wobei zu viele Faktoren schon bei der Absolvierung solcher Tests – wie im ganzen Gutachten – einfliessen. Testresultate werden mitbeeinflusst durch den Kontext der Abklärung. Einzelne oder kombinierte BVT werden deshalb als nicht *valide* betrachtet. Die Gutachtenden bezeichnen die Entscheidung bezüglich der Arbeitsfähigkeit, welche sich ja auf die Beschwerde*validität* abstützt, als eine Art Quantensprung, dem eine gewisse Unbestimmtheit inhärent ist.

4.3.6 Andere Entscheidungshilfen

Die Validität der isoliert betrachteten BVT ist ungenügend, da die Fehlerquote in vielen Situationen zu hoch sein kann. Die Kriterien von Slick et al. und Bianchini et al. (vgl. Kapitel 2.3) können bei der Verminderung der erwähnten Unbestimmtheit oder Restwahrscheinlichkeit eine Hilfe sein. Durch eine Kombination von qualitativen und quantitativen Elementen kann die Einschätzung von Verdeutlichung, Aggravation und Simulation erleichtert werden.

4.3.7 Forschungsbedarf

Grundsätzlich sind die Gutachtenden sehr offen bezüglich der weiteren Entwicklung und Forschung auf dem Gebiet von BVT. Die Gutachtenden wünschen aber nicht, dass das BSV die Einführung der BVT zu schnell vorantreibt. Die Gutachtenden behalten sich vor, weiterhin die verschiedenen Tests selber zu wählen, zu werten und deren Validität zu beurteilen. Sie interessieren sich für einen weiteren wissenschaftlichen Diskurs bezüglich BVT, kritisieren aber diesbezüglich auch vermutete politische Beweggründe. Die wissenschaftliche Validität von BVT soll weiter untersucht werden; vor allem die externe Validität sei im Moment ungenügend erforscht und belegt. BVT im körperlichen Bereich seien im Gegensatz zu kognitiven Tests noch ungenügend entwickelt.

4.3.8 Häufigkeit von Simulation

Die Prävalenz von Simulation wird von einem Grossteil der Gutachtenden als gering beschrieben. Die Mehrheit der interviewten Gutachtenden sind der Meinung, dass Simulation aufgrund deren Häufigkeit in ihrer Arbeit kein eigentliches Kernproblem darstellt. Die im Vergleich dazu höhere Grundrate in einem Teil der Malingering-Forschung wird darauf zurückgeführt, dass dort das Versi-

cherungsumfeld verschieden sei, und dass eine gänzliche Abstraktion von sozialen krankmachenden Faktoren solche Zahlen erklären könne.

5 Schriftliche Befragung von MEDAS- und RAD-Gutachtenden

5.1 Methoden

Es wurde eine strukturierte Befragung der potenziellen Anwenderinnen und Anwender von BVT durchgeführt. Dabei sollte untersucht werden, welche Beschwerdevalidierungstests von Gutachtenden bereits benutzt werden, welche Erfahrungen damit gemacht wurden und welcher Bedarf eventuell nach entsprechenden Instrumenten besteht. Von Interesse waren auch die Vor- und Nachteile sowie die Vorbehalte gegenüber dem Einsatz von Beschwerdevalidierungstests. Der Fragebogen sollte in diesem Sinne auch Problembereiche in der Anwendung von BVT in der Praxis abdecken, die sich im Rahmen der Experteninterviews verdeutlicht haben.

Hauptzielgruppe unter den Anwendenden waren Fachpersonen, die zu Handen der IV Gutachten erstellen, insbesondere Mitarbeitende der Regionalärztlichen Dienste RAD und der medizinischen Abklärungsstellen MEDAS. Über die Verantwortlichen der MEDAS (18) und der RAD (10) wurde der Fragebogen an die entsprechenden Ärztinnen und Ärzte dieser Organisationen weitergeleitet. Die Befragung war anonym.

Das Instrument enthielt geschlossene und offene Fragen zu den relevanten Punkten. Die Interviews hatten gezeigt, dass bei vielen Gutachtenden eine Inkonsistenzprüfung unbestritten war, die Identifikation von Aggravation und Simulation jedoch nicht. Deshalb wurden diese beiden Aspekte in den Fragen auseinander gehalten. Es wurde zwischen der Inkonsistenz- oder Konsistenzprüfung auf der einen Seite und der Diagnostik von Aggravation und Simulation auf der anderen Seite unterschieden. Auch wurde zwischen Aggravation und Simulation unterschieden.

In diesem Sinne wurden folgende 9 Aspekte im Fragebogen erhoben, 6 geschlossene Fragen sowie 3 offene Fragen:

Geschlossene Fragen:

1. Wie wichtig ist für Sie die Beurteilung von Inkonsistenzen bei der Beurteilung der Arbeitsfähigkeit eines Klienten oder einer Klientin?

2. Verwenden Sie standardisierte Verfahren/Tests zur Identifikation von Inkonsistenzen in Ihrer gutachterlichen Arbeit?

3. Verwenden Sie nicht-standardisierte Verfahren/Tests zur Identifikation von Inkonsistenzen in Ihrer gutachterlichen Arbeit?

4. Bei fragwürdiger Plausibilität der Befunde, aber unter Akzeptanz von Verdeutlichungstendenzen im Verhalten des Klienten, der Klientin: Inwieweit besteht Ihres Erachtens
   a. ein Zusammenhang zwischen Inkonsistenzen und Aggravation?
   b. ein Zusammenhang zwischen Inkonsistenzen und Simulation?

5. Kann im Rahmen von Begutachtungen
   a. ein Urteil über Aggravation gebildet werden?
   b. ein Urteil über Simulation gebildet werden?

6. Ist es nach Ihrer Ansicht die Aufgabe des Gutachtenden, nicht nur Inkonsistenzen aufzuzeigen, sondern auch das
   a. Ausmass von Aggravation festzustellen?
   b. Ausmass von Simulation festzustellen?

Offene Fragen:

1. Welche standardisierten Verfahren/Tests zur Beurteilung und Identifikation von Inkonsistenzen kennen Sie. Und wie häufig wenden Sie diese Verfahren an?

2. Welche standardisierten Verfahren/Tests zur Beurteilung und Identifikation von Aggravation oder Simulation kennen Sie? Und wie häufig wenden Sie diese Verfahren an?

3. Haben Sie einen zusätzlichen Kommentar zu diesem Thema?

Die Auswertung der strukturierten Befragung erfolgte mit deskriptiv-statistischen Methoden mit der Statistiksoftware SPSS.

5.2 Resultate

Insgesamt 30 Fragebogen wurden uns von Gutachtenden der RAD und der MEDAS zurückgesandt, je 15 aus der Romandie und 15 aus der Deutschschweiz. Einige MEDAS- und RAD-Stellen antworteten mit einem einzigen gemeinsamen Fragebogen, einige Stellen verteilten den Fragebogen intern an mehrere Gutachtende, die für die entsprechende Organisation für Expertisen tätig sind. Die Stichprobe ist in diesem Sinne nicht repräsentativ für alle potentiellen Anwender und Anwenderinnen von Beschwerdevalidierungstests. Tabelle 9 zeigt die detaillierten Resultate mit den absoluten und relativen Antworthäufigkeiten pro Sprachregion und gesamthaft für die geschlossenen Fragen.

Wichtigkeit Bestimmen von Inkonsistenz: 77% der antwortenden Gutachtenden fanden Inkonsistenzen sehr wichtig für das Bestimmen der Arbeitsfähigkeit, 23% für eher wichtig. Es bestand kein relevanter Unterschied zwischen der Romandie und der Deutschschweiz.

Standardisierte Tests: 43% der antwortenden Gutachtenden wenden nie standardisierte Tests an, 23% sehr häufig. Es gab keinen relevanten Unterschied zwischen den Sprachregionen.

Nicht-standardisierte Tests: Bei der Anwendung von nicht-standardisierten Tests zeigte sich ein inverses Antwortverhalten in der Deutschschweiz und der Romandie. In der Westschweiz wenden 46% nie nicht-standardisierte Tests an, in der Deutschschweiz sind dies nur 7%. Dagegen wenden 40% der antwortenden Gutachtenden in der Deutschschweiz sehr häufig nicht-standardisierte Tests an (7% in der Romandie).

Tabelle 9 Antworthäufigkeiten pro Sprachregion und total

		Sprachregion				Gesamt	
		Deutschschweiz		Romandie		Anzahl	%
		Anzahl	%	Anzahl	%	Anzahl	%
Wichtigkeit von Inkonsistenzen	eher wichtig	3	20.0%	4	26.7%	7	23.3%
	sehr wichtig	12	80.0%	11	73.3%	23	76.7%
Anwendung stand. Test	nie	7	46.7%	6	40.0%	13	43.3%
	manchmal	3	20.0%	3	20.0%	6	20.0%
	häufig	2	13.3%	2	13.3%	4	13.3%
	sehr häufig	3	20.0%	4	26.7%	7	23.3%
Anwendung nicht-stand. Test	nie	1	6.7%	7	46.7%	8	26.7%
	manchmal	2	13.3%	3	20.0%	5	16.7%
	häufig	6	40.0%	4	26.7%	10	33.3%
	sehr häufig	6	40.0%	1	6.7%	7	23.3%
Zusammenhang Inkonsistenz-Aggravation	leichter	3	23.1%	3	30.0%	6	26.1%
	mässiger	2	15.4%	3	30.0%	5	21.7%
	starker	8	61.5%	4	40.0%	12	52.2%
Zusammenhang Inkonsistenz-Simulation	keiner	1	8.3%			1	4.5%
	leichter	4	33.3%	4	40.0%	8	36.4%
	mässiger	5	41.7%	1	10.0%	6	27.3%
	starker	2	16.7%	5	50.0%	7	31.8%
Urteil möglich: Aggravation	sehr selten	1	7.1%	1	7.7%	2	7.4%
	eher selten	1	7.1%			1	3.7%
	eher häufig	11	78.6%	9	69.2%	20	74.1%
	sehr häufig	1	7.1%	3	23.1%	4	14.8%
Urteil möglich: Simulation	sehr selten	5	38.5%	4	30.8%	9	34.6%
	eher selten	4	30.8%	3	23.1%	7	26.9%
	eher häufig	4	30.8%	3	23.1%	7	26.9%
	sehr häufig			3	23.1%	3	11.5%
Gutachter-Aufgabe: Aggravation	nein	1	7.1%	1	7.7%	2	7.4%
	eher nein			2	15.4%	2	7.4%
	eher ja	9	64.3%	4	30.8%	13	48.1%
	ja	4	28.6%	6	46.2%	10	37.0%
Gutachter-Aufgabe: Simulation	nein	3	23.1%	1	7.7%	4	15.4%
	eher nein	2	15.4%	3	23.1%	5	19.2%
	eher ja	5	38.5%	5	38.5%	10	38.5%
	ja	3	23.1%	4	30.8%	7	26.9%

Inkonsistenz als Indikator für Aggravation und Simulation: 74% der antwortenden Gutachtenden sehen einen mässigen oder starken Zusammenhang zwischen bestehenden Inkonsistenzen und Aggravation, 59% sehen einen mässigen oder starken Zusammenhang zwischen bestehenden In-

konsistenzen und Simulation, mit einem umgekehrten Antwortverhalten bezüglich mässig und stark zwischen der Romandie und der Deutschschweiz.

Urteil bezüglich Aggravation und Simulation: Für 89% der antwortenden Gutachtenden ist ein Urteil bezüglich Aggravation eher häufig bis sehr häufig möglich. Für ein Urteil bezüglich Simulation sind 38% der Meinung, dass dies eher häufig bis sehr häufig möglich ist. Für 35% ist ein Urteil über Simulation sehr selten möglich.

Bestimmung von Simulation und Aggravation als Aufgabe von Gutachtenden: 85% der antwortenden Gutachtenden sind der Meinung, dass die Identifikation von Aggravation zur Aufgabe eines Gutachtenden gehört. Für Simulation sind 65% dieser Meinung.

Tabelle 10 Von den Gutachtenden genannte Tests für die Identifikation von Inkonsistenz

Genannte Tests zum Bestimmen von Inkonsistenz

		Häufigkeit	Prozent
Test	Waddell	8	22.2
	MMPI	6	16.7
	PACT	3	8.3
	HAMD	2	5.6
	Med-Spiegel	2	5.6
	Andere Labor	1	2.8
	Benton	1	2.8
	Electronystagmus	1	2.8
	HADS	1	2.8
	Indizienliste Foerster	1	2.8
	Kahn	1	2.8
	MAST	1	2.8
	Med-Sp_Ser	1	2.8
	Med-Sp_Ur	1	2.8
	MMST	1	2.8
	SKID II	1	2.8
	TOMM	1	2.8
	WMT	1	2.8
	Zahlenverbindungstest	1	2.8
	Zung	1	2.8
	Gesamt	36	100.0

Erklärung: MMPI: Minnesota Multiphasic Personality Inventory, PACT: Personal Assessment Capacity Test, HAMD: Hamilton Depression Scale, Benton: Benton Visual Retention Test, HADS: Hospital Anxiety and Depression Scale, Kahn: Kahn Test of Symbol Arrangement, MAST: Michigan Alcohol Screening Test, MMST: Mini-Mental Status-Test, SKID II: Strukturiertes Klinisches Interview für DSM-IV, TOMM: Test of Memory Malingering, WMT: Word Memory test, Zung: Zung Depression Scale

Mit den offenen Fragen wurden Tests für das Bestimmen von Inkonsistenz und für das Bestimmen von Aggravation und Simulation erfasst, die den Gutachtenden bekannt sind. Ebenfalls erfasst wurde dabei die Häufigkeit der Anwendung der genannten Tests.

Tabelle 10 (s.o.) zeigt die von den Gutachtenden genannten standardisierten Verfahren oder Tests zum Bestimmen von Inkonsistenz. Am häufigsten genannt wurden hier die Waddell-Zeichen und der MMPI. Die Anwendung der Waddell-Zeichen wurde von 7 Gutachtenden als sehr häufig und von einem Gutachter als häufig beschrieben. Der MMPI wird von je 2 Gutachtenden sehr häufig, häufig und selten eingesetzt. Der PACT wird von 3 Gutachtenden sehr häufig eingesetzt, der HAMD von 2 Gutachtenden sehr häufig. Medikamentenspiegel verschiedener Art wurden von 5 Gutachtenden angegeben, sie werden von diesen selten bis häufig eingesetzt. Alle anderen einmal genannten Tests werden häufig bis sehr häufig eingesetzt, ausser der TOMM und der MMST, die selten angewandt werden.

Tabelle 11 Von den Gutachtenden genannte Tests zum Bestimmen von Aggravation und Simulation

Genannte Tests zum Bestimmen von Aggravation und Simulation

		Häufigkeit	Prozent
Test	MMPI	3	16.7
	ASTM	2	11.1
	NP-Verfahren	2	11.1
	HADS	1	5.6
	HAMD	1	5.6
	IQ (Teile)	1	5.6
	Labor	1	5.6
	MADRS	1	5.6
	MMST	1	5.6
	MSPQ	1	5.6
	MSVT	1	5.6
	Stroop Test	1	5.6
	Waddell	1	5.6
	WMT	1	5.6
	Gesamt	18	100.0

Erklärung: MMPI: Minnesota Multiphasic Personality Inventory, ASTM: Amsterdam Short Term Memory Test, HAMD: Hamilton Depression Scale, HADS: Hospital Anxiety and Depression Scale, MADRS: Montgomery-Åsburg Depression Rating Scale, MMST: Mini-Mental Status-Test, MSPQ: Modified Somatic Perception Questionnaire, MSVT: Medical Symptom Validity Test, Stroop: Neurological Screening Test, WMT: Word Memory test

Tabelle 11 (s.o.) zeigt die den Gutachtenden bekannten standardisierten Verfahren oder Tests zum Bestimmen von Aggravation und Simulation. Drei Gutachtende nannten den MMPI, der von diesen Gutachtenden selber aber kaum angewendet wird. Genannt, aber nie bis selten angewendet wurde auch der ASTM. Von den Tests, die von einem einzigen oder von zwei Gutachtenden genannt wur-

den, werden viele − wenn überhaupt − nur selten angewendet. Ausnahmen sind der HADS (sehr häufig), der HAMD (sehr häufig), der MADRS (sehr häufig), der MMST (häufig) sowie die Waddell-Zeichen (sehr häufig).

Einige Gutachtende machten von der Möglichkeit zusätzlicher Kommentare zur Problematik von Beschwerdevalidierungstests Gebrauch. So waren die Anwendenden z.T. der Meinung, dass es keine wirklich validierten Tests für Simulation gäbe. Diese Tests würden praktisch immer unwissenschaftlich verwendet. Entscheidend sei also die klinische Beobachtung. Bemühungen, die schwierige gutachterliche Problematik der Aggravation durch objektive Tests klären zu können, führten wahrscheinlich in die Irre. Tests lieferten pseudoobjektive Daten. Aggravation sei z.T. normal und der Grad der Aggravation hänge stark vom Verhalten des Untersuchers ab. Diese Tests seien also fragwürdige Methoden.

Zudem seien nur wenige Tests bekannt, die sich für fremdsprachige Exploranden eignen. Vergleiche aus verschiedenen Quellen seien hier viel aufschlussreicher. Rund 60-70% der Exploranden beherrschten die deutsche Sprache nicht und seien schlecht sozial integriert. Dort würden standardisierte Verfahren klar an Ihre Grenzen stossen. Es wurde auch dargelegt, dass Versicherungsbetrug kein medizinisches Problem sei. Aggravation könne hingegen auf dem Hintergrund der medizinischen Parameter beurteilt werden. Die Tests seien vielfach nicht validiert für die Gutachtersituation und für fremdsprachige Personen. Letztlich ergebe dies eine Pseudoobjektivität für ein juristisches und nicht medizinisches Problem.

Inkonsistenz sei sicher ein Indikator für Aggravation und Simulation, aber Schlüsse seien delikat. Wirkliche Simulanten würden sich z.B viel konsistenter verhalten als psychisch schwache Personen. Vorhandene Inkonsistenzen könnten zwar auf Aggravation und Simulation hinweisen, jedoch auch durch möglicherweise krankheitswertige psychogene Störungen bedingt sein. Je nachdem, ob und wie gut die Psychodynamik und etwaige psychiatrische Diagnosen in der psychiatrischen Begutachtung beschrieben oder ausgeschlossen werden, sinke oder steige der Indikatorwert von Inkonsistenzen für Aggravation und Simulation.

Absolut notwendig seien vertiefte Kenntnisse der Testmethodik sowie eine umfangreiche eigene Testerfahrung in klinischen und anderen Bereichen. Die Durchführung und Interpretation könne eher durch einen Psychologen FSP oder Neuropsychologen erfolgen. Die Anwendung von BVT durch schlecht ausgebildete Expertinnen und Experten ist mit einem Fehlerrisiko behaftet.

Eigentliche Simulation sei nur durch externe Beobachtung mit ausreichend hoher Sicherheit zu identifizieren. Zur Identifikation von Aggravation hingegen seien eine grosse klinische Erfahrung und der interdisziplinäre Konsens vielfach genügend für ein hinreichend gutes Bild.

Gutachtende sollten ausdrücklich auf Inkonsistenzen hinweisen, hilfreich für die Versicherung sei die Erwähnung der Ursachen für dieses Verhalten (inkl. der Verwendung) statt Vermeidung/Tabuisierung der Begriffe Aggravation/Simulation.

5.3 Fazit

Die Rücklaufquote der Befragung war tiefer als erwartet. Gründe dafür sehen wir in einer teilweisen Tabuisierung der Thematik rund um BVT, Aggravation und Simulation. Die vorliegende Stichprobe

ist als nicht-repräsentativ für alle Anwender und Anwenderinnen von BVT anzusehen. Daher kann aus den Resultaten dieser Stichprobe sicher nicht auf die Meinung aller potentiellen Anwender und Anwenderinnen geschlossen werden. Die Daten bestätigen die bereits in den Interviews angedeutete Heterogenität der Meinungen zu Zusammenhängen zwischen Inkonsistenz, Aggravation und Simulation. Sie bestätigen auch eine gewisse Skepsis unter den Gutachtenden gegenüber BVT.

Das Erfassen von Inkonsistenzen – mehr oder weniger strukturiert – scheint unumstritten. Alle Anwender und Anwenderinnen der Stichprobe erachten das Erfassen von Inkonsistenzen wichtig für das Bestimmen der Arbeitsfähigkeit. Eine Minderheit der Gutachtenden wendet häufig standardisierte Tests zum Erfassen von Inkonsistenzen an und 43% der Gutachtenden wenden nie standardisierte Tests an.

In der Romandie ist die Anwendung von nicht-standardisierten Tests seltener als in der Deutschschweiz. Dieser Unterschied ist jedoch aufgrund der unscharfen Definition eines nicht-standardisierten Tests und aufgrund der nicht repräsentativen Stichprobe vorsichtig zu interpretieren, zumal aus der Romandie häufiger einzelne Fragebögen aus derselben Organisation zurückgeschickt wurden als aus der Deutschschweiz.

Inkonsistenz wird von 74% der Gutachtenden als ein mässiger bis starker Indikator für Aggravation und von 59% als ein mässiger bis starker Indikator für Simulation gesehen. Für 89% der Gutachtenden ist ein Urteil bezüglich Aggravation eher häufig bis sehr häufig möglich. Beim Urteil bezüglich Simulation waren 38% der Meinung, dass dies eher häufig bis sehr häufig möglich ist. Für 35% bleibt ein Urteil über Simulation sehr selten möglich. 85% der Gutachtenden waren der Meinung, dass die Identifikation von Aggravation eher bis sicher zur Aufgabe eines Gutachtenden gehört. Bezüglich Simulation waren 65% dieser Meinung.

Genannte Tests für das Bestimmen von Inkonsistenzen waren die Waddell-Zeichen, der MMPI, der PACT und der HAMD sowie verschiedene Labor- und Medikamentenspiegeltests. Die Häufigkeit der Anwendung war gross ebenfalls für die Waddell-Zeichen (sehr häufig), den PACT (sehr häufig), den MMPI (häufig) und den HAMD (sehr häufig).

Als Hilfe für das Bestimmen von Aggravation und Simulation wurden vor allem der MMPI, der ASTM und andere, nicht näher bestimmte neuropsychologische Verfahren genannt. Nur der HADS (sehr häufig), der HAMD (sehr häufig), der MADRS (sehr häufig), der MMST (häufig) sowie die Waddell-Zeichen (sehr häufig) werden aber auch mehr als selten angewendet.

Die zusätzlichen Kommentare der Gutachtenden auf die offene Frage entsprachen zum grossen Teil den Aspekten, die bereits in den Interviews thematisiert wurden. Dazu gehören eine als mangelhaft empfundene ökologische und externe Validität von BVT-Testbatterien, die Betonung der Wichtigkeit der klinischen Erfahrung, die Betonung der Wichtigkeit der Ausbildung bezüglich Testmethodik und Testinterpretationen mit der damit verbundenen Gefährlichkeit der Tests bei unsachgemässer Anwendung, die Problematik einer möglichen Verlagerung eines schliesslich juristischen und nicht medizinischen Problems in die Pseudoobjektivität, der wiederholte Hinweis auf eine zwar meistens notwendige, aber nicht hinreichende Bedingung von Inkonsistenz für die Postulierung von Simulation und Aggravation.

Eigentliche Simulation sei nur durch externe Beobachtung mit ausreichend hoher Sicherheit zu identifizieren. Zur Identifikation von Aggravation hingegen seien eine grosse klinische Erfahrung und der interdisziplinäre Konsens vielfach genügend für ein hinreichend gutes Bild.

Gutachtende sollten ausdrücklich auf Inkonsistenzen hinweisen. Hilfreich für die Versicherung sei die Erwähnung der Ursachen für das Verhalten der Exploranden, auch mit der Verwendung der Begriffe Aggravation/Simulation.

Bei den genannten Tests zur Identifikation von Inkonsistenzen und Aggravation/Simulation nannten die Anwenderinnen und Anwender auch Tests, die nicht als BVT entwickelt wurden. Die Anwender und Anwenderinnen verwenden zum Teil dieselben Tests für die Identifikation von Inkonsistenzen und für die Identifikation von Aggravation und Simulation.

Die Waddell-Zeichen z.B werden relativ häufig verwendet. Sie wurden entwickelt, um somatische von anderen Ursachen für Rückenschmerzen zu unterscheiden; sie können also im weitesten Sinne als eine Art Inkonsistenztest interpretiert werden – gewisse Beschwerden sind inkonsistent mit einem somatischen Problem. Sie sollten aber nicht als BVT angewandt werden. Andere Tests hingegen sind klar als BVT entwickelt worden wie der MMPI. Ein BVT ist gleichzeitig ein Inkonsistenztest. Dies erklärt auch, dass weniger Tests für Aggravation/Simulation genannt wurden als für Inkonsistenzen, denn BVT sind in diesem Sinn eine Untergruppe der Inkonsistenztests.

Die Ergebnisse der Anwenderbefragung weisen darauf in, dass die Begriffe Beschwerdevalidierung, Inkonsistenz, Aggravation und Simulation im gutachterlichen Alltag unscharf abgegrenzt werden. So bleibt trotz der in den Interviews und in den Kommentaren dieser Befragung häufig gemachten Abgrenzung von Inkonsistenz und Aggravation/Simulation aufgrund der Daten unklar, was genau unter Inkonsistenz verstanden wird. Vielleicht müsste man hier zwischen einer „gemessenen Inkonsistenz“ und einer „wahren Inkonsistenz“ (=Simulation) unterscheiden. Die Messung von Inkonsistenz – also die Erhebung von Widersprüchlichkeiten – kann höchstens die Reliabilität betreffen und umgeht so das Problem der Validität; daher ist sie unter Gutachtenden praktisch unbestritten.

Die meisten Gutachtenden wollen aber nicht aufgrund von einzelnen oder mehreren standardisierten Tests ein Urteil fällen über das tatsächliche Vorliegen von Beschwerden. Aussagen über die „wahre Inkonsistenz“ – oder eben über Simulation ausgehend von BVT – scheinen ihnen zu gewagt, wenn nicht gar unmöglich.

6 Synthese und Diskussion der Befunde

6.1 In Kürze: Antworten auf die zentralen Fragestellungen

Die Hauptergebnisse dieser Studie können mit Blick auf die ersten beiden Fragestellungen folgendermassen zusammengefasst werden.

1. Für die Überprüfung der Plausibilität von Beschwerden und Beeinträchtigungen (Beschwerdevalidität) im körperlichen Bereich, welche Klientinnen und Klienten im Rahmen einer Begutachtung vortragen, liegen keine standardisierten Beschwerdevalidierungstests (BVT) vor. Es muss deshalb auf nicht-standardisierte BVT zurückgegriffen werden. Dagegen liegen für eine begrenzte Anzahl psychischer Symptome standardisierte BVT vor.
2. Die Validität isoliert betrachteter Ergebnisse von BVT wird aus wissenschaftlicher Sicht im Allgemeinen als ungenügend beurteilt.
3. Von den BVT abzugrenzen sind Leitlinien. Diese definieren Kriterien zur Beurteilung der Plausibilität von Beschwerden im Gesamtzusammenhang einer Begutachtung; BVT sind soweit als möglich ein Teil dieses Vorgehens.
4. Aus der Sicht der Praxis, d.h. von (vielen) Gutachtenden und Expertinnen und Experten, wird die Überprüfung der Beschwerdevalidität (BV) im Rahmen der Begutachtung als verbesserungsfähig beurteilt. Am meisten werden nicht-standardisierte BVT verwendet. Standardisierte BVT werden bei einer Minderheit der Gutachten − vorwiegend von Psychologen − in den verschiedenen Settings der Sozial- und Unfallversicherungen (SUVA, IV, KK, Taggeld, Haftpflicht) eingesetzt. Auch die Praktikerinnen und Praktiker betonen, dass die zuverlässige Interpretation der Ergebnisse standardisierter und nicht standardisierter von BVT nur unter der Berücksichtigung aller im Rahmen einer Begutachtung erhobenen Befunde möglich ist.

Die folgenden Abschnitte dieses Kapitels liefern eine ausführlichere Synthese der Ergebnisse der vorliegenden Studie.

6.2 Methodische Grenzen der Studie

Bei der Bewertung der Ergebnisse der vorliegenden Studie muss auch auf einige methodische Grenzen hingewiesen werden:

- Die Literaturrecherche und insbesondere die Verarbeitung der wissenschaftlichen Literatur zu BVT musste aufgrund der grossen Zahl von Referenzen eingeschränkt werden sowohl in zeitlicher Hinsicht (Veröffentlichungen von 2005 bis 2007) als auch in geografischer Hinsicht (deutscher Sprachraum). Die getroffene Auswahl bildet insgesamt einen kleinen Teil der wissenschaftlichen Literatur ab, die in den letzten Jahren in anerkannten Fachzeitschriften erschienen ist. Wir gehen aber davon aus, dass es sich bei den ausgewählten Tests, die einer ausführlicheren Besprechung zugeführt wurden, nicht nur in der Wissenschaft sondern auch in der Praxis um anerkannte Verfahren handelt.

- Die schriftliche Befragung der Gutachtenden und potenziellen Anwenderinnen und Anwender von BVT ist bedingt repräsentativ, da die Rücklaufquote der Fragebögen tief war
- Die Begutachtung von Aggravation und Simulation ist auch in Bezug auf deren theoretische Grundlagen ein komplexes Unterfangen. Eine breitere Rezeption der entsprechenden Grundlagenliteratur wäre wertvoll, konnte aber im Rahmen dieses Auftrages nur in groben Zügen vorgenommen werden. Es wurde jedoch versucht, die zentralen, insbesondere auch kritischen Aspekte herauszuarbeiten.

6.3 Diskussion und Synthese der Befunde

6.3.1 Theoretische und begriffliche Grundlagen

Verdeutlichung, Aggravation, Simulation, Bewusstheit und Anreiz

Wichtig ist die Unterscheidung zwischen den Begriffen der Verdeutlichung, Aggravation und Simulation. Verdeutlichung wird als häufiges und in der Regel legitimes Verhalten in der Abklärungssituation erachtet: der Klient oder die Klientin möchte sicher gehen, ernst genommen zu werden und beschreibt seine Beschwerden deshalb leicht übertrieben. Aggravation meint dagegen die gezielte und massive Übertreibung vorhandener Beschwerden und Simulation das massive Vortäuschen nichtvorhandener Beschwerden.

Wesentlich ist dabei die Abgrenzung von Aggravation und Simulation gegenüber psychischen Erkrankungen. Eine zentrale Rolle für diese Abgrenzung spielen die Bewusstheit und die Motivierung des Verhaltens eines Klienten, einer Klientin in der Abklärungssituation. Dabei wird in der Fachliteratur postuliert, je bewusster und je stärker der Klient oder die Klientin durch externe Anreize motiviert ist, desto grösser ist die Wahrscheinlichkeit für aggravierendes oder simulierendes Verhalten.

Die Begriffe Verdeutlichung, Aggravation und Simulation sind jedoch nur bedingt trennscharf. In der englischsprachigen Fachliteratur wird deshalb zunehmend nur ein Begriff, Malingering, mit der Abstufung des Schweregrades verwendet.

Allerdings kann die Bewusstheit des Verhaltens des Exploranden in der Begutachtung aus wissenschaftlicher Sicht nicht mit Sicherheit beurteilt werden. Die Gutachtenden wissen um diese Einschränkung. Der Verzicht auf den Begriff würde als Informationsverzicht betrachtet. Die Gutachtenden formulieren ihre Beurteilung im Zusammenhang mit der Bewusstheit der Aggravation oder Simulation mit entsprechender Vorsicht und sprechen zum Beispiel von Bewusstseinsnähe.

Konsistenz und Simulation

Alle Gutachtenden befürworten die Überprüfung der Konsistenz der Befunde im Rahmen eines Gutachtens. Zurückhaltung besteht aber bei der Beurteilung der Beschwerdevalidität auf Basis der Plausibilität und Konsistenz der Beschwerden. So kann die Ursache für Inkonsistenzen auch in der ungenauen oder fehlerbehafteten Befunderhebung liegen: Diese aber darf nicht gleichgesetzt werden mit der mangelnden Validierung der Beschwerden. Die Konsistenz wird als Beurteilung der Zuverlässigkeit der Befunde betrachtet und bezieht sich primär auf das Gutachten selbst.

Die Beurteilung der Beschwerdevalidität hat Konsequenzen, die über das Gutachten hinausreichen − zum Beispiel für die Berentung − und wird deshalb als problematischer betrachtet. Deshalb wer-

den an Beschwerdevalidierungstests (BVT) hohe wissenschaftliche Anforderungen gestellt, die in der Regel von diesen zur Zeit noch nicht erfüllt werden können.

Ein möglicher weiterer Grund für die Zurückhaltung der Gutachtenden bei der Verwendung von BVT ist, dass die BVT dem Gutachtenden im Rahmen der Abklärung weniger Interpretationsspielraum belassen als die nicht standardisierte Überprüfung der BV. Positive Ergebnisse von BVT können schwer ignoriert oder wegdiskutiert werden. Nicht standardisiert erfasste Inkonsistenzen hingegen können in einem weit grösseren Ausmass nach eigenem Ermessen dargestellt und beschrieben werden. Deshalb könnte im Interesse der Förderung einer objektiveren Beurteilung der BV die häufigere Verwendung von BVT propagiert werden. Dabei warnen allerdings sowohl Gutachtende als auch wissenschaftliche Fachleute vor einer nicht verantwortbaren Zunahme falsch positiver Befunde bei übermässiger Verwendung von BVT.

6.3.2 Prävalenz von Aggravation und Simulation

Viele Studien im gutachterlichen Kontext − sowohl in Europa als auch in Amerika − zeigen, dass die Prävalenz einer reduzierten Beschwerdevalidität bei Exploranden mit schwer objektivierbaren Diagnosen beachtlich ist. In der Schweiz liegen noch keine systematisch erhobenen Zahlen über die Prävalenz einer reduzierten Beschwerdevalidität bei Exploranden mit schwer objektivierbaren Diagnosen vor. Ott et al. schätzen, dass in der IV bei den Neuberenteten 8-18% der Renten potenziell nicht zielkonform entrichtet werden (Ott et al., 2007). Höher liegen dürfte diese Rate in den Risikogruppen für Aggravation und Simulation, nämlich bei Personen mit psychischen Erkrankungen, Behinderung in Zusammenhang mit unspezifischen Schmerzen und Burn-out. Somit wäre die Prävalenz von Aggravation und Simulation hoch genug um eine allgemeine Überprüfung der BV zu befürworten.

Die tiefere Einschätzung der Prävalenz von Aggravation und Simulation durch Gutachtende in der Schweiz, bei Abwesenheit systematisch erhobener Daten, kann unterschiedliche Ursachen haben. Möglicherweise ist die Prävalenz in der Schweiz tatsächlich viel tiefer: dies könnte erklärt werden mit Unterschieden zwischen den einzelnen Ländern im Versicherungs- und gesellschaftlichen Umfeld. Möglich ist aber auch, dass die Prävalenz in der Schweiz unterschätzt wird, weil die Diagnose von Aggravation oder Simulation von den Gutachtenden hierzulande restriktiver angewendet wird.

Eine repräsentative Erhebung der BV in den Risikogruppen wäre sinnvoll, um die unklare Datenlage bezüglich der Prävalenz der Aggravation und Simulation in der Schweiz zu verbessern.

6.3.3 BVT und Leitlinien

Leitlinien

Die wissenschaftliche Literatur empfiehlt Leitlinien für die Beurteilung der BV. Am bekanntesten sind einerseits die Leitlinien von Slick et al. (1999) für die Beurteilung der BV bei Exploranden mit psychischen und kognitiven Leistungseinbussen. Bianchini et al. (2005) erweiterten diese Leitlinien für die Beurteilung der BV bei Exploranden mit Schmerzen und körperlicher Behinderung. Kernelemente der Bianchini-Leitlinien sind der Nachweis eines externen Anreizes, von Inkonsistenzen mit standar-

disierten oder nicht standardisierten Verfahren und der Ausschluss von psychischen Erkrankungen. Die genannten Leitlinien sind in der Schweiz kaum bekannt. Die Gutachtenden verwenden aber Kriterien die den Bianchini-Leitlinien sehr ähnlich sind. Einigkeit besteht auch darin, dass standardisierte BVT nicht isoliert verwendet werden sollten, sondern im Rahmen einer systematischen Vorgehensweise unter Anwendung konkret formulierte Leitlinien.

BVT, Leitlinien und Alternativen dazu

Die Anwendung von BVT ist im Rahmen der Begutachtung psychischer und kognitiver Funktionen möglich. Die Tests müssen in der Regel von Fachpersonen ausgewertet werden. Ob gewisse BVT auch von Nicht-Psychologen angewendet und ausgewertet werden können, wurde im Rahmen dieser Studie nicht untersucht.

Aus wissenschaftlicher Sicht werden hohe Anforderungen an BVT gestellt, die jedoch nur selten erfüllt werden können. Bemängelt wird in der wissenschaftlichen Literatur ausserdem, dass auch die Leitlinien (s.o.) ungenügend evaluiert sind. Mit Blick auf die kritischen Einwände gegenüber den BVT stellt sich allerdings die Frage nach Alternativen zur Verwendung von BVT und Leitlinien. Wenn sich Gutachtende nicht auf BVT und Leitlinien abstützen, dann ist das Risiko falscher Ergebnisse durch ihre Verwendung ausgeräumt. − Aber auch die Verwendung anderer Verfahren birgt das Risiko falsch-positiver Befunde, das in diesem Fall nicht einmal bekannt ist. Ausserdem aber würde auf Ergebnisse wissenschaftlich evaluierter BVT verzichtet. In dieser Situation empfiehlt sich u.E. als beste Vorgehensweise die Anwendung standardisierter und/oder nicht-standardisierter BVT als Teil eines durch systematische Leitlinien abgestützten Abklärungsprozederes.

Gutachtende befürworten einstimmig die Überprüfung der Konsistenz der Befunde im Rahmen einer Abklärung. Nicht beurteilt werden konnte in der vorliegenden Studie, wie die Konsistenz in den einzelnen Gutachten überprüft wird. Oft scheint das konkrete Vorgehen bei der Überprüfung der BV abhängig zu sein vom Verdacht des Gutachtenden auf Aggravation und Simulation. Studien zeigen, dass die subjektive Beurteilung von Aggravation und Simulation durch Gutachtende sehr unzuverlässig ist und deshalb ein Risiko besteht, dass Aggravation und Simulation nicht diagnostiziert werden.

7 Schlussfolgerungen und Empfehlungen

Folgende Empfehlungen können aus den Ergebnissen der vorliegenden Studie abgeleitet werden:

1. Standortbestimmung der Beschwerdevalidierung in der Gutachtenspraxis

Die aktuelle Qualität der Gutachten in Bezug auf die Beschwerdevalidierung ist unbekannt und kritisch zu hinterfragen. Die vorliegende Studie zeigt, dass in der Fachliteratur anerkannte Verfahren der Beschwerdevalidierung in der gutachterlichen Praxis der Schweiz nur teilweise angewendet werden. In der aktuellen Situation, in der Gutachtende oft stark individuell geprägte Vorgehensweisen bei der Abklärung anwenden, ist die Qualität nicht garantiert. Es ist nicht zuletzt im Interesse der antragstellenden Klientinnen und Klienten, wenn die Begutachtung nach einheitlicheren Massstäben erfolgt.

2. Anwendung von Beschwerdevalidierungstests und Leitlinien als Teilelement des diagnostischen Prozesses

Die Beschwerdevalidierung sollte systematisch und auf der Basis von fachlich anerkannten Leitlinien erfolgen und − soweit verfügbar − auch unter Anwendung von standardisierten BVT.

Die Gutachtenden äussern umfassende Vorbehalte gegenüber der diagnostischen Zuverlässigkeit von BVT. Es ist richtig, dass diese Tests keine absolut zuverlässigen Entscheide in Bezug auf Aggravation oder Simulation liefern können. Es ist aber zu betonen, dass eine grundsätzliche Nicht-Anwendung von BVT keineswegs zu einer zuverlässigeren Diagnose führt bzw. eine verantwortungsvollere Option darstellt. Im Unterschied nämlich zu den Tests − bei denen Aussagen zur diagnostischen Validität (Sensitivität, Spezifität) gemacht werden können − ist die diagnostische Zuverlässigkeit anderer Methoden unbekannt.

3. Aus- und Weiterbildungsangebote im Bereich Versicherungsmedizin zum Thema Beschwerdevalidierung

Aus- und Weiterbildungsangebote im Bereich der Versicherungsmedizin zu Fragen der Beschwerdevalidierung und damit verbunden zur Diagnostik schwer objektivierbarer Gesundheitsstörungen sind zu fördern. In Erwägung zu ziehen ist auch die Etablierung versicherungsmedizinischer Qualitätszirkel.

4. Bestimmung der Prävalenz von Aggravation und Simulation in der Schweiz

Kenntnisse der Prävalenz von Aggravation/Simulation in ausgewählten realen Abklärungssettings der Schweiz und bei spezifischen Störungsbildern sind eine wichtige Voraussetzung für die verantwortbare Anwendung der BVT. Denn es gilt: je niedriger die Prävalenz, desto grösser die Wahrscheinlichkeit falsch-positiver Ergebnisse bei den BVT. Ausserdem könnten Kenntnisse der Prävalenz von Aggravation und Simulation in der Schweiz auch die Bereitschaft zur Verwendung von BVT im Rahmen der Gutachtenspraxis verbessern.

5. Förderung der Validierung und Entwicklung von Beschwerdevalidierungstests mit Relevanz für die Abklärungspraxis

Die Entwicklung von standardisierten BVT steht zumindest im deutschen Sprachraum noch in den Anfängen, und für die Beurteilung der körperlichen Leistungsfähigkeit fehlen entsprechende Tests noch weitgehend. Deshalb besteht ein erheblicher Forschungsbedarf in diesem Bereich.

Bestehende BVT sind noch wenig − in der Schweiz noch praktisch gar nicht − überprüft in der Anwendung bei realen Klientenpopulationen von Sozialversicherungen.

Die Validierung ausgewählter bereits bestehender BVT an realen Populationen der Gutachtenspraxis ist zu unterstützen oder anzustossen. Die Entwicklung neuer Instrumente ist aufwändig, aber dennoch zu prüfen.

Literaturverzeichnis

Arbisi, P. A., & Butcher, J. N. (2004). Psychometric perspectives on detection of malingering of pain - Use of the Minnesota Multiphasic Personality Inventory-2. *Clinical Journal of Pain, 20*(6), 383-391.

Bachmann, L. M., ter Riet, G., Clark, T. J., Gupta, J. K., & Khan, K. S. (2003). Probability analysis for diagnosis of endometrial hyperplasia and cancer in postmenopausal bleeding: an approach for a rational diagnostic workup. *Acta obstetricia et gynecologica Scandinavica, 82*(6), 564-569.

Bianchini, K. J., Greve, K. W., & Glynn, G. (2005). On the diagnosis of malingered pain-related disability: lessons from cognitive malingering research. *Spine Journal, 5*(4), 404-417.

Blaskewitz, N. (2005). *Diagnostik der Beschwerdenvalidität.* Unveröffentlichte Diplomarbeit, Humbold Universität Berlin, Berlin.

Blaskewitz, N., & Merten, T. (2006). Validität und Reliabilität von Beschwerdenvalidierungstests und -indikatoren. Eine experimentelle Studie. *Zeitschrift für Neuropsychologie, 17*(1), 35-44.

Blaskewitz, N., & Merten, T. (2007). Diagnostik der Beschwerdenvalidität - Diagnostik bei Simulationsverdacht: ein Update 2002 bis 2005. *Fortschritte der Neurologie, Psychiatrie, 75*(3), 140-154.

Bogner, A. (2005). *Das Experteninterview: Theorie, Methode, Anwendung.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Bolan, B., Foster, J.-K., Schmand, B., & Bolan, S. (2002). A comparison of three tests to detect feigned amnesia: The effects of feedback and the measurement of response latency. *Journal of Clinical and Experimental Neuropsychology, 24*(2), 154-167.

Brockhaus, R., & Merten, T. (2004). Neuropsychologische Diagnostik suboptimalen Leistungsverhaltens mit dem Word Memory Test. *Nervenarzt, 75*(9), 882-887.

Buri, M., Härter, A., & Sottas, G. (2007). *IV-Statistik 2007*. Bern: Bundesamt für Sozialversicherungen.

Bush, S. S., Ruff, R. M., Troster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., et al. (2005). Symptom validity assessment: Practice issues and medical necessity - NAN policy & planning committee. *Archives of Clinical Neuropsychology, 20*(4), 419-426.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., et al. (2006). Diagnostik der Beschwerdenvalidität: Praktische Gesichtspunkte und medizinische Erfordernisse. *Neurologie & Rehabilitation, 12*, 69-74.

Carragee, E. J. (2008). Validity of self-reported history in patients with acute back or neck pain after motor vehicle accidents. *Spine Journal, 8*, 311-319.

Cima, M., Hollnack, S., Kremer, K., Knauer, E., Schellbach-Matties, R., Klein, B., et al. (2003). Strukturierter Fragebogen Simulierter Symptome": Die deutsche Version des "Structured Inventory of Malingered Symptomatology: SIMS". *Nervenarzt, 74*(11), 977-986.

Cima, M., Pantus, M., & Dams, L. (2007). Simulation und Dissimulation in Abhängigkeit vom strafrechtlichen Kontext und der Persönlichkeit. *Praxis der Rechtspsychologie, 17*(1), 47-62.

Cunnien, A. (1997). Psychiatric and medical syndromes associated with deception. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 23-46). New York: The Guilford Press.

Dilling, H., Mombour, W., & Schmidt, M. H. (1993). *Internationale Klassifikation psychischer Störungen. ICD-10 Kapitel V (F). Klinisch-diagnostische Leitlinien.* (2. ed.). Bern: Huber Verlag / WHO.

Dohrenbusch, R. (2007). *Begutachtung somatoformer Störungen und chronifizierter Schmerzen. Konzepte − Methoden − Beispiele*. Stuttgart: Kohlhammer.

Edens, J.-F., Poythress, N.-G., & Watkins-Clay, M. 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.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46*, 913-920.

Faust, D. (1995). The detection of deception. *Neurological Clinics, 13*, 255-265.

Freud, S. (1986). *Vorlesungen zur Einführung in die Psychoanalyse, Band 11*. Frankfurt a. M: Fischer (8. Aufl.).

Gill, D., Green, P., Flaro, L., & Pucci, T. (2007). The role of effort testing in independent medical examinations. *The Medico-Legal Journal, 75*(2), 64-71.

Green, P., Iverson, G.-L., & Allen, L. (1999). Detecting malingering in head injury litigation with the Word Memory Test. *Brain Injury, 13*(10), 813-819.

Greene, C. (2005). A direct comparison of the MMPI-2 and the PAI in the detection of malingering. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 65*(8-B).

Greve, K. W., & Bianchini, K. J. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: a methodological commentary with recommendations. *Archives of Clinical Neuropsychology, 19*(4), 533-541.

Hall, H. V., & Pritchard, D. A. (1996). *Detecting malingering and deception. Forensic Distortion Analysis (FDA)*. Delray Beach FL: St. Lucie Press.

Halligan, P. W., Bass, C., & Oakley, D. A. (2003a). Wilful deception as illness behaviour. In P. W. Halligan, C. Bass & D. A. Oakley (Eds.), *Malingering and illness deception* (pp. 3-30). Oxford: Oxford University Press.

Halligan, P. W., Bass, C., & Oakley, D. A. (Eds.). (2003b). *Malingering and illness deception*. Oxford: Oxford University Press.

Hartman, D.-E. (2002). The unexamined lie is a lie worth fibbing. Neuropsychological malingering and the Word Memory Test. *Archives of Clinical Neuropsychology, 17*(7), 709-714.

Herzberg, P. Y., & Frey, A. (2007). Testinformation: Amsterdamer Kurzzeitgedächtnistest (AKGT). *Diagnostica, 53*(4), 226-228.

Heubrock, D. (1995). Neuropsychologische Diagnostik bei Simulationsverdacht: ein Überblick über Forschungsergebnisse und Untersuchungsmethoden. *Diagnostica, 41*(4), 303-321.

Heubrock, D., Eberl, I., & Petermann, F. (2002). Neuropsychologische Diagnostik bei Simulationsverdacht: Empirische Bewährung der Bremer Symptom-Validierung als simulationssensibles Untersuchungsverfahren. *Zeitschrift für Neuropsychologie, 13*(1), 45-58.

Hurley, K.-E., & Deal, W.-P. (2006). Assessment instruments measuring malingering used with individuals who have mental retardation: Potential problems and issues. *Mental Retardation, 44*(2), 112-119.

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.

Lamnek, S. (2005). *Qualitative Sozialforschung: Lehrbuch* (4. Aufl.). Weinheim: Beltz.

Larrabee, G.-J. (2007). Introduction: Malingering, research designs, and base rates. In G.-J. Larrabee (Ed.), *Clinical assessment of malingering and deception* (pp. 3-13). Oxford: Oxford University Press.

Lechner, D. E., Bradbury, S. F., & Bradley, L. A. (1998). Detecting sincerity of effort: A summary of methods and approaches. *Physical Therapy, 78*(8), 867-888.

Lewis, J. L., Simcox, A. M., & Berry, D. T. (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*(2), 170-176.

Mayring, P. (2003). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Weinheim: Beltz.

Mayring, P., & Gläser-Zikuda, M. (2005). *Die Praxis der qualitativen Inhaltsanalyse*. Weinheim: Beltz.

Merckelbach, H., & Smith, G. P. (2003). Diagnostic accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in detecting instructed malingering. *Archives of Clinical Neuropsychology, 18*(2), 145-152.

Merten, T. (2002). Fragen der neuropsychologischen Diagnostik bei Simulationsverdacht. *Fortschritte der Neurologie und Psychiatrie, 70*(3), 126-138.

Merten, T. (2003). Authentisch oder vorgetaeuscht? Neuropsychologische Diagnostik bei Simulationsverdacht: die Testbatterie zur Forensischen Neuropsychologie (TBFN). *Report Psychologie, 28*(4), 236-240.

Merten, T. (2005). Der Stellenwert der Symptomvalidierung in der neuropsychologischen Begutachtung - Eine Positionsbestimmung. *Zeitschrift für Neuropsychologie, 16*(1), 29-45.

Merten, T. (2008, im Druck). Negative Antwortverzerrungen in der Begutachtung. In K. D. Thomann, F. Schröter & V. Grosser (Eds.), *Orthopädisch-unfallchirurgische Begutachtung*. München: Urban & Fischer.

Merten, T., Blaskewitz, N., & Stevens, A. (2007). Kann subotimale Testmotivation mit dem Aufmerksamkeits-Belasstungs-Test (Test d2) erkannt werden? *Aktuelle Neurologie, 34*(3), 134-139.

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology, 29*(3), 308-318.

Merten, T., Friedel, E., Mehren, G., & Stevens, A. (2007). Über die Validität von Persönlichkeitsprofilen in der nervenärztlichen Begutachtung. *Nervenarzt, 78*(5), 511-512.

Merten, T., Friedel, E., & Stevens, A. (2006). Eingeschränkte Kooperativität in der neurologisch-psychiatrischen Begutachtung: Schätzungen zur Auftretenshäufigkeit an einer Begutachtungspopulation. *Versicherungsmedizin, 58*(1), 19-21.

Merten, T., Friedel, E., & Stevens, A. (2007). Die Authentizität der Beschwerdenschilderung in der neurologisch-psychiatrischen Begutachtung. Eine Untersuchung mit dem Strukturierten Fragebogen Simulierter Symptome. *Praxis der Rechtspsychologie, 17*(1), 140-154.

Merten, T., Green, P., Henry, M., Blaskewitz, 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.

Merten, T., Henry, M., & Hilsabeck, R. (2004). Symptomvalidierungstests in der neuropsychologischen Diagnostik: Eine Analogstudie. *Zeitschrift für Neuropsychologie, 15*(2), 81-90.

Merten, T., Stevens, A., & Blaskewitz, N. (2007). Beschwerdenvalidität und Begutachtung: eine Einführung. *Praxis der Rechtspsychologie, 17*(1), 7-28.

Mittenberg, W., Aguila-Puentes, G., Patton, C., Canyock, E.-M., & Heilbronner, R.-L. (2002). Neuropsychological profiling of symptom exaggeration and malingering. *Journal of Forensic Neuropsychology, 3*(1-2), 227-240.

Mittenberg, W., Patton, C., Canyock, E.-M., & Condit, D.-C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24*(8), 1094-1102.

Mossman, D. (2000a). Interpreting clinical evidence of malingering: a Bayesian perspective. *Journal of the American Academy of Psychiatry and the Law, The, 28*(3), 293-302.

Mossman, D. (2000b). The meaning of malingering data: further applications of Bayes' theorem. *Behavioral Sciences And The Law, 18*(6), 761-779.

Mossman, D. (2003). Daubert, cognitive malingering, and test accuracy. *Law and Human Behavior, 27*(3), 229-249.

Ott, W., Bade, S., & Wapf, B. (2007). *Nicht zielkonforme Leistungen in der Invalidenversicherung: Bedeutung und Grössenordnung*. Bern: Bundesamt für Sozialversicherungen.

Rogers, R. (1997). Introduction. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (pp. 1-22). New York: The Guilford Press.

Rogers, R. (Ed.). (1997). *Clinical Assessment of Malingering and Deception.* (2nd ed.). New York: The Guilford Press.

Rogers, R., & Neumann, C. S. (2003). Conceptual issues and explanatory models of malingering. In P. W. Halligan, C. Bass & D. A. Oakley (Eds.), *Malingering and illness deception* (pp. 71-82). Oxford: Oxford University Press.

Rosenhan, D. L. (1973). On being sane in insane places. *Science, 179*, 250-258.

Sass, H., Wittchen, H.-U., Zaudig, M., & Houben, I. (1998). *Diagnostische Kriterien des Diagnostischen und Statistischen Manuals Psychischer Störungen DSM-IV*. Göttingen: Hogrefe.

Schagen, S., Schmand, B., deSterke, S., & Lindeboom, J. (1997). Amsterdam short-term memory test: A new procedure for the detection of feigned memory deficits. *Journal of Clinical and Experimental Neuropsychology, 19*(1), 43-51.

Schiemann, S. (2003). Development and evaluation of a test battery for the detection of malingering. *Psychology Science, 45*(Suppl3), 80-100.

Schmand, B., Lindeboom, J., Merten, T., & Millis, S. R. (2005). AKGT; Amsterdamer Kurzzeitgedaechtnistest; Amsterdam Short-Term Memory Test. from http://www.pits-online.nl/de/AKGT.html

Schmand, B., Lindeboom, J., Schagen, S., Heijt, R., Koene, T., & Hamburger, H. L. (1998). Cognitive complaints in patients after whiplash injury: the impact of malingering. *Journal of Neurology, Neurosurgery, and Psychiatry, 64*(3), 339-343.

Schmidt-Atzert, L., Bühner, M., Rischen, S., & Warkentin, V. (2004). Erkennen von Simulation und Dissimulation im Test d2. *Diagnostica, 50*(3), 124-133.

Sharpe, M. (2003). Distinguishing malingering from psychiatric disorders. In P. W. Halligan, C. Bass & D. A. Oakley (Eds.), *Malingering and illness deception* (pp. 156-170). Oxford: Oxford University Press.

Singh, J., Avasthi, A., & Grover, S. (2007). Malingering of Psychiatric Disorder: A Review. *German Journal of Psychiatry, 2007*(10), 126-132.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *Clinical Neuropsychologist, 13*(4), 545-561.

Smith, G.-P., & Burger, G.-K. (1997). Detection of malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of the American Academy of Psychiatry and the Law, 25*(2), 183-189.

Stevens, A., Friedel, E., Mehren, G., & Merten, T. (2008). Malingering and uncooperativeness in psychiatric and psychological assessment: Prevalence and effects in a German sample of claimants. *Psychiatry Research, 157*(1-3), 191-200.

Streiner, D. L., & Norman, G. R. (2006). *Health measurement scales a practical guide to their development and use* (3rd , repr. ed.). Oxford: Oxford University Press.

Vitacco, M. J., Rogers, R., Gabel, J., & Munizza, J. (2007). An evaluation of malingering screens with competency to stand trial patients: a known-groups comparison. *Law and Human Behavior, 31*(3), 249-260.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. Chichester: Wiley.

Wapf, B., & Peters, M. (2007). *Evaluation der regionalen ärztlichen Dienste (RAD)*. Bern: Bundesamt für Sozialversicherungen.