Universitätsklinikum Ulm

Klinik für Psychiatrie und Psychotherapie I

Ärztl. Direktor: Prof. Dr. med. Wolfgang Kaschka

Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie

Dissertation

zur Erlangung des Doktorgrades der Medizin

der Medizinischen Fakultät der Universität Ulm

vorgelegt von

Lennart Kirchhoff

geboren in Bad Soden am Taunus

2016

Amtierender Dekan: Prof. Dr. T. Wirth

1. Berichterstatter: Prof. Dr. T. Steinert

2. Berichterstatter: Prof. Dr. Dr. B. Widder

Tag der Promotion: 14.06.2018

Widmung

Für meine Familie.

Inhaltsverzeichnis

Abkürzungsverzeichnis..... II

1 Einleitung 1

1.1 Sozialmedizinscher Rahmen 1

1.2 Begriffsdefinitionen 2

1.3 Auftretenshäufigkeit von negativer Antwortverzerrung bzw. Simulation 4

1.4 Beschwerdevalidierungstests 6

1.5 Bedeutung der Beschwerdevalidierung 10

1.6 Hintergrund und Fragestellungen der vorliegenden Studie..... 11

2 Material und Methoden 13

2.1 Ethische Aspekte, Aufklärung und Einwilligung, Datensicherheit 13

2.2 Patientenrekrutierung 15

2.3 Studienplan..... 17

2.4 Studienprotokoll..... 19

2.5 Untersuchungsinstrumente..... 23

2.6 Statistische Methoden 33

2.7 Patientenstichprobe..... 40

3 Ergebnisse..... 44

3.1 Auswertung der Gesamtstichprobe 44

3.2 Per Protocol Analyse 84

3.3 Einzelfallbeschreibungen..... 129

4 Diskussion 134

4.1 Diskussion der Ergebnisse 134

4.2 Diskussion der Methodik..... 152

4.3 Interpretation von WMT- und SFSS-Ergebnissen (Schlussfolgerung)..... 158

5 Zusammenfassung 160

6 Literaturverzeichnis..... 162

Anhang..... 169

Danksagung..... 185

Lebenslauf..... 186

Abkürzungsverzeichnis

Abs.	**Abs**atz
AF	Skala **af**fektiver Störungen (Strukturierter Fragebogen Simulierter Symptome)
AM	Skala **am**nestischer Beschwerden (Strukturierter Fragebogen Simulierter Symptome)
AUC	**a**rea **u**nder the **c**urve
BSI	**B**rief **S**ymptom **I**nventory
BVT(s)	**B**eschwerde**v**alidierungs**t**est(**s**)
ca.	**c**irc**a**
CNS	**C**o**ns**istency (Word Memory Test)
CRF	**C**ase **R**eport **F**orm
DGPPN	**D**eutsche **G**esellschaft für **P**sychiatrie, **P**sychotherapie und **N**ervenheilkunde
DR	**D**elayed **R**ecognition (Word Memory Test)
dt.	**d**eu**t**sch(e)
FR	**F**ree **R**ecall (Word Memory Test)
GSI	**G**lobal **S**everity **I**ndex (Brief Symptom Inventory)
IR	**I**mmediate **R**ecognition (Word Memory Test)
LDFR	**L**ong **D**elay **F**ree **R**ecall (Word Memory Test)
MC	**M**ultiple **C**hoice (Word Memory Test)
Med	**Med**ian
MW	**M**ittel**w**ert

n	Stichprobenumfang
N	Skala **n**eurologischer Beschwerden (Strukturierter Fragebogen Simulierter Symptome)
NAN	**N**ational **A**cademy of **N**europsychology
NI	Skala **n**iedriger **I**ntelligenz (Strukturierter Fragebogen Simulierter Symptome)
NPW	**n**egativer **p**rädiktiver **W**ert
Nr.	**N**umme**r**
p	Signifikanzwert
P	**P**sychoseskala (Strukturierter Fragebogen Simulierter Symptome)
PA	**P**aired **A**ssociate Recall (Word Memory Test)
PPA	**P**er **P**rotocol **A**nalyse
PPW	**p**ositiver **p**rädiktiver **W**ert
r	Korrelationskoeffizient
R^2	Bestimmheitsmaß
RCT	**r**andomized **c**ontrolled **t**rial
ROC	**r**eceiver **o**perating **c**haracteristic
S.	**S**eite
s.	**s**iehe
SCIP	**S**creen for **C**ognitive **I**mpairment in **P**sychiatry
SD	Standardabweichung (Englisch: **s**tandard **d**eviation)
SFSS	**S**trukturierter **F**ragebogen **S**imulierter **S**ymptome
SGB VI	**S**ozial**g**esetz**b**uch **VI**

sog.	**sog**enannt(e)
T-GSI	**T**-Wert **GSI**
u.	**u**nd
u.a.	**u**nter **a**nderem
v.a.	**v**or **a**llem
VAS	**V**isuelle **A**nalog**s**kala
vgl.	**v**er**gl**eiche
vs.	**v**er**s**us
VwGO	**V**er**w**altungs**g**erichts**o**rdnung
WMT	**W**ord **M**emory **T**est
z.T.	**z**um **T**eil
95% KI	**95%**iges **K**onfidenz**i**ntervall

1 Einleitung

1.1 Sozialmedizinscher Rahmen

Im Jahr 2014 machten Anträge auf Erwerbsminderungsrente laut Deutscher Rentenversicherung etwa ein Fünftel aller Rentenanträge aus (Deutsche Rentenversicherung 2015, S. 69). Auffällig ist dabei, dass in den letzten 15 Jahren vor allem der Anteil an psychischen Erkrankungen als Ursache für den vorzeitigen Berentungswunsch gestiegen ist. 2005 hatte der Anteil psychischer Erkrankungen an den gewährten Erwerbsminderungsrenten noch bei 32,3% gelegen (Deutsche Rentenversicherung 2014b). 2014 betrug der Anteil bei den Rentenzugängen wegen verminderter Erwerbsfähigkeit jedoch 49,5% aller Diagnosegruppen bei den Frauen und 36,8% bei den Männern (Deutsche Rentenversicherung 2015, S. 57). Damit bildeten psychische Erkrankungen die mit Abstand größte Diagnosegruppe. Genauer betrachtet, machten die psychiatrischen ICD-10-Diagnosen F30-48, nämlich die affektiven Störungen (F30-F39) und die neurotischen, Belastungs- und somatoformen Störungen (F40-F48), 2013 etwa die Hälfte (Männer) bzw. drei Viertel (Frauen) der bei den Rentenzugängen wegen verminderter Erwerbsfähigkeit genannten psychischen Erkrankungen aus (Deutsche Rentenversicherung 2014a, S. 48ff).

Das Sozialgesetzbuch VI (SGB VI) legt zu diesem Thema fest, dass „Versicherte […] bis zum Erreichen der Regelaltersgrenze Anspruch auf Rente wegen voller Erwerbsminderung [haben], wenn sie 1. voll erwerbsgemindert sind […]“ (§43 Abs. 2 Satz 1 SGB VI). Im Internetauftritt der Deutschen Rentenversicherung heißt es: „Die medizinischen Voraussetzungen für eine Rente wegen Erwerbsminderung sind erfüllt, wenn Sie wegen Krankheit oder Behinderung weniger als sechs Stunden täglich arbeiten können, und zwar nicht nur in Ihrem, sondern in allen Berufen. Ihre Rentenversicherung prüft das anhand ärztlicher Unterlagen. Eventuell fordert sie weitere Gutachten an und stellt dann Ihr Leistungsvermögen fest.“ (Deutsche Rentenversicherung 2016). Die Rentenversicherung behält sich also die Entscheidung vor, welchem Antragssteller eine Erwerbsminderungsrente gewährt wird (Deutsche Rentenversicherung 2016). Um die Rechtmäßigkeit des Anspruchs zu prüfen, erfolgt zunächst eine Sichtung der medizinischen Befundlage eines jeden Antragsstellers. Im Rahmen dieser Befundsichtung werden z.T. auch externe Gutachter beauftragt, welche die Erwerbsfähigkeit und

Erkrankungsschwere eines Antragsstellers in einer persönlichen Begutachtung einschätzen sollen (Deutsche Rentenversicherung 2016). Gegen einen ablehnenden Bescheid der Rentenversicherung kann vom Antragssteller Widerspruch eingelegt werden (§69 VwGO), der dann eine erneute Überprüfung der Erwerbsfähigkeit durch die Rentenversicherung und gegebenenfalls weitere Begutachtungen nach sich zieht. Im Falle einer Ablehnung des Widerspruchs, kann als letzte Möglichkeit eine Klage vor einem Sozialgericht eingereicht werden (§68 VwGO), wobei auch im Verlauf des Sozialgerichtsverfahrens wiederum Gutachter beauftragt werden, die Erwerbsfähigkeit des Klägers zu bewerten.

Vor diesem Hintergrund hat die sozialmedizinische Begutachtung durch Psychiater im Rentenverfahren einen hohen Stellenwert. Psychiatrische Diagnosen aber werden nun einmal ganz überwiegend klinisch gestellt und stützen sich, gemäß den Diagnosekriterien der ICD-10, teils auf psychopathologische Befunde, die klinisch erhoben werden können (z.B. depressive Verstimmung, Wahnvorstellungen), ebenso aber auch auf berichtete Symptome, die der Arzt nicht direkt überprüfen kann (z.B. Schlafstörungen, Ängste in bestimmten Situationen). Damit stellt sich das Problem, dass die Diagnosestellung und die Beurteilung des Beschwerdeausmaßes auch in der gutachtlichen Situation zu erheblichen Teilen auf Beschwerdeschilderungen der Antragssteller beruhen, die – z.B. im Gegensatz zu Laborwerten – kaum objektiv überprüfbar sind.

1.2 Begriffsdefinitionen

Subjektive Beschwerdeschilderungen durch den Patienten können durch Antwortverzerrungen verfälscht sein (Stevens et al. 2008). Als Antwortverzerrung wird ein Verhalten bezeichnet, das Symptome und die eigenen Fähigkeiten nicht so darstellt, wie sie in Wirklichkeit sind (Merten 2014, S. 99). Antwortverzerrungen können dabei sowohl positiv als auch negativ ausgeprägt sein, je nachdem ob Symptome heruntergespielt oder übertrieben werden (Merten 2014, S. 100). Im sozialmedizinische Begutachtungskontext sind vor allem die negativen Antwortverzerrungen von Belang, denn unter diesen Begriff fällt eine generelle Tendenz, Krankheitssymptome stärker und die eigene Leistungsfähigkeit schwächer als eigentlich gegeben darzustellen (Merten 2014, S. 100). Eine mögliche Ausprägung negativer Antwortverzerrung ist die Aggravation. Unter

Aggravation wird eine „Beschwerdenübertreibung, -überhöhung und/oder -ausweitung [verstanden], die auf eine externale Zielerreichung hin ausgerichtet ist“ (Merten 2014, S. 99), wobei unter diese externalen Ziele z.B. die Gewährung einer Erwerbsminderungsrente oder Entschädigungszahlungen fallen – externale Vorteile also, die infolge einer körperlichen oder psychischen Erkrankung zugestanden werden, allgemein auch als sekundärer Krankheitsgewinn bezeichnet (Merten 2014, S. 102).

Bezüglich des Phänomens der „suboptimale[n] Leistungsanstrengung“ (Merten 2014, S. 102) gehen Blaskewitz und Merten davon aus, dass auch sie „weitergefasst […] als negative Antwortverzerrung bezeichnet werden kann“ (Blaskewitz u. Merten 2007, S. 141). Suboptimale Leistungsanstrengung liegt z.B. dann vor, wenn ein Patient im Rahmen einer sozialmedizinischen Begutachtung absichtlich Ergebnisse unterhalb seines wahren Leistungsniveaus erzielt (Merten 2014, S. 102). Dies kann jegliche Art von Leistungstests betreffen, z.B. Tests für Gedächtnis, Konzentration, oder Arbeitsgeschwindigkeit.

Die dritte mögliche Ausprägung negativer Antwortverzerrung ist die sogenannte „Verdeutlichungstendenz“ (Dohrenbusch 2009, S. 232). Sie wird als „unwillkürlich übertreibende Darstellung gesundheitlicher Beschwerden aus dem Motiv heraus, den Untersucher von der Schwere der Erkrankung zu überzeugen“ (Dohrenbusch 2009, S. 231) beschrieben. Verdeutlichungstendenzen sind also v.a. in ihrer fehlenden Vorsätzlichkeit und vorherrschenden Unbewusstheit des Handelns von den anderen beiden Formen negativer Antwortverzerrung zu unterscheiden. In einer Gutachtensituation dürfte es aber sicherlich zu einer gewissen Vermischung der beschriebenen antwortverzerrenden Phänomene kommen.

Der Begriff der Simulation hingegen muss von den vorangegangenen Definitionen – trotz stets anzunehmender fließendender Übergänge – abgegrenzt werden, denn Simulation ist laut Merten als „auf einen externalen Krankheitsgewinn ausgerichtetes, […] bewusstes Vortäuschen einer Gesundheitsstörung“ (Merten 2014, S. 102) aufzufassen. Hervorzuheben in Abgrenzung zu den Ausprägungen negativer Antwortverzerrung ist bei der Simulation also v.a. die Absicht, nicht vorhandene Symptome zu einem bestimmten Zweck vorzutäuschen, während z.B. bei der Aggravation tatsächlich vorhandene Beschwerden zu einem bestimmten Zweck übertrieben werden (Dohrenbusch 2009).

Aggravation, suboptimale Leistungsanstrengung und Verdeutlichung als

Ausprägungen negativer Antwortverzerrung sowie auch eine Simulation von Krankheitssymptomen können also theoretisch im Verlauf einer Begutachtung auftreten. Man geht allerdings davon aus, dass plumpe Simulationsversuche eher selten sind (Merten et al. 2007), da es als unwahrscheinlich erachtet wird, einen klinisch erfahrenen Gutachter mit dem Vorspielen von Symptomen täuschen zu können. Häufiger dürfte es hingegen zu einer Aggravation oder Verdeutlichung der geschilderten – und per Definition auch tatsächlich vorhandenen – Beschwerden kommen.

1.3 Auftretenshäufigkeit von negativer Antwortverzerrung bzw. Simulation

Mehrere Studien haben zu eruieren versucht, wie häufig negative Antwortverzerrung bzw. Simulation in verschiedenen sozialmedizinischen Begutachtungskontexten tatsächlich ist. So stellten Stevens et al. nach retrospektiver Auswertung von 233 deutschen Begutachtungsfällen mit sozial- oder zivilrechtlichem Hintergrund bei 44,6% der Untersuchten Hinweise auf suboptimale Leistungsanstrengung fest (Stevens et al. 2008). In diesem Zusammenhang wird häufig auch eine multizentrische amerikanische Studie von Mittenberg et al. zitiert, die verschiedene Diagnosegruppen und Gutachtenkontexte sowie eine große Fallzahl (33.531 Personen) umfasst (Mittenberg et al. 2002). Darin beschreiben Mittenberg et al. ein Auftreten von mutmaßlicher Aggravation, Simulation oder Verdeutlichung bei 30% der Berufsunfähigkeitsverfahren, bei 29% der Entschädigungskläger und weiterhin einen Prozentsatz von 19% bei strafrechtlichen Verfahren. Bei gewöhnlichen medizinischen oder psychiatrischen Begutachtungen zeigten lediglich 8% der Patienten ein auffälliges Antwortverhalten (Mittenberg et al. 2002). Kritisch anzumerken ist sowohl zur Studie von Mittenberg et al., als auch zur Studie von Stevens et al., dass es sich lediglich um grobe Schätzungen der Auftretenshäufigkeit von negativer Antwortverzerrung bzw. Simulation handeln kann. Beide Studienpopulationen setzten sich schließlich aus medizinisch-rechtlichen Gutachtenfällen zusammen, deren Beschwerdepräsentation im Verlauf der Begutachtung auf verschiedene Art und Weise (z.B. durch neuropsychologische Tests oder kritisches Hinterfragen der Anamnese etc.) hinsichtlich Aufrichtigkeit und Plausibilität evaluiert wurde (Mittenberg et al. 2002;

Stevens et al. 2008). Es ist allerdings davon auszugehen, dass keiner der Begutachteten im Anschluss an das Gutachten offen zugab, Beschwerden aggraviert, verdeutlicht oder simuliert zu haben. Daher lässt sich nicht endgültig nachprüfen, inwieweit die verwendeten Methoden zur Evaluation der Beschwerdedarstellung mit ihrer Einschätzung der Symptompräsentation des Begutachteten richtig lagen.

Dennoch erklärt z.B. auch ein Positionspapier der amerikanischen National Academy of Neuropsychology (NAN), dass negative Antwortverzerrung bzw. Simulation bei einer nicht unerheblichen Minderheit neuropsychologisch untersuchter Patienten auftritt (Bush et al. 2005) – was angesichts des bei medizinisch-rechtlichen Kontexten im Raum stehenden sekundären Krankheitsgewinns durchaus plausibel ist. Van Egmond et al. stellten beispielsweise an einer Population von 99 ambulanten Psychiatriepatienten bei 41,4% eine Erwartungshaltung bezüglich eines sekundären Krankheitsgewinns fest (van Egmond et al. 2005) – ein Anteil der ungefähr im Bereich der zuvor aufgeführten Prävalenzen von negativer Antwortverzerrung bzw. Simulation liegt. Der Schluss liegt somit nicht fern, dass der Versuch der Erzielung eines sekundären Krankheitsgewinns und negative Antwortverzerrung oder Simulation als Mittel zum Zweck vergesellschaftet sein könnten.

Die Absicht einer negativen Antwortverzerrung oder gar Simulation als solche entzieht sich aber weitgehend der direkten Erfassung, weil sie aus naheliegenden Gründen nicht eingeräumt wird. Es bietet sich daher der Ansatz an, stattdessen das Mittel zum Zweck, also die negative Antwortverzerrung oder Simulation selbst, zu erfassen. Dies ist z.B. durch eine ausgiebige und kritische Prüfung aller Befunde, Testergebnisse und Diagnosen eines Patienten hinsichtlich Plausibilität und Konsistenz möglich (Merten 2014, S. 23). Natürlich spielen bei einer solchen Einschätzung dann auch die klinische Erfahrung und die intuitive Einschätzung des Gutachters eine zentrale Rolle. Doch klinische Erfahrung variiert stark von Gutachter zu Gutachter, und auch die sehr subjektive Intuition kann nicht ernsthaft als Entscheidungskriterium in medizinisch-rechtlichen Fragen angesehen werden, zumal sie vermutlich mit Einflussfaktoren wie Sympathie, Empathiefähigkeit, Überzeugungskraft des Beschwerdevortrags, affektiven Ausdrucksmöglichkeiten und anderem mehr assoziiert ist.

1.4 Beschwerdevalidierungstests

1.4.1 Tests und Methoden

Um dem Begutachtungsverfahren mehr Transparenz und Objektivität zu verleihen, wurden v.a. gegen Ende der 80er- bzw. Anfang der 90er-Jahre (Merten 2014, S. 1) vermehrt sogenannte Beschwerdevalidierungstests (BVTs) entwickelt. Mittels spezifischer Methoden sollen BVTs „untersuchen, ob die […] in neuropsychologischen Testuntersuchungen ermittelten Leistungsprofile als authentisch oder verzerrt zu beurteilen sind“ (Merten 2014, S. 1). Mittlerweile gelten viele Tests als gut validiert, sodass das bereits erwähnte Positionspapier der NAN fordert, Ergebnissen von BVTs generell deutlich mehr Gewicht beizumessen als subjektiven Indikatoren suboptimaler Leistungsanstrengung (Bush et al. 2005), wie z.B. der bereits erwähnten Intuition.

BVTs als Oberbegriff sind laut Merten nochmals zu unterteilen in „kognitive BVT[s]“ (Merten 2014, S. 101), die vor allem suboptimale Leistungsanstrengung erfassen sollen, und „psychologische BVT[s]“ (Merten 2014, S. 101), welche die Validität der geschilderten Beschwerden überprüfen sollen (Merten 2014, S. 101). Kognitive BVTs nutzen zur Erfassung von negativer Antwortverzerrung bzw. Simulation u.a. das „Prinzip der verdeckten Leichtigkeit“ (Merten 2014, S. 24). Dabei wird ein Test, z.B. in der Testanleitung, als schwierig dargestellt, am Ende ist er aber einfacher als zunächst vorgegeben (Merten 2014, S. 26). Ein Gutachtenpatient, der dem Untersucher seine kognitiven Probleme besonders deutlich machen möchte, wird – so die Theorie – die eigentliche Schwierigkeit des Tests überschätzen (Merten 2014, S. 47) und vielleicht Fehler bzw. Schwächen präsentieren, die Vergleichspopulationen auch mit tatsächlichen kognitiven Beeinträchtigungen nicht zeigen – und somit ein auffälliges Ergebnis erzielen.

Eine weitere wichtige Methode kognitiver BVTs ist das „Alternativwahlverfahren“ (Merten 2014, S. 31), auch „Zwangswahlverfahren“ (Blaskewitz u. Merten 2007, S. 142) genannt. Bei Tests dieser Art werden, nachdem zunächst z.B. eine zu erlernende Wortliste präsentiert wurde, bei der späteren Abfrage der Liste immer zwei Antwortmöglichkeiten angeboten, eine richtige und eine falsche (Merten 2014, S. 31). Es besteht also rein rechnerisch eine 50%ige Wahrscheinlichkeit die richtige Antwort auszuwählen, und zwar auch ohne die Wortliste je gesehen zu haben (Merten 2014, S. 31). Wenn also beispielsweise 100 Testpersonen 100

Alternativentscheidungen bezüglich einer von ihnen nie gesehenen Wortliste treffen müssten, würde ihr Ergebnis um einen Wert von 50 richtigen Antworten (50%) pendeln (Merten 2014, S. 32). Mithilfe statistischer Methoden lässt sich für jede Prozentzahl richtiger Antworten berechnen, wie wahrscheinlich ein solches Ergebnis bei 100 Alternativentscheidungen ist (Merten 2014, S. 32). Ab einem gewissen Wert unter 50% richtigen Antworten ist die Wahrscheinlichkeit, das Ergebnis durch Raten erzielt zu haben, dann so gering, dass es durch tatsächliches Raten eigentlich nicht entstanden sein kann (Wahrscheinlichkeiten von 0,044 abwärts ab einem Ergebnis von ≤41% (einseitige Fragestellung); Merten 2014, S. 32). Diesen Wert bezeichnet man als „Schwelle für Unter-Zufall-Antworten“ (Merten 2014, S. 32). Wird er unterschritten, ist absichtliche Antwortmanipulation die wahrscheinlichste Erklärung für das schlechte Ergebnis (Merten 2014, S. 32). Allerdings sollte auch ein Ergebnis von 50% richtigen Antworten stutzig machen, denn es würde im Prinzip bedeuten, dass die Testperson praktisch über gar kein Gedächtnis verfügt bzw. dass dasselbe nicht besser ist als der Zufall.

Aufschlussreich kann bei kognitiven BVTs auch ein Blick auf die Leistungskurve des Gutachtenpatienten sein (Merten 2014, S. 26), v.a. dann im Vergleich zu normierten Vergleichspopulationen. Normalerweise dürfte ein kognitiv tatsächlich leicht beeinträchtigter Patient bei leichten Aufgaben keine oder geringe Abweichungen von einer gesunden Vergleichsgruppe zeigen, bei schwereren Aufgaben aber deutlich schwächer abschneiden. Verhält es sich umgekehrt, kann dies ein Hinweis auf absichtliche Antwortmanipulation sein (Merten 2014, S. 26).

Psychologische BVTs ähneln z.T. klassischen Symptomfragebögen, d.h. sie beschreiben verschiedene Symptome und der Patient soll dann das entsprechend für ihn Zutreffende selbstständig ankreuzen. Um Hinweise auf negative Antwortverzerrung oder Simulation liefern zu können, benennen psychologische BVTs – wie z.B. das Structured Inventory of Malingered Symptomatology (SIMS (Smith u. Burger 1997)) – aber vor allem atypische und groteske Symptome (Giger u. Merten 2013), die ein tatsächlich erkrankter Patient so nicht berichten würde. Von medizinisch ungebildeten negativen Antwortverzerrern bzw. Simulanten werden die atypischen Symptome jedoch als zur Erkrankung, z.B. zu einer Depression, gehörig gewertet. Sie kennen sich mit der Erkrankung schließlich

nicht wirklich aus, so die Idee hinter der Methode (Jelicic et al. 2011). Letztlich kommt es daher zu einer auffälligen und unglaubwürdigen Symptomdarstellung durch den negativen Antwortverzerrer bzw. Simulanten (Jelicic et al. 2011). Sowohl bei psychologischen als auch bei kognitiven BVTs ist, als weiterer Hinweis auf negative Antwortverzerrung bzw. Simulation, eine Konsistenzprüfung des Antwortverhaltens möglich, z.B. durch Wiederholung der Tests (Merten 2014, S. 27). Die Frage lautet dabei ganz einfach, ob sich die erzielten Testprofile bei Wiederholung ähneln oder ob Unregelmäßigkeiten auffallen.
Häufig angewandte psychologische BVTs sind zum Beispiel das bereits erwähnte Structured Inventory of Malingered Symptomatology (SIMS) bzw. seine deutsche Version, der Strukturierte Fragebogen Simulierter Symptome (SFSS (Cima et al. 2003)). Ein verbreitetes Beispiel für die Klasse der kognitiven BVTs ist der Word Memory Test (WMT (Green 2005)).

1.4.2 Cut-off

Betrachtet man die Auswertung von BVTs genauer, so ist bei den meisten Tests als zentrales Entscheidungskriterium für oder gegen eine auffällige Beschwerdepräsentation ein sogenannter „Cut-off“ (Clegg et al. 2009, S. 250) festgelegt. Dabei handelt es sich um einen statistisch festgelegten Testpunktwert, bei welchem sich für den jeweiligen Test das günstigste Verhältnis von Sensitivität und Spezifität ergab (Gaus u. Muche 2014, S. 142). Unter der Sensitivität des Tests versteht man in diesem Zusammenhang den Anteil der durch den BVT erkannten antwortverzerrenden Patienten von der Gesamtmenge der antwortverzerrenden Patienten. Eine hohe Sensitivität bedeutet also, dass viele der tatsächlich antwortverzerrenden Patienten vom BVT erkannt werden. Die Spezifität eines BVTs beschreibt den Anteil der durch den Test als Nicht-Antwortverzerrer erkannten Patienten von der Gesamtmenge der Nicht-Antwortverzerrer. Ein idealer BVT mit Werten von 100% für Spezifität und Sensitivität klassifiziert also alle Testpersonen richtig als Antwortverzerrer bzw. Nicht-Antwortverzerrer. Bei der Festlegung des Cut-offs ergibt sich meist die Frage, ob eher eine höhere Sensitivität oder eine höhere Spezifität erwünscht ist, da sich beide Werte gegenläufig verhalten und eine Erhöhung des einen Wertes meist nur zum Preis einer Erniedrigung des anderen möglich ist (Gaus u. Muche 2014, S. 143). Laut Merten wird für BVTs aber „in aller Regel eine Minimierung

von falsch-positiven Ergebnissen angestrebt, sodass eine Verringerung der Sensitivität zugunsten einer höheren Spezifität […] in Kauf genommen wird.“ (Merten 2014, S. 69).

Wird der festgelegte Cut-off unter- bzw. überschritten, je nachdem was für eine Art von Gesamtergebnis der Test liefert (z.B. Prozentwert bzw. Gesamtpunktzahl), gilt die Beschwerdedarstellung durch den Patienten – zumindest in dieser Situation (Merten 2014, S. 79) – als unzuverlässig. Dem persönlichen Ermessen des Gutachters obliegt es dann, weitere Schlüsse aus dieser Feststellung zu ziehen und auch die Tragweite eines auffälligen BVT-Ergebnisses einzuordnen. In der Literatur reichen die Meinungen zur richtigen Interpretation von BVT-Ergebnissen von der Ansicht, dass BVTs „flankierende Informationen zum Grad der Glaubwürdigkeit und Kooperativität bei einem Probanden“ (Blaskewitz u. Merten 2007, S. 147) liefern, bis hin zur Empfehlung, die gesamte Beschwerdepräsentation und sogar Lebensgeschichte eines Patienten anzuzweifeln, wenn negative Antwortverzerrung festgestellt wurde (Merten u. Merckelbach 2013).

1.4.3 Falsch-positive Klassifikationen

Falsch-positive Klassifikationen durch BVTs sind in der Beschwerdevalidierung ein viel diskutiertes Thema. Eine falsch-positive Klassifikation liegt vor, wenn ein Patient in einem BVT ein auffälliges bzw. unglaubwürdiges Ergebnis erzielt, das nicht auf negative Antwortverzerrung oder Simulation, sondern auf eine Erkrankung zurückzuführen ist, die mit der Testdurchführung interferiert (Merten 2014, S. 71). Bei kognitiven BVTs wäre es beispielsweise denkbar, dass Patienten mit tatsächlich vorliegenden, schwerwiegenden neurologischen oder psychischen Erkrankungen in den Tests – trotz ausreichender Leistungsanstrengung – ein Ergebnis unterhalb des Cut-offs erzielen. Green et al. beschrieben dieses Problem z.B. bei einer Population von Patienten mit Demenz (Green et al. 2011). Neben einer fortgeschrittenen Demenzerkrankung wäre ebenso eine schwere depressive Episode als Ursache für eine falsch-positive Klassifikation denkbar, da bei schwer depressiven Patienten häufig auch kognitive Defizite auftreten (Veiel 1997). Ferner können psychische Erkrankungen wie Schizophrenie und schwere depressive Episoden mit einer Schwächung des Antriebs und damit der Leistungsbereitschaft einhergehen, sodass in solchen Fällen der Nachweis einer mangelnden

Leistungsbereitschaft in der Testsituation nicht zwingend mit Aggravation, Verdeutlichung oder Simulation in Verbindung stehen muss. Vorkommen können falsch-positive Klassifikationen natürlich auch bei der Verwendung von psychologischen BVTs (van Impelen et al. 2014).

Um falsch-positive Klassifikationen zu vermeiden, versucht man beispielsweise, die kognitiven BVTs so einfach zu gestalten, dass selbst kognitiv stark beeinträchtigte Patienten noch ein passables Ergebnis erzielen können (z.B. auch Patienten mit hippokampaler Atrophie; Goodrich-Hunsaker u. Hopkins 2009). Eine gewisse Leistungsfähigkeit ist für die Tests aber dennoch notwendig (Carone 2014), sodass bei kognitiv stark beeinträchtigten Patienten stets die Möglichkeit einer falsch-positiven Klassifikation bedacht werden sollte.

1.5 Bedeutung der Beschwerdevalidierung

Zahlreiche Befürworter von BVTs sind der Meinung, dass Hinweise auf negative Antwortverzerrung oder Simulation bei jeder Art von Befragung mittels BVTs erfasst werden sollten – z.B. durch die Überprüfung der Leistungsanstrengung – um überhaupt valide Ergebnisse erhalten zu können (Green et al. 2001). Diese Forderung stützt sich auf Studien wie die von Green et al., bei denen der Einfluss der Leistungsanstrengung auf die Ergebnisse einer Testbatterie untersucht wurde (Green et al. 2001). Hierfür überprüften die Autoren die Leistungsanstrengung mittels BVTs (u.a. WMT) an 904 Patienten und fanden heraus, dass die Leistungsanstrengung als stärkster Einflussfaktor zwischen 49% und 54% der Gesamtvarianz aller Ergebnisse der Testbatterie erklärte (Green et al. 2001). Der Bildungsstand erklärte zum Vergleich gerade einmal 11%, das Lebensalter 4% und das Geschlecht <1% der Varianz (Green et al. 2001). Green et al. sind infolgedessen der Meinung, ein solch wichtiger Einflussfaktor wie die Leistungsanstrengung müsse stets erfasst werden, weshalb sie die ubiquitäre Verwendung von BVTs propagieren (Green et al. 2001).

Infolge solcher Feststellungen forderten auch Bush et al. von der NAN in ihrem Positionspapier von 2005, dass zur sinnvollen Interpretation der Ergebnisse neuropsychologischer Tests eine adäquate Erfassung der Antwortvalidität essenziell sei (Bush et al. 2005). Mit besonderer Bezugnahme auf den Gutachtenkontext bezeichnet auch Merten die „adäquate Diagnostik der Validität geschilderter Beschwerden“ (Merten et al. 2007, S. 511) als „Grundvoraussetzung

für ein qualifiziertes Sachverständigenurteil“ (Merten et al. 2007, S. 511). Der Gutachter handelt seiner Argumentation nach nicht im Auftrag des Patienten, wie er es sonst als therapeutisch tätiger Arzt oder Psychologe tut, sondern „im Auftrag der Gemeinschaft“ (Merten et al. 2007, S. 511) und dieser ist er die Prüfung der Symptomdarstellung auf Richtigkeit schuldig.

1.6 Hintergrund und Fragestellungen der vorliegenden Studie

Begleitforschung zu BVTs wurde bisher hauptsächlich im angloamerikanischen Sprachraum durchgeführt. Vielen Untersuchungen ist dabei gemeinsam, dass experimentelle Simulationsgruppen verwendet wurden, die – im Gegensatz zur Kontrollgruppe – eine starke und eindeutige Simulationsinstruktion erhielten (Tan et al. 2002; Merckelbach u. Smith 2003; Cima et al. 2003). Unter diesen Voraussetzungen ergaben sich hinreichend gute Werte für Sensitivität und Spezifität der Instrumente (Merckelbach u. Smith 2003; Cima et al. 2003). Es ist allerdings kritisch anzumerken, dass dieser Versuchsaufbau nicht der klinischen Realität entspricht, denn diese ist komplizierter. Typische Rentenantragsteller mit psychischen Beschwerden haben tatsächlich Symptome, neigen aber – so lassen es die Studien von Stevens et al. (Stevens et al. 2008) und Mittenberg et al. (Mittenberg et al. 2002) zumindest vermuten – z.T. zur Verdeutlichung, Aggravation oder Simulation von Beschwerden, bzw. zu suboptimaler Leistungsanstrengung. Da reine Simulation als eher selten gilt (Merten et al. 2007), handelt es sich also keineswegs um klar trennbare Zustände. Am häufigsten dürfte eine Mischung aus realen Beschwerden und einem gewissen Ausmaß negativer Antwortverzerrung in der Gutachtensituation sein. Ob BVTs geeignet sind, auch eine solche, eher mäßige Aggravation, Verdeutlichung oder suboptimale Leistungsanstrengung – bei gleichzeitigem Vorhandensein tatsächlicher Krankheitssymptome – zuverlässig zu identifizieren, wurde bisher noch nicht eingehender untersucht.

Einige BVT-Studien stützen sich zudem hauptsächlich auf Studentenpopulationen (z.B. Merckelbach u. Smith 2003; Cima et al. 2003) anstelle der realitätsnäheren, aber schwerer zu gewinnenden Patientenpopulationen. In der vorliegenden Studie soll deshalb die Leistungsfähigkeit bzw. die Klassifikationsgüte der beiden weit verbreiteten BVTs Word Memory Test (WMT) und Strukturierter Fragebogen Simulierter Symptome (SFSS) mit einer wesentlich praxisnäheren Instruktion an

Testpersonen mit tatsächlich vorhandenen Symptomen (stationäre und teilstationäre Patienten eines psychiatrischen Krankenhauses) untersucht werden. Folgende Fragestellungen sollen beantwortet werden:

1. Wie gut sind die Testverfahren WMT und SFSS in der Lage, experimentelle negative Antwortverzerrer von Studienteilnehmern mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten zu unterscheiden? Diese Einschätzung soll über die Bestimmung der Kennwerte Sensitivität, Spezifität, positiver und negativer prädiktiver Wert erfolgen.
2. Wie groß ist die Übereinstimmung von WMT und SFSS bezüglich ihrer Klassifikationen?
3. Führt die kombinierte Anwendung von WMT und SFSS zu besseren Klassifikationsergebnissen?
4. Besteht ein Zusammenhang zwischen der Ausprägung der klinischen Symptomatik bzw. dem kognitiven Leistungsniveau und den BVT-Ergebnissen?

2 Material und Methoden

2.1 Ethische Aspekte, Aufklärung und Einwilligung, Datensicherheit

Vor Beginn der Datenerhebung wurden das Studiendesign, das Patienteninformationsblatt (inklusive Einwilligungserklärung, Anhang A) und alle weiteren wichtigen Rahmenbedingungen der Studie gemäß der Prinzipien der Deklaration von Helsinki (World Medical Association 2001) der Ethikkommission der Universität Ulm in Form eines Ethikantrags zur Prüfung vorgelegt. Mit Schreiben vom 22.07.2015 (Antragsnummer: 217/15) wurde die Studie von der Ethikkommission der Universität Ulm genehmigt.

Mittels des Informationsblattes (Anhang A) wurden die teilnehmenden Patienten über das Untersuchungsprogramm sowie über die Notwendigkeit und die Risiken der Studie informiert. Das Informationsblatt wurde bezüglich der Beschreibung der verwendeten BVTs absichtlich vage gehalten, um den Studienteilnehmern vorab nicht zu viele Informationen über die Funktionsweise der Tests zu verraten. Dies hätte die Testergebnisse sonst in unerwünschter Weise verfälschen können (Bauer u. McCaffrey 2006).

Alle Patienten stimmten der Studienteilnahme in Form einer schriftlichen Einwilligungserklärung zu, sodass insgesamt der Forderung der CONSORT-Leitlinien (vgl. CONSORT 2010 (Moher et al. 2010)) nach „informed consent“ (Moher et al. 2010, S. e15) nachgekommen wurde. Die Befragung erfolgte nur, wenn die Patienten einwilligungsfähig waren. Mit Unterzeichnung der Einwilligungserklärung gaben die Patienten ihre Daten für eine anonyme Auswertung zu wissenschaftlichen Zwecken frei.

Um die Persönlichkeitsrechte der Patienten zu wahren, wurden die Patientendaten pseudonymisiert erhoben. Dazu wurde jedem Fall eine Fall-ID in Form eines dreistelligen Buchstabencodes zugewiesen. Namen wurden nicht erfasst. Die studienbezogenen Daten wurden über Fragebögen bzw. beim WMT über ein Computerprogramm erfasst. Die Fragebögen und das Computerprogramm wurden jeweils mit der Fall-ID gekennzeichnet. Die erhobenen Daten gelangten nicht in die elektronische Krankenakte und wurden den behandelnden Ärzten auch einzelfallbezogen nicht mitgeteilt. Von den Fragebögen bzw. vom WMT-Computerprogramm wurden die Daten in pseudonymisierter Form in Microsoft®

Excel® 2010-Dateien übertragen. Um spätere Überprüfungen vornehmen zu können, werden die Daten für 10 Jahre auf einem Server der Forschung in der Klinik gespeichert, bevor sie gelöscht werden.

Es handelte sich um ein Projekt, welches unter Verantwortung des Abteilungsleiters durchgeführt wurde, sodass die allgemeinen Haftungsgrundsätze galten. Als Gratifikation für ihre Teilnahme erhielten die Patienten eine Entschädigung in Höhe von 15 €.

2.2 Patientenrekrutierung

2.2.1 Einschlusskriterien

In die Studie eingeschlossen wurden Patienten mit folgenden Hauptdiagnosen bei Aufnahme:

1. leichte, mittelgradige oder schwere depressive Episode bzw. rezidivierende depressive Störung, gegenwärtig leichte, mittelgradige oder schwere Episode (ICD-10 F32.0, F32.1, F32.2, F32.3 bzw. F33.0, F33.1, F33.2, F33.3). Patienten mit der Aufnahmediagnose einer schweren Depression (F32.2, F32.3, F33.2, F33.3) wurden nur bei Besserung auf eine höchstens mittelschwere depressive Symptomatik in die Studie aufgenommen. Dazu wurde jeweils vor der Versuchsdurchführung Rücksprache mit Therapeuten und/oder Pflegepersonal gehalten.

2. Dysthymia (ICD-10 F34.1)

3. Anpassungsstörungen (ICD-10 F43.2)

4. Phobische Störungen (ICD-10 F40.x)

5. Andere Angststörungen (ICD-10 F41.x)

Es wurden nur Patienten rekrutiert, die sich zum Zeitpunkt der Datenerhebung in stationärer oder teilstationärer (= ambulanter) Behandlung befanden. Ferner wurden ausschließlich Patienten im Alter zwischen 40 und 63 Jahren in die Studie aufgenommen, da diese Altersgruppe in der Gruppe der psychischen Erkrankungen (F00 - F99) den bei weitem größten Anteil bei den Rentenzugängen wegen verminderter Erwerbsfähigkeit bildet (Deutsche Rentenversicherung 2014a, S. 48ff). Die Unterzeichnung der Einwilligungserklärung war ebenfalls Vorrausetzung für die Teilnahme.

2.2.2 Ausschlusskriterien

Von der Studie ausgeschlossen waren Patienten mit der Nebendiagnose einer hirnorganischen Erkrankung (ICD-10 F0x) oder Intelligenzminderung (ICD-10 F7x) sowie Patienten, die Deutsch nicht als Muttersprache erlernt hatten, denn es sollten nur die deutschen Versionen von WMT und SIMS (= SFSS) eingesetzt werden. Auch Patienten mit einem laufenden Rentenverfahren wurden von der Studie ausgeschlossen.

Die Patientenrekrutierung erstreckte sich vom 17.08.2015 bis zum 11.03.2016. Insgesamt wurden 60 Patienten in die Studie eingeschlossen. Zum Zeitpunkt der Datenerhebung befanden sich 55% (33 von 60 Patienten) der Patienten in teilstationärer (= ambulanter) und 45% (27 von 60 Patienten) in stationärer Behandlung in der Klinik für Psychiatrie und Psychotherapie I (Weissenau) des Universitätsklinikums Ulm. Die 60 Patienten wurden in den unterschiedlichen Abteilungen des psychiatrischen Krankenhauses rekrutiert. 30% (18 von 60 Patienten) der Patienten entstammten der Depressionsstation (Station 2011), 10% (6 von 60 Patienten) wurden in der an Station 2011 angeschlossenen Tagesklinik Depression (Station 2011-TK) rekrutiert, 15% (9 von 60 Patienten) entfielen auf die Stationen der Akutpsychiatrie (Stationen 2053, 2054, 2056), 28,3% (17 von 60 Patienten) befanden sich zum Zeitpunkt der Datenerhebung in ambulanter Behandlung in der SINOVA-Tagesklinik Ravensburg (Station 2059) und die restlichen 16,7% (10 von 60 Patienten) wurden in der Akuttagesklinik (Station 2021) der Klinik für Psychiatrie und Psychotherapie I rekrutiert. Die Datenerhebung fand jeweils in den Räumlichkeiten der oben genannten Stationen statt.

2.3 Studienplan

2.3.1 Studiendesign

Die vorliegende Studie zur Klassifikationsgüte von WMT und SFSS wurde in Form einer randomisierten kontrollierten Studie (randomized controlled trial, RCT) an Patienten der Klinik für Psychiatrie und Psychotherapie I (Weissenau) des Universitätsklinikums Ulm durchgeführt. Die Form eines RCTs wurde gewählt, da sich die Untersuchungsgruppen bei diesem Studiendesign im Idealfall nur hinsichtlich der Intervention (in diesem Fall die Testinstruktion) unterscheiden (Gaus u. Muche 2014, S. 50). Bezüglich der Störgrößen unterscheiden sich die Untersuchungsgruppen eines RCTs im Zuge der Randomisierung bestenfalls kaum und höchstens zufällig (Gaus u. Muche 2014, S. 50). Damit sind Gruppenunterschiede bei den Zielgrößen mit hoher Wahrscheinlichkeit auf die durchgeführte Intervention zurückzuführen und nicht auf die Störgrößen, was mit einer hohen internen Validität der Studienergebnisse gleichbedeutend ist (Gaus u. Muche 2014, S. 50).

Das RCT wurde in Form einer einmaligen Querschnittsuntersuchung mit Parallelgruppendesign durchgeführt. Eine verbundene Versuchsanordnung wäre, aufgrund des mit der ersten Testdurchführung erworbenen Vorwissens der Patienten, nicht möglich gewesen. Die Studienteilnehmer wurden im Verlauf der Versuchsdurchführung über ein Zufallsverfahren (= Randomisierung) in eine Interventions- bzw. Kontrollgruppe eingeteilt. Die beiden Gruppen erhielten jeweils unterschiedliche Instruktionen (s. S. 20f) bezüglich der Bearbeitung der BVTs. Bei den unterschiedlichen Instruktionen handelte es sich somit um die zwischen den Gruppen variierte Einflussgröße der vorliegenden Studie, deren Auswirkung auf die Zielgrößen (WMT- und SFSS-Ergebnisse) untersucht werden sollte (Gaus u. Muche 2014, S. 88). Die Interventionsgruppe wurde im Rahmen ihrer Instruktion gebeten, sich in die Situation eines Rentenbewerbers hineinzuversetzen, der befürchtet, dass der Gutachter ihm die Beschwerden nicht hinreichend glaubt und der seine Beschwerden deshalb besonders deutlich machen möchte. Bei der Interventionsgruppe handelte es sich also um eine Gruppe experimenteller negativer Antwortverzerrer. Die Kontrollgruppe wurde gebeten, die Tests so gut wie möglich, also mit entsprechender Anstrengung, zu bearbeiten. Um zu kontrollieren, ob die Instruktionen in den Gruppen wirksam waren, wurde im

Anschluss an die Versuchsdurchführung die subjektive Anstrengung eines jeden Patienten mittels einer visuellen Analogskala (VAS) ermittelt. Die Patienten sollten dazu auf der VAS einschätzen, wieviel Prozent ihres tatsächlichen Leistungsvermögens sie bei der Testdurchführung von WMT und SFSS gezeigt hatten.

Die Datenerhebung wurde in einfacher Verblindung auf Seiten des Versuchsleiters durchgeführt, um eine unlautere Einflussnahme des Versuchsleiters auf die Testergebnisse zu verhindern. Eine Doppelblindstudie mit zusätzlicher Verblindung der Patienten wäre bei der vorliegenden Studie nicht möglich gewesen. Entblindet wurde direkt nach Abschluss der Datenerhebung am jeweiligen Patienten.

2.3.2 Stichprobengröße

60 Patienten wurden in die Studie eingeschlossen. Da es nicht um den Nachweis klinisch relevanter Unterschiede zwischen Behandlungsgruppen ging, wurde im Vorfeld weder eine Fallzahlabschätzung noch eine Powerkalkulation durchgeführt. Die Fallzahl wurde so gewählt, dass davon auszugehen war, dass sie für statistische Mittelwertvergleiche zwischen den Gruppen ausreicht und die Kennwerte Sensitivität, Spezifität, PPW und NPW berechnet werden können.

2.3.3 Randomisierungsverfahren

Um die Störgrößen gleichmäßig auf beide Gruppen zu verteilen und um eine Einflussnahme des Versuchsleiters auf die Gruppenbesetzung zu verhindern, wurden im Zuge der Randomisierung 60 geschlossene, unbeschriftete Umschläge mit den Instruktionen A und B (je Instruktion 30 Umschläge) befüllt, in einem blickdichten Beutel verwahrt und anschließend von einer dritten Person durchmischt. Direkt vor Durchführung der Beschwerdevalidierungstests WMT und SFSS zog der Studienteilnehmer selbst einen dieser Umschläge aus dem Beutel, öffnete ihn in Abwesenheit des Versuchsleiters und erhielt somit seine Instruktion. Bei dem beschriebenen Vorgehen handelt es sich um eine einfache Randomisierung mit einer 50%igen Zuteilungswahrscheinlichkeit für jede der beiden Studiengruppen (Moher et al. 2010). Zum Zeitpunkt der Randomisierung waren die Patienten bereits unwiderruflich in die Studie eingeschlossen.

2.4 Studienprotokoll

Anhand des folgenden in ähnlicher Form verwendeten Studienprotokolls – hier mit Anmerkungen und Erklärungen – lässt sich die standardisierte Versuchsdurchführung an jedem der 60 Patienten nachvollziehen:

1. **Rekrutierung:** Potentielle Studienteilnehmer wurden, entsprechend der Ein- und Ausschlusskriterien, über die digitalisierten Patientenakten am Computer vorausgewählt und nach Rücksprache mit Therapeuten bzw. Pflegepersonal hinsichtlich emotionaler und psychischer Stabilität individuell angesprochen und über die Studie und deren Verlauf informiert. Bei mündlicher Einwilligung wurde mit jedem Teilnehmer ein individueller Termin ausgemacht. Unmittelbar vor Beginn der Versuchsdurchführung unterschrieben alle Studienteilnehmer zunächst die schriftliche Einwilligungserklärung.

2. **Pseudonymisierung:** Die Pseudonymisierung erfolgte über eine – vorab mittels Microsoft® Excel® 2010 erstellte – Pseudonymisierungsliste, bestehend aus dreistelligen Buchstabencodes. Diese dienten zudem als Patientenkürzel.

3. **Case Report Form/Basisdaten:** Zur Erhebung soziodemographischer und weiterer studienrelevanter Daten der Patienten wurde vor Studienbeginn ein Case Report Form (CRF) entworfen (s. Anhang B). Bei der Erstellung des CRFs dienten die Empfehlungen von Bellary et al. bezüglich CRFs als Orientierungsgrundlage (Bellary et al. 2014). So wurde beispielsweise eine fortlaufende Protokollnummer, der jeweilige Patientencode und eine Unterschriftenzeile mit Datum für den Testleiter eingefügt (Bellary et al. 2014). Die Unterschrift unter jedem CRF ist als nach bestem Wissen und Gewissen erfolgende Verbürgung für die Vollständigkeit und Richtigkeit der erhobenen Daten zu verstehen (Gaus u. Muche 2014, S. 66). Außerdem wurden Datenmenge und Datendarstellung im Einklang mit den Empfehlungen von Bellary et al. betont gering, einfach und standardisiert (z.B. Geburtsdatum: dd/mm/yyyy) gehalten, sodass im Idealfall nur nützliche und hochwertige Daten erhoben worden sind (Bellary et al. 2014). Freitext-Antworten wurden, wie von Gaus und Muche angeraten, nicht in das CRF integriert (Gaus u. Muche 2014, S. 68). Der Aufbau des CRFs wurde zudem so gestaltet, dass es sich nahtlos in den Studienablauf einfügen konnte, entsprechend der Empfehlung von Gaus und Muche (Gaus u. Muche 2014, S. 65). Ausgefüllt wurde das CRF in

Zusammenarbeit mit dem jeweiligen Patienten (teils in Form eines Interviews, teils als Fragebogen).

4. **Ausfüllen der ersten beiden Fragebögen/Tests: 1. BSI, 2. SCIP** (Näheres zu den Fragebögen/Tests: s. Kapitel 2.5 Untersuchungsinstrumente)

5. **Patienteninstruktion:** Ziehen und Öffnen eines vorbereiteten Briefumschlags, gepaart mit dem Hinweis, bei den folgenden Testdurchführungen auf die Instruktion zu achten und der Anweisung, dem Versuchsleiter (noch) nicht mitzuteilen, welche Instruktion gezogen wurde.

Instruktion Gruppe A (Kontrollgruppe):

Es werden nun zwei verschiedene Tests durchgeführt, die zum einen die Gedächtnisleistung prüfen und zum anderen nach dem Vorhandensein diverser psychischer Symptome fragen. Bitte strengen Sie sich bei dem Gedächtnistest an, so gut Sie können und beantworten Sie alle Fragen des zweiten Tests wahrheitsgemäß.

Bitte verraten Sie dem Testleiter vor der Untersuchung nicht, welche Anweisungen Sie erhalten haben! Sie werden anschließend gebeten, ihm diesen Zettel zu zeigen.

Instruktion Gruppe B (Interventionsgruppe):

Bitte stellen Sie sich folgende Situation vor und versuchen Sie sich während der Testdurchführung so gut wie möglich in diese hineinzuversetzen:

Sie haben einen Antrag auf Frühberentung gestellt, da Sie sich den Belastungen am Arbeitsplatz, unter anderem auf Grund Ihrer psychischen Symptome, nicht mehr gewachsen fühlen. Im Verlauf des Rentenverfahrens werden Sie zu einem ärztlichen Gutachter geschickt. Sie wissen, dass von seiner Beurteilung der Ausgang des Verfahrens abhängt. In einem vorangehenden Gespräch hatten Sie Gelegenheit, Ihre Beschwerden und Ihre Lebensgeschichte zu schildern. Anschließend fordert der Gutachter Sie auf, an zwei Tests teilzunehmen, einem zu Ihrer Gedächtnisleistung und einem zu bestimmten psychischen Symptomen. Um sicher zu gehen, dass der Gutachter Ihre Beschwerden für schwerwiegend genug hält, möchten Sie diese deutlich machen. Allerdings wollen Sie auch nicht durch zu grobe Übertreibung auffallen.

Bitte verraten Sie dem Testleiter vor der Untersuchung nicht, welche Anweisungen Sie erhalten haben! Sie werden anschließend gebeten, ihm diesen Zettel zu zeigen.

Mit dem Ziehen und Umsetzen von Instruktion A wurden die Patienten der Kontrollgruppe (Gruppe A), also den Studienteilnehmern mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten, zugeteilt.
Die absichtlich etwas offener gehaltene Instruktion B teilte die Patienten der Interventionsgruppe (Gruppe B), also den experimentellen negativen Antwortverzerrern der Studie, zu. Der Eingangssatz von Instruktion B („Sie haben einen Antrag auf Frühberentung gestellt…“) wurde in dieser Form gewählt, um möglichst realitätsnah die Situation eines Rentenbewerbers zu simulieren und damit den möglichen Spielraum von Verdeutlichung, Aggravation und suboptimaler Leistungsanstrengung bei der Durchführung von WMT und SFSS zu eröffnen. Nach genauem Durchlesen der Instruktion wurden die Teilnehmer dazu aufgefordert, diese in den Umschlag zurückzustecken, um die Verblindung zu gewährleisten.

6. **WMT, 1. Teil** (Computer-Programm)

7. **Pause** (ca. 10 Minuten)

8. **SFSS**

9. Genau 30 Minuten nach Beendigung des 1. Teils: **WMT, 2.Teil**

10. **Visuelle Analogskala (VAS):** Die visuelle Analogskala diente, als Bestandteil des CRFs, zur Überprüfung der Instruktionsumsetzung nach beendeter Versuchsdurchführung. Daher wurden die Studienteilnehmer im Anschluss an die Versuchsdurchführung dazu aufgefordert, auf einer Skala (Länge: 100 mm) von 0%-100% einzuzeichnen, wie sehr sie sich bei der Durchführung von WMT und SFSS – bezogen auf ihr tatsächliches Leistungsvermögen – angestrengt hatten. Die per Kreuz eingezeichnete „subjektive Anstrengung“ wurde dann vom Nullpunkt der Skala aus gemessen und in eine Prozentzahl umgerechnet. 1mm entsprach dabei 1% „Anstrengung“. Visuelle Analogskalen eignen sich laut Gaus und Muche gut zur Quantifizierung subjektiver Einschätzungen (Gaus u. Muche 2014, S. 86).

11. **Entblindung:** Direkt nach Beendigung der Versuchsdurchführung wurde der Umschlag mit der Instruktion geöffnet und die Gruppenzugehörigkeit des jeweiligen Studienteilnehmers auf dem CRF notiert. Die Übertragung der erhobenen Daten in eine Microsoft® Excel® 2010-Datei erfolgte somit bereits unter entblindeten Bedingungen.

12. **Erfragen und Dokumentieren, in welcher Art und Weise die jeweilige Instruktion umgesetzt wurde:** Zunächst wurden nur die Studienteilnehmer mit Instruktion B (Interventionsgruppe B) darum gebeten, im Anschluss an die Versuchsdurchführung eingehender über ihre Interpretation und Umsetzung der Instruktion während der Durchführung von WMT und SFSS zu berichten. Ab einschließlich dem 19. Studienteilnehmer mit Instruktion A wurde auch die Kontrollgruppe A zur Umsetzung ihrer Instruktion befragt. Dadurch sollte dem Gleichbehandlungsgrundsatz, der für die unterschiedlichen Gruppen eines RCTs – mit Ausnahme der Einflussgröße – gilt (Gaus u. Muche 2014, S. 50), besser entsprochen werden.

13. **Datenerhebung abschließen, Vollständigkeit prüfen**

Die Versuchsdurchführung in dieser Form dauerte pro Patient etwa 90 Minuten.

2.5 Untersuchungsinstrumente

2.5.1 Brief Symptom Inventory (BSI)

Zuallererst wurden die Studienteilnehmer darum gebeten, das Brief Symptom Inventory (BSI (Derogatis u. Melisaratos 1983), aus urheberrechtlichen Gründen wird auf einen Abdruck im Anhang verzichtet) auszufüllen. Seit dem Jahr 2000 gibt es für das BSI eine deutsche Übersetzung (Franke 2000). Entworfen wurde das „Selbstbeurteilungsverfahren“ (Franke 2000, S. 5) BSI, um in klinischen oder wissenschaftlichen Kontexten mit geringem Zeitaufwand die Art und Schwere der Symptome eines Patienten erfassen zu können (Piersma et al. 1994). Dazu fragt das BSI das Auftreten von Beschwerden verschiedener psychiatrischer Krankheitsbilder „während der vergangenen sieben Tage bis heute“ (Franke 2000, S. 8) ab. Der Untersucher erhält also ein recht aktuelles Bild der gegenwärtigen Symptomatik, bei einer Testdauer von im Schnitt 8 bis 10 Minuten (Franke 2000, S. 1).

Bei der Bearbeitung des BSI kann die Testperson bezüglich jedes der 53 genannten Items bzw. Symptome über eine Likert-Skala zwischen 5 verschiedenen Schweregraden der Belastung durch das jeweilige Symptom wählen: von 0 („überhaupt nicht“) bis 4 („sehr stark“) (Franke 2000, S. 8).

Die 53 Items des BSI können neun testpsychologisch validierten Subskalen zugeordnet werden: Somatisierung, Zwanghaftigkeit, Depressivität, Unsicherheit im Sozialkontakt, Aggressivität/Feindseligkeit, Ängstlichkeit, Phobische Angst, Paranoides Denken und Psychotizismus (Franke 2000, S. 1). Zusätzlich gibt es noch vier Items, die keiner der neun Subskalen zugeordnet werden können (Franke 2000, S. 11). Dabei handelt es sich um „Fragen nach schlechtem Appetit, Einschlafschwierigkeiten, dem Gedanken an den Tod und das Sterben und nach Schuldgefühlen“ (Franke 2000, S. 14).

Ausgewertet wird das BSI vor allem über die drei globalen Kennwerte GSI (Global Severity Index), PST (Positive Symptom Total) und PSDI (Positive Symptom Distress Index) (Franke 2000, S. 14). Der GSI wird durch Division des Gesamtsummenwertes aller Items durch 53 berechnet (Franke 2000, S. 16) und misst „die grundsätzliche psychische Belastung“ (Franke 2000, S. 14). Der PST ist gleich der Anzahl der Items, bei denen ein Schweregrad >0 angekreuzt wurde (Franke 2000, S. 16). Der PSDI wird durch Division des Gesamtsummenwertes

aller Items durch den PST berechnet (Franke 2000, S. 16). Unter diesen drei Kennwerten gilt der GSI als „der sensitivste Indikator für die psychische Belastung“ (Franke 2000, S. 19) des Patienten.

Das BSI ist für verschiedene Populationen, z.B. für Erwachsene oder Studierende, normiert (Franke 2000, S. 2). Man kann daher das Ergebnis eines Einzelfalles einordnen und vergleichen, indem man den BSI-Rohwert in einen T-Wert transformiert (Franke 2000, S. 17). Die T-Werte für Frauen und Männer unterscheiden sich, da Frauen meist höhere GSI-Werte erzielen als Männer (Urbán et al. 2014). Laut Manual gilt ein Patient „als psychisch auffällig belastet oder als ‚Fall‘, wenn: T-GSI ≥ 63 oder T-2-Skalen ≥ 63, d.h. wenn [der T-Wert des] GSI größer oder gleich 63 ist oder wenn die T-Werte bei mindestens zwei Skalen größer oder gleich 63 sind“ (Franke 2000, S. 20). Erzielt ein Patient solche Werte, sollte dies im klinischen Alltag eine eingehendere Diagnostik in Gang setzen, da der Patient augenscheinlich akut belastet ist (Franke 2000, S. 20f).

Zu den relevanten Kennwerten eines Tests gehört u.a. die interne Konsistenz, die Aussagen darüber erlaubt, in welchem Ausmaß die Items eines Tests gemeinsam dasselbe Konstrukt messen (Henson 2001). Ein häufig benutztes Maß für die interne Konsistenz ist Cronbachs Alpha (Cronbach 1951). Darüber hinaus wird oft auch die Retest-Reliabilität mitangegeben, die als Ergebnisstabilität über die Zeit aufzufassen ist und z.B. durch mehrfache Testdurchführung durch dieselbe Testperson bestimmt wird (Henson 2001). Für das BSI werden Cronbachs Alpha-Werte zwischen .71 (für die Psychotizismusskala) und .85 (für die Depressivitätsskala) (Derogatis u. Melisaratos 1983), bzw. zwischen .75 (für die Psychotizismusskala) und .89 (für die Depressivitätsskala) (Boulet u. Boss 1991) berichtet. Bei der Retest-Reliabilität ergaben sich für das BSI Werte zwischen .68 (für die Somatisierungsskala) und .91 (für die Phobische Angst-Skala) (Derogatis u. Melisaratos 1983).

Boulet und Boss stellten recht hohe Korrelationen (r = .55 bis .80) zwischen den neun Symptomskalen fest, was dafür spricht, dass die Subskalen nicht komplett unabhängig voneinander sind, sondern sich gegenseitig beeinflussen (Boulet u. Boss 1991). Dies bedeutet, dass sie sich bei der Abfrage der Symptome teilweise überschneiden. In einer abschließenden Bewertung zum BSI heißt es daher auch sinngemäß, dass das BSI wohl zur groben Erfassung der aktuellen psychischen Belastung taugt, aber über die Analyse seiner neun Subskalen keine genauere

Differentialdiagnose der Psychopathologie ermöglicht (Boulet u. Boss 1991).
In der vorliegenden Studie wurde das BSI verwendet, um die tatsächliche Symptomschwere der Patienten grob erfassen und mit den BVT-Ergebnissen in Beziehung setzen zu können.

2.5.2 Screen for Cognitive Impairment in Psychiatry (SCIP)

Entsprechend des Studienprotokolls folgte auf das BSI an zweiter Stelle das Screen for Cognitive Impairment in Psychiatry (SCIP (Purdon 2005); aus urheberrechtlichen Gründen wird auf einen Abdruck im Anhang verzichtet). Es wurde die deutsche Fassung des SCIP ((Lasser 2011), Formular 1) verwendet. Das SCIP dient dazu, die kognitiven Funktionen und das kognitive Leistungsniveau eines Patienten einzuschätzen und besteht aus fünf Aufgaben bzw. Subskalen (Czekaj et al. 2014). Bei der ersten Aufgabe wird der Testperson in drei Durchgängen jeweils die gleiche Wortliste mit 10 Wörtern vorgelesen, die jeweils so gut wie möglich wiedergegeben werden sollen. Die vierte Aufgabe fragt noch einmal dieselben 10 Wörter ab, sodass über die Aufgaben 1 und 4 das „unmittelbare[s] und [das] zeitverzögerte[s] verbale[s] Kurzzeitgedächtnis“ (Czekaj et al. 2014, S. 166) getestet wird. Aufgabe zwei prüft das Arbeitsgedächtnis, indem von der Testperson verlangt wird, sich unterschiedliche dreistellige Konsonantenkombinationen zu merken, während sie von variierten Ausgangszahlen rückwärts zählen muss (Czekaj et al. 2014). Die dritte Aufgabe eruiert die „Wortflüssigkeit“ (Czekaj et al. 2014, S. 166) der Testperson, indem diese mit unterschiedlichen Anfangsbuchstaben innerhalb von 30 Sekunden möglichst viele Wörter bilden soll. Bei der fünften und letzten Aufgabe wird die Testperson dazu aufgefordert, eine bestimmte Zeichenfolge innerhalb von 30 Sekunden möglichst schnell zu übertragen, wodurch die „Bearbeitungsgeschwindigkeit“ (Czekaj et al. 2014, S. 166) abgeschätzt wird. Durch Addition der Ergebnispunkte der Subskalen kann im Rahmen der SCIP-Auswertung ein Summenscore für den Test berechnet werden (Rojo et al. 2010).

Auch für das SCIP wurden Kennwerte wie interne Validität (Cronbachs Alpha von .72 bis .73 bei einer Gruppe schizophrener Patienten (Lasser 2011, S. 99), .73 (Pino et al. 2008) bzw. .74 (Guilera et al. 2009)) und Retest-Reliabilität (.90 (Pino et al. 2008), .87 (Guilera et al. 2009) bzw. .82 (Purdon 2005, S. 13) für den Summenscore) berechnet und für gut befunden (Pino et al. 2008). Die externe

Validität des SCIP, also die Übereinstimmung mit anderen Maßen kognitiver Funktion, wird ebenfalls als zufriedenstellend bewertet (Cuesta et al. 2011; Pino et al. 2008). Zudem kann wohl anhand der SCIP-Summenscores relativ sicher zwischen kognitiv beeinträchtigten und gesunden Personen differenziert werden (Pino et al. 2008; Cuesta et al. 2011; Lasser 2011, S. 119f). Rojo et al. berichten in diesem Zusammenhang für den SCIP-Summenscore bei einem Cut-off von <70 Punkten (<70 Punkte im SCIP-Summenscore = kognitiv beeinträchtigt) eine Sensitivität von 87,9%, eine Spezifität von 80,6% sowie einen positiven prädiktiven Wert (PPW) von 77,3% und einen negativen prädiktiven Wert (NPW) von 89,9% (Rojo et al. 2010).

Mehrere Studien geben für das SCIP übereinstimmend eine durchschnittliche Bearbeitungszeit von etwa 15 bis 16 Minuten an (Czekaj et al. 2014; Pino et al. 2008). Somit lässt es sich als Screening-Instrument relativ problemlos in den klinischen Alltag integrieren (Czekaj et al. 2014). Allerdings wird – ähnlich wie beim BSI – auch für das SCIP empfohlen, der Testdurchführung eine eingehendere Diagnostik mit weiteren Instrumenten folgen zu lassen (Rojo et al. 2010), da das SCIP nur „grobe Hinweise auf kognitive Beeinträchtigungen liefert“ (Czekaj et al. 2014, S. 167). Zu beachten ist auch, dass das SCIP v.a. an schizophrenen Patientenpopulationen validiert und normiert wurde (z.B. Pino et al. 2008; Cuesta et al. 2011; Lasser 2011, S. 87f).
In der vorliegenden Studie wurde das SCIP verwendet, um das kognitive Leistungsniveau der Studienteilnehmer festlegen und mit den BVT-Ergebnissen in Beziehung setzen zu können.

2.5.3 Strukturierter Fragebogen Simulierter Symptome (SFSS)

Der Strukturierte Fragebogen Simulierter Symptome (SFSS, wegen ungeklärter Verhältnisse bezüglich des Urheberrechts wird auf einen Abdruck im Anhang verzichtet) ist die deutsche Version des Structured Inventory of Malingered Symptomatology (SIMS) und besteht wie dieses aus 75 Items (Cima et al. 2003). In einer umfassenden Metaanalyse zum SIMS wurde u.a. die Beeinflussung der SIMS-Scores durch unterschiedliche Sprachen analysiert und festgestellt, dass in den untersuchten Studien keine signifikanten sprachbedingten Unterschiede zwischen deutschsprachigen (SFSS) und englischsprachigen (SIMS) Studiengruppen auftraten (van Impelen et al. 2014). Erkenntnisse bezüglich des

SIMS dürften daher weitestgehend auf seine deutsche Übersetzung, den SFSS, übertragbar sein.

Das SIMS gehört zu den „Selbstbeurteilungsverfahren […], die spezifisch zur Erfassung negativer Antwortverzerrungen konstruiert wurden“ (Blaskewitz u. Merten 2007, S. 145). Als Screening-Instrument soll es die Aggravation bzw. Simulation psychiatrischer Beschwerden (z.B. einer Depression) sowie kognitiver Beeinträchtigungen (z.B. Gedächtnisprobleme) erfassen (Merckelbach u. Smith 2003).

Die 75 wahr/falsch-Items des SIMS sind, den Testentwicklern Burger und Smith zufolge, z.T. an Items bereits bewährter Tests angelehnt (Smith u. Burger 1997). Einige Items wurden aber auch neu entworfen (Smith u. Burger 1997) – im Einklang mit bestehender Literatur zu den Eigenheiten von aggravierenden bzw. simulierenden Personen (z.B. Resnick 1984). Wie auch beim SIMS können beim SFSS jeweils 15 der 75 Items einer der 5 Subskalen des SFSS (affektive Störung, neurologische Beeinträchtigung, niedrige Intelligenz, Psychose und amnestische Störung (Kobelt et al. 2012)) zugeordnet werden. Im Folgenden werden sie als AF, N, NI, P und AM abgekürzt.

Um Aggravation bzw. Simulation erfassen zu können, verfolgt das SIMS – und somit auch der SFSS – verschiedene Strategien, wie z.B. die Angabe bizarrer Symptome (z.B. Item Nr. 42: „Blumen haben magische Kräfte, wie die Fähigkeit, mit Menschen sprechen zu können." (Cima et al. 2003, S. 984), P-Skala), oder unglaubwürdiger Beschwerden (z.B. Item Nr. 1: „Manchmal verliere ich alles Gefühl in meiner Hand, so als ob ich einen Handschuh anhätte." (Cima et al. 2003, S. 983), N-Skala) (Smith u. Burger 1997). Ausgewertet wird der SFSS über das Zusammenzählen aller als auffällig geltenden Antworten (Kobelt et al. 2012). Welche Antwort (ja bzw. wahr/nein bzw. falsch) bei welcher Frage als auffällig im Sinne einer Simulation oder Aggravation zu werten ist, legt die Decodierungstabelle des SFSS fest. Im Rahmen der Auswertung kann sowohl ein SFSS-Gesamtscore als auch ein Summenwert für jede einzelne der 5 Subskalen berechnet werden. Laut Burger und Smith ist aber der SIMS-Gesamtscore das beste Entscheidungskriterium für oder gegen Aggravation bzw. Simulation (Smith u. Burger 1997). Bei dem in ihrer Veröffentlichungsstudie verwendeten Cut-off von >14 Punkten im SIMS-Gesamtscore errechneten sie für den SIMS-Gesamtscore eine Sensitivität von 95,6% und eine Spezifität von 87,9% (Smith u. Burger 1997).

In van Impelen et al. SIMS-Metaanalyse wurden die Ergebnisse unterschiedlicher Studiendesigns (Erhebungen an „echten“ Antragsstellern oder Versuche mit experimentellen negativen Antwortverzerrern bzw. Simulanten und klinischen oder nicht-klinischen Kontrollgruppen) zusammengefasst (van Impelen et al. 2014). Van Impelen et al. berichten für das SIMS Sensitivitäten zwischen 63% und 100% für die gebräuchlichen Cut-offs von >14/>16 (van Impelen et al. 2014). Die für dieselben Cut-offs berechneten Spezifitäten zeigen jedoch stark schwankende Werte zwischen 23% und 100% und sind z.T. auffallend niedrig (van Impelen et al. 2014).

Van Impelen et al. fassten ferner auch die bisher errechneten Werte für den positiven (PPW) und negativen (NPW) prädiktiven Wert des SIMS zusammen (van Impelen et al. 2014). Dabei wird ersichtlich, dass beispielsweise der PPW bei Studien mit „echten“ Antragsstellern zwischen 22% und 75% variiert, während der NPW für diese Gruppe zwischen 89% und 100% schwankt, je nach basaler Auftretenshäufigkeit von negativer Antwortverzerrung bzw. Simulation in der Bevölkerung (van Impelen et al. 2014). Generell gilt nämlich, dass die Wahrscheinlichkeit bei einem positiven Testergebnis auch tatsächlich ein negativer Antwortverzerrer bzw. Simulant zu sein (PPW) mit der Auftretenshäufigkeit von negativer Antwortverzerrung bzw. Simulation in der Bevölkerung steigt. Im Umkehrschluss steigt der NPW, wenn die Auftretenshäufigkeit sinkt (van Impelen et al. 2014). Die guten Werte, die der NPW erreicht, bedeuten, dass Testpersonen mit SIMS-Gesamtscores unter dem Cut-off während der Testdurchführung auch sehr wahrscheinlich keine negative Antwortverzerrung bzw. Simulation gezeigt haben (van Impelen et al. 2014).

Als Maß für die interne Konsistenz von SFSS bzw. SIMS werden Cronbachs Alpha-Werte von .80 (Cima et al. 2003) bzw. .72 (Merckelbach u. Smith 2003) angegeben. Cima et al. beschreiben für den SFSS mit einem Wert von .97 weiterhin eine exzellente Retest-Reliabilität (Cima et al. 2003), während Merckelbach und Smith für das SIMS eine Retest-Reliabilität von .72 berechneten (Merckelbach u. Smith 2003). Die Korrelationen zwischen den 5 SIMS-Subskalen sind größtenteils gering (z.B. r = .19 zwischen Affektskala (AF) und Intelligenzskala (NI) (Smith u. Burger 1997)), d.h. die Subskalen überschneiden sich nicht zu stark in den Qualitäten, die sie erfassen sollen und sind daher nicht überflüssig.

Die externe Validität von SIMS bzw. SFSS kann durch die Stärke der Korrelation des Tests mit anderen, bewährten BVTs, wie z.B. dem Structured Interview of Reported Symptoms (SIRS (Rogers et al. 1991)), bewertet werden. Edens et al. geben für die Stärke der Korrelation des SIMS mit dem SIRS einen Wert von .81 an (Edens et al. 2007), was für einen starken gleichsinnigen Zusammenhang der beiden Tests spricht. Van Impelen et al. fiel zudem auf, dass Testpersonen mit hohen SIMS-Gesamtscores ebenfalls dazu neigen hohe Werte in gebräuchlichen klinischen Selbstbeurteilungsverfahren, wie z.B. Becks Depressionsinventar (BDI (Beck et al. 1961)), zu erzielen (van Impelen et al. 2014). Dementsprechend berichten z.B. Merten et al. für den SFSS eine Korrelation mit dem BDI im Wert von .72 (Merten et al. 2007). Angesichts dieser Ergebnisse wird ersichtlich, warum dem SIMS eine gewisse Sensitivität für tatsächlich vorliegende Psychopathologie vorgeworfen wird (van Impelen et al. 2014).
In der vorliegenden Studie wurde die von Cima et al. 2003 veröffentlichte deutsche Version des SIMS (dt. Version: SFSS) verwendet (Cima et al. 2003). Allerdings war es notwendig die Items Nr. 63 und 67 an die aktuellen Gegebenheiten anzupassen (Item Nr. 63: Bundeskanzlerin Angela Merkel statt Bundeskanzler Gerhard Schröder; Item Nr. 67: Euro und Cent statt Deutsche Mark und Pfennig).

2.5.4 Word Memory Test (WMT)

Der Word Memory Test (WMT) wurde „speziell zur Erfassung von negativen Antwortverzerrungen und Unkooperativität in der neuropsychologischen Diagnostik entwickelt“ (Merten et al. 2007, S. 512). Seit 2004 gibt es eine deutsche Version des WMT (Brockhaus u. Merten 2004). Der gute Ruf, der dem WMT v.a. in Nordamerika nachgesagt wird (Brockhaus u. Merten 2004), basiert auf zahlreichen Studien an unterschiedlichsten Populationen (Hartman 2002), in denen er validiert und normiert wurde. Zur Testdarbietung in der vorliegenden Studie wurde die deutsche Version des WMT-Windows®-Programms (Green 2005) verwendet.

Im Rahmen der WMT-Durchführung werden der Testperson am Computer in zwei Präsentationsdurchgängen 20 assoziativ miteinander verbundene Wortpaare präsentiert (z.B. Hund – Katze, Mann – Frau), also insgesamt 40 Wörter. Die Testperson wird aufgefordert, sich diese Wortpaare bzw. Wörter zu merken. In den folgenden sechs WMT-Durchgängen (Immediate Recognition, Delayed

Recognition, Multiple Choice, Paired Associate Recall, Free Recall und Long Delay Free Recall) wird die zu erlernende Wortliste abgefragt. Direkt im Anschluss an die zweimalige Präsentation der Wortpaare (Immediate Recognition (IR)) bzw. nach 30 Minuten (Delayed Recognition (DR)) wird der Testperson eines der 40 Wörter mit einem weiteren, nicht in der Testliste befindlichen Wort präsentiert und sie muss entscheiden, welches der beiden Wörter in der vorher gezeigten Liste stand. Hierbei handelt es sich um ein Alternativwahlverfahren mit Zwangswahlkomponente, da immer ein Wort ausgewählt werden muss (Merten 2014, S. 31). Das WMT-Programm berechnet anschließend automatisch die Konsistenz (CNS) zwischen den Antworten im IR- und DR-Durchgang in Prozent der Übereinstimmung (Carone 2014), um somit Hinweise auf ein möglicherweise inkonsistentes Antwortverhalten liefern zu können. IR-, DR- und CNS-Durchgang werden vom Schwierigkeitsgrad her auch als die drei „einfachen“ WMT-Durchgänge angesehen (Carone 2014) und gelten als Hauptindikatoren für die Leistungsanstrengung der Testperson (Green et al. 2003). Für den IR- und DR-Durchgang ergibt sich zudem die interessante Situation, dass es ab einem bestimmten Prozentwert unter 50% richtigen Antworten statistisch quasi unmöglich ist das Ergebnis unabsichtlich durch Raten erzielt zu haben. Diese „Schwelle für Unter-Zufall-Antworten“ (Merten 2014, S. 32) – bei Unterschreitung welcher suboptimale Leistungsanstrengung als Ursache für das schlechte Abschneiden angenommen werden sollte (Merten 2014, S. 32) – liegt für den WMT laut Tan et al. bei einem Wert von <37,5% im IR- oder DR-Durchgang (Tan et al. 2002).

Auf den DR-Durchgang folgt direkt der Multiple Choice-Durchgang (MC), bei welchem jeweils ein Wort der 20 Wortpaare präsentiert wird, während die Testperson den zugehörigen Partner aus einer Liste von acht angebotenen Wörtern wählen muss. Weiter geht es anschließend mit dem Paired Associate Recall-Durchgang (PA), bei welchem der Testleiter der Testperson das erste Wort der Paare mündlich nennt, woraufhin diese das dazugehörige Wort frei aus dem Gedächtnis erinnern soll. Auf den PA-Durchgang folgt der Free Recall-Durchgang (FR), der die freie Wiedergabe so vieler der 40 Wörter als möglich verlangt. Nach 20-minütiger Pause erfolgt abschließend der Long Delay Free Recall-Durchgang (LDFR), der in der vorliegenden Studie aus Zeitgründen weggelassen wurde. Die vier letztgenannten Durchgänge (MC, PA, FR, LDFR) gelten vom

Schwierigkeitsgrad her auch als die „schwierigen“ WMT-Durchgänge (Carone 2014), denn sie sollen tatsächlich die Gedächtnisleistung der Testperson messen (Green et al. 2003).

Ausgewertet wird der WMT vom Computerprogramm selbst, indem für jeden Durchgang der Prozentsatz richtiger Antworten angegeben wird (Goodrich-Hunsaker u. Hopkins 2009). Die Leistungsanstrengung der Testperson gilt als suboptimal, wenn bei einem der einfachen WMT-Durchgänge (IR oder DR oder CNS) der Cut-off von ≤82,5% richtigen Antworten unterschritten wird (Green 2005, S. 9). In verschiedenen Quellen werden für diesen WMT-Cut-off Sensitivitäten von 97,7% (Green et al. 2003) bis 100% (Brockhaus u. Merten 2004; Tan et al. 2002) sowie Spezifitäten von 97% (Brockhaus u. Merten 2004) bis 100% (Green et al. 2003; Tan et al. 2002) angegeben.

Green et al. berichten für die Korrelationen zwischen den einfachen WMT-Durchgängen (IR, DR, CNS) Werte zwischen .86 und .89 (Green et al. 2003), was für eine hohe interne Reliabilität der drei Durchgänge spricht. Für die Korrelationen zwischen den schwierigen WMT-Durchgängen (MC, PA, FR, LDFR) geben sie Werte zwischen .71 und .90 an (Green et al. 2003). Des Weiteren beschreibt Green eine 100%ige Übereinstimmung der WMT-Klassifikationen bei wiederholter Testdurchführung an denselben Versuchspersonen (Green 2005, S. 68) und damit eine hohe Retest-Reliabilität.

Bezüglich des Einflusses möglicher Störgrößen fanden Rienstra et al. keinen statistisch signifikanten Zusammenhang zwischen WMT-Ergebnissen und Geschlechtszugehörigkeit, allerdings war das WMT-Ergebnis statistisch signifikant von Alter und Bildungsstand der Testperson abhängig (Rienstra et al. 2009). Niedrigeres Alter und höherer Bildungsstand bedingten dabei höhere WMT-Ergebnisse (Rienstra et al. 2009). Testentwickler Green untersuchte ebenfalls den Einfluss von Bildungsstand, Alter sowie verbaler Intelligenz auf das WMT-Ergebnis, stellte dabei jedoch keine statistisch signifikanten Auswirkungen dieser Variablen auf die einfachen WMT-Durchgänge (IR, DR, CNS) fest (Green 2005, S. 62). Das Ergebnis der schwierigen WMT-Durchgänge (MC, PA, FR, LDFR) hingegen, fiel – ähnlich wie bei Rienstra et al. – je nach Bildungsstand, Alter und verbaler Intelligenz besser oder schlechter aus (Green 2005, S. 62f).

Durch seine Art der Testdarbietung nutzt der WMT das Prinzip der verdeckten Leichtigkeit (Brockhaus u. Merten 2004). Mögliche Simulanten bzw. negative

Antwortverzerrer unterschätzen angesichts der 20 Wortpaare, dass z.T. selbst Patienten mit Demenz die einfachen WMT-Durchgänge (IR, DR, CNS) bestehen (vgl. Green et al. 2011) und fallen in der Folge durch ihr schlechtes Ergebnis auf (Gervais et al. 2001). Die einfachen WMT-Durchgänge besitzen also absichtlich einen geringen Schwierigkeitsgrad, weswegen es Green zufolge nur bei extremsten Formen kognitiver Beeinträchtigung, wie z.B. eben bei einer Demenzerkrankung (Green et al. 2011), zu Werten unterhalb des Cut-offs von ≤82,5% kommen kann (Green et al. 2003). Um das individuelle WMT-Ergebnis angemessen einordnen zu können, bietet das WMT-Computerprogramm im Anschluss an die Testdurchführung 61 Vergleichsprofile unterschiedlichster Populationen (z.B. Kinder, orthopädische Patienten, depressive Patienten etc.; Green 2005, S. 54ff).

2.6 Statistische Methoden

Die statistische Auswertung der gewonnenen Daten wurde explorativ durchgeführt, da im Vorfeld der vorliegenden Studie weder eine Fallzahlabschätzung noch eine Powerkalkulation erfolgte. Bezüglich der im Ergebnisteil aufgeführten p-Werte sei also darauf hingewiesen, dass diese im Rahmen einer explorativen Datenanalyse keinerlei statistische Beweiskraft hinsichtlich der Annahme oder Ablehnung von Hypothesen besitzen, sondern bei Signifikanz lediglich Hypothesen generieren (Gaus u. Muche 2014, S. 404).

Das statistische Signifikanzniveau (α) für das explorative Testen wurde auf α = .05 festgelegt.

Ausgewertet wurden die erhobenen Daten mittels Microsoft® Excel® 2010 (Version: 14.0.7173.5000, 2010), IBM® SPSS® Statistics 21 (Version: 21.0.0.0 für Windows®, 2012) und GraphPad Software QuickCalcs (GraphPad Software 2016).

Das auch als „freezing“ (Gaus u. Muche 2014, S. 70) bezeichnete, förmliche Einfrieren der Datenbank nach Abschluss der Datenerhebung wurde für den 11.03.2016 dokumentiert. Damit sollte u.a. festgehalten werden, dass die Daten nach bestem Wissen und Gewissen erhoben wurden.
Die Datenauswertung fand unter bereits entblindeten Bedingungen statt, um Kontroll- und Interventionsgruppe formieren und statistisch miteinander vergleichen zu können. Alle relevanten erhobenen Daten wurden in eine Microsoft® Excel® 2010-Tabelle übertragen und dort zunächst deskriptiv ausgewertet. Zu diesem Zweck wurde der Mittelwert (MW) – inklusive Standardabweichung (SD) – aller relevanten Variablen berechnet. Zusätzlich wurde für alle relevanten Werte auch der Median – inklusive Variationsbreite (= Differenz aus größtem und kleinstem Wert (Bortz u. Schuster 2010, S. 32)), im Folgenden als „range“ bezeichnet – bestimmt, da der Median als robuster gegenüber Ausreißern als der Mittelwert gilt (Bortz u. Schuster 2010, S. 28). Zunächst wurde die gesamte Stichprobe und dann nach den Gruppen A und B unterteilt deskriptiv ausgewertet.

Im Anschluss wurde mittels geeigneter Verfahren (Chi-Quadrat-Test, Exakter Fisher-Test, U-Test, t-Test) geprüft, ob zwischen Interventions- und Kontrollgruppe

bezüglich der Schwere der aktuellen psychischen Symptomatik (BSI) sowie bezüglich des kognitiven Leistungsniveaus (SCIP) und der soziodemographischen Merkmale (Alter, Geschlecht, Bildungsabschluss, Beschäftigungsverhältnis, Aufnahmediagnosen, Umstände der Rekrutierung, Vorliegen abgelehnter Anträge auf Erwerbsminderungsrente) keine statistisch signifikanten Unterschiede bestanden, ob also die Randomisierung den gewünschten Effekt erzielt hatte. Die stetigen Merkmale BSI (ausgedrückt im Punktwert des GSI), SCIP-Summenscore und Alter wurden zunächst mittels des Shapiro-Wilk-Tests (Shapiro et al. 1968) auf Normalverteilung hin untersucht, um dann je nach Ergebnis ein geeignetes Testverfahren auszuwählen (Bortz u. Schuster 2010, S. 122). Zudem wurde jeweils auch die Form der Häufigkeitsverteilung per Histogramm überprüft. Im Falle einer anzunehmenden Normalverteilung wurde der t-Test (Student 1908), im Falle einer schiefen Verteilung sowie bei ordinalskalierten Merkmalen (z.B. Bildungsabschluss) der U-Test (Mann u. Whitney 1947) verwendet. Die Unabhängigkeit zweier nominalskalierter Merkmale (z.B. Geschlecht vs. Gruppenzugehörigkeit) wurde bei kleiner Fallzahl ($n \leq 60$) und insgesamt eher geringen beobachteten und erwarteten Häufigkeiten v.a. mit dem exakten Fisher-Test untersucht (Gaus u. Muche 2014, S. 290f). Bei zu geringen Zellbesetzungen der Merkmalsausprägungen wurden übergeordnete Klassen gebildet (z.B. bei den Aufnahmediagnosen).

Nachfolgend wurde geprüft, ob sich die WMT- und SFSS-Ergebnisse sowie die subjektiv anzugebende „Anstrengung“ – erfasst über die visuelle Analogskala – tatsächlich statistisch signifikant zwischen den Gruppen mit unterschiedlicher Instruktion unterschieden. Laut Larrabee neigen BVT-Ergebnisse generell eher dazu, schiefe Häufigkeitsverteilungen zu bilden, da ein Großteil der Ergebnisse von Testpersonen mit unverzerrtem Antwortverhalten im unauffälligen oberen bzw. unteren Punktzahlbereich liegt (Larrabee 2014). Zur statistischen Analyse der WMT-, SFSS- und VAS-Ergebnisse wurde daher v.a. der verteilungsfreie U-Test für zwei Parallelgruppen eingesetzt (Gaus u. Muche 2014, S. 302f). Stetige Merkmale, bei denen eine Normalverteilung angenommen werden konnte, wurden mittels t-Test auf signifikante Gruppenunterschiede hin untersucht. Da eigentlich erst ab Fallzahlen von $n > 70$-100 Aussagen über eine möglicherweise vorliegende Normalverteilung getätigt werden sollten (Gaus u. Muche 2014, S. 222), wurde die

Entscheidung für einen t-Test in der vorliegenden Studie konservativ getroffen. In allen Zweifelsfällen kamen verteilungsfreie statistische Tests zur Anwendung.

Im Falle sich signifikant zwischen Interventions- und Kontrollgruppe unterscheidender Ergebnisse von WMT, SFSS und VAS – was für die Wirksamkeit der dargebotenen Instruktionen spräche – sollten die Kennwerte Sensitivität und Spezifität sowie positiver und negativer prädiktiver Wert (PPW bzw. NPW) für WMT und SFSS berechnet werden. Die jeweilige erhaltene Patienteninstruktion galt dabei als durch die BVTs zu detektierende „Diagnose".
Die Kennwerte der BVTs sollten im Rahmen einer ROC(receiver operating characteristic)-Kurven-Analyse für unterschiedliche Cut-off-Werte berechnet werden (vgl. das Vorgehen von Benge et al. für das SIMS (Benge et al. 2012)). Bei einer ROC-Kurven-Analyse wird der Cut-off von niedrigen Grenzwerten zu hohen Grenzwerten variiert, jeweils die Sensitivität und die Spezifität des Testverfahrens berechnet und das Ganze in Form einer Kurve grafisch dargestellt (Gaus u. Muche 2014, S. 142). Die im ROC-Kurven-Schaubild standardmäßig eingezeichnete Diagonale repräsentiert „ein reines Zufallsverfahren ohne jegliche Aussagekraft" (Gaus u. Muche 2014, S. 142), das dem untersuchten Verfahren somit direkt gegenübergestellt wird. Die ROC-Kurve eines guten diagnostischen Verfahrens nähert sich idealerweise der oberen linken Ecke des Schaubildes (Gaus u. Muche 2014, S. 142). Stets mitangegeben wird auch die AUC (area under the curve) der ROC-Kurve. Ein diagnostisches Verfahren mit einer Sensitivität und Spezifität von 100% würde mit einem AUC-Wert von 1 einhergehen, ein Verfahren auf Basis einer reinen Zufallsentscheidung mit einem Wert von .5 (Hedderich u. Sachs 2016, S. 185). Daher gilt der AUC-Wert auch als „Maß für die Güte eines diagnostischen Verfahrens" (Gaus u. Muche 2014, S. 143).

Als Maß für die Güte des jeweiligen variierten Cut-offs von SFSS bzw. WMT wurde im Anschluss an die ROC-Kurven-Analyse für jeden Cut-off der Youden-Index (Youden 1950) bestimmt, der für einen „brauchbaren" Testgrenzwert möglichst Werte nahe 1 annehmen sollte (Youden-Index = Sensitivität + Spezifität – 1; Hedderich u. Sachs 2016, S. 181).
Für die unterschiedlichen Cut-offs sollten jeweils auch der positive prädiktive Wert (PPW) und der negative prädiktive Wert (NPW) der BVTs ermittelt werden. Sie

geben als Kennwerte eines Tests an, welcher Anteil der untersuchten Personen mit positivem bzw. negativem Testergebnis auch tatsächlich krank bzw. gesund ist (Hedderich u. Sachs 2016, S. 180).
Da die Berechnungen von Sensitivität, Spezifität, PPW und NPW für die Cut-offs von >16 (SFSS) bzw. ≤82,5% (WMT) als zentraler Bestandteil der Ergebnisse der vorliegenden Studie anzusehen sind, wurde für diese Werte, entsprechend einer Empfehlung der CONSORT-Leitlinien (Moher et al. 2010), das 95%ige Konfidenzintervall mitangegeben. Außerdem wurden die Klassifikationsergebnisse der BVTs für diese Cut-offs übersichtlich in Vierfeldertafeln dargestellt.

Im weiteren Verlauf der Auswertung wurde eine separate Vierfeldertafel erstellt, mit deren Hilfe die Übereinstimmung von WMT und SFSS bezüglich ihrer Klassifikationen der Studienteilnehmer untersucht werden sollte. Quantifiziert wurde die Übereinstimmung der Testverfahren über die Berechnung des Vierfelderkorrelationskoeffizienten (Φ, phi) (Gaus u. Muche 2014, S. 176) sowie des Kappa-Koeffizienten κ (Cohen 1960), der als Maß für den Grad der Übereinstimmung zweier Verfahren bezüglich einer dichotomen Entscheidung (z.B. WMT bestanden/nicht bestanden, SFSS bestanden/nicht bestanden) fungiert (Gaus u. Muche 2014, S. 184). Der Kappa-Koeffizient variiert zwischen Werten von 0 bis 1 – wobei ein Kappa-Koeffizient von 1 für eine vollständige Übereinstimmung der Verfahren spricht (Gaus u. Muche 2014, S. 184) – während der Vierfelderkorrelationskoeffizient, wie die meisten Korrelationskoeffizienten, Werte zwischen -1 und +1 annehmen kann (Gaus u. Muche 2014, S. 176).

Anhand einer weiteren Vierfeldertafel sollte anschließend untersucht werden, ob die kombinierte Anwendung von WMT und SFSS zu besseren Klassifikationsergebnissen führen würde. Zunächst mussten dafür alle Studienteilnehmer mit widersprüchlichen SFSS- und WMT-Klassifikationen wegen „Klassifikationsunsicherheit“ aus diesem Teil der Auswertung ausgeschlossen werden. Zur Berechnung von Sensitivität, Spezifität, PPW und NPW wurden für diesen Teil der Auswertung dann nur noch die Klassifikationsergebnisse der Studienteilnehmer verwendet, deren BVT-Klassifikationen übereinstimmend waren.

Über die Berechnung der Korrelationskoeffizienten zwischen den Datenpaaren BSI (GSI)/SCIP-Summenscore einerseits und WMT-/SFSS-Score andererseits,

sollte außerdem geprüft werden, ob ein Zusammenhang zwischen der Ausprägung der klinischen Symptomatik bzw. dem kognitiven Leistungsniveau (BSI bzw. SCIP) und den BVT-Ergebnissen (WMT/SFSS) bestand. Zunächst wurden die Daten wiederum mittels Shapiro-Wilk-Test und Histogrammen auf Normalverteilung hin untersucht und anschließend – je nach Ergebnis – der verteilungsfreie Rangkorrelationskoeffizient nach Spearman (Gaus u. Muche 2014, S. 176) oder der Produkt-Moment-Korrelationskoeffizient nach Pearson – der zwei normalverteilte, stetige Merkmale vorrausetzt (Gaus u. Muche 2014, S. 172) – berechnet. Beide Korrelationskoeffizienten variieren zwischen Werten von -1 bis +1(Gaus u. Muche 2014, S. 172ff; Hedderich u. Sachs 2016, S. 123). Für die in der vorliegenden Studie berechneten Korrelationskoeffizienten wurden 95%ige Konfidenzintervalle mitangegeben. Zudem wurden Punktwolken (inklusive Regressionsgeraden) zur graphischen Darstellung aller relevanten Zusammenhänge erstellt. Für jede Regressionsgerade findet sich dabei das sog. Bestimmtheitsmaß R^2 mitaufgeführt, welches angibt, „wie stark die Zielgröße von der Einflussgröße […] beeinflusst wird“ (Gaus u. Muche 2014, S. 89). R^2 kann Werte zwischen 0 (minimaler Wert, kein Einfluss der Einfluss- auf die Zielgröße) und 1 (maximaler Wert, die Zielgröße wird vollständig von der Einflussgröße bestimmt) annehmen (Gaus u. Muche 2014, S. 181).

Im ersten Teil der statistischen Analyse wurden die Daten der Gesamtstichprobe ausgewertet. Alle randomisierten Studienteilnehmer wurden in diesen Teil der Datenanalyse eingeschlossen – unabhängig davon, ob es bei ihnen zu Protokollverletzungen gekommen war oder nicht. Diese Vorgehensweise wird von den CONSORT-Leitlinien empfohlen, um die vorteilhaften Effekte zu erhalten, welche die Randomisierung mit sich bringt (Moher et al. 2010).
Wegen wiederholter schwerer Protokollverletzungen in Gruppe B – im Sinne eines fehlenden Umsetzens von Patienteninstruktion B – wurde jedoch bereits nach Beendigung der Datenerhebung entschieden, zusätzlich zur Auswertung der Gesamtstichprobe auch eine sog. Per Protocol Analyse (PPA) durchzuführen. In die Per Protocol Analyse wurden diejenigen Studienteilnehmer der Gruppe B eingeschlossen, bei denen keine schwerwiegende Protokollverletzung ersichtlich war (Gaus u. Muche 2014, S. 77) – die sich also offensichtlich an Instruktion B gehalten hatten – sowie alle in Gruppe A randomisierten Studienteilnehmer. Die

fehlende bzw. mangelhafte Umsetzung von Instruktion B durch einige Studienteilnehmer in Gruppe B war ersichtlich bei ihren Angaben auf der VAS und bei den protokollierten Antworten in der sich an die Versuchsdurchführung anschließenden Befragung zur Art und Weise des Umsetzens der Patienteninstruktion (s. Patientenkommentare nach Versuchsdurchführung, Anhang C). Letztlich mussten, nach eingehender Analyse der dokumentierten Patientenkommentare (Anhang C) in Zusammenschau mit den VAS-Angaben, 15 Studienteilnehmer in Gruppe B für die Per Protocol Analyse ausgeschlossen werden. Die verbliebene Interventionsgruppe der PPA (n = 15) wird im Folgenden als Gruppe B‘ bezeichnet.

Zielgrößen der vorliegenden Studie waren die Testergebnisse von WMT und SFSS in Kontroll- und Interventionsgruppe. Für die Auswertung der WMT-Ergebnisse wurde – zusätzlich zu den vom Computerprogramm berechneten Prozentwerten – ein separater „Score“ kreiert, indem die Prozentwerte der einfachen WMT-Durchgänge (IR- und DR-Durchgang sowie CNS-Wert) zusammengezählt wurden („Score WMT-Durchgänge IR, DR, CNS“). Dieser Score konnte theoretisch Werte zwischen 0 und 300 annehmen. Darüber hinaus wurde ein weiterer Score aus den Prozentwerten aller verwendeter WMT-Durchgänge (IR, DR, CNS, MC, PA, FR) gebildet – mit möglichen Werten zwischen 0 und 600 – um auf diese Weise das gesamte WMT-Ergebnis abbilden und in Beziehung zu anderen Tests (z.B. BSI/SCIP) setzen zu können („Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR“). Die Bildung dieser beiden Scores erwies sich sowohl für die Berechnung der Korrelationskoeffizienten als auch für die Durchführung der ROC-Kurven-Analyse des WMT als notwendig, da die von Testentwickler Green verwendete Methodik zur Berechnung des für IR-, DR- und CNS-Durchgang empfohlenen Cut-offs von ≤82,5% in keiner vorliegenden Veröffentlichung dargelegt wurde. Beim SFSS wurde für dieselben Zwecke der SFSS-Gesamtscore verwendet.

Als Einflussgröße der vorliegenden Studie ist die zwischen den beiden Gruppen A und B variierte Patienteninstruktion zu verstehen. Mögliche und im Zuge der Randomisierung bestenfalls gleichmäßig verteilte Störgrößen der vorliegenden Studie waren Alter, Geschlecht, Aufnahmediagnosen, Bildungsabschluss, Ort der Rekrutierung (stationär/ambulant), Beschäftigungsverhältnis und Unterschiede in

der Erkrankungs- oder Behandlungsdauer bzw. Behandlungsart der Patienten. Von den möglichen Störgrößen wurden Alter, Geschlecht, Aufnahmediagnosen, Bildungsabschluss, Ort der Rekrutierung und Beschäftigungsverhältnis zu Studienbeginn bei jedem Patienten erfasst.

2.7 Patientenstichprobe

Tabelle 1 zeigt die soziodemographischen Variablen und die BSI- und SCIP-Ergebnisse der Gesamtstichprobe sowie die Verteilung dieser Merkmale nach der Randomisierung in die Gruppen A und B.

Tabelle 1: Verteilung der soziodemographischen Variablen und der BSI- und SCIP-Ergebnisse in der Gesamtstichprobe aller 60 Studienteilnehmer sowie Aufteilung dieser Merkmale nach der Randomisierung in die beiden Studiengruppen A und B (je 30 Studienteilnehmer). Untersuchung der Gruppenunterschiede hinsichtlich statistischer Signifikanz

[a] Chi-Quadrat-Test, [b] t-test, [c] Exakter Fisher-Test, [d] U-Test, MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum, vs.: versus, GSI: Global Severity Index, n: Stichprobenumfang, BSI: Brief Symptom Inventory, SCIP: Screen for Cognitive Impairment in Psychiatry, absolute Häufigkeiten der Studienteilnehmer mit eingeklammerten relativen Häufigkeiten in Prozent

Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Gesamtstichprobe n = 60	**Kontrollgruppe A n = 30**	**Interventionsgruppe B n = 30**	**Signifikanz (2-seitig) der Unterschiede A vs. B**
Geschlecht	Männlich: 19 (31,7%) Weiblich: 41 (68,3%)	Männlich: 11 (36,7%) Weiblich: 19 (63,3%)	Männlich: 8 (26,7%) Weiblich: 22 (73,3%)	.580[c]
Alter (in Jahren)	MW: 50,9 (SD = 5,4) Med: 51 (range = 23)	MW: 51,5 (SD = 5,2) Med: 51,5 (range = 22)	MW: 50,3 (SD = 5,7) Med: 50,5 (range = 18)	.382[b]

Fortsetzung von Tabelle 1 auf Seite 41

Fortsetzung Tabelle 1 von Seite 40

	Gesamtstichprobe n = 60	Kontrollgruppe A n = 30	Interventionsgruppe B n = 30	Signifikanz (2-seitig) der Unterschiede A vs. B
Bildungs-abschluss	kein Abschluss: 2 (3,3%) Hauptschule: 25 (41,7%) Mittlere Reife: 20 (33,3%) Abitur: 4 (6,7%) Fachhochschule: 7 (11,7%) Universität: 2 (3,3%)	kein Abschluss: 1 (3,3%) Hauptschule: 16 (53,3%) Mittlere Reife: 8 (26,7%) Abitur: 1 (3,3%) Fachhochschule: 4 (13,3%) Universität: 0 (0%)	kein Abschluss: 1 (3,3%) Hauptschule: 9 (30%) Mittlere Reife): 12 (40%) Abitur: 3 (10%) Fachhochschule: 3 (10%) Universität: 2 (6,7%)	.103[d]
Aufnahme-diagnose(n)	Summe: 69 Diagnosen F32.x: 18 von 69 Diagnosen (26,1%) F33.x: 32 von 69 Diagnosen (46,4%) F40.x: 4 von 69 Diagnosen (5,8%) F41.x: 9 von 69 Diagnosen (13%) F43.2: 6 von 69 Diagnosen (8,7%)	Summe: 34 Diagnosen F32.x: 9 von 34 Diagnosen (26,5%) F33.x: 16 von 34 Diagnosen (47,1%) F40.x: 2 von 34 Diagnosen (5,9%) F41.x: 4 von 34 Diagnosen (11,8%) F43.2: 3 von 34 Diagnosen (8,8%)	Summe: 35 Diagnosen F32.x: 9 von 35 Diagnosen (25,7%) F33.x: 16 von 35 Diagnosen (45,7%) F40.x: 2 von 35 Diagnosen (5,7%) F41.x: 5 von 35 Diagnosen (14,3%) F43.2: 3 von 35 Diagnosen (8,6%)	 1.0[c] 1.0[c] 1.0[c] 1.0[c] 1.0[c]

Fortsetzung von Tabelle 1 auf Seite 42

Fortsetzung Tabelle 1 von Seite 41

	Gesamtstichprobe n = 60	Kontrollgruppe A n = 30	Interventionsgruppe B n = 30	Signifikanz (2-seitig) der Unterschiede A vs. B
Erwerbsminderungsrente bereits ≥ einmal abgelehnt	Ja: 5 (8,3%) Nein: 55 (91,7%)	Ja: 2 (6,7%) Nein: 28 (93,3%)	Ja: 3 (10%) Nein: 27 (90%)	1.0[c]
Rekrutierung in stationärer/ ambulanter Behandlung	Stationär: 27 (45%) Ambulant: 33 (55%)	Stationär: 13 (43,3%) Ambulant: 17 (56,7%)	Stationär: 14 (46,7%) Ambulant: 16 (53,3%)	1.0[c]
Beschäftigungsverhältnis	Erwerbstätig: 43 (71,7%) Erwerbslos: 10 (16,7%) Bereits berentet: 7 (11,7%)	Erwerbstätig: 21 (70%) Erwerbslos: 4 (13,3%) Bereits berentet: 5 (16,7%)	Erwerbstätig: 22 (73,3%) Erwerbslos: 6 (20%) Bereits berentet: 2 (6,7%)	.426[a]
Brief Symptom Inventory (BSI)	*GSI:* MW: 1,2 (SD = 0,6) Med: 1,2 (range = 2,7)	*GSI:* MW: 1,3 (SD = 0,7) Med: 1,1 (range = 2,5)	*GSI:* MW: 1,1 (SD = 0,6) Med: 1,3 (range = 2,5)	.446[b]
Screen for Cognitive Impairment in Psychiatry (SCIP)	*Summenscore:* MW: 73 (SD = 10,9) Med: 73,5 (range = 52)	*Summenscore:* MW: 72,9 (SD = 10,8) Med: 75,5 (range = 44)	*Summenscore:* MW: 73,1 (SD = 11,2) Med: 72 (range = 44)	.963[b]

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Statistisch signifikante Unterschiede zwischen den Gruppen A und B konnten für keines der in Tabelle 1 genannten Merkmale nachgewiesen werden.

In Bezug auf soziodemographische Variablen wie Alter ($p = .382$), Geschlechterverteilung ($p = .580$), Beschäftigungsverhältnis ($p = .426$), Bildungsabschluss ($p = .103$), Aufnahmediagnosen (jeweils $p = 1.0$), Umstände der Rekrutierung ($p = 1.0$), Vorliegen abgelehnter Anträge auf eine Erwerbsminderungsrente ($p = 1.0$) sowie auch hinsichtlich der aktuellen Symptomschwere – erfasst über das BSI ($p = .446$) – und des kognitiven Leistungsniveaus – erfasst über das SCIP ($p = .963$) – konnte kein statistisch signifikanter Unterschied zwischen den Gruppen A und B festgestellt werden. Es ist also davon auszugehen, dass die Randomisierung erfolgreich war und dass sich die Gruppen A und B bezüglich oben genannter Merkmale bzw. Störgrößen nicht signifikant unterscheiden.

3 Ergebnisse

An dieser Stelle sei zunächst darauf verwiesen, dass der Schwerpunkt der Arbeit auf den Ergebnissen der Per Protocol Analyse liegt. Aus methodischen Gründen werden jedoch auch die Ergebnisse der Auswertung der Gesamtstichprobe aufgeführt.

3.1 Auswertung der Gesamtstichprobe

3.1.1 SFSS-, WMT- und VAS-Ergebnisse der Gruppen A und B

Tabelle 2 fasst die in den Gruppen A und B erzielten Ergebnisse bei WMT, SFSS und VAS zusammen. Die WMT- und SFSS-Ergebnisse sind jeweils auch mit den Ergebnissen der einzelnen Durchgänge (WMT) bzw. Subskalen (SFSS) dargestellt.

Tabelle 2: Gegenüberstellung der bei WMT, SFSS und VAS erzielten Ergebnisse in den beiden Studiengruppen A und B (je 30 Studienteilnehmer). Auftrennung der Testergebnisse von WMT und SFSS in die verschiedenen Testdurchgänge (WMT) bzw. Subskalen (SFSS). Untersuchung der Gruppenunterschiede hinsichtlich statistischer Signifikanz
[a] t-Test, [b] U-Test, MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum - Minimum, vs.: versus, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MC: Multiple Choice, PA: Paired Associate Recall, FR: Free Recall, WMT: Word Memory Test, SFSS: Strukturierter Fragebogen Simulierter Symptome, VAS: Visuelle Analogskala, n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Kontrollgruppe A n = 30	**Interventionsgruppe B n = 30**	**Signifikanz (2-seitig) der Unterschiede A vs. B**
Word Memory Test (WMT)			
- **IR** (in % richtig)	MW: 96,6 (SD = 8,2) Med: 100 (range = 42)	MW: 80,5 (SD = 22,9) Med: 88 (range = 77)	.001[b***]
- **DR** (in % richtig)	MW: 96,3 (SD = 8,4) Med: 100 (range = 42)	MW: 81,5 (SD = 19,6) Med: 90 (range = 75)	< .001[b***]

Fortsetzung von Tabelle 2 auf Seite 45

Fortsetzung Tabelle 2 von Seite 44

	Kontrollgruppe A n = 30	**Interventionsgruppe B n = 30**	**Signifikanz (2-seitig) der Unterschiede A vs. B**
- **CNS** (in % richtig)	MW: 95 (SD = 10,1) Med: 100 (range = 50)	MW: 78,5 (SD = 18,8) Med: 80,5 (range = 57)	< .001[b***]
- **MC** (in % richtig)	MW: 87,8 (SD = 18,6) Med: 95 (range = 80)	MW: 69,3 (SD = 25) Med: 70 (range = 75)	.002[b**]
- **PA** (in % richtig)	MW: 85,7 (SD = 17,8) Med: 92,5 (range = 75)	MW: 63,5 (SD = 24,2) Med: 62,5 (range = 90)	< .001[b***]
- **FR** (in % richtig)	MW: 68,5 (SD = 17,7) Med: 73 (range = 75)	MW: 43,1 (SD = 17,4) Med: 40 (range = 72)	< .001[b***]
Score WMT-Durchgänge IR, DR, CNS	MW: 287,9 (SD = 26,3) Med: 300 (range = 134)	MW: 240,5 (SD = 58,1) Med: 256 (range = 174)	< .001[b***]
Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR	MW: 530 (SD = 75,5) Med: 559 (range = 364)	MW: 416,5 (SD = 109,6) Med: 417,5 (range = 356)	< .001[b***]
Strukturierter Fragebogen Simulierter Symptome (SFSS)	*Gesamtscore:* MW: 14,1 (SD = 6,5) Med: 13 (range = 25)	*Gesamtscore:* MW: 25,4 (SD = 11,7) Med: 26,5 (range = 40)	< .001[b***]

Fortsetzung von Tabelle 2 auf Seite 46

Fortsetzung Tabelle 2 von Seite 45

	Kontrollgruppe A n = 30	**Interventionsgruppe B n = 30**	**Signifikanz (2-seitig) der Unterschiede A vs. B**
SFSS-Skala neurologische Beeinträchtigung (N, 15 Items**)**	MW: 3,5 (SD = 2,3) Med: 3 (range = 9)	MW: 6,6 (SD = 3,7) Med: 6 (range = 14)	.001[b***]
SFSS-Skala affektive Störungen (AF, 15 Items**)**	MW: 6,5 (SD = 2,6) Med: 6,5 (range = 11)	MW: 8,4 (SD = 3,4) Med: 9 (range = 14)	.018[a*]
SFSS-Skala niedrige Intelligenz (NI, 15 Items**)**	MW: 1,3 (SD = 1,2) Med: 1 (range = 4)	MW: 3 (SD = 1,9) Med: 2,5 (range = 8)	< .001[b***]
SFSS-Skala Psychose (P, 15 Items**)**	MW: 0,9 (SD = 1,1) Med: 1 (range = 5)	MW: 2,3 (SD = 2,7) Med: 1 (range = 9)	.064[b]
SFSS-Skala amnestische Störung (AM, 15 Items**)**	MW: 1,9 (SD = 2,2) Med: 1 (range = 8)	MW: 5,1 (SD = 3,5) Med: 5,5 (range = 11)	< .001[b***]
Visuelle Analogskala (VAS), subjektive Anstrengung in %	MW: 99,7 (SD = 1,1) Med: 100 (range = 6)	MW: 74,8 (SD = 24,4) Med: 77,5 (range = 73)	< .001[b***]

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Für beinahe alle in Tabelle 2 beschriebenen Merkmale zeigten sich jeweils statistisch signifikante Unterschiede zwischen den Gruppen A und B. Einzig und allein der Unterschied der auf der SFSS-P-Skala erzielten Werte verfehlte das

Signifikanzniveau knapp ($p = .064$). Hervorzuheben ist vor allem, dass die von Studieneilnehmern in Gruppe B über die VAS angegebene subjektive Anstrengung signifikant niedriger war als die subjektiv angegebene Anstrengung in Gruppe A ($p < .001$). In Gruppe A lag die subjektiv angegebene Anstrengung während der Testdurchführung im Mittel bei 99,7% (Median: 100%), gegenüber einem Mittelwert von 74,8% in Gruppe B (Median: 77,5%). Studienteilnehmer in Gruppe A erzielten außerdem signifikant niedrigere SFSS-Gesamtscores als Studienteilnehmer in Gruppe B (Mittelwert: 14,1 in Gruppe A gegenüber Mittelwert: 25,4 in Gruppe B, $p < .001$). Sowohl die Werte im Score WMT-Durchgänge IR, DR, CNS (Mittelwert: 287,9 in Gruppe A gegenüber Mittelwert: 240,5 in Gruppe B) als auch die Werte im Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR (Mittelwert: 530 in Gruppe A gegenüber Mittelwert: 416,5 in Gruppe B) waren hingegen in Gruppe A signifikant höher als in Gruppe B ($p < .001$ bzw. $p < .001$). Aufgrund dieser Ergebnisse wurde angenommen, dass die Instruktionen A und B in den Gruppen wirksam waren und alle weiteren geplanten statistischen Analysen konnten durchgeführt werden.

3.1.1.1 Analyse der SFSS-Ergebnisse

Eine eingehendere deskriptive Analyse der SFSS-Ergebnisse in den Gruppen A und B findet sich in Tabelle 3.

Tabelle 3: Deskriptive Darstellung und Analyse der SFSS-Ergebnisse aller 60 Studieneilnehmer sowie Gegenüberstellung der in den beiden Studiengruppen A und B (je 30 Studienteilnehmer) erzielten SFSS-Ergebnisse. Ein SFSS-Gesamtscore von >16 Punkten wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests. Auflistung bestimmter häufig von Studienteilnehmern angekreuzter Testitems
MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum, SFSS: Strukturierter Fragebogen Simulierter Symptome, AF-Skala: Skala für affektive Störungen (SFSS), Nr.: Nummer, n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Gesamt-stichprobe n = 60	**Kontrollgruppe A n = 30**	**Interventions-gruppe B n = 30**
SFSS-Gesamtscore >16	31 von 60 (51,7%) Studienteil-nehmern	9 von 30 (30%) Studienteilnehmern	22 von 30 (73,3%) Studienteilnehmern
Anteil AF-Skala-Ergebnis an SFSS-Gesamtscore (Anteile angegeben in %)	MW: 41,6% (SD = 14,4%) Med: 41,2% (range = 54,8%)	MW: 48,3% (SD = 14,2%) Med: 49,7% (range = 54,8%)	MW: 34,8% (SD = 11,1%) Med: 32,9% (range = 44%)

Fortsetzung von Tabelle 3 auf Seite 49

Fortsetzung Tabelle 3 von Seite 48

	Gesamt-stichprobe n = 60	**Kontrollgruppe A n = 30**	**Interventions-gruppe B n = 30**
Item Nr. 2: Wenn meine Depression zu stark wird, mache ich lange Spaziergänge oder Sport, um die Spannung zu vermindern.	27 von 60 (45%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	20 von 30 (66,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“	7 von 30 (23,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“
Item Nr. 19: Manchmal bin ich so deprimiert, dass ich mich nach meinem Bett sehne, um die schlechte Stimmung weg-zuschlafen.	43 von 60 (71,7%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	20 von 30 (66,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“	23 von 30 (76,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“
Item Nr. 32: Ich kann schlecht schlafen.	44 von 60 (73,3%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	20 von 30 (66,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“	24 von 30 (80%) Studienteilnehmern beantworteten dieses Item mit „Ja“

Fortsetzung von Tabelle 3 auf Seite 50

Fortsetzung Tabelle 3 von Seite 49

	Gesamt-stichprobe n = 60	**Kontrollgruppe A n = 30**	**Interventions-gruppe B n = 30**
Item Nr. 35: Manchmal sind meine Muskeln ohne ersichtlichen Grund so schlaff und kraftlos, dass sich meine Arme und Beine anfühlen, als ob sie zentnerschwer wären.	41 von 60 (68,3%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	19 von 30 (63,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“	22 von 30 (73,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“
Item Nr. 52: Es scheint, dass ich nicht mehr die Energie habe, die ich früher immer gehabt habe.	55 von 60 (91,7%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	28 von 30 (93,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“	27 von 30 (90%) Studienteilnehmern beantworteten dieses Item mit „Ja“

Bei der Betrachtung von Tabelle 3 fällt auf, dass über die Hälfte (51,7%, 31 von 60 Studienteilnehmern) aller Studienteilnehmer den SFSS bei seinem üblichen Cut-off von >16 nicht bestand, also einen höheren SFSS-Gesamtscore als 16 erzielte. Neben 22 von 30 Teilnehmern (73,3%) der Interventionsgruppe B bestanden auch 9 von 30 Studienteilnehmern (30%) in Kontrollgruppe A den SFSS für den Cut-off von >16 nicht. Der Anteil des AF-Skala-Ergebnisses am SFSS-Gesamtscore betrug in Kontrollgruppe A im Mittel 48,3% (vgl. 34,8% in Gruppe B). Tabelle 3

listet zudem die SFSS-Items Nr. 2, 19, 32, 35 und 52 auf, bei denen besonders viele Studienteilnehmer eine – laut SFSS-Decodierungstabelle – als auffällig einzustufende Antwort ankreuzten (bei diesen Items jeweils die Antwort „Ja“). Bei allen dieser Items lag der Anteil der Studienteilnehmer in Kontrollgruppe A, die das Item bejaht hatten, bei 63,3% und höher. Item Nr. 52 wurde sogar von 93,3% (28 von 30 Studienteilnehmern) der Studienteilnehmer in Gruppe A bejaht.

3.1.1.2 Analyse der WMT-Ergebnisse

Eine eingehendere deskriptive Analyse der WMT-Ergebnisse in den Gruppen A und B zeigt Tabelle 4.

Tabelle 4: Deskriptive Darstellung und Analyse der WMT-Ergebnisse aller 60 Studieneilnehmer sowie Gegenüberstellung der in den beiden Studiengruppen A und B (je 30 Studienteilnehmer) erzielten WMT-Ergebnisse. Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests. Betrachtung weiterer relevanter Ergebnisbereiche des WMT
IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, WMT: Word Memory Test, n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Gesamt-stichprobe n = 60	**Kontrollgruppe A n = 30**	**Interventionsgruppe B n = 30**
IR- oder DR- oder CNS-Ergebnis (WMT) ≤82,5%	17 von 60 (28,3%) Studienteilnehmern	2 von 30 (6,7%) Studienteilnehmern	15 von 30 (50%) Studienteilnehmern
IR- oder DR- oder CNS-Ergebnis (WMT) ≥50% - ≤60%	7 von 60 (11,7%) Studienteilnehmern	1 von 30 (3,3%) Studienteilnehmern	6 von 30 (20%) Studienteilnehmern
IR- oder DR- oder CNS-Ergebnis (WMT) <50%	5 von 60 (8,3%) Studienteilnehmern dabei: 3 von 60 (5%) Studienteilnehmern mit IR- oder DR-Ergebnis <37,5%	0 von 30 (0%) Studienteilnehmern dabei: 0 von 30 (0%) Studienteilnehmern mit IR- oder DR-Ergebnis <37,5%	5 von 30 (16,7%) Studienteilnehmern dabei: 3 von 30 (10%) Studienteilnehmern mit IR- oder DR-Ergebnis <37,5%

Aus Tabelle 4 wird ersichtlich, dass 2 von 30 Studienteilnehmern (6,7%) in Kontrollgruppe A den WMT bei Anwendung des üblichen Cut-offs von ≤82,5% nicht bestanden, während 15 von 30 Studienteilnehmern (50%) in der Interventionsgruppe B ein Ergebnis unterhalb dieses Cut-offs erzielten. Weiterhin fällt auf, dass 7 von 60 Studienteilnehmern der Gesamtstichprobe (11,7%) in den „einfachen“ WMT-Durchgängen (IR, DR, CNS) ein Ergebnis zwischen 50% und 60% erzielten – davon 6 Studienteilnehmer in Gruppe B. Allerdings schloss auch einer von 30 Studienteilnehmern (3,3%) in Kontrollgruppe A die „einfachen“ WMT-Durchgänge mit einem Ergebnis zwischen 50% und 60% ab. Ein WMT-Ergebnis <50% bzw. <37,5% (= Unter-Zufall-Schwelle) im IR-, DR- oder CNS-Durchgang erzielten ausschließlich Studienteilnehmer der Interventionsgruppe B (5 von 30 Studienteilnehmern (16,7%) bzw. 3 von 30 Studienteilnehmern (10%)).

3.1.2 Sensitivität, Spezifität, PPW und NPW von SFSS und WMT

Die Tabellen 5 und 6 zeigen jene Vierfeldertafeln, die zur Berechnung von Sensitivität, Spezifität, PPW und NPW von SFSS und WMT in der Gesamtstichprobe für die gebräuchlichen Cut-offs (>16 (SFSS) bzw. ≤82,5% (WMT)) benutzt wurden. Die Vierfeldertafeln sollen beispielhaft darstellen, nach welchem Prinzip das Statistikprogramm IBM® SPSS® Statistics 21 im Rahmen der ROC-Kurven-Analyse für jeden variierten Cut-off Sensitivität und Spezifität errechnete und zudem eine Übersicht über die BVT-Klassifikationen in Kontroll- und Interventionsgruppe der Gesamtstichprobe bieten.

Tabelle 5: Darstellung der Klassifikationen aller 60 Studienteilnehmer für den WMT-Cut-off von ≤82,5% in Form einer Vierfeldertafel. Unterteilung der Klassifikationen nach den beiden Studiengruppen A und B (je 30 Studienteilnehmer). Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests
WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	WMT-Ergebnis ≤82,5% (auffällig)	**WMT-Ergebnis >82,5% (unauffällig)**	***Summe***
experimentelle negative Antwortverzerrer (Gruppe B)	15	15	30
subjektiv bestmögliche Leistungsanstrengung/ unverzerrtes Antwort-verhalten (Gruppe A)	2	28	30
Summe	17	43	60

Für den üblichen WMT-Cut-off von ≤82,5% bzw. obige Vierfeldertafel ergaben sich folgende Testkennwerte:

Sensitivität: 50% (95% KI = 32,7% - 67,3%)

Spezifität: 93,3% (95% KI = 80,8% - 98,9%)

Positiver prädiktiver Wert (PPW): 88,2% (95% KI = 67,9% - 97,9%)

Negativer prädiktiver Wert (NPW): 65,1% (95% KI = 50,3% - 78,2%).

Tabelle 6 stellt die Klassifikationsergebnisse des SFSS in der Gesamtstichprobe für den Cut-off von >16 dar.

Tabelle 6: Darstellung der Klassifikationen aller 60 Studienteilnehmer für den SFSS-Cut-off von >16 in Form einer Vierfeldertafel. Unterteilung der Klassifikationen nach den beiden Studiengruppen A und B (je 30 Studienteilnehmer). Ein SFSS-Gesamtscore von >16 Punkten wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests
SFSS: Strukturierter Fragebogen Simulierter Symptome
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	SFSS-Ergebnis >16 (auffällig)	**SFSS-Ergebnis ≤16 (unauffällig)**	***Summe***
experimentelle negative Antwortverzerrer (Gruppe B)	22	8	30
subjektiv bestmögliche Leistungsanstrengung/ unverzerrtes Antwort-verhalten (Gruppe A)	9	21	30
Summe	31	29	60

Für den üblichen SFSS-Cut-off von >16 bzw. obige Vierfeldertafel ergaben sich folgende Testkennwerte:

Sensitivität: 73,3% (95% KI = 56,0% - 86,8%)

Spezifität: 70% (95% KI = 52,4% - 84,3%)

Positiver prädiktiver Wert (PPW): 71% (95% KI = 53,8% - 84,8%)

Negativer prädiktiver Wert (NPW): 72,4% (95% KI = 54,7% - 86,3%).

3.1.2.1 Falsch-positive SFSS-Klassifikationen

Tabelle 7 fasst einige Merkmale von Studienteilnehmern mit falsch-positiver SFSS-Klassifikation in Gruppe A (bei einem Cut-off von >16) zusammen. Gegenübergestellt werden den falsch-positiv Klassifizierten die richtig-negativ klassifizierten Studienteilnehmer der Kontrollgruppe A. Die Teilgruppe A (n = 30) kann nur falsch-positive und richtig-negative Klassifikationen enthalten.

Tabelle 7: Darstellung der Häufigkeit falsch-positiver und richtig-negativer Klassifikationen durch den SFSS bei einem Cut-off von >16 in Teilgruppe A (beinhaltet 30 der insgesamt 60 Studienteilnehmer). Eingehen auf die SCIP- und BSI-Ergebnisse der falsch-positiv klassifizierten Studienteilnehmer und Gegenüberstellung der Ergebnisse der richtig-negativ klassifizierten Studienteilnehmer in Teilgruppe A. Ein SFSS-Gesamtscore von >16 Punkten wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests
SCIP: Screen for Cognitive Impairment in Psychiatry, BSI: Brief Symptom Inventory, GSI: Global Severity Index, SFSS: Strukturierter Fragebogen Simulierter Symptome, MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Falsch-positive SFSS-Klassifikationen in Gruppe A	**Richtig-negative SFSS-Klassifikationen in Gruppe A**
Häufigkeit	9 von 30 (30%) Studienteilnehmern	21 von 30 (70%) Studienteilnehmern
Screen for Cognitive Impairment in Psychiatry (SCIP)	*Summenscore:* MW: 68,9 (SD = 12,2) Med: 70 (range = 42)	*Summenscore:* MW: 74,7 (SD = 9,9) Med: 77 (range = 33)
Brief Symptom Inventory (BSI)	*GSI:* MW: 1,8 (SD = 0,6) Med: 1,5 (range = 1,9)	*GSI:* MW: 1,1 (SD = 0,6) Med: 0,9 (range = 2,5)

Aus Tabelle 7 wird ersichtlich, dass durch den SFSS 9 von 30 Studienteilnehmern (30%) in Gruppe A bei einem Cut-off von >16 falsch-positiv klassifiziert wurden, also ein Ergebnis über diesem Cut-off erzielten. Die falsch-positiv Klassifizierten wiesen mit 68,9 (Median: 70) einen im Mittel niedrigeren SCIP-Summenscore auf

als die richtig-negativ klassifizierten Studienteilnehmer (Mittelwert: 74,7; Median: 77). Der mittlere GSI-Wert der falsch-positiv Klassifizierten war mit 1,8 (Median: 1,5) hingegen höher als der mittlere GSI-Wert der richtig-negativ klassifizierten Studienteilnehmer in Gruppe A (Mittelwert: 1,1; Median: 0,9).

3.1.2.2 Falsch-positive WMT-Klassifikationen

In Tabelle 8 werden die Merkmale der beiden Studienteilnehmer in Gruppe A, die durch den WMT falsch-positiv klassifiziert wurden (bei einem Cut-off von ≤82,5%), den Merkmalen der richtig-negativ klassifizierten Studienteilnehmer der Kontrollgruppe A gegenübergestellt. Die Teilgruppe A (n = 30) kann nur falsch-positive und richtig-negative Klassifikationen enthalten.

Tabelle 8: Darstellung der Häufigkeit falsch-positiver und richtig-negativer Klassifikationen durch den WMT bei einem Cut-off von ≤82,5% in Teilgruppe A (beinhaltet 30 der insgesamt 60 Studienteilnehmer). Eingehen auf die SCIP- und BSI-Ergebnisse der falsch-positiv klassifizierten Studienteilnehmer und Gegenüberstellung der Ergebnisse der richtig-negativ klassifizierten Studienteilnehmer in Teilgruppe A. Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests
SCIP: Screen for Cognitive Impairment in Psychiatry, BSI: Brief Symptom Inventory, GSI: Global Severity Index, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Falsch-positive WMT-Klassifikationen in Gruppe A	**Richtig-negative WMT-Klassifikationen in Gruppe A**
Häufigkeit	2 von 30 (6,7%) Studienteilnehmern	28 von 30 (93,3%) Studienteilnehmern
Screen for Cognitive Impairment in Psychiatry (SCIP)	*Summenscore:* MW: 49,5 (SD = 7,8) Med: 49,5 (range = 11)	*Summenscore:* MW: 74,6 (SD = 8,9) Med: 76,5 (range = 30)
Brief Symptom Inventory (BSI)	*GSI:* MW: 1,3 (SD = 1,3) Med: 1,3 (range = 1,9)	*GSI:* MW: 1,3 (SD = 0,6) Med: 1,1 (range = 2,5)

Tabelle 8 zeigt, dass 2 von 30 Studienteilnehmern (6,7%) in Gruppe A durch den WMT bei einem Cut-off von ≤82,5% falsch-positiv klassifiziert wurden. Sie erzielten also ein IR-, DR-, oder CNS-Ergebnis unter dem Cut-off von ≤82,5%. Die beiden falsch-positiv klassifizierten Studienteilnehmer wiesen mit 49,5 (Median: 49,5) einen im Mittel deutlich niedrigeren SCIP-Summenscore auf als die richtig-negativ klassifizierten Studienteilnehmer der Kontrollgruppe A (Mittelwert: 74,6; Median: 76,5). Der mittlere GSI-Wert der falsch-positiv Klassifizierten nahm mit 1,3 (Median: 1,3) hingegen den gleichen Wert an wie der mittlere GSI-Wert der richtig-negativ klassifizierten Studienteilnehmer in Gruppe A (Mittelwert: 1,3; Median: 1,1).

3.1.3 ROC-Kurven-Analyse

3.1.3.1 ROC-Kurven-Analyse SFSS

Abbildung 1 zeigt den Verlauf der ROC-Kurve des SFSS bei n = 60 (Auswertung der Gesamtstichprobe). In Grün gegenübergestellt wird der SFSS-ROC-Kurve eine Diagonale, die den Verlauf der ROC-Kurve eines reinen Zufallsverfahrens wiedergibt.

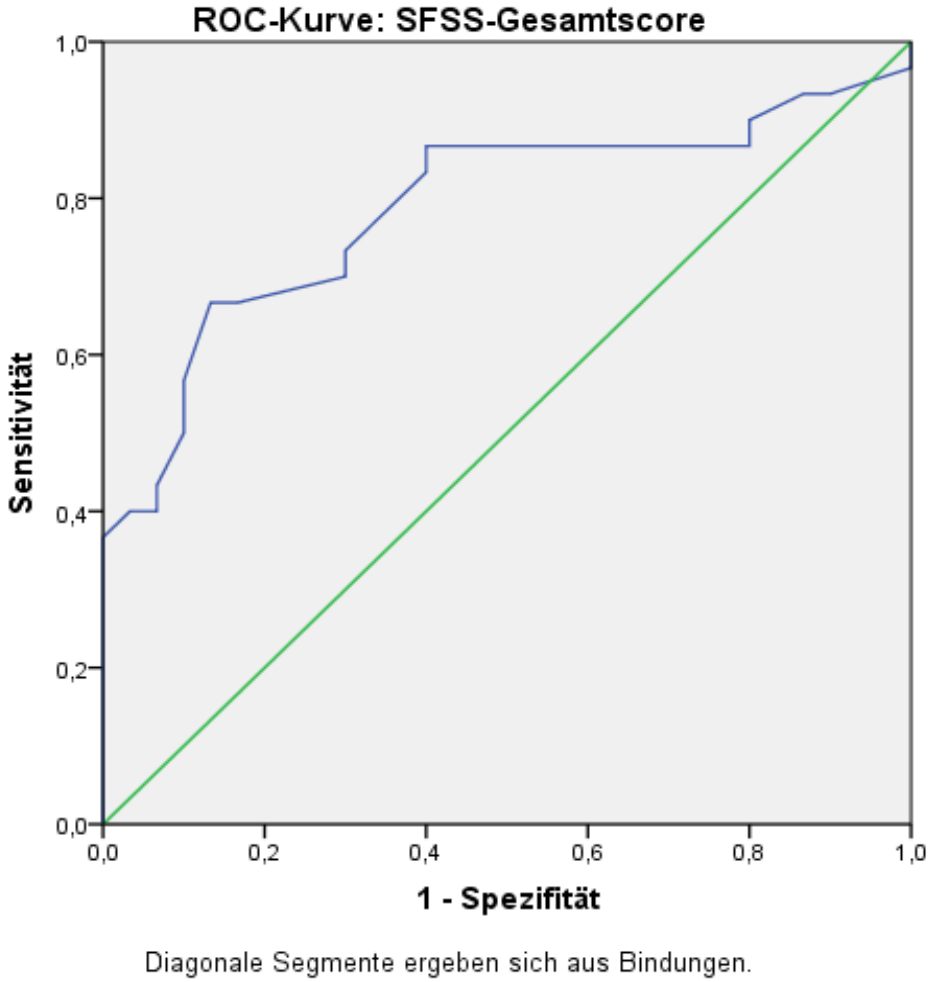

Abbildung 1: ROC-Kurven-Analyse für den SFSS-Gesamtscore aller 60 Studienteilnehmer. Die ROC-Kurve eines diagnostischen Zufallsverfahrens (grün, diagonal) ist der ROC-Kurve des SFSS (blau) zu Vergleichszwecken gegenübergestellt. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde sich als Punkt in der oberen linken Ecke darstellen
SFSS: Strukturierter Fragebogen Simulierter Symptome, ROC: receiver operating characteristic
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Die in Abbildung 1 blau eingezeichnete ROC-Kurve des SFSS setzt sich größtenteils deutlich von der in grün eingezeichneten ROC-Kurve des Zufallsverfahrens ab. Bei sehr hohen Sensitivitäten allerdings – die laut Abbildung

1 mit äußerst niedrigen Spezifitäten einhergehen – unterschreitet die SFSS-ROC-Kurve die ROC-Kurve des Zufallsverfahrens. Insgesamt spricht das Schaubild in Abbildung 1 dennoch dafür, dass der SFSS als Unterscheidungsverfahren zwischen den in der vorliegenden Studie gebildeten Zuständen „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“ und „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ – rein graphisch beurteilt – bei den meisten Cut-offs aussagekräftiger ist als ein Zufallsverfahren.

3.1.3.1.1 Fläche unter der ROC-Kurve (SFSS)

In Tabelle 9 findet sich die für das Schaubild in Abbildung 1 berechnete Fläche unter der ROC-Kurve (AUC, maximaler Wert = 1) mit 95%igem Konfidenzintervall. Sie ist ein Maß für die Güte des untersuchten diagnostischen Verfahrens.

Tabelle 9: Darstellung des für die Fläche unter der SFSS-ROC-Kurve aller 60 Studienteilnehmer berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde einen AUC-Wert von 1 erzielen

[a] Unter der nicht-parametrischen Annahme, [b] Nullhypothese: Wahrheitsfläche = .5, SFSS: Strukturierter Fragebogen Simulierter Symptome, ROC: receiver operating characteristic, AUC: area under the curve

Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Fläche unter ROC-Kurve des SFSS (AUC)	**Standardfehler[a]**	**Asymptotische Signifikanz[b]**	**Asymptotisches 95% Konfidenzintervall**	
			Untergrenze	Obergrenze
.79	.06	< .001***	.67	.91

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Der für die AUC der SFSS-ROC-Kurve berechnete Wert von .79 unterschied sich statistisch signifikant (p < .001) vom Wert der AUC des Zufallsverfahrens (AUC = .5). Dieser Umstand spricht für den SFSS als brauchbares Unterscheidungsverfahren zwischen den Zuständen „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ und „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“.

3.1.3.1.2 Koordinaten der ROC-Kurve (SFSS)

Tabelle 10 enthält die Koordinaten der in Abbildung 1 dargestellten SFSS-ROC-Kurve und listet dazu in der ersten Spalte die verschiedenen SFSS-Cut-offs auf, für die jeweils Sensitivität, Spezifität, PPW und NPW des SFSS in der vorliegenden Gesamtstichprobe berechnet wurden. Ein in der ersten Spalte angegebener Cut-off von 6,5 z.B. bedeutet, dass der zugehörige SFSS-Cut-off für eine Klassifikation in den Zustand „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ effektiv bei >6 liegt, da eigentlich nur ganzzahlige SFSS-Gesamtscores möglich sind. Die letzte Spalte von Tabelle 10 enthält das ebenfalls für jeden Cut-off berechnete Maß für die Güte des jeweiligen Cut-offs (Youden-Index, maximaler Wert = 1). Die für die Berechnung von PPW und NPW angenommene Basalrate negativer Antwortverzerrung in der Population lag bei 50% (30 von 60 Studienteilnehmern).

Tabelle 10: Auflistung der Koordinaten der SFSS-ROC-Kurve aller 60 Studienteilnehmer. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten SFSS-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind. Die variierten Cut-offs beziehen sich auf den SFSS-Gesamtscore. Der höchste Youden-Index-Wert ist rot hervorgehoben
Basalrate: Basalrate negativer Antwortverzerrung in der untersuchten Population, SFSS: Strukturierter Fragebogen Simulierter Symptome, PPW: positiver prädiktiver Wert, NPW: negativer prädiktiver Wert, ROC: receiver operating characteristic, Youden-Index: Sensitivität + Spezifität – 1
[a] Der kleinste Cut-off entspricht dem kleinsten beobachteten Testwert minus 1, der größte Cut-off dem größten beobachteten Testwert plus 1. Alle anderen Cut-offs sind Mittelwerte zweier aufeinanderfolgender, geordneter beobachteter Testwerte
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Positives Test-ergebnis, wenn größer oder gleich[a] (Variable: SFSS-Gesamtscore)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 50%)*	**NPW (%)** *(Basalrate: 50%)*	**Youden-Index**
4	100	0	-	-	0
5,5	96,7	0	49,2	0	-0,033
6	93,3	10	50,9	59,9	0,033

Fortsetzung von Tabelle 10 auf Seite 62

Fortsetzung Tabelle 10 von Seite 61

Positives Test-ergebnis, wenn größer oder gleich[a] (Variable: SFSS-Gesamtscore)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 50%)*	**NPW (%)** *(Basalrate: 50%)*	**Youden-Index**
6,5	93,3	13,3	51,8	66,5	0,066
8	90	20	52,9	66,7	0,1
9,1	86,7	20	52	60,1	0,067
10,1	86,7	23,3	53,1	63,7	0,1
11,5	86,7	26,7	54,2	66,8	0,134
12,5	86,7	43,3	60,5	76,5	0,3
13,1	86,7	60	68,4	81,9	0,467
13,6	83,3	60	67,6	78,2	0,433
14,5	80	63,3	68,6	76	0,433
15,5	76,7	66,7	69,7	74,1	0,434
16,1	73,3	70	71	72,4	0,433
16,6	70	70	70	70	0,4
18	66,7	83,3	80	71,4	0,5
20,5	66,7	86,7	83,4	72,3	**0,534**
23,5	56,7	90	85	67,5	0,467
25,5	53,3	90	84,2	65,8	0,433
26,5	50	90	83,3	64,3	0,4
27,5	43,3	93,3	86,6	62,2	0,366
28,5	40	93,3	85,7	60,9	0,333
30	40	96,7	92,4	61,7	0,367

Fortsetzung von Tabelle 10 auf Seite 63

Fortsetzung Tabelle 10 von Seite 62

Positives Test-ergebnis, wenn größer oder gleich[a] (Variable: SFSS-Gesamtscore)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 50%)*	**NPW (%)** *(Basalrate: 50%)*	**Youden-Index**
31,5	36,7	100	~100	61,2	0,367
33	33,3	100	~100	60	0,333
35	23,3	100	~100	56,6	0,233
36,5	20	100	~100	55,5	0,2
37,5	16,7	100	~100	54,5	0,167
38,5	13,3	100	~100	53,5	0,133
40,5	10	100	~100	52,6	0,1
43	6,7	100	~100	51,7	0,067
44,5	3,3	100	~100	50,8	0,033
46	0	100	-	-	0

Der höchste Wert für den Youden-Index (= 0,534) ergab sich laut Tabelle 10 für einen SFSS-Cut-off von ≥20,5, also effektiv >20, da beim SFSS eigentlich nur ganzzahlige Gesamtscores möglich sind. Dieser Cut-off ging mit einer Sensitivität von 66,7%, einer Spezifität von 86,7% und einem PPW bzw. NPW von 83,4% bzw. 72,3% einher. Wenn andere Sensitivitäts- oder Spezifitätswerte gewünscht sind, kann der SFSS-Cut-off entsprechend Tabelle 10 variiert werden. Eine Spezifität von >90% wurde laut Tabelle 10 erst bei SFSS-Cut-offs von ≥27,5 (effektiv >27) und höher erreicht, bei gleichzeitiger Sensitivität von ≤43,3%. Falsch-positive SFSS-Klassifikationen kamen in der vorliegenden Gesamtstichprobe bei einem Cut-off von ≥31,5 (effektiv >31) und höher nicht mehr vor (Spezifität = 100%, Sensitivität ≤36,7%). Der übliche SFSS-Cut-off von >16 erzielte eine Sensitivität und Spezifität von 73,3% bzw. 70% (vgl. Vierfeldertafel/Tabelle 6).

3.1.3.2 ROC-Kurven-Analyse WMT

In Abbildung 2 ist der Verlauf der ROC-Kurve des WMT für n = 60 (Auswertung der Gesamtstichprobe) dargestellt. Dabei wird der WMT-ROC-Kurve (blau) eine Diagonale (grün) gegenübergestellt, die den Verlauf der ROC-Kurve eines reinen Zufallsverfahrens wiedergibt.

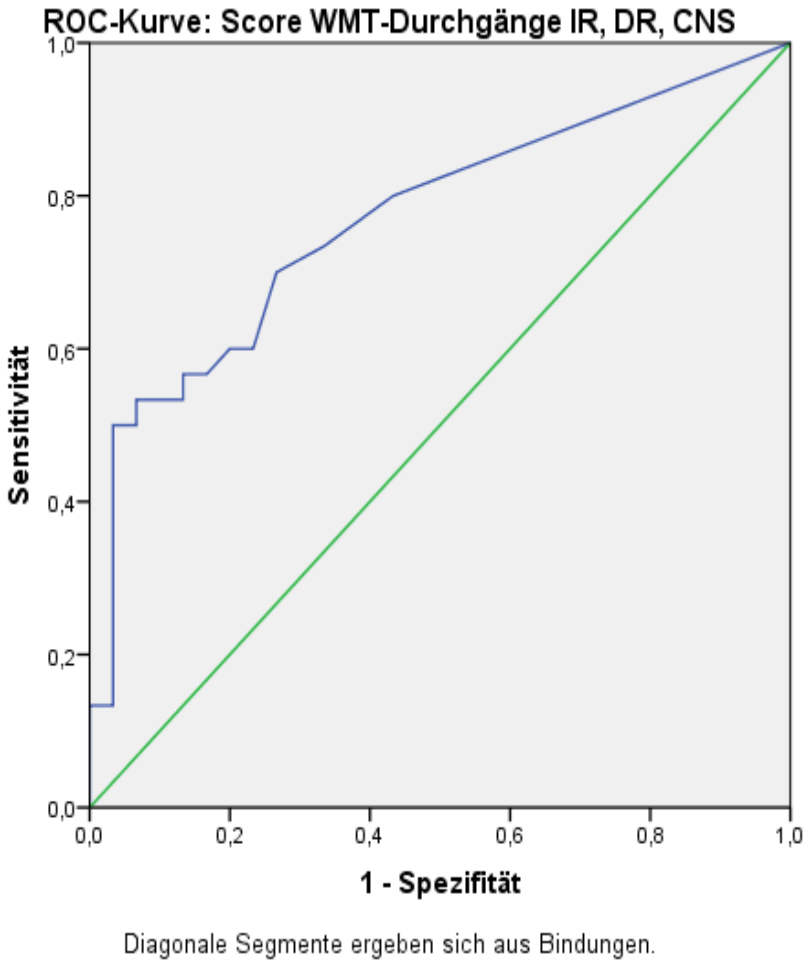

Abbildung 2: ROC-Kurven-Analyse für den Score WMT-Durchgänge IR, DR, CNS aller 60 Studienteilnehmer. Die ROC-Kurve eines diagnostischen Zufallsverfahrens (grün, diagonal) ist der ROC-Kurve des WMT (blau) zu Vergleichszwecken gegenübergestellt. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde sich als Punkt in der oberen linken Ecke darstellen. Der Score WMT-Durchgänge IR, DR, CNS entstand durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte
WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, ROC: receiver operating characteristic
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Die in Abbildung 2 blau eingezeichnete ROC-Kurve des WMT setzt sich größtenteils deutlich von der in grün eingezeichneten ROC-Kurve des Zufallsverfahrens ab. Nur bei sehr hohen und sehr niedrigen Sensitivitäten bzw.

Spezifitäten nähert sich die WMT-ROC-Kurve der ROC-Kurve des Zufallsverfahrens an. Bei einem WMT-Cut-off mit einer Sensitivität von 100% und einer Spezifität von 0% schneidet die WMT-ROC-Kurve die ROC-Kurve des Zufallsverfahrens. Das Schaubild in Abbildung 2 spricht dennoch insgesamt dafür, dass der WMT als Unterscheidungsverfahren zwischen den in der vorliegenden Studie gebildeten Zuständen „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“ und „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ – rein graphisch beurteilt – bei fast allen Cut-offs aussagekräftiger ist als ein Zufallsverfahren.

3.1.3.2.1 Fläche unter der ROC-Kurve (WMT)

Quantifizieren lässt sich die graphisch basierte Beurteilung der in Abbildung 2 dargestellten WMT-ROC-Kurve über die in Tabelle 11 angegebene Fläche unter der ROC-Kurve (AUC, maximaler Wert = 1). Sie ist ein Maß für die Güte des untersuchten diagnostischen Verfahrens.

Tabelle 11: Darstellung des für die Fläche unter der WMT-ROC-Kurve aller 60 Studienteilnehmer berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde einen AUC-Wert von 1 erzielen

[a] Unter der nicht-parametrischen Annahme, [b] Nullhypothese: Wahrheitsfläche = .5, WMT: Word Memory Test, ROC: receiver operating characteristic, AUC: area under the curve
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Fläche unter ROC-Kurve des WMT (AUC)	**Standardfehler[a]**	**Asymptotische Signifikanz[b]**	**Asymptotisches 95% Konfidenzintervall**	
			Untergrenze	Obergrenze
.77	.06	< .001***	.65	.89

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Der AUC-Wert (AUC WMT = .77) der WMT-ROC-Kurve unterschied sich statistisch signifikant (p < .001) vom AUC-Wert des Zufallsverfahrens (AUC = .5) und spricht daher für den WMT als brauchbares Unterscheidungsverfahren zwischen den Zuständen „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ und „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“.

3.1.3.2.2 Koordinaten der ROC-Kurve (WMT)

Tabelle 12 enthält die Koordinaten der in Abbildung 2 dargestellten WMT-ROC-Kurve und listet dazu in der ersten Spalte die verschiedenen variierten WMT-Cut-offs auf, für die jeweils Sensitivität, Spezifität, PPW und NPW des WMT in der vorliegenden Gesamtstichprobe berechnet wurden. Ein in der ersten Spalte angegebener Cut-off von 133,5 z.B. bedeutet, dass der zugehörige WMT-Cut-off für eine Klassifikation in den Zustand „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ effektiv bei <134 liegt, da es nur ganzzahlige Ergebnisse für den Score WMT-Durchgänge IR, DR, CNS gibt. Die letzte Spalte von Tabelle 12 enthält wiederum das für jeden Cut-off berechnete Maß für die Güte des jeweiligen Cut-offs (Youden-Index, maximaler Wert = 1). Die zur Berechnung von PPW und NPW angenommene Basalrate negativer Antwortverzerrung in der Population lag bei 50% (30 von 60 Studienteilnehmern).

Tabelle 12: Auflistung der Koordinaten der WMT-ROC-Kurve aller 60 Studienteilnehmer. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten WMT-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind. Die variierten Cut-offs beziehen sich auf den Score WMT-Durchgänge IR, DR, CNS, welcher durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte entstand. Der höchste Youden-Index-Wert ist rot hervorgehoben
Basalrate: Basalrate negativer Antwortverzerrung in der untersuchten Population, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, PPW: positiver prädiktiver Wert, NPW: negativer prädiktiver Wert, ROC: receiver operating characteristic, Youden-Index: Sensitivität + Spezifität – 1
[a] Der kleinste Cut-off entspricht dem kleinsten beobachteten Testwert minus 1, der größte Cut-off dem größten beobachteten Testwert plus 1. Alle anderen Cut-offs sind Mittelwerte zweier aufeinanderfolgender, geordneter beobachteter Testwerte
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Positives Testergebnis, wenn kleiner oder gleich[a] (Variable: Score WMT-Durchgänge IR, DR, CNS)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 50%)*	**NPW (%)** *(Basalrate: 50%)*	**Youden-Index**
125	0	100	-	-	0
133,5	6,7	100	~100	51,7	0,067

Fortsetzung von Tabelle 12 auf Seite 67

Fortsetzung Tabelle 12 von Seite 66

Positives Testergebnis, wenn kleiner oder gleich[a] (Variable: Score WMT-Durchgänge IR, DR, CNS)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 50%)*	**NPW (%)** *(Basalrate: 50%)*	**Youden-Index**
151	10	100	~100	52,6	0,1
163,5	13,3	100	~100	53,6	0,133
168,5	13,3	96,7	80,1	52,7	0,1
178,5	16,7	96,7	83,5	53,7	0,134
188,5	20	96,7	85,8	54,7	0,167
193,5	23,3	96,7	87,6	55,8	0,2
198,5	26,7	96,7	89	56,9	0,234
203,5	30	96,7	90,1	58	0,267
208	33,3	96,7	91	59,2	0,3
218	36,7	96,7	91,8	60,4	0,334
236	43,3	96,7	92,9	63	0,4
248,5	46,7	96,7	93,4	64,5	0,434
253	50	96,7	93,8	65,9	**0,467**
258	50	93,3	88,2	65,1	0,433
263	53,3	93,3	88,8	66,6	0,466
265,5	53,3	90	84,2	65,8	0,433
268	53,3	86,7	80	65	0,4
270,5	56,7	86,7	81	66,7	0,434
275,5	56,7	83,3	77,3	65,8	0,4

Fortsetzung von Tabelle 12 auf Seite 68

Fortsetzung Tabelle 12 von Seite 67

Positives Testergebnis, wenn kleiner oder gleich[a] (Variable: Score WMT-Durchgänge IR, DR, CNS)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 50%)*	**NPW (%)** *(Basalrate: 50%)*	**Youden-Index**
280,5	60	80	75	66,7	0,4
283,5	60	76,7	72	65,7	0,367
288	70	73,3	72,4	71	0,433
293	73,3	66,7	68,8	71,4	0,4
298	80	56,7	64,9	73,9	0,367
301	100	0	-	-	0

Der höchste Youden-Index-Wert (= 0,467) in Tabelle 12 ergab sich für einen WMT-Cut-off von ≤253, oder einfacher ausgedrückt <254. Der Cut-off von ≤253 (effektiv <254) ging mit einer Sensitivität von 50%, einer Spezifität von 96,7% und einem PPW bzw. NPW von 93,8% bzw. 65,9% einher. Sind andere Sensitivitäts- oder Spezifitätswerte gewünscht, kann der WMT-Cut-off entsprechend Tabelle 12 variiert werden. Eine Sensitivität von >90% wurde laut Tabelle 12 für den WMT nur erreicht, wenn man den WMT-Cut-off auf den maximal erreichbaren Wert von ≤300 festlegte (effektiv <301, aber dieses Ergebnis ist nicht möglich), was dann mit einer Sensitivität von 100% einherging – bei gleichzeitiger Spezifität von 0%. Falsch-positive WMT-Klassifikationen kamen in der vorliegenden Gesamtstichprobe bei einem Cut-off von ≤163,5 (effektiv <164) und niedriger nicht mehr vor (Spezifität = 100%, Sensitivität ≤13,3%).

3.1.4 Übereinstimmung zwischen den BVT-Klassifikationen

In Tabelle 13 ist die Übereinstimmung der Klassifikationsergebnisse von WMT und SFSS bei Verwendung der gebräuchlichen Cut-offs von >16 (SFSS) bzw. ≤82,5% (WMT) dargestellt. Als Maßzahlen für die Übereinstimmung beider Testverfahren finden sich unter Tabelle 13 der Vierfelderkorrelationskoeffizient (Φ, phi) sowie der Kappa-Koeffizient κ.

Tabelle 13: Darstellung der Übereinstimmung der Klassifikationsergebnisse von SFSS und WMT (bestanden/nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für die Gesamtstichprobe aller 60 Studienteilnehmer. Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang bzw. ein SFSS-Gesamtscore >16 Punkten wird jeweils als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des jeweiligen Tests
WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency SFSS: Strukturierter Fragebogen Simulierter Symptome
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

		WMT		
		bestanden	**nicht bestanden**	***Summe***
SFSS	**bestanden**	28	1	29
	nicht bestanden	15	16	31
	Summe	43	17	60

Vierfelderkorrelationskoeffizient (Φ, phi) = .53 (p < .001, 95% KI = .34 - .7)

Kappa-Koeffizient κ: .47 (p < .001, 95% KI = .28 - .66).

Tabelle 13 zeigt, dass 44 der 60 Studienteilnehmer (73,3%) der Gesamtstichprobe von beiden Beschwerdevalidierungstests gleich klassifiziert wurden (jeweils bestanden bzw. nicht bestanden). 15 der 60 Studienteilnehmer (25%) bestanden zwar den WMT, nicht aber den SFSS. Bei einem der 60 Studienteilnehmer (1,7%) verhielt es sich genau anders herum (WMT nicht bestanden, SFSS bestanden). Sowohl der für den Vierfelderkorrelationskoeffizienten als auch der für den Kappa-Koeffizienten berechnete Wert war statistisch signifikant (p < .001). In der

vorliegenden Studie ist also von einer statistisch signifikanten Übereinstimmung zwischen den Klassifikationen der beiden Beschwerdevalidierungstests auszugehen.

3.1.5 Klassifikationen bei Verwendung von SFSS *und* WMT

Tabelle 14 (Seite 71) zeigt welche Klassifikationen und Testkennwerte sich in der vorliegenden Gesamtstichprobe für die Cut-offs von >16 (SFSS) bzw. ≤82,5% (WMT) ergeben, wenn folgendes Vorgehen bei der Verwendung der BVTs festgelegt wird:

Regeln: 1. Beide BVTs müssen bestanden oder nicht-bestanden sein für eine Klassifikation. 2. Studienteilnehmer mit nicht übereinstimmenden Klassifikationen werden in beiden Studiengruppen wegen Unsicherheit ausgeschlossen (Gruppe A: 9 Studienteilnehmer ausgeschlossen, Gruppe B: 7 Studienteilnehmer ausgeschlossen).

Tabelle 14: Darstellung der Klassifikationsergebnisse von SFSS und WMT bei kombinierter Anwendung der beiden BVTs (Tests beide bestanden/beide nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für die Gesamtstichprobe aller 60 Studienteilnehmer. Bei der kombinierten Anwendung der beiden BVTs SFSS und WMT galten folgende Regeln: 1. Beide Tests müssen bestanden oder nicht-bestanden sein für eine Klassifikation. 2. Studienteilnehmer mit nicht übereinstimmenden Klassifikationen werden wegen Unsicherheit ausgeschlossen. Letztlich mussten für diese Art der Auswertung 16 der insgesamt 60 Studienteilnehmer wegen differierender BVT-Klassifikationen (z.B. WMT bestanden/SFSS nicht bestanden) ausgeschlossen werden. Es ergab sich eine angepasste Gesamtstichprobe von 44 Studienteilnehmern. Unterteilung der Klassifikationen nach den beiden Studiengruppen A und B (ursprünglich je 30 Studienteilnehmer). Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang bzw. ein SFSS-Gesamtscore >16 Punkten wird jeweils als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des jeweiligen Tests
BVT: Beschwerdevalidierungstest, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, SFSS: Strukturierter Fragebogen Simulierter Symptome
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	BVT-Ergebnisse (WMT *und* SFSS) auffällig	**BVT-Ergebnisse (WMT *und* SFSS) unauffällig**	***Summe***
experimentelle negative Antwortverzerrer (Gruppe B)	15	8	23
subjektiv bestmögliche Leistungsanstrengung/ unverzerrtes Antwort-verhalten (Gruppe A)	1	20	21
Summe	16	28	44

Bei kombinierter Verwendung beider BVTs bzw. für obige Vierfeldertafel ergaben sich folgende Testkennwerte:

Sensitivität: 65,2% (95% KI = 44,9% - 82,3%)

Spezifität: 95,2% (95% KI = 80,7% - 99,7%)

Positiver prädiktiver Wert (PPW): 93,8% (95% KI = 75,3% - 99,6%)

Negativer prädiktiver Wert (NPW): 71,4% (95% KI = 53,3% - 85,8%).

Bei oben beschriebenem Vorgehen mussten zwar 16 von 60 Studienteilnehmern (26,7%) der Gesamtstichprobe als „nicht klassifizierbar“ akzeptiert und für diesen Teil der Auswertung ausgeschlossen werden, für die verbleibenden 44 Studienteilnehmer ergab sich jedoch eine Spezifität von 95,2% - bei gleichzeitiger Sensitivität von 65,2%. Kontrollgruppe A wies bei diesem Vorgehen nur noch eine falsch-positive Klassifikation durch die BVTs auf. In der Interventionsgruppe B kam es zu 8 falsch-negativen Klassifikationen (vgl. Tabelle 14).

3.1.6 Korrelationen zwischen WMT/SFSS und BSI/SCIP

3.1.6.1 Punktwolke und Korrelationen BSI/WMT

Abbildung 3 zeigt den Zusammenhang zwischen BSI-Ergebnis (GSI) und WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS) in Form einer Punktwolke, aufgeschlüsselt nach den Gruppen A und B.

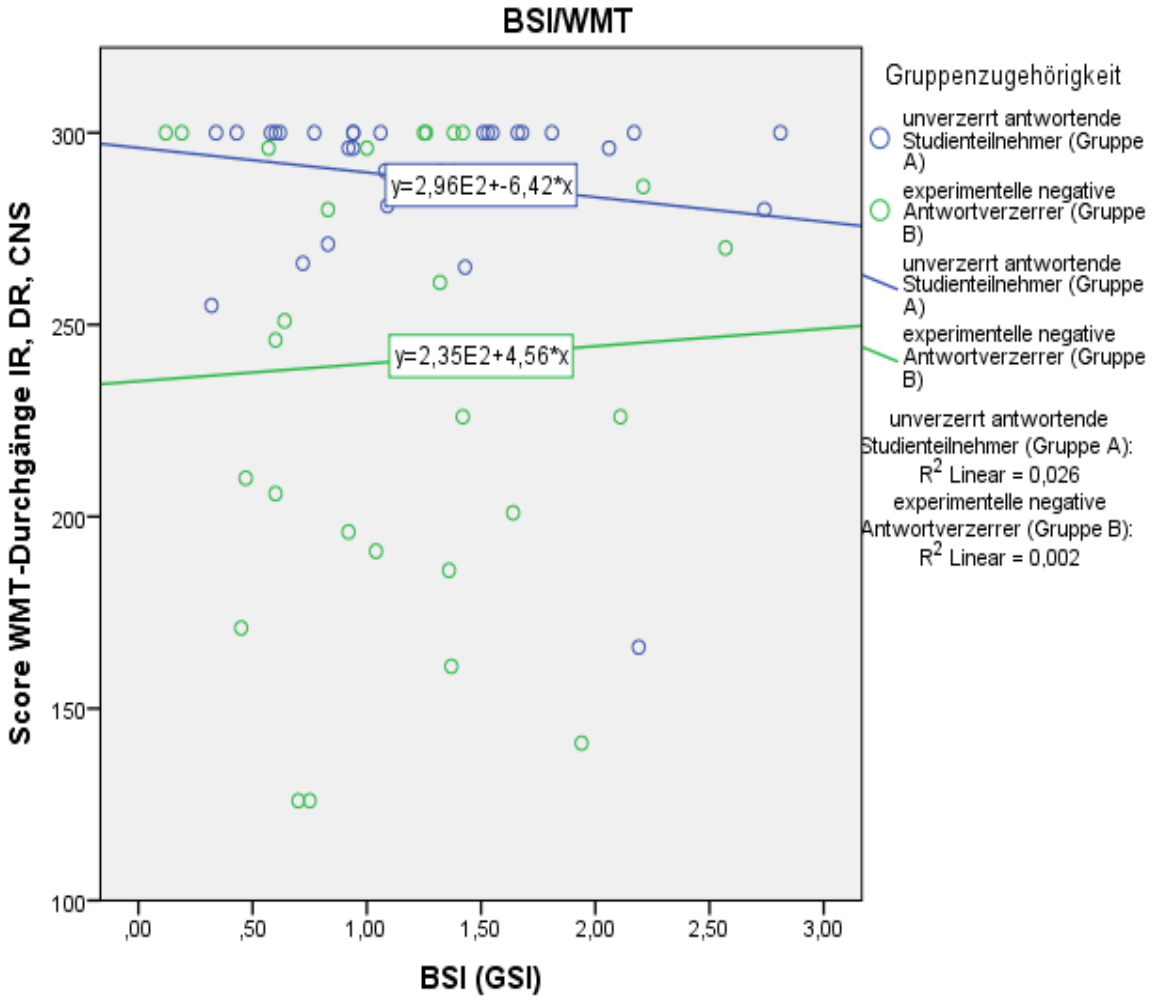

Abbildung 3: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 60 Studienteilnehmer gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B (grün, 30 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2(Maximalwert = 1). Der Score WMT-Durchgänge IR, DR, CNS entstand durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte
BSI: Brief Symptom Inventory, GSI: Global Severity Index, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Die Regressionsgerade von Gruppe B (grün) lässt in Abbildung 3 – rein graphisch betrachtet – einen geringen gleichsinnigen Zusammenhang zwischen BSI-Ergebnis (aktuelle Symptomschwere) und WMT-Ergebnis in Gruppe B vermuten. Die Punktwolke von Gruppe B weist allerdings keine annähernd lineare Anordnung auf und auch das für Gruppe B berechnete R^2 nimmt einen Wert nahe 0 an (R^2 = 0,002). Der Einfluss des BSI-Ergebnisses auf das WMT-Ergebnis in Gruppe B ist hier folglich als gering einzuschätzen. In Gruppe A (blau) scheint das WMT-Ergebnis graphisch betrachtet ebenfalls kaum vom BSI-Ergebnis beeinflusst zu werden. Die in blau eingezeichnete Regressionsgerade von Gruppe A erweckt zwar den Anschein eines leicht gegenläufigen Zusammenhangs zwischen den Merkmalen BSI-Ergebnis und WMT-Ergebnis, das R^2 in Gruppe A nimmt allerdings einen Wert nahe 0 an (R^2 = 0,026). Der Einfluss des BSI-Ergebnisses auf das WMT-Ergebnis ist also in Gruppe A ebenfalls als gering einzuschätzen.

In Tabelle 15 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 3 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe als auch für die Gruppen A und B getrennt berechnet. Auch die Korrelationskoeffizienten zwischen BSI-Ergebnis (GSI) und „gesamtem“ WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) werden in Tabelle 15 zusätzlich zu den Korrelationskoeffizienten zwischen BSI und „einfachen“ WMT-Durchgängen (Score WMT-Durchgänge IR, DR, CNS) angegeben.

Tabelle 15: Auflistung der für den Zusammenhang zwischen BSI- und WMT-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 60 Studienteilnehmer als auch für die beiden Studiengruppen A und B (je 30 Studienteilnehmer) getrennt berechnet. Der Zusammenhang zwischen BSI und WMT wurde für den Score WMT-Durchgänge IR, DR, CNS sowie zusätzlich auch für den Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR („gesamter WMT“) berechnet. Beide WMT-Scores entstanden durch Addition der in den jeweiligen WMT-Durchgängen erzielten Prozentwerte
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MC: Multiple Choice, PA: Paired Associate Recall, FR: Free Recall, BSI: Brief Symptom Inventory, Spearman: Rangkorrelationskoeffizient nach Spearman
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation BSI/WMT	r	n	p	95% KI
Spearman (Gesamtstichprobe)	.01	60	.969	-.27 - .28

Fortsetzung von Tabelle 15 auf Seite 75

Fortsetzung Tabelle 15 von Seite 74

Korrelation BSI/WMT	r	n	p	95% KI
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gesamtstichprobe)	.07	60	.573	-.21 - .37
Spearman (Gruppe A)	-.03	30	.874	-.42 - .39
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe A)	.06	30	.752	-.33 - .47
Spearman (Gruppe B)	-.001	30	.997	-.36 - .38
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe B)	.08	30	.691	-.31 - .48

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

In Gruppe A ergab sich zwischen dem BSI-Ergebnis und dem Score der „einfachen“ WMT-Durchgänge sowie dem Score des „gesamten“ WMT ein sehr geringer gegenläufiger bzw. gleichsinniger linearer Zusammenhang nahe 0 (IR, DR, CNS: r = -.03. p = .874; IR, DR, CNS, MC, PA, FR: r = .06, p = .752). Ein hohes oder niedriges BSI-Ergebnis ging also in Gruppe A weder mit einem stark erhöhten noch mit einem stark erniedrigten WMT-Ergebnis einher. In Gruppe B ergab sich ein ebenfalls sehr geringer gegenläufiger bzw. gleichsinniger linearer Zusammenhang zwischen BSI- und WMT-Ergebnis (IR, DR, CNS: r = -.001, p = .997; IR, DR, CNS, MC, PA, FR: r = .08, p = .691). Keine der in Tabelle 15 aufgeführten Korrelationen war statistisch signifikant.

3.1.6.2 Punktwolke und Korrelationen SCIP/WMT

Abbildung 4 stellt den Zusammenhang zwischen SCIP-Ergebnis (Summenscore) und WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS) – unterteilt nach den Gruppen A und B – graphisch dar.

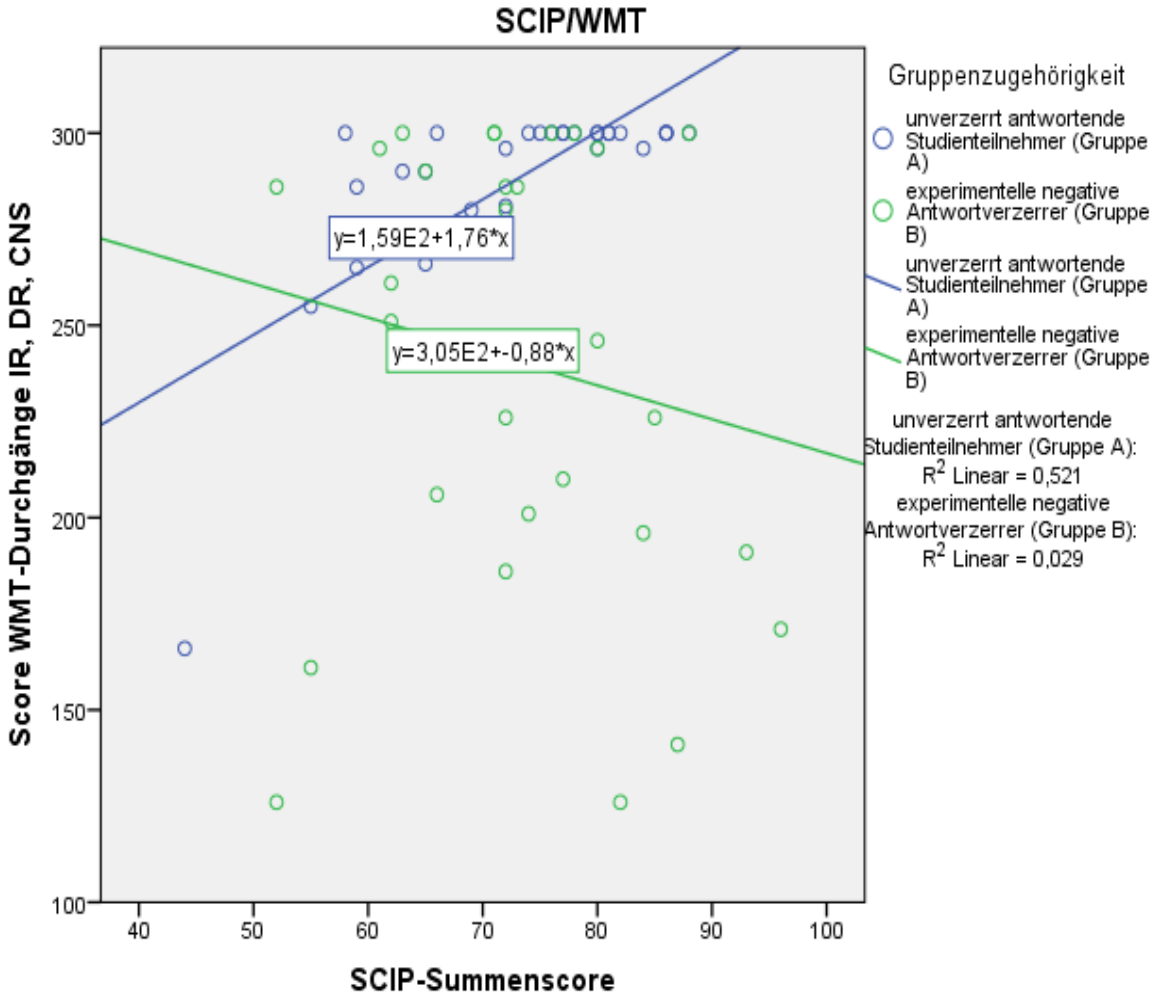

Abbildung 4: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 60 Studienteilnehmer gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B (grün, 30 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1). Der Score WMT-Durchgänge IR, DR, CNS entstand durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte
SCIP: Screen for Cognitive Impairment in Psychiatry, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Für Gruppe A (blau) zeigt sich in Abbildung 4 graphisch ein deutlicher gleichsinniger Zusammenhang zwischen SCIP-Summenscore (kognitives

Leistungsniveau) und WMT-Ergebnis (R^2 = 0,521). Die in grün eingezeichnete Regressionsgerade von Gruppe B erweckt den Anschein eines gegenläufigen Zusammenhangs zwischen den Merkmalen SCIP-Ergebnis und WMT-Ergebnis in Gruppe B. Das mitangegebene R^2 nimmt in Gruppe B jedoch einen Wert nahe 0 an (R^2 = 0,029) – was bedeutet, dass der Einfluss des SCIP-Ergebnisses auf das WMT-Ergebnis in Gruppe B als eher gering einzuschätzen ist.

In Tabelle 16 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 4 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe als auch für die Gruppen A und B getrennt berechnet. Auch die Korrelationskoeffizienten zwischen SCIP-Ergebnis (Summenscore) und „gesamtem“ WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) wurden zusätzlich zu den Koeffizienten zwischen SCIP und „einfachen“ WMT-Durchgängen (Score WMT-Durchgänge IR, DR, CNS) berechnet.

Tabelle 16: Auflistung der für den Zusammenhang zwischen SCIP- und WMT-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 60 Studienteilnehmer als auch für die beiden Studiengruppen A und B (je 30 Studienteilnehmer) getrennt berechnet. Der Zusammenhang zwischen SCIP und WMT wurde für den Score WMT-Durchgänge IR, DR, CNS sowie zusätzlich auch für den Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR („gesamter WMT“) berechnet. Beide WMT-Scores entstanden durch Addition der in den jeweiligen WMT-Durchgängen erzielten Prozentwerte
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MC: Multiple Choice, PA: Paired Associate Recall, FR: Free Recall, SCIP: Screen for Cognitive Impairment in Psychiatry, Spearman: Rangkorrelationskoeffizient nach Spearman
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation SCIP/WMT	r	n	p	95% KI
Spearman (Gesamtstichprobe)	.23	60	.083	-.06 - .48
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gesamtstichprobe)	.29	60	.026*	.01 - .53
Spearman (Gruppe A)	.68	30	< .001***	41 - .87

Fortsetzung von Tabelle 16 auf Seite 78

Fortsetzung Tabelle 16 von Seite 77

Korrelation SCIP/WMT	r	n	p	95% KI
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe A)	.66	30	< .001***	.4 - .83
Spearman (Gruppe B)	-.18	30	.346	-.6 - .29
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe B)	-.13	30	.479	-.57 - .33

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Zwischen dem SCIP-Ergebnis und dem Score der „einfachen“ WMT-Durchgänge sowie dem Score des „gesamten“ WMT ergab sich in Gruppe A ein mittelstarker gleichsinniger linearer Zusammenhang (IR, DR, CNS: r = .68, p < .001; IR, DR, CNS, MC, PA, FR: r = .66, p < .001). Ein hohes SCIP-Ergebnis ging also in Gruppe A tendenziell auch mit einem hohen WMT-Ergebnis einher. In Gruppe B hingegen ergab sich ein sehr geringer gegenläufiger linearer Zusammenhang zwischen SCIP- und WMT-Ergebnis (IR, DR, CNS: r = -.18, p = .346; IR, DR, CNS, MC, PA, FR: r = -.13, p = .479). Mit steigendem SCIP-Ergebnis sank das WMT-Ergebnis in Gruppe B also tendenziell. Ein statistisch signifikanter, geringer gleichsinniger linearer Zusammenhang ergab sich zudem zwischen SCIP-Ergebnis und „gesamtem“ WMT-Ergebnis für die Gesamtstichprobe (r = .29, p = .026). Mit steigendem SCIP-Summenscore erhöhte sich also tendenziell in der gesamten Stichprobe der Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR.

3.1.6.3 Punktwolke und Korrelationen BSI/SFSS

Abbildung 5 zeigt den Zusammenhang zwischen BSI-Ergebnis (GSI) und SFSS-Ergebnis (Gesamtscore) in Form einer Punktwolke, aufgeschlüsselt nach den Gruppen A und B.

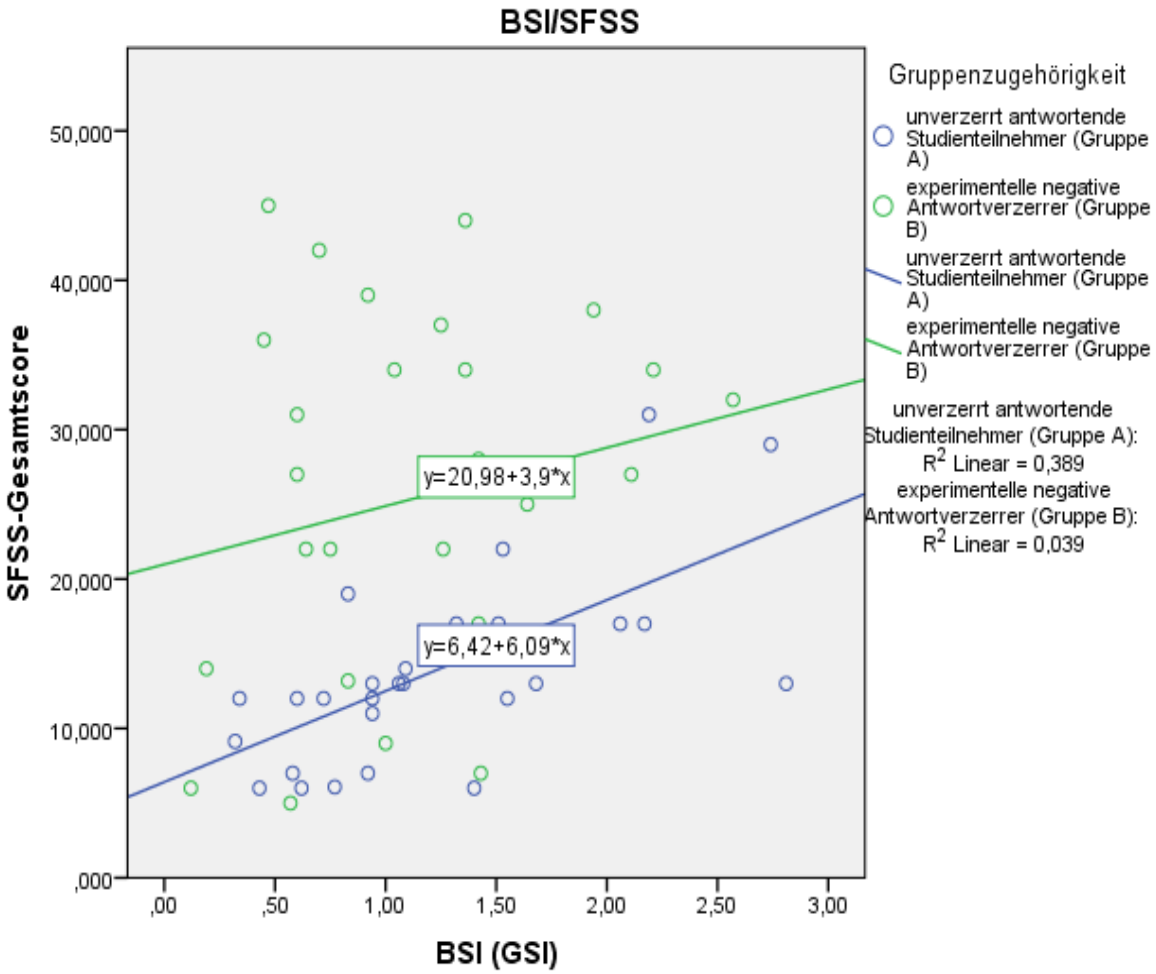

Abbildung 5: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 60 Studienteilnehmer gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B (grün, 30 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1)
BSI: Brief Symptom Inventory, GSI: Global Severity Index, SFSS: Strukturierter Fragebogen Simulierter Symptome, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

In Abbildung 5 zeigt sich für Gruppe A (blau) graphisch ein deutlicher gleichsinniger Zusammenhang zwischen BSI-Ergebnis (aktuelle Symptomschwere) und SFSS-Ergebnis (R^2 = 0,389). Die in grün eingezeichnete

Regressionsgerade von Gruppe B deutet auf einen ebenfalls gleichsinnigen Zusammenhang zwischen den Merkmalen BSI-Ergebnis und SFSS-Ergebnis in Gruppe B hin, auch wenn das mitangegebene R^2 in Gruppe B einen Wert nahe 0 annimmt (R^2 = 0,039) – womit der Einfluss des BSI-Ergebnisses auf das SFSS-Ergebnis in Gruppe B zumindest als gering einzuschätzen ist.

In Tabelle 17 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 5 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe als auch für die Gruppen A und B getrennt berechnet.

Tabelle 17: Auflistung der für den Zusammenhang zwischen BSI- und SFSS-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 60 Studienteilnehmer als auch für die beiden Studiengruppen A und B (je 30 Studienteilnehmer) getrennt berechnet
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, SFSS: Strukturierter Fragebogen Simulierter Symptome, BSI: Brief Symptom Inventory, Spearman: Rangkorrelationskoeffizient nach Spearman, Pearson: Produkt-Moment-Korrelationskoeffizient nach Pearson
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation BSI/SFSS	r	n	p	95% KI
Spearman (Gesamtstichprobe)	.3	60	.018*	.03 - .55
Spearman (Gruppe A)	.68	30	< .001***	.45 - .84
Spearman (Gruppe B)	.1	30	.596	-.33 - .48
Pearson (Gruppe B)	.2	30	.296	-.17 - .53

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

In Gruppe A ergab sich zwischen dem BSI-Ergebnis und dem SFSS-Ergebnis bzw. –Gesamtscore ein mittelstarker gleichsinniger linearer Zusammenhang (r = .68, p < .001). Ein hohes BSI-Ergebnis ging also in Gruppe A tendenziell auch mit einem hohen SFSS-Ergebnis einher. In Gruppe B ergab sich ein sehr geringer gleichsinniger linearer Zusammenhang zwischen BSI-Ergebnis und SFSS-Ergebnis (r = .1, p = .596/Pearson: r = .2, p = .296). Mit steigendem BSI-Ergebnis erhöhte sich das SFSS-Ergebnis bzw. der SFSS-Gesamtscore in Gruppe B also ebenfalls tendenziell. Ein statistisch signifikanter, geringer gleichsinniger linearer Zusammenhang ergab sich zwischen BSI- und SFSS-Ergebnis zudem für die Gesamtstichprobe (r = .3, p = .018).

3.1.6.4 Punktwolke und Korrelationen SCIP/SFSS

Abbildung 6 stellt den Zusammenhang zwischen SCIP-Ergebnis (Summenscore) und SFSS-Ergebnis (Gesamtscore) – unterteilt nach den Gruppen A und B – graphisch dar.

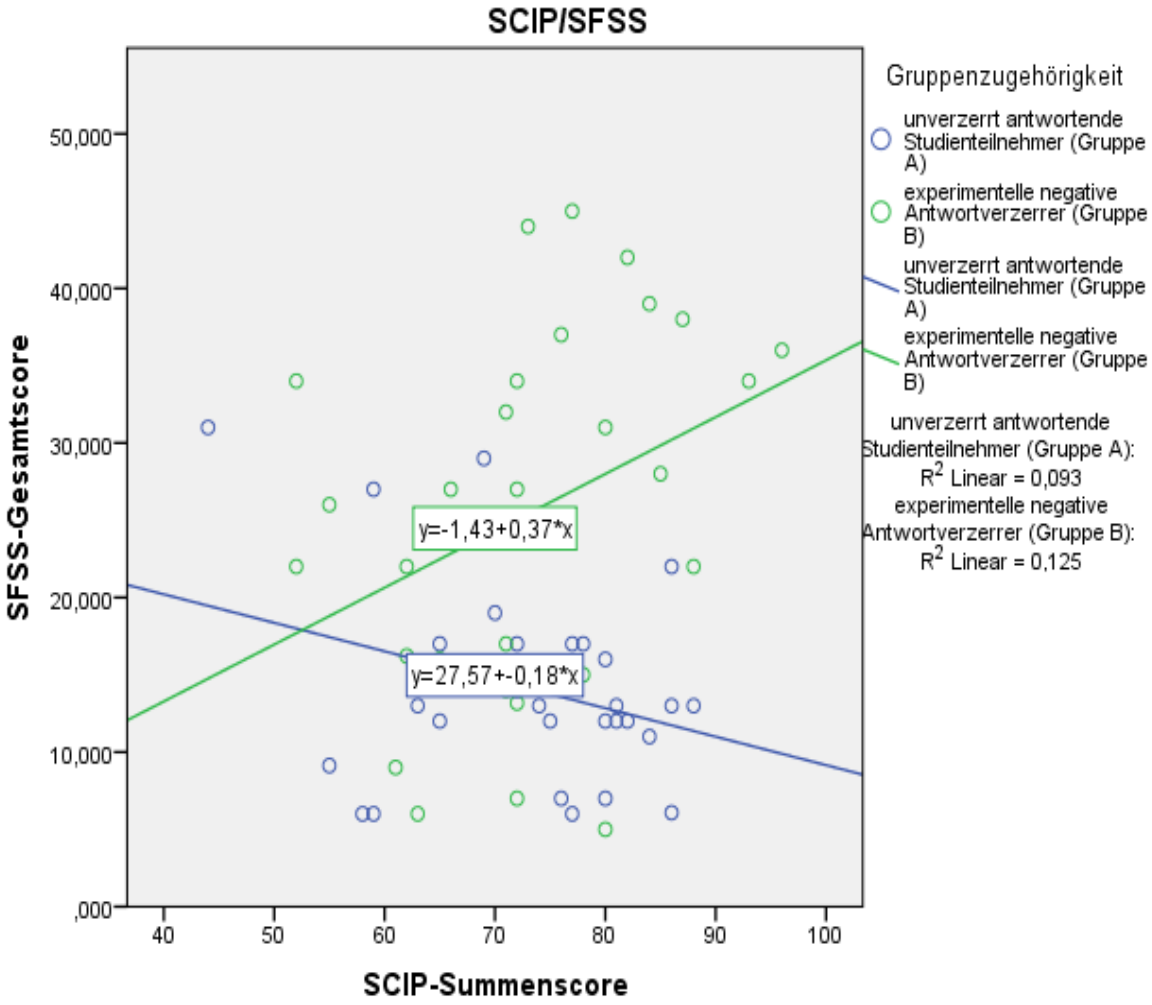

Abbildung 6: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 60 Studienteilnehmer gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B (grün, 30 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1)
SCIP: Screen for Cognitive Impairment in Psychiatry, SFSS: Strukturierter Fragebogen Simulierter Symptome, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Graphisch zeigt sich in Abbildung 6 ein leicht gegenläufiger Zusammenhang zwischen SCIP-Ergebnis (kognitives Leistungsniveau) und SFSS-Ergebnis in Gruppe A (blau) – allerdings bei recht geringem R^2 (R^2 = 0,093). In Gruppe B

(grün) ist ein tendenziell gleichsinniger Zusammenhang zwischen SCIP-Ergebnis und SFSS-Ergebnis zu erkennen (R^2 = 0,125).

In Tabelle 18 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 6 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe als auch für die Gruppen A und B getrennt berechnet.

Tabelle 18: Auflistung der für den Zusammenhang zwischen SCIP- und SFSS-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 60 Studienteilnehmer als auch für die beiden Studiengruppen A und B (je 30 Studienteilnehmer) getrennt berechnet
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, SFSS: Strukturierter Fragebogen Simulierter Symptome, SCIP: Screen for Cognitive Impairment in Psychiatry, Spearman: Rangkorrelationskoeffizient nach Spearman, Pearson: Produkt-Moment-Korrelationskoeffizient nach Pearson
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation SCIP/SFSS	r	n	p	95% KI
Spearman (Gesamtstichprobe)	.08	60	.558	-.18 - .33
Spearman (Gruppe A)	-.14	30	.464	-.56 - .31
Spearman (Gruppe B)	.42	30	.020*	.14 - .63
Pearson (Gruppe B)	.35	30	.055	.07 - .62

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Ein geringer gleichsinniger linearer Zusammenhang zwischen SCIP-Ergebnis und SFSS-Ergebnis ergab sich in Gruppe B (r = .42, p = .020/Pearson: r = .35, p = .055). Ein hohes SCIP-Ergebnis ging in Gruppe B also auch tendenziell mit einem hohen SFSS-Ergebnis bzw. SFSS-Gesamtscore einher. In Gruppe A ergab sich hingegen ein sehr geringer gegenläufiger linearer Zusammenhang (r = -.14, p = .464) zwischen SCIP-Ergebnis und SFSS-Ergebnis. Mit steigendem SCIP-Summenscore sank also der SFSS-Gesamtscore in Gruppe A tendenziell.

3.1.7 Bericht über die Datenqualität (Gesamtstichprobe)

Von allen 60 Studienteilnehmern der Gesamtstichprobe lagen vor Beginn der Auswertung alle Datenerhebungsbögen nahezu vollständig ausgefüllt vor. Die CRFs waren bei allen 60 Patienten lückenlos ausgefüllt. Für das BSI fehlten von 6 Patienten insgesamt 7 Werte, was aber nur geringe GSI-Veränderungen (Größenordnung 1/100) verursachte und daher vernachlässigt wurde. Das SCIP lag für alle 60 Patienten lückenlos ausgefüllt vor. Beim WMT gab es ebenfalls keine fehlenden Werte. Vom SFSS lagen von 56 Patienten vollständig ausgefüllte Bögen vor, bei 4 Patienten fehlten insgesamt 6 Angaben. Da es sich beim SFSS-Gesamtscore um eine der Zielgrößen der vorliegenden Studie handelte, wurde im Falle unausgefüllter SFSS-Items die jeweilige durchschnittliche „SFSS-Fehlerquote“ (SFSS-Gesamtscore des jeweiligen Studienteilnehmers geteilt durch die Anzahl ausgefüllter Items) zum SFSS-Gesamtscore addiert. In 56 Fällen wurde die Studie strikt entsprechend des Studienprotokolls an einem Termin durchgeführt. In 4 Fällen wurde die Datenerhebung zweizeitig, mit jeweils eintägiger Pause (nach Punkt 4. des Studienprotokolls), durchgeführt. In einem Fall konnte der zweite Teil des Studienprotokolls (ab Punkt 4. des Studienprotokolls) wegen bereits erfolgter Entlassung aus der ambulanten Behandlung nicht mehr durchgeführt werden. Eine Randomisierung hatte noch nicht stattgefunden und es lagen nur die Ergebnisse von BSI und SCIP vor, weshalb der Code neu vergeben wurde und der Fall auch nicht im Flussdiagramm eingeschlossener Patienten auftaucht (vgl. Abbildung 7). In die Auswertung der Gesamtstichprobe wurden alle 60 randomisierten Patienten eingeschlossen.

3.2 Per Protocol Analyse

3.2.1 Flussdiagramm ein- und ausgeschlossener Patienten

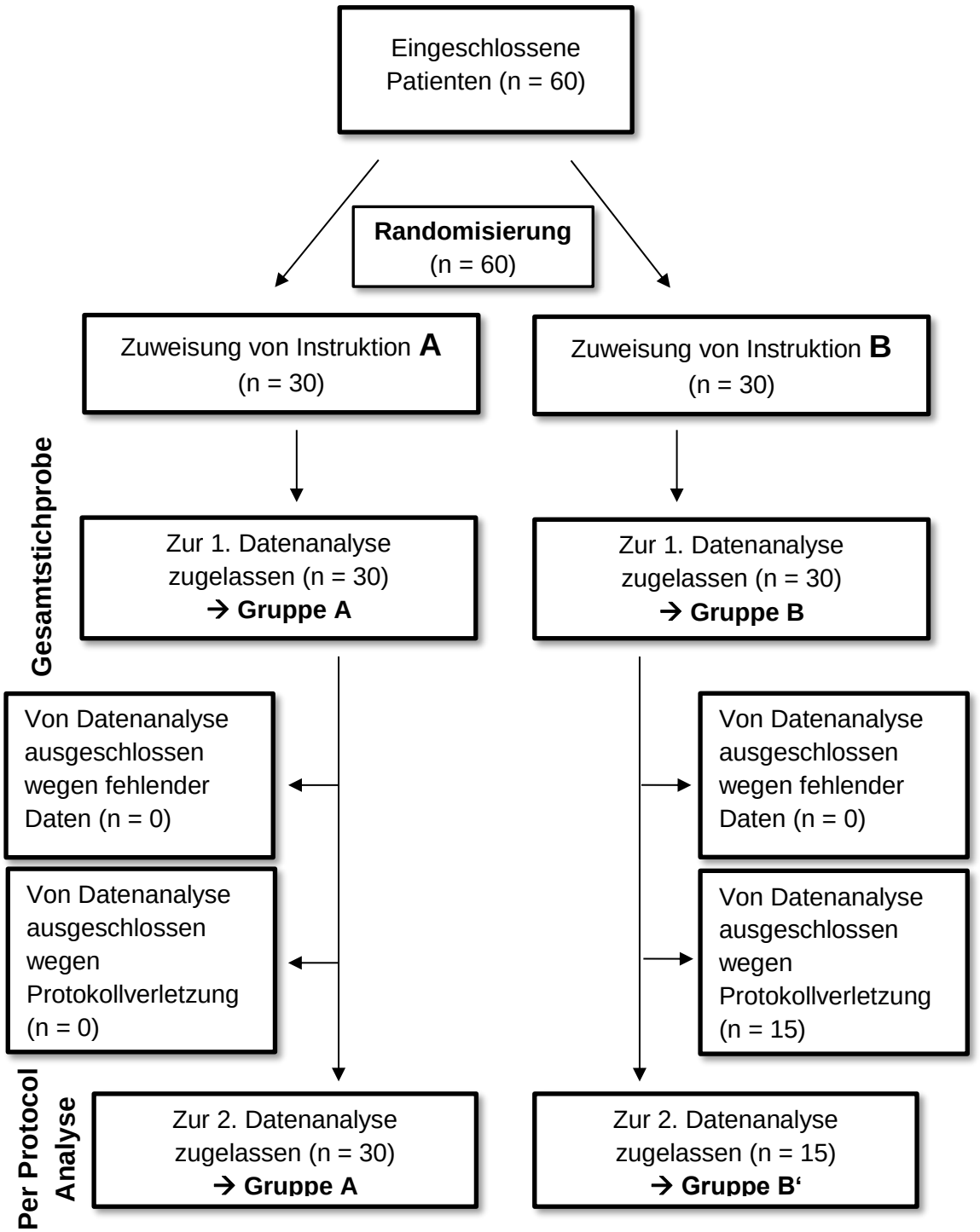

Abbildung 7: Darstellung des Ein- bzw. Ausschlusses der insgesamt 60 Studienteilnehmer in Form eines Flussdiagramms. Unterteilung in Analyse der Gesamtstichprobe (60 Teilnehmer) und Per Protocol Analyse (45 Teilnehmer). Notwendig wurde die Per Protocol Analyse aufgrund mehrfacher Protokollverletzungen in Studiengruppe B (30 Teilnehmer). n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Für die Per Protocol Analyse wurden aus Gruppe B 15 Studienteilnehmer ausgeschlossen (vgl. Abbildung 7). Ihre Protokollverletzung – im Sinne eines fehlenden Umsetzens von Patienteninstruktion B – war ersichtlich bei den VAS-Angaben und bei den nach Versuchsdurchführung protokollierten Aussagen in der Patientenbefragung zur Umsetzung der Instruktion (Anhang C). Aus Studiengruppe B entstand somit Studiengruppe B‘ mit 15 verbliebenen Studienteilnehmern.

3.2.2 Patientenstichprobe der Per Protocol Analyse

Tabelle 19 zeigt die soziodemographischen Variablen und die BSI- und SCIP-Ergebnisse der Gesamtstichprobe der PPA sowie die Verteilung dieser Merkmale in den Gruppen A und B‘.

Tabelle 19: Verteilung der soziodemographischen Variablen und der BSI- und SCIP-Ergebnisse in der Gesamtstichprobe der Per Protocol Analyse (45 Studienteilnehmer) sowie Aufteilung dieser Merkmale nach der Randomisierung in die beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer). Untersuchung der Gruppenunterschiede hinsichtlich statistischer Signifikanz

[a] Chi-Quadrat-Test, [b] t-test, [c] Exakter Fisher-Test, [d] U-Test, MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum, vs.: versus, GSI: Global Severity Index, n: Stichprobenumfang, BSI: Brief Symptom Inventory, SCIP: Screen for Cognitive Impairment in Psychiatry, absolute Häufigkeiten der Studienteilnehmer mit eingeklammerten relativen Häufigkeiten in Prozent

Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Gesamtstichprobe Per Protocol Analyse n = 45	Kontrollgruppe A n = 30	Interventionsgruppe B‘ n = 15	Signifikanz (2-seitig) der Unterschiede A vs. B‘
Geschlecht	Männlich: 15 (33,3%) Weiblich: 30 (66,7%)	Männlich: 11 (36,7%) Weiblich: 19 (63,3%)	Männlich: 4 (26,7%) Weiblich: 11 (73,3%)	.738[c]
Alter (in Jahren)	MW: 50,9 (SD = 5,6) Med: 51 (range = 23)	MW: 51,5 (SD = 5,2) Med: 51,5 (range = 22)	MW: 49,6 (SD = 6,4) Med: 51 (range = 18)	.280[b]

Fortsetzung von Tabelle 19 auf Seite 87

Fortsetzung Tabelle 19 von Seite 86

	Gesamtstichprobe Per Protocol Analyse n = 45	Kontrollgruppe A n = 30	Interventionsgruppe B‘ n = 15	Signifikanz (2-seitig) der Unterschiede A vs. B‘
Bildungs-abschluss	kein Abschluss: 1 (2,2%) Hauptschule: 21 (46,7%) Mittlere Reife: 13 (28,9%) Abitur: 4 (8,9%) Fachhochschule: 6 (13,3%) Universität: 0 (0%)	kein Abschluss: 1 (3,3%) Hauptschule: 16 (53,3%) Mittlere Reife: 8 (26,7%) Abitur: 1 (3,3%) Fachhochschule: 4 (13,3%) Universität: 0 (0%)	kein Abschluss: 0 (0%) Hauptschule: 5 (33,3%) Mittlere Reife: 5 (33,3%) Abitur: 3 (20%) Fachhochschule: 2 (13,3%) Universität: 0 (0%)	.138[d]
Aufnahme-diagnose(n)	Summe: 53 Diagnosen F32.x: 13 von 53 Diagnosen (24,5%) F33.x: 24 von 53 Diagnosen (45,3%) F40.x: 3 von 53 Diagnosen (5,7%) F41.x: 8 von 53 Diagnosen (15,1%) F43.2: 5 von 53 Diagnosen (9,4%)	Summe: 34 Diagnosen F32.x: 9 von 34 Diagnosen (26,5%) F33.x: 16 von 34 Diagnosen (47,1%) F40.x: 2 von 34 Diagnosen (5,9%) F41.x: 4 von 34 Diagnosen (11,8%) F43.2: 3 von 34 Diagnosen (8,8%)	Summe: 19 Diagnosen F32.x: 4 von 19 Diagnosen (21,1%) F33.x: 8 von 19 Diagnosen (42,1%) F40.x: 1 von 19 Diagnosen (5,3%) F41.x: 4 von 19 Diagnosen (21,1%) F43.2: 2 von 19 Diagnosen (10,5%)	 1.0[c] 1.0[c] 1.0[c] .410[c] 1.0[c]

Fortsetzung von Tabelle 19 auf Seite 88

Fortsetzung Tabelle 19 von Seite 87

	Gesamtstichprobe Per Protocol Analyse n = 45	Kontrollgruppe A n = 30	Interventionsgruppe B‘ n = 15	Signifikanz (2-seitig) der Unterschiede A vs. B‘
Erwerbsminderungsrente bereits ≥ einmal abgelehnt	Ja: 3 (6,7%) Nein: 42 (93,3%)	Ja: 2 (6,7%) Nein: 28 (93,3%)	Ja: 1 (6,7%) Nein: 14 (93,3%)	1.0[c]
Rekrutierung in stationärer/ ambulanter Behandlung	Stationär: 16 (35,6%) Ambulant: 29 (64,4%)	Stationär: 13 (43,3%) Ambulant: 17 (56,7%)	Stationär: 3 (20%) Ambulant: 12 (80%)	.189[c]
Beschäftigungsverhältnis	Erwerbstätig: 33 (73,3%) Erwerbslos: 6 (13,3%) Bereits berentet: 6 (13,3%)	Erwerbstätig: 21 (70%) Erwerbslos: 4 (13,3%) Bereits berentet: 5 (16,7%)	Erwerbstätig: 12 (80%) Erwerbslos: 2 (13,3%) Bereits berentet: 1 (6,7%)	.642[a]
Brief Symptom Inventory (BSI)	*GSI:* MW: 1,2 (SD = 0,6) Med: 1,1 (range = 2,5)	*GSI:* MW: 1,3 (SD = 0,7) Med: 1,1 (range = 2,5)	*GSI:* MW: 1,2 (SD = 0,5) Med: 1,4 (range = 1,8)	.636[b]
Screen for Cognitive Impairment in Psychiatry (SCIP)	*Summenscore:* MW: 73,4 (SD = 11,8) Med: 75 (range = 52)	*Summenscore:* MW: 72,9 (SD = 10,8) Med: 75,5 (range = 44)	*Summenscore:* MW: 74,2 (SD = 14,1) Med: 74 (range = 44)	.739[b]

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Ein statistisch signifikanter Unterschied zwischen den Gruppen A und B‘ konnte für keines der in Tabelle 19 genannten Merkmale nachgewiesen werden.
Hinsichtlich soziodemographischer Variablen wie Alter (p = .280), Geschlechterverteilung (p = .738), Beschäftigungsverhältnis (p = .642), Bildungsabschluss (p = .138), Aufnahmediagnosen (jeweils p = 1.0, nur F41.x: p = .410), Umstände der Rekrutierung (p = .189), Vorliegen abgelehnter Anträge auf eine Erwerbsminderungsrente (p = 1.0) sowie auch hinsichtlich des kognitiven Leistungsniveaus – erfasst über das SCIP (p = .739) – und der aktuellen Symptomschwere – erfasst über das BSI (p = .636) – konnte zwischen den Gruppen A und B‘ kein statistisch signifikanter Unterschied festgestellt werden. Es ist also anzunehmen, dass die gleichmäßige Verteilung oben genannter Merkmale bzw. Störgrößen auf die Gruppen A und B‘ im Zuge der Randomisierung trotz des Ausschlusses von 15 Studienteilnehmern aus Gruppe B größtenteils erhalten geblieben ist.

3.2.3 SFSS-, WMT- und VAS-Ergebnisse der Gruppen A und B‘

Tabelle 20 fasst die in den Gruppen A und B‘ erzielten Ergebnisse bei WMT, SFSS und VAS zusammen. Die WMT- und SFSS-Ergebnisse sind jeweils auch mit den Ergebnissen der einzelnen Durchgänge (WMT) bzw. Subskalen (SFSS) dargestellt.

Tabelle 20: Gegenüberstellung der bei WMT, SFSS und VAS erzielten Ergebnisse in den beiden Studiengruppen A und B‘ (30 bzw. 15 Studienteilnehmer). Auftrennung der Testergebnisse von WMT und SFSS in die verschiedenen Testdurchgänge (WMT) bzw. Subskalen (SFSS). Untersuchung der Gruppenunterschiede hinsichtlich statistischer Signifikanz
[a] t-Test, [b] U-Test, MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum, vs.: versus, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MC: Multiple Choice, PA: Paired Associate Recall, FR: Free Recall, WMT: Word Memory Test, SFSS: Strukturierter Fragebogen Simulierter Symptome, VAS: Visuelle Analogskala, n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Kontrollgruppe A n = 30	**Interventionsgruppe B‘ n = 15**	**Signifikanz (2-seitig) der Unterschiede A vs. B‘**
Word Memory Test (WMT)			
- **IR** (in % richtig)	MW: 96,6 (SD = 8,2) Med: 100 (range = 42)	MW: 66,8 (SD = 24,9) Med: 73 (range = 77)	< .001[b***]
- **DR** (in % richtig)	MW: 96,3 (SD = 8,4) Med: 100 (range = 42)	MW: 67,9 (SD = 18,8) Med: 65 (range = 70)	< .001[b***]
- **CNS** (in % richtig)	MW: 95 (SD = 10,1) Med: 100 (range = 50)	MW: 65,5 (SD = 16) Med: 63 (range = 52)	< .001[b***]
- **MC** (in % richtig)	MW: 87,8 (SD = 18,6) Med: 95 (range = 80)	MW: 57,7 (SD = 21,9) Med: 60 (range = 75)	< .001[b***]

Fortsetzung von Tabelle 20 auf Seite 91

Fortsetzung Tabelle 20 von Seite 90

	Kontrollgruppe A n = 30	**Interventionsgruppe B‘ n = 15**	**Signifikanz (2-seitig) der Unterschiede A vs. B‘**
- **PA** (in % richtig)	MW: 85,7 (SD = 17,8) Med: 92,5 (range = 75)	MW: 54,3 (SD = 20,3) Med: 55 (range = 80)	< .001[b***]
- **FR** (in % richtig)	MW: 68,5 (SD = 17,7) Med: 73 (range = 75)	MW: 37,3 (SD = 14,3) Med: 40 (range = 47)	< .001[b***]
Score WMT-Durchgänge IR, DR, CNS	MW: 287,9 (SD = 26,3) Med: 300 (range = 134)	MW: 200,2 (SD = 54) Med: 196 (range = 164)	< .001[b***]
Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR	MW: 530 (SD = 75,5) Med: 559 (range = 364)	MW: 349,5 (SD = 95,2) Med: 318 (range = 312)	< .001[b***]
Strukturierter Fragebogen Simulierter Symptome (SFSS)	*Gesamtscore:* MW: 14,1 (SD = 6,5) Med: 13 (range = 25)	*Gesamtscore:* MW: 32,7 (SD = 8,5) Med: 34 (range = 29)	< .001[b***]
SFSS-Skala neurologische Beeinträchtigung (N, 15 Items**)**	MW: 3,5 (SD = 2,3) Med: 3 (range = 9)	MW: 8,3 (SD = 3,2) Med: 9 (range = 10)	< .001[b***]

Fortsetzung von Tabelle 20 auf Seite 92

Fortsetzung Tabelle 20 von Seite 91

	Kontrollgruppe A n = 30	**Interventionsgruppe B‘ n = 15**	**Signifikanz (2-seitig) der Unterschiede A vs. B‘**
SFSS-Skala affektive Störungen (AF, 15 Items**)**	MW: 6,5 (SD = 2,6) Med: 6,5 (range = 11)	MW: 10,1 (SD = 2,2) Med: 10 (range = 8)	< .001[a***]
SFSS-Skala niedrige Intelligenz (NI, 15 Items**)**	MW: 1,3 (SD = 1,2) Med: 1 (range = 4)	MW: 3,7 (SD = 2,3) Med: 3 (range = 8)	< .001[b***]
SFSS-Skala Psychose (P, 15 Items**)**	MW: 0,9 (SD = 1,1) Med: 1 (range = 5)	MW: 3,3 (SD = 3,2) Med: 2 (range = 9)	.006[b**]
SFSS-Skala amnestische Störung (AM, 15 Items**)**	MW: 1,9 (SD = 2,2) Med: 1 (range = 8)	MW: 7,3 (SD = 2,5) Med: 7 (range = 8)	< .001[b***]
Visuelle Analogskala (VAS), subjektive Anstrengung in %	MW: 99,7 (SD = 1,1) Med: 100 (range = 6)	MW: 54,7 (SD = 15,9) Med: 56 (range = 53)	< .001[b***]

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Für alle in Tabelle 20 beschriebenen Merkmale zeigten sich jeweils statistisch signifikante Unterschiede zwischen den Gruppen A und B‘. Hervorzuheben ist vor allem, dass die von Studieneilnehmern in Gruppe B‘ über die VAS angegebene subjektive Anstrengung signifikant niedriger war als die subjektiv angegebene Anstrengung in Gruppe A (p < .001). In Gruppe A lag die subjektiv angegebene Anstrengung während der Testdurchführung im Mittel bei 99,7% (Median: 100%), gegenüber einem Mittelwert von 54,7% in Gruppe B‘ (Median: 56%). Studienteilnehmer in Gruppe A erzielten außerdem signifikant niedrigere SFSS-

Gesamtscores als Studienteilnehmer in Gruppe B‘ (Mittelwert: 14,1 in Gruppe A gegenüber Mittelwert: 32,7 in Gruppe B‘, $p < .001$). Sowohl die Werte im Score WMT-Durchgänge IR, DR, CNS (Mittelwert: 287,9 in Gruppe A gegenüber Mittelwert: 200,2 in Gruppe B‘) als auch die Werte im Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR (Mittelwert: 530 in Gruppe A gegenüber Mittelwert: 349,5 in Gruppe B‘) waren hingegen in Gruppe A signifikant höher als in Gruppe B‘ ($p < .001$ bzw. $p < .001$). Aufgrund dieser Ergebnisse wurde angenommen, dass die Instruktionen A und B in den Gruppen wirksam waren und alle weiteren geplanten statistischen Analysen konnten durchgeführt werden.

3.2.3.1 Analyse der SFSS-Ergebnisse

Eine eingehendere deskriptive Analyse der SFSS-Ergebnisse in den Gruppen A und B‘ findet sich in Tabelle 21.

Tabelle 21: Deskriptive Darstellung und Analyse der SFSS-Ergebnisse aller 45 Studieneilnehmer der Per Protocol Analyse sowie Gegenüberstellung der in den beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) erzielten SFSS-Ergebnisse. Ein SFSS-Gesamtscore von >16 Punkten wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests. Auflistung bestimmter häufig von Studienteilnehmern angekreuzter Testitems
MW: Mittelwert, SD: Standardabweichung, Med: Median, range: Maximum – Minimum, PPA: Per Protocol Analyse, SFSS: Strukturierter Fragebogen Simulierter Symptome, AF-Skala: Skala für affektive Störungen (SFSS), Nr.: Nummer, n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Gesamt-stichprobe PPA n = 45	**Kontrollgruppe A n = 30**	**Interventions-gruppe B‘ n = 15**
SFSS-Gesamtscore >16	23 von 45 (51,1%) Studienteil-nehmern	9 von 30 (30%) Studienteilnehmern	14 von 15 (93,3%) Studienteilnehmern
Anteil AF-Skala-Ergebnis an SFSS-Gesamtscore (Anteile angegeben in %)	MW: 43% (SD = 14,5%) Med: 41,4% (range = 54,8%)	MW: 48,3% (SD = 14,2%) Med: 49,7% (range = 54,8%)	MW: 32,3% (SD = 7,6%) Med: 32,4% (range = 24,9%)

Fortsetzung von Tabelle 21 auf Seite 95

Fortsetzung Tabelle 21 von Seite 94

	Gesamt-stichprobe PPA n = 45	**Kontrollgruppe A n = 30**	**Interventions-gruppe B‘ n = 15**
Item Nr. 2: Wenn meine Depression zu stark wird, mache ich lange Spaziergänge oder Sport, um die Spannung zu vermindern.	23 von 45 (51,1%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	20 von 30 (66,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“	3 von 15 (20%) Studienteilnehmern beantworteten dieses Item mit „Ja“
Item Nr. 19: Manchmal bin ich so deprimiert, dass ich mich nach meinem Bett sehne, um die schlechte Stimmung weg-zuschlafen.	34 von 45 (75,6%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	20 von 30 (66,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“	14 von 15 (93,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“
Item Nr. 32: Ich kann schlecht schlafen.	34 von 45 (75,6%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	20 von 30 (66,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“	14 von 15 (93,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“

Fortsetzung von Tabelle 21 auf Seite 96

Fortsetzung Tabelle 21 von Seite 95

	Gesamt-stichprobe PPA n = 45	**Kontrollgruppe A n = 30**	**Interventions-gruppe B‘ n = 15**
Item Nr. 35: Manchmal sind meine Muskeln ohne ersichtlichen Grund so schlaff und kraftlos, dass sich meine Arme und Beine anfühlen, als ob sie zentnerschwer wären.	32 von 45 (71,1%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	19 von 30 (63,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“	13 von 15 (86,7%) Studienteilnehmern beantworteten dieses Item mit „Ja“
Item Nr. 52: Es scheint, dass ich nicht mehr die Energie habe, die ich früher immer gehabt habe.	43 von 45 (95,6%) Studien-teilnehmern beantworteten dieses Item mit „Ja“	28 von 30 (93,3%) Studienteilnehmern beantworteten dieses Item mit „Ja“	15 von 15 (100%) Studienteilnehmern beantworteten dieses Item mit „Ja“

Bei der Betrachtung von Tabelle 21 fällt auf, dass über die Hälfte (51,1%, 23 von 45 Studienteilnehmern) der Studienteilnehmer der PPA den SFSS bei seinem üblichen Cut-off von >16 nicht bestand, also einen höheren SFSS-Gesamtscore als 16 erzielte. Neben 14 von 15 (93,3%) Teilnehmern der Interventionsgruppe B‘ bestanden auch 9 von 30 Studienteilnehmern (30%) in Kontrollgruppe A den SFSS für den Cut-off von >16 nicht. Der Anteil des AF-Skala-Ergebnisses am SFSS-Gesamtscore betrug in Kontrollgruppe A im Mittel 48,3% (vgl. 32,3% in

Gruppe B‘). Tabelle 21 listet zudem die SFSS-Items Nr. 2, 19, 32, 35 und 52 auf, bei denen besonders viele Studienteilnehmer eine – laut SFSS-Decodierungstabelle – als auffällig einzustufende Antwort ankreuzten (bei diesen Items jeweils die Antwort „Ja“). Bei allen dieser Items lag der Anteil der Studienteilnehmer in Kontrollgruppe A, die das Item bejaht hatten, bei 63,3% und höher. Item Nr. 52 wurde sogar von 93,3% (28 von 30 Studienteilnehmern) der Studienteilnehmer in Gruppe A bejaht.

3.2.3.2 Analyse der WMT-Ergebnisse

Eine eingehendere deskriptive Analyse der WMT-Ergebnisse in den Gruppen A und B‘ zeigt Tabelle 22.

Tabelle 22: Deskriptive Darstellung und Analyse der WMT-Ergebnisse aller 45 Studieneilnehmer der Per Protocol Analyse sowie Gegenüberstellung der in den beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) erzielten WMT-Ergebnisse. Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests. Betrachtung weiterer relevanter Ergebnisbereiche des WMT
PPA: Per Protocol Analyse, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, WMT: Word Memory Test, n: Stichprobenumfang
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	Gesamt-stichprobe PPA n = 45	**Kontrollgruppe A n = 30**	**Interventionsgruppe B‘ n = 15**
IR- oder DR- oder CNS-Ergebnis (WMT) ≤82,5%	14 von 45 (31,1%) Studienteilnehmern	2 von 30 (6,7%) Studienteilnehmern	12 von 15 (80%) Studienteilnehmern
IR- oder DR- oder CNS-Ergebnis (WMT) ≥50% - ≤60%	7 von 45 (15,6%) Studienteilnehmern	1 von 30 (3,3%) Studienteilnehmern	6 von 15 (40%) Studienteilnehmern
IR- oder DR- oder CNS-Ergebnis (WMT) <50%	5 von 45 (11,1%) Studienteilnehmern dabei: 3 von 45 (6,7%) Studienteilnehmern mit IR- oder DR-Ergebnis <37,5%	0 von 30 (0%) Studienteilnehmern dabei: 0 von 30 (0%) Studienteilnehmern mit IR- oder DR-Ergebnis <37,5%	5 von 15 (33,3%) Studienteilnehmern dabei: 3 von 15 (20%) Studienteilnehmern mit IR- oder DR-Ergebnis <37,5%

Aus Tabelle 22 wird ersichtlich, dass 2 von 30 Studienteilnehmern (6,7%) in Kontrollgruppe A den WMT bei Anwendung des üblichen Cut-offs von ≤82,5% nicht bestanden, während 12 von 15 Studienteilnehmern (80%) in der Interventionsgruppe B‘ ein Ergebnis unterhalb dieses Cut-offs erzielten. Weiterhin fällt auf, dass 7 von 45 Studienteilnehmern der Gesamtstichprobe der PPA (15,6%) in den „einfachen“ WMT-Durchgängen (IR, DR, CNS) einen Wert zwischen 50% und 60% erzielten – davon 6 Studienteilnehmer in Gruppe B‘. Allerdings schloss auch einer von 30 Studienteilnehmern (3,3%) in Kontrollgruppe A die „einfachen“ WMT-Durchgänge mit einem Ergebnis zwischen 50% und 60% ab. Ein WMT-Ergebnis <50% bzw. <37,5% (= Unter-Zufall-Schwelle) im IR-, DR- oder CNS-Durchgang erzielten ausschließlich Studienteilnehmer der Interventionsgruppe B‘ (5 von 15 Studienteilnehmer (33,3%) bzw. 3 von 15 Studienteilnehmern (20%)).

3.2.4 Sensitivität, Spezifität, PPW und NPW von SFSS und WMT

Die Tabellen 23 und 24 zeigen jene Vierfeldertafeln, die zur Berechnung von Sensitivität, Spezifität, PPW und NPW von SFSS und WMT in der Gesamtstichprobe der PPA für die gebräuchlichen Cut-offs (>16 (SFSS) bzw. ≤82,5% (WMT)) benutzt wurden. Die Vierfeldertafeln sollen beispielhaft darstellen, nach welchem Prinzip das Statistikprogramm IBM® SPSS® Statistics 21 im Rahmen der ROC-Kurven-Analyse für jeden variierten Cut-off Sensitivität und Spezifität errechnete und zudem eine Übersicht über die BVT-Klassifikationen in Kontroll- und Interventionsgruppe der PPA bieten.

Tabelle 23: Darstellung der Klassifikationen aller 45 Studienteilnehmer der Per Protocol Analyse für den WMT-Cut-off von ≤82,5% in Form einer Vierfeldertafel. Unterteilung der Klassifikationen nach den beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer). Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests
WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	WMT-Ergebnis ≤82,5% (auffällig)	**WMT-Ergebnis >82,5% (unauffällig)**	***Summe***
experimentelle negative Antwortverzerrer (Gruppe B‘)	12	3	15
subjektiv bestmögliche Leistungsanstrengung/ unverzerrtes Antwort-verhalten (Gruppe A)	2	28	30
Summe	14	31	45

Für den üblichen WMT-Cut-off von ≤82,5% bzw. obige Vierfeldertafel ergaben sich folgende Testkennwerte:

Sensitivität: 80% (95% KI = 56,0% - 94,6%)

Spezifität: 93,3% (95% KI = 80,0% - 98,9%)

Positiver prädiktiver Wert (PPW): 85,7% (95% KI = 62,1% - 97,5%)

Negativer prädiktiver Wert (NPW): 90,3% (95% KI = 76,8% - 97,5%).

Tabelle 24 stellt die Klassifikationsergebnisse des SFSS in der Gesamtstichprobe der PPA für den Cut-off von >16 dar.

Tabelle 24: Darstellung der Klassifikationen aller 45 Studienteilnehmer der Per Protocol Analyse für den SFSS-Cut-off von >16 in Form einer Vierfeldertafel. Unterteilung der Klassifikationen nach den beiden Studiengruppen A und B' (30 bzw. 15 Teilnehmer). Ein SFSS-Gesamtscore von >16 Punkten wird als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des Tests
SFSS: Strukturierter Fragebogen Simulierter Symptome
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	SFSS-Ergebnis >16 (auffällig)	**SFSS-Ergebnis ≤16 (unauffällig)**	***Summe***
experimentelle negative Antwortverzerrer (Gruppe B')	14	1	15
subjektiv bestmögliche Leistungsanstrengung/ unverzerrtes Antwort-verhalten (Gruppe A)	9	21	30
Summe	23	22	45

Für den üblichen SFSS-Cut-off von >16 bzw. obige Vierfeldertafel ergaben sich folgende Testkennwerte:

Sensitivität: 93,3% (95% KI = 73,8% - 99,6%)

Spezifität: 70% (95% KI = 52,4% - 84,3%)

Positiver prädiktiver Wert (PPW): 60,9% (95% KI = 40,6% - 78,9%)

Negativer prädiktiver Wert (NPW): 95,5% (95% KI = 81,5% - 99,7%).

3.2.4.1 Falsch-positive SFSS-/WMT-Klassifikationen

An dieser Stelle folgte im Abschnitt 3.1 (Auswertung der Gesamtstichprobe) eine genauere deskriptive Analyse der falsch-positiven und richtig-negativen Klassifikationen von SFSS und WMT in Kontrollgruppe A. Da sich die Zusammensetzung von Gruppe A für die PPA aber nicht veränderte, die Analysen also identisch wären, sei hier auf die Tabellen 7 und 8 im Abschnitt 3.1 verwiesen.

3.2.5 ROC-Kurven-Analyse

3.2.5.1 ROC-Kurven-Analyse SFSS

Abbildung 8 zeigt den Verlauf der ROC-Kurve des SFSS bei n = 45 (PPA). In Grün gegenübergestellt wird der SFSS-ROC-Kurve eine Diagonale, die den Verlauf der ROC-Kurve eines reinen Zufallsverfahrens wiedergibt.

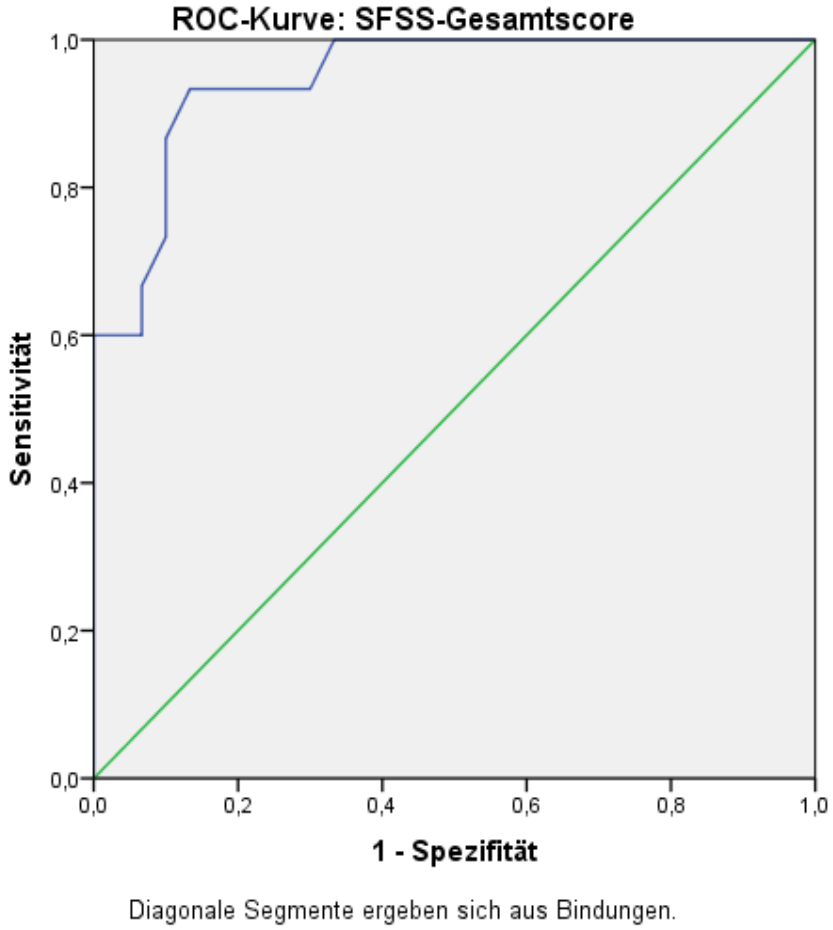

Abbildung 8: ROC-Kurven-Analyse für den SFSS-Gesamtscore aller 45 Studienteilnehmer der Per Protocol Analyse. Die ROC-Kurve eines diagnostischen Zufallsverfahrens (grün, diagonal) ist der ROC-Kurve des SFSS (blau) zu Vergleichszwecken gegenübergestellt. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde sich als Punkt in der oberen linken Ecke darstellen
SFSS: Strukturierter Fragebogen Simulierter Symptome, ROC: receiver operating characteristic
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Die in Abbildung 8 blau eingezeichnete ROC-Kurve des SFSS setzt sich deutlich von der in grün eingezeichneten ROC-Kurve des Zufallsverfahrens ab und nähert sich der oberen linken Ecke des Schaubildes, in der ein perfektes diagnostisches Verfahren mit einer Sensitivität und Spezifität von 100% liegen würde. Das

Schaubild in Abbildung 8 spricht also dafür, dass der SFSS als Unterscheidungsverfahren zwischen den in der vorliegenden Studie gebildeten Zuständen „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“ und „experimentelle negative Antwortverzerrer der Interventionsgruppe B‘“ – rein graphisch beurteilt – deutlich aussagekräftiger ist als ein Zufallsverfahren.

3.2.5.1.1 Fläche unter der ROC-Kurve (SFSS)

In Tabelle 25 findet sich die für das Schaubild in Abbildung 8 berechnete Fläche unter der ROC-Kurve (AUC, maximaler Wert = 1) mit 95%igem Konfidenzintervall. Sie ist ein Maß für die Güte des untersuchten diagnostischen Verfahrens.

Tabelle 25: Darstellung des für die Fläche unter der SFSS-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde einen AUC-Wert von 1 erzielen
[a] Unter der nicht-parametrischen Annahme, [b] Nullhypothese: Wahrheitsfläche = .5, SFSS: Strukturierter Fragebogen Simulierter Symptome, ROC: receiver operating characteristic, AUC: area under the curve
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Fläche unter ROC-Kurve des SFSS (AUC)	**Standardfehler[a]**	**Asymptotische Signifikanz[b]**	**Asymptotisches 95% Konfidenzintervall**	
			Untergrenze	Obergrenze
.95	.03	< .001***	.89	1

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Der für die AUC der SFSS-ROC-Kurve berechnete Wert von .95 liegt nahe dem maximal möglichen AUC-Wert von 1 und spricht daher für den SFSS als brauchbares Unterscheidungsverfahren zwischen den Zuständen „experimentelle negative Antwortverzerrer der Interventionsgruppe B‘“ und „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“. Die AUC der SFSS-ROC-Kurve unterschied sich zudem statistisch signifikant (p < .001) von der AUC des Zufallsverfahrens (AUC = .5). Als Untergrenze des 95%igen Konfidenzintervalls ergab sich ein SFSS-AUC-Wert von .89.

3.2.5.1.2 Koordinaten der ROC-Kurve (SFSS)

Tabelle 26 enthält die Koordinaten der in Abbildung 8 dargestellten SFSS-ROC-Kurve und listet dazu in der ersten Spalte die verschiedenen SFSS-Cut-offs auf, für die jeweils Sensitivität, Spezifität, PPW und NPW des SFSS in der Gesamtstichprobe der PPA berechnet wurden. Ein in der ersten Spalte angegebener Cut-off von 6,5 z.B. bedeutet, dass der zugehörige SFSS-Cut-off für eine Klassifikation in den Zustand „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ effektiv bei >6 liegt, da eigentlich nur ganzzahlige SFSS-Gesamtscores möglich sind. Die letzte Spalte von Tabelle 26 enthält das ebenfalls für jeden Cut-off berechnete Maß für die Güte des jeweiligen Cut-offs (Youden-Index, maximaler Wert = 1). Die für die Berechnung von PPW und NPW angenommene Basalrate negativer Antwortverzerrung in der Population lag bei 33,3% (15 von 45 Studienteilnehmern).

Tabelle 26: Auflistung der Koordinaten der SFSS-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten SFSS-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind. Die variierten Cut-offs beziehen sich auf den SFSS-Gesamtscore. Der höchste Youden-Index-Wert ist rot hervorgehoben
Basalrate: Basalrate negativer Antwortverzerrung in der untersuchten Population, SFSS: Strukturierter Fragebogen Simulierter Symptome, PPW: positiver prädiktiver Wert, NPW: negativer prädiktiver Wert, ROC: receiver operating characteristic, Youden-Index: Sensitivität + Spezifität – 1
[a] Der kleinste Cut-off entspricht dem kleinsten beobachteten Testwert minus 1, der größte Cut-off dem größten beobachteten Testwert plus 1. Alle anderen Cut-offs sind Mittelwerte zweier aufeinanderfolgender, geordneter beobachteter Testwerte
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Positives Test-ergebnis, wenn größer oder gleich[a] (Variable: SFSS-Gesamtscore)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 33,3%)*	**NPW (%)** *(Basalrate: 33,3%)*	**Youden-Index**
5	100	0	-	-	0
6	100	10	35,7	~100	10
6,5	100	13,3	36,6	~100	0,133
8,1	100	20	38,5	~100	0,2

Fortsetzung von Tabelle 26 auf Seite 106

Fortsetzung Tabelle 26 von Seite 105

Positives Test-ergebnis, wenn größer oder gleich[a] (Variable: SFSS-Gesamtscore)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 33,3%)*	**NPW (%)** *(Basalrate: 33,3%)*	**Youden-Index**
10,1	100	23,3	39,5	~100	0,233
11,5	100	26,7	40,6	~100	0,267
12,5	100	43,3	46,9	~100	0,433
13,5	100	60	55,6	~100	0,6
14,5	100	63,3	57,7	~100	0,633
15,5	100	66,7	60	~100	0,667
16,5	93,3	70	60,9	95,4	0,633
18	93,3	83,3	73,6	96,1	0,766
20,5	93,3	86,7	77,8	96,3	**0,8**
23,5	86,7	90	81,3	93,1	0,767
25,5	80	90	80	90	0,7
26,5	73,3	90	78,6	87,1	0,633
27,5	66,7	93,3	83,3	84,9	0,6
28,5	60	93,3	81,7	82,4	0,533
30	60	96,7	90,1	82,9	0,567
32,5	60	100	~100	83,3	0,6
35	40	100	~100	76,9	0,4
37	33,3	100	~100	75	0,333
38,5	26,7	100	~100	73,2	0,267
40,5	20	100	~100	71,4	0,2

Fortsetzung von Tabelle 26 auf Seite 107

Fortsetzung Tabelle 26 von Seite 106

Positives Test-ergebnis, wenn größer oder gleich[a] (Variable: SFSS-Gesamtscore)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 33,3%)*	**NPW (%)** *(Basalrate: 33,3%)*	**Youden-Index**
43	13,3	100	~100	69,8	0,133
44,5	6,7	100	~100	68,2	0,067
46	0	100	-	-	0

Der höchste Wert für den Youden-Index (= 0,8) ergab sich laut Tabelle 26 für einen SFSS-Cut-off von ≥20,5, also effektiv >20, da beim SFSS eigentlich nur ganzzahlige Gesamtscores möglich sind. Dieser Cut-off ging mit einer Sensitivität von 93,3%, einer Spezifität von 86,7% und einem PPW bzw. NPW von 77,8% bzw. 96,3% einher. Wenn andere Sensitivitäts- oder Spezifitätswerte gewünscht sind, kann der SFSS-Cut-off entsprechend Tabelle 26 variiert werden. Eine Spezifität von >90% wurde laut Tabelle 26 erst bei SFSS-Cut-offs von ≥27,5 (effektiv >27) und höher erreicht, bei gleichzeitiger Sensitivität von ≤66,7%. Falsch-positive SFSS-Klassifikationen kamen in der vorliegenden Gesamtstichprobe der PPA bei einem Cut-off von ≥32,5 (effektiv >32) und höher nicht mehr vor (Spezifität = 100%, Sensitivität ≤60%). Der übliche SFSS-Cut-off von >16 erzielte laut Tabelle 26 eine Sensitivität und Spezifität von 93,3% bzw. 70% (vgl. Vierfeldertafel/Tabelle 24).

3.2.5.2 ROC-Kurven-Analyse WMT

In Abbildung 9 ist der Verlauf der ROC-Kurve des WMT für n = 45 (PPA) dargestellt. Dabei wird der WMT-ROC-Kurve (blau) eine Diagonale (grün) gegenübergestellt, die den Verlauf der ROC-Kurve eines reinen Zufallsverfahrens wiedergibt.

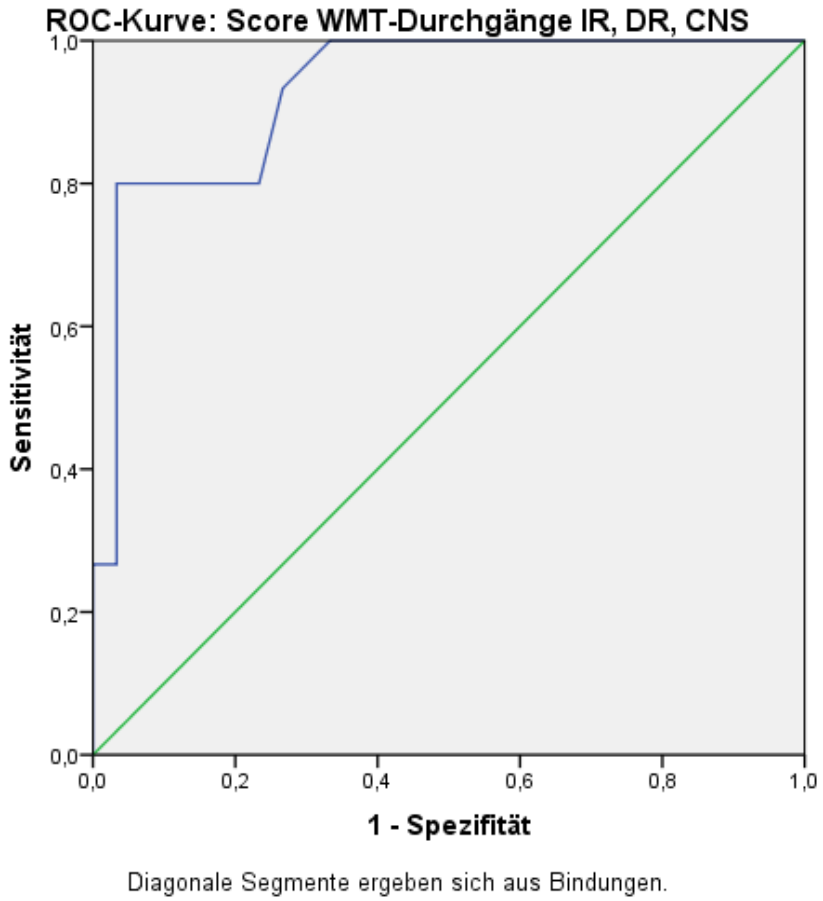

Abbildung 9: ROC-Kurven-Analyse für den Score WMT-Durchgänge IR, DR, CNS aller 45 Studienteilnehmer der Per Protocol Analyse. Die ROC-Kurve eines diagnostischen Zufallsverfahrens (grün, diagonal) ist der ROC-Kurve des WMT (blau) zu Vergleichszwecken gegenübergestellt. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde sich als Punkt in der oberen linken Ecke darstellen. Der Score WMT-Durchgänge IR, DR, CNS entstand durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte
WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, ROC: receiver operating characteristic
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Die in Abbildung 9 blau eingezeichnete ROC-Kurve des WMT setzt sich deutlich von der in grün eingezeichneten ROC-Kurve des Zufallsverfahrens ab und nähert sich tendenziell der oberen linken Ecke des Schaubildes. Dies spricht dafür, dass

der WMT als Unterscheidungsverfahren zwischen den in der vorliegenden Studie gebildeten Zuständen „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“ und „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ – rein graphisch beurteilt – deutlich aussagekräftiger ist als ein Zufallsverfahren.

3.2.5.2.1 Fläche unter der ROC-Kurve (WMT)

Quantifizieren lässt sich die graphisch basierte Beurteilung der in Abbildung 9 dargestellten WMT-ROC-Kurve über die in Tabelle 27 angegebene Fläche unter der ROC-Kurve (AUC, maximaler Wert = 1). Sie ist ein Maß für die Güte des untersuchten diagnostischen Verfahrens.

Tabelle 27: Darstellung des für die Fläche unter der WMT-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall. Ein perfektes diagnostisches Verfahren mit 100% Sensitivität und Spezifität würde einen AUC-Wert von 1 erzielen
[a] Unter der nicht-parametrischen Annahme, [b] Nullhypothese: Wahrheitsfläche = .5, WMT: Word Memory Test, ROC: receiver operating characteristic, AUC: area under the curve
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Fläche unter ROC-Kurve des WMT (AUC)	**Standardfehler[a]**	**Asymptotische Signifikanz[b]**	**Asymptotisches 95% Konfidenzintervall**	
			Untergrenze	Obergrenze
.93	.04	< .001***	.85	1

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Tabelle 27 gibt für die AUC der WMT-ROC-Kurve einen Wert von .93 – also nahe dem maximal möglichen AUC-Wert von 1 – an. Dieses Ergebnis spricht für den WMT als brauchbares Unterscheidungsverfahren zwischen den Zuständen „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ und „Kontrollgruppe A mit subjektiv bestmöglicher Leistungsanstrengung“. Zudem war der Unterschied zwischen der AUC der WMT-ROC-Kurve (AUC = .93) und der AUC des Zufallsverfahrens (AUC = .5) statistisch signifikant (p < .001). Als Untergrenze des 95%igen Konfidenzintervalls ergab sich ein WMT-AUC-Wert von .85.

3.2.5.2.2 Koordinaten der ROC-Kurve (WMT)

Tabelle 28 enthält die Koordinaten der in Abbildung 9 dargestellten WMT-ROC-Kurve und listet dazu in der ersten Spalte die verschiedenen variierten WMT-Cut-offs auf, für die jeweils Sensitivität, Spezifität, PPW und NPW des WMT in der Gesamtstichprobe der PPA berechnet wurden. Ein in der ersten Spalte angegebener Cut-off von 133,5 z.B. bedeutet, dass der zugehörige WMT-Cut-off für eine Klassifikation in den Zustand „experimentelle negative Antwortverzerrer der Interventionsgruppe B“ effektiv bei <134 liegt, da es nur ganzzahlige Ergebnisse für den Score WMT-Durchgänge IR, DR, CNS gibt. Die letzte Spalte von Tabelle 28 enthält wiederum das für jeden Cut-off berechnete Maß für die Güte des jeweiligen Cut-offs (Youden-Index, maximaler Wert = 1). Die zur Berechnung von PPW und NPW angenommene Basalrate negativer Antwortverzerrung in der Population lag bei 33,3% (15 von 45 Studienteilnehmern).

Tabelle 28: Auflistung der Koordinaten der WMT-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten WMT-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind. Die variierten Cut-offs beziehen sich auf den Score WMT-Durchgänge IR, DR, CNS, welcher durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte entstand. Der höchste Youden-Index-Wert ist rot hervorgehoben

Basalrate: Basalrate negativer Antwortverzerrung in der untersuchten Population, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, PPW: positiver prädiktiver Wert, NPW: negativer prädiktiver Wert, ROC: receiver operating characteristic, Youden-Index: Sensitivität + Spezifität – 1

[a] Der kleinste Cut-off entspricht dem kleinsten beobachteten Testwert minus 1, der größte Cut-off dem größten beobachteten Testwert plus 1. Alle anderen Cut-offs sind Mittelwerte zweier aufeinanderfolgender, geordneter beobachteter Testwerte

Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Positives Testergebnis, wenn kleiner oder gleich[a] (Variable: Score WMT-Durchgänge IR, DR, CNS)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 33,3%)*	**NPW (%)** *(Basalrate: 33,3%)*	**Youden-Index**
125	0	100	-	-	0

Fortsetzung von Tabelle 28 auf Seite 111

Fortsetzung Tabelle 28 von Seite 110

Positives Testergebnis, wenn kleiner oder gleich[a] (Variable: Score WMT-Durchgänge IR, DR, CNS)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 33,3%)*	**NPW (%)** *(Basalrate: 33,3%)*	**Youden-Index**
133,5	13,3	100	~100	69,8	0,133
151	20	100	~100	71,4	0,2
163,5	26,7	100	~100	73,2	0,267
168,5	26,7	96,7	80,2	72,5	0,234
178,5	33,3	96,7	83,5	74,4	0,3
188,5	40	96,7	85,8	76,3	0,367
193,5	46,7	96,7	87,6	78,4	0,434
198,5	53,3	96,7	89	80,6	0,5
203,5	60	96,7	90,1	82,9	0,567
208	66,7	96,7	91	85,3	0,634
218	73,3	96,7	91,7	87,9	0,7
240,5	80	96,7	92,4	90,6	**0,767**
260	80	93,3	85,7	90,3	0,733
265,5	80	90	80	90	0,7
268,5	80	86,7	75	89,7	0,667
275,5	80	83,3	70,5	89,3	0,633
280,5	80	80	66,7	88,9	0,6
283,5	80	76,7	63,2	88,5	0,567
288	93,3	73,3	63,6	95,6	0,666

Fortsetzung von Tabelle 28 auf Seite 112

Fortsetzung Tabelle 28 von Seite 111

Positives Testergebnis, wenn kleiner oder gleich[a] (Variable: Score WMT-Durchgänge IR, DR, CNS)	**Sensitivität (%)**	**Spezifität (%)**	**PPW (%)** *(Basalrate: 33,3%)*	**NPW (%)** *(Basalrate: 33,3%)*	**Youden-Index**
293	100	66,7	60	~100	0,667
298	100	56,7	53,6	~100	0,567
301	100	0	-	-	0

Der höchste Youden-Index-Wert (= 0,767) in Tabelle 28 ergab sich für einen WMT-Cut-off von ≤240,5, also effektiv <241, da für den Score WMT-Durchgänge IR, DR, CNS nur ganzzahlige Ergebnisse möglich sind. Der Cut-off von ≤240,5 (effektiv <241) ging mit einer Sensitivität von 80%, einer Spezifität von 96,7% und einem PPW bzw. NPW von 92,4% bzw. 90,6% einher. Sind andere Sensitivitäts- oder Spezifitätswerte gewünscht, kann der WMT-Cut-off entsprechend Tabelle 28 variiert werden. Eine Sensitivität von >90% wurde laut Tabelle 28 erst bei Cut-offs von ≤288 (einfacher ausgedrückt <289) und höher erreicht, bei gleichzeitiger Spezifität von ≤73,3%. Falsch-positive WMT-Klassifikationen kamen in der vorliegenden Gesamtstichprobe der PPA bei einem Cut-off von ≤163,5 (effektiv <164) und niedriger nicht mehr vor (Spezifität = 100%, Sensitivität ≤26,7%).

3.2.6 Übereinstimmung zwischen den BVT-Klassifikationen

In Tabelle 29 ist die Übereinstimmung der Klassifikationsergebnisse von WMT und SFSS bei Verwendung der gebräuchlichen Cut-offs von >16 (SFSS) bzw. ≤82,5% (WMT) dargestellt. Als Maßzahlen für die Übereinstimmung beider Testverfahren finden sich unter Tabelle 29 der Vierfelderkorrelationskoeffizient (Φ, phi) sowie der Kappa-Koeffizient κ.

Tabelle 29: Darstellung der Übereinstimmung der Klassifikationsergebnisse von SFSS und WMT (bestanden/nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für alle 45 Studienteilnehmer der Per Protocol Analyse. Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang bzw. ein SFSS-Gesamtscore >16 Punkten wird jeweils als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des jeweiligen Tests
WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, SFSS: Strukturierter Fragebogen Simulierter Symptome
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

		WMT		
		bestanden	**nicht bestanden**	***Summe***
SFSS	**bestanden**	21	1	22
	nicht bestanden	10	13	23
	Summe	31	14	45

Vierfelderkorrelationskoeffizient (Φ, phi) = .56 (p < .001, 95% KI = .34 - .76)

Kappa-Koeffizient κ: .52 (p < .001, 95% KI = .3 - .73).

Tabelle 29 zeigt, dass 34 der 45 Studienteilnehmer (75,5%) der PPA von beiden Beschwerdevalidierungstests gleich klassifiziert wurden (jeweils bestanden bzw. nicht bestanden). 10 der 45 Studienteilnehmer (22,2%) bestanden zwar den WMT, nicht aber den SFSS. Bei einem der 45 Studienteilnehmer (2,2%) verhielt es sich genau anders herum (WMT nicht bestanden, SFSS bestanden). Sowohl der für den Vierfelderkorrelationskoeffizienten als auch der für den Kappa-Koeffizienten berechnete Wert war statistisch signifikant (p < .001). In der vorliegenden Studie

ist also von einer statistisch signifikanten Übereinstimmung zwischen den Klassifikationen der beiden Beschwerdevalidierungstests auszugehen.

3.2.7 Klassifikationen bei Verwendung von SFSS *und* WMT

Tabelle 30 (Seite 115) zeigt, welche Klassifikationen und Testkennwerte sich in der vorliegenden Gesamtstichprobe der PPA für die Cut-offs von >16 (SFSS) bzw. ≤82,5% (WMT) ergeben, wenn folgendes Vorgehen bei der Verwendung der BVTs festgelegt wird:

Regeln: 1. Beide BVTs müssen bestanden oder nicht-bestanden sein für eine Klassifikation. 2. Studienteilnehmer mit nicht übereinstimmenden Klassifikationen werden in beiden Studiengruppen wegen Unsicherheit ausgeschlossen (Gruppe A: 9 Studienteilnehmer ausgeschlossen, Gruppe B‘: 2 Studienteilnehmer ausgeschlossen).

Tabelle 30: Darstellung der Klassifikationsergebnisse von SFSS und WMT bei kombinierter Anwendung der beiden BVTs (Tests beide bestanden/beide nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für alle 45 Studienteilnehmer der Per Protocol Analyse. Bei der kombinierten Anwendung der beiden BVTs SFSS und WMT galten folgende Regeln: 1. Beide Tests müssen bestanden oder nicht-bestanden sein für eine Klassifikation. 2. Studienteilnehmer mit nicht übereinstimmenden Klassifikationen werden wegen Unsicherheit ausgeschlossen. Letztlich mussten für diese Art der Auswertung 11 der insgesamt 45 Studienteilnehmer der Per Protocol Analyse wegen differierender BVT-Klassifikationen (z.B. WMT bestanden/SFSS nicht bestanden) ausgeschlossen werden. Es ergab sich eine angepasste Stichprobe von 34 Studienteilnehmern. Unterteilung der Klassifikationen nach den beiden Studiengruppen A und B' (ursprünglich 30 bzw. 15 Studienteilnehmer). Ein WMT-Ergebnis ≤82,5% im IR-, DR- oder CNS-Durchgang bzw. ein SFSS-Gesamtscore >16 Punkten wird jeweils als auffälliges Antwortverhalten im Sinne von negativer Antwortverzerrung bzw. Simulation gewertet und führt zum Nichtbestehen des jeweiligen Tests
BVT: Beschwerdevalidierungstest, WMT: Word Memory Test, SFSS: Strukturierter Fragebogen Simulierter Symptome, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

	BVT-Ergebnisse (WMT *und* SFSS) auffällig	**BVT-Ergebnisse (WMT *und* SFSS) unauffällig**	***Summe***
Experimentelle negative Antwortverzerrer (Gruppe B')	12	1	13
subjektiv bestmögliche Leistungsanstrengung/ unverzerrtes Antwort-verhalten (Gruppe A)	1	20	21
Summe	13	21	34

Bei kombinierter Verwendung beider BVTs bzw. für obige Vierfeldertafel ergaben sich folgende Testkennwerte:

Sensitivität: 92,3% (95% KI = 70,3% - 99,5%)

Spezifität: 95,2% (95% KI = 80,7% - 99,7%)

Positiver prädiktiver Wert (PPW): 92,3% (95% KI = 70,3% - 99,5%)

Negativer prädiktiver Wert (NPW): 95,2% (95% KI = 80,7% - 99,7%).

Bei oben beschriebenem Vorgehen mussten zwar 11 von 45 Studienteilnehmern (24,4%) der PPA als „nicht klassifizierbar“ akzeptiert und für diesen Teil der Auswertung ausgeschlossen werden, für die verbleibenden 34 Studienteilnehmer ergaben sich jedoch ausschließlich Testkennwerte jenseits von 90%. Sowohl Kontrollgruppe A als auch Interventionsgruppe B‘ wiesen nur noch je eine falsch-positive bzw. falsch-negative Klassifikation durch die BVTs auf (vgl. Tabelle 30).

3.2.8 Korrelationen zwischen WMT/SFSS und BSI/SCIP

3.2.8.1 Punktwolke und Korrelationen BSI/WMT

Abbildung 10 zeigt den Zusammenhang zwischen BSI-Ergebnis (GSI) und WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS) in Form einer Punktwolke, aufgeschlüsselt nach den Gruppen A und B‘.

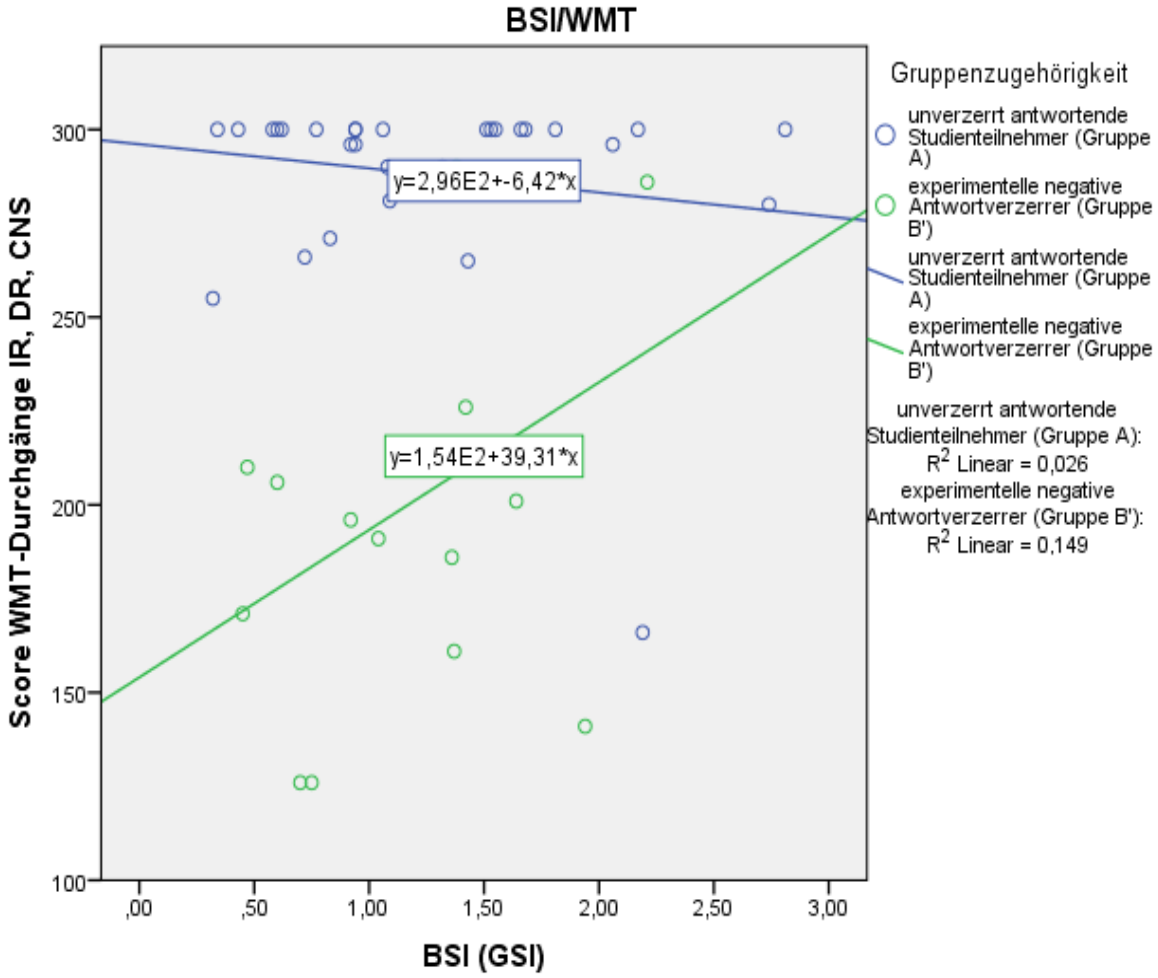

Abbildung 10: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B‘ (grün, 15 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1). Der Score WMT-Durchgänge IR, DR, CNS entstand durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte
BSI: Brief Symptom Inventory, GSI: Global Severity Index, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

In Abbildung 10 stellt sich für die Punktwolke von Gruppe B‘ (grün) – rein graphisch betrachtet – ein entfernt gleichsinniger Zusammenhang zwischen BSI-Ergebnis (aktuelle Symptomschwere) und WMT-Ergebnis dar (R^2 = 0,149). In Gruppe A (blau) hingegen scheint das WMT-Ergebnis graphisch betrachtet kaum vom BSI-Ergebnis beeinflusst zu werden. Die in blau eingezeichnete Regressionsgerade von Gruppe A erweckt zwar den Anschein eines leicht gegenläufigen Zusammenhangs zwischen den Merkmalen BSI-Ergebnis und WMT-Ergebnis, das R^2 in Gruppe A nimmt allerdings einen Wert nahe 0 an (R^2 = 0,026). Der Einfluss des BSI-Ergebnisses auf das WMT-Ergebnis in Gruppe A ist also als gering einzuschätzen.

In Tabelle 31 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 10 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe der PPA als auch für die Gruppen A und B‘ getrennt berechnet. Auch die Korrelationskoeffizienten zwischen BSI-Ergebnis (GSI) und „gesamtem“ WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) werden in Tabelle 31 zusätzlich zu den Korrelationskoeffizienten zwischen BSI und „einfachen“ WMT-Durchgängen (Score WMT-Durchgänge IR, DR, CNS) angegeben.

Tabelle 31: Auflistung der für den Zusammenhang zwischen BSI- und WMT-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 45 Studienteilnehmer der Per Protocol Analyse als auch für die beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) getrennt berechnet. Der Zusammenhang zwischen BSI und WMT wurde für den Score WMT-Durchgänge IR, DR, CNS sowie zusätzlich auch für den Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR („gesamter WMT“) berechnet. Beide WMT-Scores entstanden durch Addition der in den jeweiligen WMT-Durchgängen erzielten Prozentwerte
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MC: Multiple Choice, PA: Paired Associate Recall, FR: Free Recall, BSI: Brief Symptom Inventory, Spearman: Rangkorrelationskoeffizient nach Spearman, Pearson: Produkt-Moment-Korrelationskoeffizient nach Pearson, PPA: Per Protocol Analyse
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation BSI/WMT	r	n	p	95% KI
Spearman (Gesamtstichprobe PPA)	.07	45	.635	-.25 - .39

Fortsetzung von Tabelle 31 auf Seite 119

Fortsetzung Tabelle 31 von Seite 118

Korrelation BSI/WMT	r	n	p	95% KI
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gesamtstichprobe PPA)	.13	45	.405	-.21 - .43
Spearman (Gruppe A)	-.03	30	.874	-.42 - .39
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe A)	.06	30	.752	-.33 - .47
Spearman (Gruppe B‘)	.3	15	.277	-.34 - .75
Pearson (Gruppe B‘)	.39	15	.156	-.16 - .76
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe B‘)	.47	15	.078	-.06 - .84
Pearson für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe B‘)	.51	15	.052	.05 - .81

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

In Gruppe A ergab sich zwischen dem BSI-Ergebnis und dem Score der „einfachen“ WMT-Durchgänge sowie dem Score des „gesamten“ WMT ein sehr geringer gegenläufiger bzw. gleichsinniger linearer Zusammenhang nahe 0 (IR, DR, CNS: r = -.03. p = .874; IR, DR, CNS, MC, PA, FR: r = .06, p = .752). Ein hohes oder niedriges BSI-Ergebnis ging also in Gruppe A weder mit einem stark erhöhten noch mit einem stark erniedrigten WMT-Ergebnis einher. In Gruppe B‘ hingegen ergab sich ein geringer gleichsinniger linearer Zusammenhang zwischen BSI- und WMT-Ergebnis (IR, DR, CNS: r = .3, p = .277/Pearson: r = .39, p = .156;

IR, DR, CNS, MC, PA, FR: r = .47, p = .078/Pearson: r = .51, p = .052). Mit steigendem BSI-Ergebnis erhöhte sich also in Gruppe B‘ tendenziell auch das WMT-Ergebnis. Keine der in Tabelle 31 aufgeführten Korrelationen war statistisch signifikant.

3.2.8.2 Punktwolke und Korrelationen SCIP/WMT

Abbildung 11 stellt den Zusammenhang zwischen SCIP-Ergebnis (Summenscore) und WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS) – unterteilt nach den Gruppen A und B‘ – graphisch dar.

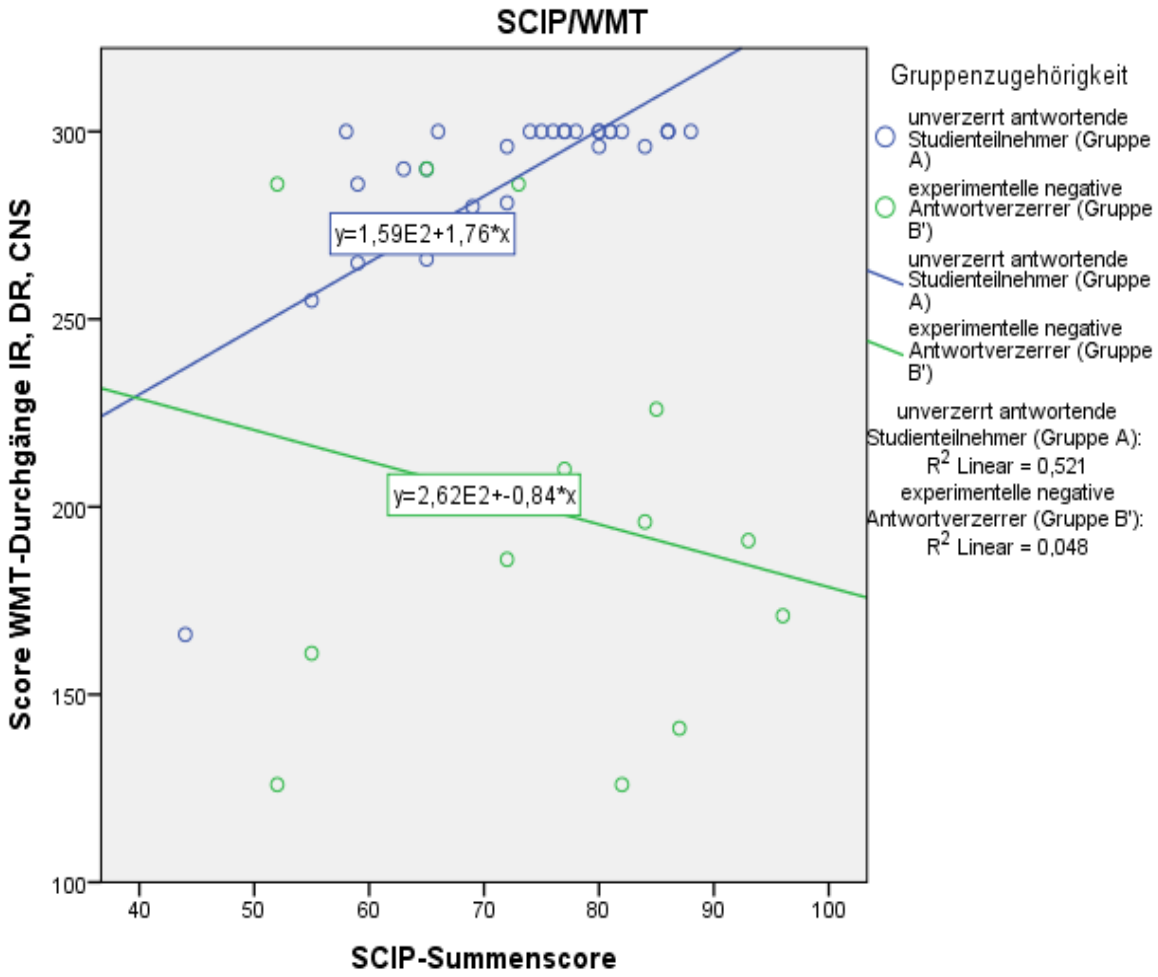

Abbildung 11: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B‘ (grün, 15 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1). Der Score WMT-Durchgänge IR, DR, CNS entstand durch Addition der in den WMT-Durchgängen IR, DR und CNS erzielten Prozentwerte
SCIP: Screen for Cognitive Impairment in Psychiatry, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Für Gruppe A (blau) zeigt sich in Abbildung 11 graphisch ein deutlicher gleichsinniger Zusammenhang zwischen SCIP-Summenscore (kognitives

Leistungsniveau) und WMT-Ergebnis (R^2 = 0,521). Die in grün eingezeichnete Regressionsgerade von Gruppe B‘ erweckt den Anschein eines gegenläufigen Zusammenhangs zwischen den Merkmalen SCIP-Ergebnis und WMT-Ergebnis in Gruppe B‘. Das mitangegebene R^2 nimmt in Gruppe B‘ jedoch einen Wert nahe 0 an (R^2 = 0,048) – was bedeutet, dass der Einfluss des SCIP-Ergebnisses auf das WMT-Ergebnis in Gruppe B‘ als gering einzuschätzen ist.

In Tabelle 32 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 11 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe der PPA als auch für die Gruppen A und B‘ getrennt berechnet. Auch die Korrelationskoeffizienten zwischen SCIP-Ergebnis (Summenscore) und „gesamtem“ WMT-Ergebnis (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) wurden zusätzlich zu den Koeffizienten zwischen SCIP und „einfachen“ WMT-Durchgängen (Score WMT-Durchgänge IR, DR, CNS) berechnet.

Tabelle 32: Auflistung der für den Zusammenhang zwischen SCIP- und WMT-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 45 Studienteilnehmer der Per Protocol Analyse als auch für die beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) getrennt berechnet. Der Zusammenhang zwischen SCIP und WMT wurde für den Score WMT-Durchgänge IR, DR, CNS sowie zusätzlich auch für den Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR („gesamter WMT“) berechnet. Beide WMT-Scores entstanden durch Addition der in den jeweiligen WMT-Durchgängen erzielten Prozentwerte
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, WMT: Word Memory Test, IR: Immediate Recognition, DR: Delayed Recognition, CNS: Consistency, MC: Multiple Choice, PA: Paired Associate Recall, FR: Free Recall, SCIP: Screen for Cognitive Impairment in Psychiatry, Spearman: Rangkorrelationskoeffizient nach Spearman, Pearson: Produkt-Moment-Korrelationskoeffizient nach Pearson, PPA: Per Protocol Analyse
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation SCIP/WMT	r	n	p	95% KI
Spearman (Gesamtstichprobe PPA)	.23	45	.131	-.13 - .56
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gesamtstichprobe PPA)	.24	45	.118	-.11 - .58

Fortsetzung von Tabelle 32 auf Seite 123

Fortsetzung Tabelle 32 von Seite 122

Korrelation SCIP/WMT	r	n	p	95% KI
Spearman (Gruppe A)	.68	30	< .001***	.41 - .87
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe A)	.66	30	< .001***	.4 - .83
Spearman (Gruppe B‘)	-.2	15	.482	-.82 - .52
Pearson (Gruppe B‘)	-.22	15	.434	-.76 - .38
Spearman für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe B‘)	-.3	15	.283	-.82 - .37
Pearson für Korrelation mit „gesamtem“ WMT (Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR) (Gruppe B‘)	-.41	15	.125	-.8 - .04

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Zwischen dem SCIP-Ergebnis und dem Score der „einfachen“ WMT-Durchgänge sowie dem Score des „gesamten“ WMT ergab sich in Gruppe A ein mittelstarker gleichsinniger linearer Zusammenhang (IR, DR, CNS: r = .68, p < .001; IR, DR, CNS, MC, PA, FR: r = .66, p < .001). Ein hohes SCIP-Ergebnis ging also in Gruppe A tendenziell auch mit einem hohen WMT-Ergebnis einher. In Gruppe B‘ hingegen ergab sich ein geringer gegenläufiger linearer Zusammenhang zwischen SCIP- und WMT-Ergebnis (IR, DR, CNS: r = -.2, p = .482/Pearson: r = -.22, p = .434; IR, DR, CNS, MC, PA, FR: r = -.3, p = .283/Pearson: r = -.41, p = .125). Mit steigendem SCIP-Ergebnis sank das WMT-Ergebnis in Gruppe B‘ also tendenziell.

3.2.8.3 Punktwolke und Korrelationen BSI/SFSS

Abbildung 12 zeigt den Zusammenhang zwischen BSI-Ergebnis (GSI) und SFSS-Ergebnis (Gesamtscore) in Form einer Punktwolke, aufgeschlüsselt nach den Gruppen A und B‘.

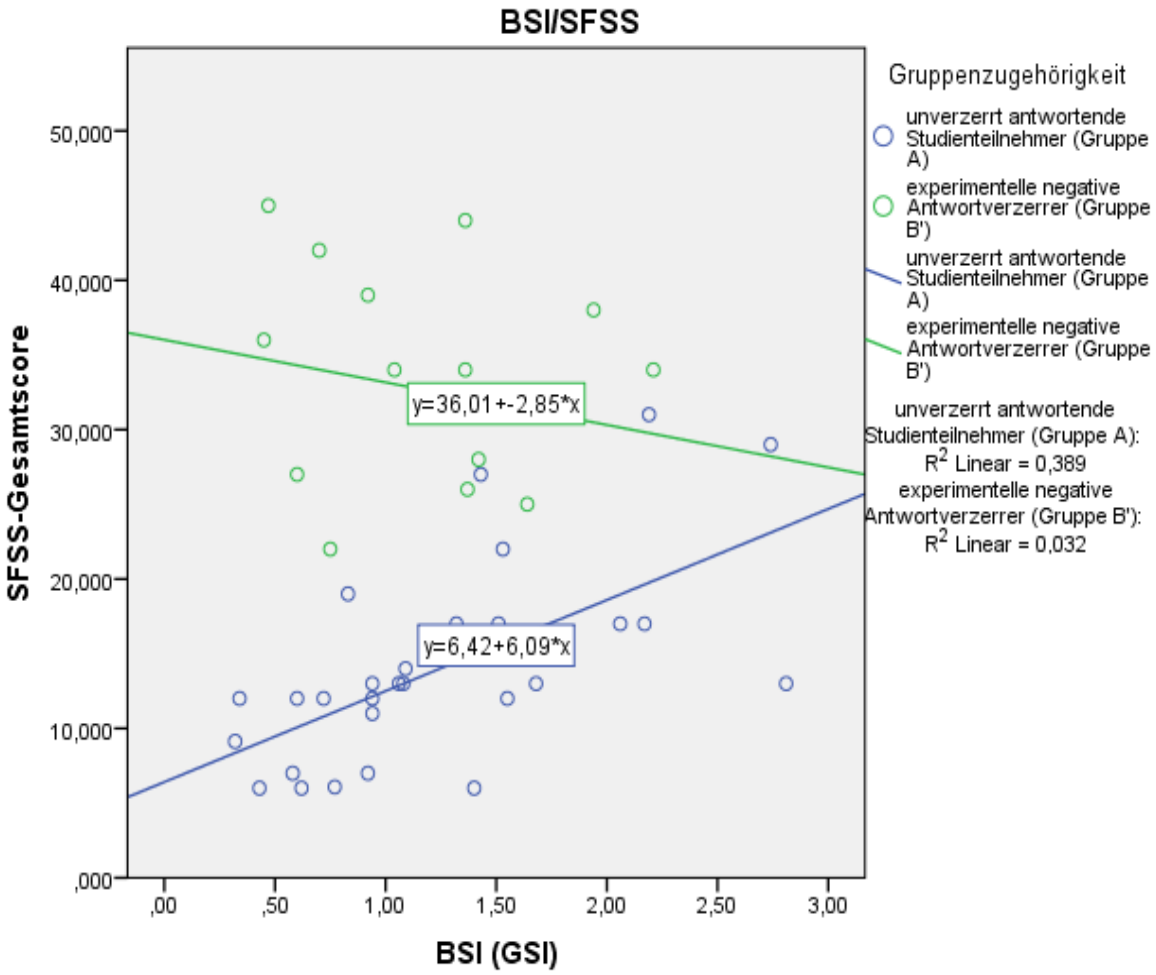

Abbildung 12: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B‘ (grün, 15 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1)
BSI: Brief Symptom Inventory, GSI: Global Severity Index, SFSS: Strukturierter Fragebogen Simulierter Symptome, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

In Abbildung 12 zeigt sich für Gruppe A (blau) graphisch ein deutlicher gleichsinniger Zusammenhang zwischen BSI-Ergebnis (aktuelle Symptomschwere) und SFSS-Ergebnis (R^2 = 0,389). Die in grün eingezeichnete

Regressionsgerade von Gruppe B‘ deutet auf einen leichten gegenläufigen Zusammenhang zwischen den Merkmalen BSI-Ergebnis und SFSS-Ergebnis in Gruppe B‘ hin. Das mitangegebene R^2 in Gruppe B‘ nimmt jedoch einen Wert nahe 0 an (R^2 = 0,032) – womit der Einfluss des BSI-Ergebnisses auf das SFSS-Ergebnis in Gruppe B‘ als gering einzuschätzen ist.

In Tabelle 33 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 12 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe der PPA als auch für die Gruppen A und B‘ getrennt berechnet.

Tabelle 33: Auflistung der für den Zusammenhang zwischen BSI- und SFSS-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 45 Studienteilnehmer der Per Protocol Analyse als auch für die beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) getrennt berechnet
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, SFSS: Strukturierter Fragebogen Simulierter Symptome, BSI: Brief Symptom Inventory, Spearman: Rangkorrelationskoeffizient nach Spearman, Pearson: Produkt-Moment-Korrelationskoeffizient nach Pearson, PPA: Per Protocol Analyse
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation BSI/SFSS	r	n	p	95% KI
Spearman (Gesamtstichprobe PPA)	.27	45	.074	-.05 - .57
Spearman (Gruppe A)	.68	30	< .001***	.45 - .84
Spearman (Gruppe B‘)	-.33	15	.230	-.73 - .14
Pearson (Gruppe B‘)	-.18	15	.525	-.59 - .22

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

In Gruppe A ergab sich zwischen dem BSI-Ergebnis und dem SFSS-Ergebnis bzw. –Gesamtscore ein mittelstarker gleichsinniger linearer Zusammenhang (r = .68, p < .001). Ein hohes BSI-Ergebnis ging also in Gruppe A tendenziell auch mit einem hohen SFSS-Ergebnis einher. In Gruppe B‘ hingegen ergab sich ein geringer gegenläufiger linearer Zusammenhang zwischen BSI-Ergebnis und SFSS-Ergebnis (r = -.33, p = .230/Pearson: r = -.18, p = .525). Mit steigendem BSI-Ergebnis sank das SFSS-Ergebnis bzw. der SFSS-Gesamtscore in Gruppe B‘ also tendenziell.

3.2.8.4 Punktwolke und Korrelationen SCIP/SFSS

Abbildung 13 stellt den Zusammenhang zwischen SCIP-Ergebnis (Summenscore) und SFSS-Ergebnis (Gesamtscore) – unterteilt nach den Gruppen A und B‘ – graphisch dar.

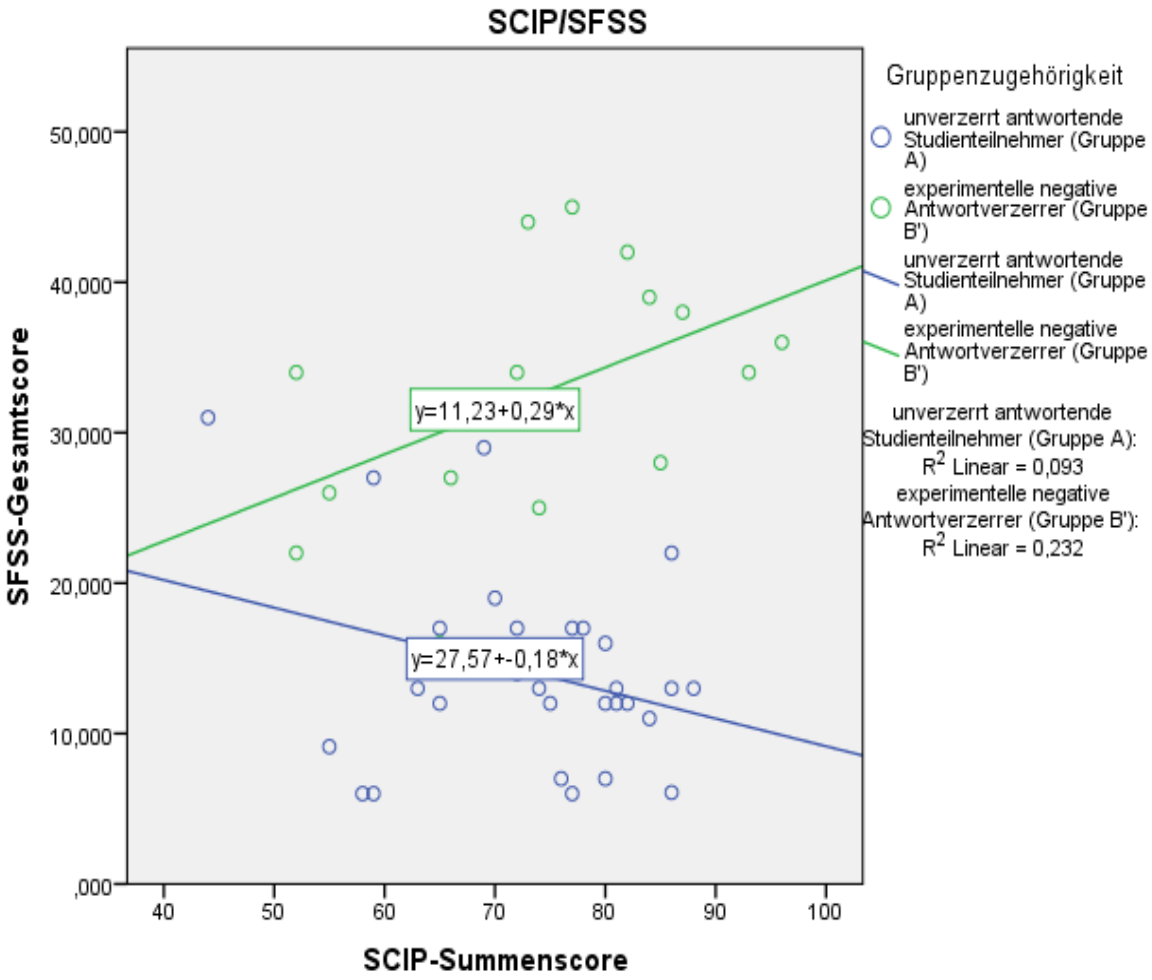

Abbildung 13: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke. Farbliche Unterteilung nach den beiden Studiengruppen A (blau, 30 Studienteilnehmer) und B‘ (grün, 15 Studienteilnehmer), Darstellung der Regressionsgeraden beider Gruppen und Auflistung des jeweiligen Bestimmtheitsmaßes R^2 (Maximalwert = 1)
SCIP: Screen for Cognitive Impairment in Psychiatry, SFSS: Strukturierter Fragebogen Simulierter Symptome, R^2: Bestimmtheitsmaß
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Graphisch zeigt sich in Abbildung 13 ein leicht gegenläufiger Zusammenhang zwischen SCIP-Ergebnis (kognitives Leistungsniveau) und SFSS-Ergebnis in Gruppe A (blau) – allerdings bei geringem R^2 (R^2 = 0,093). In Gruppe B‘ (grün) ist

ein tendenziell gleichsinniger Zusammenhang zwischen SCIP-Ergebnis und SFSS-Ergebnis zu erkennen (R^2 = 0,232).

In Tabelle 34 finden sich u.a. die Korrelationskoeffizienten zu den in Abbildung 13 graphisch dargestellten Zusammenhängen. Die Koeffizienten wurden sowohl für die Gesamtstichprobe der PPA als auch für die Gruppen A und B‘ getrennt berechnet.

Tabelle 34: Auflistung der für den Zusammenhang zwischen SCIP- und SFSS-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten. Die Koeffizienten wurden sowohl für die Gesamtstichprobe aller 45 Studienteilnehmer der Per Protocol Analyse als auch für die beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) getrennt berechnet
r: Korrelationskoeffizient, n: Stichprobenumfang, p: Signifikanz (2-seitig), 95% KI: 95%iges Konfidenzintervall, SFSS: Strukturierter Fragebogen Simulierter Symptome, SCIP: Screen for Cognitive Impairment in Psychiatry, Spearman: Rangkorrelationskoeffizient nach Spearman, Pearson: Produkt-Moment-Korrelationskoeffizient nach Pearson, PPA: Per Protocol Analyse
Aus: Klassifikationsgüte der Beschwerdevalidierungstests Word Memory Test und Strukturierter Fragebogen Simulierter Symptome in einer klinischen Stichprobe – Eine randomisierte kontrollierte Studie, Ravensburg 2015-16

Korrelation SCIP/SFSS	r	n	p	95% KI
Spearman (Gesamtstichprobe PPA)	.05	45	.755	-.25 - .32
Spearman (Gruppe A)	-.14	30	.464	-.56 - .31
Spearman (Gruppe B‘)	.5	15	.056	.1 - .75
Pearson (Gruppe B‘)	.48	15	.069	.12 - .77

** p ≤ .05, ** p ≤ .01, *** p ≤ .001*

Ein geringer gleichsinniger linearer Zusammenhang zwischen SCIP-Ergebnis und SFSS-Ergebnis (r = .5, p = .056/Pearson: r = .48, p = .069) ergab sich in Gruppe B‘. Ein hohes SCIP-Ergebnis ging in Gruppe B‘ also auch tendenziell mit einem hohen SFSS-Ergebnis bzw. SFSS-Gesamtscore einher. In Gruppe A ergab sich hingegen ein sehr geringer gegenläufiger linearer Zusammenhang (r = -.14, p = .464) zwischen SCIP-Ergebnis und SFSS-Ergebnis. Mit steigendem SCIP-Summenscore sank also der SFSS-Gesamtscore in Gruppe A tendenziell. Keine der in Tabelle 34 aufgeführten Korrelationen war statistisch signifikant.

3.2.9 Bericht über die Datenqualität (PPA)

Von allen 45 Studienteilnehmern der Per Protocol Analyse lagen vor Beginn der PPA-Auswertung alle Datenerhebungsbögen nahezu vollständig ausgefüllt vor. Die CRFs waren bei allen 45 Patienten vollständig ausgefüllt. Beim BSI fehlten bei 3 Patienten insgesamt 3 Werte, was aber wiederum (vgl. Kapitel 3.1.7) wegen geringer GSI-Veränderungen (Größenordnung 1/100) vernachlässigt wurde. Das SCIP lag von allen 45 Patienten lückenlos ausgefüllt vor. Beim WMT gab es ebenfalls keine fehlenden Werte. Vom SFSS lagen von 43 Patienten vollständig ausgefüllte Bögen vor, bei 2 Patienten fehlten insgesamt 4 Angaben. Im Falle unausgefüllter SFSS-Items wurde wiederum (vgl. Kapitel 3.1.7) die jeweilige durchschnittliche „SFSS-Fehlerquote“ (SFSS-Gesamtscore des jeweiligen Studienteilnehmers geteilt durch die Anzahl ausgefüllter Items) zum SFSS-Gesamtscore addiert. In 42 Fällen wurde die Studie strikt entsprechend des Studienprotokolls an einem Termin durchgeführt. In 3 Fällen wurde die Datenerhebung zweizeitig, mit jeweils eintägiger Pause (nach Punkt 4. des Studienprotokolls), durchgeführt. Bei keinem der 15 Studienteilnehmer in Gruppe B‘ bestand nunmehr eine schwere Protokollverletzung – zumindest laut den Angaben auf der VAS sowie laut den nach Versuchsdurchführung dokumentierten Patientenkommentaren (s. Anhang C). Auswertbar im Sinne der Per Protocol Analyse waren daher alle 45 Studienteilnehmer der Gesamtstichprobe der PPA.

3.3 Einzelfallbeschreibungen

KYF (Gruppe A): 55 Jahre alt, weiblich. Diagnosen: rezidivierende depressive Störung, gegenwärtig mittelgradige Episode (F33.1); Panikstörung (F41.0); histrionische Persönlichkeitsstörung (F60.4). Bereits berentet. Kein Schulabschluss. VAS: 100%

Die Studienteilnehmerin erzählte während der Versuchsdurchführung, sie habe ihre Kindheit in einem Heim verbracht, der Vater sei Alkoholiker gewesen, die Zustände zu Hause untragbar. Lesen und Schreiben habe sie erst als 30-Jährige richtig gelernt. Seit 26 Jahren sei sie psychisch krank. Zwischen den Tests brach die Studienteilnehmerin z.T. in Tränen aus, sobald sie sich an Situationen aus Kindheit und frühem Erwachsenenalter erinnerte. Am Schluss gab sie zu Protokoll, dass sie bei der Testdurchführung „alles gezeigt“ habe, mehr habe sie nicht leisten können.

WMT-Ergebnis: IR: 100%, DR: 100%, CNS: 100%, MC: 100%, PA: 90%, FR: 73%, Score WMT-Durchgänge IR, DR, CNS: 300; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 563. **WMT (Cut-off: ≤82,5%): bestanden**

SFSS-Ergebnis: N: 3, AF: 3, P: 2, NI: 3, AM: 2; Gesamt: 13. **SFSS (Cut-off: >16): bestanden**

Qhh (Gruppe A): 52 Jahre alt, männlich. Diagnosen: mittelgradige depressive Episode (F32.1); schlaffe Paraparese und Paraplegie: chronische inkomplette Querschnittslähmung (G82.03); Ataxie, nicht näher bezeichnet (R27.0). Bereits berentet. Hauptschulabschluss. VAS: 100%

Der Studienteilnehmer erzielte einen SFSS-Gesamtscore von 17 und lag damit über dem häufig verwendeten Cut-off von 16, obwohl er angab 100% seines tatsächlichen Leistungsvermögens gezeigt zu haben. Da er bereits Rente bezog und die Datenerfassung anonym erfolgte, ist es unwahrscheinlich, dass ein sekundärer Krankheitsgewinn – und damit ein Anreiz für negative Antwortverzerrung bzw. Simulation – im Raum gestanden haben könnte. 2001 wurde bei dem Studienteilnehmer eine Querschnittsmyelitis diagnostiziert. Entsprechend klagte er vor allem über neurologische Beschwerden und erzielte recht hohe Werte auf der N-Skala des SFSS (7 von 15 möglichen Punkten).

WMT-Ergebnis: IR: 95%, DR: 100%, CNS: 95%, MC: 95%, PA: 95%, FR: 73%,

Score WMT-Durchgänge IR, DR, CNS: 290; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 553. **WMT (Cut-off: ≤82,5%): bestanden**
SFSS-Ergebnis: N: 7, AF: 7, P: 0, NI: 2, AM: 1; Gesamt: 17. **SFSS (Cut-off: >16): nicht bestanden**

XHs (Gruppe A): 54 Jahre alt, weiblich. Diagnosen: schwere depressive Episode mit psychotischen Symptomen (F32.3, *bei Aufnahme*). Erwerbstätig. Hauptschulabschluss. VAS: 100%
Die Studienteilnehmerin gab beim PA-Durchgang (Paired Associate Recall) des WMT auf die Wortvorgabe „Messer“ durch den Testleiter die Antwort „Löffel“, anstatt der richtigen Antwort „Gabel“, welche eigentlich stärker mit dem Wort Messer assoziiert sein sollte. Ebenso antwortete die Studienteilnehmerin auf die Wortvorgabe „Feuer“ mit „Kerze“, anstatt der richtigen Antwort „Wasser“. Auch hier ist durch das Gegensatzpaar „Feuer-Wasser“ eigentlich von einer leicht zu merkenden Assoziation auszugehen. Da die Studienteilnehmerin bei der Testdurchführung laut Angabe auf der VAS 100% ihres tatsächlichen Leistungsvermögens zeigte, ist davon auszugehen, dass die oben geschilderten Assoziationen bei ihr tatsächlich auftraten.
WMT-Ergebnis: IR: 100%, DR: 95%, CNS: 95%, MC: 70%, PA: 70%, FR: 48%, Score WMT-Durchgänge IR, DR, CNS: 290; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 478. **WMT (Cut-off: ≤82,5%): bestanden**
SFSS-Ergebnis: N: 2, AF: 9, P: 0, NI: 2, AM: 0; Gesamt: 13. **SFSS (Cut-off: >16): bestanden**

TaJ (Gruppe A): 55 Jahre alt, weiblich. Diagnosen: rezidivierende depressive Störung, gegenwärtig schwere Episode ohne psychotische Symptome (F33.2, *bei Aufnahme*). Bereits berentet. Hauptschulabschluss. VAS: 100%
Die Studienteilnehmerin erzielte im SFSS einen Gesamtscore von 19, womit sie deutlich über dem gängigen Cut-off von 16 lag. Laut VAS und eigener Angaben beantwortete sie die Testitems aber ehrlich und zeigte ihre subjektiv bestmögliche Leistungsanstrengung. Die Studienteilnehmerin war bereits berentet, auch bei ihr dürfte also keinerlei Anreiz zu negativer Antwortverzerrung bzw. Simulation

bestanden haben.
WMT-Ergebnis: IR: 90%, DR: 88%, CNS: 93%, MC: 80%, PA: 75%, FR: 55%, Score WMT-Durchgänge IR, DR, CNS: 271; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 481. **WMT (Cut-off: ≤82,5%): bestanden**
SFSS-Ergebnis: N: 5, AF: 10, P: 1, NI: 3, AM: 0; Gesamt: 19. **SFSS (Cut-off: >16): nicht bestanden**

OvJ (Gruppe A): 62 Jahre alt, männlich. Diagnosen: rezidivierende depressive Störung, gegenwärtig schwere Episode ohne psychotische Symptome (F33.2, *bei Aufnahme*). Erwerbstätig. Hauptschulabschluss. VAS: 100%
Der Studienteilnehmer gab nach der Versuchsdurchführung an, so ehrlich wie möglich geantwortet zu haben, er habe keine Symptome übertrieben. Nächstes Jahr gehe er nach 45 Arbeitsjahren in Altersrente. Den WMT bestand er bei seinem üblichen Cut-off von ≤82,5% jedoch nicht. Eine Pflegekraft hatte vor der Versuchsdurchführung bereits berichtet, dass der Studienteilnehmer Probleme bei Gedächtnisaufgaben im Rahmen der Therapiegruppe gezeigt habe. Der Studienteilnehmer selbst war sich des Problems nicht wirklich bewusst. Auf der SFSS-Skala für amnestische Störungen erzielte er einen Wert von 0.
WMT-Ergebnis: IR: 90%, DR: 85%, CNS: 80%, MC: 40%, PA: 45%, FR: 28%, Score WMT-Durchgänge IR, DR, CNS: 255; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 368. **WMT (Cut-off: ≤82,5%): nicht bestanden**
SFSS-Ergebnis: N: 1, AF: 6, P: 1, NI: 1, AM: 0; Gesamt: 9. **SFSS (Cut-off: >16): bestanden**

WBi (Gruppe A): 45 Jahre alt, männlich. Diagnosen: Anpassungsstörungen (F43.2). Bereits berentet. Hauptschulabschluss. VAS: 100%
Der Studienteilnehmer gab im Anschluss an die Versuchsdurchführung an, er habe sich angestrengt alles richtig zu beantworten. Dies versicherte er mit Nachdruck und wirkte dabei glaubhaft. Er habe nach bestem Wissen und Gewissen geantwortet, mehr sei zurzeit nicht möglich. Beim WMT-IR- und DR-Durchgang erzielte er dennoch nur ein Ergebnis von 58% und lag damit nur knapp über dem statistischen Zufallswert von 50% richtigen Antworten. Auch der SFSS-

Gesamtscore war bei diesem Studienteilnehmer stark erhöht (31).
WMT-Ergebnis: IR: 58%, DR: 58%, CNS: 50%, MC: 20%, PA: 25%, FR: 20%, Score WMT-Durchgänge IR, DR, CNS: 166; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 231. **WMT (Cut-off: ≤82,5%): nicht bestanden**
SFSS-Ergebnis: N: 9, AF: 9, P: 5, NI: 4, AM: 4; Gesamt: 31. **SFSS (Cut-off: >16): nicht bestanden**

Pzv (Gruppe B): 46 Jahre alt, weiblich. Diagnosen: rezidivierende depressive Störung, gegenwärtig schwere Episode ohne psychotische Symptome (F33.2, *bei Aufnahme*). Erwerbstätig. Abitur. VAS: 65%
Die Studienteilnehmerin gab an, sie hätte „natürlich“ auch besser bei den Tests abschneiden können, habe aber versucht, Bereiche, die ihr tatsächlich schwer fielen, etwas zu übertreiben, um sie deutlich zu machen. Wenn sie sich bei einer Antwort nicht sicher gewesen sei, habe sie darüber auch nicht mehr richtig nachgedacht.
WMT-Ergebnis: IR: 78%, DR: 78%, CNS: 70%, MC: 60%, PA: 55%, FR: 30%, Score WMT-Durchgänge IR, DR, CNS: 226; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 371. **WMT (Cut-off: ≤82,5%): nicht bestanden**
SFSS-Ergebnis: N: 9, AF: 10, P: 0, NI: 1, AM: 8; Gesamt: 28. **SFSS (Cut-off: >16): nicht bestanden**

Wbi (Gruppe B): 51 Jahre alt, weiblich. Diagnosen: Anpassungsstörungen (F43.2). Erwerbstätig. Mittlere Reife. VAS: 36%
Die Studienteilnehmerin erzählte im Anschluss an die Versuchsdurchführung, sie habe sich bei den Tests „dümmer“ gestellt als sie sei und habe versucht, sich ihre Fehler im WMT zu merken, um diese wiederholen zu können. Beim SFSS, wie auch beim WMT, habe sie versucht, etwa 40% der Fragen ehrlich bzw. richtig zu beantworten, um nicht zu sehr aufzufallen. Die restlichen Fragen habe sie absichtlich falsch bzw. unaufrichtig beantwortet.
WMT-Ergebnis: IR: 25%, DR: 53%, CNS: 63%, MC: 60%, PA: 55%, FR: 50%, Score WMT-Durchgänge IR, DR, CNS: 141; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 306. **WMT (Cut-off: ≤82,5%): nicht bestanden**

SFSS-Ergebnis: N: 7, AF: 11, P: 8, NI: 5, AM: 7; Gesamt: 38. **SFSS (Cut-off: >16): nicht bestanden**

fRC (Gruppe B): 53 Jahre alt, männlich. Diagnosen: rezidivierende depressive Störung, gegenwärtig schwere Episode ohne psychotische Symptome (F33.2, *bei Aufnahme*). Erwerbslos. Mittlere Reife. VAS: 100%

Der Studienteilnehmer berichtete, er habe beim SFSS zwar versucht, die Instruktion umzusetzen, beim WMT sei aber „das Ego durchgekommen“. Da habe er einfach sehen wollen, ob das Kurzzeitgedächtnis noch funktioniert. Für die Per Protocol Analyse wurde der Studienteilnehmer daher aus Gruppe B ausgeschlossen.

WMT-Ergebnis: IR: 100%, DR: 100%, CNS: 100%, MC: 95%, PA: 95%, FR: 85%, Score WMT-Durchgänge IR, DR, CNS: 300; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 575. **WMT (Cut-off: ≤82,5%): bestanden**

SFSS-Ergebnis: N: 6, AF: 10, P: 2, NI: 3, AM: 1; Gesamt: 22. **SFSS (Cut-off: >16): nicht bestanden**

pTJ (Gruppe B): 58 Jahre alt, weiblich. Diagnosen: rezidivierende depressive Störung, gegenwärtig mittelgradige Episode (F33.1). Erwerbstätig. Abitur. VAS: 80%

Die Studienteilnehmerin gab an beim WMT hin und wieder absichtlich etwas falsch gesagt bzw. „vergessen“ zu haben, um ihre Beschwerden deutlich zu machen. Sie hätte beim WMT eigentlich mehr gewusst, als sie letztlich gezeigt habe. Beim Ausfüllen des SFSS habe sie versucht nicht zu stark zu übertreiben und ihre Beschwerden glaubhaft darzustellen, „so wie es halt gefordert war“ in der Instruktion.

WMT-Ergebnis: IR: 100%, DR: 95%, CNS: 95%, MC: 95%, PA: 90%, FR: 38%, Score WMT-Durchgänge IR, DR, CNS: 290; Score WMT-Durchgänge IR, DR, CNS, MC, PA, FR: 513. **WMT (Cut-off: ≤82,5%): bestanden**

SFSS-Ergebnis: N: 5, AF: 7, P: 0, NI: 1, AM: 3; Gesamt: 16. **SFSS (Cut-off: >16): bestanden**

4 Diskussion

4.1 Diskussion der Ergebnisse

In der vorliegenden Studie sollte die Klassifikationsgüte der deutschen Versionen von SIMS (dt. Version: SFSS) und WMT in einer klinischen Stichprobe von 60 Patienten untersucht werden. Dazu wurden je 30 Studienteilnehmer über den zufälligen Erhalt einer vorgefertigten Patienteninstruktion in eine Kontrollgruppe (Gruppe A, Studienteilnehmer mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten) bzw. in eine Interventionsgruppe (Gruppe B, experimentelle negative Antwortverzerrer) randomisiert. Anschließend wurden die Kennwerte Sensitivität, Spezifität, PPW und NPW für verschiedene Test-Cut-offs bestimmt. Die zunächst vorgenommene Auswertung der Daten der Gesamtstichprobe schloss in Gruppe B jedoch 15 Studienteilnehmer mit ein, deren fehlende Umsetzung von Patienteninstruktion B bei ihren Angaben auf der visuellen Analogskala und bei den nach Versuchsdurchführung protokollierten Aussagen in der Patientenbefragung (Anhang C) klar ersichtlich war. Um jedem Vorwurf unlauterer Beeinflussung der Ergebnisse zuvorzukommen, wurde beschlossen, diese 15 Patienten nicht von vornherein von der Auswertung auszuschließen und die Auswertung der Gesamtstichprobe im Ergebnisteil dieser Arbeit mit aufzuführen. Es sei jedoch darauf hingewiesen, dass der Schwerpunkt der Arbeit auf der konservativen Auswertung „im Sinne der Tests“ – also unter Ausschluss aller zweifelhafter Patienten – liegt, weshalb im Folgenden nur die Ergebnisse der Per Protocol Analyse (PPA) diskutiert werden.

In der PPA ergab sich für WMT und SFSS bei Anwendung der üblichen Cut-offs von ≤82,5% (WMT) bzw. >16 (SFSS) eine Sensitivität von 80% (WMT) bzw. 93,3% (SFSS) und eine Spezifität von 93,3% (WMT) bzw. 70% (SFSS). PPW und NPW nahmen für diese Cut-offs Werte von 85,7% (PPW WMT) und 90,3% (NPW WMT) bzw. 60,9% (PPW SFSS) und 95,5% (NPW SFSS) an.
Als Maß für die Übereinstimmung der Klassifikationen von SFSS und WMT wurden in der PPA ein Kappa-Koeffizient κ von .52 ($p < .001$) sowie ein Vierfelderkorrelationskoeffizient Φ von .56 ($p < .001$) berechnet.
Die kombinierte Anwendung von WMT und SFSS führte zu einer Sensitivität von 92,3% und einer Spezifität von 95,2% (für die Cut-offs von ≤82,5% bzw. >16).

PPW und NPW nahmen bei kombinierter Anwendung der Tests Werte von 92,3% (PPW) bzw. 95,2% (NPW) an. Allerdings mussten bei dieser Art der kombinierten Testdurchführung 11 von 45 Studienteilnehmern der PPA (24,4%) wegen differierender BVT-Klassifikationen als „nicht klassifizierbar“ akzeptiert und für die Berechnung der Kennwerte ausgeschlossen werden.

Die Korrelation zwischen dem Score der WMT-Durchgänge IR, DR und CNS und dem BSI nahm in der PPA einen Wert von r = -.03 (p = .874, Spearman) in Kontrollgruppe A an, die Korrelation zwischen dem Score der WMT-Durchgänge IR, DR und CNS und dem SCIP einen Wert von r = .68 (p < .001, Spearman, Kontrollgruppe A). Für die Korrelation zwischen SFSS und BSI ergab sich in Kontrollgruppe A ein Wert von r = .68 (p < .001, Spearman), während die Korrelation zwischen SFSS und SCIP in Kontrollgruppe A einen Wert von r = -.14 (p = .464, Spearman) annahm.

4.1.1 Sensitivität und Spezifität von WMT und SFSS

Bei der Bewertung der für die üblichen Cut-offs von ≤82,5% bzw. >16 berechneten Sensitivitäten und Spezifitäten von 80% (Sensitivität WMT) und 93,3% (Sensitivität SFSS) bzw. 93,3% (Spezifität WMT) und 70% (Spezifität SFSS) ist zunächst folgende Überlegung wichtig: Ein diagnostisches Zufallsverfahren, bei dem die Befunde z.B. ausgewürfelt werden (z.B. gerade Zahl = krank, ungerade Zahl = gesund), weist eine Sensitivität und Spezifität von 50% auf (Gaus u. Muche 2014, S. 140). Jedes ernstzunehmende diagnostische Verfahren sollte also bezüglich seiner Sensitivität und Spezifität deutlich über diesem Wert liegen (Gaus u. Muche 2014, S. 140). Daher lässt sich zunächst festhalten, dass WMT und SFSS in der vorliegenden Studie bei Verwendung der üblichen Cut-offs klar bessere Kennwerte als ein reines Zufallsverfahren erzielten, womit ihnen eine gewisse Trennschärfe für die Unterscheidung der Zustände „experimenteller negativer Antwortverzerrer“ (Studiengruppe B) vs. „Studienteilnehmer mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten“ (Studiengruppe A) attestiert werden kann. Dafür sprechen auch die hohen Werte der Flächen unter den ROC-Kurven (AUC), die als Maße der Güte eines untersuchten diagnostischen Verfahrens mit Werten von .95 (SFSS) beziehungsweise .93 (WMT) nahe dem Idealwert von 1 liegen.

Dennoch fallen die berechneten Testkennwerte im Vergleich mit den Werten der bisherigen Literatur tendenziell eher ab. So kommt beispielsweise eine SIMS-Studie mit 56 Studenten bei einem Cut-off von >16 auf eine Sensitivität von 93% und eine Spezifität von 100% (Jelicic et al. 2011). Cima et al. geben in ihrer Begleitstudie zur Erstpublikation der deutschen Version des SIMS (dt. Version: SFSS) mit 266 Studienteilnehmern (204 Studenten, 62 forensisch-psychiatrische Patienten) für einen Cut-off von >16 eine Gesamt-Sensitivität von 87% und eine Gesamt-Spezifität von 86% an (Cima et al. 2003). Allerdings fällt bei genauer Betrachtung der Ergebnisse auf, dass die mit 86% angegebene Spezifität des SFSS v.a. durch die hohen Werte (100%) in der Studentenkontrollgruppe zustande kommt, während der SFSS für das aufrichtig antwortende Patientenkollektiv nur eine Spezifität von 73% erzielte (Cima et al. 2003). Damit läge der Spezifitätswert der Patientenpopulation etwa im selben Bereich wie der in der vorliegenden Studie berechnete Wert von 70%. Eine der weiteren seltenen Patientenstudien zum SIMS (dt. Version: SFSS) gibt für eine Studienpopulation von 41 schizophrenen Patienten und einen Cut-off von >16 eine Gesamt-Spezifität von 71% - und damit ebenfalls in ähnlicher Größenordnung – an (Peters et al. 2013).

Vergleicht man die für den WMT berechneten Kennwerte mit den Angaben in der Literatur fallen auch hier zunächst hohe Werte für Sensitivität und Spezifität auf. So berichten z.B. Tan et al. über eine Studie mit 52 Studenten, bei welcher der WMT bei einem Cut-off von <82,5% eine Sensitivität von 100% und eine Spezifität von ebenfalls 100% erzielte (Tan et al. 2002). Auch Testentwickler Green berichtet Kennwerte in diesem Bereich (Sensitivität: 97,7%, Spezifität: 100%, bei einem Cut-off von ≤82,5% (Green et al. 2003)). Brockhaus und Merten geben in der Erstpublikation der deutschen WMT-Version für den ersten, an 127 gesunden Probanden durchgeführten Teil ihrer Studie bei einem Cut-off von ≤82,5% eine Sensitivität und Spezifität von jeweils 100% an (Brockhaus u. Merten 2004). Der zweite Teil ihrer Studie umfasste die Durchführung des WMT mit 32 geistig behinderten Patienten. Dabei ergab sich eine Spezifität von 97% (Brockhaus u. Merten 2004).

Insgesamt unterscheiden sich die oben genannten Werte meist deutlich von den 80% (WMT) bzw. 93,3% (SFSS) Sensitivität und 93,3% (WMT) bzw. 70% (SFSS) Spezifität, die WMT und SFSS in der vorliegenden Studie erzielten. Die Ursachen

hierfür dürften neben der Tatsache, dass sich Kontroll- und Interventionsgruppe der vorliegenden Studie aus einer reinen Patientenstichprobe zusammensetzten, auch in der Methodik begründet liegen und werden dementsprechend auch im Kapitel der Methodendiskussion analysiert.

Bezüglich der für Sensitivität und Spezifität berechneten Werte sollte angemerkt werden, dass es sich hierbei lediglich um statistische Schätzwerte handelt, die im Einzelfall keine sicheren Aussagen erlauben und mit denen daher entsprechend vorsichtig umgegangen werden sollte. Da es in Gutachtensituationen stets um die Beurteilung von Einzelfällen geht, sollte Klarheit darüber herrschen, was die in der vorliegenden Studie berechneten Sensitivitäten und Spezifitäten für eine fiktive Population von beispielsweise 100 Rentenbewerbern tatsächlich bedeuten: Eine WMT-Sensitivität von 80% hieße, dass 20 von 100 eigentlich antwortverzerrenden Rentenbewerbern mit dem WMT falsch-negativ als Testpersonen mit unauffälliger Leistungsanstrengung klassifiziert würden. Andererseits wäre die Feststellung suboptimaler Leistungsanstrengung bei einer WMT-Spezifität von 93,3% bei ca. 7 von 100 Rentenbewerbern mit subjektiv optimaler Leistungsanstrengung falsch-positiv (Cut-off ≤82,5%). 7 von 100 Rentenbewerbern würde also fälschlicherweise suboptimale Leistungsanstrengung unterstellt werden, obwohl sie den WMT wahrscheinlich eher aufgrund schwerer kognitiver Beeinträchtigungen nicht bestanden hätten, also besonders stark in ihrer Leistungsfähigkeit eingeschränkt wären und damit am ehesten der Leistungen der Sozialkassen bedürften. Falsch-positive Resultate sind deshalb als ethisch sehr bedenklich einzuordnen und legen eine grundsätzlich zurückhaltende Interpretation der Testergebnisse nahe.
Der SFSS würde bei einer Sensitivität von 93,3% ca. 7 von 100 negativen Antwortverzerrern falsch-negativ als unauffällig im Antwortverhalten klassifizieren. Eine Spezifität von 70% hieße, dass mit dem SFSS 30 von 100 aufrichtig antwortenden Rentenbewerbern fälschlicherweise als negative Antwortverzerrer klassifiziert werden würden – also fast ein Drittel der „aufrichtigen“ Rentenbewerber (Cut-off >16).

4.1.1.1 Falsch-positive Klassifikationen

Allgemein sollten vor allem die falsch-positiven Klassifikationen eines BVTs als besonders schwerwiegend angesehen werden, da diese vermutlich überdurchschnittlich stark eingeschränkte Patienten als negative Antwortverzerrer

darstellen und ihnen somit schwer benötigte finanzielle Leistungen verwehren. Im Themenbereich der Beschwerdevalidierung darf daher auf keinen Fall ein genaueres Eingehen auf mögliche falsch-positive Klassifikationen der untersuchten BVTs fehlen.

Larrabee zufolge sollte die Beurteilung der falsch-positiv-Rate eines BVTs an einer klinischen Stichprobe in einer Situation fern jedes sekundären Krankheitsgewinns erfolgen, um glaubwürdige und unverfälschte falsch-positive Klassifikationen erhalten zu können (Larrabee 2014). Diese Anforderungen erfüllt die vorliegende Studie weitestgehend, denn die Datenerhebung erfolgte anonym und wurde nicht von einem Gutachter oder fertig ausgebildeten Psychiater bzw. Therapeuten, sondern von einem Doktoranden durchgeführt, vor welchem sich die Patienten unbefangen präsentiert haben dürften. Zudem befand sich keiner der Studienteilnehmer in einem laufenden Rentenverfahren und den Studienteilnehmern wurde darüber hinaus unmissverständlich klar gemacht, dass ihre Testergebnisse lediglich zur statistischen Auswertung im Rahmen einer wissenschaftlichen Studie verwendet werden würden. Aus diesen Gründen sind externe Verstärker als Ursache für eine durch die BVTs festgestellte negative Antwortverzerrung in Kontrollgruppe A der vorliegenden Studie praktisch auszuschließen. Wenn Patienten der Kontrollgruppe den WMT und/oder SFSS bei den gewählten Cut-offs also tatsächlich nicht bestanden, sind die einzig logischen Folgerungen, dass die Cut-offs der Tests nicht hoch (SFSS) bzw. niedrig (WMT) genug gewählt wurden, um alle Patienten richtig zu klassifizieren und dass es sich um falsch-positive Klassifikationen der BVTs für diese Cut-offs handeln muss. 9 von 30 Patienten (30%) in Gruppe A bestanden den SFSS bei einem Cut-off von >16 nicht. 2 von 30 Patienten (6,7%) in Kontrollgruppe A bestanden den WMT bei einem Cut-off von ≤82,5% nicht. Sie sind somit als falsch-positive Klassifikationen der vorliegenden Studie anzusehen.

Vor allem in der WMT-Literatur wird das Auftreten falsch-positiver Klassifikationen bei Patienten mit Diagnosen aus dem depressiven Spektrum jedoch stark angezweifelt (Green et al. 2011). Stets wird darauf hingewiesen, dass selbst kognitiv schwerst-beeinträchtigte Patienten die einfachen WMT-Durchgänge (IR, DR, CNS) bestehen können – z.B. auch Patienten mit bilateraler hippokampaler Atrophie (Goodrich-Hunsaker u. Hopkins 2009) – weshalb bei depressiven

Patienten suboptimale Leistungsanstrengung als Ursache für das Abschneiden unterhalb des Cut-offs von ≤82,5% angenommen werden müsse (Green et al. 2011). Schließlich seien kognitive Beeinträchtigungen im Rahmen depressiver Erkrankungen nicht so schwerwiegend wie solche, die bei bilateraler hippokampaler Schädigung aufträten, so die Argumentation (Green et al. 2011). Auch wenn kognitive Beeinträchtigungen im Rahmen depressiver Erkrankungen in der Fachliteratur belegt (Veiel 1997) und wieder angezweifelt (Rohling et al. 2002) worden sind, ist bei den beiden vorliegenden falsch-positiven WMT-Klassifikationen dennoch davon auszugehen, dass situationsbedingt keinerlei sekundärer Krankheitsgewinn im Raum gestanden haben kann, sodass das Abschneiden unterhalb des WMT-Cut-offs am ehesten auf tatsächlich vorliegende kognitive Beeinträchtigungen zum Zeitpunkt der Testdurchführung zurückgeführt werden sollte. Bei der Beurteilung eines auffälligen WMT-Ergebnisses sollte daher im Einzelfall – zumindest für den WMT-Cut-off von ≤82,5% - immer die Möglichkeit falsch-positiver Klassifikationen bedacht werden.

Mit an Sicherheit grenzender Wahrscheinlichkeit kann man vorsätzliche suboptimale Leistungsanstrengung im WMT sowieso nur bei Unterschreitung der – im Rahmen des Alternativwahlverfahrens im IR-/DR-Durchgang mathematisch quasi unmöglich zu unterschreitenden – „Schwelle für Unter-Zufall-Antworten“ (Merten 2014, S. 32) unterstellen. Diese gilt als unterschritten, wenn ein Ergebnis von <37,5% im IR- oder DR-Durchgang des WMT erzielt wurde (Tan et al. 2002). Bei einem solchen Ergebnis gilt es als sehr wahrscheinlich, dass die Antworten absichtlich falsch ausgewählt wurden (Merten 2014, S. 32), schließlich bestehen im IR- und DR-Durchgang jeweils 50%ige Chancen das richtige Wort zu erraten, auch ohne die Wortliste vorher je gesehen zu haben. 3 von 15 (20%) Studienteilnehmern in Gruppe B‘ erzielten im IR- oder DR-Durchgang Ergebnisse unter 37,5% und waren damit eindeutig als negative Antwortverzerrer zu identifizieren. Kein Studienteilnehmer in Gruppe A kam auch nur annähernd in den Bereich eines solchen Ergebnisses, weshalb sich die Schwelle für Unter-Zufall-Antworten tatsächlich als sicheres Unterscheidungskriterium zwischen negativen Antwortverzerrern und Studienteilnehmern mit unverzerrtem Antwortverhalten eignen dürfte. Allerdings wurde sie auch nur in 3 von 45 Fällen der PPA unterschritten und würde somit helfen gerade einmal 6,7% der Studienteilnehmer

der PPA richtig zu klassifizieren. Auch ein IR-, DR- oder CNS-Ergebnis um 50% richtige Antworten sollte allerdings stutzig machen, denn es würde bedeuten, dass das Gedächtnis der Testperson kaum besser ist als der Zufall. 6 von 15 (40%) Studienteilnehmern in Gruppe B‘ erzielten ein IR-, DR- oder CNS-Ergebnis von ≥50% - ≤60%. Anzumerken ist, dass auch einer von 30 (3,3%) Studienteilnehmern in Kontrollgruppe A ein Ergebnis in diesem Bereich erzielte.

Bei Anwendung des häufig verwendeten SFSS-Cut-offs von >16 bestanden 9 von 30 Patienten (30%) in Kontrollgruppe A den SFSS nicht. Aufgrund der jeweiligen VAS-Angaben sowie aufgrund der jeweiligen Patientenkommentare nach Versuchsdurchführung (Anhang C) muss jedoch vermutet werden, dass die 9 betroffenen Patienten ihre Beschwerden bei der Testdurchführung unverzerrt dargestellt haben. Mitunter ursächlich für die 9 also höchstwahrscheinlich falsch-positiven SFSS-Klassifikationen dürfte ein Umstand sein, der bereits mehrfach in der SFSS- bzw. SIMS-Literatur kritisiert und angesprochen wurde. Neben Clegg et al. (Clegg et al. 2009) bemängelten nämlich auch van Impelen et al. in ihrer SIMS-Metaanalyse die Empfindlichkeit des SIMS für tatsächlich vorliegende Psychopathologie (van Impelen et al. 2014). Van Impelen et al. fiel diesbezüglich auf, dass nicht-klinische Kontrollgruppen im SIMS signifikant niedrigere Gesamtscores erzielten als klinische Kontrollgruppen (van Impelen et al. 2014). Der durchschnittliche SFSS-Gesamtscore der in der vorliegenden Studie aufgeführten klinischen Kontrollgruppe (Gruppe A) lag beispielsweise bei 14,1. Clegg et al. hingegen berichten für ihre nicht-klinische, unverzerrt antwortende Kontrollgruppe mit 30 Studienteilnehmern einen durchschnittlichen SIMS-Gesamtscore von 7,73, während ihre unverzerrt antwortende klinische Kontrollgruppe mit 35 Studienteilnehmern einen Durchschnittswert von 17,13 erzielte (Clegg et al. 2009). Die nicht-klinischen Kontrollgruppen von Jelicic et al. erzielten sogar nur durchschnittliche SIMS-Gesamtscores von 4,6 (29 Studienteilnehmer (Jelicic et al. 2011)) bzw. 5,0 (30 Studienteilnehmer (Jelicic et al. 2007)). Cima et al. berichten in ihrer Patientenkontrollgruppe übrigens ebenfalls einen vergleichsweise hohen mittleren SFSS-Gesamtscore von 14,13 – gegenüber einem Mittelwert von 8,09 in der nicht-klinischen Studentenkontrollgruppe (Cima et al. 2003). Anhand dieser vielfältigen Hinweise lässt sich also tatsächlich eine gewisse Empfindlichkeit des SFSS für „echte“

Psychopathologie vermuten. Auf diese Weise ließen sich auch die vergleichsweise niedrigen Spezifitäten des SFSS bzw. SIMS in klinischen Kontrollgruppen in den Studien von Cima et al. (73% (Cima et al. 2003)), Peters et al. (71% (Peters et al. 2013)) sowie in der vorliegenden Studie (70%) erklären.

Verantwortlich für die beschriebene Empfindlichkeit des SFSS für tatsächlich vorliegende Psychopathologie könnten u.a. einige strittige Testitems sein. Eigentlich sollte das SIMS laut Smith und Burger vor allem aus bizarren und unglaubwürdigen Beschwerden zusammengesetzt sein (Smith u. Burger 1997). Doch bereits van Impelen et al. kritisierten, dass das SIMS z.T. auch authentische psychiatrische Beschwerden benennt (van Impelen et al. 2014). Zu den von van Impelen et al. diesbezüglich bemängelten Items gehören v.a. die Items Nr. 32 („Ich kann schlecht schlafen." (Cima et al. 2003, S. 984)) und Nr. 52 („Es scheint, dass ich nicht mehr die Energie habe, die ich früher immer gehabt habe." (Cima et al. 2003, S. 984)) (van Impelen et al. 2014). Tatsächlich wurden diese beiden Items auch in der vorliegenden Studie von 66,7% (20 von 30 Studienteilnehmern, Item Nr. 32) bzw. 93,3% (28 von 30 Studienteilnehmern, Item Nr. 52) aller aufrichtig antwortenden Studienteilnehmer in Kontrollgruppe A bejaht – was laut SFSS-Decodierungstabelle als auffälliges Antwortverhalten zu werten war. Da z.B. Schlafstörungen typischerweise zum Symptomenkomplex einer Depression gehören, ist dieses Ergebnis wenig verwunderlich. Problematisch ist letztlich aber, dass Patienten durch das Bejahen solcher authentischer Items verstärkt Gefahr laufen den Cut-off zu überschreiten und falsch-positiv als negative Antwortverzerrer klassifiziert zu werden (van Impelen et al. 2014). Weitere Items, die von knapp zwei Drittel der Studienteilnehmer in Gruppe A bejaht wurden, waren die Items Nr. 2 (20 Mal bejaht, „Wenn meine Depression zu stark wird, mache ich lange Spaziergänge oder Sport, um die Spannung zu vermindern." (Cima et al. 2003, S. 983)), Nr. 19 (20 Mal bejaht, „Manchmal bin ich so deprimiert, dass ich mich nach meinem Bett sehne, um die schlechte Stimmung wegzuschlafen." (Cima et al. 2003, S. 983)) und Nr. 35 (19 Mal bejaht, „Manchmal sind meine Muskeln ohne ersichtlichen Grund so schlaff und kraftlos, dass sich meine Arme und Beine anfühlen, als ob sie zentnerschwer wären." (Cima et al. 2003, S. 984)). Da die genannten Items von einer solchen Mehrheit an Studienteilnehmern der aufrichtig antwortenden Kontrollgruppe A als zutreffend

angekreuzt wurden, ist davon auszugehen, dass auch sie zu stark mit tatsächlich auftretenden Symptomen übereinstimmen.

Vergleicht man die SCIP- und BSI-Ergebnisse der 9 mittels SFSS falsch-positiv klassifizierten Studienteilnehmer mit den Ergebnissen der 21 mittels SFSS richtig-negativ klassifizierten Studienteilnehmer in Gruppe A, fällt bei den falsch-positiv Klassifizierten ein im Mittel niedrigerer SCIP-Summenscore (Mittelwert SCIP-Summenscore: 68,9) sowie ein erhöhter mittlerer BSI-Wert (Mittelwert BSI (GSI): 1,8) auf (richtig-negativ klassifizierte Studienteilnehmer im Vergleich: Mittelwert SCIP-Summencore: 74,7; Mittelwert BSI (GSI): 1,1). Auch wenn es sich bei diesen Werten um bloße deskriptive Statistik handelt, sollte ersichtlich sein, dass gerade die falsch-positiv klassifizierten Studienteilnehmer kognitiv stärker beeinträchtigt (SCIP) und psychisch stärker belastet (BSI) waren als die richtig-negativ klassifizierten Studienteilnehmer.
Der Vergleich der mittels WMT falsch-positiv und richtig-negativ klassifizierten Studieneilnehmer in Gruppe A weist ähnliche Tendenzen auf. Die beiden mittels WMT falsch-positiv klassifizierten Studienteilnehmer erzielten mit einem mittleren SCIP-Summenscore von 49,5 ein niedrigeren Wert als die 28 richtig-negativ klassifizierten Studienteilnehmer in Kontrollgruppe A (Mittelwert SCIP-Summenscore: 74,6). Die mittleren BSI-Werte beider Untergruppen nahmen den gleichen Wert an (Mittelwert BSI (GSI) jeweils: 1,3). Auch wenn der mittlere SCIP-Summenscore der falsch-positiv Klassifizierten sich nur aus zwei Werten zusammensetzt und rein deskriptiver Natur ist, sollte wiederum ersichtlich sein, dass vor allem kognitiv überdurchschnittlich beeinträchtige Patienten durch den WMT falsch-positiv klassifiziert wurden. Angesichts der übereinstimmenden mittleren BSI-Ergebnisse in beiden Untergruppen scheint aber die Schwere der klinischen Symptomatik – anders als bei den falsch-positiven SFSS-Klassifikationen – für die falsch-positiven WMT-Klassifikationen unerheblich gewesen zu sein.
Für die Beschwerdevalidierungstests WMT und SFSS sind die oben genannten, im vorliegenden Fall rein deskriptiven Beobachtungen folgenschwer, denn sie bedeuten, dass die Tests gerade denjenigen Patienten finanzielle Leistungen durch eine Klassifikation als negative Antwortverzerrer verbauen können, die kognitiv und/oder symptomatisch besonders schwer beeinträchtigt sind. Bei der

Verwendung von WMT und SFSS mit den üblichen Cut-offs von ≤82,5% bzw. >16 in Gutachtensituationen müssen diese erheblichen Einschränkungen der Tests unbedingt berücksichtigt werden.

4.1.2 Positiver und negativer prädiktiver Wert

Positiver und negativer prädiktiver Wert des SFSS nahmen in der vorliegenden Studie für den Cut-off von >16 Werte von 60,9% (PPW SFSS) bzw. 95,5% (NPW SFSS) an. Damit unterscheiden sie sich teils deutlich von den Ergebnissen der deutschen SFSS-Erstpublikation. Cima et al. geben für ihre Gesamtstichprobe Werte von 87% für den PPW und 86% für den NPW an (Cima et al. 2003). Van Impelen et al. hingegen berechneten in ihrer Metaanalyse für eine zusammengefasste Teilstichprobe unverzerrt antwortender Patienten – bei einer Basalrate negativer Antwortverzerrung in der Gesamtstichprobe von 30% - einen PPW von 49% sowie einen NPW von 95% (van Impelen et al. 2014). In diesem Bereich liegen in etwa auch die in der vorliegenden Studie berechneten Werte – bei einer Basalrate negativer Antwortverzerrung in der Studienpopulation der PPA von 33,3% (15 von 45 Studienteilnehmern). Die Angabe der Basalrate negativer Antwortverzerrung bzw. von Simulation in der Studienpopulation ist bei der Beurteilung der Ergebnisse von PPW und NPW wichtig, da die Werte von PPW und NPW von der Höhe dieser Basalrate beeinflusst werden (van Impelen et al. 2014).

Ein PPW von 60,9% bedeutet, dass ein Studienteilnehmer mit einem positiven Testergebnis – also einem SFSS-Gesamtscore über dem Cut-off von 16 – „nur“ mit einer Wahrscheinlichkeit von 60,9% auch tatsächlich ein experimenteller negativer Antwortverzerrer ist, während ein NPW von 95,5% bedeutet, dass ein Studienteilnehmer mit negativen Testergebnis – also einem SFSS-Gesamtscore unter dem Cut-off – mit 95,5%iger Wahrscheinlichkeit auch tatsächlich ein unverzerrt antwortender Studienteilnehmer ist. Die für PPW und NPW berechneten Werte sprechen damit am ehesten für eine Verwendung des SFSS als Screening-Instrument (vgl. Merckelbach u. Smith 2003), da negative Antwortverzerrung aufgrund des relativ hohen NPW bei einem Testergebnis unter dem Cut-off von >16 mit hoher Wahrscheinlichkeit ausgeschlossen werden kann, während ein positives bzw. auffälliges Testergebnis aufgrund des relativ niedrigen PPW stets eine eingehendere Diagnostik nach sich ziehen sollte. Dies entspricht

übrigens in etwa dem Vorgehen, das auch die Testentwickler Smith und Burger für das SIMS ursprünglich vorsahen (Smith u. Burger 1997). Sie empfahlen, das SIMS als Teil einer Beschwerdevalidierungsprozedur zu verwenden, bei welcher nach einem auffälligen SIMS-Ergebnis weitere Tests und Untersuchungen angestrengt werden sollten, bevor Aggravation bzw. Simulation als Erklärung für die auffälligen Testergebnisse in Erwägung gezogen werden dürfe (Smith u. Burger 1997). Im Laufe der Jahre scheint diese Idee der Anwendung des SIMS bzw. SFSS jedoch von einem unkritischeren Einsatz des Tests abgelöst worden zu sein.

4.1.3 ROC-Kurven-Analyse

Weit verbreitet sind ROC-Kurven-Analysen in der SFSS- und WMT-Literatur nicht. Bei den meisten bisher durchgeführten Studien wurde der zu verwendende Cut-off vor Studienbeginn festgelegt und anschließend wurden dann, anhand unterschiedlicher Studienpopulationen, Sensitivität und Spezifität berechnet (z.B. Tan et al. 2002; Cima et al. 2003). Die ROC-Kurven-Analyse bietet jedoch die Möglichkeit Sensitivität und Spezifität für unterschiedliche Cut-offs zu berechnen, um somit den „besten“ Cut-off ausfindig machen zu können. Dennoch wird in der gesamten gesichteten WMT-Literatur keine einzige ROC-Kurven-Analyse erwähnt. Auch Testentwickler Green ging – soweit beurteilbar – nie genauer auf die ursprüngliche Berechnung seines ubiquitär empfohlenen Cut-offs von ≤82,5% ein. Eine ROC-Kurven-Analyse bei ähnlichem Studiendesign (u.a. experimentelle negative Antwortverzerrer vs. Kontrollgruppe(n)) fand sich für das SIMS nur in einer Studie von Edens et al. (Edens et al. 2007). Der für ihre Gesamtstichprobe berechnete AUC-Wert von .81 spricht jedoch für eine geringere Trennschärfe des SIMS zwischen negativen Antwortverzerrern und Kontrollen als es der in der vorliegenden Studie für den SFSS berechnete Wert von .95 vermuten lässt.

Vergleichswerte für die Youden-Indizes fanden sich nicht. Der höchste Youden-Index ergab sich für den WMT in der vorliegenden Studie bei einem Cut-off von <241 (für den Score WMT-Durchgänge IR, DR, CNS; Youden-Index: 0,767) und für den SFSS bei einem Cut-off von >20 (für den SFSS-Gesamtscore; Youden-Index: 0,8). Die mit diesen Cut-offs einhergehenden Sensitivitäten und Spezifitäten betrugen in der vorliegenden Patientenstichprobe 93,3% bzw. 86,7% (SFSS) und 80% bzw. 96,7% (WMT). Mit der Erwähnung dieser Cut-offs soll allerdings keine

unbedingte Empfehlung für einen der beiden Grenzwerte einhergehen. Anhand der ROC-Tabellen im Ergebnisteil (Kapitel 3.2.5) ist es jedem Anwender vielmehr selbst möglich, je nach Anforderungen an Sensitivität und Spezifität einen Cut-off zu wählen.

4.1.4 Übereinstimmung zwischen den BVT-Klassifikationen

Über die Berechnung des Viefelderkorrelationskoeffizienten Φ und des Kappa-Koeffizienten κ wurde die Übereinstimmung der Klassifikationen von WMT und SFSS auf einen statistisch messbaren Zusammenhang hin untersucht. Deskriptiv ließ sich zunächst festhalten, dass 75,5% (34 von 45 Studienteilnehmern) der Studienteilnehmer der PPA von beiden Beschwerdevalidierungstests übereinstimmend klassifiziert wurden (d.h. WMT und SFSS wurden jeweils bestanden bzw. nicht bestanden). 22,2% (10 von 45 Studienteilnehmern) der Studienteilnehmer bestanden zwar den WMT, nicht aber den SFSS. 2,2% (1 von 45 Studienteilnehmern) der Teilnehmer bestanden zwar den SFSS, nicht aber den WMT. Damit ergaben sich ein Vierfelderkorrelationskoeffizient (phi) von .56 und ein Kappa-Koeffizient von .52 (jeweils $p < .001$). Kappa-Werte im Bereich zwischen .41 und .6 sind laut Hedderich und Sachs als „deutliche Übereinstimmung“ (Hedderich u. Sachs 2016, S. 728) zu werten, sodass man für WMT und SFSS insgesamt eine ordentliche Übereinstimmung bezüglich ihrer Klassifikationen annehmen kann. Merten et al. berichten einen etwas geringeren aber dennoch ähnlichen Phi-Koeffizienten von .44 für den Zusammenhang der WMT- und SFSS-Klassifikationen in ihrer Studienpopulation (Merten et al. 2007). Stevens et al. geben einen Kappa-Koeffizienten von .4 für die Übereinstimmung der Klassifikationen von SIMS und WMT an (Stevens et al. 2008). Dass sich SIMS und WMT nicht stärker bezüglich ihrer Klassifikationen überschneiden, ist laut Stevens et al. zum Teil auf die unterschiedliche zugrundeliegende Methodik der Tests zurückzuführen (Stevens et al. 2008). Schließlich erfasst der WMT hauptsächlich suboptimale Leistungsanstrengung (Brockhaus u. Merten 2004), während das SIMS bzw. der SFSS als Selbstbeurteilungsverfahren v.a. die Aggravation bzw. Simulation von psychiatrischen und/oder kognitiven Beschwerden erkennen soll (Merckelbach u. Smith 2003). Eine Testperson, die Beschwerden aggraviert und den SFSS-Cut-off überschreitet, muss somit nicht unbedingt auch ein auffälliges WMT-Ergebnis erzielen. Die lediglich deutliche

Übereinstimmung (κ = .52) der Klassifikationen von WMT und SFSS ist in diesem Maße also sogar erwünscht. Wäre die Übereinstimmung höher bzw. sehr hoch, würde die kombinierte Anwendung beider Tests keinen Sinn machen. So aber bietet sich die Verwendung von WMT und SFSS in Kombination zur Erzielung verlässlicherer Klassifikationsergebnisse an.

4.1.5 Klassifikationen bei Verwendung beider BVTs

Eine vielversprechende Art der kombinierten Anwendung von WMT und SFSS ergibt sich, wenn folgendes Vorgehen festgelegt wird: 1. Beide BVTs müssen bestanden oder nicht-bestanden sein für eine Klassifikation. 2. Studienteilnehmer mit nicht übereinstimmenden Klassifikationen werden in beiden Studiengruppen wegen Unsicherheit ausgeschlossen. In Gruppe A mussten bei diesem Vorgehen 9 Studienteilnehmer für die Berechnung der Kennwerte ausgeschlossen werden, in Gruppe B‘ waren es 2 Studienteilnehmer. Somit mussten insgesamt 11 von 45 Studienteilnehmern (24,4%) der PPA als „nicht klassifizierbar“ akzeptiert werden. Für die verbliebene Studienpopulation (n = 34) ergab die auf diese Weise kombinierte Anwendung von WMT und SFSS jedoch eine Sensitivität von 92,3% und eine Spezifität von 95,2% (für die Cut-offs von ≤82,5% bzw. >16). Die für PPW und NPW berechneten Werte lagen ebenfalls bei 92,3% bzw. 95,2%. Die kombinierte Anwendung von SFSS und WMT führte also im Vergleich zum Einzelgebrauch der BVTs zu deutlich besseren Klassifikationsergebnissen und Kennwerten. Für das knappe Viertel (24,4%) an Studienteilnehmern der PPA, das wegen differierender BVT-Klassifikationen als „nicht klassifizierbar“ ausgeschlossen werden musste, würden sich eingehendere Untersuchungen – z.B. mit weiteren BVTs – anbieten bzw. in solchen Fällen sollte der klinischen Beurteilung, das heißt z.B. der Wertung von Konsistenz von Beschwerdeschilderung und Befunden oder des Verhaltens bei der körperlichen Untersuchung, der ihr tatsächlich gebührende Wert zugemessen werden.

Die Ergebnisse der kombinierten Anwendung von WMT und SFSS unterstützen die Empfehlung mehrerer Autoren, in Gutachtensituationen stets mehr als einen BVT einzusetzen (Gervais et al. 2004; Jelicic et al. 2011). Konkret wird dabei häufig die Praxis propagiert, negative Antwortverzerrung bzw. Simulation erst bei ≥2 nicht-bestandenen BVTs als wahrscheinlich anzusehen (Larrabee 2014). Giger und Merten empfehlen zusätzlich, dass die BVTs bei der kombinierten Anwendung

mehrerer Tests möglichst geringfügig miteinander korrelieren (Giger u. Merten 2013), schließlich sollen im Idealfall unterschiedliche Ausprägungen negativer Antwortverzerrung erfasst werden. Larrabee befürwortet daher die kombinierte Anwendung mehrerer kognitiver sowie psychologischer BVTs (Larrabee 2014).
Den vorangegangenen Empfehlungen wurde in der vorliegenden Studie mit der Durchführung sowohl eines kognitiven (WMT) als auch eines psychologischen BVTs (SFSS) entsprochen. Letztlich gelang es 94,1% (32 von 34 Studienteilnehmern) der „klassifizierbaren“ Studienteilnehmer der PPA dem richtigen Zustand (experimentelle negative Antwortverzerrer, Gruppe B‘ vs. Studienteilnehmer mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten, Gruppe A) zuzuordnen. Diese Ergebnisse unterstützen damit einerseits die Praxis der Anwendung von mehr als einem BVT in Gutachtensituationen, andererseits dürften sie gerade auch die Kombination von WMT und SFSS als besonders geeignet herausstellen.

4.1.6 Korrelationen zwischen WMT/SFSS und BSI/SCIP

Zur Erfassung der Art und Schwere der aktuellen klinischen Symptomatik wurde in der vorliegenden Studie das BSI verwendet. Das kognitive Leistungsniveau wurde mittels SCIP bestimmt. Im Rahmen der statistischen Auswertung wurden die Korrelationskoeffizienten für den Zusammenhang zwischen WMT-(Score WMT-Durchgänge IR, DR, CNS) und SFSS-(Gesamtscore)Ergebnissen einerseits und BSI-(GSI) bzw. SCIP-(Summenscore)Ergebnissen andererseits berechnet. Die für Interventionsgruppe B‘ berechneten Korrelationskoeffizienten dürften durch das Auftreten negativer Antwortverzerrung in dieser Gruppe stark verfälscht sein, weshalb im Folgenden nur auf die in Kontrollgruppe A berechneten Korrelationskoeffizienten eingegangen wird. Zwischen der Schwere der klinischen Symptomatik (BSI-Ergebnisse) und den WMT-Ergebnissen bestand in Gruppe A ein minimaler gegenläufiger Zusammenhang (Spearmans $r = -.03$, $p = .874$). Diese sehr geringe Korrelation mit dem BSI spricht allerdings klar für den WMT, denn sie bedeutet, dass das WMT-Ergebnis größtenteils unabhängig von der Art und Schwere der klinischen Symptomatik war – eine wichtige Eigenschaft für einen BVT. Für die Korrelation zwischen kognitivem Leistungsniveau (SCIP-Ergebnisse) und den WMT-Ergebnissen in Gruppe A ergab sich ein statistisch signifikanter, mittelstarker gleichsinniger Zusammenhang (Spearmans $r = .68$, $p <$

.001). Auch wenn es vielleicht wenig überrascht, dass die WMT-Ergebnisse mit den Ergebnissen eines kognitiven Tests korrelieren, weist dieser deutliche Zusammenhang auf die Gefahr hin, dass kognitiv schwer beeinträchtigte Patienten in den einfachen WMT-Durchgängen (IR, DR, CNS) schlecht abschneiden und evtl. auch ein Ergebnis unterhalb des Cut-offs erzielen können.
Sullivan et al. stellten ebenfalls statistisch signifikante, gleichsinnige Korrelationen zwischen den einfachen WMT-Durchgängen (IR, DR, CNS) und den Ergebnissen kognitiver Tests fest (r = bis zu .76 (Sullivan et al. 2007)). Zudem beschrieben sie signifikante gegenläufige Korrelationen zwischen den einfachen WMT-Durchgängen und den Ergebnissen ausgewählter Selbstbeurteilungsskalen (r = bis zu -.66 (Sullivan et al. 2007)). Kognitiv gut leistungsfähige und symptomatisch gering belastete Studienteilnehmer erzielten also laut ihrer Erhebung tendenziell bessere Ergebnisse in den einfachen WMT-Durchgängen. Diese durchaus kritisch zu betrachtenden Zusammenhänge konnten mit den vorliegenden Ergebnissen jedoch nur zum Teil bestätigt werden, da sich kein deutlicher gegenläufiger Zusammenhang zwischen den einfachen WMT-Durchgängen und dem Selbstbeurteilungsverfahren BSI feststellen ließ.

Für die Korrelation zwischen der Schwere der klinischen Symptomatik (BSI-Ergebnisse) und den SFSS-Gesamtscores ergab sich in Gruppe A ein statistisch signifikanter, mittelstarker gleichsinniger Zusammenhang (Spearmans r = .68, p < .001). Angesichts der bereits angesprochenen vermuteten Empfindlichkeit des SFSS für tatsächlich vorliegende Psychopathologie, verwundert dieser deutliche Zusammenhang nicht. Ganz im Gegenteil, er liefert weitere Argumente zur Unterstützung dieser These. In der SIMS- bzw. SFSS-Literatur berichten einige Autoren von ähnlich starken Zusammenhängen. So geben z.B. Edens et al. für die Korrelation zwischen einem von ihnen verwendeten Selbstbeurteilungsverfahren und dem SIMS-Gesamtscore einen mittelstarken gleichsinnigen Zusammenhang von r = .52 an (Edens et al. 1999). Van Impelen et al. berichten von Korrelationskoeffizienten zwischen r = .50 und r = .72 für den Zusammenhang des SIMS mit Selbstbeurteilungsverfahren, wie z.B. Becks Depressionsinventar (BDI (Beck et al. 1961)) (van Impelen et al. 2014).
Diese Werte sollten jedem SFSS-Anwender nochmals verdeutlichen, dass die Warnungen vor falsch-positiven Klassifikationen psychisch kranker Patienten

durch das SIMS bzw. den SFSS – zumindest für die häufig verwendete Cut-offs von >14/>16 – ernst zu nehmen sind.

Zwischen kognitivem Leistungsniveau (SCIP-Ergebnisse) und SFSS-Gesamtscores bestand in Gruppe A ein sehr geringer gegenläufiger Zusammenhang (Spearmans r = -.14, p = .464). Die gegenläufige Tendenz dieses Zusammenhangs ist kritisch zu betrachten, denn sie weist darauf hin, dass der SFSS kognitiv weniger leistungsfähige Patienten eventuell benachteiligen könnte. Angesichts der geringen Größe des Korrelationskoeffizienten soll es hier aber bei der Erwähnung einer möglichen ungünstigen Tendenz bleiben.

Zusammenfassend sei gesagt, dass zwischen den BVT-Ergebnissen und der Schwere der klinischen Symptomatik bzw. dem kognitiven Leistungsniveau – erfasst über BSI und SCIP – teils deutliche, statistisch signifikante Zusammenhänge bestanden, die bei der Beurteilung eines individuellen Klassifikationsergebnisses als mögliche Fehlerquellen berücksichtigt werden sollten.

4.1.7 Analyse möglicher Störgrößen und Bewertung der internen und externen Validität

Vor Beginn der Versuchsdurchführung wurden bei jedem teilnehmenden Patienten die Merkmale und möglichen Störgrößen Alter, Geschlecht, Aufnahmediagnosen, Bildungsabschluss, Ort der Rekrutierung (stationär/ambulant) und Beschäftigungsverhältnis erfasst. Statistische Analysen zeigten, dass sich die Gruppen A und B‘ nicht signifikant bezüglich der Verteilung dieser Störgrößen unterschieden, weshalb der Einfluss dieser Störgrößen auf die Zielgrößen (WMT- und SFSS-Ergebnisse) in beiden Gruppen als etwa gleich groß einzuschätzen ist. Nicht erfasst wurden die möglichen Störgrößen Erkrankungs- und Behandlungsdauer sowie Behandlungsart (medikamentös, psychotherapeutisch etc.) der Patienten. Diese Merkmale dürften zwar einen gewissen Einfluss auf die Schwere der klinischen Symptomatik und auf das kognitive Leistungsniveau der Patienten gehabt haben – und somit indirekt auch auf die Zielgrößen – zur Beantwortung der wissenschaftlichen Fragestellungen war ihre Erhebung jedoch nicht relevant. Abgesehen davon können sie im Rahmen der Randomisierung des RCTs als zwischen den Gruppen gleichverteilt angesehen werden (Gaus u. Muche 2014, S. 50). Aus Sicht von Gaus und Muche liefern randomisierte

kontrollierte Studien daher „die bestmögliche interne Validität“ (Gaus u. Muche 2014, S. 50). Für eine hohe interne Validität der vorliegenden Studie spricht der Umstand, dass sich die Gruppen A und B‘ nicht statistisch signifikant bezüglich der erfassten Störgrößen unterschieden, sodass von einer sehr guten Vergleichbarkeit bzw. Strukturgleichheit der Gruppen auszugehen ist. Durch die Anwendung eines standardisierten Studienprotokolls und angesichts der Tatsache, dass nur ein Testleiter alle Studienteilnehmer durch das Protokoll führte, kann zudem eine weitgehende Behandlungsgleichheit der Studienteilnehmer angenommen werden. Klar formulierte Ein- und Ausschlusskriterien stärken die interne Validität der vorliegenden Studie auf der einen Seite weiter (Gaus u. Muche 2014, S. 52), verdeutlichen andererseits aber auch die Grenzen der externen Validität, d.h. der Übertragbarkeit der Studienergebnisse auf andere Populationen (Gaus u. Muche 2014, S. 78). Verallgemeinerbar sind die erzielten Ergebnisse folglich nur für psychiatrische Patienten zwischen 40 und 63 Jahren, mit den in den Einschlusskriterien festgelegten Diagnosen sowie den weiteren in der Charakterisierung der Stichprobe beschriebenen demographischen Merkmalen. Bei der Bewertung der Repräsentativität bzw. externen Validität der vorliegenden Studienpopulation bzw. Studie fällt zudem auf, dass deutlich mehr Frauen (Gesamtstichprobe PPA: 30 Studienteilnehmerinnen, 66,7%) als Männer (Gesamtstichprobe PPA: 15 Studienteilnehmer, 33,3%) an der Datenerhebung teilnahmen. Dies ist zum einen auf die subjektive Beobachtung zurückzuführen, dass männliche Patienten eine Teilnahme häufiger ablehnten. Zum anderen ist die Lebenszeitprävalenz einer Diagnose aus dem depressiven Spektrum bei Frauen in Deutschland ca. doppelt so hoch wie bei Männern (Busch et al. 2013). Daher dürfte auch in der Grundgesamtheit – die alle Antragssteller auf eine Erwerbsminderungsrente in Deutschland umfasst – ein gewisses Geschlechterungleichgewicht bezüglich dieses Diagnosespektrums bestehen, womit die externe Validität der vorliegenden Studie eher noch gestärkt würde. Im Übrigen ist es unwahrscheinlich, dass der Überschuss an weiblichen Teilnehmern die BVT-Ergebnisse stark beeinflusst haben könnte, denn Rienstra et al. beispielsweise fanden keinen signifikanten Zusammenhang zwischen den Ergebnissen der WMT-Durchgänge und der Geschlechtszugehörigkeit (Rienstra et al. 2009). Für das SIMS konnten Giger und Merten ebenfalls keinen Einfluss der Geschlechtszugehörigkeit auf die SIMS-Ergebnisse feststellen (Giger u. Merten

2013). Dass die Datenerhebung an einer klinischen Stichprobe erfolgte, dürfte der externen Validität der Studie weiter zuträglich sein, da Anträge auf eine Erwerbsminderungsrente letztlich vor allem in dieser Population gestellt werden – d.h. von aktuellen oder zumindest ehemaligen Patienten psychiatrischer oder psychosomatischer Krankenhäuser.

4.2 Diskussion der Methodik

4.2.1 Stichprobenzusammensetzung/Ein- und Ausschlusskriterien

Um die Klassifikationsgüte von BVTs angemessen abschätzen zu können, ist es laut van Impelen et al. wichtig, Patientenstichproben mit unverzerrtem Antwortverhalten als Kontrollgruppen zu verwenden und nicht bloß Stichproben gesunder Probanden (van Impelen et al. 2014). Nur auf diese Weise dürfte sich die Eignung der Tests für Personen, die möglicherweise Beschwerden aggravieren oder verdeutlichen, aber dennoch krank sind – was in der Praxis wohl die häufigste Situation ist – richtig einschätzen lassen. Tatsächlich basiert die angeblich hohe Sensitivität und Spezifität von WMT und SFSS aber oft auf Studien an überwiegend nicht-klinischen Kontrollgruppen (z.B. Cima et al. 2003; Brockhaus u. Merten 2004).

Die Stichprobe der vorliegenden Studie wurde bezüglich Alter und Einschlussdiagnosen so gewählt, dass sie näherungsweise der Grundgesamtheit der Antragssteller auf eine Erwerbsminderungsrente ähneln dürfte. Aus den Statistiken der Deutschen Rentenversicherung lässt sich ableiten, dass die ICD-Diagnosen F30-F39 (affektive Störungen) im Jahr 2013 ca. 35% (Männer) bzw. 51% (Frauen) der gesamten F-Skala-Diagnosen bei den Rentenzugängen aufgrund verminderter Erwerbsfähigkeit ausmachten (Deutsche Rentenversicherung 2014a, S. 48ff). Die ICD-Diagnosen F40-F48 (neurotische, Belastungs- und somatoforme Störungen) stellten 2013 ca. 16% (Männer) bzw. 25% (Frauen) der F-Skala-Diagnosen (Deutsche Rentenversicherung 2014a, S. 49ff). Circa 84% (Männer) bzw. 86% (Frauen) der Rentenzugänge aufgrund verminderter Erwerbsfähigkeit mit F-Skala-Diagnosen gehörten zudem der Altersgruppe zwischen 40 und ≥ 60 Jahren an, bei einem Durchschnittsalter von 48,94 (Männer) bzw. 49,27 (Frauen) Jahren (Deutsche Rentenversicherung 2014a, S. 48ff). Die Einschlusskriterien der vorliegenden Studie (Aufnahmediagnose(n): F32.x, F33.x, F34.1, F40.x, F41.x, F43.2; Alter: 40 bis 63 Jahre) waren dementsprechend angepasst. Um die Stichprobe nicht zu heterogen zu gestalten, wurde auf den Einschluss weiterer Diagnosen verzichtet.

4.2.2 Studiendesign

Das Studiendesign des RCTs ist sicherlich als Stärke der vorliegenden Studie anzusehen, da sich Unterschiede der Zielgrößen in den Untersuchungsgruppen eines RCTs im Zuge der Randomisierung praktisch ausschließlich auf die variierte Einflussgröße zurückführen lassen (Gaus u. Muche 2014, S. 50). Durch die statistisch anzunehmende Gleichverteilung der Störgrößen auf die Untersuchungsgruppen im Rahmen der Randomisierung, erreicht man bei einem RCT eine hohe Ergebnisqualität (Bortz u. Schuster 2010, S. 8). Die CONSORT-Leitlinien empfehlen allerdings, darauf hinzuweisen, dass es sich bei einer Per Protocol Analyse – wie sie in der vorliegenden Studie letztlich auch durchgeführt werden musste – um einen nunmehr nicht-randomisierten Vergleich zwischen den Studiengruppen handelt, da der Ausschluss von bereits randomisierten Studienteilnehmern die positiven Effekte der Randomisierung beeinträchtigt und somit prinzipiell zu einem Bias der Ergebnisse führen kann (Moher et al. 2010). Da sich nach Ausschluss von 15 Studienteilnehmern für die PPA jedoch weiterhin keine statistisch signifikanten Unterschiede zwischen den Gruppen A und B‘ ergaben, sollte die oben beschriebene Einschränkung der Studienergebnisse nicht überbewertet werden.

Um die Verteilung der Störgrößen noch besser zu kontrollieren, hätte man die Stichprobe auch in unterschiedliche Subgruppen bzw. Schichten (z.B. nach Geschlecht, Altersklasse oder Diagnose) unterteilen können (= Stratifizierung) (Gaus u. Muche 2014, S. 232). Eine Stratifizierung der Stichprobe wäre allerdings bei einem Stichprobenumfang von 60 Patienten – bzw. 30 Patienten pro Gruppe – nicht sinnvoll gewesen, denn sie hätte zu kleinen Teilgruppen geführt, die statistisch wenig aussagekräftig gewesen wären. In der vorliegenden Studie wurde die Verteilung der Störgrößen daher dem Zufall (= Randomisierung) überlassen. Dennoch waren die Störgrößen letztlich relativ ausgeglichen verteilt.

Das verwendete Randomisierungsverfahren ist am ehesten als einfache Randomisierung zu bezeichnen (Moher et al. 2010). Die einfache Randomisierung ist – wie in den CONSORT-Leitlinien von einem solchen Verfahren gefordert – nicht vorhersagbar und völlig zufällig und anderen Randomisierungsverfahren in dieser Hinsicht sogar überlegen (Moher et al. 2010).

Zweifel daran, ob die wahren Verhältnisse einer Begutachtung mit dem vorliegenden Studiendesign nachgebildet werden konnten, sind zunächst einmal berechtigt. Von außen betrachtet wurden lediglich 30 Patienten in Gruppe B per Instruktion dazu aufgefordert, sich in einen an einer psychischen Erkrankung leidenden Rentenbewerber hineinzuversetzen, der einem Gutachter seine Einschränkungen und Symptome deutlich machen möchte. Ob ein Studienteilnehmer, der sich bloß in die Situation eines Rentenbewerbers hineinversetzt, genauso antwortet wie ein tatsächlicher Rentenbewerber, kann natürlich infrage gestellt werden. Allerdings ist genau dieser Ansatz weit verbreitet in der dringend notwendigen Forschung zu BVTs (z.B. Jelicic et al. 2011; Cima et al. 2003), denn er ist die wohl beste Annäherung an die Situation. Aus einer „echten“ Gutachtensituation kann schließlich gerade nicht der gewünschte Erkenntnisgewinn gezogen werden. Ob der Antragssteller Beschwerden aggraviert, verdeutlicht oder simuliert oder ob er seine Beschwerden unverzerrt darstellt, würde sich – der Situation geschuldet – nie abschließend klären lassen. Daher dürfte eine möglichst realitätsgetreue Nachbildung der Rentenbewerberpopulation, gepaart mit der Aufforderung an Patienten, sich so zu verhalten, als sei man in einer tatsächlichen Gutachtensituation, am ehesten verwertbare Ergebnisse liefern. Clegg et al. berichten immerhin, dass sich die SIMS-Ergebnisse experimenteller Aggravanten bzw. Simulanten und vermutlich aggravierender bzw. simulierender Antragssteller in ihrer Studie nicht signifikant unterschieden (Clegg et al. 2009), was darauf hinweist, dass ein Studiendesign mit experimentellen Aggravanten bzw. Simulanten durchaus brauchbare Ergebnisse liefern kann. Dennoch stellten auch sie klar heraus, dass die externe Validität von Studien mit experimentellen negativen Antwortverzerrern bzw. Simulanten in gewisser Weise stets fragwürdig sei (Clegg et al. 2009). Nies und Sweet empfahlen daher, bei Studien diesen Typs – zumindest von der Art der Zuwendung her – ähnliche Anreize wie in tatsächlichen Gutachtensituationen zu schaffen (Nies u. Sweet 1994), weshalb die Studienteilnehmer im vorliegenden Fall nach abgeschlossener Versuchsdurchführung mit einer monetären Aufwandsentschädigung in Höhe von 15 € entlohnt wurden.

Die in Kapitel 3.2.4 aufgeführten Werte für die Sensitivität von WMT und SFSS sollten angesichts der beschriebenen Einschränkungen der Methodik in der

Gruppe der experimentellen negativen Antwortverzerrer (Gruppe B‘) etwas zurückhaltender interpretiert werden, während die an Kontrollgruppe A berechneten Spezifitäten eine gute Annäherung an die „wahren“ Testspezifitäten in einer Patientenpopulation darstellen dürften.

4.2.3 Entblindung

Kritisch zu beurteilen ist die vorzeitige Entblindung der Studienteilnehmer. Während die Datenerhebung noch verblindet erfolgte, war die Gruppenzugehörigkeit der Studienteilnehmer bei der Übertragung sowie der Auswertung der Daten bereits bekannt. Zumindest die Übertragung der Studiendaten in die Microsoft® Excel® 2010-Tabelle hätte noch verblindet erfolgen können bzw. müssen. Anstatt die Briefumschläge mit den Patienteninstruktionen also direkt nach der Versuchsdurchführung einzusehen, hätte man sie auch noch bis nach der Berechnung der deskriptiven Statistik der Gesamtstichprobe unter Verschluss halten können. Die Berechnung von Sensitivität und Spezifität aber, mit der Gruppenzugehörigkeit als zu detektierender „Diagnose“, wäre so oder so nicht unter verblindeten Bedingungen möglich gewesen. Dennoch kann das vorzeitige Wissen über die Gruppenzugehörigkeiten bei der Auswertung durch eine willkürliche oder unwillkürliche Einflussnahme theoretisch zu einem Bias der Ergebnisse geführt haben (Moher et al. 2010).

4.2.4 Per Protocol Analyse

Der Ausschluss von 15 Studienteilnehmern aus Gruppe B für die Per Protocol Analyse darf natürlich nicht unkommentiert bleiben. Zunächst sei versichert, dass kein Studienteilnehmer willkürlich oder aus Gründen vorsätzlicher Ergebnisbeeinflussung ausgeschlossen wurde. Anhang C enthält die in der Patientenbefragung getätigten Aussagen aller Studienteilnehmer in Gruppe B, inklusive Kennzeichnung der für die PPA ausgeschlossenen Patienten. Unter Berücksichtigung dieser Aussagen in Zusammenschau mit den VAS-Angaben wurde die Entscheidung über den Ausschluss von Studienteilnehmern für die PPA getroffen. Auch wenn dieses Vorgehen insgesamt sicherlich als „im Sinne der Tests“ zu werten ist – da sich in der PPA deutlich bessere Testkennwerte ergaben – sollte man anerkennen, dass einige der letztlich aus Gruppe B

ausgeschlossenen Patienten offen zugaben, sich bei der Testdurchführung entgegen ihrer Instruktion bestmöglich angestrengt zu haben. Diese Patienten der Gruppe der experimentellen negativen Antwortverzerrer zuzuordnen, wäre schlichtweg falsch. Um unverfälschte Ergebnisse zu erhalten, war der Ausschluss der 15 Patienten also zwingend notwendig.

Die 15 Protokollverletzungen waren allesamt dadurch charakterisiert, dass Instruktion B aus unterschiedlichen Gründen nicht befolgt worden war – z.T. aus „Ehrlichkeit“ (z.B. Patient HSf, s. Anhang C), z.T. weil beispielsweise bei der Durchführung des WMT „das Ego“ durchkam (Patient fRC, s. Anhang C) und zum Teil auch weil vermutlich die Instruktion nicht richtig verstanden wurde (z.B. Patient oYE, s. Anhang C). Es handelte sich bei diesen Fällen also keineswegs um falsch-negative Klassifikationen der BVTs, die als möglicherweise unerwünschte Ergebnisse ausgeschlossen wurden, um die Sensitivität zu schönen.

4.2.5 Patienteninstruktionen

Die vielfachen Protokollverletzungen in Gruppe B sind zumindest teilweise auf die, im Vergleich zu anderen Studien (z.B. Jelicic et al. 2011), relativ „weich“ formulierte Patienteninstruktion B zurückzuführen. Patienteninstruktion B entstand aus der Annahme heraus, dass der Großteil der Rentenbewerber in einer Gutachtensituation wahrscheinlich keine Beschwerden simulieren oder hinzuerfinden wird, sondern am ehesten versucht, tatsächlich vorhandene Beschwerden deutlich zu machen, damit der Gutachter das Ausmaß der jeweiligen Beeinträchtigung nachvollziehen kann. Ein gewisses unglückliches Spannungsfeld ergibt sich in der Gutachtenpraxis nämlich dadurch, dass die Antragssteller bzw. Patienten natürlich einerseits finanzielle Leistungen zuerkannt bekommen wollen, im Rahmen der Begutachtung aber dazu angehalten sind, ihre bestmögliche Leistungsfähigkeit zu zeigen (z.B. bei der Durchführung des WMT). Mit der eher weichen Instruktion – die individuellen Interpretationsspielraum offen ließ – wurde versucht, diese sich im Laufe einer Begutachtung ergebenden Dynamiken möglichst realistisch nachzubilden.

Eventuell ist die weniger deutlich formulierte Aggravationsinstruktion ein Grund für die im Vergleich zu anderen Studien geringere Sensitivität von WMT und SFSS in der vorliegenden Studie. Dafür ist sie aber viel realitätsnäher als eine eindeutige

Simulationsaufforderung, die der klinischen gutachtlichen Situation in aller Regel nicht entspricht, aber zu unangemessen guten Testkennwerten führt.

Abschließend sei nochmals betont, dass es sich bei den in Kapitel 1.6 gestellten und im Verlauf von Kapitel 4.1 beantworteten wissenschaftlichen Fragestellungen streng genommen nicht um mithilfe statistischer Methoden bestätigte Hypothesen handelt. Im Zuge der in der vorliegenden Studie durchgeführten explorativen Datenanalyse ist die Signifikanz eines statistischen Tests als Generierung einer Hypothese zu werten und erlaubt keine Aussage hinsichtlich der Annahme oder Ablehnung von Hypothesen (vgl. konfirmatorisches Testen) (Gaus u. Muche 2014, S. 404). Dies muss bei der Einordnung der Ergebnisse unbedingt berücksichtigt werden.

4.3 Interpretation von WMT- und SFSS-Ergebnissen (Schlussfolgerung)

In einem Positionspapier der Deutschen Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN) zum Gebrauch von BVTs in der psychiatrischen Gutachtenpraxis heißt es: „Keinesfalls darf ein auffälliger Befund in einem BVT von vornherein mit Simulation gleichgesetzt werden." (Dressing et al. 2011, S. 389). Meistens wird daher dazu geraten, den Schluss auf Aggravation bzw. Simulation nicht vorschnell infolge eines einzelnen BVT-Ergebnisses zu ziehen, sondern eher infolge übereinstimmender Auffälligkeiten (Iverson 2006), die im Rahmen einer umfassenden Begutachtung verschiedener Informationsquellen festgestellt wurden, z.B. auch in ärztlichen Befunden oder im gutachtlichen Gespräch (Clegg et al. 2009). Oder wie es die DGPPN formulierte: „BVT können […] in begründeten Einzelfällen zusätzliche Informationen liefern, wenn die Ergebnisse in einer umfassenden psychiatrischen Gesamtbeurteilung gewürdigt werden. Sie können diese umfassende Gesamtbeurteilung aber in keinem Fall ersetzen." (Dressing et al. 2011, S. 389).

Die vorliegende Studie unterstützt diese Auffassung, da gezeigt werden konnte, dass die WMT- und SFSS-Klassifikationen einer Patientenpopulation bei Verwendung der üblichen Cut-offs (≤82,5% bzw. >16) im Einzelfall keine sicheren Aussagen erlauben – auch wenn die Trennschärfe der Tests zwischen den Zuständen experimentelle negative Antwortverzerrer (Gruppe B') vs. Studienteilnehmer mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten (Gruppe A) insgesamt zufriedenstellend war. WMT-Ergebnisse unter dem Cut-off von ≤82,5% traten jedoch auch bei optimaler Leistungsanstrengung auf und sogar knapp ein Drittel der aufrichtig antwortenden Patienten in Kontrollgruppe A erzielte ein SFSS-Ergebnis über dem Cut-off von >16.

Mit Sicherheit kann suboptimale Leistungsanstrengung im WMT eigentlich erst bei einem Ergebnis unterhalb der Schwelle für Unter-Zufall-Antworten (<37,5% im IR- oder DR-Durchgang) angenommen werden. Ergebnisse darüber sollten trotz einer insgesamt überzeugenden Spezifität von 93,3% (für den Cut-off ≤82,5%) vorsichtig und unter Berücksichtigung des Gesamtkontextes interpretiert werden, da sich u.a. eine deutliche Korrelation (r = .68) der WMT-Ergebnisse mit den

Ergebnissen eines kognitiven Tests (SCIP) zeigte. Die kognitive Leistungsfähigkeit scheint also Auswirkungen auf das WMT-Ergebnis zu haben, womit kognitiv stark beeinträchtigte Patienten evtl. durch den WMT benachteiligt werden könnten. Auf die Schwere der klinischen Symptomatik (in der vorliegenden Studie mittels BSI erfasst) scheint der WMT angesichts eines minimalen Korrelationskoeffizienten (r = -.03) allerdings weniger empfindlich zu reagieren als z.B. der SFSS. Generell sollte bei einem auffälligem WMT-Ergebnis höchstens geschlossen werden, dass die Testperson bei der Durchführung des WMT ihre tatsächliche Leistungsfähigkeit nicht gezeigt hat – und nicht, dass das Rentengesuch unberechtigt ist.

Der SFSS fiel auch in der vorliegenden Studie durch seine ausgeprägte Empfindlichkeit für tatsächlich vorliegende Psychopathologie auf. Deutlich wurde dies u.a. durch die hohe Korrelation (r = .68) des SFSS-Gesamtscores mit den Ergebnissen eines klinischen Symptomfragebogens (BSI). Der Einzelgebrauch des SFSS mit dem häufig verwendeten Cut-off von >16 scheint daher in einer Patientenpopulation nicht gerechtfertigt zu sein. Er führte in der vorliegenden Studie zu einer hohen Rate falsch-positiver Klassifikationen (30%). Dennoch bietet sich der SFSS als Screening-Verfahren an, solange dem Test bei auffälligen Befunden eingehendere Untersuchungen folgen.

Sowohl beim WMT als auch beim SFSS scheint tendenziell die Gefahr zu bestehen, dass kognitiv besonders stark beeinträchtigte bzw. psychisch besonders stark belastete Patienten die Cut-offs leichter über- (SFSS) bzw. unterschreiten (WMT). Auffällige WMT- und SFSS-Ergebnisse sollten daher unbedingt vorsichtig interpretiert werden. In Gutachtensituationen kann der Einzelgebrauch von WMT bzw. SFSS mit den bekannten Cut-offs (≤82,5% bzw. >16) letztlich nicht empfohlen werden. Den Erkenntnissen der vorliegenden Studie zufolge gelingt es jedoch, durch die kombinierte Anwendung von WMT und SFSS überzeugende Klassifikationsergebnisse zu erzielen (Sensitivität: 92,3%; Spezifität: 95,2%; PPW: 92,3%; NPW: 95,2% für die Cut-offs von ≤82,5% bzw. >16) – auch wenn bei dieser Art der Testanwendung etwa ein Viertel (24,4%) der Studienteilnehmer der PPA wegen divergierender BVT-Klassifikationen als „nicht klassifizierbar“ akzeptiert werden musste. Die kombinierte Anwendung von WMT und SFSS sollte dem Einzelgebrauch der Tests dennoch jederzeit vorgezogen werden.

5 Zusammenfassung

In der vorliegenden Studie sollte die Klassifikationsgüte der Beschwerdevalidierungstests (BVTs) Word Memory Test (WMT) und Strukturierter Fragebogen Simulierter Symptome (SFSS) in einer klinischen Stichprobe bestimmt werden. Dazu wurde zunächst untersucht, wie gut WMT und SFSS in der Lage sind, zwischen experimentellen negativen Antwortverzerrern und Studienteilnehmern mit subjektiv bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten zu unterscheiden. Weiterhin sollten die Übereinstimmung von WMT und SFSS bezüglich ihrer Klassifikationen sowie die Korrelation der BVTs mit der Ausprägung der klinischen Symptomatik bzw. dem kognitiven Leistungsniveau berechnet werden. Auch die Frage, ob die kombinierte Anwendung von WMT und SFSS zu besseren Klassifikationsergebnissen führen würde, sollte mithilfe der vorliegenden Studie beantwortet werden.

Die Studie wurde in Form einer randomisierten kontrollierten Studie (randomized controlled trial, RCT) an 60 ambulanten und stationären Patienten der Klinik für Psychiatrie und Psychotherapie I (Weissenau) des Universitätsklinikums Ulm durchgeführt. Die 60 Studienteilnehmer wurden über ein Zufallsverfahren in zwei Gruppen mit jeweils unterschiedlicher Patienteninstruktion randomisiert. Kontrollgruppe A erhielt die Instruktion, die BVTs WMT und SFSS mit bestmöglicher Leistungsanstrengung bzw. unverzerrtem Antwortverhalten anzugehen, während Interventionsgruppe B die Anweisung erhielt, sich in die Rolle eines Rentenbewerbers hineinzuversetzen, der seine Beschwerden einem Gutachter besonders deutlich machen möchte. Die Gruppenzugehörigkeit war dabei der durch die Tests zu detektierende Zustand. Wegen wiederholter schwerer Protokollverletzungen, im Sinne eines Nichtbeachtens von Patienteninstruktion B, mussten 15 Studienteilnehmer in Gruppe B für die Durchführung einer Per Protocol Analyse (PPA) ausgeschlossen werden. Die Ergebnisse der Auswertung der Gesamtstichprobe finden sich zwar ebenfalls aufgeführt, allerdings sind nur die Ergebnisse der Per Protocol Analyse als durch die schweren Protokollverletzungen unbelastet anzusehen.

Für die häufig verwendeten Cut-offs von >16 (SFSS) und ≤82,5% (WMT) ergaben sich Sensitivitäten und Spezifitäten von 93,3% bzw. 70% (SFSS) und 80% bzw.

93,3% (WMT). Positiver prädiktiver Wert (PPW) und negativer prädiktiver Wert (NPW) lagen bei diesen Cut-offs bei 60,9% bzw. 95,5% (SFSS) und bei 85,7% bzw. 90,3% (WMT). Den höchsten Youden-Index-Wert erhielt man für den SFSS-Gesamtscore bei einem Cut-off von >20 und für den Gesamtscore der WMT-Durchgänge Immediate Recognition, Delayed Recognition und Consistency bei einem Cut-off von <241. Als Maß für die Übereinstimmung der WMT- und SFSS-Klassifikationen wurden ein Vierfelderkorrelationskoeffizient (phi) von .56 und ein Kappa-Koeffizient von .52 berechnet. Die kombinierte Anwendung der BVTs führte zu einer Sensitivität und Spezifität von 92,3% bzw. 95,2% (für die Cut-offs von ≤82,5% bzw. >16). PPW und NPW lagen bei kombinierter Testanwendung ebenfalls bei 92,3% bzw. 95,2%. Allerdings musste bei dieser Art der Auswertung etwa ein Viertel (24,4%) der Studienteilnehmer der PPA aufgrund divergierender BVT-Klassifikationen als „nicht klassifizierbar“ akzeptiert werden. Statistisch signifikante Korrelationen konnten zwischen Brief Symptom Inventory (BSI) und SFSS (Korrelationskoeffizient $r = .68$, $p < .001$) sowie zwischen dem Screen for Cognitive Impairment in Psychiatry (SCIP) und dem WMT (Korrelationskoeffizient $r = .68$, $p < .001$) festgestellt werden.

Die in der vorliegenden Studie berechneten Testkennwerte (Sensitivität, Spezifität etc.) unterschieden sich teils deutlich von den in der bisherigen Literatur angegebenen Werten. Oft bestanden die jeweiligen Studienpopulationen allerdings auch mehrheitlich aus Studenten oder gesunden Probanden und es wurden starke Simulationsinstruktionen gegeben, die nicht die gutachtliche Realität wiederspiegeln dürften. Einige wiederholt bemängelte Schwachstellen des SFSS traten auch in der vorliegenden Studie deutlich hervor.

Festgehalten werden muss letztlich, dass bei unkritischer Verwendung von WMT und SFSS – vor allem bei den Cut-offs >16 (SFSS) und ≤82,5% (WMT) – die Gefahr gutachtlicher Fehlbeurteilungen auf Kosten tatsächlich kranker Menschen besteht (falsch-positive Klassifikationen). Zur umsichtigen Beurteilung der Testergebnisse von WMT und SFSS sollte daher stets eine ausführliche Einordnung des Einzelergebnisses unter Berücksichtigung des Gesamtkontextes erfolgen. Aufgrund der in der vorliegenden Studie gewonnenen Erkenntnisse ist für die Gutachtenpraxis insbesondere die kombinierte Anwendung von WMT und SFSS zu empfehlen.

6 Literaturverzeichnis

1. Bauer L, McCaffrey RJ: Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and Word Memory Test on the Internet: Is test security threatened? Arch Clin Neuropsychol 21: 121–126 (2006)
2. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J: An Inventory for Measuring Depression. Arch Gen Psychiatry 4: 561–571 (1961)
3. Bellary S, Krishnankutty B, Latha MS: Basics of case report form designing in clinical research. Perspect Clin Res 5: 159–166 (2014)
4. Benge JF, Wisdom NM, Collin RL, Franks R, LeMaire A, Chen DK: Diagnostic utility of the Structured Inventory of Malingered Symptomatology for identifying psychogenic non-epileptic events. Epilepsy Behav 24: 439–444 (2012)
5. Blaskewitz N, Merten T: Diagnostik der Beschwerdenvalidität. Fortschr Neurol Psychiatr 75: 140–154 (2007)
6. Bortz J, Schuster C: Statistik für Human- und Sozialwissenschaftler, 7. Aufl, Springer, Berlin Heidelberg New York, S. 8–122 (2010)
7. Boulet J, Boss MW: Reliability and Validity of the Brief Symptom Inventory. Psychol Assess 3: 433–437 (1991)
8. Brockhaus R, Merten T: Neuropsychologische Diagnostik suboptimalen Leistungsverhaltens mit dem Word Memory Test. Nervenarzt 75: 882–887 (2004)
9. Busch MA, Maske UE, Ryl L, Schlack R, Hapke U: Prävalenz von depressiver Symptomatik und diagnostizierter Depression bei Erwachsenen in Deutschland. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 56: 733–739 (2013)
10. Bush SS, Ruff RM, Tröster AI, Barth JT, Koffler SP, Pliskin NH, Reynolds CR, Silver CH: Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. Arch Clin Neuropsychol 20: 419–426 (2005)
11. Carone DA: Young Child With Severe Brain Volume Loss Easily Passes the Word Memory Test and Medical Symptom Validity Test: Implications for Mild TBI. Clin Neuropsychol 28: 146–162 (2014)
12. Cima M, Hollnack S, Kremer K, Knauer E, Schellbach-Matties R, Klein B, Merckelbach H: „Strukturierter Fragebogen Simulierter Symptome". Nervenarzt 74: 977–986 (2003)
13. Clegg C, Fremouw W, Mogge N: Utility of the Structured Inventory of Malingered Symptomatology (SIMS) and the Assessment of Depression

Inventory (ADI) in screening for malingering among outpatients seeking to claim disability. J Forens Psychiatry Psychol 20: 239–254 (2009)

14. Cohen J: A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20: 37–46 (1960)

15. Cronbach LJ: Coefficient alpha and the internal structure of tests. Psychometrika 16: 297–334 (1951)

16. Cuesta MJ, Pino O, Guilera G, Rojo JE, Gómez-Benito J, Purdon SE, Franco M, Martínez-Arán A, Segarra N, Tabarés-Seisdedos R, Vieta E, Bernardo M, Crespo-Facorro B, Mesa F, Rejas J: Brief cognitive assessment instruments in schizophrenia and bipolar patients, and healthy control subjects: A comparison study between the Brief Cognitive Assessment Tool for Schizophrenia (B-CATS) and the Screen for Cognitive Impairment in Psychiatry (SCIP). Schizophr Res 130: 137–142 (2011)

17. Czekaj A, Steinert T, Uhlmann C: Klinische Routinediagnostik kognitiver Beeinträchtigungen. Psychiatr Prax 41: 166–167 (2014)

18. Das Sechste Buch Sozialgesetzbuch – Gesetzliche Rentenversicherung – in der Fassung der Bekanntmachung vom 19. Februar 2002 (BGBl. I S. 754, 1404, 3384), das zuletzt durch Artikel 7 des Gesetzes vom 21. Dezember 2015 (BGBl. I S. 2517) geändert worden ist. Fundstelle: http://www.sozialgesetzbuch-sgb.de/sgbvi/43.html (30.09.2016)

19. Derogatis LR, Melisaratos N: The Brief Symptom Inventory: an introductory report. Psychol Med 13: 595–605 (1983)

20. Deutsche Rentenversicherung: Statistik der Deutschen Rentenversicherung: Rentenzugang 2013, einschließlich Rentenwegfall, Rentenänderung/Änderung des Teilrentenanteils, Rentenarten, Rentenhöhe, zeitlicher Umfang der angerechneten Zeiten, Erwerbsminderungsdiagnosen, Zugangsalter, Wegfallsalter, Rentenbezugsdauer, Bd 198, Deutsche Rentenversicherung Bund, Berlin, S. 48–59 (2014a)

21. Deutsche Rentenversicherung: Statistische Analysen Bereich 0760: Indikatoren zu Erwerbsminderungsrenten (EM-Renten) im Zeitablauf, Stand: Mai 2014, Rentenanträge und ihre Erledigungen, Rentenarten und -höhen (Zugang, Bestand), Zugangsalter, Abschläge und Diagnosen, weitere ausgewählte Indikatoren und Ausgaben. Deutsche Rentenversicherung Bund, Berlin (2014b)

22. Deutsche Rentenversicherung: Statistik der Deutschen Rentenversicherung: Rentenversicherung in Zahlen 2015, aktuelle Ergebnisse, Stand: 12. Juni 2015, Entwicklung der Daten bis heute, Faustdaten, Werte, Finanzen, Versicherte, Rentenbestand, Rentenzugang, Rehabilitation. Deutsche Rentenversicherung Bund, Berlin, S. 57–69 (2015)

23. Deutsche Rentenversicherung: http://www.deutsche-rentenversicherung.de/Allgemein/de/Navigation/2_Rente_Reha/01_Rente/01_a

llgemeines/03_rentenarten_und_leistungen/08_erwerbsminderungsrente_node.html (30.09.2016)

24. Dohrenbusch R: Symptom- und Beschwerdevalidierung chronifizierter Schmerzen in sozialmedizinischer Begutachtung. Teil I: Terminologische und methodologische Zugänge. Schmerz 23: 231–240 (2009)
25. Dressing H, Foerster K, Widder B, Schneider F, Falkai P: Zur Anwendung von Beschwerdenvalidierungstests in der psychiatrischen Begutachtung. Stellungnahme der Deutschen Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN). Nervenarzt 82: 388–390 (2011)
26. Edens JF, Otto RK, Dwyer T: Utility of the Structured Inventory of Malingered Symptomatology in Identifying Persons Motivated to Malinger Psychopathology. J Am Acad Psychiatry Law 27: 387–396 (1999)
27. Edens JF, Poythress NG, Watkins-Clay MM: Detection of Malingering in Psychiatric Unit and General Population Prison Inmates: A Comparison of the PAI, SIMS, and SIRS. J Pers Assess 88: 33–42 (2007)
28. Franke GH: BSI: Brief Symptom Inventory von L.R. Derogatis (Kurzform der SCL-90-R), deutsche Version, Manual. Beltz Test, Göttingen, S. 1–21 (2000)
29. Gaus W, Muche R: Medizinische Statistik. Angewandte Biometrie für Ärzte und Gesundheitsberufe. Schattauer, Stuttgart, S. 50–404 (2014)
30. Gervais RO, Green P, Allen III LM, Iverson GL: Effects of Coaching on Symptom Validity Testing in Chronic Pain Patients Presenting for Disability Assessments. Journal of Forensic Neuropsychology 2: 1–19 (2001)
31. Gervais RO, Rohling ML, Green P, Ford W: A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. Arch Clin Neuropsychol 19: 475–487 (2004)
32. Giger P, Merten T: Swiss population-based reference data for six symptom validity tests. Clínica y Salud 24: 153–159 (2013)
33. Goodrich-Hunsaker NJ, Hopkins RO: Word Memory Test Performance in Amnesic Patients With Hippocampal Damage. Neuropsychology 23: 529–534 (2009)
34. GraphPad Software: http://www.graphpad.com/quickcalcs/clinTest1/ (30.09.2016)
35. Green P, Rohling ML, Lees-Haley PR, Allen III LM: Effort has a greater effect on test scores than severe brain injury in compensation claimants. Brain Inj 15: 1045–1060 (2001)
36. Green P, Lees-Haley PR, Allen III LM: The Word Memory Test and the Validity of Neuropsychological Test Scores. Journal of Forensic Neuropsychology 2: 97–124 (2003)

37. Green P: Green's Word Memory Test for Microsoft Windows®, User's Manual, Revised June 2005. Green's Publishing, Edmonton, S. 9–68 (2005)

38. Green P, Montijo J, Brockhaus R: High Specificity of the Word Memory Test and Medical Symptom Validity Test in Groups with Severe Verbal Memory Impairment. Appl Neuropsychol 18: 86–94 (2011)

39. Guilera G, Pino O, Gómez-Benito J, Rojo JE, Vieta E, Tabarés-Seisdedos R, Segarra N, Martínez-Arán A, Franco M, Cuesta MJ, Crespo-Facorro B, Bernardo M, Purdon SE, Díez T, Rejas J for the Spanish Working Group in Cognitive Function: Clinical usefulness of the screen for cognitive impairment in psychiatry (SCIP-S) scale in patients with type I bipolar disorder. Health Qual Life Outcomes 7: 1–10 (2009)

40. Hartman DE: The unexamined lie is a lie worth fibbing: Neuropsychological malingering and the Word Memory Test. Arch Clin Neuropsychol 17: 709–714 (2002)

41. Hedderich J, Sachs L: Angewandte Statistik. Methodensammlung mit R, 15. Aufl, Springer, Berlin Heidelberg, S. 123–728 (2016)

42. Henson RK: Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. Meas Eval Couns Dev 34: 177–189 (2001)

43. Iverson GL: Ethical Issues Associated With the Assessment of Exaggeration, Poor Effort, and Malingering. Appl Neuropsychol 13: 77–90 (2006)

44. Jelicic M, Merckelbach H, Candel I, Geraerts E: Detection of feigned cognitive dysfunction using special malinger tests: a simulation study in naïve and coached malingerers. Int J Neurosci 117: 1185–1192 (2007)

45. Jelicic M, Ceunen E, Peters MJV, Merckelbach H: Detecting Coached Feigning Using the Test of Memory Malingering (TOMM) and the Structured Inventory of Malingered Symptomatology (SIMS). J Clin Psychol 67: 850–855 (2011)

46. Kobelt A, Göbber J, Bassler M, Petermann F: Beschwerdenvalidität im Rahmen stationärer psychosomatischer Rehabilitation. Rehabilitation (Stuttg) 51: 349–355 (2012)

47. Larrabee GJ: False-Positive Rates Associated with the Use of Multiple Performance and Symptom Validity Tests. Arch Clin Neuropsychol 29: 364–373 (2014)

48. Lasser I: Die deutsche Version des Screen for Cognitive Impairment in Psychiatry (SCIP-D): Psychometrische Eigenschaften einer kurzen Skala zur Erfassung kognitiver Leistungsfähigkeit bei schizophrenen Patienten. Psychologische Diplomarbeit, Universität Wien (2011)

49. Mann HB, Whitney DR: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 18: 50–60 (1947)

50. Merckelbach H, Smith GP: Diagnostic accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in detecting instructed malingering. Arch Clin Neuropsychol 18: 145–152 (2003)

51. Merten T, Friedel E, Mehren G, Stevens A: Uber die Validität von Persönlichkeitsprofilen in der nervenärztlichen Begutachtung. Nervenarzt 78: 511–520 (2007)

52. Merten T, Merckelbach H: Symptom Validity Testing in Somatoform and Dissociative Disorders: A Critical Review. Psychol Inj Law 6: 122–137 (2013)

53. Merten T: Beschwerdenvalidierung. Reihe: Fortschritte der Neuropsychologie, Bd 14, Thöne-Otto A, Flor H, Gauggel S, Lautenbacher S, Niemann H (Hrsg). Hogrefe, Göttingen Bern Wien, S. 1–102 (2014)

54. Mittenberg W, Patton C, Canyock EM, Condit DC: Base Rates of Malingering and Symptom Exaggeration. J Clin Exp Neuropsychol 24: 1094–1102 (2002)

55. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG: CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol 63: e1–e37 (2010)

56. Nies KJ, Sweet JJ: Neuropsychological Assessment and Malingering: A Critical Review of Past and Present Strategies. Arch Clin Neuropsychol 9: 501–552 (1994)

57. Peters MJV, Jelicic M, Moritz S, Hauschildt M, Jelinek L: Assessing the Boundaries of Symptom Over-Reporting Using the Structured Inventory of Malingered Symptomatology in a Clinical Schizophrenia Sample: Its Relation to Symptomatology and Neurocognitive Dysfunctions. J Exp Psychopathol 4: 64–77 (2013)

58. Piersma HL, Reaume WM, Boes JL: The Brief Symptom Inventory (BSI) as an outcome measure for adult psychiatric inpatients. J Clin Psychol 50: 555–563 (1994)

59. Pino O, Guilera G, Rojo JE, Gómez-Benito J, Bernardo M, Crespo-Facorro B, Cuesta MJ, Franco M, Martinez-Aran A, Segarra N, Tabarés-Seisdedos R, Vieta E, Purdon SE, Díez T, Rejas J on behalf of the Spanish Working Group in Cognitive Function: Spanish version of the Screen for Cognitive Impairment in Psychiatry (SCIP-S): Psychometric properties of a brief scale for cognitive evaluation in schizophrenia. Schizophr Res 99: 139–148 (2008)

60. Purdon SE: The Screen for Cognitive Impairment in Psychiatry: Administration and Psychometric Properties. PNL, Edmonton, S. 13 (2005)

61. Resnick PJ: The Detection of Malingered Mental Illness. Behav Sci Law 2: 21–38 (1984)

62. Rienstra A, Spaan PEJ, Schmand B: Reference Data for the Word Memory Test. Arch Clin Neuropsychol 24: 255–262 (2009)

63. Rogers R, Gillis JR, Dickens SE, Bagby RM: Standardized Assessment of Malingering: Validation of the Structured Interview of Reported Symptoms. Psychol Assess 3: 89–96 (1991)

64. Rohling ML, Green P, Allen III LM, Iverson GL: Depressive symptoms and neurocognitive test scores in patients passing symptom validity tests. Arch Clin Neuropsychol 17: 205–222 (2002)

65. Rojo E, Pino O, Guilera G, Gómez-Benito J, Purdon SE, Crespo-Facorro B, Cuesta MJ, Franco M, Martínez-Arán A, Segarra N, Tabarés-Seisdedos R, Vieta E, Bernardo M, Mesa F, Rejas J and on behalf of the Spanish Working Group in Cognitive Function: Neurocognitive diagnosis and cut-off scores of the Screen for Cognitive Impairment in Psychiatry (SCIP-S). Schizophr Res 116: 243–251 (2010)

66. Shapiro SS, Wilk MB, Chen HJ: A Comparative Study of Various Tests for Normality. J Am Stat Assoc 63: 1343–1372 (1968)

67. Smith GP, Burger GK: Detection of Malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS). J Am Acad Psychiatry Law 25: 183–189 (1997)

68. Stevens A, Friedel E, Mehren G, Merten T: Malingering and uncooperativeness in psychiatric and psychological assessment: Prevalence and effects in a German sample of claimants. Psychiatry Res 157: 191–200 (2008)

69. Student: The Probable Error of a Mean. Biometrika 6: 1–25 (1908).

70. Sullivan BK, May K, Galbally L: Symptom Exaggeration by College Adults in Attention-Deficit Hyperactivity Disorder and Learning Disorder Assessments. Appl Neuropsychol 14: 189–207 (2007)

71. Tan JE, Slick DJ, Strauss E, Hultsch DF: How'd They Do It? Malingering Strategies on Symptom Validity Tests. Clin Neuropsychol 16: 495–505 (2002)

72. Urbán R, Kun B, Farkas J, Paksi B, Kökönyei G, Unoka Z, Felvinczi K, Oláh A, Demetrovics Z: Bifactor structural model of symptom checklists: SCL-90-R and Brief Symptom Inventory (BSI) in a non-clinical community sample. Psychiatry Res 216: 146–154 (2014)

73. Van Egmond J, Kummeling I, aan Balkom T: Secondary gain as hidden motive for getting psychiatric treatment. Eur Psychiatry 20: 416–421 (2005)

74. van Impelen A, Merckelbach H, Jelicic M, Merten T: The Structured Inventory of Malingered Symptomatology (SIMS): A Systematic Review and Meta-Analysis. Clin Neuropsychol 28: 1336–1365 (2014)

75. Veiel HOF: A Preliminary Profile of Neuropsychological Deficits Associated with Major Depression. J Clin Exp Neuropsychol 19: 587–603 (1997)

76. Verwaltungsgerichtsordnung in der Fassung der Bekanntmachung vom 19. März 1991 (BGBl. I S. 686), die zuletzt durch Artikel 3 des Gesetzes vom 21. Dezember 2015 (BGBl. I S. 2490) geändert worden ist. Fundstelle: http://www.landesrecht-bw.de/jportal/portal/t/5iy/page/bsbawueprod.psml?pid=Dokumentanzeige&showdoccase=1&js_peid=Trefferliste&documentnumber=2&numberofresults=10&fromdoctodoc=yes&doc.id=BJNR000170960BJNE009002307&doc.part=S&doc.price=0.0#focuspoint (01.10.2016)

77. Verwaltungsgerichtsordnung in der Fassung der Bekanntmachung vom 19. März 1991 (BGBl. I S. 686), die zuletzt durch Artikel 3 des Gesetzes vom 21. Dezember 2015 (BGBl. I S. 2490) geändert worden ist. Fundstelle: http://www.landesrecht-bw.de/jportal/portal/t/9a3/page/bsbawueprod.psml;jsessionid=AC6B93EC6DB76EB029F79ECE5E8D9F9D.jp91?pid=Dokumentanzeige&showdoccase=1&js_peid=Trefferliste&documentnumber=2&numberofresults=10&fromdoctodoc=yes&doc.id=BJNR000170960BJNE009101308&doc.part=S&doc.price=0.0#focuspoint (01.10.2016)

78. World Medical Association: World Medical Association Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects. Bull World Health Organ 79: 373–374 (2001)

79. Youden WJ: Index for rating diagnostic tests. Cancer 3: 32–35 (1950)

Anhang

Anhang A: S. 170	Auszug aus der Patienteninformation
Anhang B: S. 172	Case Report Form (CRF)
Anhang C: S. 174	Patientenkommentare nach Versuchsdurchführung
Anhang D: S. 180	Tabellenverzeichnis
Anhang E: S. 183	Abbildungsverzeichnis

Anhang A

Auszug aus der Patienteninformation (Abdruck mit freundlicher Genehmigung von Prof. Dr. Tilman Steinert, Layout angepasst):

Patienteninformation

Einfluss von unterschiedlichen Instruktionen auf die Ergebnisse in Beschwerdevalidierungstests

Sehr geehrte Patientin, sehr geehrter Patient,

Wir möchten Sie bitten, an einer wissenschaftlichen Studie teilzunehmen. Es handelt sich um eine Studie im Rahmen einer medizinischen Dissertation, die von dem Doktoranden Lennart Kirchhoff durchgeführt wird. Die Studie hat nichts mit Ihrer persönlichen Behandlungssituation zu tun, Ihre Fragebögen kommen nicht in die Krankenakte und die Ergebnisse werden den Behandlern auch nicht mitgeteilt. Die Studie dient ausschließlich wissenschaftlichen Zwecken; Sie wurden dafür als mögliche Teilnehmerin/möglicher Teilnehmer angesprochen, weil Sie auf einer Station für Patienten mit depressiven Beschwerden behandelt werden.

Im Rahmen dieser Studie werden die Teilnehmer gebeten, zwei sogenannte Beschwerdevalidierungstests, zwei Fragebögen zu ihrer psychischen Befindlichkeit sowie zu Konzentration und Gedächtnis und einen kurzen Fragebogen zu ihren persönlichen Daten auszufüllen. Beschwerdevalidierungstests werden in Rentenverfahren verwendet. Sie sollen eine Aussage darüber erlauben, ob die von Rentenantragstellern geschilderten Beschwerden auch tatsächlich vorhanden sind. Wir wollen herausfinden, ob diese Tests sich tatsächlich dazu eignen zu überprüfen, ob Beschwerden zutreffend geschildert werden. Das wichtigste Prinzip dieser Studie ist, dass Sie über ein Zufallsverfahren eine von zwei verschiedenen Anweisungen erhalten werden, mit welcher inneren Einstellung Sie die Fragebögen ausfüllen sollen. Es ist von großer Wichtigkeit, dass Sie dies beim Ausfüllen der Fragebögen berücksichtigen. Bei

dem einen Fragebogen werden Sie nach Erhalt Ihrer Anweisungen gefragt, ob bestimmte Symptome bei Ihnen vorhanden sind. Bei dem anderen Test, der am Laptop durchgeführt wird, handelt es sich um einen Test der Merkfähigkeit für Wörter. Im Gegensatz zu diesen Beschwerdevalidierungstests dienen die anderen Fragebögen, die Sie davor ausfüllen, dazu, Ihre tatsächlichen Beschwerden zu erfassen. Das ist erforderlich, um nachher zu überprüfen, ob und wie die Ergebnisse der Beschwerdevalidierungstests auch von den tatsächlichen Beschwerden beeinflusst werden. Der gesamte Ablauf wird in etwa 90 Minuten in Anspruch nehmen. Risiken bei der Studienteilnahme bestehen nach vernünftigem Ermessen nicht.

Für die Auswertung speichern wir zunächst Ihre Daten unter einem Nummerncode (sog. Pseudonymisierung). Ihr Name wird nicht erfasst. Die Ergebnisse der Studie werden dann in einer medizinischen Doktorarbeit verwendet und in einer Fachzeitschrift in anonymisierter Form als statistische Daten veröffentlicht. Ein Rückschluss auf Einzelpersonen ist somit also ausgeschlossen.

Als Gratifikation für Ihre Teilnahme erhalten Sie eine Entschädigung in Höhe von 15 €.

Freiwilligkeit:

An diesem Forschungsprojekt nehmen Sie freiwillig teil. Ihr Einverständnis können Sie jederzeit und ohne Angabe von Gründen widerrufen, dann werden alle bis dahin studienbedingt erhobenen Daten gelöscht und die entsprechenden Proben werden vernichtet. Dieser eventuelle Widerruf hat keine Auswirkungen auf Ihre medizinische Betreuung. […]

Anhang B

Case Report Form (CRF):

Patienten-ID: ☐☐☐ Protokoll-Nr.: ☐☐

Basisdaten

Im Falle einer Korrektur: Falsches einfach durchstreichen und Korrektur mit Datum und Unterschrift am Rand anfügen.

Geburtsdatum (DD/MM/YYYY): ☐☐ /☐☐ /☐☐☐☐ Alter (Jahre): ☐☐

Geschlecht: ☐ männlich ☐ weiblich

Bildungsabschluss: (Höchstes Zutreffendes ankreuzen) ☐ Hauptschule ☐ Mittlere Reife ☐ Abitur ☐ Universität ☐ Fachhochschule ☐ kein Abschluss

Studienrelevante Hauptdiagnosen nach ICD-10: ☐ F 32.0 ☐ F32.1 ☐ F33.0 ☐ F33.1 ☐ F34.1 ☐ F43.2 ☐ F40 ☐ F41

Rekrutierung in: ☐ stationärer Behandlung ☐ ambulanter Behandlung

Beschäftigungsverhältnis (aktuell): ☐ erwerbstätig ☐ erwerbslos

Erwerbsminderungsrente bereits ≥ einmal abgelehnt worden: ☐ ja ☐ nein

Bereits berentet: ☐ ja ☐ nein

Testdurchführung

1. BSI (siehe Anhang)

2. SCIP (siehe Anhang, inkl. Anleitung)

3. Patienteninstruktionen

4. WMT (1. Teil) (Computer-Programm)

5. Pause (ca. 10 min)

6. SFSS (siehe Anhang)

7. WMT (2.Teil)

Visuelle Analogskala

Bitte kreuzen Sie auf der unten folgenden Skala an, wieviel Prozent Ihres tatsächlichen Leistungsvermögens Sie bei der Testdurchführung von WMT und SFSS gezeigt haben!

0% Leistung 100% Leistung

Entblindung

Proband gehörte zur Gruppe: ☐ A ☐ B

Unterschrift des Testleiters
(Lennart Kirchhoff)

Ort, Datum

Anhang C

Patientenkommentare nach Versuchsdurchführung:

uHJ (Gruppe B): die Teilnehmerin fand die Instruktion zwar recht schwierig umzusetzen, war aber nicht überfordert und konnte sich gut mit ihrer „Rolle“ identifizieren.

lJQ (Gruppe B): die Teilnehmerin konnte sich nach eigenen Angaben gut in die in der Instruktion geschilderte Situation hineinversetzen, fand es aber „furchtbar schwer“ die Tests entsprechend der Instruktion zu beantworten und gleichzeitig nicht als „Simulant“ aufzufallen. Sie bezeichnete die Testdurchführung in dieser Form als „sehr anspruchsvoll". Zudem war sie der Meinung, dass sie sich noch mehr Fehler hätte erlauben können. Nach eigenem Bekunden war sie am Ende besser als beabsichtigt.

ddO (Gruppe B): Der Teilnehmerin widerstrebte es nach eigenen Angaben, sich beim WMT an die Instruktion B zu halten und absichtliche Falsch-Antworten zu geben, schließlich sei der Test „so einfach“. Sie habe aber letztlich doch versucht, die Instruktion so gut wie möglich umzusetzen.

AFc (Gruppe B): Auch diesem Teilnehmer fiel es nach eigenem Bekunden schwer die Instruktion B umzusetzen. Es sei ihm „etwas gegen den Strich“ gegangen absichtlich falsche Antworten einzubauen. Dennoch habe er absichtlich falsche Antworten eingestreut und gleichzeitig versucht, immer dieselben Fehler zu machen, um konsistent zu antworten.

PZV (Gruppe B): Die Teilnehmerin gab an, sie habe bei den Tests „natürlich“ besser abschneiden können. Bei anspruchsvolleren Aufgaben habe sie absichtlich falsche Antworten gegeben. „Wenn ich bei einer Antwort sowieso nicht sicher war, habe ich das deutlich gemacht und nicht mehr 100%ig nachgedacht. Lügen ist nicht meine Sache, aber dieses ‚Deutlich machen‘ fiel mir leicht.“

SVE (Gruppe B): Die Teilnehmerin erzählte, sie habe versucht, sich an die Aggravationsinstruktion zu halten und diese gut umzusetzen, auch wenn es nach eigenen Angaben „ganz schön fordernd“ für sie war, da sie sich habe überlegen müssen, wann und wie sie „falsche“ Antworten gebe.

OlP (Gruppe B): Die Teilnehmerin sagte im Anschluss an die Versuchsdurchführung: „Wenn ich beim einem Gutachter wäre, dann würde ich wieder heimgeschickt werden wollen als kerngesund, denn ich möchte vor allem wieder gesund werden. Ich könnte den Gutachter nicht vorsätzlich täuschen, damit er mich schlechter einschätzt.“ Daher habe sie die Tests auch ehrlich ausgefüllt. ***Für die PPA ausgeschlossen!***

TuP (Gruppe B): Nach der Versuchsdurchführung gab die Teilnehmerin zu Protokoll: „Wenn ich etwas kann, dann kann ich das. Bei den Tests sollte ich mich jetzt zwar etwas doof stellen, aber das war mir in der Situation nicht recht. Wenn ich bei einer Begutachtung etwas nicht wüsste, dann würde ich das immer nachfragen. Daher habe ich auch hier immer alles gesagt, was ich wusste und war ansonsten ehrlich.“ ***Für die PPA ausgeschlossen!***

fRC (Gruppe B): Der Teilnehmer meinte nach der Versuchsdurchführung, er habe beim SFSS „ein bisschen“ versucht die Anweisung umzusetzen, beim WMT sei ihm aber einfach „das Ego durchgekommen“. Er habe sehen wollen, ob sein Kurzzeitgedächtnis noch funktioniert und die Instruktion in der Situation gar nicht mehr beachtet. ***Für die PPA ausgeschlossen!***

Zwv (Gruppe B): Der Studienteilnehmerin fiel es nach eigenen Angaben eher schwer absichtlich falsche Antworten zu geben. Für sie sei es aber in der Studiensituation auch „um nichts gegangen“. Sie könne sich gut vorstellen, dass man in einer „echten“ Gutachtensituation verstärkt Beschwerden zeigt und „mehr schummelt“. Hin und wieder habe sie nach eigenen Angaben falsche Antworten eingestreut, z.B. beim WMT statt „Schwein-Speck“, „Schwein-Schinken“.

FrJ (Gruppe B): Der Studienteilnehmer gab an, er habe versucht, die Instruktion auf sich zu beziehen und sich in die Rolle hineinzuversetzen. Theoretisch hätte er schon auch Symptome übertreiben können, aber so schlimm sei es bei ihm von der Symptomatik einfach nicht, weshalb er letztlich nichts übertrieben habe, sondern aufrichtig geantwortet habe. ***Für die PPA ausgeschlossen!***

wBc (Gruppe B): Die Studienteilnehmerin gab zu Protokoll, sie könne sich nicht verstellen und schon gar nicht könne sie lügen. Selbst wenn sie sich schlechter machen solle, um ihren Rentenantrag „durchzukriegen“, könne sie das nicht, das

widerspräche ihrer „Einstellung“. Auch in einer „realen“ Situation würde sie das nicht machen. ***Für die PPA ausgeschlossen!***

zML (Gruppe B): Die Studienteilnehmerin erklärte im Anschluss an die Versuchsdurchführung: „Ich gebe immer 100% und verstelle mich nicht, auch nicht in einer Gutachtensituation. Ich kann so etwas nicht. Da gehe ich hin, so wie ich bin.“ ***Für die PPA ausgeschlossen!***

oYE (Gruppe B): Die Studienteilnehmerin zeigte während der Versuchsdurchführung ein sehr schlechtes Konzentrationsvermögen. Am Ende kam heraus, dass sie nicht verstanden hatte, dass die Instruktion auf die Tests zu beziehen war. Sie dachte, sie solle die Instruktion einfach nur auswendig lernen. Nach eigenen Angaben strengte sie sich bei den Tests an. ***Für die PPA ausgeschlossen!***

pva (Gruppe B): Die Studienteilnehmerin gab an, sich an die Instruktion gehalten zu haben, sagte aber gleichzeitig, sie habe bei den Tests 100% gegeben und versucht, alles wahrheitsgemäß zu beantworten. Es ist stark zu anzunehmen, dass die Instruktion nicht richtig verstanden wurde! ***Für die PPA ausgeschlossen!***

roT (Gruppe A): Die Studienteilnehmerin erklärte, sie habe versucht, sich an die Instruktion zu halten. Es sei kein Problem gewesen diese umzusetzen. Sie habe alles wahrheitsgemäß beantwortet.

SVp (Gruppe B): Der Studienteilnehmer gab zu Protokoll: „Wenn etwas falsch beantwortet ist, dann nicht mit Absicht.“ Er habe versucht, alles „so ehrlich wie möglich“ auszufüllen. ***Für die PPA ausgeschlossen!***

qKe (Gruppe A): Der Studienteilnehmer sagte, er habe versucht, sich an die Instruktion zu halten und ehrlich zu antworten.

FSF (Gruppe B): Der Studienteilnehmer erklärte im Anschluss an die Versuchsdurchführung: „Ich habe bei beiden Tests extra mal etwas falsch gemacht. Ich habe nicht alles gezeigt, was ich gewusst hätte.“

Xhw (Gruppe B): Die Studienteilnehmerin fand es nach eigenem Bekunden schwierig die Instruktion umzusetzen. Der WMT sei für sie „Kinderkram“ gewesen, da würde sie „in echt auch keine Fehler einbauen", denn Leistung habe nichts mit

der Psyche zu tun. Der WMT sei also eine leichte Übung für sie gewesen, beim SFSS habe sie dann vielleicht etwas übertrieben, sie sei aber eigentlich ein ehrlicher Mensch. ***Für die PPA ausgeschlossen!***

HSf (Gruppe B): Die Studienteilnehmerin gab im Anschluss an die Versuchsdurchführung zu Protokoll, sie gehöre „zur ehrlichen Spezies“ und würde nie mit „Versicherungsbetrug oder Ähnlichem“ zu tun haben wollen. „Ich habe das Ehrlichkeits-Gen.“ Deshalb habe sie auch alles ehrlich ausgefüllt. ***Für die PPA ausgeschlossen!***

hTd (Gruppe B): Die Studienteilnehmerin erklärte, sie habe sich zu 100% angestrengt alles richtig zu beantworten. Sie habe auch Symptome nicht schwerer dargestellt als sie seien und sei bei den Tests immer ehrlich gewesen. ***Für die PPA ausgeschlossen!***

qWf (Gruppe A): Der Studienteilnehmer gab an, er habe stets ehrlich geantwortet und sich an die „Aufgabe“ gehalten. Er habe sich „angestrengt, das alles richtig zu machen“.

eOp (Gruppe B): Die Studienteilnehmerin sagte nach der Versuchsdurchführung, sie habe versucht, den Schweregrad ihrer Beeinträchtigung darzustellen, indem sie ihre Symptome übertrieben dargestellt habe. Nach eigenen Angaben habe sie diese „sehr schwerwiegend dargestellt“. Beim WMT habe sie absichtlich Fehler eingebaut.

Qpg (Gruppe A): Die Studienteilnehmerin erklärte, sie habe „die Fragen nach bestem Wissen und Gewissen beantwortet“ und sich „voll angestrengt“.

OvJ (Gruppe A): Der Studienteilnehmer gab zu Protokoll, er habe versucht, immer so ehrlich wie möglich zu antworten und habe keine Symptome übertrieben.

pNY (Gruppe A): Die Studienteilnehmerin gab an, dass sie versucht habe, die Tests wahrheitsgemäß und so gut es ihr möglich war zu beantworten.

Wbi (Gruppe B): Die Studienteilnehmerin sagte im Anschluss an die Versuchsdurchführung, sie habe sich bei den Tests „doofer gestellt“ als sie sei und habe auch versucht, sich ihre Fehler zu merken und diese zu wiederholen. Beim SFSS habe sie versucht 40% der Fragen ehrlich und die restlichen 60% Prozent nicht-ehrlich zu beantworten.

lPr (Gruppe B): Die Studienteilnehmerin erklärte, sie habe sich genau überlegt, wie sie die falschen Antworten auswählen solle und abgeschätzt, wie viele Fragen es in etwa seien. „Mindestens die Hälfte" der Fragen habe sie falsch beantworten wollen, um ihre „Beeinträchtigung deutlich zu machen“.

ijI (Gruppe B): Die Studienteilnehmerin gab an, mittlerweile „nur noch 60%“ ihres ursprünglichen Leistungsvermögens zu haben. Von diesen 60% habe sie aber bei der Testdurchführung 100% gezeigt und sich angestrengt, alles richtig zu beantworten, so gut wie sie es sich eben merken konnte. ***Für die PPA ausgeschlossen!***

xcb (Gruppe A): Die Studienteilnehmerin meinte im Anschluss, sie habe sich angestrengt, immer ehrlich zu antworten und beim SFSS habe sie versucht, alles so gut wie möglich anzukreuzen.

jFk (Gruppe B): Die Studienteilnehmerin erweckte durch ihre Kommentare den Eindruck, nicht abstrahieren zu können, dass die Instruktion im Briefumschlag Auswirkungen auf die Testdurchführung haben könnte. Sie erklärte, sich angestrengt zu haben, um die Aufgaben richtig zu beantworten. „Wenn man sich nicht anstrengt, dann fliegst ja gleich durch.“ ***Für die PPA ausgeschlossen!***

RoX (Gruppe A): Die Studienteilnehmerin sagte nach der Versuchsdurchführung, sie habe bei den Tests alles gegeben, was zurzeit möglich sei. Vor vier Wochen wären die Tests ihrer Einschätzung nach wahrscheinlich deutlich schlechter ausgefallen, da sie zu der Zeit deutlich eingeschränkter gewesen sei.

KbI (Gruppe A): Die Studienteilnehmerin erklärte, sie habe sich „zu 100% angestrengt“, die Fragen richtig zu beantworten und habe nichts absichtlich falsch beantwortet. Sie könne nicht lügen, weil ihr das jeder sofort ansehe.

OBe (Gruppe B): Der Studienteilnehmer gab im Anschluss an die Versuchsdurchführung zu Protokoll, er habe versucht, „etwa 30% der Fragen“ falsch zu beantworten, „damit die auf der Rentenstelle nicht gleich merken, dass man schummelt“. Alle anderen falschen Antworten seien vermutlich wirklich falsch, aufgrund seiner derzeitigen schlechten Verfassung.

pTJ (Gruppe B): Bezüglich des WMT gab die Studienteilnehmerin an, sie habe hin und wieder auch absichtlich etwas Falsches gesagt bzw. „vergessen“, um ihre Beschwerden deutlich zu machen. Sonst hätte sie beim WMT „schon noch mehr

gewusst“. Bezüglich des SFSS erklärte sie, nicht völlig übertrieben zu haben. Sie habe versucht alles so darzustellen, dass es glaubhaft erscheine, „so wie es halt gefordert war“.

bxP (Gruppe A): Der Studienteilnehmer meinte im Anschluss an die Versuchsdurchführung, er habe „alles gegeben und alles gezeigt", was er gewusst habe. Beim Multiple-Choice-Durchgang des WMT sei er aber kurzzeitig abgelenkt gewesen und habe daher dort ein paar unbeabsichtigte Fehler gemacht.

WBi (Gruppe A): Der Studienteilnehmer erklärte, er habe sich „auf jeden Fall“ angestrengt, alles richtig zu beantworten. Mehr gehe „zur Zeit halt nicht“ bei ihm.

TRi (Gruppe A): Der Studienteilnehmer gab zu Protokoll, er habe nichts absichtlich falsch beantwortet. Vielleicht habe er die eine oder andere Frage falsch verstanden, ansonsten habe er sich aber wirklich angestrengt. Er sei immer ehrlich erzogen worden und sei daher eine „ehrliche Haut“.

vbp (Gruppe B): Der Studienteilnehmer sagte im Anschluss, er habe „extra“ auch falsche Antworten beim WMT eingebaut, allerdings ohne eine bestimmte Strategie zu verfolgen. Beim SFSS habe er ebenfalls absichtlich Fehler gemacht bzw. Fragen „anders beantwortet als normalerweise“. Er habe also auch Antworten gegeben, die vielleicht so nicht zutreffen.

piE (Gruppe B): Der Studienteilnehmer meinte, er habe bei beiden Tests alles gesagt, was er gewusst habe und nichts absichtlich falsch beantwortet. „Ich bin ein ehrlicher Mensch, auch wenn die Anweisung eine andere war“. ***Für die PPA ausgeschlossen!***

uQc (Gruppe B): Die Studienteilnehmerin gab nach der Versuchsdurchführung an, sie habe entsprechend der Instruktion absichtlich falsch geantwortet und sich „extra dumm gestellt“. Beim WMT habe sie sich ihre Fehler mit Absicht nicht gemerkt, weil ihr „Gedächtnis ja so schlecht“ sei. Sie habe beim WMT gleich versucht, sich andere Assoziationen zu merken (z.B.: „Salz – Pfeffer“). Beim SFSS habe sie dann auch „in Richtung Gedächtnisstörungen simuliert“. Die SFSS-Items, bei denen es um Stimmenhören ging, habe sie verneint, das sei nicht ihr „Problemfeld“ gewesen.

Anhang D

Tabellenverzeichnis:

Tabelle 1: Verteilung der soziodemographischen Variablen und der BSI- und SCIP-Ergebnisse in der Gesamtstichprobe aller 60 Studienteilnehmer sowie Aufteilung dieser Merkmale nach der Randomisierung in die beiden Studiengruppen A und B (je 30 Studienteilnehmer) 40
Tabelle 2: Gegenüberstellung der bei WMT, SFSS und VAS erzielten Ergebnisse in den beiden Studiengruppen A und B (je 30 Studienteilnehmer) 44
Tabelle 3: Deskriptive Darstellung und Analyse der SFSS-Ergebnisse aller 60 Studieneilnehmer sowie Gegenüberstellung der in den beiden Studiengruppen A und B (je 30 Studienteilnehmer) erzielten SFSS-Ergebnisse 48
Tabelle 4: Deskriptive Darstellung und Analyse der WMT-Ergebnisse aller 60 Studieneilnehmer sowie Gegenüberstellung der in den beiden Studiengruppen A und B (je 30 Studienteilnehmer) erzielten WMT-Ergebnisse 52
Tabelle 5: Darstellung der Klassifikationen aller 60 Studienteilnehmer für den WMT-Cut-off von ≤82,5% in Form einer Vierfeldertafel 54
Tabelle 6: Darstellung der Klassifikationen aller 60 Studienteilnehmer für den SFSS-Cut-off von >16 in Form einer Vierfeldertafel 55
Tabelle 7: Darstellung der Häufigkeit falsch-positiver und richtig-negativer Klassifikationen durch den SFSS bei einem Cut-off von >16 in Teilgruppe A (beinhaltet 30 der insgesamt 60 Studienteilnehmer) 56
Tabelle 8: Darstellung der Häufigkeit falsch-positiver und richtig-negativer Klassifikationen durch den WMT bei einem Cut-off von ≤82,5% in Teilgruppe A (beinhaltet 30 der insgesamt 60 Studienteilnehmer) 57
Tabelle 9: Darstellung des für die Fläche unter der SFSS-ROC-Kurve aller 60 Studienteilnehmer berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall 60
Tabelle 10: Auflistung der Koordinaten der SFSS-ROC-Kurve aller 60 Studienteilnehmer. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten SFSS-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind 61
Tabelle 11: Darstellung des für die Fläche unter der WMT-ROC-Kurve aller 60 Studienteilnehmer berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall 65
Tabelle 12: Auflistung der Koordinaten der WMT-ROC-Kurve aller 60 Studienteilnehmer. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten WMT-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind 66

Tabelle 13: Darstellung der Übereinstimmung der Klassifikationsergebnisse von SFSS und WMT (bestanden/nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für die Gesamtstichprobe aller 60 Studienteilnehmer.. 69
Tabelle 14: Darstellung der Klassifikationsergebnisse von SFSS und WMT bei kombinierter Anwendung der beiden BVTs (Tests beide bestanden/beide nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für die Gesamtstichprobe aller 60 Studienteilnehmer..... 71
Tabelle 15: Auflistung der für den Zusammenhang zwischen BSI- und WMT-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten.. 74
Tabelle 16: Auflistung der für den Zusammenhang zwischen SCIP- und WMT-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten.. 77
Tabelle 17: Auflistung der für den Zusammenhang zwischen BSI- und SFSS-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten.. 80
Tabelle 18: Auflistung der für den Zusammenhang zwischen SCIP- und SFSS-Ergebnissen aller 60 Studienteilnehmer berechneten Korrelationskoeffizienten.. 82
Tabelle 19: Verteilung der soziodemographischen Variablen und der BSI- und SCIP-Ergebnisse in der Gesamtstichprobe der Per Protocol Analyse (45 Studienteilnehmer) sowie Aufteilung dieser Merkmale nach der Randomisierung in die beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer)..... 86
Tabelle 20: Gegenüberstellung der bei WMT, SFSS und VAS erzielten Ergebnisse in den beiden Studiengruppen A und B‘ (30 bzw. 15 Studienteilnehmer)..... 90
Tabelle 21: Deskriptive Darstellung und Analyse der SFSS-Ergebnisse aller 45 Studieneilnehmer der Per Protocol Analyse sowie Gegenüberstellung der in den beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) erzielten SFSS-Ergebnisse 94
Tabelle 22: Deskriptive Darstellung und Analyse der WMT-Ergebnisse aller 45 Studieneilnehmer der Per Protocol Analyse sowie Gegenüberstellung der in den beiden Studiengruppen A und B‘ (30 bzw. 15 Teilnehmer) erzielten WMT-Ergebnisse 98
Tabelle 23: Darstellung der Klassifikationen aller 45 Studienteilnehmer der Per Protocol Analyse für den WMT-Cut-off von ≤82,5% in Form einer Vierfeldertafel 100
Tabelle 24: Darstellung der Klassifikationen aller 45 Studienteilnehmer der Per Protocol Analyse für den SFSS-Cut-off von >16 in Form einer Vierfeldertafel ... 101
Tabelle 25: Darstellung des für die Fläche unter der SFSS-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall..... 104
Tabelle 26: Auflistung der Koordinaten der SFSS-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten SFSS-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind 105

Tabelle 27: Darstellung des für die Fläche unter der WMT-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Wertes (AUC) mit Standardfehler, Signifikanz des Unterschiedes zur Fläche eines diagnostischen Zufallsverfahrens (AUC = .5) und 95%igem Konfidenzintervall 109
Tabelle 28: Auflistung der Koordinaten der WMT-ROC-Kurve aller 45 Studienteilnehmer der Per Protocol Analyse. Die erste Spalte benennt dazu die im Rahmen der ROC-Kurven-Analyse variierten WMT-Cut-offs, für welche in den weiteren Spalten die jeweils berechneten Kennwerte Sensitivität, Spezifität, PPW, NPW und Youden-Index mitangegeben sind 110
Tabelle 29: Darstellung der Übereinstimmung der Klassifikationsergebnisse von SFSS und WMT (bestanden/nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für alle 45 Studienteilnehmer der Per Protocol Analyse 113
Tabelle 30: Darstellung der Klassifikationsergebnisse von SFSS und WMT bei kombinierter Anwendung der beiden BVTs (Tests beide bestanden/beide nicht bestanden, für die Cut-offs: >16 bzw. ≤82,5%) in Form einer Vierfeldertafel für alle 45 Studienteilnehmer der Per Protocol Analyse 115
Tabelle 31: Auflistung der für den Zusammenhang zwischen BSI- und WMT-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten 118
Tabelle 32: Auflistung der für den Zusammenhang zwischen SCIP- und WMT-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten 122
Tabelle 33: Auflistung der für den Zusammenhang zwischen BSI- und SFSS-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten 125
Tabelle 34: Auflistung der für den Zusammenhang zwischen SCIP- und SFSS-Ergebnissen aller 45 Studienteilnehmer der Per Protocol Analyse berechneten Korrelationskoeffizienten 127

Anhang E

Abbildungsverzeichnis:

Abbildung 1: ROC-Kurven-Analyse für den SFSS-Gesamtscore aller 60 Studienteilnehmer 59
Abbildung 2: ROC-Kurven-Analyse für den Score WMT-Durchgänge IR, DR, CNS aller 60 Studienteilnehmer..... 64
Abbildung 3: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 60 Studienteilnehmer gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke..... 73
Abbildung 4: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 60 Studienteilnehmer gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke. 76
Abbildung 5: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 60 Studienteilnehmer gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke ... 79
Abbildung 6: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 60 Studienteilnehmer gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 60 Studienteilnehmer in Form einer Punktwolke..... 81
Abbildung 7: Darstellung des Ein- bzw. Ausschlusses der insgesamt 60 Studienteilnehmer in Form eines Flussdiagramms..... 84
Abbildung 8: ROC-Kurven-Analyse für den SFSS-Gesamtscore aller 45 Studienteilnehmer der Per Protocol Analyse..... 103
Abbildung 9: ROC-Kurven-Analyse für den Score WMT-Durchgänge IR, DR, CNS aller 45 Studienteilnehmer der Per Protocol Analyse 108
Abbildung 10: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke..... 117
Abbildung 11: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die WMT-Ergebnisse (Score WMT-Durchgänge IR, DR, CNS; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke..... 121
Abbildung 12: Auftragung der BSI-Ergebnisse (ausgedrückt im Punktwert des GSI; Abszisse) aller 45 Studienteilnehmer der Per Protocol Analyse gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke 124
Abbildung 13: Auftragung der SCIP-Ergebnisse (ausgedrückt im Punktwert des SCIP-Summenscores; Abszisse) aller 45 Studienteilnehmer der Per Protocol

Analyse gegen die SFSS-Ergebnisse (SFSS-Gesamtscore; Ordinate) aller 45 Studienteilnehmer der Per Protocol Analyse in Form einer Punktwolke..... 126

Danksagung

An dieser Stelle möchte ich zunächst Herrn Prof. Dr. Steinert für die außergewöhnlich gute Betreuung meiner Arbeit sowie die Idee für die oben aufgeführte Studie danken. Weiterhin danke ich Maria Weithmann für die ausführliche und verständliche Einführung in den Ablauf des Rentenverfahrens. Ein großer Dank gilt auch meinen Eltern, meiner Freundin und meinen Freunden für das Korrekturlesen, den steten produktiven Gedankenaustausch und vieles mehr.

Lebenslauf

Der Lebenslauf wurde aus Gründen des Datenschutzes entfernt.