Medizinische Fakultät der Martin-Luther-Universität Halle-Wittenberg

Der Zusammenhang zwischen Intelligenz und der Messbarkeit negativer Antwortverzerrung in der Begutachtung psychischer Erkrankungen

Dissertation

zur Erlangung des akademischen Grades

Doktor der Medizin (Dr. med.)

.....

vorgelegt

der Medizinischen Fakultät

der Martin-Luther-Universität Halle-Wittenberg

von Benjamin Reufsteck

geboren am 26.06.1987 in Oberhausen

Betreuer: apl. Prof. Dr. Stefan Watzke

Gutachterin/Gutachter:

Prof. Dr. phil. habil. Bernhard Strauß (Jena)

PD Dr. med. Frank Pillmann

02.04.2019

04.12.2019

Referat

Zielsetzung: Diese Arbeit hat zum Ziel, die Korrelation zwischen der Intelligenz Begutachteter und der Testgüte eines Beschwerdenvalidierungstests zu untersuchen. Es sollen Erkenntnisse gewonnen werden, ob die Messung bzw. Schätzung des Intellekts einer Probandin eine nützliche Zusatzinformation für die Bewertung des Testergebnisses liefern kann und ob es einen signifikanten Zusammenhang zwischen der Funktionalität des Tests und dem Intellekt der einzelnen Probandin gibt.

Probandinnen und Methodik: Die untersuchte Stichprobe umfasst n=60 Probandinnen, die randomisiert zwei Untersuchungsbedingungen (authentische / aggravierte Beschwerdendarlegung) zugeteilt wurden. Die Erhebung der Beschwerdenvalidität erfolgte durch die deutschen Version des *SIRS-2* (Schmidt et al., 2019). Die Intelligenz der Probandinnen wurde anhand der Untertests 1-3 des *Leistungsprüfsystems* (Horn, 1983) sowie mithilfe einer Sozialformel (nach Jahn et al., 2013) erfasst bzw. geschätzt.

Ergebnisse: Ein systematischer Gruppenunterschied der Beurteilungsqualität des *SIRS-2* zwischen *korrekter* vs. *nicht korrekter* Zuordnung ließ sich hinsichtlich keiner der Intelligenzmaße nachweisen. Bei der Teilgruppe von durch den *SIRS-2* in ihrer Beschwerdenvalidität nicht eindeutig Begutachteten ließ sich jedoch ein signifikanter Gruppenunterschied hinsichtlich der geschätzten Intelligenz finden. Übertreibende hatten deskriptiv einen höher geschätzten IQ.

Schlussfolgerung: Die Ergebnisse dieser Untersuchung bieten Hinweise auf Zusammenhänge zwischen Intelligenz der Probandinnen und der Messbarkeit negativer Antwortverzerrung. Für dieses Studienkollektiv stellte insbesondere die geschätzte Intelligenz einen hilfreichen Zusatzparameter zu den Ergebnissen des *SIRS-2* dar. Zukünftige Forschung sollte diese Interaktionen weiterführend untersuchen.

Reufsteck, Benjamin: Der Zusammenhang zwischen Intelligenz und der Messbarkeit negativer Antwortverzerrung in der Begutachtung psychischer Erkrankungen, Halle (Saale), Univ., Med. Fak., Diss., 80 Seiten, 2019

Inhaltsverzeichnis

Abkürzungsverzeichnis

AMDP	Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie
AUC	Area Under Curve
Beschw.	Beschwerden
BVT	Beschwerdenvalidierungstest
df	degrees of freedom (Freiheitsgrade)
DSM	Diagnostic and Statistical Manual of Mental Disorders
F	F-Teststatistik
F*xx.x*	Klassifikation psychiatrischer Erkrankungen gemäß *ICD-10*
(HA)WIE	(Hamburg-)Wechsler-Intelligenztest für Erwachsene
ICD	International Classification of Diseases
IQ	Intelligenzquotient
LPS	Leistungsprüfsystem
M	Mittelwert
max	maximaler Wert
min	minimaler Wert
MMPI	Minnesota Multiphasic Personality Inventory
p	probability – Signifikanzniveau
ROC	Receiver Operating Characteristic
SD	Standard Deviation (Standardabweichung)
SIRS	Structured Interview of Reported Symptoms
t	t-Test
WHO	World Health Organization (Weltgesundheitsorganisation)
$\chi 2$	Chi-Quadrat-Test

Des Weiteren werden in dieser Arbeit Abkürzungen gemäß Dudenkonvention verwendet (vgl. hierzu Steinhauer, 2011).

Hinweis: Zur sprachlichen Vereinfachung wird in dieser Dissertation bei der Benennung von gemischtgeschlechtlichen Personengruppen immer ausschließlich die weibliche Form stellvertretend für beide Geschlechter verwendet.

1 Einleitung

Innerhalb des Tätigkeitsbereichs der Begutachtung bergen psychische Erkrankungen besondere Herausforderungen für die Gutachterinnen. Anamnestisch erfragtes Erleben von Beschwerden stellt hier die Hauptinformationsquelle im Prozess der Begutachtung dar und ist anfällig gegenüber Störvariablen. Daher muss die Authentizität dieser Informationen geprüft werden, wofür seit einiger Zeit Beschwerdenvalidierungstests – zunächst im Englischen, zunehmend auch im Deutschen – entwickelt werden. Diese haben zum Ziel, negative Antwortverzerrung, also die bewusste oder unbewusste Täuschung in der Darlegung von Beschwerden, sicher zu detektieren. Da es sich bei Beschwerdenvalidität allerdings um ein schwer messbares Konstrukt handelt und Begutachtete möglicherweise gezielt auf ein bestimmtes Gutachtenergebnis – z.B. eine Verrentung – abzielen, ist der Prozess fehleranfällig. Den Gutachterinnen stehen zwar validierte Testverfahren zur Verfügung, jedoch müssen deren Ergebnisse zur Nutzung für das Gutachten im klinischen Kontext beurteilt werden.

Bis dato ist der Einfluss der Intelligenz Begutachteter auf die Testgütekriterien eingesetzter Beschwerdenvalidierungstests nicht hinreichend erforscht. Weiterführend stellt sich die Frage, ob Intelligenz in bestimmten Konstellationen eine wertvolle Zusatzvariable für die Einordnung eines Testergebnisses zur Beschwerdenvalidierung in den klinischen Kontext darstellt. Diesen Fragen zum Zusammenhang von Intelligenz und der Messbarkeit negativer Antwortverzerrung soll in dieser Arbeit nachgegangen werden.

Als Teil einer übergeordneten Studie zur Validierung der deutschen Version des *Structured Interview of Reported Symptoms* (*SIRS-2*; Schmidt et al., 2019) wird die Beschwerdenvalidität von Probandinnen untersucht, die randomisiert in eine authentische und eine aggravierende (übertreibende) Gruppe aufgeteilt wurden. Ebenso erfolgt ein orientierendes Assessment der Intelligenz mithilfe verschiedener Herangehensweisen, um die Verwendbarkeit schnell ermittelbarer Intelligenzmess- und -schätzwerte im Kontext der Beschwerdenvalidierung zu untersuchen.

2 Theoretischer Hintergrund und Zielstellung

2.1 Gutachten in der Medizin

Unter einem medizinischen Gutachten versteht man die „Anwendung medizinischer Erkenntnisse und Erfahrungen auf einen Einzelfall im Hinblick auf eine (oft aus rechtlichen Gründen notwendige) Fragestellung“ (Lippert et al., 2015, S.5). In Abgrenzung von Befundberichten und Attesten, die meist von Behörden und Versicherungen angefordert werden und juristisch gesehen einer schriftlichen Zeugenaussage gleichkommen, gilt es als Wesensmerkmal des Gutachtens, dass eine wissenschaftliche Schlussfolgerung hinsichtlich der Fragestellung abgeleitet wird.

In Abhängigkeit vom Kontext, in dem das Gutachten erforderlich wird, unterscheidet man Gerichtsgutachten, Verwaltungsgutachten und Privatgutachten. Besonders häufig werden in der Medizin Verwaltungsgutachten erforderlich, bei denen es darum geht, dass eine Bürgerin Leistungen aus dem staatlichen Sozialversicherungs- bzw. Sozialversorgungssystem oder aus einer Haftpflichtversicherung für sich in Anspruch nehmen möchte (Eisenmenger et al., 2015; Lippert et al., 2015).

Die ein Gutachten in Auftrag gebenden Instanzen, z.B. Gerichte, Behörden und Versicherungsträger, wie z.B. Krankenkassen (Krankenversicherung) oder Berufsgenossenschaften (gesetzliche Unfallversicherung), wenden sich mit einer konkreten medizinischen Fragestellung an Sachverständige, die über jenes Fachwissen verfügen, das zur Beantwortung der Frage notwendig ist. Die große Mehrheit dieser Sachverständigen gehört der leitenden Ärzteschaft und somit dem öffentlichen Dienst an. Sie erledigt ihre gutachterliche Arbeit im Rahmen ihrer Dienstaufgaben oder als Nebentätigkeit (Lippert et al., 2015), wobei es ihre Aufgabe ist, medizinische Befunde unparteiisch in rechtliche Kontexte (wie z.B. das Sozialversicherungsrecht) einzuordnen. Juristisch gesehen gelten diese Sachverständigen als Gehilfinnen oder fachkundige Beraterinnen des Gerichts oder sonstiger Dritter (Deutsche Gesellschaft für Neurowissenschaftliche Begutachtung, 2013).

Ärztinnen haben hier also, wenn sie zusätzlich zu ihrer regulären Arbeit Gutachten erstellen, neben den kurativen Aufgaben an Kranken auch einen sozialmedizinischen Auftrag, der durch die Begutachtung im Krankheitsfall sowie die Gesundheitsförderung und Prävention bei

gesunden Personen erfüllt wird (Fritze und Fritze, 2012). Der Antrieb einer sachgemäßen Begutachtung ist allerdings keinesfalls das altruistische Bestreben, der Patientin helfen zu wollen. Stattdessen sind auf dem Boden wissenschaftlicher Evidenz eine oder mehrere konkrete Fragestellungen präzise und unbefangen zu beantworten, sofern das nach medizinisch-wissenschaftlichen Kriterien unter Anwendung gesicherten Wissens möglich ist.

Zur Erstellung und Weiterleitung eines Gutachtens muss die Gutachterin von der ärztlichen Schweigepflicht entbunden werden. Dies ist nach entsprechender Aufklärung nur durch eine schriftliche Einwilligung der Patientin möglich. Ärztinnen sollten dabei nicht gleichzeitig an einer durch sie behandelten Patientin gutachterlich tätig werden, da zum einen von Befangenheit auszugehen ist, zum anderen die Trennung zwischen für die Begutachtung erhobenen Daten und im Rahmen der Behandlung der Schweigepflicht unterstellten Informationen schwer fallen dürfte (Fritze und Fritze, 2012). Im Umkehrschluss bedeutet dies, dass im Regelfall Gutachterin und Patientin in der Gutachtensituation erstmalig aufeinandertreffen und die Ärztin im Gegensatz zur klassischen Diagnostik- oder Behandlungssituation kein Vorwissen über die Motivlage der Befragten hat und diese auch im weiteren Prozess nicht handlungsleitend in ihre Tätigkeit einbeziehen muss und darf.

Die formale Regelung der gerichtlich angeordneten Begutachtung sieht wie folgt aus: Mit der Approbation ist jede Ärztin verpflichtet, auf gerichtliche Aufforderung hin als Sachverständige Gutachten zu erstellen. Diese Aufgabe ist nicht delegierbar und ihre Verweigerung sowie ein Verstreichen der Frist können zur Verhängung von Ordnungsgeldern führen. Gutachtenaufträge durch gesetzliche oder private Versicherungsträger hingegen dürfen ohne Angabe von Gründen abgelehnt werden (Dreßing et al., 2018).

Ein Gutachten kann grundsätzlich sowohl schriftlich als auch mündlich vor Gericht erfolgen. Bei schriftlichen Gutachten unterscheidet man das Formulargutachten, das vor allem von privaten Unfall- und Lebensversicherungen eingeholt wird und aus geschlossenen Fragen besteht, und das freie Gutachten, das nach dem Ermessen der Gutachterin gegliedert und formuliert wird (Fritze und Fritze, 2012).

Häufig steht die Diagnostik, also das sichere Erkennen bzw. der Ausschluss krankhafter Veränderung, im Mittelpunkt der Begutachtung. Hierzu bedient sich die Gutachterin – wie auch bei der Behandlung von Patientinnen – diagnostischer Methoden, die sich im Groben aus der Befragung der betroffenen Person (Anamnese), der körperlichen Untersuchung und technisch-

apparativer Diagnostik zusammensetzt (Fritze und Fritze, 2012). Je nach Fachbereich und konkreter Fragestellung können die Anteile dieser drei Komponenten stark variieren. In ihrer *S2k-Leitlinie* zu allgemeinen Grundlagen der medizinischen Begutachtung definiert die Deutsche Gesellschaft für neurowissenschaftliche Begutachtung (2013) zwölf Anforderungen an die Gutachterin, die hier nur kurze Erwähnung finden sollen: Unparteilichkeit und Unabhängigkeit, Eigenverantwortlichkeit, Kompetenz, Beachtung der Rechtsgrundlage, vollständige Erfassung der Sachverhalte, Vermeidung von Interaktionsfehlern, Klarheit und gutachterliche Relevanz der Darstellungen und Aussagen, Beschränkung auf die von der Auftraggeberin gestellten Fragen, termingerechte Erstellung, Beachtung der Schweigepflicht, Beachtung der Rechte der zu Begutachtenden und die Beachtung der Aufbewahrungsfristen.

Die gesellschaftliche Bedeutung von Gutachten in der Medizin nimmt derzeit stetig zu. In den Jahren 1995 bis 2015 sind die Gesundheitsausgaben für Gutachten in Deutschland sukzessive von insgesamt 822 Millionen Euro auf 1,234 Milliarden Euro pro Jahr gestiegen (Statistisches Bundesamt, 2017). Allein für die gesetzliche Unfallversicherung werden jährlich ca. 100.000 Verwaltungsgutachten erstellt (Kater, 2011), die derzeitige Anzahl der sozialmedizinischen Stellungnahmen durch den Medizinischen Dienst der Krankenkassen gegenüber den Trägern von Kranken- und Pflegeversicherung ist um ein Vielfaches höher (Medizinischer Dienst des Spitzenverbandes Bund der Krankenkassen, 2019). Die Ursachen für diese Gesamtentwicklung sind vielfältig. Zum einen findet in Industrieländern als Folge zunehmender Behandlungsmöglichkeiten einer voranschreitenden Medizin eine Verlängerung der durchschnittlichen Lebensdauer (laut Fritze und Fritze 18 Monate je Dekade) und in der Folge eine stärkere Repräsentation chronischer gegenüber akuten Leiden statt, sofern Methoden der Prävention dem nicht entgegenwirken. Dieses Phänomen wird als Expansion der Morbidität bezeichnet (Olshansky et al., 1991; Graham et al., 2004; Buser et al., 2007). Darüber hinaus stellen technische Entwicklungen in allen Lebensbereichen Menschen vor Herausforderungen, denen ihre biologische Ausstattung nicht zwingend gewachsen ist. Ermüdung, Überforderung und ein ungesunder Lebensstil sind daraus resultierende Noxen, die das Auftreten von körperlichen sowie psychischen Erkrankungen und Unfallvoraussetzungen begünstigen (Fritze und Fritze, 2012). Diese Entwicklungen zeigen sich in der Sozialmedizin unter anderem in Form von mehr Begutachtungsanlässen.

Zusammenfassend kann festgehalten werden, dass die Begutachtung in der Medizin einen Tätigkeitsbereich mit besonderen Anforderungen und Zielstellungen darstellt, der klaren Regularien unterworfen ist und der im Zuge gesellschaftlicher Entwicklungen an Bedeutung gewinnt.

2.2 Begutachtung psychischer Erkrankungen

2.2.1 Relevanz und Unterschiede zu anderen Fachrichtungen der Medizin

Im medizinischen Fachgebiet der Psychiatrie, Psychotherapie und Psychosomatik hat das Aufgabenfeld der Gutachtenerstellung einen wesentlich höheren Stellenwert als in anderen Disziplinen (Nedopil et al., 2012). Dabei kommt der Begutachtung eine hohe Bedeutung für die Betroffenen, die Gesellschaft und auch für den Fachbereich Psychiatrie, Psychotherapie und Psychosomatik zu.

Fragestellungen an Gutachterinnen psychischer Erkrankungen beziehen sich letztlich auf die Einschränkung kognitiver und voluntativer Fähigkeiten und deren Auswirkung auf die rechtliche Verantwortlichkeit, die rechtliche Willenserklärung oder die berufliche Leistungsfähigkeit (Förster und Dreßing, 2015). Das daraus resultierende Grundproblem besteht darin, dass eine Patientin in einem derartigen Verfahren grundsätzlich besondere Interessen verfolgt und die Gutachterin sich unter Umständen im Spannungsfeld zwischen gegensätzlichen Erwartungen verschiedener Parteien sieht. Wichtig ist hier, dass es keineswegs Aufgabe der Gutachterin ist, die Rechtsfrage selbst zu beantworten, sondern das entscheidende Gericht in konkreten Fachfragen zu beraten. Diese Expertise zu beurteilen und daraus die juristischen Konsequenzen zu ziehen bleibt Aufgabe der Rechtsprechung (Förster und Dreßing, 2015). In etwa 95% der Fälle folgt das Gericht jedoch letztendlich dem Sachverständigengutachten (Schneider, Frister, et al., 2015).

Die Psychiatrie unterscheidet sich bei der Befunderhebung insofern grundlegend von anderen Fachrichtungen, als dass die Hauptinformationsquelle, auf der die Bestandsaufnahme einer Psychopathologie fußt, die mündlichen und schriftlichen Aussagen der Patientin sind. Es kann hier meist nur in geringerem Umfang als in den primär somatischen Fachrichtungen ein Befund zusätzlich apparativ oder in der körperlichen Untersuchung gesichert werden. Neben wenigen erfassbaren physikalischen Maßen wie der Konzentration eines Medikaments im

Körper (Erfassung der Therapietreue) oder physiologischen Parametern (z.B. bei der Konfrontation mit einem symptomauslösenden Stimulus) stellen psychometrische Testverfahren eine valide Methode dar, um Funktionseinschränkungen und das Erleben der Patientin zusätzlich zu operationalisieren. Dabei helfen solche klinischen Verfahren in der Begutachtung insbesondere bei der Schweregradbestimmung von Symptomen und der Leistungsmessung, ersetzen jedoch keinesfalls die Diagnostikkriterien nach *ICD-10* (Schneider, Frister, et al., 2015). Inhalt der Begutachtung psychischer Erkrankungen ist in Abhängigkeit von der konkreten Fragestellung meist ein allgemeiner psychopathologischer Befund, bei dem zum Beispiel das *AMDP-System* als Leitfaden der Befunderhebung dienen kann (Fähndrich und Stieglitz, 2018).

Einen entscheidenden Einfluss auf die Bedeutung von Gutachten psychischer Erkrankungen hat die allgemein zunehmende Relevanz solcher Krankheiten in der Bevölkerung. Epidemiologische Erhebungen haben für alle Erkrankungen aus dem Fachbereich der Psychiatrie eine 12-Monats-Prävalenz von ca. 31% ergeben (z.B. Jacobi et al., 2004). Die WHO-Publikation *Global Burden of Diseases* nennt währenddessen sieben psychische Erkrankungen mit einem Prävalenzanteil von 52% unter den 20 wichtigsten Ursachen für langfristige Behinderung in Industrienationen (Mathers et al., 2008; Whiteford et al., 2013). Auch über die Zeit nimmt die Bedeutung psychiatrischer Erkrankungen sukzessive zu. So hat sich die Zahl der Fehltage durch Arbeitsunfähigkeit aufgrund einer psychischen Erkrankung von 1997 bis 2012 fast verdreifacht (DAK Gesundheit, 2013). Hinsichtlich der dauerhaften Verminderung der Erwerbsfähigkeit zeigt sich dieselbe Tendenz: Erkrankungen der Psyche kommt die größte Bedeutung zu und sie führen mit 42% der zur Verrentung führenden Krankheiten das Feld gegenüber Krankheiten anderer Fachbereiche deutlich an (Bundespsychotherapeutenkammer, 2013). Darüber hinaus scheint sich diese Entwicklung zu verschärfen: Während Verrentungen aufgrund somatischer Erkrankungen über die Zeit rückläufig sind, haben sich von 1993 bis 2015 die jährlichen Rentenzugänge aufgrund psychiatrischer Diagnosen in Deutschland von 41.409 auf 74.234 fast verdoppelt (Stadtland und Nedopil, 2015; Deutsche Rentenversicherung, 2017). Es lässt sich parallel zu den ansteigenden Zahlen außerdem beobachten, dass die geforderten Gutachten in der Psychiatrie zunehmend differenziert sein müssen, um besser zwischen gerechtfertigten und ungerechtfertigten Ansprüchen unterscheiden zu können (Stadtland und Nedopil, 2015).

2.2.2 Juristische Kontexte der Begutachtung

Es gibt zahlreiche Situationen mit unterschiedlichsten (verwaltungs-)rechtlichen Konsequenzen, in denen Begutachtung erforderlich werden kann. Hier kann man grundsätzlich zwischen sozial-, zivil- und strafrechtlichen Kontexten unterscheiden. Hinsichtlich des zu begutachtenden Konstrukts lassen sich verschiedene Fragestellungen differenzieren: Schuldfähigkeit, Betreuungsrecht, Einwilligungs- und Geschäftsfähigkeit, Rentenversicherungs- und Entschädigungsrecht, Berufsunfähigkeit, soziales Entschädigungsrecht, gesetzliche Unfallversicherung etc. (Schneider und Weber-Papen, 2015). Die genannten Bereiche unterscheiden sich in rechtlichen Rahmenbedingungen und Anforderungen an die Gutachterin. Eine umfassende Darstellung dessen würde den Rahmen dieser Arbeit jedoch sprengen, wobei hier auf die ausführliche medizinische und juristische Fachliteratur zu diesem Thema verwiesen sei (z.B. Mehrhoff et al., 2012).

2.2.3 Psychiatrische Diagnosen im Fokus von Begutachtung

Grundsätzlich kann jede psychische Erkrankung im Kontext einer rechtlichen Frage das Objekt einer Begutachtung werden. Wird allerdings verglichen, welche Diagnosen z.B. als Grund für Arbeitsunfähigkeit (AU) am häufigsten gestellt werden, zeigt sich ein klares Bild: Mit einem Durchschnitt von 114,3 AU-Tagen je 100 Versichertenjahren liegen depressive Erkrankungen (depressive Episode *F32*, rezidivierende depressive Störung *F33* nach *ICD-10*) deutlich vor Reaktionen auf schwere Belastungen und Anpassungsstörungen (*F43*) mit 44, anderen neurotischen Störungen (*F48*) mit 21,7, anderen Angststörungen (*F41*) mit 16,6 und somatoformen Störungen (*F45*) mit 16,3 AU-Tagen je 100 Versichertenjahren (Marschall et al., 2016). Die Statistik für die Frühverrentung aufgrund psychiatrischer Erkrankungen zeigt ebenfalls, dass im Jahr 2012 mit 38,5% aller Fälle die depressiven Diagnosen (*F3*) als Ursache führten, gefolgt von neurotischen, Belastungs- und somatoformen Störungen (*F4*) mit 20,9% und Schizophrenie (*F2*) mit 10,3% der Berentungen (Bundespsychotherapeutenkammer, 2012). Auch in Hinblick auf die mit Krankheit verbrachten Lebensjahre (*years lived with disabilities - YLDs*) sind depressive Störungen mit 42,5% die häufigste ursächliche Pathologie innerhalb der Psychiatrie (Whiteford et al., 2013).

Depression spielt also als Erkrankungsgruppe in der Psychiatrie und auch für die Sozialleistungen eine große Rolle. Im Folgenden soll daher eine kurze Charakterisierung dieses Störungsbildes erfolgen.

Mit einer Punktprävalenz von 5% und einer Lebenszeitprävalenz von 10-20% stellt die unipolare Depression ein häufiges Krankheitsbild dar (Tölle und Windgassen, 2014). Gleichzeitig handelt es sich um eine schwere Erkrankung, die das Denken und Handeln der Betroffenen maßgeblich negativ beeinflusst. Hauptkennzeichen der Erkrankung sind eine generell negative Grundstimmung, Interessenverlust, Freudlosigkeit und Antriebsminderung. Zusätzlich kann es zu reduziertem Selbstwertgefühl, Störungen der Konzentration und Aufmerksamkeit, Schuldgefühlen, Gedanken an oder Versuch der Selbsttötung, Appetitminderung, einer pessimistischen Zukunftsperspektive und körperlichen Symptomen wie z.B. Schlafstörungen oder Appetitminderung kommen (Härter et al., 2016). Auch Suchtmittelkonsum als Folge depressiver Störungen ist beschrieben, wobei hier im Einzelfall auch eine umgekehrte Kausalität vermutet werden kann (Soyka et al., 1996).
Depression manifestiert sich in der Regel in Episoden (*F32*), die man nach *ICD-10* als leichte (*F32.0*), mittelschwere (*F32.1)* und schwere Episoden (*F32.2/3*) klassifiziert. Wenn diese Phasen depressiven Erlebens wiederholt auftreten, spricht man von einer rezidivierenden depressiven Störung (*F33*). Der Begriff Dysthymia (*F.34.1*) bezeichnet hingegen eine anhaltende depressive Verstimmung ohne den Schweregrad einer manifesten Depression. In der Regel dauert die Symptomatik hier jedoch länger als bei einer klassischen depressiven Episode an. Als affektive Störung kann Depression nicht nur unipolar, sondern auch wechselnd mit manischen Episoden als Teil einer bipolaren Erkrankung (*F31*) auftreten (Dilling und Freyberger, 2012).
Depression kann junge sowie alte Menschen, Frauen sowie Männer betreffen. Allerdings lässt sich feststellen, dass Frauen ungefähr doppelt so häufig die Diagnose Depression erhalten (Tölle und Windgassen, 2014).
Die WHO geht davon aus, dass Depression bis 2030 weltweit die Erkrankung sein wird, die das Leben am meisten beeinträchtigt und verkürzt – und zwar noch vor anderen „Volkskrankheiten“ wie der koronaren Herzkrankheit, dem Bluthochdruck oder Diabetes Mellitus (DGPPN et al., 2015). Daraus ergibt sich nicht nur großes persönliches Leid der Patientinnen, sondern ebenso eine weitreichende Konsequenz für die gesellschaftliche Bewertung dieser Erkrankung

sowie eine wachsende Bedeutung für die Begutachtung. Zusätzlich ist zukünftig auch ein größeres Angebot von Hilfsangeboten gefragt.

2.2.4 Verfälschungstendenzen in der Begutachtung psychischer Erkrankungen

Grundsätzlich ist in der Begutachtung psychischer Erkrankungen eine Reihe systematischer Verzerrungen denkbar, die bei der Bewertung des Befunds berücksichtigt werden müssen. Die Ursachen dieser Verzerrungen können sowohl in der Persönlichkeit der Diagnostikerin liegen (z.B. Wahrnehmungsabwehr, Interaktionsfehler wie vermutete Ähnlichkeit, Kontrastfehler und Übertragungsfehler), als auch der Begutachtungssituation oder der zu begutachtenden Person zugeschrieben werden (Westhoff und Kluck, 2014). Die Deutsche Gesellschaft für Neurowissenschaftliche Begutachtung (DGNB) weist in ihrer Richtlinie zur medizinischen Begutachtung ausdrücklich darauf hin, dass die Beziehung zwischen Gutachterin und Patientin ein Gefahrenpotenzial für Interaktionsfehler darstellt. Eine unfreundliche und ablehnende Haltung der Gutachterin beispielsweise könne bei der Begutachteten zu Verdeutlichungstendenzen führen, die dann als negative Antwortverzerrung missinterpretiert werden könnten (Deutsche Gesellschaft für Neurowissenschaftliche Begutachtung, 2013).

Im Hinblick auf diese Arbeit sei hier jedoch vor allem die Verzerrung der Beschwerden im Sinne einer nichtauthentischen Kommunikation durch die zu begutachtende Person herausgestellt. Es wird bei einer derartigen Verfälschung von berichteten Beschwerden zwischen einer unbewussten und einer bewussten Verzerrung unterschieden (Dreßing et al., 2018).

Bei einer *unbewussten Verzerrung* ist der Patientin die Verzerrung, geschweige denn die Ursache der Verzerrung, nicht klar. Man kann die unbewusste Verzerrung in die somatoforme (Motiv und Symptombildung unbewusst) und die artifizielle Störung (bewusstes Motiv nicht vermutet, Symptome jedoch bewusst hervorgerufen) untergliedern (Schiltenwolf und Henningsen, 2006; American Psychiatric Association, 2013; Dreßing et al., 2018). Von einer somatoformen Verzerrung spricht man dementsprechend beispielsweise bei einer im Vergleich zum Durchschnitt geringeren Symptomtoleranz, also einer tatsächlich schlimmer wahrgenommenen Symptomschwere (z.B. von Schmerzen).

Bewusste Verzerrung hingegen geschieht mit Vorsatz und dient der Verfolgung sekundärer Motive. Es ist den Personen durchaus klar, dass sie ihre Beschwerden nicht authentisch schildern.

Dabei ist häufig ein sekundärer Krankheitsgewinn, beispielsweise als Kranke nicht arbeiten zu müssen oder besonders liebevoll umsorgt zu werden, ein Motiv bewusster Verzerrung (Buser et al., 2007). Solch ein sekundärer Benefit kann sowohl durch bewusste, als auch durch unbewusste Prozesse angestrebt werden. Im Allgemeinen tritt die bewusste Täuschung in der Medizin jedoch seltener auf als meist angenommen (Schiltenwolf und Henningsen, 2006). Für den Bereich der Begutachtung stellt sie jedoch eine häufige und oft übersehene Fehlerquelle dar. Das folgende Kapitel soll daher einen Überblick über die Formen und die Möglichkeiten der Diagnostik bewusster Verzerrung im Kontext der gutachterlichen Tätigkeit geben.

2.3 Bewusste Täuschung in der Begutachtung

2.3.1 Bedeutung für die Psychiatrie und Begriffsdefinitionen

Wie in Kapitel 2.2.1 bereits erwähnt, unterscheidet sich die Psychiatrie insofern grundsätzlich von anderen Bereichen der Medizin, als dass die subjektiven und schwer überprüfbaren Schilderungen der Patientin den entscheidenden Anteil erhobener Befunde darstellen. Während man außerdem im Kontext der medizinischen Behandlung davon ausgehen kann, dass Beschwerden im Sinne einer optimalen Diagnostik und Therapie weitestgehend authentisch geschildert werden, muss im Kontext der Begutachtung die Möglichkeit einer absichtlichen zweckgerichteten Täuschung (engl.: *deception, faking*) oder anderer Verfälschungen in Betracht gezogen werden. Die Ursache hierfür kann im Rahmen bewusster Verzerrung in den möglichen Konsequenzen des Gutachtens für die Betroffene gesehen werden (Schmidt et al., 2011). Es muss dabei davon ausgegangen werden, dass Personen in Gutachtensituationen bestimmte Interessen hinsichtlich des Ergebnisses verfolgen. Verfälschungen lassen sich daher vor allem im Kontext von Schuldfähigkeitsbeurteilung und Verrentungsverfahren beobachten (Schneider, Frister, et al., 2015). Hier kann es durch den Begutachtungsbefund einer besonders schweren Erkrankung zu einem sekundären Krankheitsgewinn, z.B. einer verminderten Schuldfähigkeit bzw. einer Berentung, kommen (Faller und Lang, 2010).

Merten fasst diese Phänomene unter dem Sammelbegriff *negative Antwortverzerrung* zusammen, wobei hierbei ungeachtet einer Motivationslage die nicht-valide Symptomdokumentation gemeint ist (Merten, 2011).

Eine mögliche Strategie, um in der Begutachtung besonders belastet zu erscheinen, kann z.B. die *Simulation* (engl.: *malingering*) von Beschwerden sein. Sie wird im medizinischen Kontext als die „bewusste u. absichtliche, evtl. betrügerische Vortäuschung von – meist funktionellen – Krankheitszuständen mit bestimmter Zweckabsicht (z.B. Rentenbegehren)" (Reiche et al., 2003, S. 1704) definiert. Eine simulierende Patientin schildert hierbei also Symptome, die sie nicht hat, um bewusst das Begutachtungsergebnis zugunsten eines sekundären Krankheitsgewinns zu manipulieren (Dreßing et al., 2018).

Als *Aggravation* hingegen wird die „unangemessen übertriebene, unter Umständen zweckgerichtete Präsentation von Schmerzen, Symptomen oder Einschränkungen" (Margraf, 2019) einer Patientin definiert. Die Beschwerden existieren im Ansatz also tatsächlich, werden jedoch in ihrem Ausmaß bewusst übertrieben (Buser et al., 2007; Dreßing et al., 2018). Bei der Lektüre englischsprachiger Fachliteratur fällt auf, dass hier die Konstrukte Simulation und Aggravation nicht differenziert, sondern als *Malingering* zusammengefasst werden. Auch die deutsche Version des *DSM-5* trifft diese Unterscheidung nicht (American Psychiatric Association, 2013). Der Vollständigkeit halber sei hier auch ergänzend die *Dissimulation* erwähnt, bei der es sich um das „bewusste Verheimlichen von Krankheitssymptomen oder die bewusst herunterspielende Darstellung von (psychischen) Erkrankungen und Beschwerden" (Stadtland und Nedopil, 2015) handelt, also eine Verzerrung der geschilderten Symptomatik in Richtung geringerer Symptomschwere. Dabei wird ein sekundärer Benefit durch eine Vertuschung bzw. Abschwächung von Krankheitssymptomen angestrebt.

Es stellt sich nun also aus Sicht der Gutachterinnen die Frage, inwieweit die Beschwerdenvalidität, also die einer Symptomdarstellung zugrundeliegende Authentizität, überprüft werden kann (Merten, 2011). Die zu diesem Zwecke entwickelten Strategien und Testverfahren sollen im Folgenden kurz beschrieben werden (für eine ergänzende Übersicht vgl. Schmidt et al., 2011).

2.3.2 Messung und Beurteilung negativer Antwortverzerrung

In der Literatur werden bereits seit langem auf Verfälschungstendenzen hinweisende Verhaltensmuster und Testbefunde beschrieben (z.B. Jones et al., 1917). Bestimmte klinische Auffälligkeiten bei Testungen gelten daher hinsichtlich systematischer Verzerrung als verdächtig. So werden z.B. Inkonsistenzen zwischen verschiedenen Untertests von Intelligenzmessungen und zwischen demonstrativ unbeholfenem Verhalten während der Testung und unbeeinträchtigtem Lösen von Aufgaben außerhalb des eigentlichen Testverfahrens (z.B. im Alltag) dokumentiert (Rist und Dirksmeier, 2001). Außerdem finden sich in der Literatur konkrete Phänomene als Hinweise für Übertreibung: ein Nebeneinander übertriebener Genauigkeit und mangelnder Präzision in der Symptombeschreibung (v.a. in Bezug auf den zeitlichen Verlauf der Beschwerden), rasches Auftauchen und Verschwinden der Störungen ohne graduierte Veränderung, Dominanz von Positiv- gegenüber Negativsymptomen in der Darstellung psychotischer Beschwerden, kaum subtile Symptome, Widersprüche und Inkonsistenzen in der wiederholten Befragung und viele mehr (Rogers, 1988; zitiert nach Birck, 2002). Merten und Dohrenbusch geben eine ausführliche Übersicht über aus neuropsychologischer Perspektive erarbeitete strategische Ansätze zur Diagnostik der Beschwerdenvalidität (2016, Tabelle 7-3).

Ohne die Anwendung standardisierter Methoden ist jedoch von einer geringen Validität und Reliabilität der Begutachtung von Beschwerdenauthentizität auszugehen (Schmidt et al., 2011). Untersuchungen diesbezüglich ergeben für die unstandardisierte klinische Einschätzung, selbst unter Berücksichtigung der oben erwähnten Hinweise, lediglich Ergebnisse knapp über dem Zufall (Hall und Poirier, 2000). Aus diesem Grund wurde die Entwicklung und Verbesserung von Beschwerdenvalidierungstests, also Testverfahren zur Klärung der Authentizität von geschilderten Symptomen, zuerst in Nordamerika (siehe hierzu z.B. Hall und Poirier, 2000; Rogers, 2008) und etwa seit der Jahrtausendwende auch im deutschen Sprachraum vorangetrieben. Sie sollen eine psychometrische Überprüfung der Beschwerdenvalidität unter wissenschaftlichen Standards, also unter Berücksichtigung überprüfbarer Testgütekriterien wie *Validität*, *Reliabilität* und *Objektivität*, ermöglichen. Ein derartiger Test ist valide, wenn er tatsächlich das Konstrukt misst, das durch ihn evaluiert werden soll. Die Reliabilität ist gegeben, wenn wiederholte Messungen gleiche Ergebnisse erbringen und ein Test gilt dann als objektiv, wenn die Ergebnisse unabhängig von der jeweiligen Untersucherin gleich sind (Buser et al., 2007; Faller und Lang, 2010). Des Weiteren können bei standardisierten Testverfahren *Sensitivität*,

Spezifität und die *positiven* sowie *negativen Prädiktionswerte* bestimmt und unter Veränderung der *Cut-off-Werte* des Tests angepasst werden. Es muss hier unbedingt berücksichtigt werden, dass im Falle eines Testverfahrens mit Screeningcharakter – sprich einer hohen Sensitivität auf Kosten der Spezifität und des positiv-prädiktiven Werts (z.B. Coin-in-the-Hand-Test, SFSS – siehe hierzu Kapitel 2.3.3) – weitere sichernde Diagnostik mit höherer Spezifität folgen muss. Nur so kann verhindert werden, dass eine authentische Patientin fälschlicherweise als Übertreibende klassifiziert wird (zu ausführlichen epidemiologischen Begriffsdefinitionen siehe z.B. Faller und Lang, 2010).

Trotz bereits bestehender standardisierter Beschwerdenvalidierungstests sind die Empfehlungen in den Richtlinien der Fachgesellschaften noch zurückhaltend. Dieser Umstand ist darauf zurückzuführen, dass die Entwicklung und Validierung dieser Verfahren momentan noch eine gewisse Herausforderung darstellt und die Testgüte von publizierten Beschwerdenvalidierungstests nicht durchgehend als ausreichend beurteilt wird. Zudem mag eine grundsätzliche Zurückhaltung der Fachgesellschaften gegenüber psychologischen Testverfahren aktuell noch eine Rolle spielen. Der Leitfaden *Begutachtung der beruflichen Leistungsfähigkeit bei psychischen und psychosomatischen Erkrankungen* der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) konstatiert jedoch klar, dass berichtete Beschwerden „grundsätzlich mit den unterschiedlichen zur Verfügung stehenden geeigneten Methoden validiert werden müssen" (Schneider et al., 2016, S. 480 f.) und listet unter weiteren Methoden auch Symptomvalidierungstests auf. Die DGNB empfiehlt in ihrer *S2k-Leitlinie* von 07/2013 ebenfalls ausdrücklich den Einsatz von Beschwerdenvalidierungstests zur Überprüfung der Authentizität beklagter Defizite. Gleichzeitig grenzt sie ein, dass solche Verfahren keinen Vollbeweis für Manipulationsversuche bieten könnten, sondern immer im Kontext von Verhaltensbeobachtung und qualitativer sowie quantitativer Analyse von Untersuchungsergebnissen interpretiert werden müssten (Deutsche Gesellschaft für Neurowissenschaftliche Begutachtung, 2013). Auch in der *Stellungnahme der Deutschen Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde* (DGPPN) zur Anwendung von Beschwerdenvalidierungstests in der psychiatrischen Begutachtung wird betont, dass diese „kein obligates Qualitätsmerkmal psychiatrischer Gutachten darstellt" (Dreßing et al., 2011, S. 1).

Von diesen Standpunkten kann abgeleitet werden, dass ein Einsatz von Verfahren zur Beschwerdenvalidierung zwar einen sinnvollen Bestandteil von Gutachten psychischer

Erkrankungen darstellt, die Ergebnisse jedoch keineswegs als alleiniges Urteil über die Beschwerdenvalidität missbraucht werden sollten. Insbesondere die Beurteilung einer detektierten negativen Antwortverzerrung als Simulation bzw. Aggravation kann nur in einem zweiten Schritt unter Berücksichtigung des klinischen Kontextes erfolgen (Dreßing et al., 2018). Darüber hinaus ist die Entwicklung valider, reliabler und objektiver Beschwerdenvalidierungstests für den deutschen Sprachraum gerade erst im Gange und sicherlich noch nicht abgeschlossen. Dies hebt die Notwendigkeit von Forschung im Bereich der Beschwerdenvalidierung hervor, wodurch eine bessere, wissenschaftlich fundierte Methodik verfügbar gemacht würde.
Im Folgenden sollen einige häufig verwendete und bekannte Beschwerdenvalidierungstests kurz beschrieben werden. Eine ergänzende Auflistung findet man z.B. bei Merten und Dohrenbusch (2016).

2.3.3 Beschwerdenvalidierungstests

Als Beispiel einfacher, effektiver und am Patientenbett einsetzbarer Tests zur Detektion aggravierender und simulierender Patientinnen kann der *Coin-in-the-Hand-Test* genannt werden. Zuerst nimmt die Diagnostikerin unter Beobachtung der Patientin eine Münze in eine Hand und schließt diese. Dann wird die Patientin aufgefordert, nach zehn Sekunden des Zurückzählens bei geschlossenen Augen zu benennen, in welcher Hand die Untersucherin die Münze genommen hat. Es handelt sich also um eine ausgesprochen einfache Aufgabe, die zehn Mal wiederholt wird. Die Ergebnisse aggravierender Probandinnen waren signifikant schlechter (durchschnittlich 4,1 von 10 richtigen Tipps) als die authentischer Personen (durchschnittlich 9,95 von 10 richtigen Tipps) (Kapur, 1994; Hall und Poirier, 2000). Das hier angewandte *Prinzip der verdeckten Leichtigkeit* ist eine übliche Strategie, um Aggravation und Simulation zu erkennen (Baker et al., 1993). Probandinnen erhalten dabei Aufgaben, die auch bei stärkster Symptomschwere in der Regel problemlos bewältigt werden können, jedoch ggf. je nach Testverfahren auf den ersten Blick sehr schwer erscheinen. Ein weiteres Testverfahren, das dieses Prinzip nutzt, ist der *Word-Memory-Test* (*WMT*; Green, 2005). Die Probandinnen werden vor die Aufgabe gestellt, sich 20 Wortpaare zu merken und daraufhin in zwei Durchgängen zu reproduzieren. Diese Wortpaare lassen sich jedoch im Allgemeinen gut assoziieren. Hierbei kommt es für die Auswertung sowohl auf die richtige Nennung der Worte als auch auf die

Übereinstimmung der Antworten in beiden Abfragen an. In einer ersten Validierungsstudie wurden eine hohe Sensitivität sowie Spezifität für die Detektion negativer Antwortverzerrung ermittelt (Brockhaus und Merten, 2004).

Andere Testverfahren zielen auf eine differenziertere Operationalisierung der Beschwerdenauthentizität ab und dauern daher wesentlich länger. Das in den USA am meisten genutzte Verfahren ist der bereits in den 1940er Jahren erstmals veröffentlichte Persönlichkeitsfragebogen *Minnesota Multiphasic Personality Inventory* (*MMPI*; Hathaway und McKinley, 1943). Hierbei handelt es sich nicht um einen primär für Beschwerdenvalidierung konzipierten Test, sondern um ein Multiskaleninventar zu einer Vielzahl klinischer Fragestellungen. In insgesamt 566 Items werden verschiedene Dimensionen der Persönlichkeit abgefragt und in Skalen zusammengefasst, von denen vor allem die *Lügenskala* und die *Seltenheitsskala* für die Beurteilung der Beschwerdenvalidität relevant sind (Amelang und Schmidt-Atzert, 2006). Es liegt inzwischen eine deutsche Version des revidierten Selbstbeurteilungsverfahrens vor (*MMPI-2*; Hathaway et al., 2000), die jedoch aufgrund unzureichender psychometrischer Qualität häufig kritisiert wurde und sich in Deutschland nicht wirklich etablieren konnte (Walter et al., 2012).

Auch andere Persönlichkeitstest wie z.B. die Revision des *Freiburger Persönlichkeitsinventars* (*FPI-R*; Fahrenberg et al., 2010) bieten Skalen zur Offenheit der Probandin an. Diese können allerdings maximal Hinweise auf negative Antwortverzerrung geben und müssen in dem Fall dringend durch Beschwerdenvalidierungstests abgesichert werden, da sich gezeigt hat, dass die *FPI*-Offenheitsskala nicht mit Ergebnissen etablierter Beschwerdenvalidierungstests wie des *SFSS* und des *WMT* korreliert (Merten et al., 2007).

Ein weiterer Test im Selbstbeurteilungsformat ist das *Structured Inventory of Malingered Symptomatology* (*SIMS*; Smith und Burger, 1997), das in seiner deutschen Übersetzung *Strukturierter Fragebogen Simulierter Symptome* (*SFSS*; Cima et al., 2003) heißt. Hierfür ergaben sich Spezifitäts-Werte von 80-90%, das bedeutet, dass über 10% der authentischen Probandinnen nicht eindeutig als solche klassifiziert werden konnten. Diese Ergebnisse sind zwar im Sinne eines Screenings tolerabel, jedoch für die Begutachtung als nicht ausreichend zu bewerten (Schmidt et al., 2011).

Schließlich existiert seit Kurzem im Englischen eine überarbeitete Version des *Structured Interview of Reported Symptoms* (*SIRS-2*; Rogers et al., 2010). Dieser Beschwerdenvalidierungstest im Fremdbeurteilungsformat (strukturiertes Interview) geht über die Möglichkeiten der oben

genannten Fragebögen bzw. Leistungstests hinaus. Zusammenfassend kann man beim *SIRS-2* von einem hohen methodischen und konzeptionellem Standard mit guter Validität und Anwendbarkeit für den Kontext der Begutachtung psychischer Erkrankungen ausgehen (Schmidt et al., 2011).

Da die deutsche Version des *SIRS-2* (Schmidt et al., 2019) in dieser Studie zur Evaluation der Beschwerdenvalidität verwendet wird, wird es im Methodenteil (Kapitel 0) ausführlich beschrieben.

2.3.4 Einflüsse auf negative Antwortverzerrung

Es stellt sich nun die Frage, welche möglichen Einflussfaktoren negativ auf die Authentizität der Symptomdarlegung einwirken.

Hier sei neben den bereits genannten motivationalen Verzerrungen (siehe Kapitel 2.3.1) vor allem die bei psychiatrischen Patientinnen zum Negativen veränderte Selbstwahrnehmung herausgestellt. Im Rahmen von verzerrter Kognition bei psychischen Erkrankungen wie Depression und Persönlichkeitsstörungen ist eine Ausweitung krankhaft veränderter Sichtweisen auf die Wahrnehmung der eigenen Person möglich, wie es in Aaron Becks *Kognitiver Triade* beschrieben ist (Tölle und Windgassen, 2014). Beck sieht zentrale Antwortverzerrungen gerade bei Patientinnen mit Depressionen in der negativen Beurteilung ihrer eigenen Person, der Umwelt und ihrer Zukunft (1979). So kommt es unter verschiedenen kognitiven Abweichungen von der Norm auch zu übertriebenen Wahrnehmungen der eigenen Symptomschwere. Eine derart abweichende Selbstwahrnehmung kann auch in der Begutachtung zu einer Fehldarstellung der Symptomschwere führen. In der jüngeren Fachliteratur zu Symptomauthentizität ist ebenfalls beschrieben, dass bestimmte Erkrankungen verstärkt mit negativer Antwortverzerrung einhergehen. So wurde z.B. herausgestellt, dass Depression und ein geringes Selbstwertgefühl bei Patientinnen als Bewältigungsmechanismus unbewusst eine Selbst- und Fremdtäuschung auslösen können (Hall und Poirier, 2000), was sich durch die oben erläuterte kognitive Triade erklären lässt. Allerdings liegt der verzerrten Schilderung in diesem Fall keinesfalls eine bewusste Täuschung, wie in Kapitel 2.3.1 beschrieben, zugrunde, was die Gutachterin vor die Herausforderung stellt, diese unterschiedlichen Quellen nichtauthentischer Beschwerdendarlegung zu differenzieren.

Zusammenfassend kann daher festgestellt werden: Psychische Erkrankungen sind unter anderem durch kognitive Veränderungen gekennzeichnet, die sich auf die Sicht auf die eigene Person, die Zukunft und die Umwelt beziehen. Im Sinne dieser kognitiven Triade ist es bei Depressionen also wahrscheinlich, dass persönliche Eigenschaften, Fähigkeiten und Merkmale negativer bewertet werden als es objektiv der Fall ist (Beck, 1979). Gleiches gilt für die Einschätzung der eigenen Zukunftsperspektive und der Absichten von Menschen im sozialen Umfeld der Person.

In einer Gutachtensituation sollte dieses Wissen integriert werden. Die Gutachterin muss in der Lage sein, im Rahmen der Erkrankung zu erwartende subjektiv verzerrte Symptomschilderungen von motivational bedingten Täuschungen abzugrenzen. Für diesen Zweck ist es daher notwendig, differenzierte Instrumente zu verwenden, die den komplexen Gegebenheiten der Erkrankungen gerecht werden und die die so erlangten Befunde stets im Kontext der klinischen Einschätzung von Patientin und Diagnose beurteilen und bewerten. Beschwerdenvalidierungstests müssen diese Fehlerquelle berücksichtigen und im Kontext der Integration des Testergebnisses in die gesamte Begutachtung den Versuch einer trennscharfen Differenzierung unbewusster und krankheitsbedingter von vorsätzlicher und nutzenorientierter negativer Antwortverzerrung ermöglichen. Dieser Erwartungshorizont soll auch für die Testentwicklung und Testverbesserung berücksichtigt werden und geeignete Verfahren zur Unterstützung der Gutachterin in diesem Prozess sollten verfügbar gemacht werden.

2.3.5 Einfluss intellektueller Fähigkeiten auf negative Antwortverzerrung

Wie im vorangegangenen Abschnitt geschildert, beinhaltet insbesondere die depressive Symptomatik unter anderem eine systematische Unterschätzung der eigenen Ressourcen und Fähigkeiten einer Betroffenen. Diese wahrgenommene Reduktion der eigenen Leistungsfähigkeit ist bei manifesten Depressionen zusätzlich mit tatsächlichen Einschränkungen kognitiver Leistungen assoziiert.

Der Zusammenhang zwischen Depression und kognitiver Leistungsminderung sowie zwischen einem erhöhten Depressionsrisiko bei vergleichsweise niedrigem intellektuellem Ausgangsniveau ist vielfach untersucht. Es ist hierbei hinlänglich bekannt, dass ein niedriger IQ im Allgemeinen mit Psychopathologien vergesellschaftet ist (siehe hierzu auch Kapitel 2.4.4).

Im Speziellen ist dieser Zusammenhang zwischen niedriger Intelligenz, emotionalem Dysstress und dem Erkrankungsrisiko für Depressionen in zahlreichen Studien belegt worden (z.B. Navrady et al., 2017). Die Kausalität ist dabei bidirektional zu vermuten: Menschen mit niedrigem IQ fehlen eher die Ressourcen, dem Aufkommen bestimmter psychischer Belastungen bewusst entgegenzuwirken und Copingstrategien zu entwickeln bzw. auszuführen (Zammit et al., 2004), während Depression aufgrund der Hemmung kognitiver Funktionen das messbare Intelligenzlevel, und zwar vor allem Ergebnismaße der fluiden Intelligenz wie z.B. den Handlungs-IQ, signifikant senkt (Sackeim et al., 1992). Im Umkehrschluss ist also grundsätzlich davon auszugehen, dass eine erhöhte Intelligenz mit einem niedrigeren Depressionsrisiko assoziiert ist.

Vor dem Hintergrund, dass – wie in Kapitel 2.3.4 ausgeführt – bei psychischen Erkrankungen wie Depressionen von einem negativ verzerrten Antwortverhalten auszugehen ist, könnte nun geschlussfolgert werden, dass auch eine Assoziation von geringerem IQ und negativer Antwortverzerrung vorliegt – natürlich neben den oben genannten motivationalen Ursachen von bewusster negativer Antwortverzerrung. Hier existiert jedoch kaum wissenschaftliche Evidenz. Es zeigte sich z.B. lediglich, dass simulierende und aggravierende Probandinnen schlechtere Ergebnisse in Tests zur kristallinen Intelligenz produzierten als die tatsächlich stark betroffene Kontrollgruppe (Rist und Dirksmeier, 2001), doch man kann hier auch den Schluss ziehen, dass bei geringerer kristalliner Intelligenz die Detektion der negativen Antwortverzerrung durch Beschwerdenvalidierungstests erleichtert ist.

Unbeantwortet bleibt hierbei außerdem die Frage, ob auch die Messbarkeit negativer Antwortverzerrung von der Intelligenz der Patientin abhängt. Bevor dieser Frage jedoch nachgegangen wird, soll im folgenden Kapitel vorerst ein Überblick des theoretischen Hintergrunds zum Thema Intelligenz erarbeitet werden.

2.4 Intelligenz

2.4.1 Definition des Intelligenzbegriffs

Intelligenz ist ein in der Literatur auf verschiedene Arten und nicht endgültig definierter Begriff. So beschreiben Lexika und Fachbücher der Psychologie Intelligenz häufig als Gruppe von Fähigkeiten, die die Bewältigung neuartiger Situationen durch problemlösendes Verhalten ermöglichen oder allgemeiner als Sammelbegriff für die kognitive Leistungsfähigkeit des Menschen (Faller und Lang, 2010; Fleischhacker und Hinterhuber, 2012).

Da die für problemlösendes Verhalten essentiellen kognitiven Fähigkeiten schwer zu fassen und abhängig vom Kontext des zu lösenden Problems sind, hat sich Edwin Borings tautologische, operationale Prägung des Begriffs Intelligenz in der Psychologie als das „was Intelligenztests messen" etabliert (Boring, 1923; Buser et al., 2007). Eine wichtige Schlussfolgerung daraus ist, dass Intelligenz stark von den konkreten zur Messung eingesetzten Testanforderungen abhängt und keine im naturwissenschaftlichen Sinne direkt messbare Größe ist (Schneider, Niebling, et al., 2015). Vielmehr muss Intelligenz als ein zeitlich relativ konstantes theoretisches Konstrukt betrachtet werden, das bestimmte Leistungen ermöglicht und nur durch die Messung dieser Leistung abgeschätzt werden kann (Faller und Lang, 2010).

2.4.2 Intelligenzkonzepte

Zur Beschreibung der Intelligenz hat es seit Beginn des 20. Jahrhunderts verschiedene Modelle gegeben, von denen die meistzitierten und für diese Arbeit relevanten hier kurz aufgeführt werden sollen.

Das 1904 von Charles Spearman formulierte *General-* oder *Zweifaktorenmodell* beschreibt einen Generalfaktor (g-Faktor) und mehrere verschieden stark ausgebildete spezifische Faktoren (s-Faktoren) für Intelligenz. Konkrete in Lebenssituationen oder psychometrischen Testungen erforderliche Fähigkeiten greifen dabei immer auf die Kombination des generellen mit einem spezifischen Intelligenzfaktor zurück (Spearman, 1904). In den 1920er und -30er Jahren bildete sich – vorangetrieben durch Edward Lee Thorndike und Louis Leon Thurstone – eine alternative Ansicht heraus, die von multiplen und gleichrangigen Faktoren ausgeht. Beim Lösen intelligenzfordernder Aufgaben kommt demnach mindestens einer dieser Faktoren, teilweise

auch mehrere, zum Einsatz (Rost, 2013). Thurstone formulierte auf Basis breiter Untersuchungen an Studierenden sieben *Primärfaktoren*, die die grundlegenden intellektuellen Fähigkeiten abdecken sollten (Thurstone, 1938).

Wenige Jahre darauf publizierte Raymond Cattell ein neues Gedankenmodell, das mentale Kapazität in zwei Bereiche einteilt: fluide und kristalline Intelligenz. Die *fluide Intelligenz* meint hier das abstrakte logische Problemlösungsvermögen, das zum Erschließen und der Analyse neuer Konzepte und Sachverhalte notwendig ist. Sie kann auch durch den Begriff „Denkfähigkeit" charakterisiert werden und steigt im Laufe der Kindheit und Jugend kontinuierlich an, um dann im Erwachsenenalter flach und schließlich im höheren Alter steiler abzufallen. Als die *kristalline Intelligenz* hingegen beschreibt Cattell die problemlösende Kompetenz, die aus Erlerntem und Erfahrung resultiert. Sie kann vereinfacht auch als „Wissen" bezeichnet werden und steigt im Laufe der Kindheit und Jugend rapide und im weiteren Verlauf eines Lebens kontinuierlich leicht an (Cattell, 1941; Preckel und Brüll, 2008; Rost, 2013). Der Vollständigkeit halber soll erwähnt werden, dass in neueren Intelligenztheorien vor allem versucht wird, die Vorgänge zu analysieren, die intelligentem Verhalten in allen denkbaren Lebenssituationen zugrunde liegen und dies nicht nur unter Testbedingungen. So hat beispielsweise Howard Gardner zu Beginn der 1980er Jahre durch seine *Theorie der multiplen Intelligenzen* ein Modell erschaffen, das auch Fähigkeiten in den Intelligenzbegriff miteinbezieht, die zuvor als unabhängig davon angesehen wurden, wie z.B. eine musikalische, eine körperlich-kinästhetische und eine interpersonale Intelligenz. Diese können aber laut Gardner ebenso als Prädiktoren für ein in weiten Bereichen erfolgreiches Leben herangezogen werden. Mit der *triarchischen Intelligenztheorie* von Robert Sternberg wird stattdessen durch eine Dreiteilung intellektueller Fähigkeiten in eine analytische, eine kreativ-kognitive und eine praktisch-kognitive Intelligenz versucht, auch schöpferische Fähigkeiten und praktische, anwendungsnahe Problemlösungen im Alltag in den Intelligenzbegriff zu inkludieren (Buser et al., 2007; Rost, 2013).

2.4.3 Intelligenzdiagnostik

Zur Operationalisierung von Intelligenz wurden ab dem Ende des 19. Jahrhunderts evidenzbasierte Intelligenztests entwickelt. Ziel war es, eine Messung intellektueller Fähigkeitspotenziale unabhängig von außerhalb des Tests gezeigter Leistung vorzunehmen, um auf dieser

Basis Schätzungen bezüglich der künftigen Leistungsfähigkeit zu machen. So beabsichtigte man, bei der Auswahl von Offiziersanwärtern oder der Voraussage von Schullaufbahnen objektive Vergleichswerte zu produzieren (Linden et al., 2015). Die Messverfahren sollten den Testgütekriterien Validität, Objektivität und Reliabilität entsprechen (zur Begriffserklärung siehe Kapitel 0). Alfred Binet prägte die frühen Intelligenztests maßgeblich, indem er die intellektuelle Leistung von Kindern mit standardisierten Testverfahren maß und ihnen auf dieser Grundlage ein Intelligenzgrundalter zuwies, das im Durchschnitt und beim normal begabten Kind dem Lebensalter entsprach, bei intelligenteren Kindern allerdings höher und bei unterdurchschnittlich intelligenten Kindern niedriger war. Vor dem Hintergrund der Annahme, dass Intelligenz normalverteilt sei, prägte William Stern 1912 weiterführend den Begriff des *Intelligenzquotienten* (*IQ*). Der IQ wird errechnet, indem man das Intelligenzgrundalter durch das Lebensalter dividiert und mit 100 multipliziert. Im Durchschnitt beträgt der IQ dabei 100, eine Standardabweichung 15 Punkte. Auch heute noch wird der Intelligenzquotient als Gesamtgröße für Intelligenz verwendet (z.B. bei den Wechsler-Intelligenztests), allerdings nicht in Bezug auf das Alter, sondern als Abweichungsvariable um das Verhältnis zwischen der Leistungsfähigkeit des Einzelnen und der Leistungsfähigkeit der Gesamtheit zu charakterisieren. Es sind hier jedoch auch andere Skalierungen wie z.B. mithilfe von *z-Werten* (Mittelwert = 0, Standardabweichung = 1), *c-Werten* (Mittelwert = 5, Standardabweichung = 2) oder bei nicht gegebener Normalverteilung anhand von *Prozentrangwerten* möglich und üblich (Holling et al., 2004; Faller und Lang, 2010).

Intelligenztests sind meist, jedoch nicht immer, auf der Basis konkreter Intelligenzmodelle entworfen (Buser et al., 2007). Die folgende Übersicht soll die bekanntesten Intelligenztests kurz beschreiben und orientiert sich dabei in der Auswahl und der Reihenfolge an zwei Umfragen aus den 1990er Jahren, in denen die am häufigsten verwendeten Testverfahren ermittelt wurden (Schorr, 1995; Steck, 1997).

Eine Reihe viel verwendeter Intelligenztests basiert auf den Arbeiten von David Wechsler. Dieser publizierte erste Intelligenzskalen bereits in den 1930er Jahren und schuf 1955 mit der *Wechsler Adult Intelligence Scale* (*WAIS*; Wechsler, 1955) ein Testverfahren, dessen Weiterentwicklungen und Übersetzungen sich als die meistbenutzten und -zitierten Intelligenztests in der zweiten Hälfte des 20. Jahrhunderts herausgestellt haben (Schorr, 1995; Steck, 1997). Sie wurden im Verlauf der folgenden Jahrzehnte mehrmals auf Basis neuer

Normierungsstichproben für die englischsprachige Bevölkerung wiederaufgelegt, bis schließlich vor wenigen Jahren die 4. Revision erschien (*WAIS-IV*; Wechsler, 2012). Auch deutsche Adaptationen sind im Verlauf erschienen, von denen der *Hamburg-Wechsler-Intelligenztest für Erwachsene* (*HAWIE;* Wechsler, 1964) und die Revision (*HAWIE-R;* Tewes, 1991) sowie die deutsche Version des *WAIS-IV* (Petermann, 2012), zu nennen sind. Auch für das Intelligenzassessment bei Kindern zwischen 6 und 16 Jahren sind diverse Versionen (z.B. *HAWIK-III*; Tewes und Rossmann, 2002) erschienen (Holling et al., 2004; Amelang und Schmidt-Atzert, 2006). Wechsler orientierte sich bei der Konzeption seines Intelligenztests an dem Generalfaktor-Modell von Spearman und versuchte, über verschiedene Aufgaben eine allgemeine Intelligenz im Sinne des *g-Faktors* zu ermitteln. Aufgrund dieser undifferenzierten Betrachtung von Intelligenz ließ sich das Testergebnis also in einem einzigen Wert ausdrücken, dem oben genannten Intelligenz-Quotienten (Buser et al., 2007). Jene IQ-Darstellung wird aus diesem Grund häufig mit Wechsler in Verbindung gebracht. Der *HAWIE* und seine Revisionen teilen sich in einen Verbalteil und einen nichtverbalen Teil auf und umfassen folgende Untertests: Allgemeinwissen, Zahlen nachsprechen, Wortschatztest, rechnerisches Denken, allgemeines Verständnis, Gemeinsamkeiten finden, Bilder ergänzen, Mosaiktest, Figuren legen und Zahl-Symbol-Test (Tewes, 1991).

Die populärsten Vertreter figuraler Matrizen sind *Ravens Matrizentests* (Horn, 2009). Sie basieren ebenfalls auf Spearmans Annahme eines Generalfaktors, bieten jedoch die Möglichkeit einer ausschließlich sprachfreien und somit kultur- und bildungsunabhängigen Intelligenztestung, indem Probandinnen ausschließlich Zeichen und Muster analysieren und ergänzen müssen. Matrizentests werden inzwischen als Messverfahren für die *fluide* Intelligenz nach Cattell angesehen und testen somit analytisch-logische Intelligenz (Heller et al., 1998; Holling et al., 2004).

Als Gegenmodell können hier Wortschatztests, wie z.B. der *Mehrfach-Wortschatztest* (*MWT-B*; Lehrl, 1999), genannt werden, da sie fast ausschließlich die kristalline Intelligenz erheben. Dabei muss ein umgangs-, bildungs- oder wissenschaftssprachliches bekanntes Wort unter vier nichtexistierenden sinnlosen Wörter identifiziert werden (Halsband und Unterrainer, 2001). In weiteren Wortschatztests werden andere Strategien verfolgt, wie beispielsweise in den Untertests 1 und 2 des *Leistungsprüfsystems* (siehe unten), bei denen in existierenden Wörtern unterschiedlicher Alltagsnähe ein falscher Buchstabe detektiert werden muss.

Viele jüngere Intelligenztests orientieren sich am oben genannten Modell multipler Intelligenzfaktoren. Hier seien die *Intelligenz-Struktur-Tests* (z.B. *I-S-T 2000 R*; Liepmann et al., 2007) zu erwähnen, die in neun bis elf Aufgabengruppen (Skalen) neben verbalen, figural-räumlichen und rechnerischen Fähigkeiten auch die Merkfähigkeit testet. Sie versuchen kristalline sowie fluide Intelligenz zu evaluieren und bieten neben der Profilauswertung (Intelligenzstruktur) auch die Möglichkeit, einen Gesamtwert der allgemeinen Intelligenz (IQ) zu ermitteln.
Grundintelligenztests wie die Skalen des *Culture Fair Tests* (z.B. *CFT-20*; Weiß, 1998) zielen genau wie Matrizentests auf die ausschließliche Messung der fluiden Intelligenz nach Cattell, also der Grundintelligenz einer Person, ab. Die verschiedenen Tests der *CFT*-Reihe unterscheiden sich im Schwierigkeitsniveau und der Zielgruppe (Holling et al., 2004).
Während die bereits aufgeführten Testverfahren sich an Spearman und Cattell orientieren, wurde mit dem *Leistungsprüfsystem* (*LPS*; Horn, 1983) ein Intelligenztest konstruiert, dessen Aufgabengruppen den Primärfaktoren nach Thurstone zugeordnet werden können. Da in der hier vorliegenden Studie Teile des *LPS* verwendet wurden, sind ausführlichere Informationen zu diesem Intelligenztest unter 3.3.1 aufgeführt.
Für die Ergänzung von Beschwerdenvalidierungstests um die Erfassung intellektueller Funktionen steht somit eine Vielzahl validierter Testverfahren zur Verfügung, die Intelligenz im Sinne einer Gesamtleistung nach Spearman oder in Form von Subfacetten intellektueller Leistung zu erfassen versuchen.

2.4.4 Einfluss psychiatrischer Krankheiten auf Intelligenz

Psychische Erkrankungen hängen häufig mit Abweichungen kognitiver Teilleistungen und des Intelligenzquotienten von der Norm und prämorbiden Messwerten zusammen.
In Untersuchungen konnte man signifikante und spezifische Veränderung in den Ergebnissen multimodaler Intelligenztests als charakteristisch für bestimmte Psycho- und Neuropathologien erkennen. So kann eine Intelligenztestung wichtige differentialdiagnostische Hinweise geben, ob z.B. eine Verhaltens- oder Persönlichkeitsstörung als Folge einer hirnorganischen Störung auftritt oder bei der Beantwortung der Frage helfen, inwieweit eine verminderte intellektuelle Leistungsfähigkeit von einer dementiellen Entwicklung oder einer Depression herrührt

(Holling et al., 2004). Weiterhin kann die Beeinträchtigung definierter kognitiver Teilleistungen bei Depressiven sehr differenziert beobachtet werden (Martinez-Aran et al., 2002).

Wie bereits in Kapitel 2.3.5 erwähnt, haben viele Studien außerdem gezeigt, dass psychiatrische Erkrankungen mit einem allgemein unterdurchschnittlichen Intelligenzlevel korrelieren. Diese Ergebnisse, z.B. den Zusammenhang von niedriger Intelligenz mit späteren psychiatrischen Störungen bei Kindern und Jugendlichen betreffend (Schmidt, 2000), legen jedoch nah, dass die Intelligenzminderung die Erkrankung bedingt und nicht umgekehrt. Wenige Studien versuchen eine gegenläufige Kausalität herzustellen: So gelang es Sackeim et al. (1992) in einer longitudinalen Studie nachzuweisen, dass der Handlungs-IQ von Patientinnen in aktuellen depressiven Episoden signifikant niedriger war als vor und nach der Krankheitsphase. Keine Unterschiede ließen sich jedoch beim Verbal-IQ feststellen. Andere Studien belegen sogar einen irreversiblen strukturellen Schaden des Gehirns infolge depressiver Grundleiden (Adachi et al., 2011; Ahdidan et al., 2013). Davon kann abgeleitet und verallgemeinert werden, dass vor allem die fluide Intelligenz bei psychischen Erkrankung, hirnorganischen Schädigungen und Abbauprozessen nachlässt, nicht jedoch die kristalline (Holling et al., 2004). Das erscheint einerseits schlüssig, da für die Durchführung eines Intelligenztests ein Grundmaß an Antrieb vorhanden sein muss, welches bei starker Depression und anderen Erkrankungen nicht erreicht wird (Tölle und Windgassen, 2014). Andererseits erklärt sich diese Beobachtung durch die als *Matthäus-Effekt* bezeichnete Annahme Cattells, dass die kristalline Intelligenz die Konsequenz der früheren fluiden Intelligenz sei (Preckel und Brüll, 2008). Halsband und Unterrainer (2001) benennen explizit die Depression als ungünstig für die Intelligenzdiagnostik. Eine durch Depression hervorgerufene Minderung der kognitiven Leistungsfähigkeit wird als Pseudodemenz bezeichnet (Fellgiebel, 2017). Bei Intelligenztestungen Depressiver muss daher stets vor dem Hintergrund der aktuellen Krankheitsschwere geprüft werden, ob das Assessment formal durch die Patientin durchgeführt werden kann. Um Fehlern vorzubeugen, die bei der Anwendung üblicher Intelligenztests bei psychisch Kranken entstehen, wurde bereits 1968 eine Kurzform des *HAWIE* (reduzierter Wechsler-Intelligenztest, *WIP*; Dahl, 1986) erstellt, die sich in späteren Normierungen als gut mit den *HAWIE*-Ergebnissen korrespondierend herausstellte, obwohl die Testung hier deutlich kürzer ist.

Zusammenfassend lässt sich feststellen: Psychische Krankheiten wie Depression beeinträchtigen die fluide Intelligenz signifikant, was sich auf die Suppression kognitiver Teilfunktionen

zurückführen lässt. Dies muss in der akuten Krankheitsphase im Zuge eines Intelligenzassessments bei der Auswahl der Testverfahren berücksichtigt werden. Eine Methode zur Intelligenzerhebung, die in geringerem Ausmaß als konventionelle IQ-Tests von der aktuellen kognitiven Leistungsfähigkeit abhängt, ist die Intelligenzschätzung anhand sozialer Parameter, die im folgenden Kapitel kurz erläutert werden soll.

2.4.5 Abschätzung (prämorbider) Intelligenz

Neben den in Kapitel 2.4.3 dargestellten Testverfahren existiert noch die deutlich zeitsparendere, jedoch ungenauere Methode der Intelligenzschätzung. Dieses Verfahren wird in der Regel für die retrospektive Diagnostik der prämorbiden Intelligenz von Patientinnen mit Unfällen oder mit intelligenzmindernden Erkrankungen verwendet, da es gegenüber akuten kognitiven Beeinträchtigungen stabiler ist. Es wird dabei das aktuelle (morbide) Intelligenzniveau mithilfe von Tests gemessen und mit geschätzten Intelligenzmaßen vor Krankheitsbeginn (prämorbide) verglichen, um einerseits das Ausmaß der erkrankungsbedingten Intelligenzminderung zu eruieren und zusätzlich konkurrierende Kausalfaktoren zu beurteilen (Schneider et al., 2016).
Eine gängige Methode ist, in Wechsler-Tests den individuell besten Untertest als prämorbide und den schlechtesten als aktuelle Intelligenz zu werten (*Methode der Leistungsspitzen*). Dieses Vorgehen gilt allerdings aus verschiedenen Gründen nicht als hinreichend valide (Holling et al., 2004; Jahn et al., 2013).
Ebenfalls möglich ist die Errechnung eines Abbauindexes, wie es für den *HAWIE* von Wechsler vorgeschlagen und erfolgreich angewandt wurde (Crookes, 1961). Dabei werden die Wertpunkte stabiler (änderungsresistenter) mit denen instabiler (änderungssensitiver) Untertests hinsichtlich mentaler Abbauprozesse verglichen. Reliabilität und Validität dieses Verfahrens werden allerdings infrage gestellt, da es nie in einer Längsschnitt-Studie (im Vorher-Nachher-Vergleich) validiert wurde (Holling et al., 2004).
Außerdem Anwendung gefunden hat der Vergleich fluider mit kristalliner Intelligenz nach Cattell. Wie oben beschrieben leidet vor allem die fluide Intelligenz unter psychischen Störungen, weshalb man zur Erhebung des gesunden Intelligenzniveaus häufig die rein kristalline Intelligenz misst und auf Abweichungen zur fluiden untersucht.

Schließlich ist auch die Schätzung prämorbider Intelligenz anhand biographischer Daten möglich. Die hierzu verwendeten Sozialformeln entstehen aus der Analyse großer Untersuchungsstichproben, bei denen Korrelationen zwischen gemessenem Intelligenzquotienten und Angaben zu biographischen Daten, Gewohnheiten und Lebensumständen der Probandinnen ermittelt werden. Anhand der errechneten Formeln kann dann zukünftig der IQ bei anderen Probandinnen abgeschätzt werden. Diese Methode wurde erstmals in den 1970er Jahren zur Abschätzung von Wechsler-Intelligenztest-Ergebnissen verwendet (Wilson et al., 1978). Für eine Darstellung bereits publizierter Intelligenzschätzer im Englischen sowie Deutschen siehe Jahn et al. (2013). Ein relativ aktuelles Schätzverfahren dieser Art ist die in dieser Studie verwendete *Sozialformel* nach Jahn et al., deren Konzeption unter Kapitel 3.3.1 genauer beschrieben wird.

2.5 Aktueller Stand der Wissenschaft zum Zusammenhang von Intelligenz und negativer Antwortverzerrung

In Kapitel 2.3.5 wurde bereits dargelegt, dass Krankheiten der Psyche im Sinne der kognitiven Triade einen Einfluss auf die Authentizität der Symptomschilderung haben. Davon kann man jedoch nicht die Aussage ableiten, dass Personen mit niedrigerem IQ in Begutachtungssituationen bewusst häufiger übertreiben. Die Abweichung der Selbstwahrnehmung bei psychisch Kranken ist eine unbewusste nicht-motivationale.

Auch der aktuelle Forschungsstand bestätigt dies. Eine Studie aus dem Jahr 2011 vergleicht die IQ-Werte verschieden motivierter Gruppen mit ihren Ergebnissen in Beschwerdenvalidierungstests. Hier zeigte sich, dass vor allem die Motivation ausschlaggebend für Aggravationstendenzen war, für den IQ ließ sich keine prädiktive Aussagekraft bezüglich des Ergebnisses feststellen (Chafetz et al., 2011). Demakis et al. (2015) beobachteten des Weiteren in einer Untersuchung von 92 Studierenden keine signifikanten Zusammenhänge der mittels eines Wechsler-Lesetests bestimmten Intelligenz und der Ergebnisse verschiedener Beschwerdenvalidierungstests, wie z.B. des Word-Memory-Tests . Weitere Forschung zu diesem Thema steht noch aus.

Von diesen Ergebnissen unbeantwortet bleibt jedoch die Frage, inwieweit das Ausmaß, in dem eine bestehende Antwortverzerrung durch ein Testverfahren detektiert werden kann, auch durch Intelligenzmaße beeinflusst wird. Sie ist bisher wissenschaftlich gar nicht behandelt

worden. Es lassen sich keine Studien finden, in denen Probandinnen randomisiert angewiesen werden, *authentisch* bzw. *aggraviert* Beschwerden zu schildern und in denen die Ergebnisse eines Beschwerdenvalidierungstests dann mit ihrem IQ auf Zusammenhänge untersucht werden. Nur durch eine solche Instruktion (zu übertreiben bzw. sich authentisch zu äußern) könnte man die interindividuell unterschiedliche Motivation zur negativen Antwortverzerrung als Störvariable eliminieren (Schmidt et al., 2011).

2.6 Ableitung der Fragestellung

In Begutachtungssituationen kann das Ergebnis eines Beschwerdenvalidierungstests immer nur als ein Baustein im klinischen Gesamtbild betrachtet werden. Die Gutachterin muss dabei ihre Beurteilung anhand eines multimethodalen und multimodalen Assessments erstellen (Dohrenbusch et al., 2011; Dreßing et al., 2018). Dabei spielen auch intellektuelle Fähigkeiten der Patientin eine Rolle, vor deren Hintergrund die Möglichkeit einer negativen Antwortverzerrung eruiert werden muss. Nicht selten steht die Gutachterin hier vor der Frage, ob eine Patientin intelligent genug ist, die Strategien eines Beschwerdenvalidierungstests zu durchschauen und zu überlisten. Sinnvoll wäre hier eine kurze, valide Intelligenztestung, die auch bei erkrankten Patientinnen ein differenziertes Assessment des Ausmaßes von Intelligenz liefert und eine übergeordnete Beurteilung der getesteten Beschwerdenvalidität ermöglicht.
Es stellen sich nun also folgende Fragen: Gibt es einen Zusammenhang zwischen der Intelligenz der Begutachteten und der Qualität, mit der ein Beschwerdenvalidierungstest negative Antwortverzerrungen (z.B. Aggravation und Simulation) detektieren kann? Hängt die Testgüte eines Beschwerdenvalidierungstests antiproportional mit der Intelligenz der Begutachteten zusammen? Stellt die ermittelte Intelligenz einen nutzbaren Zusatzparameter zur Beurteilung der Beschwerdenvalidität dar? Und welche Dimensionen der Intelligenz nach Cattell (*kristallin* vs. *fluid*) betrifft dieser Zusammenhang am ehesten?

3 Material und Methoden

Die hier vorliegende Arbeit ist Bestandteil eines Validierungsprojektes zur Etablierung einer deutschsprachigen Fassung (Schmidt et al., 2019) des englischsprachigen Beschwerdenvalidierungstests *Structured Interview of Reported Symptoms, 2nd Edition* (Rogers et al., 2010).
Im Folgenden werden die für diese Arbeit relevante Stichprobe, die eingesetzten Instrumente sowie der Untersuchungsablauf beschrieben.
Abschließend soll die genannte Fragestellung in prüfbare Hypothesen übersetzt und entsprechende inferenzstatistische Verfahren zu ihrer Bewertung beschrieben werden.

3.1 Beschreibung der Stichprobe

3.1.1 Auswahl der Untersuchungsstichprobe und Rekrutierung

Die Studie fand im Rahmen eines Kooperationsprojektes mit dem BG-Klinikum *Bergmannstrost* in Halle statt. Sowohl die Gesamtstudie als auch die hier dargestellte Teilerhebung wurden von der Ethikkommission der Medizinischen Fakultät der *Martin-Luther-Universität Halle-Wittenberg* geprüft und genehmigt. Alle Studienteilnehmerinnen wurden über den Inhalt der Untersuchung informiert und gaben schriftlich ihre Zustimmung zur Teilnahme ab.
Eine erste Auswahl potenziell geeigneter Kandidatinnen für die vorliegende Untersuchung erfolgte quartalsweise anhand einer Datenbank über die an der Universitätspoliklinik für Psychiatrie, Psychotherapie und Psychosomatik innerhalb des Zeitraums April 2015 bis Oktober 2016 behandelten Patientinnen. Diese Datenbank umfasste Daten von insgesamt 1.525 ambulant versorgten Personen und lieferte Informationen über den Namen, die Behandlungsdiagnosen, das Geburtsdatum sowie die Kontaktdaten der jeweiligen Patientin.
Die Kriterien der Auswahl geeigneter Probandinnen bestanden zunächst in der gestellten Diagnose einer depressiven Erkrankung (depressive Anpassungsstörung *F43.2*, depressive Episoden *F32*, rezidivierende depressive Störung *F33*, Dysthymia *F34.1*). Zudem sollten die Patientinnen mindestens 18, höchsten aber 65 Jahre alt sein und Deutsch als Muttersprache sprechen. Als Gründe für die obere Altersgrenze sind statistisch häufigeres Auftreten dementieller Abbauprozesse und anderer psychopathologischer Phänomene als Störvariablen bei Probandinnen jenseits der 65 zu sehen. Außerdem sollte ein repräsentatives Probandinnenkollektiv

für die spätere Anwendbarkeit des *SIRS-2* in Verrentungsverfahren gewählt werden. Ausschlusskriterien stellten das Vorliegen komorbider hirnorganischer Erkrankungen (*F0*), akuter Substanzabhängigkeitserkrankungen (*F1x.0*, *F1x.2-F1x.9*), psychotischer Störungen (*F2*), manischer Episoden bzw. bipolar-affektiver Erkrankungen (*F30*, *F31*), emotional instabiler Persönlichkeitsstörungen (*F60.3*) und Intelligenzminderungen (*F7*) dar.

Aus der Datenbank gingen zunächst 188 Personen (davon 128 Frauen [68,1%]; Alter: M=45,70 Jahre; SD=12,66; min=20; max=65) als potenziell geeignete Studienteilnehmerinnen hervor.

Die eigentliche Rekrutierung der Untersuchungsstichprobe erfolgte parallel im Zeitraum Mai 2015 bis November 2016. Es wurde zunächst sukzessive versucht, telefonisch einen Kontakt mit den Patientinnen aufzunehmen, um das Interesse an der Studienteilnahme zu eruieren und ggf. einen Untersuchungstermin zu vereinbaren. Gelang dies, wurde nachfolgend eine Terminbestätigung inklusive der im Zuge der Studienvorbereitung ausgegebenen Dokumente (Informationsblatt, Einwilligungserklärung) postalisch an die Probandinnen versandt.

60 Personen wurden telefonisch auch nach mehreren Kontaktversuchen nicht erreicht. 53 Patientinnen lehnten eine Teilnahme aus verschiedenen Gründen ab (zu lange Anfahrt, keine zeitlichen Ressourcen, kein Interesse etc.). Von den Patientinnen, die zum Termin zugesagt hatten, erschienen zehn Personen in der Folge jedoch nicht zum vereinbarten Interviewtermin. Fünf interviewte Probandinnen mussten nachträglich aufgrund vorliegender Ausschlusskriterien (andere Muttersprache als Deutsch) aus dem Studienkollektiv exkludiert werden.

3.1.2 Deskriptive Stichprobenbeschreibung

Somit standen für die hier berichtete Arbeit 60 Probandinnen zur Verfügung (n=60). Der Einschlusszeitraum endete, nachdem die in dieser Höhe angestrebte Stichprobengröße erreicht worden war.

Die letztlich in die Studie eingeschlossenen Personen unterschieden sich mit einer Quote weiblicher Teilnehmerinnen von 70% (n=42) hinsichtlich der Geschlechterverteilung nicht signifikant von der Ausgangsgruppe aller n=188 laut Aktenlage potenziell geeigneten Personen (χ^2[df=1]=0,15; p=0,700). Auch bei der Altersstruktur der eingeschlossenen Probandinnen ließ sich kein signifikanter Unterschied zu den als grundsätzlich geeignet identifizierten Patientinnen feststellen (t[df=186]=-0,076; p=0,939).

Tabelle 1 zeigt die soziodemographischen Merkmale der Studienpopulation.

Tabelle 1. Stichprobenbeschreibung der eingeschlossenen Untersuchungsgruppe

	Gesamtstichprobe (n=60)
	M ± SD [min-max] bzw. n (%)
Alter (Jahre)	45,8 ± 12,06 [24-65]
Geschlecht	
weiblich	42 (70%)
männlich	18 (30%)
Familienstand	
ledig / alleine lebend	15 (25,0%)
verheiratet oder in Partnerschaft	26 (43,3%)
geschieden / getrennt	15 (25,0%)
verwitwet	4 (6,7%)
Bildung	
Hauptschulabschluss oder <10 Schuljahre	8 (13,3%)
Realschulabschluss oder <12 Schuljahre	31 (51,6%)
Abitur ohne abgeschl. Studium	6 (10%)
Abitur und abgeschl. Studium	15 (25%)
Diagnosen (psych. Störung)	
Depressive Anpassungsstörung	3 (5,0%)
Depressive Episode	32 (53,3%)
Rezidivierende depressive Störung	25 (41,7%)
Komorbide Persönlichkeitsstörung	5 (8,3%)

3.2 Durchführung der Untersuchung und Untersuchungsdesign

Die Probandinnen wurden auf dem Postweg schriftlich über die genauen Studieninhalte informiert. Im Anschluss erschienen sie zum vereinbarten Termin. Zu Beginn wurden die unterschriebenen Einwilligungen zur Studienteilnahme entgegengenommen, die bereits im Vorfeld per Post zugeschickt worden waren. Dann erfolgte die Aufnahme sozialer Daten zur späteren IQ-Abschätzung mithilfe einer *Sozialformel* (Jahn et al., 2013) und die Durchführung der Untertests 1-3 des *Leistungsprüfsystems* entsprechend dem Testmanual (Horn, 1983). Nach abgeschlossener Intelligenztestung bzw. -schätzung folgte dann die Zuordnung zu einer

Untersuchungsbedingung (siehe Kapitel 3.3.1). Dann wechselten die Probandinnen den Raum zu einer anderen Untersucherin, die das *SIRS-2* ohne Wissen über die Untersuchungsgruppenzugehörigkeit der Probandin durchführte. Erst zum Ende der Testung wurde die Gruppenzugehörigkeit aufgedeckt und die Selbsteinschätzung der Probandin bezüglich der Instruktionstreue erfragt.

Bei der dargestellten Studie handelt es sich um eine experimentelle Untersuchung, die über Fragebögen, Tests und Interviewinstrumente realisiert wurde.
Im folgenden Kapitel werden die eingesetzten Untersuchungsverfahren detailliert beschrieben.

3.3 Operationalisierung

Um zu ermitteln, ob die intellektuelle Leistungsfähigkeit einer Probandin mit der Messbarkeit der negativen Antwortverzerrung zusammenhängt, musste neben der Beschwerdenvalidität auch die Intelligenz gemessen bzw. geschätzt werden. Im Folgenden sind kurz die in dieser Studie eingesetzten Testverfahren bzw. Schätzformeln dargestellt.

3.3.1 Unabhängige Variablen: Untersuchungsinstruktion und Intelligenz

Unabhängige Einflussgrößen in dieser Studie sind zum einen die randomisiert zugewiesene Untersuchungsinstruktion (*authentisch* vs. *aggraviert*) und zum anderen die intellektuelle Leistungsfähigkeit, die mithilfe der Untertests 1-3 des *Leistungsprüfsystems* (Horn, 1983) gemessen und durch Anwendung einer *Sozialformel* (Jahn et al., 2013) abgeschätzt wurde.

Durchführung der Untersuchungsinstruktion
Vor der Durchführung des *SIRS-2* erfolgte die Gruppenzuteilung der Studienteilnehmerinnen, die randomisiert mittels Losverfahren (ohne Zurücklegen) durchgeführt wurde. Dabei bekam die Hälfte der Probandinnen die Anweisung, in der nachfolgenden Begutachtung mithilfe des *SIRS-2* ihre Beschwerden authentisch darzustellen, während die andere Hälfte instruiert wurde, bestehende Beschwerden zu aggravieren. Diese Instruktion erfolgte mithilfe standardisierter schriftlicher Anweisungstexte (vgl. hierzu Anlage B2 des SIRS-2-Testmanuals, S. 104, Schmidt et al., 2019). Die Intelligenztestung und Gruppenzuweisung der Probandinnen war so

organisiert, dass die Untersuchung mithilfe des *SIRS-2* daraufhin durch eine andere Untersucherin erfolgen konnte. Man kann also von einer Verblindung der Gutachterin hinsichtlich der Untersuchungsbedingung ausgehen.

Intelligenzmessung mithilfe des Leistungsprüfsystems (LPS)

Das *Leistungsprüfsystem* ist ein erstmals 1962 von Horn anhand einer Stichprobe von 10.000 Personen normiertes Testverfahren zur Ermittlung intellektueller Fähigkeiten, das später erweitert und verbessert wurde (Horn, 1983; Kreuzpointner et al., 2013). Die theoretische Grundlage hinsichtlich der Teststrukturierung bilden die Veröffentlichungen von Thurstone, die von sieben primären Gruppenfaktoren der Intelligenz ausgehen: *verbal comprehension* (Wortverständnis), *reasoning* (logisches Denken), *word fluency* (Wortflüssigkeit), *rote memory* (Merkfähigkeit), *space* (räumliches Vorstellungsvermögen), *perceptual speed* (Auffassungsschnelligkeit) und *number* (Rechengewandtheit) (Thurstone, 1938, 1946; Rost, 2013). Von diesen grundlegenden intellektuellen Kompetenzen ausgehend hat Horn 15 Untertests konzipiert. Zwei Untertests mit jeweils mind. 40 Aufgaben testen gemeinsam einen primären Gruppenfaktor.

In dieser Untersuchung wurde den Probandinnen ein Set aus drei Untertests (*LPS-1, -2 und* -3) der Auflage des *Leistungsprüfsystems* von 1983 vorgelegt. Die ersten beiden dieser Untertests (*LPS-1, LPS-2*) prüfen, wie gut die Probandinnen bestimmte real existierende Wörter erkennen können (*verbal comprehension*). Dadurch wird Aufschluss über die kristalline bzw. verbale Intelligenz erlangt. Untertest *LPS-3* hingegen erfordert, die innere Logik einer Reihe von Zeichen zu erkennen und das nicht in diese Reihe passende Zeichen zu markieren. Er zielt somit auf die Testung der fluiden Intelligenz ab (*reasoning*) (Amelang und Schmidt-Atzert, 2006). Auf Wunsch wurde den Probandinnen zur Vereinfachung statt des regulären *LPS*-Testbogens die vergrößerte Variante des *LPS-50+* (Sturm et al., 1993) ausgehändigt. Die hier verwendete Auswahl an Untertests ermöglichte ein knappes und ökonomisches Intelligenzassessment innerhalb von etwa 10 Minuten. Anhand des Untersuchungsmanuals ließen sich für die fluide und kristalline Intelligenz je ein populationsbezogener C-Wert auf der Normwertskala ermitteln (Horn, 1983), der im Verlauf in die IQ-Darstellung nach Stern und Wechsler (Wechsler, 1956) überführt und auf Korrelation mit der abhängigen Variablen untersucht werden konnte (siehe Kapitel 4).

Intelligenzschätzung mithilfe einer Sozialformel

Neben der testpsychologischen Beurteilung der Intelligenz mithilfe des *Leistungsprüfsystems* wurde zusätzlich eine Abschätzung der Intelligenz anhand einer *Sozialformel* vorgenommen. Dieses ursprünglich zur retrospektiven Einschätzung prämorbider Intelligenzniveaus bei Patientinnen mit Hirnverletzungen und Hirnfunktionsstörungen entwickelte Verfahren ermöglicht eine Abschätzung von intellektuellen Fähigkeiten anhand erfragter Variablen zur Person und deren Lebensumständen.

In der diesem Verfahren zugrundeliegenden Arbeit (Jahn et al., 2013) wurden bei einer bevölkerungsrepräsentativen Quotenstichprobe von 612 Probandinnen grundlegende soziodemographische Daten wie Alter, Geschlecht, Familienstand und sonstige Variablen wie Anzahl der Geschwister, Position in der Geschwisterreihe, Schulabschluss mit Durchschnittsnote, typische Fachnoten, Hochschulabschluss, höchste erreichte Berufsstellung, Mediennutzung (Presse, Bücher, Internet), Einwohnerzahl des Wohnorts und das Spielen eines Instrumentes erhoben. Diese Daten hatten sich in vorherigen Publikationen als geeignete Prädiktorvariablen für Intelligenzschätzung herausgestellt (Wilson et al., 1978; Barona und Chastain, 1986; Leplow und Friege, 1998). Jahn et. al (2013) unterzog alle teilnehmenden Probandinnen zusätzlich der Revision des *Hamburg-Wechsler-Intelligenztests für Erwachsene* (*HAWIE-R;* Tewes, 1991), der Werte für den *Verbal-IQ*, den *Handlungs-IQ* und den *Gesamt-IQ* lieferte. Im Verlauf der Datenanalyse wurden die Einflussvariablen auf Korrelationen mit den Intelligenzmaßen überprüft. Dabei ließen sich zumindest für den *Gesamt-IQ* und den *Verbal-IQ* Formeln entwickeln, um die Intelligenz effektiv zu schätzen (siehe Anhang in Jahn et al., 2013). Es wird für den *Gesamt-IQ* angegeben, dass 51,8% der Schätzungen innerhalb eines Intervalls von ≤ 7,5 IQ-Punkten Abweichung liegen. Für den *Verbal-IQ* war dieser Präzisionsparameter sogar bei 53,3% (Jahn et al., 2013).

3.3.2 Abhängige Variable: Beschwerdenvalidität

Als von den oben beschriebenen Größen abhängige Variable wurde das Ausmaß einer negativ verzerrten Beschwerdendarstellung mithilfe der Subskalen und des Klassifikationsergebnisses des *SIRS-2* erfasst (zur Testdurchführung vgl. das Testmanual: Schmidt et al., 2019).

Aufbau des SIRS-2

Die englische Ursprungsversion *SIRS-2* wurde im Jahr 2010 von Rogers et al. publiziert (Rogers et al., 2010). Sie besteht aus 171 Items, die in Form eines strukturierten Interviews erhoben werden. Die meisten davon sind Entscheidungsfragen und prüfen das Vorhandensein eines bestimmten Symptoms oder einer Symptomkombination. Hierbei unterteilt sich das *SIRS-2* in die *Allgemeine Befragung*, bei der die Antwortmöglichkeiten „Ja", „Nein" und „eingeschränktes Ja / manchmal" möglich sind und die *Detaillierte Befragung* vom geschlossen dichotomen Fragestil (nur „Ja" und „Nein"). Bei den Items der detaillierten Befragung wird zu jedem geschilderten Symptom zusätzlich erhoben, ob dieses als „unerträglich" empfunden wird. Außerdem werden die Items der detaillierten Befragung im Laufe des Interviews einmal wiederholt. Es gibt neben diesen geschlossenen Fragen zusätzlich zwei Testabschnitte, in denen die Probandinnen Gegenteile und Reime zu vorgegebenen Wörtern bilden sollen. Neben der Befragung gehört es außerdem zur Aufgabe der Gutachterin, das Verhalten der Probandin während des Interviews zu beobachten, um hier mögliche Inkonsistenzen herauszufinden.

Skalen des SIRS-2

Das *SIRS-2* verwendet verschiedene Strategien, um Antwortverzerrungen bei den Probandinnen zu detektieren. Diese Methoden wurden von Rogers aus bereits bestehenden Beschwerdenvalidierungstests wie z.B. dem *MMPI* (Hathaway und McKinley, 1943) und üblichen klinischen Interviews (Rogers, 1984) übernommen. Die dabei erhobenen Antworten münden dann in die nachfolgend samt ihren Herangehensweisen aufgeführten Primär- und Zusatzskalen (zu den einzelnen Items und Cut-off-Werten vgl. Rogers et al., 2010). Als auffällig gilt jeweils, wer eine bestimmte Anzahl an positiven Antworten überschreitet:

Primärskalen

- **Seltene Beschwerden – Rare Symptoms (*RS und RS-total*)**

  Es handelt sich um Symptome, die in echten klinischen Gruppen sehr selten sind, vor allem Wahrnehmungs- und psychotische Phänomene.

- **Symptomkombinationen – Symptom Combinations (*SC*)**

  Hierbei werden Paare von Beschwerden erfragt, die einzeln häufig vorkommen, aber selten gemeinsam.

- **Unglaubhafte oder absurde Beschwerden – Improbable or Absurd Symptoms (*IA)***

  Diese Beschwerden sind besonders absurde, bizarre und fantastische Beschwerden, die so gut wie nie in echten klinischen Gruppen angegeben werden.

- **Offenkundige Beschwerden – Blatant Symptoms (*BL*)**

  Es handelt sich hierbei um charakteristische Beschwerden, die von der Allgemeinbevölkerung ohne Fachwissen schweren psychischen Störungen zugeschrieben werden.

- **Subtile Beschwerden – Subtle Symptoms (*SU*)**

  Hierbei handelt es sich um Beschwerden, die von der Allgemeinbevölkerung als nicht pathologisch angesehen werden, jedoch in Maßen in klinischen Gruppen bei psychischen Erkrankungen auftreten. Eine moderate Erhöhung der positiven Antworten ist jedoch auch bei posttraumatischen Störungen beobachtet worden.

- **Selektivität der Beschwerden – Selectivity of Symptoms (*SEL*)**

  Probandinnen, die allgemein häufiger als in echten klinischen Stichproben beobachtet, Fragen nach konkreten Beschwerden bejahen, zeigen eine generell verstärkte Darstellung von Problemen, die Hinweis auf negative Antwortverzerrung geben kann.

- **Schweregrad der Beschwerden – Severity of Symptoms (*SEV*)**

  Diese Subskala misst, wie häufig in der detaillierten Befragung konkrete Beschwerden als „unerträglich“ angegeben werden. Der Vergleich zu klinischen Stichproben hat ergeben, dass viele als derart belastend angegebene Beschwerden für eine authentische Schilderung von Beschwerden untypisch sind. Eine Ausnahme bildeten dabei ehrliche Patientinnen, die unter einem akuten Disstress-Erleben leiden.

- **Geschilderte vs. Beobachtete Beschwerden – Reported vs. Observed Symptoms (*RO*)**

  Hierbei wird erhoben, inwieweit die Schilderung von konkreten sichtbaren Symptomen mit dem tatsächlichen Verhalten der Probandin während des Interviews übereinstimmt.

Häufige nicht-authentische Symptomschilderungen und mehr als eine plötzliche Adaptation des abgefragten Verhaltens nach Abfrage des Items durch die Probandin treten fast nie bei authentisch antwortenden Probandinnen auf.

Zusatzskalen

- **Direkte Einschätzung von Ehrlichkeit – Direct Appraisal of Honesty (*DA*)**
  Es handelt sich um Items, die die Ehrlichkeit im Umgang mit dem psychiatrischen Behandlungsteam direkt erfragen.
- **Defensive Beschwerden – Defensive Symptoms (*DS*)**
  Bei diesen Fragen geht es um alltägliche Probleme, die sogar viele Probandinnen aus gesunden Stichproben schildern. Geringe Werte auf dieser Skala sprechen oft für eine Dissimulation.
- **Übergenaue Beschwerdenschilderung– Overly Specified Symptoms (*OS*)**
  Es handelt sich um häufige Beschwerden, bei denen jedoch eine spezifische (unwahrscheinliche) Ausprägung erfragt wird.
- **Unwahrscheinliche Fehler – Improbable Failure (*IF*)**
  Hier werden den Probandinnen einfache Aufgaben gestellt, die auch in klinischen Gruppen psychisch Kranker in der Regel bewältigt werden (Gegenteile, Reime bilden). Eine hohe Rate von Fehlern erhöht die Wahrscheinlichkeit des Übertreibens, stellt aber keinesfalls einen Beweis dar, da auch ein geringes Intelligenzniveau die Falschantworten erklären kann.
- **Inkonsistente Symptome – Inconsistency of Symptoms (*INC*)**
  Bei der wiederholten detaillierten Befragung kommt es auch bei authentischen Beschwerdenschilderungen zu veränderten Antworten. Wenn hier allerdings besonders häufig anders als beim ersten Mal geantwortet wird, gibt das einen Hinweis auf nicht-authentische Symptomschilderung.

Ergebniskategorien des SIRS-2

Je nach Wert kann für jede Skala ermittelt werden, wie wahrscheinlich die jeweiligen Antworten auf eine verzerrte Symptomschilderung hinweist. Rogers legt für jede Skala einen eigenen Cut-off-Wert fest, ab dem Übertreiben wahrscheinlich bis definitiv ist. Daraufhin ist zur

Feststellung der Ergebniskategorie ein Flussdiagramm anzuwenden, bei dem aus den Skalenergebnissen und weiteren daraus errechneten Indizes zu einer von fünf im Folgenden kurz beschriebenen Ergebnisklassen gelangt wird (siehe Abbildung 4.1, S. 38, in Rogers et al., 2010): Bei der Kategorie *Feigning* kann bei einer Falsch-positiv-Rate von 2,5% relativ sicher von einer negativen Antwortverzerrung ausgegangen werden, da in den Primärskalen mehrfach auffällige Werte ermittelt wurden.

Probandinnen mit dem *SIRS-2*-Ergebnis *Indetermininate: Evaluate* übertreiben in über der Hälfte der Fälle. Das Testergebnis ist allerdings nicht eindeutig genug, sodass ein weiteres Assessment zur Sicherung der Befunde erfolgen muss.

Wenn die Skalenergebnisse ein heterogenes unstimmiges Gesamtbild mit Hinweisen auf negative Antwortverzerrung sowie authentische Symptomschilderung ergeben, kann von einem Disengagement (Kategorie *Disengagement: Indeterminate Evaluate)* ausgegangen werden. Auch hier ist die Wahrscheinlichkeit, dass es sich um eine übertreibende Probandin handelt, über 50%.

Bei der Klassifikation *Indeterminate: General* ist weder eine negative Antwortverzerrung noch eine authentische Symptomschilderung eindeutig festzustellen. Die Rate an tatsächlich Übertreibenden beträgt hier 34,3%.

Wenn als Ergebnis des *SIRS-2* die Kategorie *Genuine Responding* ermittelt wird, liegt mit hoher Wahrscheinlichkeit eine authentische Symptomschilderung vor.

Erstellung und Gütekriterien der deutschen Übersetzung des SIRS-2

Im März 2019 wurde die deutsche Version des *SIRS-2* bei dem *Hogrefe*-Verlag publiziert, wobei die in dieser Arbeit durchgeführten Interviews zur Validierung des übersetzten Testverfahrens beitrugen (Schmidt et al., 2019). Durch die Autoren dieses übergeordneten Studienprojektes erfolgte nach Einwilligungen des Erstautors und des Verlages der Originalversion zunächst ein Übersetzungsprozess nach wissenschaftlichen Standards. Dabei wurden die Instruktionen und Items der Originalversion ins Deutsche überführt, die Übersetzungen mit unabhängigen Sprachwissenschaftlern diskutiert und einer Stichprobe von Patientinnen zur Prüfung der Anwendbarkeit vorgelegt. Anschließend erfolgte eine Rückübersetzung durch zwei bilinguale Übersetzer, die schließlich durch den Autor des Originalinstrumentes validiert und freigegeben wurde. Für eine differenzierte Übersicht zur Skalenübersetzung siehe Schmidt et al. (2019).

Die Validierung der deutschen Übersetzung des *SIRS-2* anhand einer Stichprobe von 143 Probandinnen hat ergeben, dass sich für die Primärskalen eine mittlere Effektstärke von d=2,36 errechnen ließ, was die Validität der Skalenauswahl bescheinigt. Wiederholte unabhängige Durchführungen des *SIRS-2* durch verschiedene Untersucherinnen haben hier außerdem eine hohe Interrater-Reliabilität ergeben (mittleres r=.99). Bei einer moderaten bis hohen Retest-Reliabilität wiesen Probandinnen, die das Interview mehrmals durchlaufen haben, hinsichtlich der Cut-off-Werte in allen Skalen eine hohe Übereinstimmungen zwischen den unterschiedlichen Testzeitpunkten auf (mittlere Konkordanzen der Klassifikation: 96,4%). Bei einer Sensitivität von 97,3%, einer Spezifität von 100%, einem positiv-prädiktivem Wert von 100% und einem negativ-prädiktivem Wert von 98,6% kann außerdem von sehr guten Klassifikationsgütekriterien ausgegangen werden (Schmidt et al., 2019). In einer Begutachtung erfolgen fälschliche Einschätzungen durch den Test dementsprechend praktisch ausschließlich zugunsten der Probandin – also in Richtung einer geringeren negativen Antwortverzerrung, was vor dem Hintergrund der Folgen denkbarer Falschergebnisse für die Gesellschaft und die einzelne Probandin auf eine sinnvolle Festlegung der Cut-off-Werte schließen lässt.

3.3.3 Umgang mit und Erfassung von potenziellen Störgrößen

Grundsätzlich ist davon auszugehen, dass nicht allen Probandinnen gleichermaßen das systematische Übertreiben der Beschwerden gelingt. Außerdem ist denkbar, dass die Instruktion nicht korrekt verstanden wurde. Um diese Unsicherheiten zu klären, wurde bei jeder Probandin nach durchgeführtem Beschwerdenvalidierungstest ein *Adherence-Check* durchgeführt. Sie wurde gebeten, die Instruktion in eigenen Worten wiederzugeben und außerdem einzuschätzen, inwieweit es ihr gelungen ist, den Anweisungen Folge zu leisten (*Adherence* von 0-100%). Um außerdem sicherzustellen, dass die Untersuchungsgruppen sich in sozioökonomischen Merkmalen nicht signifikant unterscheiden, wurden verschiedene potenzielle Störgrößen erhoben und für die jeweiligen Gruppen verglichen. Diese Ergebnisse sind in Kapitel 4.1 aufgeführt.

3.4 Auswertungsplan und statistische Hypothesen

3.4.1 Hypothesen

Hinsichtlich der Ergebnisse dieser Studie werden folgende Hypothesen aufgestellt:

1. Die Experimentalbedingungen (*authentisch* vs. *aggraviert*) schlagen sich in signifikanten Gruppenunterschieden der *SIRS-2*-Subskalen nieder.
2. Die korrekte Detektion der bei den Probandinnen a priori festgelegten Experimentalbedingungen durch das *SIRS-2* ist nicht mit deren Intelligenzleistung assoziiert. Es wird erwartet, dass sich Personen mit *korrekter* vs. *nicht korrekter* Klassifikation des *SIRS-2* nicht systematisch in ihren intellektuellen Fähigkeiten (nach *LPS* Untertests 1-3 und *Sozialformel*) unterscheiden.
3. Im Gegensatz zu Hypothese 2 wird jedoch angenommen, dass sich eine signifikante Interaktion zwischen den aus Experimentalbedingung (*authentisch* vs. *aggraviert*) und Klassifikationsgüte des *SIRS-2* (*korrekt* vs. *nicht korrekt*) gebildeten Gruppen und den Intelligenzmaßen der Probandinnen *(LPS* Untertests 1-3 und *Sozialformel)* findet.
4. Die eingesetzten Intelligenzmaße eignen sich zur Vorhersage der Experimentalbedingung *authentisch* vs. *aggraviert* bei allen Probandinnen nicht eindeutiger *SIRS-2*-Klassifikation. Es lässt sich ein signifikantes lineares Regressionsmodell erstellen.
5. Hypothese 5 postuliert signifikante korrelative Zusammenhänge zwischen den Subskalen des *SIRS-2* und den eingesetzten Intelligenzmaßen für die Gesamtstichprobe sowie für beide Experimentalbedingungen getrennt.

3.4.2 Statistische Verfahren

Die statistische Auswertung aller erhobenen Daten wurde mit der Statistiksoftware *SPSS* (Version 17.0) vorgenommen.

Dabei erfolgte zunächst eine deskriptive Darstellung der Merkmale der Untersuchungsgruppe anhand von Mittelwerten und Standardabweichung bzw. absoluten und relativen Häufigkeiten. Ein Vergleich zur Referenzpopulation stationär behandelter Patientinnen sowie die Prüfung der Vergleichbarkeit von randomisiert zugewiesener Gruppenzugehörigkeit wurde inferenzstatistisch mittels X^2 und t-Test realisiert.

Die Hypothesentestung erfolgte zunächst unter Verwendung varianzanalytischer Verfahren (Hypothesen 1-3), die Einzelgruppenvergleiche unter Hypothese 3 erfolgten zunächst mittels t-Tests (Bonferroni-korrigiert), aufgrund der kleiner Teilgruppengrößen und nicht normalverteilter Variablen wurden diese Befunde non-parametrisch abgesichert (Kruskall-Wallis-Test).

Hypothese 4 wurde mittels eines linearen Regressionsmodells geprüft. Aufgrund der hohen Korrelationen zwischen den eingesetzten Prädiktoren (Kollinearität) wurde ein schrittweises Modell angewendet ($p_{in} \leq 0{,}05$; $p_{out} \geq 0{,}10$). Zusätzlich kam eine ROC-Analyse zur Bestimmung der Zuordnungsgüte zum Einsatz. Zur Bestimmung der korrelativen Zusammenhänge zwischen den Intelligenzmaßen und den *SIRS-2*-Skalen wurden Produktmomentkorrelationen berechnet (Hypothese 5).

4 Ergebnisse

4.1 Randomisierung der Untersuchungsbedingungen

Die randomisierte Zuweisung der eingeschlossenen Probandinnen resultierte in zwei Unterstichproben (*authentische* vs. *aggravierte* Symptomdarstellung) mit je n=30 Personen. In keinem der untersuchten Fälle zeigte die systematische Abfrage des Instruktionsverständnisses (*Adherence*-Check) eine Notwendigkeit an, Probandinnen von der Untersuchung auszuschließen.
Beide Gruppen unterschieden sich statistisch nicht signifikant hinsichtlich ihres Alters ($t[df=58]=-0{,}661$; $p=0{,}511$), Geschlechts ($\chi^2[df=1]=0{,}000$; $p>0{,}999$), Familienstands ($\chi^2[df=3]=5{,}320$; $p=0{,}150$) und ihrer Bildung ($\chi^2[df=3]=3{,}747$; $p=0{,}290$). Gleichsam unterschieden sich die Probandinnen nicht bezüglich der Häufigkeit depressiver Störungen ($\chi^2[df=2]=4{,}485$; $p=0{,}106$) sowie komorbider Persönlichkeitsstörungen ($\chi^2[df=1]=218$; $p=0{,}640$).
Bezüglich der in der vorliegenden Arbeit untersuchten unabhängigen Variablen der Intelligenz zeigte sich das in Tabelle 2 dargestellte Muster.

Tabelle 2. Intelligenzmaße in den Untersuchungsgruppen

	M ± SD		**Teststatistik**
Intelligenzmaß	*authentisch* (n=30)	*aggraviert* (n=30)	t[df=58]; p
Sozialformel			
Gesamt-IQ	110,75 ± 12,70	109,10 ± 11,66	-0,525; p=0,602
Verbal-IQ	109,51 ± 12,43	108,78 ± 10,06	-0,249; p=0,804
LPS			
1+2	101,75 ± 8,48	100,75 ± 9,19	-0,438; p=0,663
3	105,05 ± 12,04	102,50 ± 10,65	-0,869; p=0,389

Wie zu sehen ist, unterscheiden sich die Untersuchungsgruppen in keinem der untersuchten Intelligenzmaße signifikant voneinander.

4.2 *SIRS*-2-Subskalen und Instruktionstreue in den Untersuchungsgruppen

Um zu ermitteln, ob die Probandinnen den Instruktionen der Untersuchungsbedingungen Folge geleistet haben, wurde unter Hypothese 1 zunächst geprüft, ob die Zuweisung zu den Experimentalgruppen (*authentisch* / *aggraviert*) zu signifikanten Gruppenunterschieden in den *SIRS-2*-Subskalen führen. In der folgenden Tabelle 3 werden die Skalenwerte der Primär- und Zusatzskalen des *SIRS-2* nach Zuweisung zu einer der beiden Untersuchungsinstruktionen getrennt aufgeführt und auf Gruppenunterschiede untersucht.

Tabelle 3. *SIRS-2*-Subskalen nach Untersuchungsgruppen

	M ± SD		**Teststatistik**
***SIRS*-2-Skalen**	*authentisch* (n=30)	*aggraviert* (n=30)	F[df=1]; p
Primärskalen			
Seltene Beschwerden (RS)	1,17 ± 1,98	5,37 ± 2,67	47,79; p<0,001
Symptomkombinationen (SC)	1,00 ± 0,98	4,60 ± 2,54	52,39; p<0,001
Unglaubhafte oder absurde Beschw. (IA)	0,67 ± 1,12	3,57 ± 2,08	45,16; p<0,001
Offenkundige Beschwerden (BL)	3,80 ± 3,25	14,83 ± 4,36	123,59; p<0,001
Subtile Beschwerden (SU)	8,77 ± 5,28	23,00 ± 5,15	111,60; p<0,001
Selektivität der Beschwerden (SEL)	8,90 ± 4,77	20,23 ± 3,32	114,02; p<0,001
Schweregrad der Beschwerden (SEV)	3,67 ± 4,06	17,60 ± 4,61	154,41; p<0,001
Geschilderte vs. beobachtete Beschw. (RO)	1,50 ± 1,72	5,57 ± 2,87	44,29; p<0,001
RS-total	0,93 ± 1,78	5,80 ± 5,14	24,05; p<0,001
Zusatzskalen			
Direkte Einschätzung der Ehrlichkeit (DA)	2,43 ± 1,31	3,37 ± 3,29	2,09; p=0,153
Defensive Beschwerden (DS)	23,57 ± 8,38	32,37 ± 5,03	24,34; p<0,001
Unwahrscheinliche Fehler (IF)	0,53 ± 1,70	2,10 ± 2,86	6,67; p=0,012
Übergenaue Beschwerdenschilderung (OS)	1,03 ± 1,40	2,77 ± 2,46	11,25; p=0,001
Inkonsistente Symptome (INC)	1,37 ± 1,30	3,90 ± 3,12	16,84; p<0,001
***Adherence* (%)**	94,48 ± 10,02	68,11 ± 17,61	48,74; p<0,001

Tabelle 3 zeigt, dass sich die beiden Untersuchungsgruppen in ihren Ergebnissen der *SIRS-2*-Skalen systematisch und signifikant unterscheiden. Lediglich in der Subskala *Direkte Einschätzung der Ehrlichkeit* ergab sich kein Gruppenunterschied (p=0,153). Von dieser Ausnahme abgesehen lässt sich Hypothese 1 also bestätigen.

Der jeweils am Ende der Untersuchung durchgeführte *Adherence-Check* zeigte, dass es einen signifikanten Unterschied in der Selbsteinschätzung der Instruktionstreue zwischen den Gruppen gab. Deskriptiv fanden sich höhere Werte, also eine bessere selbsteingeschätzte Instruktionstreue, bei den Probandinnen der Gruppe *authentische* Symptomdarstellung.

4.3 *SIRS-2*-Klassifikation und Intelligenz in den Untersuchungsgruppen

Nachfolgend wird in Tabelle 4 aufgeführt, wie viele Probandinnen in den beiden Untersuchungsgruppen durch das *SIRS-2* welcher Ergebniskategorie zugeordnet wurden.

Tabelle 4. Klassifikation der Beschwerdenvalidität im *SIRS-2* nach Untersuchungsgruppen

	Untersuchungsgruppe	
***SIRS*-2-Klassifikation**	*authentisch* (n=30)	*aggraviert* (n=30)
Feigning	0	16
Indeterminate Evaluate	0	9
Disengagement: Indeterminate Evaluate	1	0
Indeterminate General	4	4
Genuine Responding	25	1

Entsprechend Tabelle 4 ist bei 16 Probandinnen eine negative Antwortverzerrung *(Feigning)*, bei 25 die authentische Symptomschilderung *(Genuine Responding)* richtig detektiert worden. Diese 41 Personen wurden durch das *SIRS-2* eindeutig und korrekt klassifiziert. Eine aggravierende Probandin wurde fälschlicherweise als ehrlich klassifiziert. Bei den restlichen 18 Teilnehmerinnen ist das *SIRS-2* zu keinem eindeutigen Ergebnis gekommen und hat das Antwortverhalten als *Indeterminate Evaluate*, *Disengagement: Indeterminate Evaluate* oder *Indeterminate General* klassifiziert.

In einem ersten analytischen Schritt wurde nun analog zu Hypothese 2 untersucht, ob die Intelligenz der Probandinnen systematisch mit der Beurteilbarkeit der Beschwerdenvalidität durch das *SIRS-2* zusammenhängt. Hierzu wurden zwei Gruppen gebildet:

Als ***korrekt*** bezeichnet werden alle 41 Probandinnen, bei denen das *SIRS-2* die Untersuchungsbedingung eindeutig bestimmt hat.

Nicht korrekt eingestuft sind die restlichen 19 Probandinnen, bei denen ein uneindeutiges oder falsches Klassifikationsergebnis ermittelt wurde.

Entsprechend dieser zwei Gruppen werden in der folgenden Tabelle 5 die erhobenen Intelligenzmaße aufgeführt und auf Gruppenunterschiede untersucht:

Tabelle 5. Intelligenzmaße nach Beurteilungsgüte

Intelligenzmaß (M ± SD)	*korrekt* (n=41)	*nicht korrekt* (n=19)	**Teststatistik** F[df=3]; p
Sozialformel			
Gesamt-IQ	109,83 ± 12,15	110,12 ± 11,39	0,007; p=0,932
Verbal-IQ	109,17 ± 11,53	109,14 ± 11,22	0,001; p=0,977
LPS			
1+2	100,48 ± 8,89	102,92 ± 8,52	1,008; p=0,320
3	102,34 ± 11,55	106,86 ± 10,50	2,106; p=0,152

Wie Tabelle 5 zeigt, ließ sich ein systematischer Gruppenunterschied der Beurteilungsqualität des *SIRS-2* zwischen *korrekter* vs. *nicht korrekter* Zuordnung hinsichtlich keiner der Intelligenzmaße nachweisen. Hypothese 2 hat sich somit bestätigt.

Weiterführend wurde die Untersuchung des IQ zur Beantwortung von Hypothese 3 in enger gefassten Gruppen vorgenommen. Aufgrund der kleinen Zellbesetzungen der uneindeutigen Zuordnungen (siehe Tabelle 4) von insgesamt n=13 Personen bei der *aggravierten* und n=5 Personen bei der *authentischen* Untersuchungsbedingung werden die Kategorien im Folgenden je nach Untersuchungsinstruktion zu zwei korrekten und zwei uneindeutigen Klassen zusammengefasst. Es ergeben sich folgende vier Gruppen:

Korrekt authentisch bezeichnet alle n=25 richtig als authentisch klassifizierte Probandinnen.

Nicht eindeutig authentisch sind alle n=5 Probandinnen, die die Instruktion erhalten hatten, die Wahrheit zu sagen, aber vom *SIRS-2* inkorrekt als *Disengagement: Indeterminate Evaluate* oder *Indeterminate General* klassifiziert wurden.

Korrekt aggraviert bezeichnet die n=16 richtig als Übertreibende klassifizierten Probandinnen.

Nicht eindeutig aggraviert sind die n=13 Übertreibende, die vom *SIRS-2* fälschlicherweise als *Disengagement: Indeterminate Evaluate* und *Indeterminate General* klassifiziert wurden.

Tabelle 6 stellt nun die Intelligenzmaße in diesen Klassifikationsgruppen dar:

Tabelle 6. Intelligenzmaße in den Untersuchungsgruppen

Intelligenzmaß (M ± SD)	*korrekt authentisch* (n=25)	*korrekt aggraviert* (n=16)	*Nicht eindeutig authentisch* (n=5)	*Nicht eindeutig aggraviert* (n=13)	**Test-statistik** F[df=3]; p
Sozialformel					
Gesamt-IQ	113,34 ± 12,17	104,36 ± 11,50	97,82 ± 5,44	115,06 ± 9,73	5,02; p=0,004
Verbal-IQ	112,04 ± 11,86	104,69 ± 9,73	96,85 ± 6,00	113,90 ± 8,70	4,94; p=0,004
LPS					
1+2	102,94 ± 8,53	96,63 ± 8,26	95,80 ± 5,55	106,00 ± 8,10	4,26; p=0,009
3	104,62 ± 11,73	98,78 ± 10,67	107,20 ± 14,79	107,15 ± 9,49	1,64; p=0,190

Diese Daten sind in den folgenden Abbildungen 1-4 (siehe folgende Seiten) in Form von Boxplot-Diagrammen unter Angabe der p-Werte bei signifikantem Unterschied der Intelligenzmaße zwischen den Gruppen veranschaulicht.

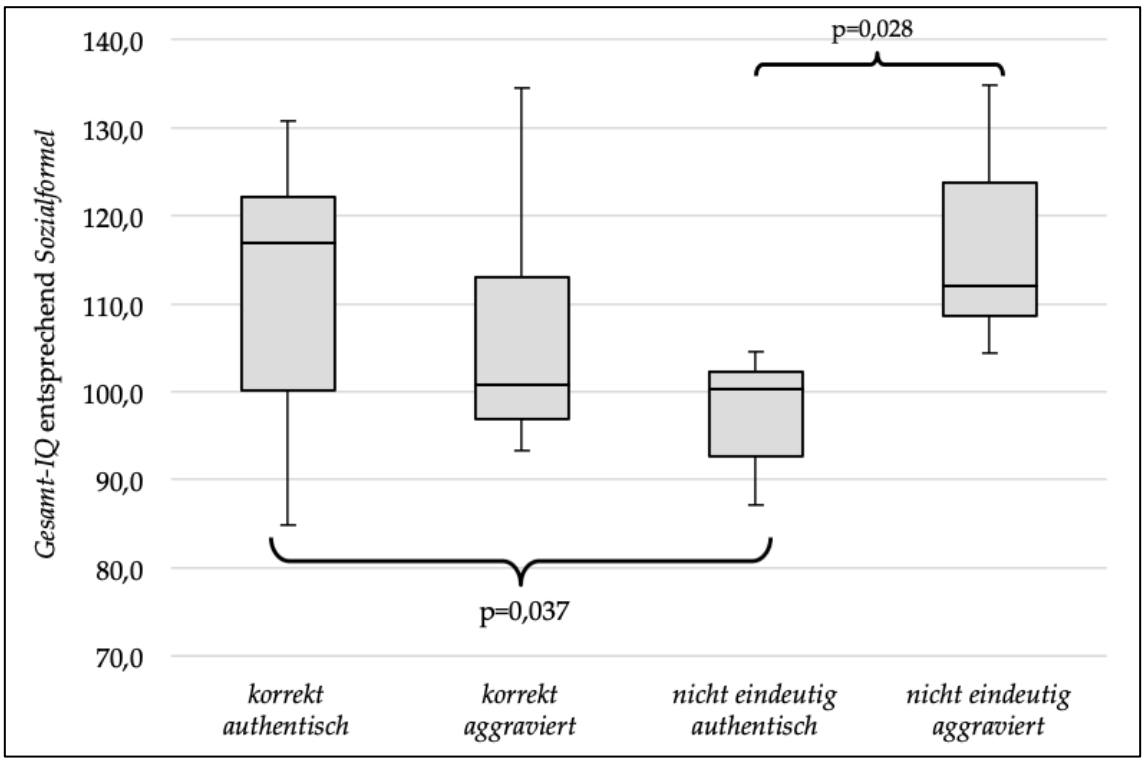

Abbildung 1. *Gesamt-IQ* der *Sozialformel* in den Untersuchungsgruppen

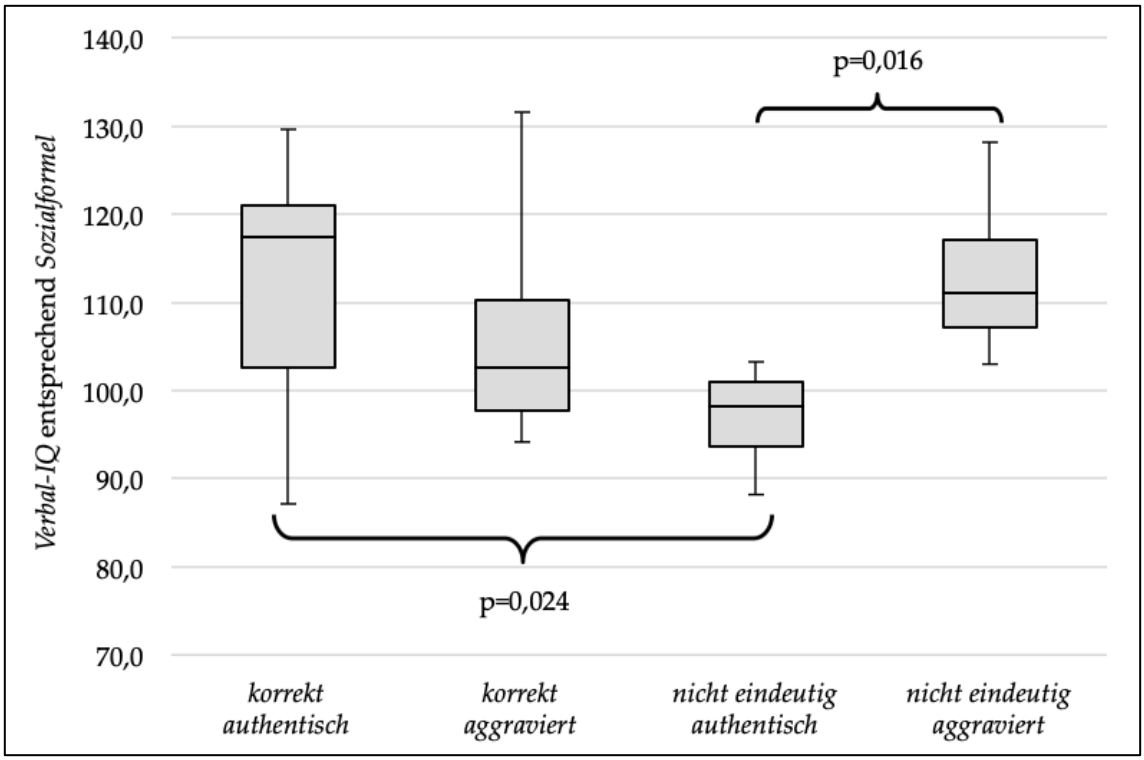

Abbildung 2. *Verbal-IQ* der *Sozialformel* in den Untersuchungsgruppen

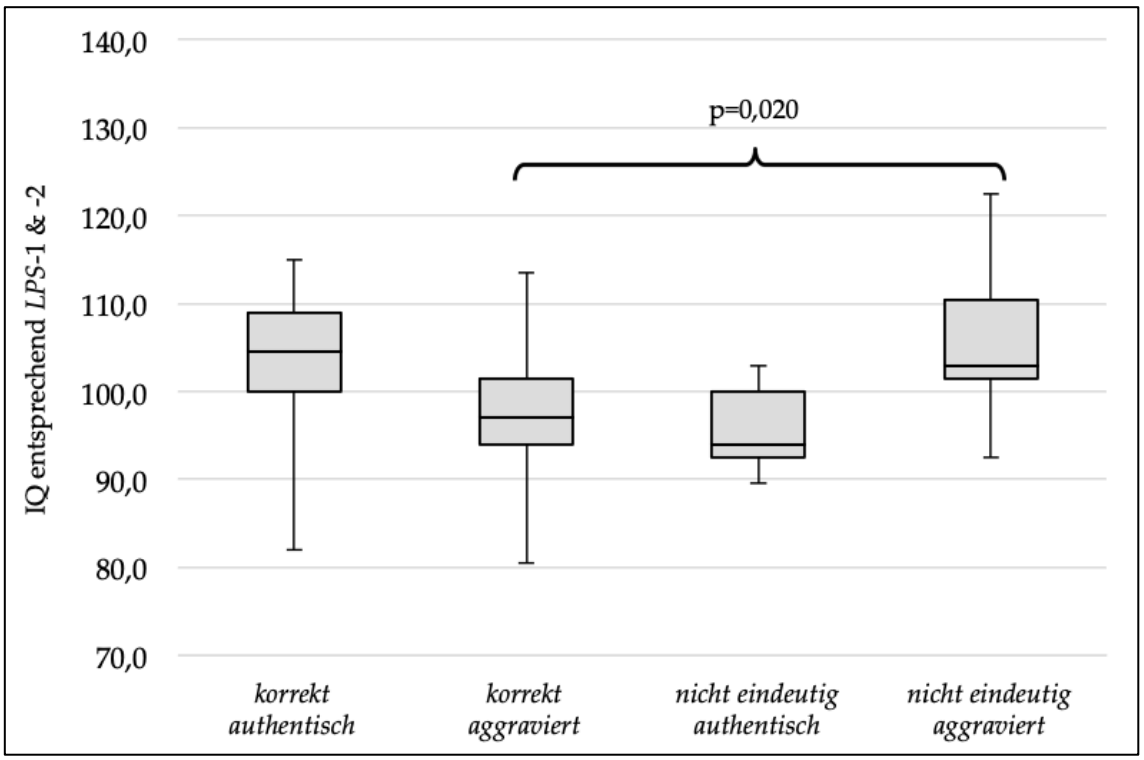

Abbildung 3. IQ nach *LPS*-Untertests 1+2 in den Untersuchungsgruppen

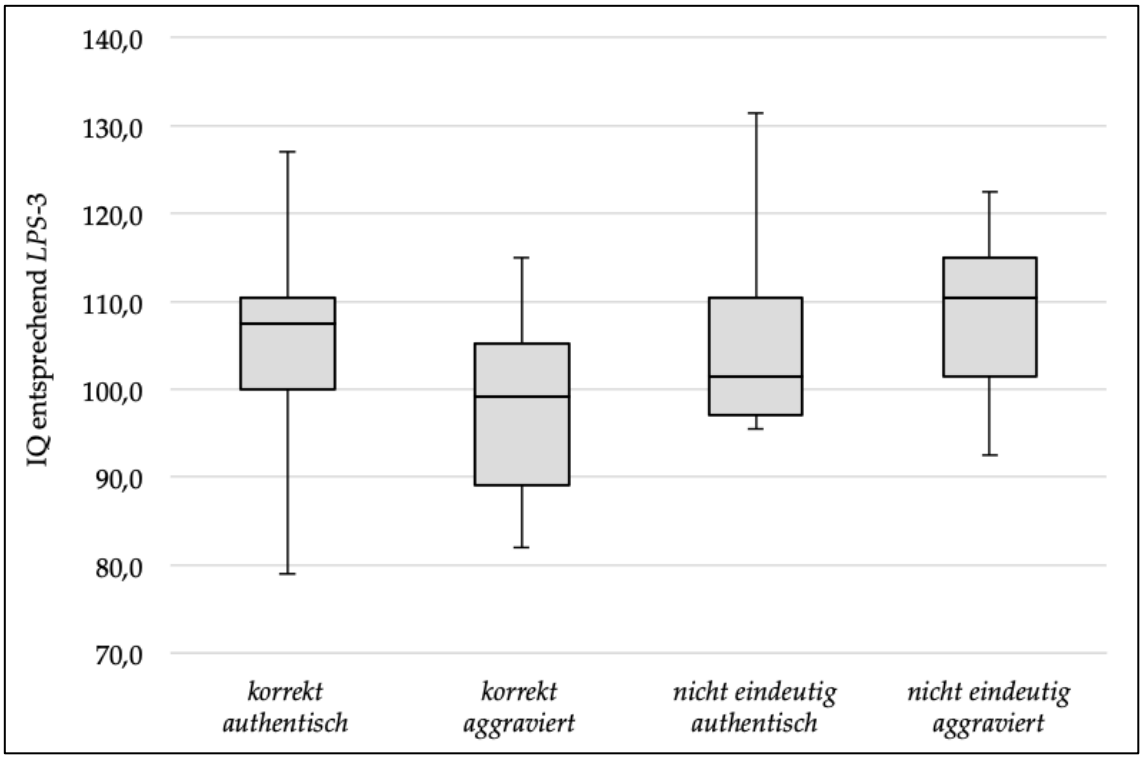

Abbildung 4. IQ nach *LPS*-Untertest 3 in den Untersuchungsgruppen

Wie die Abbildungen 1-4 (siehe vorherige Seiten) darstellen, finden sich signifikante Unterschiede der Intelligenzmaße *Gesamt-IQ*, *Verbal-IQ* (*Sozialformel*) und *LPS*-Untertests 1+2 zwischen einzelnen Gruppen.

Post hoc (Bonferroni-Korrektur des kumulierten α-Niveaus) zeigten sich in der Variable *Sozialformel Gesamt-IQ* signifikante Einzelunterschiede zwischen den Gruppen *korrekt authentisch* und *nicht eindeutig authentisch* (p=0,037, deskriptiv IQ *korrekt authentisch* > *nicht eindeutig authentisch*) sowie zwischen *nicht eindeutig authentisch* und *nicht eindeutig aggraviert* (p=0,028, deskriptiv IQ *nicht eindeutig aggraviert* > *nicht eindeutig authentisch*). Alle übrigen Einzelvergleiche erwiesen sich für dieses Intelligenzmaß als statistisch nicht signifikant.

Für die Variable *Sozialformel Verbal-IQ* fanden sich ebenfalls signifikante Einzelvergleiche zwischen den Gruppen *korrekt authentisch* und *nicht eindeutig authentisch* (p=0,024, deskriptiv IQ *korrekt authentisch* > *nicht eindeutig authentisch*) sowie zwischen *nicht eindeutig authentisch* und *nicht eindeutig aggraviert* (p=0,016, deskriptiv IQ *nicht eindeutig aggraviert* > *nicht eindeutig authentisch*). Alle übrigen Einzelvergleiche erwiesen sich auch hier als statistisch nicht signifikant.

Für die Variable *LPS* 1+2 unterschieden sich lediglich die Gruppen *korrekt aggraviert* und *nicht eindeutig aggraviert* signifikant (p=0,020, deskriptiv IQ *nicht eindeutig aggraviert* > *korrekt aggraviert*).

Hypothese 3 wird somit für beide Maße der *Sozialformel* und *LPS* Untertests 1 + 2 bestätigt. Ausschließlich für *LPS* Untertest 3 ließ sich keine signifikante Interaktion finden.

In einem nächsten Auswertungsschritt wurde der Frage nachgegangen, welche der eingesetzten Intelligenzmaße die höchste Aussagekraft in Bezug auf die Trennung von *authentischen* vs. *aggravierten* Symptomschilderungen aufweist (siehe Hypothese 4). Zu diesem Zweck wurde ein schrittweises lineares Regressionsmodell zur Vorhersage der Zugehörigkeit zu den Untersuchungsgruppen (*authentisch* vs. *aggraviert*) mit allen vier Intelligenzmaßen als Prädiktoren für diejenigen Probandinnen berechnet, bei denen das *SIRS-2* in einer unbestimmten Einordnung resultierte (n=18).

Ein solches Modell schloss lediglich die Variable *Verbal-IQ* der *Sozialformel* als bedeutsamen Prädiktor ein (β=0,707; p=0,001) und erreichte eine Varianzaufklärung von R^2_{corr}=0,468.

Aus den empirischen Daten der vorliegenden Studie ist für diese Studiensubgruppe mit uneindeutigem *SIRS*-2-Ergebnis ein Cut-Off von *Verbal-IQ* ≥ 104 (ROC AUC=0,962) plausibel,

um zwischen den Untersuchungsbedingungen *authentisch* vs. *aggraviert* zu trennen. Die Verteilung der Zuordnung findet sich in **Tabelle 7**.

Tabelle 7. Uneindeutiges *SIRS-2*-Ergebnis (n=18) in Relation zum *Verbal-IQ* der *Sozialformel*

Verbal-IQ	*authentisch* (n=5)	*aggraviert* (n=13)
<104	5	1
≥ 104	0	12

Es zeigt sich also, dass Hypothese 4 durch die Identifikation des signifikanten Prädiktors *Verbal-IQ* entsprechend der verwendeten *Sozialformel* bestätigt wird.

4.4 *SIRS-2*-Subskalen und Intelligenz in den Untersuchungsgruppen

Zur Klärung der Frage, welche der verwendeten Intelligenzmaße mit den Subskalen des *SIRS-2* assoziiert sind, wird im Folgenden zur Bearbeitung von Hypothese 5 die Korrelation der eingesetzten IQ-Maße mit den *SIRS-2*-Subskalen (Produktmomentkorrelation) für die Gesamtpopulation sowie für die zugeteilten Untersuchungsbedingungen *authentisch* und *aggraviert* berichtet (siehe
Tabelle 8 bis **Tabelle 10**, auf den folgenden Seiten).

Tabelle 8. *SIRS-2*-Skalen und Intelligenz in der Gesamtgruppe (n=60)

	Sozialformel		***LPS***	
***SIRS*-2-Skalen**	*Gesamt-IQ*	*Verbal-IQ*	1+2	3
Primärskalen				
Seltene Beschwerden (RS)	-0,253	-0,230	-0,197	-0,127
Symptomkombinationen (SC)	-0,237	-0,226	-0,226	-0,191
Unglaubhafte oder absurde Beschw. (IA)	-0,261*	-0,224	-0,258*	-0,072
Offenkundige Beschwerden (BL)	-0,193	-0,189	-0,096	-0,082
Subtile Beschwerden (SU)	-0,090	-0,120	0,055	-0,025
Selektivität der Beschwerden (SEL)	-0,104	-0,103	0,013	-0,006
Schweregrad der Beschwerden (SEV)	-0,165	-0,197	-0,030	-0,088
Geschilderte vs. beobachtete Beschw. (RO)	-0,247	-0,219	-0,200	-0,109
RS-total	-0,361**	-0,309*	-0,347**	-0,315*
Zusatzskalen				
Direkte Einschätzung der Ehrlichkeit (DA)	-0,082	-0,168	-0,024	-0,085
Defensive Beschwerden (DS)	0,218	0,192	0,324*	0,126
Unwahrscheinliche Fehler (IF)	-0,319*	-0,324*	-0,370**	-0,302*
Übergenaue Beschwerdenschilderung (OS)	-0,485***	-0,371**	-0,432***	-0,390**
Inkonsistente Symptome (INC)	-0,273*	-0,258*	-0,424***	-0,317*
***Adherence* (%)**	0,101	0,078	0,072	0,188

* p<0,05; ** p<0,01; *** p<0,001

In der Gesamtgruppe ging laut Tabelle 8 die Schilderung *unglaubhafter* oder *absurder Symptome* (IA) mit einer niedrigen Intelligenz einher (*Gesamt-IQ*, *LPS* Untertests 1+2). Gleichermaßen korrelierten Werte der Subskalen *Seltene Beschwerden* (RS-total), *Übergenaue Beschwerdenschilderung* (OS), *Unwahrscheinliche Fehler* (IF) und *Inkonsistente Symptome* (INC) signifikant negativ mit allen erhobenen Intelligenzmaßen.

Tabelle 9. *SIRS-2*-Skalen und Intelligenz bei *authentischer* Symptomdarlegung (n=30)

	Sozialformel		***LPS***	
***SIRS*-2-Skalen**	*Gesamt-IQ*	*Verbal-IQ*	1+2	3
Primärskalen				
Seltene Beschwerden (RS)	-0,245	-0,224	-0,113	-0,067
Symptomkombinationen (SC)	-0,088	-0,100	-0,099	0,101
Unglaubhafte oder absurde Beschw. (IA)	-0,352	-0,342	-0,279	0,258
Offenkundige Beschwerden (BL)	-0,384*	-0,398*	0,000	0,171
Subtile Beschwerden (SU)	-0,312	-0,302	0,026	0,061
Selektivität der Beschwerden (SEL)	-0,214	-0,188	0,090	0,188
Schweregrad der Beschwerden (SEV)	-0,462**	-0,489**	-0,073	-0,005
Geschilderte vs. beobachtete Beschw. (RO)	-0,380*	-0,304	-0,222	0,086
RS-total	-0,394*	-0,346	-0,362*	-0,372*
Zusatzskalen				
Direkte Einschätzung der Ehrlichkeit (DA)	-0,247	-0,287	-0,052	-0,328
Defensive Beschwerden (DS)	0,201	0,197	0,362*	0,086
Unwahrscheinliche Fehler (IF)	-0,430*	-0,363*	-0,474**	-0,412*
Übergenaue Beschwerdenschilderung (OS)	-0,394*	-0,389*	-0,362*	-0,283
Inkonsistente Symptome (INC)	-0,115	-0,101	-0,168	-0,304
***Adherence* (%)**	0,513**	0,389*	0,433*	0,443*

* p<0,05; ** p<0,01; *** p<0,001

Für die Gruppe der *authentischen* Symptomschilderungen fanden sich negative Zusammenhänge zwischen *Gesamt-* und *Verbal-IQ* der *Sozialformel* und der Subskala *Offenkundige Beschwerden* (BL) sowie *Schweregrad der Beschwerden* (SEV). Der oben berichtete Zusammenhang zwischen *RS-total* und den Intelligenzmaßen fand sich bei der authentischen Symptomschilderung ebenfalls (exklusive *Verbal-IQ*). Auch *Übergenaue Beschwerdenschilderung* (OS) und *Unwahrscheinliche Fehler* (IF) wiesen im Wesentlichen die oben berichteten Zusammenhänge auf. Lediglich für *Inkonsistente Symptome* (INC) fand sich in dieser Gruppe kein signifikanter Zusammenhang. Zusätzlich korrelierte die *Adherence* positiv mit allen Intelligenzmaßen (siehe Tabelle 9).

Tabelle 10. *SIRS-2*-Skalen und Intelligenz bei *aggravierter* Symptomdarlegung (n=30)

	Sozialformel		***LPS***	
***SIRS*-2-Skalen**	*Gesamt-IQ*	*Verbal-IQ*	1+2	3
Primärskalen				
Seltene Beschwerden (RS)	-0,316	-0,353	-0,286	-0,186
Symptomkombinationen (SC)	-0,378*	-0,430*	-0,341	0,298
Unglaubhafte oder absurde Beschw. (IA)	-0,276	-0,261	-0,316	0,150
Offenkundige Beschwerden (BL)	-0,135	-0,203	-0,147	-0,104
Subtile Beschwerden (SU)	0,226	0,016	0,315	0,178
Selektivität der Beschwerden (SEL)	0,118	-0,030	0,125	0,088
Schweregrad der Beschwerden (SEV)	0,040	-0,152	0,123	0,037
Geschilderte vs. beobachtete Beschw. (RO)	-0,215	-0,261	-0,219	-0,139
RS-total	-0,456*	-0,436*	-0,425*	-0,339
Zusatzskalen				
Direkte Einschätzung der Ehrlichkeit (DA)	-0,006	-0,141	-0,001	0,041
Defensive Beschwerden (DS)	0,513**	0,368*	0,565***	0,499**
Unwahrscheinliche Fehler (IF)	-0,564***	-0,438*	-0,433*	-0,382*
Übergenaue Beschwerdenschilderung (OS)	-0,299	-0,347*	-0,404*	-0,307
Inkonsistente Symptome (INC)	-0,387*	-0,424*	-0,601***	-0,348
***Adherence* (%)**	-0,202	-0,133	-0,164	-0,046

* p<0,05; ** p<0,01; *** p<0,001

Für die Gruppe der *aggravierten* Symptombeschreibung fanden sich entsprechend Tabelle 10 negative Zusammenhänge zwischen den IQ-Maßen der *Sozialformel* mit der Subskala *Symptomkombinationen* (SC). Bei den Subskalen *Offenkundige Beschwerden* (BL) und *Unwahrscheinliche Fehler* (IF) hingegen fanden sich keine Korrelationen. In der Subskala *RS-total* gingen niedrige Skalenwerte mit hohen Werten in den Maßen *Gesamt-IQ*, *Verbal-IQ* und *LPS* Untertest 1+2 einher. Wie zuvor berichtet finden sich auch hier enge Zusammenhänge zwischen *Defensiven Beschwerden* (DS), *Übergenauer Beschwerdenschilderung* (OS), *Unwahrscheinlichen Fehlern* (IF) und *Inkonsistenten Symptomen* (INC) mit den Intelligenzmaßen.

Aus dieser Übersicht ergeben sich zusammenfassend folgende differenzierte Zusammenhänge: Während bei aggravierter Symptomdarstellung höher ausgeprägte Intelligenz mit niedriger Ausprägung bezüglich angegebenen *Symptomkombinationen* (SC) und erhöhten Werten bezüglich *Defensiver Beschwerden* (DS) einhergehen, ist bei authentischer Symptomdarstellung ein höher ausgeprägtes Intelligenzniveau mit niedrigen Werten der Skalen *Offenkundige Beschwerden* (BL), *Schweregrad der Beschwerden* (SEV) und *Geschilderte vs. beobachtete Beschwerden* (RO), sowie einer erhöhten *Adherence* assoziiert. Hypothese 5 ist somit zu verifizieren.

5 Diskussion

Die vorliegende Arbeit hat zum Ziel, Zusammenhänge zwischen der intellektuellen Leistungsfähigkeit von Probandinnen und der Messbarkeit negativer Antwortverzerrung zu untersuchen. In diesem Kapitel werden die zuvor dargestellten Ergebnisse dieser Fragestellung im Licht des Theorieteils beurteilt und kritisch hinterfragt.

5.1 Limitationen der Arbeit

Zu Beginn werden Stichprobe, Studiendesign und Methoden der Operationalisierung kritisch hinterfragt und Verbesserungsvorschläge im Hinblick auf zukünftige Folgestudien herausgearbeitet.

5.1.1 Rekrutierung und Stichprobe

Wie in Kapitel 3.1.1 beschrieben, ist die dieser Arbeit zugrundeliegende Untersuchung Teil einer übergeordneten Validierungsstudie. Es waren daher von vorneherein zeitliche sowie methodische Rahmenanforderungen gesetzt, denen bei der Planung dieser Studie Rechnung getragen werden musste.

Die Auswahl geeigneter Probandinnen über die Ambulanzdatenbank und die telefonische Kontaktaufnahme boten sich als unkomplizierteste und erfolgversprechendste Herangehensweise an. Von den Personen, die die entsprechenden Teilnahmekriterien der Vorauswahl erfüllten, ließen sich jedoch nur ca. 32% davon zur Teilnahme an der Untersuchung bewegen. Gründe dafür sind in der telefonischen Erreichbarkeit, in der Distanz zwischen Wohnort und Klinik sowie in der Motivationslage und Termintreue der Personen zu sehen. Hier muss von einer systematischen Stichprobenverzerrung ausgegangen werden, da die Motivationslage, ohne monetäre Anreize teilzunehmen, sowie die Compliance nach der Terminvereinbarung vermutlich mit motivationalen Aspekten und den intellektuellen Fähigkeiten der jeweiligen Person korrelieren. Zukünftige Arbeiten könnten versuchen, dieses Selektions-Bias zu minimieren, indem man den Aufwand zur Teilnahme an der Untersuchung minimiert oder durch eine Aufwandsentschädigung egalisiert.

Aufgrund enger zeitlicher Vorgaben der Rahmenstudie wurde mit n=60 Probandinnen ein Stichprobenumfang angestrebt und erreicht, der sich personell und logistisch gut bewältigen ließ und gleichzeitig auf der Basis einer zuvor durchgeführten Stichprobenschätzung und Power-Analyse auch die Darstellung signifikanter Zusammenhänge bei mittleren Effektstärken ermöglichte. Künftige Arbeiten sollten dennoch größere Stichproben untersuchen, um auch kleinere Effekte statistisch signifikant abbilden zu können.

Die Altersspannweite sowie das Durchschnittsalter der getesteten Probandinnen bilden mit 24-65 bzw. 45,8 Jahren ein Kollektiv ab, das sich zum einen in der typischen Altersspanne für depressive Erkrankungen befindet und zum anderen ein für Erwerbsfähigkeit und Verrentungsanträge repräsentatives Alter hat (Busch et al., 2013).

Die Dominanz weiblicher Personen (70%) in der Untersuchungsstichprobe erklärt sich dadurch, dass bereits im nach Ein- und Ausschlusskriterien vorselektierten Grundkollektiv ein erhöhter Anteil von Frauen bestanden hat. Dies kann darauf zurückgeführt werden, dass ihnen im Allgemeinen ca. doppelt so häufig wie Männern die Diagnose Depression gestellt wird (Tölle und Windgassen, 2014). Auch verglichen mit der Geschlechterverteilung bei Arbeitsunfähigkeiten aufgrund psychischer Störungen bildet die erfasste Stichprobe eine Näherung an die Grundgesamtheit ab, in der Frauen mit ca. 63% überrepräsentiert sind (BKK Dachverband, 2017).

Hinsichtlich sonstiger sozioökonomischer Variablen ließen sich keine diskussionswürdigen Besonderheiten der Stichproben verzeichnen.

5.1.2 Studiendesign und Versuchsaufbau

Von einer Sicherung wissenschaftlich-ethischer Standards kann nach der Prüfung des Studiendesigns durch die Ethikkommission der Medizinischen Fakultät der *Martin-Luther-Universität Halle-Wittenberg* ausgegangen werden. Die Aufklärung über die Studieninhalte und die Einwilligung durch die Probandinnen erfolgten schriftlich und mit ausreichendem zeitlichen Vorlauf von mindestens drei Tagen.

Bei der eigentlichen Untersuchung wurde zuerst eine Randomisierung zur Untersuchungsbedingung (*authentisch* vs. *aggraviert*) ohne Zurücklegen der gezogenen Lose durchgeführt. Dadurch konnten systematische Unterschiede zwischen den Gruppen ausgeschlossen und eine

gleiche Gruppengröße erzielt werden. Darüber hinaus erwiesen sich die beiden erstellten Experimentalgruppen hinsichtlich ihrer soziodemografischen Merkmale als vergleichbar.

Die Instruktion mithilfe standardisierter Anweisungstexte und die zweizeitige mündliche Kontrolle des Instruktionsverständnisses vor und nach der Maßnahme zur Beschwerdenvalidierung sicherten das Verständnis der Untersuchungsbedingung durch die Probandinnen. Eine große Schwierigkeit stellte allerdings die Frage dar, inwieweit die standardisierte Instruktion, zu aggravieren, tatsächlich adäquate schauspielerische Leistungen hervorrief. Zwar wurden allen Probandinnen dieselben Informationen bezüglich des anzunehmenden Szenarios gegeben, jedoch ist davon auszugehen, dass die einzelnen Personen sich unterschiedlich gut in diese Situationen hineinversetzen konnten. Es wurde im Rahmen des *Adherence*-Checks zwar erfragt, inwieweit den Instruktionen Folge geleistet wurde, jedoch kann nicht erwartet werden, dass die Probandinnen hier einen objektiven Blick auf ihre eigene Leistung haben. Im Rahmen der logistisch machbaren Möglichkeiten ließ sich dieses Motivations-Bias weder evaluieren noch reduzieren und musste in Kauf genommen werden. Dieser Umstand ist bei der Beurteilung der Ergebnisse zu berücksichtigen. Inwieweit also davon ausgegangen werden kann, ob die in dieser Untersuchung erhobene Beschwerdenvalidität den Daten realer Begutachtungssituationen entspricht, müssen Folgearbeiten unter Einbezug echter Begutachtungsstichproben (bzw. unter Zuhilfenahme einer sogenannten *known-groups-Methode* mit Probandinnen, die sich in der Praxis als tatsächliche Übertreibende herausgestellt haben) oder unter Einbezug von ausgebildeten Schauspielpatientinnen zeigen.

Bei der Auswertung der erfragten Instruktionstreue (*Adherence*) zeigten sich zwar Unterschiede zwischen den Gruppen *authentischer* vs. *aggravierter* Symptomdarstellung mit höheren Werten in der ersten Teilstichprobe, jedoch wurde keine Angabe zur *Adherence* als derart unzureichend beurteilt, dass die Teilnehmerin aufgrund mangelnden Instruktionsverständnisses und -treue aus der Studie ausgeschlossen werden musste. Da es sich hierbei um eine subjektive Selbsteinschätzung der Probandinnen auf einer Ratingskala handelt, sind die produzierten Werte nicht miteinander vergleichbar und können nicht quantitativ (z.B. im Sinne einer Gewichtung) verwertet werden.

Durch die Aufteilung der Befunderhebung auf zwei Untersucherinnen und die Trennung zwischen den intelligenzbezogenen Vortests sowie der Zuweisung zu einer Untersuchungsgruppe einerseits und der Beschwerdenvalidierung mithilfe des *SIRS-2* andererseits wurde für eine

Verblindung der zweiten Untersucherin hinsichtlich der Gruppe gesorgt. Es konnten daher durch Erwartungen der Diagnostikerin hinsichtlich der Probandinnenmotivation bedingte Verzerrungen (Detektions-Bias) verhindert werden. Eine zusätzliche Verblindung der Probandin (Doppel-Verblindung), wie sie zur Optimierung der Studienqualität z.B. bei Medikamentenstudien durchgeführt wird, war wegen der notwendigen Instruktion der Motivationslage nicht möglich.

Hinsichtlich der erzielten Datenstruktur und deren Vollständigkeit ist das in der vorliegenden Arbeit gewählte Untersuchungsdesign als effektiv zu bewerten, da durch die geschulten Versuchsleiterinnen die unter ökonomischen Gesichtspunkten erhobene Stichprobe und das ausschließlich aus standardisierten und objektiven Testverfahren bestehende Instrumentarium ein Datensatz gewonnen wurde, der durch keinerlei fehlende Werte gekennzeichnet war.

5.1.3 Instrumentarium

Assessment der Intelligenz

Eine zentrale Herausforderung dieser Arbeit war es, innerhalb kurzer Zeit (ca. 15 Minuten) und mithilfe einfacher Verfahren eine valide Abschätzung der Intelligenz der Probandin zu erreichen. Hierzu wurden zum einen drei Untertests des *Leistungsprüfsystems* (Horn, 1983) und die Abschätzung anhand sozioökonomischer Variablen mithilfe einer *Sozialformel* (Jahn et al., 2013) durchgeführt. Es stellte sich während der Untersuchungsplanung die Frage, ob statt der Ursprungsversion von 1983 die zum Studienzeitpunkt seit kurzem veröffentlichte Neuauflage des *Leistungsprüfsystems* (*LPS-2*; Kreuzpointner et al., 2013) verwendet werden sollte. Der Vorteil davon wäre eine aktuellere Normierung gewesen, die die Auswirkungen des weiter unten diskutierten *Flynn-Effektes* (Zunahme gemessener Intelligenzwerte einer Bevölkerung über die Zeit) reduziert hätte. Es wurde dennoch die erste Auflage des *Leistungsprüfsystems* gewählt. Zum einen sollte der zu bestimmende Intelligenzquotient keineswegs als absolut mit anderweitig bestimmten IQ-Werten vergleichbares Intelligenzmaß dienen, sondern lediglich als stichprobeninterner Vergleichswert. Außerdem war die 1983 normierte Ursprungsstichprobe mit einem Umfang von n=10.000 Probandinnen etwa viermal so groß wie die Stichprobe der neueren Normierung. Man kann also bei der ersten Auflage von einer größeren Trennschärfe in den Extrembereichen der IQ-Verteilung ausgehen. Zusätzlich wurde die Neuauflage von

2013 insbesondere für Kinder und Jugendliche konzipiert, passt in dieser Hinsicht also nicht optimal zum in dieser Untersuchung verwendeten Studienkollektiv.

Im Zuge einer kritischen Betrachtung der Intelligenzschätzung mithilfe einer *Sozialformel* ist herauszustellen, dass Äußerungen der Probandinnen zu ihren Lebensumständen und Gewohnheiten durch Angehörige des Untersuchungsteams beurteilt und klassifiziert werden mussten. So wurden z.B. die in der Befragung angegebene Buchlektüre in durch Jahn et al. (2013) vorgegebene Kategorien eingeteilt, was grundsätzlich Beurteilungsfehler ermöglicht. Diese Klassifikation erfolgte jedoch in dieser Studie mit zeitlichem Abstand zur Befragung unter Verblindung der Probandinnenidentität und für alle Fälle durch dieselben zwei Personen, wodurch die systemische Verzerrung der Befunde auf ein Geringes reduziert sein sollte. Es wurde weiterführend darauf verzichtet, die Nachweise der angegebenen Bildungsparameter (Schulabschlusszeugnis, Fächernoten) einzuholen, da dies für die Probandinnen einen unverhältnismäßig großen Aufwand bedeutet hätte. Stattdessen wurde auf die Glaubhaftigkeit der gegebenen Informationen vertraut und eine vermutete leichte Verfälschung in Richtung höherer Intelligenz in Kauf genommen.

Bei Durchsicht der IQ-Werte lässt sich vor allem bei den Schätzungen mithilfe der *Sozialformel* bemerken, dass die Mittelwerte 9 bis 10 IQ-Punkte über dem Normwert von 100 (also dem Mittelwert der Normierungsstichprobe des *HAWIE-R*) liegen. Dieses Phänomen fiel schon bei den 2005 im Rahmen der Validierungsstudie gemessenen IQ-Werten auf (Jahn et al., 2013). Vor dem Hintergrund, dass die IQ-Wert-Normierung des *HAWIE-R* bereits 1988, also 17 Jahre vor der Datenerhebung durch Jahn et al. und ca. 28 Jahre vor Durchführung der hier berichteten Untersuchung erfolgt ist, kann man die Differenz des durchschnittlichen IQ-Werts als eine Auswirkung des *Flynn-Effekts* betrachten. Dieser besagt, dass der messbare IQ aus bisher nur unzureichend geklärten Gründen von Generation zu Generation steigt, nämlich – je nach Untersuchung – etwa 3 bis 7 IQ-Punkte pro Dekade (Flynn, 2007, 2012; Pietschnig und Voracek, 2015; Shenk, 2017). Eine Anwendung von Äquivalenztabellen zur Anpassung der IQ-Werte (z.B. Erzberger und Engel, 2010) wäre zwar möglich, ist aber für die hier zu beantwortende Fragestellung nach einer stichprobeninternen Korrelation mit dem *SIRS-2*-Ergebnis nicht notwendig.

Es muss dringend angemerkt werden, dass man hier insgesamt bei den beiden angewandten Verfahren zur Intelligenzbestimmung nur von einer orientierenden Abschätzung der

intellektuellen Fähigkeiten ausgehen kann, keinesfalls von einer umfassenden und genauen Testung, die mehrere Stunden in Anspruch genommen hätte. Die Ergebnisse sind daher als Richtwerte zu verstehen und liefern ausschließlich im Kontext der gesamten Stichprobe verwertbare Daten. Weiterführende Studien könnten hier zur genaueren IQ-Bestimmung ausführlichere und aktuellere Verfahren wie den *Wechsler-Intelligenztest für Erwachsene* (Wechsler, 2012) oder den *Intelligenz-Struktur-Test* (Liepmann et al., 2007) anwenden, da bei letzterem auch eine exakte Trennung multipler Intelligenzfaktoren und in deren Folge eine genauere Unterscheidung *fluider* von *kristalliner* Intelligenz möglich wäre.

Beschwerdenvalidierung mithilfe des SIRS-2

Die Anwendung der deutschen Übersetzung des *SIRS-2* wurde entsprechend dem Testmanual der englischen Ursprungsversion (Rogers et al., 2010) durchgeführt. Vor Beginn der Studie wurden alle eingesetzten Gutachterinnen durch das Team des übergeordneten Studienprojektes (Schmidt et al., 2019) in der Durchführung und Auswertung des Testverfahrens geschult.

Während der *allgemeinen Befragung* wurde ausgesprochen selten ein „eingeschränktes Ja" bzw. „manchmal" als Antwort angegeben. Das kann daran gelegen haben, dass die geschlossene Frageform ein dichotomes Antwortverhalten gefördert hat, was durchaus im Sinne der Teststrategie ist.

Durch die Abfrage der Items in Form eines strukturierten Interviews ließ sich gut überprüfen, ob die Fragen von der Probandin verstanden wurden. Hier konnte die Unsicherheit dadurch, dass die deutsche Version des *SIRS-2* noch kein etabliertes Testverfahren und potenziell missverständlich ist, weitestgehend minimiert werden (zusätzlich zu der nach der Testübersetzung durchgeführten Validierung mit Hilfe von Testprobandinnen).

Die Auswertung erfolgte unter Verblindung der Untersuchungsbedingung sowie der Probandinnenidentität durch einen geschulten Gutachter.

Für eine umfassende kritische Würdigung des *SIRS-2* sei an dieser Stelle auf die dem Validierungsprojekt übergeordnete Arbeit verwiesen (Schmidt et al., 2019).

5.2 Interpretation der Ergebnisse

Die in Kapitel 4 herausgearbeiteten Ergebnisse sollen im Folgenden in ihrer Bedeutung bewertet und in theoretische Aussagen überführt werden.

5.2.1 *SIRS-2*-Subskalen und Instruktionstreue in den Untersuchungsgruppen

Bei Betrachtung der Subskalenwerte in den Untersuchungsgruppen zeigen sich bei den aggravierten Symptomschilderungen für alle Primär- und Zusatzskalen signifikant höhere Werte außer in der Zusatzskala *Direkte Einschätzung der Ehrlichkeit*. Hier gibt es nur geringe Skalenunterschiede zwischen Übertreibenden und Authentischen. Diese Skala wird ausführlich im Testmanual (Schmidt et al., 2019) diskutiert. Es handelt sich hier um eine Zusatzskala, deren Ausprägung nicht unmittelbar in den Unterscheidungsalgorithmus einfließt. Mit ihrer Aussagekraft über die allgemeine Offenheit der Probandin gegenüber Behandelnden gibt sie eher eine Zusatzinformation für die Gutachterin. Im Kontext der Validierungsstudie mit instruierten Probandinnen wäre denkbar, dass die direkte Abfrage von Authentizität der berichteten Beschwerden in so engem zeitlichen Zusammenhang zur Instruktion glaubhaft zu aggravieren steht, dass hier häufiger eine hohe Authentizität bekundet wird, als das in der klinischen Vergleichsgruppe der Fall war.
Die Befunde der Instruktionstreue (*Adherence*) zeigen, dass es den authentischen Probandinnen signifikant leichter fiel glaubwürdig zu erscheinen, als den aggravierenden Probandinnen zu übertreiben. Das erscheint nachvollziehbar, da das Übertreiben – und nicht die authentische Beschwerdenschilderung – die herausfordernde schauspielerische Leistung verlangte.

5.2.2 *SIRS-2*-Klassifikation und Intelligenz in den Untersuchungsgruppen

Tabelle 4 gibt einen Hinweis darauf, wie valide das *SIRS-2* die Beschwerdenvalidität operationalisiert: Von 60 Studienteilnehmerinnen wurden mehr als $^2/_3$ korrekt als ehrliche oder übertreibende Probandinnen klassifiziert, bei 18 Personen ergab sich ein uneindeutiges Klassifikationsergebnis, und nur eine Probandin wurde durch das *SIRS-2* falsch beurteilt.
Es wurde ermittelt, dass sich die IQ-Werte der korrekt und nicht korrekt klassifizierten Probandinnen nicht signifikant unterschieden (siehe Tabelle 5). Dementsprechend ist nicht davon

auszugehen, dass die Messbarkeit von negativer Antwortverzerrung pauschal und systematisch durch den IQ bestimmt wird.

Die vergleichende Betrachtung der IQ-Werte für die korrekt und die nicht eindeutig klassifizierten Probandinnen je nach Untersuchungsbedingung (Tabelle 6, Abbildung 1 bis Abbildung 4) zeichnet dagegen ein anderes Bild:

In den Intelligenzmaßen der *Sozialformel* nach Jahn (*Gesamt-IQ* und *Verbal-IQ*, Abbildung 1 und Abbildung 2) haben korrekt authentisch klassifizierte Probandinnen signifikant höhere IQ-Werte als die nicht eindeutig klassifizierten. Man kann daraus ableiten, dass die Beschwerdenvalidierung ehrlicher Probandinnen mithilfe des *SIRS-2*, also die korrekte Detektion authentischer Symptomschilderung, bei höherer abgeschätzter Intelligenz besser funktioniert als bei niedrigeren IQ-Werten. Gleichzeitig haben unter den uneindeutig klassifizierten Studienteilnehmerinnen die Übertreibenden deutlich höhere IQ-Werte als die Authentischen. Wenn also ein uneindeutiges Klassifikationsergebnis vorliegt, ist ein höherer IQ in der Abschätzung mithilfe der *Sozialformel* eher mit einer aggravierten Symptomdarstellung assoziiert, während eine geringere Intelligenz eher bei authentischer Symptomschilderung zu finden war. Es zeigten sich in diesem Zug in den vier Untersuchungsgruppen (*korrekt authentisch*, *korrekt aggraviert*, *nicht eindeutig authentisch*, *nicht eindeutig aggraviert*) signifikant unterschiedliche IQ-Werte (Tabelle 6).

Hier wurde anhand der verfügbaren Daten von Probandinnen mit uneindeutigem *SIRS-2*-Ergebnis post hoc ein Trennwert intellektueller Leistungsfähigkeit ermittelt, der am ehesten geeignet ist, zwischen Personen mit authentischer und aggravierter Symptomschilderung zu differenzieren. Dieser Trennwert erwies sich bei einem *Verbal-IQ* von 104 als gut geeignet, ist aber aufgrund der geringen Fallzahl und der oben genannten Einschränkungen in der Operationalisierung von Intelligenz nur mit großer Vorsicht zu generalisieren (siehe Tabelle 7). Folgestudien sollten daher unter Einbeziehung größerer Stichproben eine derartige Analyse wiederholen, um durch einfache Schätzung der Intelligenz z.B. mittels *Sozialformel* eine Methode anzubieten, die gerade in unsicheren Fällen eine Differenzierung erlaubt, in welchen Fällen eine zusätzliche Diagnostik hinsichtlich negativer Antwortverzerrung naheliegt und in welchen eher davon abgesehen werden kann.

Für die Intelligenzmaße des *Leistungsprüfsystems* sind die Befunde als weniger aufschlussreich zu beurteilen. Lediglich in den Untertests 1 und 2 haben unter den Übertreibenden die korrekt

klassifizierten einen signifikant geringeren IQ, als die uneindeutig klassifizierten Personen. Die Beschwerdenvalidierung durch das *SIRS-2* funktioniert bei Aggravierenden mit geringerer Intelligenz also besser als bei intelligenteren Probandinnen.

Da es sich bei dem *Verbal-IQ* um ein Maß *kristalliner* Intelligenz handelt (Wechsler, 1955) und gerade dieser eine starke Korrelation mit der Beschwerdenvalidierungstestgüte aufweist, während klassische Maße *fluider* Intelligenz nach Cattell (z.B. der *LPS* Untertest 3) einen weniger deutlichen Zusammenhang zeigen, muss davon ausgegangen werden, dass vor allem die *kristalline* Intelligenz Probandinnen dabei hilft, Teststrategien des *SIRS-2* zu erkennen und zu beeinflussen. Hier kann man eine allgemeine verbale Befähigung zum Umgang mit kritischen Fragen und das Erkennen der übergeordneten Teststrategie des *SIRS-2* als Schlüsselkompetenzen zum „Überlisten" des Tests vermuten. Auch wäre eine bessere Kenntnis der typischen Symptome der aggravierten Erkrankung durch intelligentere Probandinnen ein Erklärungsansatz.

Zusammenfassend lässt sich für die klinische Verwendung der hier dargestellten Erkenntnisse feststellen, dass die *Sozialformel* nach Jahn (2013) ein zeitsparendes und valides Verfahren darstellt, um Ergebnisse der Beschwerdenvalidierung abzusichern. Sie kann insbesondere für den Fall eines uneindeutigen Testergebnisses des *SIRS-2* als wichtige zusätzliche Informationsquelle für die Bewertung der Befunde dienen. Weitere Studien zur Erweiterung dieses und anderer Beschwerdenvalidierungstests um die Intelligenzschätzung mithilfe einer Sozialformel wären wünschenswert. Hier sei auch insbesondere auf neue Verfahren der Detektion negativer Antwortverzerrung verwiesen, die möglicherweise von einer zusätzlichen Intelligenzmessung bzw. -schätzung profitieren könnten (z.B. *SRSI*, Merten et al., 2016; *BEVA*, Walter et al., 2017).

5.2.3 *SIRS*-2-Subskalen und Intelligenz in den Untersuchungsgruppen

Die in Kapitel 4.4 dargestellten Zusammenhänge zwischen Subskalenergebnissen und Intelligenzmaßen in der Gesamtgruppe sowie bei den authentischen und aggravierenden Probandinnen kann für die zukünftige Überarbeitung des *SIRS-2* wichtige Hinweise auf systematische Verzerrungen geben. So haben z.B. authentische Probandinnen in der Subskala

Unwahrscheinliche Fehler bei einem niedrigeren IQ einen höheren Skalenwert als bei einem höheren IQ. Dieses Phänomen ist bereits bei Rogers et al. (2010) beschrieben und lässt sich durch die vom IQ abhängige Lösungskompetenz der gestellten Aufgaben (Gegenteile, Reime bilden) erklären. Auch diese Skala fließt im Übrigen als Zusatzskala nicht in die Klassifikation ein, sondern soll lediglich Hinweise auf eine zusätzliche Aggravation kognitiver Defizite liefern.
Bei der aggravierten Symptomdarstellung haben eher Probandinnen mit einem höheren IQ in der Skala *Defensive Beschwerden* hohe Werte produziert, während Personen mit einem geringen IQ in den Skalen *RS-total*, *Unwahrscheinliche Fehler* und *Inkonsistente Symptome* über dem Gesamtgruppendurchschnitt lagen. Die verschiedenen Strategien, die das *SIRS-2* bei der Identifikation von negativer Antwortverzerrung verfolgt, funktionieren also jeweils in Abhängigkeit vom IQ der Probandin unterschiedlich gut. Dies rechtfertigt die multiplen Teststrategien des *SIRS-2*, wodurch eine Anwendung auch bei vom Durchschnitt abweichender Intelligenz bis hin zu leichter geistiger Behinderung möglich wird (Rogers et al., 2010). Die oben genannten Korrelationen können für die Testweiterentwicklung interessant sein, für die konkrete Begutachtungssituation ist vor der Anwendbarkeit der Erkenntnisse jedoch dringend noch eine Absicherung anhand einer größeren Stichprobe notwendig.
Bei Betrachtung der angegebenen Instruktionstreue (*Adherence*) zeigte sich bei den authentischen Personen eine deutliche Korrelation mit der Intelligenz – vor allem für den *Gesamt-IQ* der *Sozialformel*. Intelligentere Probandinnen gaben an, dass sie der Instruktion, ihre Beschwerden authentisch zu schildern, besser Folge geleistet haben. Dahingegen lässt sich ein genereller signifikanter Zusammenhang zwischen Intelligenz und *Adherence* bei den Übertreibenden nicht finden. Eine mögliche Erklärung könnte sein, dass Menschen mit geringeren intellektuellen Fähigkeiten im Sinne der kognitiven Triaden nach Beck (1979) eine grundsätzlich eher negativ verzerrte Selbsteinschätzung ihrer Leistung vornehmen – möglicherweise infolge erlebten negativen Feedbacks. Dies fällt im Vergleich mit den weniger „bescheidenen" intelligenteren Personen vor allem im Zuge der Selbstevaluation bei der leichteren Aufgabe (Symptome *authentisch* zu schildern) auf. Bei der schwierigeren Aufgabe (zu aggravieren) beurteilen sich die intelligenteren Probandinnen deutlich kritischer, weshalb sich hier im Resultat ein geringerer Gruppenunterschied ergab.

5.3 Die Befunde im Licht des theoretischen Kontextes

In der theoretischen Konzeption der vorliegenden Arbeit zeigte sich, dass zum Zusammenhang von Intelligenz und der Messbarkeit negativer Antwortverzerrung bisher wenige gesicherte Erkenntnisse vorliegen. Die Ergebnisse dieser Arbeit ergänzen das verfügbare Wissen substanziell, da erstmals infolge einer randomisierten Zuweisung zu einer Motivationslage (*authentisch* vs. *aggraviert*) die Beschwerdenvalidierungstestgüte und die Intelligenz der Probandinnen auf Zusammenhänge untersucht wurden. Die Arbeit liefert also Befunde, die die Störvariable der interindividuell verschiedenen Motivationslagen ausschließen.

Für die konkrete Begutachtungssituation, in der eine Sachverständige anhand eines kurzen Intelligenzassessments das Ergebnis des *SIRS-2* besser zu beurteilen und einzuordnen versucht, können die dargestellten Ergebnisse vor allem folgenden Schluss zulassen: Je nach Klassifikationsergebnis des *SIRS-2* liefert die Intelligenz wichtige Hinweise für die Bewertung der Befunde. Hier ist es vor allem die Intelligenzschätzung des *Verbal-IQ* mithilfe der *Sozialformel* nach Jahn et al. (2013), die auf Grundlage der dargestellten Daten als hilfreich herausgestellt werden kann. Die von den Fachgesellschaften empfohlene multimethodale Absicherung der Beschwerdenvalidität (siehe Kapitel 0) kann hiermit um ein weiteres testpsychologisch ermittelbares Maß erweitert werden. Diese Ergebnisse dürfen jedoch nicht ohne weiteres für andere Beschwerdenvalidierungs- und Intelligenztests bzw. -schätzformeln generalisiert werden. Bei der Testung mithilfe des *SIRS-2* und der Intelligenzschätzung anhand der *Sozialformel* nach Jahn et al. kann jedoch Folgendes festgehalten werden: Bei uneindeutigen Ergebnissen des *SIRS-2* stellen vor allem der *Verbal-IQ*, aber auch der *Gesamt-IQ* Maße dar, die einen wichtigen Hinweis auf die Authentizität der berichteten Beschwerden liefern können. Als intelligenter eingeschätzte Probandinnen sind hier signifikant häufiger Übertreibende.

Für die Intelligenzforschung liefert die hier beschriebene Untersuchung einen weiteren Verwendungsbereich für Sozialformeln im Allgemeinen. Die einfache und zeitsparende Durchführbarkeit qualifiziert die IQ-Schätzverfahren für den Kontext der Begutachtung und erweitert das Instrumentarium um standardisierte und gut validierte Intelligenzschätzmaße.

Eine weitere Erkenntnis ist, dass höhere Werte vor allem der *kristallinen* Intelligenzmaße nach Cattell (1941) mit schlechterer Messbarkeit in der Beschwerdenvalidierung korrelieren. Sie bilden also zuverlässiger als die *fluide* Intelligenz die für das Überlisten von

Beschwerdenvalidierungstests notwendige intellektuelle Kompetenz ab. In der vorliegenden Studie haben sich für die eher *kristallinen* Maße, also die beiden mithilfe der *Sozialformel* abgeschätzten Werte und die *LPS*-Untertests 1 und 2 signifikante Zusammenhänge zu den *SIRS-2*-Ergebnissen ergeben. Der einzige verwendete Teiltest in Hinblick auf die fluide Intelligenz, der *LPS*-Untertest 3, hingegen ergab keine nutzbaren eindeutigen Ergebnisse.

Auch diese Beobachtung muss anhand größerer Stichproben und ausreichend zwischen *fluiden* und *kristallinen* Maßen differenzierender Intelligenzmessung gefestigt werden, bevor man theoretische Aussagen sicher ableiten kann. Hier können unter anderem Tests zur Erhebung rein *kristalliner* Intelligenz wie z.B. *Mehrfach-Wortschatztests* mit *fluiden* Intelligenzmaßen aus z.B. *Raven-Matrizentests* in ihrem Zusammenhang zur Messbarkeit der Antwortverzerrung verglichen werden, um möglichst valide Daten zu erhalten. Ein wichtiger Vorteil bei der Nutzung *kristalliner* Intelligenzmaße für die Begutachtung von Beschwerdenvalidität ist, dass diese weniger abhängig von der aktuellen Symptomschwere psychischer Erkrankungen sind (siehe Kapitel 2.4.4) und z.B. bei Depressionen besser eingesetzt werden können.

Für die Messung der Beschwerdenauthentizität hat sich das *SIRS-2* in Verbindung mit der Erfassung intellektueller Leistungsfähigkeit im Rahmen dieser Studie als wertvolles Instrument herausgestellt. Es liegt als weiterer Teilerfolg einer auch in Deutschland voranschreitenden Forschung im Bereich standardisierter Beschwerdenvalidierung nun also infolge der übergeordneten Studie eine gut validierte deutsche Version des *SIRS-2* (Rogers et al., 2010) vor (Schmidt et al., 2019). Der in Kapitel 2.3 dargestellten hohen Bedeutung einer validen Detektion bewusster Antwortverzerrung im Bereich der Psychiatrie wird also Rechnung getragen.

Im Kontext von Begutachtung psychischer Erkrankungen kann die gesellschaftliche Bedeutung hochwertiger, valider Tests nur noch einmal unterstrichen werden. Hier leistet die vorliegende Studie einen Beitrag zu der von den Fachgesellschaften geforderten Erarbeitung und Nutzung valider Beschwerdenvalidierungstests (Deutsche Gesellschaft für Neurowissenschaftliche Begutachtung, 2013; Schneider et al., 2016, S. 480 f.).

Es lässt sich zusammenfassen, dass in dieser Arbeit dem übergeordneten Ziel einer fairen sowie wirtschaftlichen Begutachtung psychischer Erkrankungen durch die Evaluation eines validen Testverfahrens und den Miteinbezug nutzbarer Zusatzvariablen entgegengearbeitet wurde.

5.4 Ausblick

Die hier durchgeführte Studie kann wichtige Hinweise auf die Rolle von Intelligenz für die Messbarkeit negativer Antwortverzerrung bieten. Die Ergebnisse sollten jedoch in Zukunft mit anderen, größeren Stichproben und für andere Testverfahren repliziert werden, bevor man davon ausgehen kann, dass es sich um gesicherte Erkenntnisse handelt. Auch für Menschen mit anderen psychiatrischen Erkrankungen, für andere Settings von Begutachtung (z.B. bei der Frage nach Schuldfähigkeit in der Forensik), und auf der Grundlage anderer Studiendesigns sollte zusätzliche Forschung erfolgen. Denkbar wäre ein der vorliegenden Untersuchung ähnliches experimentelles Studiendesign, bei dem eine größere Anzahl von Probandinnen rekrutiert wird, die sich aus geschulten Schauspielerinnen oder bereits klinisch auffällig gewordenen Übertreibenden zusammensetzt. So könnte die in dieser Arbeit nicht zu verhindernde suboptimale Instruktionstreue verbessert werden. Neben einer ausführlichen Rollenbeschreibung mit der konkreten Instruktion, Beschwerden einer schweren psychischen Erkrankung zu schildern und zu verkörpern, könnte man die Intelligenz mithilfe eines differenzierten Intelligenztests, wie z.B. dem *Wechsler-Intelligenztest für Erwachsene* (Wechsler, 2012) oder dem *Intelligenz-Struktur-Test* (Liepmann et al., 2007) messen. Zusätzlich würde die Durchführung des *SIRS-2* oder eines vergleichbaren Beschwerdenvalidierungstests erfolgen und die Befunde schließlich auf signifikante Zusammenhänge überprüft werden.

Die Ergebnisse würden klären, ob von einem Zusammenhang zwischen Intelligenz und den Testgütekriterien von Beschwerdenvalidierungstests allgemein ausgegangen werden kann. Anzumerken ist jedoch, dass bei dem Einsatz von Schauspielerinnen an negativer Antwortverzerrung vor allem die *Simulation* gemessen würde, weniger die *Aggravation*, da es sich vermutlich schwierig gestalten würde, ein Studienkollektiv psychisch kranker (z.B. depressiver) Schauspielerinnen zu rekrutieren. Die Ergebnisse wären dann mit der hier vorliegenden Arbeit also nicht uneingeschränkt vergleichbar, da es sich streng genommen nicht um dasselbe gemessene Konstrukt handelt.

Sollten sich die hier angedeuteten Einflüsse in Folgestudien bestätigen lassen, würde dies den wissenschaftlichen Erkenntnisstand substantiell erweitern. Für die Praxis könnte die Intelligenzschätzung einen hilfreichen Zusatzbefund zur Detektion negativer Antwortverzerrung

darstellen, der anzeigt, in welchen Fällen die Beschwerdenvalidität noch differenzierter untersucht werden sollte oder wo darauf verzichtet werden kann.
Seitens der Testentwicklerinnen wäre die kritische Auseinandersetzung mit und Analyse der Validierungsstudien von Beschwerdenvalidierungstests hinsichtlich der Störgröße Intelligenz wünschenswert. Es wäre hier möglich, die Testgütekriterien Sensitivität, Spezifität sowie die Falsch-positiv- und -negativrate für unterschiedliche Intelligenzniveaus unabhängig zu bestimmen. Weiterführend könnten geschätzte oder gemessene Intelligenzmaße in Korrekturformeln für die Beschwerdenvalidierungstests eingehen, gegebenenfalls sogar innerhalb der Testverfahren erhoben werden. Dadurch würde bei Bestimmung der Ergebniskategorie des Beschwerdenvalidierungstests der gemessene IQ als Faktor berücksichtigt werden, um durch höhere Testgüte validere Ergebnisse zu produzieren.

Die oben genannten Befunde weisen auf ein übergeordnetes Problem der Begutachtung von Beschwerdenvalidität hin: Es handelt sich hierbei um ein schwierig messbares Konstrukt, bei dessen Erhebung trotz wachsenden Instrumentariums zur Detektion immer Störvariablen interagieren und Fehler nicht auszuschließen sind. Solche Fehler wiegen umso mehr, da das Begutachtungssetting die dichotome Entscheidung zwischen glaubwürdigen und unehrlichen Probandinnen nahelegt. Diese Entscheidung ist jedoch aus oben genannten Gründen schwierig zu treffen. Das *SIRS-2* macht hier durch die unsicheren Ergebniszwischenkategorien einen Versuch, die Ergebnisse zu differenzieren. Zukünftig muss der Fehleranfälligkeit insofern Rechnung getragen werden, als dass neue, weiterführend differenzierte Testverfahren entwickelt werden.
Gleichzeitig darf sowohl bei der Testkonzeption als auch bei der Durchführung nie vergessen werden, dass die in der Begutachtung ermittelte Beschwerdenvalidität hohe Bedeutung für das begutachtete Individuum und für die Gesellschaft haben. Hier muss die Störanfälligkeit der verfügbaren Testverfahren gewürdigt werden und im Zweifel auf eindeutige Ergebnisse der Beschwerdenvalidierung verzichtet werden.

Schließlich zeigt sich an dieser Arbeit auch die Schwierigkeit, die die Rolle der Gutachterin mit sich bringt. Während die Aufgabe im Sozialsystem klar definiert ist (Detektion nicht-authentischer Beschwerdendarstellung), stellt sich den begutachtenden Heilberuflerinnen auch die

Frage der persönlichen Motivation der Begutachteten. Häufig lassen sich bei Übertreibenden schwierige persönliche Situationen und schwere Schicksale erkennen, aus denen z.B. eine Verrentung einen Ausweg darzustellen scheint. Hier steht die Gutachterin im Spannungsfeld zwischen ihrer gutachterlichen Aufgabe und der zwischenmenschlichen Verantwortung, was im Einzelfall in einen Interrollenkonflikt münden kann. Wünschenswert wäre die Aufdeckung der verborgenen Motivation und die Integration der Antragstellerin in die Gesellschaft ohne den gleichzeitigen Missbrauch einer Sozialleistung. Dies wird jedoch in vielen Fällen im Rahmen einer Begutachtung nicht leistbar sein.

Es muss also weiterhin auch auf die Verbesserung unseres Versorgungssystems für Menschen in schwierigen Lebenssituationen gesetzt werden. Eine bessere Verfügbarkeit seelsorgerischer und psychotherapeutischer Leistungen kann ihren Leidensdruck senken und Hilfestellungen bei persönlichen und beruflichen Schwierigkeiten geben, wodurch das Begehren eines sekundären Benefits (z.B. einer Verrentung) möglicherweise gar nicht erst entstünde.

6 Zusammenfassung

Bei einer derzeit stetig zunehmenden Bedeutung von Gutachten psychischer Erkrankungen sind die Gutachterinnen vor die Herausforderung gestellt, negative Antwortverzerrung zu erkennen und hinsichtlich des Gutachtenergebnisses zu berücksichtigen. Negative Antwortverzerrung bezeichnet eine Gruppe bewusster sowie unbewusster Täuschungsphänomene durch die Patientinnen, die häufig dadurch erklärbar sind, dass als krank begutachtete Probandinnen Zugang zu attraktiven Entschädigungen (z.B. Rentenzahlungen) erhalten. Zur Detektion negativer Antwortverzerrung existieren in Form von Beschwerdenvalidierungstest (BVT) mittlerweile Verfahren mit akzeptablen Testgütekriterien.

Intelligenz kann als theoretisches Konstrukt nur durch die Messung solcher Leistungen erfasst werden, die mithilfe der Intelligenz erbracht werden. Ergänzend hat man darüber hinaus Schätzformeln entwickelt, die alternativ zu Intelligenztests eine zeitsparende, jedoch ungenauere Schätzung des Intellekts anhand sozialer Parameter ermöglichen.

Bislang sind noch keine Erkenntnisse zur Interaktion von Intelligenz und Testgütekriterien von BVT erlangt worden. Auch ist die Frage unbeantwortet, ob der IQ Begutachteter als nützliche Zusatzvariable zu BVT verwendet werden kann.

Die dieser Arbeit zugrundeliegende Untersuchung teilt das 60 Probandinnen umfassende Studienkollektiv randomisiert zu zwei Untersuchungsbedingungen (*authentisch* / *aggraviert*) zu. Die jeweilige Verhaltensanweisung während des BVT wurde durch standardisierte Instruktionstexte kommuniziert.

Zur Beschwerdenvalidierung wurde die deutsche Version des *Structured Interview of Reported Symptoms 2* (*SIRS-2*) verwendet (Schmidt et al., 2019), während die Intelligenz mithilfe der Untertests 1-3 des *Leistungsprüfsystems* (Horn, 1983) sowie einer Sozialformel (Jahn et al., 2013) erhoben bzw. abgeschätzt wurde.

Eine systematische Verzerrung der BVT-Testgüte in Abhängigkeit von der Intelligenz der Probandinnen ließ sich nicht feststellen. Allerdings zeigte sich, dass die Intelligenz der Begutachteten, bei denen das *SIRS-2* keine eindeutige Zuordnung zu authentischer Symptomdarstellung bzw. negativer Antwortverzerrung vornehmen konnte, eine Aussagekraft hinsichtlich der Authentizität der Beschwerden hatte. Übertreibende innerhalb dieser Teilgruppe wiesen einen signifikant höher geschätzten IQ auf als die authentischen Probandinnen. Ein sinnvoller Cut-

off zur Annäherung an die verborgene Beschwerdenvalidität ließ sich bei einem *Verbal-IQ* (Sozialformel) von ≥ 104 bestimmen.

Abschließend bleibt festzustellen, dass diese Studie zwar einen wichtigen Hinweis auf, jedoch keinen schlussendlichen Beweis der vermuteten Zusammenhänge liefert. Folgestudien sollten weitere Erkenntnisse zur Interaktion von Probandinnenintelligenz und Testmechanismen sowie deren Erfolg untersuchen, um für die zukünftige Begutachtung psychischer Erkrankungen ein valides, reliables und objektives Instrumentarium mit dem Ziel der Erfassung von Beschwerdenvalidität verfügbar zu machen.

7 Literaturverzeichnis

Adachi M, Shibata A, Sato T, Kawaguchi E (2011) Smaller brain size likely in young adults (< 40 years old) with depressive symptoms compared to healthy controls: a retrospective study. JPN J RADIOL 29:19-24.

Ahdidan J, Foldager L, Rosenberg R, Rodell A, Videbech P, Mors O (2013) Hippocampal volume and serotonin transporter polymorphism in major depressive disorder. Acta Neuropsychiatr. 25:206-214.

Amelang M, Schmidt-Atzert L: Psychologische Diagnostik und Intervention. 4. Aufl. Springer-Verlag, Berlin-Heidelberg, 2006.

American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub, 2013.

Baker G, Hanley J, Jackson H, Kimmance S, Slade P (1993) Detecting the faking of amnesia: Performance differences between simulators and patients with memory impairment. J Clin Exp Neuropsychol 15:668-684.

Barona A, Chastain RL (1986) An improved estimate of premorbid IQ for blacks and whites on the WAIS-R. International Journal of Clinical Neuropsychology 8:169-173.

Beck AT: Cognitive Therapy and the Emotional Disorders. Plume, New York, NY, USA, 1979.

Birck A (2002) Echte und vorgetäuschte posttraumatische Belastungsstörungen. Psychotraumatologie 3:26.

BKK Dachverband (2017): Arbeitsunfähigkeit aufgrund von psychischen Störungen in Deutschland nach Geschlecht im Zeitraum von 1994 bis 2015. Abgerufen von „https://de.statista.com/statistik/daten/studie/251325/umfrage/arbeitsunfaehigkeit-aufgrund-von-psychischen-stoerungen-nach-geschlecht/“ am 13.03.2019.

Boring EG (1923) Intelligence as the tests test it. New Repub 36:35-37.

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

Bundespsychotherapeutenkammer (2012): Verteilung der psychisch verursachten Frühverrentungen in Deutschland nach Diagnosegruppe und Geschlecht im Jahr 2012. Abgerufen von „https://de.statista.com/statistik/daten/studie/318159/umfrage/fruehrente-psychisch-verursachten-fruehverrentungen-nach-diagnose-und-geschlecht/“ am 13.03.2019.

Bundespsychotherapeutenkammer (2013): BPtK-Studie zur Arbeits- und Erwerbsunfähigkeit: Psychische Erkrankungen und gesundheitsbedingte Frühverrentung. Abgerufen von „https://www.bptk.de/uploads/media/20140128_BPtK-Studie_zur_Arbeits-und_Erwerbsunfaehigkeit_2013_1.pdf“ am 13.03.2019.

Busch M, Maske U, Ryl L, Schlack R, Hapke U (2013) Prävalenz von depressiver Symptomatik und diagnostizierter Depression bei Erwachsenen in Deutschland. Bundesgesundheitsblatt 56:733-739.

Buser K, Schneller T, Wildgrube K, Dangl S: Medizinische Psychologie, medizinische Soziologie. 6., überarb. Aufl. Elsevier, Urban & Fischer, München-Jena, 2007.

Cattell RB (1941) Some theoretical issues in adult intelligence testing. Psychological Bulletin 38(592):10.

Chafetz MD, Prentkowski E, Rao A (2011) To work or not to work: motivation (not low IQ) determines symptom validity test findings. Arch Clin Neuropsychol 26:306-313.

Cima M, Hollnack S, Kremer K, Knauer E, Schellbach-Matties R, Klein B, Merckelbach H (2003) Strukturierter Fragebogen Simulierter Symptome. Nervenarzt 74:977-986.

Crookes T (1961) Wechsler's deterioration ratio in clinical practice. J Consult Psychol 25:234.

Dahl G: WIP: Handbuch zum Reduzierten Wechsler-Intelligenztest. Hain, Königstein, 1986.

DAK Gesundheit (2013): DAK Gesundheitsreport 2013: Schwerpunktthema psychische Erkrankungen. Abgerufen von „https://www.dak.de/dak/download/vollstaendiger-bundesweiter-gesundheitsreport-2013-1318306.pdf“ am 13.03.2019.

Demakis G, Rimland C, Reeve C, Ward J (2015) Intelligence and psychopathy do not influence malingering. Appl Neuropsychol Adult 22:262-270.

Deutsche Gesellschaft für Neurowissenschaftliche Begutachtung (2013): S2k-Leitlinie: Allgemeine Grundlagen der medizinischen Begutachtung (Stand: 07/2013). Abgerufen von „https://www.awmf.org/uploads/tx_szleitlinien/094-001l_S2k_Allgemeine_Grundlagen_der_medizinischen_Begutachtung_2013-07-abgelaufen.pdf“ am 13.03.2019.

Deutsche Rentenversicherung (2017): Rentenzugang wegen verminderter Erwerbsfähigkeit aufgrund psychischer Erkrankungen in Deutschland von 1993 bis 2015. Abgerufen von „https://de.statista.com/statistik/daten/studie/6960/umfrage/verrentung-wegen-psychischen-erkrankungen-seit-1993/“ am 13.03.2019.

DGPPN, BÄK, KBV, AWMF, AkdÄ, BPtK, BApK, DAGSHG, DEGAM, DGPM, DGPs, DGRW (Hrsg.) (2015): S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression - Langfassung, 2. Auflage, Version 1. Abgerufen von „https://www.dgppn.de/fileadmin/user_upload/_medien/download/pdf/kurzversion-leitlinien/S3-NVLdepression-lang_2015.pdf“ am 28. Februar 2017.

Dilling H, Freyberger HJ: Taschenführer zur ICD-10-Klassifikation psychischer Störungen. 5., überarb. Aufl. Huber, Bern, 2012.

Dohrenbusch R, Henningsen P, Merten T (2011) Die Beurteilung von Aggravation und Dissimulation in der Begutachtung psychischer und psychosomatischer Störungen. Versicherungsmedizin 63:81-85.

Dreßing H, Foerster K, Widder B, Schneider F, Falkai P (2011) Zur Anwendung von Beschwerdenvalidierungstests in der psychiatrischen Begutachtung: Stellungnahme der Deutschen Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN). Nervenarzt 82:388-389.

Dreßing H, Förster K, Schneider F (2018) Begutachtung psychischer Erkrankungen in der gesetzlichen Renten-und Unfallversicherung. Fortschr Neurol Psychiatr 86:422-427.

Eisenmenger W, Lippert H-D, Wandl U: Grundbegriffe der Begutachtung. In: Dörfler H, Eisenmenger W, Lippert H-D, Wandl U (Hrsg): Medizinische Gutachten. 2. Aufl. Springer, Berlin-Heidelberg, 2015, S. 41-66.

Erzberger CS, Engel RR (2010) Zur Äquivalenz der Normen des Wechsler-Intelligenztests für Erwachsene (WIE) mit denen des Hamburg-Wechsler-Intelligenztests für Erwachsene–Revision (HAWIE-R). Z Neuropsychol 21:25-37.

Fähndrich E, Stieglitz R-D: Leitfaden zur Erfassung des psychopathologischen Befundes: halbstrukturiertes Interview anhand des AMDP-Systems. 5., korrigierte Aufl. Hogrefe, Göttingen, 2018.

Fahrenberg J, Hampel R, Selg H: Das Freiburger Persönlichkeitsinventar (FPI-R). 8. Aufl. Hogrefe, Göttingen, 2010.

Faller H, Lang H: Medizinische Psychologie und Soziologie. 3. Aufl. Springer-Verlag, Berlin-Heidelberg, 2010.

Fellgiebel A: Pseudodemenz: Abgrenzung Altersdepression – Demenz. In: Fellgiebel A, Hautzinger M (Hrsg): Altersdepression. 1. Aufl. Springer, Wien-New York, 2017, S. 51-55.

Fleischhacker W, Hinterhuber H: Einführung. In: Fleischhacker W, Hinterhuber H (Hrsg): Lehrbuch Psychiatrie. 1. Aufl. Springer, Wien-New York, 2012, S. 1-20.

Flynn JR: What is Intelligence?: Beyond the Flynn Effect. 1. Aufl. Cambridge University Press, Cambridge, 2007.

Flynn JR: Are we Getting Smarter?: Rising IQ in the Twenty-First Century. 1. Aufl. Cambridge University Press, Cambridge, 2012.

Förster K, Dreßing H: Aufgaben und Stellung des psychiatrischen Sachverständigen. In: Dreßing H, Habermeyer E, Venzlaff U, Foerster K (Hrsg): Psychiatrische Begutachtung: ein praktisches Handbuch für Ärzte und Juristen. 6., neu bearb. und erw. Aufl. Elsevier, Urban&Fischer, 2015, S. 3-13.

Fritze E, Fritze J: Allgemeine Grundlagen. In: Fritze J, Mehrhoff F (Hrsg): Die ärztliche Begutachtung: Rechtsfragen, Funktionsprüfungen, Beurteilungen. 8. Aufl. Springer, Berlin-Heidelberg, 2012, S. 2-12.

Graham P, Blakely T, Davis P, Sporle A, Pearce N (2004) Compression, expansion, or dynamic equilibrium? The evolution of health expectancy in New Zealand. J Epidemiol Commun H 58:659-666.

Green P: Green's Word Memory Test for Microsoft Windows: User's manual. Green's Publishing, Edmonton, Canada, 2005.

Hall HV, Poirier J: Detecting malingering and deception: Forensic distortion analysis. 2. Aufl. CRC Press, Boca Raton, FL, USA, 2000.

Halsband U, Unterrainer J: Neuropsychologische Funktionsdiagnostik. In: Stieglitz R-D, Baumann U, Freyberger HJ (Hrsg): Psychodiagnostik in klinischer Psychologie, Psychiatrie und Psychotherapie. 2. überarb. und erw. Aufl. Thieme, Stuttgart, 2001, S. 159-182.

Härter M, Möller O, Schneider F, Niebling W: Affektive Störungen. In: Schneider F (Hrsg): Klinikmanual Psychiatrie, Psychosomatik und Psychotherapie. 2., aktualisierte Aufl. Springer, Berlin-Heidelberg, 2016, S. 365-400.

Hathaway S, McKinley J: Manual for administering and scoring the MMPI. University of Minnesota Press, Minneapolis, MN, USA, 1943.

Hathaway S, McKinley J, Engel R: Minnesota Multiphasic Personality Inventory: Manual zum Deutschen MMPI-2. Huber, Göttingen, 2000.

Heller K, Kratzmeier H, Lengfelder A: Matrizen Test Manual: Ein Handbuch mit Deutschen Normen zu den Standard Progressive Matrices (SPM; Raven, JC, 1958). Beltz Test GmbH, Göttingen, 1998.

Holling H, Preckel F, Vock M: Intelligenzdiagnostik. 1. Aufl. Hogrefe, Göttingen, 2004.

Horn R: SPM: Raven's Standard Progressive Matrices (Raven, JC). Pearson Assessments, Frankfurt/Main, San Antonio, TX, USA, 2009.

Horn W: Leistungsprüfsystem L-P-S: Handanweisung für die Durchführung, Auswertung und Interpretation. 2., erw. u. verb. Aufl. (1.Aufl. 1962). Hogrefe, Göttingen, 1983.

Jacobi F, Klose M, Wittchen H-U (2004) Psychische Störungen in der deutschen Allgemeinbevölkerung: Inanspruchnahme von Gesundheitsleistungen und Ausfalltage. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz 47:736-744.

Jahn T, Beitlich D, Hepp S, Knecht R, Köhler K, Ortner C, Sperger E, Kerkhoff G (2013) Drei Sozialformeln zur Schätzung der (prämorbiden) Intelligenzquotienten nach Wechsler. Z Neuropsychol 24:7-24.

Jones AB, Llewellyn LJ, Beaumont W: Malingering or the simulation of disease. 1. Aufl. W. Heinemann, 1917.

Kapur N (1994) The coin-in-the-hand test: a new "bed-side" test for the detection of malingering in patients with suspected memory disorder. J Neurol Neurosurg Psychiatry Res 57:385.

Kater H: Das ärztliche Gutachten im sozialgerichtlichen Verfahren: die schwierige Kommunikation zwischen Juristen und Medizinern. 2., neu bearb. Aufl. Erich Schmidt Verlag, Berlin, 2011.

Kreuzpointner L, Lukesch H, Horn W: Leistungsprüfsystem 2 (LPS-2): Manual. Hogrefe, Göttingen, 2013.

Lehrl S: Mehrfachwahl-Wortschatz-Intelligenztest: MWT-B. PERIMED-Spitta, Balingen, 1999.

Leplow B, Friege L (1998) Eine Sozialformel zur Schätzung der prämorbiden Intelligenz. Z Klin Psychol 27:1-8.

Liepmann D, Beauducel A, Brocke B, Amthauer R: I-S-T 2000 R: Intelligenz-Struktur-Test 2000 R. 2. Aufl. Hogrefe, Göttingen, 2007.

Linden M, Baron S, Muschalla B, Ostholt-Corsten M: Fähigkeitsbeeinträchtigungen bei psychischen Erkrankungen: Diagnostik, Therapie und sozialmedizinische Beurteilung in Anlehnung an das Mini-ICF-APP. 1. Aufl. Huber, Bern, 2015.

Lippert H-D, Zahrl J, Bollag Y: Rechtliche Grundlagen. In: Dörfler H, Eisenmenger W, Lippert H-D, Wandl U (Hrsg): Medizinische Gutachten. 2. Aufl. Springer, Berlin-Heidelberg, 2015, S. 3-40.

Margraf J (2019): Pschyrembel online. Abgerufen von „https://www.pschyrembel.de/Aggravation/K01TH/doc/“ am 06. März 2019.

Marschall J, Hildebrandt S, Sydow H, Nolting H-D (2016): DAK Gesundheitsreport 2016. Abgerufen von „https://www.dak.de/dak/download/Gesundheitsreport_2016_-_Warum_Frauen_und_Maenner_anders_krank_sind-1782660.pdf“ am 15.03.2019.

Martinez-Aran A, Vieta E, Colom F, Reinares M, Benabarre A, Torrent C, Goikolea JM, Corbella B, Sánchez-Moreno J, Salamero M (2002) Neuropsychological Performance in Depressed and Euthymic Bipolar Patients. Neuropsychobiology 46:16-21.

Mathers C, Fat DM, Boerma JT: The global burden of disease: 2004 update. World Health Organization, Geneve, Switzerland, 2008.

Medizinischer Dienst des Spitzenverbandes Bund der Krankenkassen (2019): Begutachtungen der Medizinischen Dienste der Krankenkassen (MDK) für die Soziale Pflegeversicherung in den Jahren 2013 bis 2015. Abgerufen von „https://de.statista.com/statistik/daten/studie/273346/umfrage/begutachtungen-der-mdk-fuer-die-soziale-pflegeversicherung/.“ am 15.03.2019.

Mehrhoff F, Meeßen A, Cibis W, Dünn S, Diedrich U, Rombach W, Raddatz G, Losch E, Fritze E, Fritze J, Nedopil N, Lümmen D, Bahemann A, Dirschedl P, Ostendorf G-M, Link M: Rechtsgrundlagen der Auftraggeber von ärztlichen Gutachten. In: Fritze J, Mehrhoff F (Hrsg): Die ärztliche Begutachtung: Rechtsfragen, Funktionsprüfungen, Beurteilungen. 8. Aufl. Springer, Berlin-Heidelberg, 2012, S. 13-98.

Merten T (2011) Beschwerdenvalidierung bei der Begutachtung kognitiver und psychischer Störungen. Fortschr Neurol Psychiatr 79:102-116.

Merten T, Dohrenbusch R: Psychologische Methoden der Beschwerdenvalidierung. In: Schneider W, Henningsen P, Dohrenbusch R, Freyberger HJ, Irle H, Köllner V, Widder B (Hrsg): Begutachtung bei psychischen und psychosomatischen Erkrankungen: autorisierte Leitlinien und Kommentare. 2. überarb. und erw. Aufl. Hogrefe, Bern, 2016, S. 152-188.

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

Merten T, Merckelbach H, Giger P, Stevens A (2016) The Self-Report Symptom Inventory (SRSI): A new instrument for the assessment of distorted symptom endorsement. Psychol Inj Law 9:102-111.

Navrady L, Ritchie S, Chan S, Kerr D, Adams M, Hawkins E, Porteous D, Deary I, Gale C, Batty G (2017) Intelligence and neuroticism in relation to depression and psychological distress: Evidence from two large population cohorts. Eur Psychiatry 43:58-65.

Nedopil N, Trott G, Lodemann E, Scherbaum N: Psychische Krankheiten und Störungen. In: Fritze J, Mehrhoff F (Hrsg): Die ärztliche Begutachtung: Rechtsfragen, Funktionsprüfungen, Beurteilungen. 8. Aufl. Springer, Berlin-Heidelberg, 2012, S. 690-738.

Olshansky SJ, Rudberg MA, Carnes BA, Cassel CK, Brody JA (1991) Trading off longer life for worsening health: the expansion of morbidity hypothesis. J Aging Health 3:194-216.

Petermann F: WAIS-IV - Wechsler Adult Intelligence Scale - Fourth Edition. Deutschsprachige Adaptation des WAIS-IV von D. Wechsler. 4. Aufl. Pearson Assessment, Frankfurt/Main, 2012.

Pietschnig J, Voracek M (2015) One century of global IQ gains: a formal meta-analysis of the flynn effect (1909-2013). Perspect Psychol Sci 10:282-306.

Preckel F, Brüll M: Intelligenztests. 1. Aufl. Reinhardt, München-Basel, 2008.

Reiche D, Bindig M, Boss N, Wangerin G: Roche-Lexikon Medizin. 5., neu bearb. und erw. Aufl. Urban&Fischer, München-Jena, 2003.

Rist F, Dirksmeier C: Leistungsdiagnostik bei psychischen Störungen. In: Stieglitz R-D, Baumann U, Freyberger HJ (Hrsg): Psychodiagnostik in klinischer Psychologie, Psychiatrie, Psychotherapie. 2. überarb. und erw. Aufl. Thieme, Stuttgart, 2001, S. 145-158.

Rogers R (1984) Towards an empirical model of malingering and deception. Behav Sci Law 2:93-111.

Rogers R: Clinical assessment of malingering and deception. 3. Aufl. Guilford Press, New York, NY, USA, 2008.

Rogers R, Sewell KW, Gillard ND: Structured Interview of Reported Symptoms, 2nd Edition: professional Manual. 2. Aufl. PAR, Inc., Lutz, FL, USA, 2010.

Rost DH: Handbuch Intelligenz. 1. Aufl. Beltz, Weinheim, 2013.

Sackeim HA, Freeman J, McElhiney M, Coleman E, Prudic J, Devanand D (1992) Effects of major depression on estimates of intelligence. J Clin Exp Neuropsychol 14:268-288.

Schiltenwolf M, Henningsen P: Muskuloskelettale Schmerzen: Diagnostizieren und Therapieren nach biopsychosozialem Konzept. 1. Aufl. Deutscher Ärzteverlag, Köln, 2006.

Schmidt MH: Psychische Störungen infolge von Intelligenzminderungen. In: Petersmann F (Hrsg): Lehrbuch der Klinischen Kinderpsychologie und -psychotherapie. 4. Aufl. Hogrefe, Göttingen, 2000, S. 359-380.

Schmidt T, Lanquillon S, Ullmann U (2011) Kontroverse zu Beschwerdenvalidierungsverfahren bei der Begutachtung psychischer Störungen. Forens Psychiatr Psychol Kriminol 5:177-183.

Schmidt T, Watzke S, Lanquillon S, Stieglitz R-D: SIRS-2. Deutschsprachige Adaptation des Structured Interview of Reported Symptoms, 2nd edition, von Richard Rogers, Kenneth W. Sewell und Nathan D. Gillard. Hogrefe, Bern, 2019.

Schneider F, Frister H, Olzen D: Begutachtung psychischer Störungen. 3., vollst. überarb. und aktualisierte Aufl. Springer, Berlin-Heidelberg, 2015.

Schneider F, Niebling W, Habel U, Nickl-Jockschat T: Diagnostik. In: Schneider F (Hrsg): Klinikmanual Psychiatrie, Psychosomatik und Psychotherapie. 2., aktualisierte Aufl. Springer, Berlin-Heidelberg, 2015, S. 19-67.

Schneider F, Weber-Papen S: Begutachtung. In: Schneider F (Hrsg): Klinikmanual Psychiatrie, Psychosomatik und Psychotherapie. 2., aktualisierte Aufl. Springer, Berlin-Heidelberg, 2015, S. 637-658.

Schneider W, Dohrenbusch R, Freyberger HJ, Gündel H, Henningsen P, Köllner V, Barth J, Becker D, Kowalewsky S, Schickel S: Manual zum Leitfaden „Begutachtung der beruflichen Leistungsfähigkeit bei psychischen und psychosomatischen Erkrankungen". In: Schneider W, Henningsen P, Dohrenbusch R, Freyberger HJ, Irle H, Köllner V, Widder B (Hrsg): Begutachtung bei psychischen und psychosomatischen Erkrankungen: autorisierte Leitlinien und Kommentare. 2. überarb. und erw. Aufl. Hogrefe, Bern, 2016, S. 475-548.

Schorr A (1995) Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. Diagnostica 41:3-20.

Shenk D (2017) What is the Flynn Effect, and how does it change our understanding of IQ? Wiley Interdiscip Rev Cogn Sci 8:e1366.

Smith GP, Burger GK (1997) Detection of malingering: validation of the Structured Inventory of Malingered Symptomatology (SIMS). J Am Acad Psychiatry Law 25:183-189.

Soyka M, Hollweg M, Naber D (1996) Alkoholabhängigkeit und Depression: Klassifikation, Komorbidität, genetische und neurobiologische Aspekte. Nervenarzt 67:896-904.

Spearman C (1904) "General Intelligence": objectively determined and measured. Am J Psychol 15:201-292.

Stadtland C, Nedopil N: Psychiatrische Begutachtung. In: Dörfler H, Eisenmenger W, Lippert H-D, Wandl U (Hrsg): Medizinische Gutachten. 2. Aufl. Springer, Berlin-Heidelberg, 2015, S. 557-594.

Statistisches Bundesamt (2017): Gesundheit Ausgaben: 1995-2015 (Fachserie 12 Reihe 7.1.2). Abgerufen von „https://www.destatis.de/DE/Publikationen/Thematisch/Gesundheit/Gesundheitsausgaben/AusgabenGesundheitLangeReihePDF_2120712.pdf?__blob=publicationFile" am 15.03.2019.

Steck P (1997) Psychologische Testverfahren in der Praxis: Ergebnisse einer Umfrage unter Testanwendern. Diagnostica 43:267-284.

Steinhauer A: Duden - Das Wörterbuch der Abkürzungen: Über 50.000 nationale und internationale Abkürzungen und Kurzwörter mit ihren Bedeutungen. 6., überarb. und erw. Aufl. Dudenverlag, Mannheim-Zürich, 2011.

Sturm W, Willmes K, Horn W: Leistungsprüfsystem für 50-90-jährige (LPS 50+). Hogrefe, Göttingen [u.a.], 1993.

Tewes U: Revision des Hamburg-Wechsler Intelligenztest für Erwachsene (HAWIE-R). Huber, Bern, 1991.

Tewes U, Rossmann P: HAWIK-III: Hamburg-Wechsler-Intelligenztest für Kinder-dritte Auflage: Manual: Übersetzung und Adaptation des WISC-III Wechsler Intelligence Scale for Children-von David Wechsler. 3. Aufl. Huber, Bern, 2002.

Thurstone LL: Primary mental abilities. 1. Aufl. University of Chicago Press, Chicago, IL, USA, 1938.

Thurstone LL (1946) Theories of intelligence. Sci Mon 62:101-112.

Tölle R, Windgassen K: Psychiatrie: Einschließlich Psychotherapie. 17., überarb. und erg. Aufl. Springer, Berlin-Heidelberg, 2014.

Walter F, Lid N, Petermann F, Kobelt A (2017) Einsatz des Beschwerdenvalidierungstests BEVA in der sozialmedizinischen Begutachtung. Die Rehabilitation 56:173-180.

Walter F, Petermann F, Kobelt A (2012) Beschwerdenvalidierung: ein aktueller Überblick. Die Rehabilitation 51:342-348.

Wechsler D: Manual for the Wechsler Adult Intelligence Scale. Psychological Corporation, New York, NY, USA, 1955.

Wechsler D: Die Messung der Intelligenz Erwachsener: Textband zum Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE). Huber, Bern, 1956.

Wechsler D: Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE): Deutsche Bearbeitung: Hardesty, A, Lauber, H. Herausgegeben von Bondy, C, Bern. 3. Aufl. Hans Huber Verlag, Stuttgart-Wien, 1964.

Wechsler D: Wechsler Adult Intelligence Scale – fourth edition: Deutsche Bearbeitung: F. Petermann. Pearson, Frankfurt/Main, 2012.

Weiß RH: Grundintelligenztest Skala 2 mit Wortschatztest (WS) und Zahlenfolgentest (ZF) (CFT 20). Hogrefe, Göttingen, 1998.

Westhoff K, Kluck M-L: Psychologische Gutachten schreiben und beurteilen. 6., vollst. überarb. und erw. Aufl. Springer, Berlin-Heidelberg, 2014.

Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, Charlson FJ, Norman RE, Flaxman AD, Johns N (2013) Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 382:1575-1586.

Wilson RS, Rosenbaum G, Brown G, Rourke D, Whitman D, Grisell J (1978) An index of premorbid intelligence. J Consult Clin Psych 46:1554.

Zammit S, Allebeck P, David AS, et al. (2004) A longitudinal study of premorbid iq score and risk of developing schizophrenia, bipolar disorder, severe depression, and other nonaffective psychoses. Arch Gen Psychiatry 61:354-360.

8 Thesen der Arbeit

1. Die zunehmend an Bedeutung gewinnende Begutachtung psychischer Erkrankungen unterscheidet sich insofern grundsätzlich von der körperlichen Begutachtung, als dass die Diagnostik weniger auf physikalischen Messverfahren, sondern zum größeren Teil auf anamnestisch erhobenem Erleben der Patientinnen beruht.
2. Für Gutachten psychischer Erkrankungen stellt negative Antwortverzerrung – also die bewusste und unbewusste Täuschung geschilderten Krankheitserlebens durch die Patientinnen – eine schwer zu enttarnende Fehlerquelle dar. Es existieren in Form von Beschwerdenvalidierungstests (BVT) zunehmend Assessmentinstrumente, die sich um eine valide Erfassung negativer Antwortverzerrung bemühen.
3. Intelligenz ist ein zeitlich relativ konstantes theoretisches Konstrukt, welches das Individuum zu bestimmten Leistungen befähigt und über die Prüfung dieser Leistungen annähernd bestimmt werden kann. Auf der Basis großer Stichproben sind darüber hinaus Schätzformeln entwickelt worden, die alternativ zu Intelligenztests eine zeitsparende jedoch ungenauere Schätzung des Intellekts anhand sozialer Parameter ermöglichen.
4. Bisher haben kaum wissenschaftliche Studien die Fragen behandelt, ob die Intelligenz von Probandinnen signifikant mit der Messbarkeit von Beschwerdenvalidität durch BVT korreliert und ob der IQ (bzw. ein Schätzwert) eine nützliche Zusatzinformation im Kontext einer Beschwerdenvalidierung darstellen könnte.
5. In dieser Arbeit wurde der Zusammenhang zwischen Intelligenz und der Messbarkeit negativer Antwortverzerrung an einer Stichprobe von n=60 Probandinnen untersucht. Dieses Studienkollektiv wurde randomisiert zwei Untersuchungsbedingungen (authentische / aggravierte Beschwerdendarlegung) zugeteilt und durchlief daraufhin ein Assessment zur Beschwerdenvalidierung.
6. Die Erhebung der Beschwerdenvalidität wurde durch die deutschen Version des *Structured Interview of Reported Symptoms 2* (*SIRS-2*) (Schmidt et al., 2019), ein strukturiertes Interview, realisiert. Die Intelligenz der Probandinnen wurde anhand der Untertests 1-3 des *Leistungsprüfsystems* (Horn, 1983) sowie mithilfe einer Sozialformel (nach Jahn et al., 2013) erfasst bzw. geschätzt.

7. Ein systematischer Gruppenunterschied der Beurteilungsqualität des *SIRS-2* zwischen *korrekter* vs. *nicht korrekter* Zuordnung ließ sich hinsichtlich keiner der Intelligenzmaße nachweisen.
8. In der Gruppe der Probandinnen, bei denen das *SIRS-2* zu keiner eindeutigen Beurteilung der Beschwerdenvalidität kam, ließ sich jedoch ein signifikanter Gruppenunterschied hinsichtlich der geschätzten Intelligenz finden. Die intelligenteren Probandinnen waren dabei signifikant häufiger unentdeckte Übertreibende, während bei den weniger intelligenten Probandinnen häufiger eine authentische Symptomdarstellung vorgelegen hatte.
9. Aus den empirischen Daten der vorliegenden Studie ist für diese Studiensubgruppe mit uneindeutigem *SIRS-2*-Ergebnis ein Cut-Off von *Verbal-IQ* ≥ 104 plausibel, um zwischen den Untersuchungsbedingungen *authentisch* vs. *aggraviert* zu trennen. Ein geschätzter Verbal-IQ von ≥ 104 wies dabei auf eine aggravierte Symptompräsentation, eine Verbal-IQ von <104 auf eine authentische hin.
10. Die hier durchgeführte Studie erweitert den bisherigen Erkenntnisstand zum Zusammenhang zwischen Intelligenz und der Messbarkeit negativer Antwortverzerrung substantiell und kann wichtige Hinweise auf die Rolle von Intelligenz für die Entwicklung und Anwendung von BVT bieten. Die Ergebnisse sollten jedoch in Zukunft mit anderen, größeren Stichproben und auch für andere Testverfahren repliziert werden, bevor man davon ausgehen kann, dass es sich um gesicherte Erkenntnisse handelt.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich diese Dissertation selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Halle (Saale), den 10.04.2019

_____

Benjamin Reufsteck

Erklärung über frühere Promotionsversuche

Hiermit erkläre ich, dass ich an keiner anderen Fakultät oder Universität jemals in ein anderes als das im Kontext dieser Arbeit an der Medizinischen Fakultät der *Martin-Luther-Universität Halle-Wittenberg* eröffnete Dissertationsverfahren involviert war.
Darüber hinaus habe ich nie zuvor einen anderen Promotionsversuch an einer anderen wissenschaftlichen Einrichtung unternommen.

Halle (Saale), den 10.04.2019

_____

Benjamin Reufsteck

Lebenslauf

Persönliche Angaben

Name:	Benjamin Reufsteck
Geburtstag und –ort:	26. Juni 1987 in Oberhausen
Staatsangehörigkeit:	deutsch

Ausbildung / Schulbildung

10/2012 – 03/2019	Studium der Humanmedizin an der *Martin-Luther-Universität Halle-Wittenberg* (MLU) (abgeschlossen mit Erhalt der Approbation 01/2019)
10/2007 - 09/2010	Ausbildung zum Gesundheits- und Krankenpfleger am *Luisenhospital Aachen* (abgeschlossen mit Staatsexamen)
08/1997 - 06/2006	*Städt. Gymnasium Straelen* mit Erwerb der Allgemeinen Hochschulreife
08/1993 - 06/1997	*Katharinen-Grundschule Straelen*

Berufliche Erfahrungen

11/2017 - 10/2018	Praktisches Jahr
07/2018 - 10/2018	Wahltertial Pädiatrie am *Kinderspital Zürich*, Schweiz
03/2018 - 07/2018	Tertial Chirurgie am *BG Klinikum Bergmannstrost Halle*
11/2017 - 03/2018	Tertial Innere Medizin am *Universitätsklinikum Halle* (UKH)
10/2010 - 08/2011	Vollzeitbeschäftigung als Gesundheits- und Krankenpfleger auf der medizinischen Intensivstation der *Universitätsklinik Bonn*

Sonstige praktische Tätigkeiten

08/2006 - 04/2007	Zivildienst in der ambulanten Pflege bei der *Caritas Sozialstation Straelen*

_____

Benjamin Reufsteck

Danksagung

Zunächst danke ich allen, die zur Erstellung dieser Doktorarbeit beigetragen haben, insbesondere den 60 Probandinnen und Probanden, die sich die Zeit genommen und die Mühe gemacht haben, unentgeltlich an dieser Studie teilzunehmen.

Mein größter Dank gilt apl. Prof. Dr. Stefan Watzke, der mir als Betreuer mit seiner wissenschaftlichen Erfahrung, seinem statistischen Können, seinen konstruktiven Anregungen und seinem unerschütterlichen Optimismus während der gesamten Zeit meiner Arbeit zur Seite stand. Ich danke außerdem Thomas Schmidt für die vielen fachlichen und methodischen Hilfestellungen. Beiden danke ich für ihre ständige Erreichbarkeit, die investierte Zeit und ihre beharrliche Hilfsbereitschaft.
Meinen beiden Kommilitoninnen Ruth-Sophia Ebert und Wiebke Heinemann danke ich für die angenehme und unkomplizierte Zusammenarbeit bei der Probandinnenrekrutierung und der Durchführung der Studie.
Ich danke darüber hinaus Christian Kaden für die sprachlichen Anregungen und Verbesserungsvorschläge zu meiner Dissertation.

Ein besonderer Dank gilt meinen Eltern Lydia und Günther, meiner Schwester Esther, meiner Freundin Mirjam sowie meinem besten Freund Daniel, ohne deren verlässliche Unterstützung während der letzten 13 Jahre mein Ausbildungsweg so nicht möglich gewesen wäre.