Medizinische Fakultät der

Martin-Luther-Universität Halle-Wittenberg

PSYCHOMETRISCHE ANALYSE DES BECK DEPRESSIONS-INVENTARS-II MITTELS ITEM RESPONSE THEORIE

Dissertation zur Erlangung des akademischen Grades

Doktor der Medizin (Dr. med.)

vorgelegt

der Medizinischen Fakultät

der Martin-Luther-Universität Halle-Wittenberg

von Johanna Wittmann

geboren am 27.01.1993 in München

Betreuer: apl. Prof. Dr. Stefan Watzke

Gutachter:

- apl. Prof. U. Preuss, Herborn
- Prof. D. Ostwald, Magdeburg

12.07.2021

09.02.2022

REFERAT:

Zielsetzung: Das Beck Depressions-Inventar-II (BDI-II) [1, 2], das in Deutschland zu den am häufigsten zur Einschätzung des Schweregrads depressiver Symptomatik eingesetzten Selbstbeurteilungsfragebögen zählt [3], wurde auf Basis der klassischen Testtheorie (KTT) entwickelt und validiert, während Erkenntnisse auf Basis der Item Response Theorie (IRT) bisher weitestgehend auf fremdsprachigen Studien beruhen. Ziel der vorliegenden Arbeit war es daher, die psychometrischen Charakteristika der deutschen Version des BDI-II mittels IRT näher zu analysieren.

Material und Methodik: Das BDI-II wurde anhand einer Stichprobe junger Menschen in der Ausbildungsphase ($n=2419$) auf seine Eignung für IRT-Analysen hin überprüft. Die anschließende Auswertung erfolgte mit dem IRT-Modell, das in der Überprüfung die beste Anpassungsgüte für den Datensatz aufwies und umfasste neben einer Auswertung der allgemeinen psychometrischen Charakteristika des BDI-II auch die Analyse auf Differential Item Functioning (DIF) bezüglich Alter und Geschlecht.

Ergebnisse: Die Analysen, die bei guter Modellanpassung mit dem unidimensionalen Graded Response Modell durchgeführt wurden, zeigten für alle Items eine mind. moderate, für neun Items sogar eine sehr hohe Diskriminationsfähigkeit. Hierbei zeigte sich das Item 'Wertlosigkeit' ($\alpha_{14} = 2{,}8$) als das am besten, und das Item 'Appetitveränderungen' ($\alpha_{18} = 1{,}0$) als das am schlechtesten diskriminierende Item. Die Items deckten mit ihren Schwierigkeitsparametern ($\text{LP von } 1{,}4 - 3{,}9$) ein sehr breites Spektrum der Depressivität ab, wobei dem Item 'Ermüdbarkeit' von den Probanden am leichtesten und dem Item 'Appetitveränderungen' am schwierigsten zugestimmt werden konnte. Das BDI-II zeigte im Bereich durchschnittlicher bis ausgeprägter Depressivität eine sehr hohe Reliabilität, während sich im Bereich niedriger Depressivität Schwächen zeigten. Während sich für das Item 'Weinen' geschlechtsabhängig signifikant unterschiedliches Antwortverhalten abzeichnete, das jedoch keinen signifikanten Einfluss auf den Testscore ausübte, zeigten sich bezüglich des Alters keine derartigen Hinweise.

Schlussfolgerung: Anhand der Ergebnisse konnten die soliden psychometrischen Charakteristika des BDI-II bestätigt werden. Es kann gefolgert werden, dass das BDI-II gleichermaßen gut zur Einschätzung des Schweregrads depressiver Symptomatik im klinischen Setting eingesetzt werden kann, wie als Screening-Instrument in der Allgemeinbevölkerung. Hierbei trägt das Item 'Wertlosigkeit' im Bereich durchschnittlicher bis hoher Depressivität am stärksten zur Diskrimination der Depressionsschwere bei. Anhand der DIF-Analyse kann geschlossen werden, dass es zulässig ist, die Depressivität zwischen Probanden unterschiedlichen Geschlechts, sowie unterschiedlichen Alters anhand ihres BDI-II-Testscores miteinander zu vergleichen.

INHALTSVERZEICHNIS

ABKÜRZUNGSVERZEICHNIS

2-PL-Modell	2-Parameter-logistisches-Modell
AIC	Akaike's Information Criterion
BDI-II	Beck Depressions-Inventar – Version II
BDI-FS	Beck Depressions-Inventar – Fast Screen for medical patients
BIC	Schwarz Bayesian Information Criterion
CAT	Computerbasiertes adaptives Testen
CES-D	Center of Epidemiologic Studies Depression Scale
CFI	Comparative Fit Index
CIDI	Composite International Diagnostic Interview
DIF	Differential Item Functioning
DSM	Diagnostic and Statistical Manual of Mental Disorders
ECV	Explained Common Variance of the General Factor
GOF	Goodness-of-Fit = Anpassungsgüte
GPCM	Generalized Partial Credit Modell
GRM	Graded Response Model
ICC	Item Characteristic Curve
ICD-10	International Statistical Classification of Diseases and Related Health Problems
IRF	Item Response Funktion
IRT	Item Response Theorie
KTT	Klassische Testtheorie
MML	Marginal-Maximum-Likelihood
OCC	Option Characteristic Curve
PCM	Partial Credit Modell
PHQ-2 bzw. PHQ-9	Patient Health Questionnaire-2 bzw. -9
PROMIS	Patient Reported Outcome Measure Information System
RMSEA	Root mean square error of approximation
SRMR	Standardized root mean residuals
TLI	Tucker Lewis Index
WHO	World Health Organisation = Weltgesundheitsorganisation

VERZEICHNIS DER IN DEN FORMELN VERWENDETEN SYMBOLE

IN DER KLASSISCHEN TESTTHEORIE VERWENDETE SYMBOLE:

X	Beobachteter Testscore (=klassischer Summenscore)
T	True Score bzw. wahre Merkmalsausprägung
E	Messfehler
v	Proband v
m	Anzahl paralleler Testformen (=Anzahl der Items eines Tests)
SD	Standardmessfehler
Rel	Reliabilität

IN DER ITEM RESPONSE THEORIE VERWENDETE SYMBOLE:

N	Stichprobengröße
n	Anzahl der Items eines Tests
v	Proband v
i	Item i
k_i	Antwortkategorie k bei Item i
m_i	Anzahl Antwortkategorien von Item i
x_{vi}	Antwort x von Proband v auf Item i
$p(x_{vi})$	Wahrscheinlichkeit von Antwort x von Proband v auf Item i
Ψ	Logistische Verteilungsfunktion
Ξ	Item- und Personenparameter
θ_v	Personenparameter von Proband v
α_i	Diskriminationsparameter (=Steigungsparameter) von Item i
δ_{ik}	Schwellenparameter von Kategorie k bei Item i
γ_i	Schnittpunkt der logistischen Regressionskurve von Item i
τ_i	Schnittpunkt benachbarter OCCs von Item i
w	Anzahl univariater Residuen
λ^2_{gen}	Ladungsmatrix des Generalfaktors
λ^2_{grp}	Standardisierte Ladungen des Gruppenfaktors
h^2	Kommunalität
$T(\theta_v)\ bzw.\ I(\theta_v)$	Test- bzw. Iteminformationsfunktion
$SEE(\theta)$	Standard Error of Estimation (=Standardmessfehler)

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

1. Einleitung

Depressionen gehören zu den häufigsten und dennoch meistunterschätzten Erkrankungen, obwohl sie massive Auswirkungen auf die Betroffenen und ihr direktes Umfeld, sowie immense Bedeutung für die Volkswirtschaft haben [4]. Laut Weltgesundheitsorganisation (WHO) führen depressive Erkrankungen zu den gravierendsten gesundheitsbedingten Einschränkungen für die Betroffenen. Außerdem stellen sie einen maßgeblichen Faktor der globalen Krankheitslast dar. Aufgrund der weltweiten demographischen Entwicklung muss in den nächsten Jahren zudem von einer weiteren Verschärfung der Situation ausgegangen werden [5].

Für die Diagnosestellung einer depressiven Störung wird das ärztliche bzw. psychotherapeutische 1:1-Gespräch, das sich an den Kriterien des ICD-10 (International Statistical Classification of Diseases and Related Health Problems) und des DSM-V (Diagnostic and Statistical Manual of Mental Disorders) orientiert, als Goldstandard angesehen [3]. Wenn Punktprävalenzen in großen Stichproben bestimmt werden sollen, stellt neben ärztlicher oder psychotherapeutischer Expertise insbesondere der Zeitfaktor eine limitierende Rolle dar, sodass alternative Verfahren benötigt werden. Ein häufig eingesetztes Instrument hierfür sind Selbstbeurteilungsfragebögen, die insgesamt ein günstiges Kosten-Nutzen-Profil aufweisen [5]. Die im deutschsprachigen Raum am häufigsten eingesetzten Selbstbeurteilungsfragebögen zur Erfassung des Schweregrads depressiver Symptome sind der Center of Epidemiologic Studies Depression Scale (CES-D) und das Beck Depressions-Inventar-II (BDI-II) [3, 5].

Derartige klinische Fragebögen wurden in der Regel nach den Prinzipien der klassischen Testtheorie (KTT) entwickelt, die im Rahmen der Fragebogenkonstruktion jedoch einige Probleme aufweist: So geht die klassische Testtheorie von einem linearen Zusammenhang zwischen dem beobachteten Testscore und der Ausprägung der latenten Variablen aus, der nur in seltenen Fällen den reell vorliegenden Zusammenhang beschreibt. Zudem kann die wahre Ausprägung der latenten Variablen nur indirekt abgeschätzt werden und die bestimmten Parameter, wie bspw. die Reliabilität oder Item-Schwierigkeit, sind von der zur Validierung genutzten Stichprobe abhängig [6].

Durch Verwendung der Item Response Theorie (IRT), einem probabilistischen Testmodell, kann den zuvor beschriebenen Limitationen der klassischen Testtheorie begegnet werden. Gleichwohl ermöglicht die IRT zusätzlich die differenzierte Analyse eines Instrumentes auf Item-Ebene, um so spezifischere Informationen, wie bspw. Differential Item Functioning (DIF), ableiten zu können [6–8].

Die vorliegende Studie nutzt daher eine umfangreiche Stichprobe junger Menschen in der Ausbildungsphase, um an dieser den Goldstandard der Messung depressiver Symptomschwere, den Beck Depression Inventar-II, mittels IRT näher zu analysieren. Bevor genauer auf die Durchführung der vorliegenden Studie eingegangen wird, sollen zunächst einige theoretische Hintergründe näher erläutert werden.

1.1 Depressionen

1.1.1 Definition

Der Begriff 'Depression' stammt von dem lateinischen Wort 'deprimere' (niederdrücken) ab und wird im klinischen Zusammenhang für ein breites Spektrum von Krankheitsbildern verwendet [4, 5]. In der aktuellen deutschen S3-Leitlinie zur unipolaren Depression [4, Seite 17] findet sich folgende Definition:

'Depressionen sind psychische Störungen, die durch einen Zustand deutlich gedrückter Stimmung, Interesselosigkeit und Antriebsminderung über einen längeren Zeitraum gekennzeichnet sind. Damit verbunden treten häufig verschiedenste körperliche Beschwerden auf [...]. Depressive Menschen sind durch ihre Erkrankung meist in ihrer gesamten Lebensführung beeinträchtigt. Es gelingt ihnen nicht oder nur schwer, alltägliche Aufgaben zu bewältigen, sie leiden unter starken Selbstzweifeln, Konzentrationsstörungen und Grübelneigung. Depressionen gehen wie kaum eine andere Erkrankung mit hohem Leidensdruck einher, da diese Erkrankung in zentraler Weise das Wohlbefinden und das Selbstwertgefühl von Patienten beeinträchtigt'.

Depressionen werden im internationalen Diagnosemanual ICD-10 daher zu den affektiven Störungen (F30–F39) gezählt, deren Gemeinsamkeit in der pathologischen Veränderung der Stimmungslage der Betroffenen liegt [4, 5].

1.1.2 Prävalenz

Aktuellen Schätzungen zufolge leiden 322 Millionen Menschen weltweit unter depressiven Störungen, dies entspricht einer Punktprävalenz von circa 4,4% der Weltbevölkerung. Die WHO bezeichnet depressive Störungen daher als eine der wichtigsten Volkskrankheiten weltweit und geht davon aus, dass die Erkrankung aufgrund der demographischen Entwicklung in den nächsten Jahren zunehmend an Bedeutung gewinnen wird [5]. Aktuell wird die Lebenszeitprävalenz international bereits auf 16-20%, die 12-Monats-Prävalenz auf 7,7% geschätzt. Alleine in Deutschland leiden somit pro Jahr circa 6,2 Millionen Menschen an einer depressiven Störung [4].

Die depressiven Störungen gehen dabei mit schwerwiegenden Einschnitten für die Betroffenen einher. Eine mögliche Maßeinheit zur Einschätzung der Lebensjahre, die durch Behinderung oder vorzeitigem Tod aufgrund einer Erkrankung verloren gehen,

sind die ‘Disability adjusted life years’ (DALYs). 2004 lag die unipolare depressive Störung im Ranking noch auf dem dritten Rang, bis zum Jahr 2030 geht die WHO davon aus, dass die depressiven Störungen den ersten Rang, noch vor ischämischen Herzerkrankungen und Demenz einnehmen werden [4].

Damit einhergehend ist auch eine deutliche Zunahme der Krankheitskosten zu erwarten. Noch stärker als die direkten Therapiekosten, die sich aus medizinischen und nicht-medizinischen Behandlungen zusammensetzen, fallen dabei die sekundären Folgekosten durch Arbeitsunfähigkeit, Frühpensionierungen und Krankheitstage ins Gewicht, sodass die Erkrankung neben den Betroffenen und deren sozialem Umfeld auch für das Gesundheitssystem eine große Belastung darstellt [4, 5]. Aus diesen Gründen kommt einer validen und ökonomischen Diagnostik gesamtgesellschaftlich eine hohe Bedeutung zu, weshalb sich im folgenden Abschnitt den Messmöglichkeiten von Depressivität zugewandt wird.

1.1.3 Messmöglichkeiten von Depressivität

Zur klinischen Diagnostik einer depressiven Störung gibt es anhand der beiden internationalen Diagnosemanuale – ICD-10 und DSM-V – festgelegte Kriterien, anhand derer eine Einteilung in Schweregrade und entsprechend auch die Entscheidung einer Therapiebedürftigkeit festgelegt werden. Als Hauptsymptome einer Depression gelten hierbei eine depressive, gedrückte Stimmung, ein Verlust von Freude und Interesse (Anhedonie), sowie eine Antriebsminderung bzw. erhöhte Ermüdbarkeit. Für eine Diagnosestellung müssen neben weiteren Zusatzkriterien mind. zwei dieser Hauptsymptome für mehr als zwei Wochen anhaltend vorliegen [4].
Als Goldstandard der Diagnostik werden im klinischen Alltag in der Regel standardisierte diagnostische Interviews, wie das Composite International Diagnostic Interview (CIDI) genutzt, die all diese Kriterien in strukturierter Weise abfragen [3, 4].

Liegt der Fokus allerdings nicht auf der möglichst validen Depressions-Diagnose des Einzelnen, sondern auf der Einschätzung der Prävalenz depressiver Symptome in einer großen Stichprobe, ist die Durchführung strukturierter klinischer Interviews durch medizinisches Fachpersonal häufig zu zeit- und kostenaufwendig, weswegen nach möglichen Alternativen gesucht wurde [5].

Eine häufig für Forschungszwecke eingesetzte Möglichkeit stellt hierbei die Nutzung von Fragebögen als Screening-Instrument dar [5]. Diese dienen primär dazu, Personen mit erhöhtem Risiko für das Vorliegen einer psychischen Erkrankung zu identifizieren und nicht dazu, valide Diagnosen zu stellen [4]. Das Robert-Koch-Institut empfiehlt daher ‘*auf der Grundlage von Fragebogendiagnostik besser von depressiven Syndromen […] als von Diagnosen im engeren Sinne’* [3, Seite 13] zu sprechen.

Neben dem Center of Epidemiologic Studies Depression Scale (CES-D) zählt das Beck Depressions-Inventar-II (BDI-II) zu den weltweit am häufigsten zur Einschätzung des Schweregrads depressiver Symptomatik eingesetzten Selbstbeurteilungsfragebögen [3]. Auf das BDI-II soll im folgenden Abschnitt näher eingegangen werden, da es das zentrale Instrument der vorliegenden Arbeit repräsentiert.

1.2 Beck Depressions-Inventar

Um eine reliable und valible Messung von depressiven Symptomen zu ermöglichen, führte Beck systematische Beobachtungen im Rahmen von psychoanalytischen Therapiesitzungen an depressiven Patienten durch und entwickelte hieraus das aus 21 Items bestehende Beck Depressions-Inventar, das folgende Symptome abfragte [9]:

> Traurigkeit, Pessimismus, Versagensgefühle, Verlust von Freude, Schuldgefühle, Bestrafungsgefühle, Selbstablehnung, Selbstvorwürfe, Suizidalität, Weinen, Reizbarkeit, sozialer Rückzug, Entschlussunfähigkeit, Negatives Selbstbild, Arbeitsunfähigkeit, Schlafstörungen, Ermüdbarkeit, Appetitveränderungen, Gewichtsveränderungen, Hypochondrie, Libidoverlust

Bereits das ursprüngliche Beck Depressions-Inventar (BDI-I) [9] entwickelte sich nach seiner Veröffentlichung zu einem der weltweit am häufigsten eingesetzten Selbstbeurteilungsfragebögen zur Evaluation der Depressivität. Um den Kriterien des DSM-IV besser zu entsprechen, wurde das BDI-I 1996 nochmals überarbeitet [2]. Hierfür wurde die Zeitspanne, auf die sich die Beantwortung der Fragen stützen soll, auf zwei Wochen erhöht. Die Items wurden zudem zum Teil umformuliert und die vier Items ‘Negatives Selbstbild’, ‘Arbeitsunfähigkeit’, ‘Gewichtsveränderung’ und ‘Hypochondrie’ durch ‘Unruhe’, ‘Wertlosigkeit’, ‘Verlust an Energie’ und ‘Konzentrationsschwierigkeiten’ ersetzt, die den Diagnosekriterien des DSM-IV besser entsprechen. Durch diese Veränderungen entstand das Beck Depressions-Inventar–II (BDI-II) [2], das im Laufe der Jahre den ursprünglichen Fragebogen weitestgehend abgelöst hat, in sämtliche Sprachen übersetzt wurde und heute weltweit zu den am häufigsten eingesetzten Selbstbeurteilungsfragebögen zur Einschätzung der Depressionsschwere zählt [1, 2].

Im Gegensatz zu anderen gängigen Depressions-Fragebögen erfolgt die Beantwortung der Items beim BDI-II nicht durch Item-Stämme, sondern durch die Wahl hochdeskriptiver Antwortoptionen in Form von graduellen Antwortskalen – auch bezeichnet als Likert-Skalen –, wobei höhere Kategorien mit Aussagen verknüpft wurden, die schwerere depressive Symptome repräsentieren [10]. Die Auswertung und damit die Bewertung des individuellen Schweregrades der depressiven Symptomatik erfolgt klassischerweise anhand des einfachen Summenscores aller Einzelitems [9].

1.3 Möglichkeiten der Testkonstruktion

Der folgende Abschnitt wendet sich nun den Grundlagen der Fragebogenkonstruktion zu. Hierzu wird mit einer kurzen Einführung in die Grundlagen der klassischen Testtheorie, auf deren Basis das BDI-II konstruiert wurde, begonnen. Ein spezieller Fokus soll dabei auf den mit der klassischen Testtheorie verbundenen Limitationen, insbesondere in Bezug auf das BDI-II, liegen. Im Anschluss werden dann die Grundzüge der IRT eingeführt, die auf genau diesen Limitationen beruht und hierfür Lösungsstrategien verspricht [6, 11].

1.3.1 Definition eines Tests

Lienert und Raatz [12, Seite 1] definieren einen 'Test' als '*ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung*'. Die Definition verdeutlicht, dass für einen wissenschaftlich fundierten Test die zu messende Variable sehr genau definiert sein muss und entsprechende Qualitätsansprüche an den Fragebogen gestellt werden. Diese werden in der Regel in Form von sogenannten Testgütekriterien empirisch überprüft und bei wissenschaftlich fundierten Fragebögen die Ergebnisse im beiliegenden Testmanual berichtet. Die drei wichtigsten Testgütekriterien stellen hierbei die Objektivität, die Reliabilität und die Validität dar [11].

Psychologische Tests basieren zudem meist auf einer Testtheorie, die den mathematischen Zusammenhang zwischen dem Testverhalten und der zu messenden individuellen Merkmalsausprägung beschreibt. McDonald [13] beschrieb eine Testtheorie als Sammlung von mathematischen Konzepten, die bestimmte Fragen über die Konstruktion und Nutzung von Tests operationalisieren und entsprechende Lösungsstrategien zur Verfügung stellen. Das hierin dominierende Verfahren war im letzten Jahrhundert klar die klassische Testtheorie, die sich zum Goldstandard der Testkonstruktion entwickelte. In den letzten Jahrzehnten erhielt die klassische Testtheorie jedoch Konkurrenz durch ein neu entwickeltes Verfahren, das als Item Response Theorie bezeichnet wird [6]. Bevor näher auf die Entwicklungen durch die Item Response Theorie eingegangen wird, sollen zunächst die Grundzüge der klassischen Testtheorie, sowie deren Limitationen beleuchtet werden, die zur Entwicklung der Item Respone Theorie beigetragen haben [14, 15].

1.3.2 Klassische Testtheorie (KTT)

Das Problem bei der Evaluation von psychologischen Eigenschaften ist, dass sie nicht direkt messbar sind. Man benötigt also ein Messinstrument, wie zum Beispiel einen Fragebogen, um die Probanden anhand ihrer Merkmalsausprägung einordnen zu

können. Da kein Test die wahre Ausprägung perfekt abbilden kann, muss ein zufälliger Messfehler mit in die Betrachtung einbezogen werden [11]. Die Basis der klassischen Testtheorie, deren Prinzipien von Gulliksen [16], Lord und Novick [17] entwickelt wurde, bildet daher die Formel:

$$X = T + E, \tag{1}$$

wobei X den beobachteten Testscore, T die wahre Merkmalsausprägung, die nicht direkt beobachtet werden kann, und E den Messfehler (ebenfalls unbekannt) darstellt [6, 11]. Da diese Gleichung für jeden Probanden primär zwei Unbekannte enthält, ist sie nur dann lösbar, wenn zusätzlich vereinfachende Annahmen gemacht werden. Diese Annahmen umfassen, dass (a) die wahre Merkmalsausprägung und der Messfehler unabhängig voneinander sind, (b) der mittlere Messfehler in der Probandenpopulation Null ist, und (c) die Messfehler in parallelen Tests unkorreliert sind [18].

Sind diese Annahmen erfüllt, kann davon ausgegangen werden, dass durch eine unendliche Anzahl von Messungen die wahre Merkmalsausprägung eines Probanden abgeschätzt werden kann, sofern die gemessene Merkmalsausprägung über alle Messungen hinweg konstant ist [18]. Im Rahmen psychologischer Tests ist die Applikation desselben Tests mehrmals nacheinander jedoch problematisch. Um dies zu umgehen, werden mehrere parallele Testformen nacheinander benutzt, um so den Standardmessfehler zu neutralisieren und die wahre Merkmalsausprägung abzuschätzen. Bei parallelen Testformen ‚*handelt es sich [dabei] um Testformen, bei denen man nach empirischer Prüfung davon ausgehen kann, dass sie (trotz unterschiedlicher Items) zu gleichen wahren Werten und gleichen Varianzen der Testwerte führen*‘ [11, Seite 29].

Der beobachtete Testscore X eines Probanden v setzt sich dann aus der Summe der m Testscores zusammen und wird daher auch als Summenscore bezeichnet [11]:

$$X_v = \sum_{i=1}^{m} x_{vi} \tag{2}$$

Nach mathematischer Umformung entspricht der beobachtete Testscore X_v einer Punktschätzung $\widehat{t_v}$ der wahren Merkmalsausprägung von Proband v. Da es sich jedoch nur um eine empirische Schätzung handelt, ist es erstrebenswert, ein Konfidenzintervall anzugeben, das die Unsicherheit der Schätzung berücksichtigt [11]. Dies lässt sich mit dem Standardmessfehler realisieren, der sich mit der Formel

$$SD(E) = SD(x)\sqrt{1 - Rel} \tag{3}$$

berechnen lässt, wobei Rel die Reliabilität des Fragebogens darstellt. Anhand dieser Formel ist der Zusammenhang nachvollziehbar, dass der Standardmessfehler umso kleiner wird, je höher die Reliabilität ausfällt [6, 11].

Das Konfidenzintervall um die Ausprägung t_v lässt sich dann entsprechend der Formel

$$\widehat{t_v} - z\alpha_{/2} * SD(E) \leq t_v \leq \widehat{t_v} + z\alpha_{/2} * SD(E) \quad (4)$$

darstellen, wobei der wahre Wert der Merkmalsausprägung t_v mit einer Wahrscheinlichkeit von *(1-α)* in diesem Bereich liegt [11].

Die Reliabilität zur Berechnung lässt sich entweder dem Testmanual eines bestehenden Fragebogens entnehmen, oder mittels Parallel-Test-, Re-Test-, Split-Half-Reliabilität oder Interner Konsistenz abschätzen [11]. Die Reliabilität kann dabei gesteigert werden, wenn zusätzliche parallele Testteile – auf einen Fragebogen bezogen entsprechend zusätzliche Testitems – hinzugenommen werden [18]. Hierdurch entstanden in der Praxis Fragebögen, die zwar sehr reliable Messungen ermöglichten, aber dafür aus sehr vielen Items bestanden. Dies geht mit einer erheblichen Belastung der Testteilnehmer einher, insbesondere bei Beantwortung von mehr als nur einem Fragebogen.
Neben den so resultierenden langen Fragebögen, weist die KTT weitere Limitationen auf. So wird bspw. von einem linearen Zusammenhang zwischen dem beobachteten Testscore und der Ausprägung der latenten Variablen ausgegangen, der jedoch nur in seltenen Fällen den reell vorliegenden Zusammenhang beschreibt. Zudem lässt sich die Ausprägung der latenten Variable nur durch vereinfachende Annahmen, die jedoch häufig in der Realität schwer zu erfüllen sind, bestimmen. Durch die Annahme eines über alle Facetten der latenten Variable konstanten Messfehlers sind zudem keine individuellen Bewertungen eines Instrumentes in Bezug auf die Merkmalsausprägung möglich. Auch die Stichprobenabhängigkeit der Testcharakteristika, bspw. der Reliabilität oder Itemschwierigkeit, sowie die Itemabhängigkeit der Summenscores schränkt die Anwendungsgebiete der klassischen Testtheorie ein [6, 11, 15].

Überträgt man diese Limitationen der klassischen Testtheorie nun konkret auf das BDI-II, das mithilfe dieser Testtheorie entwickelt wurde, ergeben sich in der praktischen Anwendung folgende Einschränkungen:
Durch die Annahme eines linearen Zusammenhangs zwischen dem beobachteten Testscore und der Ausprägung der latenten Variablen ‚Depressivität' erfolgt die Auswertung des BDI-II klassischerweise mithilfe des einfachen Summenscores. Hierbei wird die erreichte Punktzahl pro Item zu einem Testscore aufsummiert, der dann zur Einordnung der Depressivität genutzt wird [2]. Mithilfe einer Normierungsstichprobe werden Cut-off-Werte für den Testscore definiert, ab welchen Hinweise auf eine Depression vorliegen bzw. die eine Einteilung der Depressivität in Schweregrade ermöglichen [19]. Die Annahme paralleler Testformen gibt hierbei vor, dass alle Items gleichgewichtet in den Testscore eingehen und somit bspw. die Zustimmung zu 'Suizidalität' in gleichem Verhältnis in die Bewertung der Depressionsschwere mit eingeht,

wie die Zustimmung zu ‘Ermüdbarkeit’. Die im Antwortmuster enthaltenen Informationen gehen somit für die Einschätzung der Depressivität in der KTT verloren [6, 18].
Die Itemabhängigkeit des Summenscores bedeutet auf das BDI-II bezogen, dass der Testscore nur dann interpretierbar ist, wenn der Fragebogen in unveränderter Form angewendet wurde. Eine Verkürzung bzw. Veränderung des Fragebogens ist nur nach erneuter Normierung zulässig und erlaubt anhand des Testscores keinen direkten Vergleich mit Probanden, die den ursprünglichen Fragebogen bearbeitet haben. Dies erschwert insbesondere die Entwicklung von adaptiven Testverfahren, bei denen Probanden – abhängig von ihrer individuellen Merkmalsausprägung – eine ausprägungsangepasste Auswahl von Items eines Itempools vorgelegt werden [18].
Mithilfe des Standardmessfehlers, der in der klassischen Testtheorie über alle Bereiche der latenten Variable, auf den BDI-II bezogen über alle Facetten der Depressiviät hinweg, als konstant angenommen wird, lässt sich ein Konfidenzintervall um den Testscore errechnen, das jedoch für jeden Probanden gleich weit gefasst ist. Dementsprechend kann für einen einzelnen Probanden nicht abgeleitet werden, wie präzise der BDI-II in dem individuellen Ausprägungsbereich der Depressivität des Probanden diskriminieren kann und erschwert insbesondere die Entscheidung, ob Veränderungen im BDI-II Testscore nach einer Therapiemaßnahme eine signifikante Veränderung anzeigen, oder im Bereich der Messungenauigkeit des Fragebogens liegen [6, 11, 15].

Die Item Response Theorie verspricht, die genannten Probleme der KTT zu lösen, sodass in den nachfolgenden Abschnitten die Grundzüge der IRT und die damit verbundenen Erwartungen für das Beck Depressions-Inventar-II beleuchtet werden sollen.

1.3.3 Item Response Theorie (IRT)

Die ersten Grundlagen der Item Response Theorie gehen auf Rasch [20] und Birnbaum [21] zurück. Das Buch ‘Statistical theories of mental test scores’ [17] von Lord und Novick, durch das die IRT deutlich an Bekanntheit zunahm, bildete den Auftakt für eine Reihe von Veröffentlichungen, die sich sowohl mit Anwendungshinweisen, als auch mit technischen Weiterentwicklungen der IRT befassten [18].

Unter der IRT wird eine Gruppe von Modellen verstanden, die in probabilistischer Art und Weise die Beziehung zwischen dem Verhalten einer Person und dem Level der latenten Variable beschreiben [22]. Es wird hierbei die Existenz einer oder mehrerer nicht direkt beobachtbarer Merkmale angenommen, die als ‘latente Variablen’ bezeichnet werden und sich durch beobachtbares Verhalten, zum Beispiel das Antwortverhalten in einem Fragebogen, als ‘manifeste Variablen’ abbilden lassen [11, 15]. Die Basis der IRT bilden dann explizite Annahmen über die Beziehung zwischen der latenten Variablen, der Antwort auf ein einzelnes Item und der Beziehung zu den Antworten

auf die restlichen Items. Diese Annahmen bilden die mathematische Grundlage, um Aussagen über die latente Variable auf Basis der Item-Antworten ableiten zu können [23]. Mittlerweile existieren hierfür bereits mehr als 100 unterschiedliche Modelle [22]. Bevor auf einzelne dieser Modelle näher eingegangen wird, sollen zunächst die Grundlagen hierfür durch Herleitung eines General-Modells, auf dem alle anderen Modelle fußen, erörtert werden [15].

(1) General-Modell

Die Wahrscheinlichkeit p einer Antwort x auf ein Item i lässt sich mit der Funktion

$$p(x_i) = f(\Xi) \tag{5}$$

darstellen, wobei Ξ die Item- und Personenparameter verkörpert. Für die Variable x_i unterscheidet man hierbei dichotome Formate, bei denen nur zwei Antwortmöglichkeiten, zum Beispiel richtig/ falsch oder ja/ nein, zur Verfügung stehen, von polytomen Antwortformaten, bei denen mehr als zwei Antwortmöglichkeiten existieren, beispielsweise in Form von Likert-Skalen. Entsprechend stellen dichotome Modelle beschränkte Varianten von polytomen Modellen dar [15].

Betrachtet man sich die Funktion $f(\Xi)$ näher, existieren prinzipiell verschiedene Ansätze. Der am häufigsten verwendete Ansatz nutzt hierfür die logistische Verteilungsfunktion Ψ[15], wobei e für die Euler`sche Zahl (≈ 2,718) steht [11]:

$$p(x_i) = f(\Xi) = \Psi(\Xi) = \frac{1}{1 + e^{-\Xi}} = \frac{e^{\Xi}}{1 + e^{\Xi}} \tag{6}$$

Diese Funktion, bezeichnet als Item Response Funktion (IRF), beschreibt die nonlineare Beziehung zwischen der Wahrscheinlichkeit eines manifesten Antwortverhaltens in Abhängigkeit von der Ausprägung der zugrunde liegenden latenten Variablen der Person und stellt die Grundlage aller im Weiteren vorgestellten Modelle dar [15, 24].

(2) 2-Parameter-logistisches-Modell (2-PL-Modell)

Bevor auf die komplexeren polytomen Modelle eingegangen wird, soll zunächst das 2-PL-Modell nach Birnbaum [21], das binäre Antwortkategorien voraussetzt, näher betrachtet werden. Es stellt eine wichtige Grundlage für die im Weiteren beschriebenen Modelle dar, bei denen es sich jeweils um erweiterte Varianten dieses Modells handelt.

Die Item- und Personenparameter Ξ lassen sich im 2-PL-Modell mit

$$\Xi = \gamma_i + \alpha_i\theta_v = \alpha_i(\theta_v - \delta_i) \tag{7}$$

darstellen, wobei θ_v der Ausprägung der latenten Variablen von Proband v, sowie α_i der Steigung, γ_i dem Schnittpunkt der logistischen Regressionslinie von Item i und δ_i der Lage von Item i auf dem Kontinuum der latenten Variablen entsprechen [15].

Setzt man nun Ξ in die Formel des General-Modells (Formel 6) ein, erhält man die Item Response Funktion des 2-PL-Modells:

$$p(x_i|\theta_v) = \Psi[\alpha_i(\theta_v - \delta_i)] = \frac{e^{\alpha_i(\theta_v-\delta_i)}}{1+e^{\alpha_i(\theta_v-\delta_i)}} = \frac{e^{\gamma_i+\alpha_i\theta_v}}{1+e^{\gamma_i+\alpha_i\theta_v}} \quad (8)$$

Bei der latenten Variablen θ_v sprechen höhere Werte für eine hochgradigere Schwere, wohingegen niedrigere Werte für eine geringere Ausprägung der latenten Variablen sprechen. Anhand der Formel $p(x_i|\theta_v) = \Psi[\alpha_i(\theta_v - \delta_i)]$ ist erkennbar, dass die Wahrscheinlichkeit einer bestimmten Antwort $p(x_i)$ maßgeblich durch die gewichtete (α_i) Differenz zwischen Personen (θ_v)- und Item (δ_i)-Location bestimmt wird [15].

Die graphische Repräsentation der IRF wird als Item Characteristic Curve (ICC) bezeichnet (Abbildung 1) und verdeutlicht, dass der Locationparameter δ_i eines Items auf demselben Kontinuum liegt, wie die Ausprägung der latenten Variablen θ_v eines Probanden (‚joint scale') [11]. Dies ermöglicht direkte Vergleiche zwischen dem Personenparameter θ_v und dem Itemparameter δ_i, der im 2-PL-Modell als Schwierigkeitsparameter eines Items fungiert. Der Steigungsparameter α_i, auch bezeichnet als Diskriminationsparameter, lässt sich anhand der Steigung der Tangentiallinie am Punkt δ_i ablesen [15].

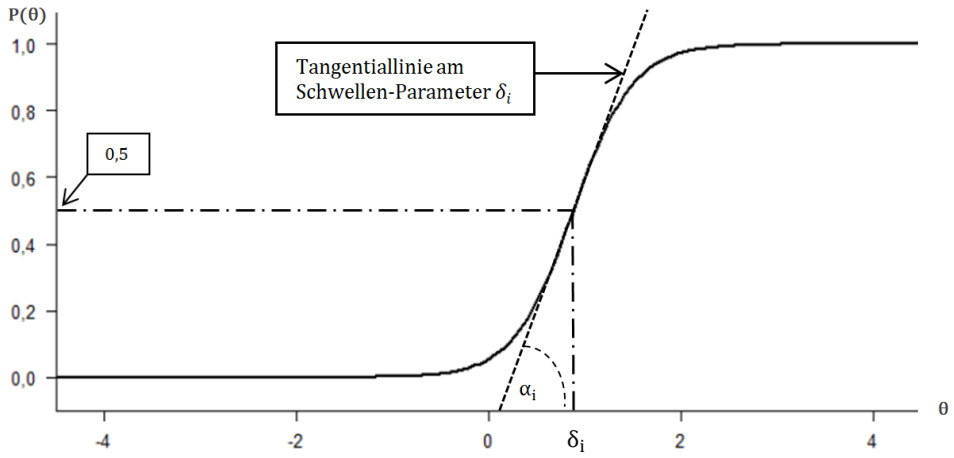


Abbildung 1: Item Characteristic Curve im 2-PL-Modell (angelehnt an de Ayala [15])

Um auch polytome Fragebögen mittels IRT evaluieren zu können, muss das 2-PL-Modell nun erweitert werden [25]. Werden geordnete Antwortkategorien genutzt, wie bspw. die Likert-Skala beim BDI-II, sind die bekanntesten Modelle das Partial Credit Modell (PCM) [26], das Generalized Partial Credit Modell (GPCM) [27] und das Graded Response Modell (GRM) [28], die im Folgenden näher erläutert werden.

(2) (Generalized) Partial Credit Modell (GPCM)

Beim Generalized Partial Credit Modell nach Muraki [27] lässt sich die Wahrscheinlichkeit, bei einem Item i Antwortkategorie k anstatt der vorhergehenden Antwortkate-

gorie *k-1* zu wählen, durch das 2-PL-Modell beschreiben [15]. Die Item- und Personenparameter Ξ werden hierbei folgendermaßen definiert:

$$\Xi = \sum_{h=1}^{k} \alpha_i(\theta_v - \delta_{ih}) \tag{9}$$

Insgesamt ergibt sich für das GPCM somit die folgende Item Response Funktion:

$$p\left(x_{ik}|\theta_v, \alpha_i, \underline{\delta_i}\right) = \frac{exp\left[\sum_{h=1}^{k} \alpha_i(\theta_v - \delta_{ih})\right]}{\sum_{c=1}^{m_i} exp[\sum_{h=1}^{c} \alpha_i(\theta_v - \delta_{ih})]} \tag{10}$$

Hierbei entspricht $p\left(x_{ik}|\theta_v, \alpha_i, \underline{\delta_i}\right)$ der Wahrscheinlichkeit, dass eine Person v mit der Ausprägung der latenten Variablen von θ_v bei Item i, das durch die Parameter α_i und $\underline{\delta_i}$ beschrieben wird, von m_i möglichen Kategorien genau Kategorie k wählt. Während α_i – wie beim 2-PL-Modell – der Steigung der IRF entspricht, repräsentiert der Schwellenparameter δ_{ik} beim GPCM jeweils den Punkt auf der latenten Variablen, an dem die Likelihoods der jeweiligen Antwortkategorien identisch sind. Entsprechend müssen die Schwellenparameter δ_{ik} keiner vorgegeben Ordnung folgen [15, 25].

Die Iteminformation $I(\theta_v)$ lässt sich im (Generalized) Parial Credit Modell aufbauend auf der vorausgehenden Formel (Formel 10) durch folgende Funktion beschreiben [25]:

$$I(\theta_v, x_i) = {\alpha_i}^2 \sum_{k=0}^{m_i} [k - E(x_{iv}|\theta_v)]^2 p(x_{ik}|\theta_v) \tag{11}$$

Das Partial Credit Modell (PCM) nach Masters [26], das annimmt, dass alle Items denselben Steigungsparameter haben, also entsprechend nur ein Parameter α für alle Items existiert, stellt dabei eine beschränkte Variante des GPCM dar [15].

(3) Graded Response Modell (GRM)

Ein weiteres, häufig eingesetztes Modell für polytome Daten, dem das 2-PL-Modell zugrunde liegt, ist das Graded Response Modell nach Samejima [28]. Das GRM nimmt an, dass die Beantwortung eines Items eine bestimmte Anzahl von Einzelschritten erfordert, bei denen das erfolgreiche Absolvieren eines Einzelschrittes das erfolgreiche Absolvieren aller vorhergehenden Schritte erfordert. Wird ein Schritt k erfolgreich absolviert, wird entsprechend angenommen, dass alle vorhergehenden Schritte ebenfalls erfolgreich absolviert wurden [25].

Die kumulative Wahrscheinlichkeit für Person v, Kategorie k oder höher zu wählen, wird dann mit Hilfe des 2-PL-Modells dargestellt [29]:

$$p^*(k) = \frac{exp[\alpha_i(\theta_v - \delta_{ik})]}{1 + exp[\alpha_i(\theta_v - \delta_{ik})]} = \frac{1}{1 + \exp[-\alpha_i(\theta_v - \delta_{ik})]} \tag{12}$$

Entsprechend wird die Wahrscheinlichkeit, genau Kategorie k zu wählen $[p^*(x_{ik})]$, als Differenz zwischen der Wahrscheinlichkeit Kategorie k oder höher zu wählen $[p^*(k)]$ und der Wahrscheinlichkeit Kategorie *k+1* oder höher zu wählen $[p^*(k+1)]$, berechnet

$$p^*(x_{ik}|\theta_v, \alpha_i, \delta_{ik}) = p^*(k) - p^*(k+1) \tag{13}$$

$$= \frac{1}{1+\exp[-\alpha_i(\theta_v - \delta_{ik})]} - \frac{1}{1+\exp\left[-\alpha_i\left(\theta_v - \delta_{i(k+1)}\right)\right]}$$

Hierbei gilt für $p^*(x_{ik}=0|\theta_v)=1$ und $p^*(x_{ik}=m_i|\theta_v)=0$. Auf Basis der vorhergehenden Formel lässt sich die Iteminformation folgendermaßen bestimmen [25]:

$$I(\theta_v, x_i) = \sum_{k=1}^{m_i+1} \frac{\left[\alpha_i\left(P^*_{ik-1}(\theta_v)\right) * \left(1 - P^*_{ik-1}(\theta_v)\right) - \alpha_i\left(P^*_{ik}(\theta_v)\right) * \left(1 - P^*_{ik}(\theta_v)\right)\right]^2}{P^*_{ik-1}(\theta_v) - P^*_{ik}(\theta_v)} \tag{14}$$

Anders als im (Generalized) Partial Credit Modell repräsentiert der Schwellenparameter δ_{ik} im GRM den Punkt auf der latenten Variablen, bei dem die Wahrscheinlichkeit Kategorie k oder höher zu wählen, 0,5 beträgt. Hieraus ergibt sich auch, dass die Schwellenparameter (δ_{ik}) im GRM definitionsgemäß geordnet sein müssen, da die Wahrscheinlichkeit in Kategorie k zu antworten nicht niedriger sein kann, als in Kategorie *k+1* zu antworten [25, 29].

Die Wahrscheinlichkeit in Abhängigkeit von der Ausprägung der latenten Variablen θ eine bestimmte Antwortkategorie eines Items zu wählen, lässt sich sowohl im GPCM als auch im GRM graphisch mittels einer Option Characteristic Curve (OCC) darstellen. Während beim (Generalized) Partial Credit Modell die Schwellenparameter δ_{ik} den Schnittpunkten benachbarter OCCs (τ_{ik}) entsprechen, stimmen diese beim Graded Response Modell nicht miteinander überein (Abbildung 2).

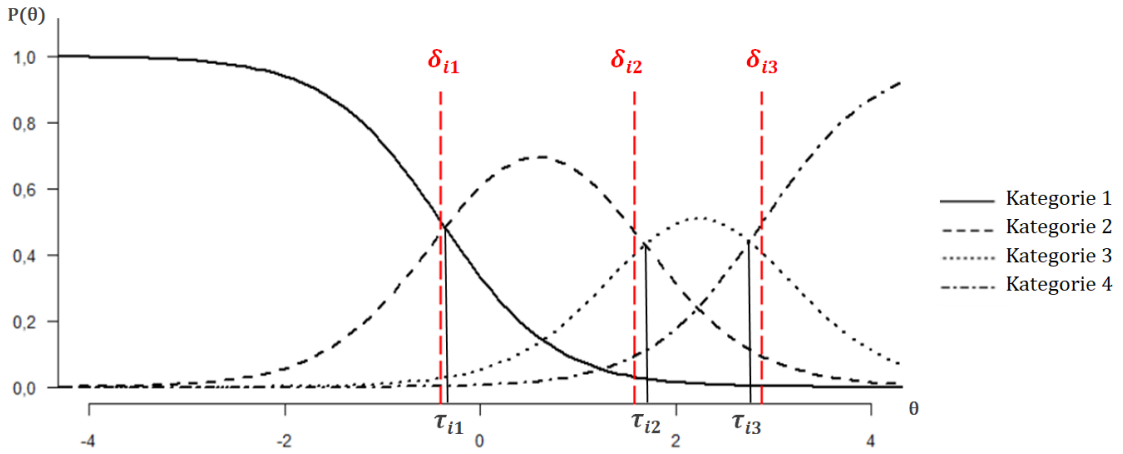


Abbildung 2: OCC eines Items im GRM (angelehnt an García-Pérez und de Ayala [15, 29])

Im Gegensatz zu den Schwellenparameter δ_{ik}, die im GRM geordnet sein müssen, müssen die Schnittpunkte der OCCs (τ_{ik}) bei beiden Modellen keiner vorgegebenen Ordnung folgen [25, 29].

Den bisher genannten Modellen gemeinsam sind die strengen Voraussetzungen für deren Anwendung. So fordern alle vier Modelle das Vorliegen von Monotonie und lokaler Unabhängigkeit [22]. Auf diese Kriterien soll daher in Abschnitt 3.4.2 näher eingegangen werden. Nur wenn diese hinreichend erfüllt werden können, ist eine unverzerrte Parameterschätzung mittels des gewählten Modells möglich.

Hierfür müssen für einen Datensatz, bspw. bestehend aus den Antworten einer Stichprobe auf einen Fragebogen wie den BDI-II, diejenigen Item- und Personenparameter gefunden werden, die am wahrscheinlichsten eine Datenmatrix wie die vorliegende erzeugen würden [11]. Hierbei unterscheidet man grundsätzlich Verfahren, die auf Bayes`sche Schätzverfahren zurückgreifen, von Maximum-Likelihood-basierten Ansätzen, bei denen die Wahrscheinlichkeit der beobachteten Antworten in Abhängigkeit von den unbekannten Personen- und Itemparametern als Likelihood-Funktion umschrieben werden. Während eine ungünstige Wahl der unbekannten Parameter niedrige Werte der Likelihood mit sich bringen, führt die bestmögliche Kombination von Item- und Personenparametern zu einer Maximierung der Likelihood. '*Durch systematische Veränderung der Werte versucht man [entsprechend], das erzielbare Maximum der Likelihood zu finden (ML-Verfahren). Als beste Schätzer für die unbekannten Parameter gelten jene Werte, bei denen die Likelihood unter Annahme lokaler stochastischer Unabhängigkeit ihren relativen Maximalwert erreicht*' [11, Seite 390].

Für mehrparametrische IRT-Modelle wird häufig die Marginal-Maximum-Likelihood (MML)-Schätzung eingesetzt, bei der neben den Itemparametern nicht die individuellen Personenparameter ($\hat{\theta}_v$), sondern primär lediglich die Verteilungsparameter der latenten Varaiblen geschätzt werden. Zur Approximation der im Rahmen der MML-Schätzung entstehenden Gleichungen, die zumeist nicht analytisch zu berechnen sind, kann der auf adaptive Verfahren zur numerischen Quadratur basierende Expectation-Maximization (EM)-Algorithmus nach Bock und Aitkin [30] verwendet werden [11].

Nach dem Schätzen der Itemparameter können dann in einem separaten zweiten Schritt die individuellen Personenparameter $\hat{\theta}_v$ geschätzt werden. Auch bei der Personenparameterschätzung unterscheidet man hierbei Maximum-Likelihood basierte Verfahren, auch bezeichnet als ‚ML-Scoring', und Bayes'sche Schätzer.

Beim ML-Scoring ist der Personenparameterschätzer $\hat{\theta}_v$ einer Person der Wert der latenten Variablen, bei dem die Antwortmusterwahrscheinlichkeit ‚*auf Basis der initial geschätzten Itemparameter*' maximal wird [11, Seite 484]. Allerdings erlaubt das ML-Scoring keine Punktschätzung für extreme Antwortmuster, d. h. für Probanden, die alle Items mit Kategorie 0, bzw. alle Items mit Kategorie 3 beantwortet haben.

Ein alternatives Vorgehen stellt der Bayes'sche Ansatz dar, bei dem zusätzlich zum vorliegenden Antwortmuster auf die A-priori-Verteilung der latenten Variable zurückgegriffen wird. Die Verteilungsparameter dieser A-priori-Verteilung, die in der Regel auf der Normalverteilung $N(\mu_\theta, \sigma_\theta)$ basieren, wurden parallel mit den Itemparametern im ersten Schritt geschätzt. Die individuelle A-posteriori-Verteilung der latenten Variablen lässt sich dann als normiertes Produkt von Likelihood und A-priori-Verteilung errechnen. Der Erwartungswert dieser A-posteriori-Verteilung, bezeichnet als Expected-a-posteriori (EAP)-Schätzung, eignet sich insbesondere bei unidimensionalen Modellen als Personenparameterschätzer, während der Maximalwert – die Maximum-a-posteriori (MAP)-Schätzung – vor allem im Bereich höherdimensionaler Modelle effizient ist. Beide Verfahren führen gleichermaßen zu präzisen Personenparameterschätzern, tendieren jedoch bei über- bzw. unterdurchschnittlichen Werten $\hat{\theta}_v$ zu einer leichten Verzerrung (‚Shinkage') in Richtung des Erwartungswertes der A-priori-Verteilung [11].

Aufgrund des Umfangs und der Komplexität der vorgenannten Verfahren wird an dieser Stelle auf eine ausführlichere Darstellung verzichtet und für nähere Informationen auf die ausführliche Zusammenfassung von Moosbrugger und Kelava [11], sowie die dort aufgeführte entsprechende Originalliteratur verwiesen.

1.3.4 Integration von klassischer Testtheorie und Item Response Theorie

Betrachtet man sich nun die Limitationen der klassischen Testtheorie, die zur Entwicklung der Item Response Theorie beigetragen haben, ist erkennbar, dass durch die IRT viele dieser Probleme gelöst werden konnten [6]. Dem Problem der Stichprobenabhängigkeit der Testcharakteristika begegnet die IRT dabei mit einer stichprobenunabhängigen Schätzung der Itemparameter ('personfree item estimation'). Entsprechend wird bspw. die Schwierigkeit eines Items gleich bewertet, unabhängig davon, ob sie in einer Stichprobe mit einer im Schnitt hohen oder einer durchschnittlich niedrigen Ausprägung der latenten Variablen geschätzt wird [15, 18] und ermöglicht somit Vergleiche auch bei Durchführung von Studien in verschiedenen Populationen. Zudem sind in der IRT die Personenparameter $\hat{\theta}_v$ unabhängig von den zur Errechnung genutzten Items ('itemfree person estimation') und somit auch zwischen Probanden vergleichbar, die einen völlig verschiedenen Item-Satz bearbeitet haben [6, 15]. Dies stellt eine zentrale Grundvoraussetzung für die Entwicklung von adaptiven Testverfahren dar [18]. Im Gegensatz dazu sind im Rahmen der klassischen Testtheorie Probanden nur dann vergleichbar, wenn sie den exakt gleichen Fragebogen vorgelegt bekommen. Dies erscheint logisch, wenn man sich folgendes Szenario vorstellt: Der erste Proband erhält das aus 21 Items bestehende BDI-II und erreicht bspw. einen Summenscore von 8 Punkten. Ein zweiter Proband erhält hingegen eine gekürzte Version des BDI-II, die

nur aus 10 Items besteht, und erhält ebenfalls einen Summenscore von 8 Punkten. Ein direkter Vergleich dieser Probanden ist anhand des einfachen Summenscores nicht möglich. Nutzt man hingegen zur Bewertung der Depressivität den Personenparameterschätzer $\hat{\theta}_v$, der itemunabhängig errechnet wird, ist ein Vergleich anhand dieses Parameters problemlos möglich – auch bei Bearbeitung eines vollständig unterschiedlichen Itemsatzes. Durch die Berechnung des Personenparameters $\hat{\theta}_v$ anstelle des klassischen Summenscores geht zudem die im Antwortmuster enthaltene Information durch Gewichtung anhand der Itemparameter direkt in die Bewertung der Depressionsschwere mit ein und erlaubt eine noch differenziertere Einschätzung der Depressivität.

Das IRT-basierte Linking mehrerer Instrumente, die dieselbe latente Variable erfassen, ermöglicht zudem, diese auf einer gemeinsamen Metrik zu verorten und somit direkte Vergleiche auch bei Bearbeitung verschiedener Fragebögen zu erlauben [31–33]. Auf die Depressionsschwere bezogen wäre somit denkbar, die beiden in Deutschland am häufigsten zur Evaluation der Depressivität genutzten Fragebögen – das BDI-II und den CES-D – miteinander zu verlinken. Anhand des Personenparameters $\hat{\theta}_v$ könnte somit die Depressionsschwere von Patienten, die den BDI-II bearbeiteten, direkt mit Patienten verglichen werden, die den CES-D vorgelegt bekamen, während bei Anwendung des Summenscores hier keine vergleichenden Aussagen möglich wären [31–33].

Der Einschränkung der klassischen Testtheorie, dass der Standardmessfehler über alle Facetten der latenten Variable als konstant angenommen wird, begegnet die IRT mit der Berechnung von Messfehlerstatistiken, die einen Index für die Genauigkeit der Personenparameterschätzung darstellen [15] und für jeden einzelnen Probanden ermöglichen, ein individuelles Konfidenzintervall um den Personenparameterschätzer $\hat{\theta}_v$ anzugeben [11]. Die Angabe dieses individuellen Konfidenzintervalls kann in der klinischen Anwendung wichtige Informationen über die Präzision der Parameterschätzung liefern und so entscheidende Anhaltspunkte für die klinische Entscheidungsfindung bieten, bspw. bei der Festlegung, ob ein Therapieeffekt – evaluiert mittels BDI-II – das Signifikanzniveau übersteigt oder im Bereich der Messungenauigkeit liegt. Die Errechnung von Testinformationskurven, die ebenfalls auf dem Standardmessfehler basieren, ermöglicht zudem die Veranschaulichung, in welchem Bereich der latenten Variablen das Instrument als Ganzes Stärken bzw. Schwächen aufweist. Dies ist bspw. für die Frage, ob ein Fragebogen gewinnbringender zum Screening oder im klinischen Setting eingesetzt werden sollte, essentiell. Wurden mehrere Instrumente, wie oben beschrieben, verlinkt, kann zudem ein direkter Vergleich dieser bezüglich der Aussagekraft in einem bestimmten Bereich der latenten Variablen erfolgen.

2. Zielstellung

Über die letzten Jahrzehnte wurde eine Vielzahl von internationalen Studien veröffentlicht, die dem BDI-II eine starke diagnostische Performance, sowie auf Basis der klassischen Testtheorie solide psychometrische Charakteristika bescheinigten. Die von Wang und Gorenstein [34] veröffentlichte systematische, länderübergreifende Datenbank-Analyse, die Studien zu den psychometrischen Charakteristika des BDI-II zwischen 1996 und 2012 berücksichtigte, ergab allerdings, dass von den 118 hier eingeschlossenen Studien lediglich drei Studien mit insgesamt ca. 600 Probanden die deutsche Version des BDI-II verwendeten. Dies betrifft insbesondere IRT-basierte Studien, die bisher weitestgehend auf fremdsprachigen Versionen des BDI-II basieren. Zu den wenigen IRT-basierten Studien an der deutschen Version des BDI-II zählt die von Alexandrowicz et al. [8] veröffentlichte Studie, die auf Basis der IRT die sichere Anwendbarkeit der deutschen Version des BDI-II sowohl in klinischen, als auch nicht-klinischen Populationen bestätigen konnte. Auch zu nennen wäre hier die Arbeit von Wahl et al. [33], die elf häufig in Deutschland verwendete Depressionsinventare – u. a. das BDI-I und das BDI-II – mittels IRT auf einer gemeinsamen Metrik vereinte. Eine strukturierte Überprüfung der grundlegenden Eignung für unidimensionale IRT-Analysen, sowie die Evaluation der allgemeinen psychometrischen Charakteristika, wie dies bspw. von de Sá Junior et al. [35, 36] für die portugiesische Version des BDI-II durchgeführt wurde, steht für die deutsche Version des BDI-II aktuell noch aus. Insbesondere die Erkenntnisse bezüglich Differential Item Functioning (DIF) gehen bisher ausschließlich auf fremdsprachige Versionen des BDI-II zurück, sodass hier noch weiterer Forschungsbedarf besteht.

Ziel der vorliegenden Arbeit ist es daher, anhand des Antwortverhaltens einer Stichprobe junger Menschen in der Ausbildungsphase die deutsche Version des BDI-II [1, 2] systematisch auf seine Eignung für unidimensionale IRT-Analysen hin zu überprüfen, auf Basis der IRT die allgemeinen psychometrischen Charakteristika des BDI-II zu analysieren, sowie Aussagen zu DIF bezüglich Alter und Geschlecht abzuleiten.

Durch eine Item Faktor Analyse soll hierfür zunächst die Dimensionalität des BDI-II in der vorliegenden Population näher untersucht werden, die die Grundlage für die Wahl eines geeigneten IRT-Modells – unidimensional vs. multidimensional – darstellt. Bisher veröffentlichte Studien zur Faktorenstruktur des BDI-II, wie z.B. Subica et al. [37] oder Faro et al. [38], konnten keine einheitliche Struktur nachweisen, legen aber die Erwartung nahe, dass der Großteil der BDI-II-Item-Varianz durch den Generalfaktor ‚De-

pressivität‘ erklärt wird. Bestätigt sich diese Annahme in der vorliegenden Stichprobe, sollen die weiteren Analysen mittels unidimensionaler IRT-Modelle erfolgen.

Da inferenzstatistische Aussagen über die latente Variable eines Probanden auf Grundlage der IRT jedoch nur zulässig sind, wenn das gewählte IRT-Modell die Daten in adäquater Weise abbilden kann [23], muss vor der Interpretation der Ergebnisse zunächst die Anpassung an den vorliegenden Datensatz für die in Frage kommenden Modelle überprüft werden. Hierfür sollen Goodness-of-Fit-Statistiken auf Modell- und Item-Ebene zum Einsatz kommen, um so das IRT-Modell zu finden, das den vorliegenden Datensatz am präzisesten abbilden kann. Um mögliche Fehlerquellen, die zu einem Modell-Misfit führen können, auszuschließen, sollen für alle Items die Kriterien der Monotonie und lokalen Unabhängigkeit überprüft werden.

Um aberrantes Antwortverhalten von Probanden zu detektieren, soll die Personfit-Statistik Anwendung finden. Auf diese Weise sollen Probanden detektiert werden, für die der errechnete IRT-Score keine zuverlässige Einschätzung der Depressivität erlaubt. Klinisch bietet dies bspw. die Möglichkeit, Probanden zu identifizieren, die durch (Dis-)Simulation ihr Testergebnis zu manipulieren versuchen.

Die IRT-Analyse soll dann stichprobenunabhängig eine Einschätzung der Diskriminationsfähigkeit und Schwierigkeit aller Items des BDI-II erlauben. Mittels Analyse der Iteminformationskurven sollen zudem diejenigen Items identifiziert werden, die eine besonders hohe Diskriminationsfähigkeit aufweisen. Über die Testinformationskurve mit zugehörigem Standardfehler soll abgeschätzt werden, in welchem Bereich der latenten Variablen das BDI-II seine maximale Aussagekraft und somit seine maximale Präzision besitzt, bzw. in welchem Bereich das Instrument Schwächen aufweist.

Um einen Test sicher anwenden zu können, muss zudem gewährleistet sein, dass der Test die latente Variable für alle Probanden gleichartig misst, unabhängig von verschiedenen demographischen Eigenschaften, wie bspw. dem Alter. Ist dies nicht der Fall, spricht man von DIF. Für den deutschsprachigen Raum fehlen bisher strukturierte IRT-basierte DIF-Analysen für den BDI-II. Da der Nachweis der Äquivalenz psychometrischer Charakteristika zwischen den sprachlichen Fassungen für den BDI-II noch aussteht und für den BDI-IA im Rahmen der ODIN-Studie nicht nachgewiesen werden konnte [39], ist eine Übertragung der international gewonnenen Erkenntnisse auf die deutsche Version nur mit Vorbehalt möglich. Insbesondere die Analyse auf DIF bezüglich des Geschlechtes ist jedoch ganz zentral, da in diversen Untersuchungen eine Tendenz zu höheren BDI-II-Testscores bei weiblichen im Vergleich zu männlichen Probanden beobachtet werden konnte [3, 40], die erst nach dem Ausschluss von DIF als real vorliegende Geschlechtsunterschiede in der Depressionsschwere interpretiert werden können, sodass die vorliegende Arbeit versucht, diese Lücke zu schließen.

3. Material und Methodik

3.1 Stichprobe

3.1.1 Auswahl der Stichprobe und Rekrutierung

Vor der ersten Datenerhebung im Rahmen des Studienkomplexes ‘Psychische Gesundheit von Studierenden’ erfolgte eine ausführliche Beurteilung der Studie durch die Ethikkommission der Martin-Luther-Universität Halle-Wittenberg. Vor Ausweitung der Befragungen auf Studierende und Berufsschüler weiterer Fachrichtungen erfolgte jeweils eine erneute Beurteilung durch die Ethikkommission, die die Studien in der vorgestellten Form als ethisch unbedenklich einstufte und ein positives Votum (2017-138) ausstellte. Die Probanden nahmen nach ausführlicher Aufklärung über die Ziele der Studie, den Umgang mit den Daten (Pseudonomisierung), sowie der Zusicherung der Freiwilligkeit und insbesondere der Konsequenzenfreiheit bei Nichtteilnahme, an der Studie teil. Die im Rahmen der Arbeit verwendete Formulierung ‘Probanden’ soll geschlechtsneutral sowohl weibliche, als auch männliche Teilnehmer einschließen. Da minderjährige Probanden nur nach Einverständnis der Eltern an der Befragung teilnehmen könnten und dies mit der Anonymität der Teilnahme kollidierte, wurden diese von einer Studienteilnahme ausgeschlossen.

Die Befragungen der Medizin- und Zahnmedizinstudierenden der Martin-Luther-Universität Halle-Wittenberg erfolgten im Zeitraum vom Wintersemester 2017/18 bis zum Sommersemester 2020 im Rahmen von Pflichtveranstaltungen an den jeweiligen Fakultäten. Von den circa 1150 Medizinstudierenden der befragten Semester nahmen 1107 (96,3%) und von den 184 Zahnmedizinstudierenden nahmen 155 (84,2%) freiwillig an der Studie teil.

Die Befragungen der Studierenden der Rechtswissenschaften und der Psychologie fanden im Wintersemester 2018/19 im Rahmen von freiwilligen Lehrveranstaltungen an der Martin-Luther-Universität Halle-Wittenberg statt. Von den circa 1050 Jura-Studierenden der befragten Semester konnten circa 480 im Rahmen der Veranstaltungen angetroffen und zur Befragung eingeladen werden und 306 (63,8%) davon nahmen an der Studie teil. Bei den Psychologie-Studierenden entschieden sich von den 310 angetroffenen Studierenden 109 (35,2%) zur Studienteilnahme.

Bei den Berufsschülern und -schülerinnen der Berufsbildenden Schule V in Halle (Saale) fand die Befragung im Wintersemester 2019/20 im Rahmen von Pflichtveranstaltungen der verschiedenen Ausbildungsgänge statt. Insgesamt kamen 680 volljährige Personen für die Studie in Frage. Hiervon entschieden sich 555 (81,6%) für die Teilnahme an der Befragung.

Die Erhebung an der Friedrich-Schiller-Universität Jena und der Fachhochschule für Gesundheit in Gera erfolgte über Online-Fragebögen. Die Studierenden wurden über den Universitätsverteiler bzw. den Verteiler der Fachhochschule über die Studie informiert und zur Teilnahme eingeladen. Von der Friedrich-Schiller-Universität Jena nahmen 86 Psychologie- und 78 Medizin-Studierende, sowie von der Fachhochschule in Gera 75 Psychologie-Studierende teil.

3.1.2 Deskriptive Charakteristika der Stichprobe

Insgesamt nahmen auf diese Weise 2471 junge Menschen an der Befragung teil. Nach Ausschluss der miterfassten Minderjährigen verblieben noch 2457 Probanden, von denen jedoch 81 den Teilbereich des BDI-II nicht vollständig ausgefüllt hatten. Wertet man die Verteilung der nicht beantworteten Items aus, zeigte sich eine homogene Verteilung über die Items 1 bis 20 von ca. 0,01%, während sich die Rate an fehlenden Antworten für Item 21 (‘Verlust des Interesses an Sex’) mit 0,02% etwas abhob.

Bei den 43 Probanden, die weniger als 3 Fragen des BDI-II unbeantwortet ließen, wurden die Antworten auf diese Items mittels der ‚k-nearest-neighbors‘- (KNN)-Technik ersetzt. Hierfür wurden für die betroffenen Probanden jeweils die k Probanden herausgefiltert, deren Antwortmuster dem mit fehlenden Angaben am ähnlichsten waren. Die fehlende Antwort wurde dann durch die Antwortkategorie ersetzt, die am häufigsten von den k ausgewählten Probanden auf das nicht beantwortete Item gewählt wurde [41, 42]. Für die Berechnung wurde bei der vorliegenden Arbeit ein k von fünf Probanden festgelegt.
Da bei mehr als drei fehlenden Antworten nicht mehr von einem gewissenhaften Ausfüllen des Fragebogens ausgegangen werden kann, wurden die 38 Probanden, bei denen mehr als drei Items unbeantwortet blieben, aus den weiteren Analysen ausgeschlossen.

Somit entstand die endgültige Stichprobe, die sich aus 2419 jungen Menschen in der Ausbildungsphase zusammensetzte. Diese wiesen im Mittel einen BDI-II-Summenscore von 10,35 (SD=8,53) auf. Als klinisch unauffällig, einem BDI-II-Summenscore von 0 bis 13 Punkten entsprechend, wurden 73,1% (n=1768) eingestuft, wohingegen 13,4% (n=325) Hinweise auf ein mildes (14 bis 19 Punkte), 9,3% (n=226) Hinweise auf ein moderates (20 bis 29 Punkte) und 4,1% (n=100) Hinweise auf ein schweres depressives Syndrom (BDI-Summenscore >30 Punkte) zeigten. Angaben zur Geschlechtsverteilung, der Altersstruktur und der Prävalenz psychischer Erkrankungen in der vorliegenden Stichprobe finden sich in Tabelle 1.

Tabelle 1: Deskriptive Charakteristika der Stichprobe

Geschlecht	
- weiblich	68,2 (1650)
- männlich	31,6 (765)
Alter :	
- Mittelwert ± Standardabweichung	22,59 ± 4,54
Psychische Erkrankung in der Familie	
- Nicht vorhanden	67,5 (1633)
- Vorhanden	26,2 (634)
Depression in der Familie	
- Nicht vorhanden	65,2(1578)
- Vorhanden	18,2 (439)
Eigene Psychische Erkrankung bekannt	
- Ja	16,3 (394)
- Nein	76,4 (1848)
Eigene Depression diagnostiziert	
- Ja	7,2 (175)
- Nein	85,4 (2065)

Die Differenz zu 100% entsteht jeweils durch fehlende Angaben der Probanden zu den jeweiligen Kategorien.

3.2 Messinstrument

Die vorliegende Arbeit ist Teil eines größeren Studienkomplexes [43–45], der sich mit der Prävalenz und den Risikofaktoren depressiver Symptome bei Studierenden und Berufsschülern beschäftigte. Im Rahmen dieser Studie erhielten die an einer Teilnahme Interessierten ein Informationsschreiben zu Zielen und Ablauf der Studie, Umgang mit den Daten, sowie den expliziten Hinweis auf die Freiwilligkeit der Studienteilnahme. Um an der Studie teilzunehmen, wurden die Probanden gebeten, den Fragebogen in Ruhe und in Einzelarbeit so ehrlich wie möglich zu beantworten und anschließend in einem verschlossenen Umschlag in die hierfür deponierten Boxen einzuwerfen. Die Rückgabe des ausgefüllten Fragebogens wurde als Einwilligung zur Studienteilnahme gewertet.

Im Folgenden sollen nun die Abschnitte des verwendeten Fragebogens näher vorgestellt werden, die für die vorliegende Arbeit Relevanz haben. Informationen zu den übrigen Bestandteilen sind ausführlich in der Arbeit von Kindt et al. [43] dargestellt.

Der erste Teil des Fragebogens erfasste Angaben zu den soziodemographischen Eigenschaften der Probanden, wie beispielsweise Alter, Geschlecht und Familienstand. Als zweites wurden die Probanden gebeten, die aus 21 Items bestehende deutsche Fassung des Beck Depressions-Inventars-II [1, 2] zu beantworten. Außer den Items 16 (‘Schlafveränderungen’) und 18 (‘Appetitveränderungen’), die jeweils 7 Antwortkategorien bieten und anschließend mit null bis drei Punkten bewertet werden, erfolgt die

Beantwortung der restlichen 19 Items mittels 4-Punkt-Likert-Skala. Entsprechend sind Summenscores von 0 - 63 Punkten möglich, wobei höhere Werte für eine stärkere Ausprägung der depressiven Symptomatik sprechen. Dem DSM-IV folgend beziehen sich alle Fragen auf einen Zeitraum, der die letzten beiden Wochen inklusive des Tages der Bearbeitung des BDI-II umfasst. Zur Beurteilung wird meist die Unterteilung anhand folgender Kriterien genutzt: Als klinisch unauffällig gelten Probanden mit einem Summenscore zwischen 0 und 13 Punkten. Ein Score zwischen 14 und 19 weist auf eine mildes, zwischen 20 und 29 auf eine moderates und zwischen 30 und 63 Punkten auf ein schweres depressives Syndrom hin [2, 19].

3.3 Software

Alle im folgenden Kapitel erläuterten Analysen wurden mithilfe des Statistikprogramms R [46], sowie der im Folgenden aufgeführten R-Pakete durchgeführt. Um die fehlenden Antworten im Teilbereich des BDI-II zu ersetzen, wurde das Paket ‘VIM’ [42] genutzt. Die Berechnungen auf Grundlage der klassischen Testtheorie basieren auf dem Paket ‘psych’ [47], die IRT-Analysen auf dem Paket ‘mirt’ [48]. Für die Berechnung der non-parametrischen IRT-Modelle wurde das ‘KernSmoothIRT’ [49]-, für die Prüfung auf Normalverteilung das ‘userfriendlyscience’ [50]-, für die Prüfung auf Tau-Äquivalenz das ‘coeffizientalpha’ [51], und für die IRT-basierten DIF-Analysen das ‘lordif’ [7]- Paket genutzt. Die am Ende der Arbeit vorgestellte Online-Applikation nutzt zur Gestaltung neben den hier aufgeführten Paketen noch Weitere, die explizit dort aufgeführt sind.

3.4 Statistische Methoden

Als Grundlage für die strukturierte Analyse des BDI-II wurde der ‚Leitfaden mit den wissenschaftlichen Standards der Fragebogen-Entwicklung und Validierung‘ [52] der PROMIS-Forschungsinitiative genutzt. Das ‚Patient Reported Outcome Measurement Information System‘ (PROMIS), das von den National Instituts of Health (NIH) finanziert wird, verfolgt das Ziel, effiziente, präzise und valide Selbstbeurteilungs-fragebögen zum Thema Gesundheit und Wohlbefinden zur Verfügung zu stellen. Die PROMIS-Forschungsinitiative unterstützt klinisch Tätige, sowie Wissenschaftler mit aktuellen Standards zur Entwicklung und Validierung von Fragebögen [22, 52]. Diese Standards wurden daher als Leitfaden für die vorliegende Arbeit genutzt. Die durch-geführten Analysen lassen sich hierbei grob in zwei Teilbereiche untergliedern.

Begonnen wird mit einer strukturierten Analyse des BDI-II, die das Ziel verfolgt, dasjenige IRT-Modell zu identifizieren, das den vorliegenden Datensatz am treff-sichersten abzubilden vermag. Dies stellt die grundlegende Voraussetzung dafür dar,

im zweiten Teilbereich der Arbeit anhand des gewählten IRT-Modells Aussagen über die Informationsstruktur des BDI-II ableiten zu können.

3.4.1 Klassische Item Analyse

Um dem Ziel näher zu kommen, die psychometrischen Eigenschaften des BDI-II zu evaluieren, wurde zunächst eine klassische Item Analyse mit:

- absoluter und prozentualer Verteilung der Antworten pro Kategorie,
- Mittelwerten der Itemscores mit entsprechender Standardabweichung,
- mittlerem Testscore mit Standardabweichung,
- erreichtem Wertebereich der Testscores,
- Inter-Item- und Item-Total-Korrelationen und
- Abschätzung der Reliabilität

durchgeführt [52]. Aufgrund der in der Literatur vermehrt angeführten Kritik, dass Cronbach α nur bei normalverteilten und essentiell τ-äquivalenten Daten einen sinnvoll interpretierbaren unteren Grenzwert der Reliabilität angibt [17], wurde im vorliegenden Datensatz zunächst die Annahme der essentiellen τ-Äquivalenz geprüft. Diese besagt, dass – Unidimensionalität vorausgesetzt – die Faktorladungen aller Items denselben Wert aufweisen. Hierfür wurde eine robuste F-Statistik mit der Nullhypothese, dass der Datensatz sich essentiell τ-äquivalent verhält, durchgeführt. Bei einem signifikanten Ergebnis muss die Nullhypothese der essentiellen τ-Äquivalenz abgelehnt werden [51]. Als nächstes wurde überprüft, ob die Daten einer Normalverteilung folgen. Hierfür wurde neben der Berechnung von Skewness (Schiefe) und Kurtosis (Wölbung) des Tests, die optische Auswertung mittels Histogramm und QQ-Plot (quantil-quantil-Plot), sowie die Berechnung des Shapiro-Wilk-Tests durchgeführt. Ein signifikantes Ergebnis spricht hier für nicht-normalverteilte Daten [53].
Bei Verletzung der essentiellen τ-Äquivalenz und Vorliegen einer Normalverteilung wäre McDonald's ω_t, das lediglich τ-Kongenerität voraussetzt, eine sinnvolle Alternative zu Cronbach α, wohingegen bei nicht-normalverteilten und nicht-τ-äquivalenten Daten der Greatest Lower Bound (GLB) die bessere Alternative darstellt, da dieser sich robuster gegen eine Verletzung der Normalverteilung zeigt [54]. Als unterer Grenzwert der Reliabilität wird – unabhängig vom gewählten Reliabilitätskoeffizienten – in der Regel 0,90 empfohlen [22, 52].

3.4.2 IRT-Modellauswahl

Nach der Auswertung der generellen Eigenschaften des BDI-II stand als nächster Schritt die Auswahl eines IRT-Modells an. Dieser Modell-Wahl kommt dabei eine ganz zentrale Rolle zu, da ein passendes IRT-Modell die essentielle Grundlage dafür

darstellt, im weiteren Verlauf der Arbeit auf Grundlage dieses Modells allgemeingültige Aussagen über das BDI-II ableiten zu können. Dies ist nur zulässig, wenn das Modell den vorliegenden Datensatz treffsicher abbilden kann [23].
In der vorliegenden Arbeit wurde die Modell-Wahl daher schrittweise mithilfe eines Stufenplanes erarbeitet und überprüft. Den ersten Schritt bildete hierbei die Untersuchung der dem BDI-II zugrunde liegenden Faktorstruktur, die im weiteren Verlauf ein zentrales Element der Modell-Wahl darstellt.

(1) Evaluation der Dimensionalität des BDI-II

Grundsätzlich unterscheidet man bei der Dimensionalität unidimensionale Instrumente, bei denen nur eine einzige latente Variable Einfluss auf die Beantwortung aller Items nimmt, von multidimensionalen Fragebögen, bei denen mehr als eine latente Variable Einfluss ausübt [22]. Um die einem Fragebogen zugrunde liegende Dimensionalität zu evaluieren, bestehen prinzipiell zwei äquivalente Ansätze:

- die Faktorenanalyse, die der klassischen Testtheorie zugerechnet wird und
- die Item Faktor Analyse [55], die auf der IRT basiert [56].

Aufgrund der Äquivalenz der Ansätze wurde für die vorliegende Arbeit im Folgenden die Item Faktor Analyse verwendet, da sie analog zur restlichen Arbeit auf den Grundlagen der IRT basiert. Hierbei unterscheidet man – wie bei der klassischen Faktorenanalyse – einen explorativen von einem konfirmatorischen Ansatz. Die explorative Item Faktor Analyse dient dabei dem Zweck, bei unbekannter Struktur eines Instrumentes Hypothesen über mögliche zugrundeliegende Faktorstrukturen zu generieren [56].

Da seit der Veröffentlichung des BDI-II eine Vielzahl von Studien zu dessen Faktorstruktur durchgeführt wurde, existieren hierüber bereits einige Hypothesen (Tabelle 2, nach [37, 38]). Die am häufigsten nachgewiesene klassische Struktur basiert hierbei auf den zwei Faktoren ‚kognitiv' und ‚somatisch', wobei sich die Itemzuordnung zu diesen Faktoren jedoch studienabhängig unterschied. Vereinzelt wurden auch Einfaktoren- oder Dreifaktorenlösungen beschrieben, allerdings auch hier nicht mit einheitlicher Zuordnung der Items zu den Faktoren [37, 38]. Wang und Gorenstein [34] bieten eine gute Übersicht bisheriger Studien, die sich mit der klassischen Faktorstruktur des BDI-II befassen.

In den letzten Jahren gewann die Anwendung von sogenannten Bifaktormodellen zur Beschreibung von psychologischen Instrumenten zunehmend an Bedeutung. Hierbei wird ein latentes Konstrukt sowohl durch einen Generalfaktor beschrieben, der den größten Teil zur erklärten Varianz beiträgt, als auch durch mehrere Gruppenfaktoren, die nur einen begrenzten Anteil an der gesamten erklärten Varianz beitragen [57].

Auch für den BDI-II konnte in einigen Arbeiten, beispielsweise von Faro et al. [38] und Subica et al. [37], eine Überlegenheit der Bifaktor-Modelle gegenüber den klassischen Faktorenstrukturen nachgewiesen werden, wobei die überprüften Gruppenfaktoren den in Tabelle 2 dargestellten klassischen Faktorstrukturen entsprachen.

Tabelle 2: Häufig gefundene Faktorstrukturen nach [37, 38]

Faktorstruktur	Autor*	Extrahierte Faktoren	Itemzuordnung
Eindimensional	Steer et al., 1999	Depressivität	1-21
Zweidimensional	Beck et al., 1996	Somatisch-affektiv	4,10-13, 15-21
		Kognitiv	1-3, 5-9, 14
	Dozois et al., 1998	Kognitiv-affektiv	1-3, 5-9, 13-14
		Somatisch-vegetativ	4, 10-12, 15-21
	Gorenstein et al., 2011	Kognitiv-affektiv	1-10, 12, 14
		Somatisch-affektiv	11, 13, 15-21
Dreidimensional	Osman et al., 1997	Negative Einstellung	1-3, 5-10, 14
		Performanceprobleme	4, 12-13, 15, 17, 19-20
		Somatische Probleme	11, 16, 18, 21
	Beck et al., 2002	Kognitiv	3, 5-8, 13-14
		Somatisch	10-11, 15-21
		Affektiv	1,2,4,9,12

**Die Quellenangaben der Originalarbeiten finden sich in [37, 38]*

Aufgrund dieser bereits bestehenden Hypothesen zur Faktorenstruktur des BDI-II wurde in der vorliegenden Arbeit auf eine erneute explorative Item Faktor Analyse verzichtet. Stattdessen wurden die bisher am häufigsten nachgewiesenen Faktorenstrukturen (Tabelle 2), sowie die bei den multidimensionalen Modellen korrespondierenden Bifaktor-Modelle, als Hypothesen verwendet und mittels konfirmatorischer Item Faktor Analyse auf ihre Anwendbarkeit für den vorliegenden Datensatz hin überprüft. Als Kriterien zur Bewertung wurden hierfür die beiden relativen Fit-Indices Tucker Lewis Index (TLI) und Comparative Fit Index (CFI), sowie die beiden absoluten Fit-Indices Standardized root mean residual (SRMR) und Root mean square error of approximation (RMSEA) verwendet. Für die jeweilige Bedeutung und die Formeln zur Berechnung dieser Indices wird auf die Arbeit von Hu und Bentler verwiesen [58].
In der Regel wird als Grenzwert beim TLI und CFI ein Wert >0,95 angesehen, während bei den absoluten Fit-Indices ein Wert von SRMR < 0,08, sowie RMSEA < 0,06 darauf hinweisen, dass die untersuchte Faktorstruktur die dem Fragebogen tatsächlich zugrunde liegende gut beschreiben kann [22, 59].

Zusammenfassend ermöglicht die Item Faktor Analyse, die dem BDI-II zugrunde liegende Faktorstruktur herauszuarbeiten und somit die Faktoren zu identifizieren, die einen relevanten Einfluss auf das Antwortverhalten der Probanden ausüben.

Für die Auswahl eines passenden IRT-Modells ist die nachgewiesene Faktorstruktur von zentraler Bedeutung, da hiervon abhängt, ob ein unidimensionales IRT-Modell verwendet werden kann, oder ob die vorliegende Struktur die Anwendung der deutlich komplexeren multidimensionalen IRT-Modelle erfordert [57].

(2) Festlegung der Dimensionalität der geplanten IRT-Analyse

Die o. g. Fragestellung lässt sich – abhängig von der nachgewiesenen Faktorstruktur – nicht immer eindeutig beantworten. Dies gilt insbesondere dann, wenn sich die Faktorstruktur eines Instrumentes am besten mithilfe eines Bifaktor-Modells beschreiben lässt. Für diese Fälle empfiehlt die PROMIS-Forschungsinitiative die Anwendung von unidimensionalen IRT-Modellen, wenn die Kriterien für essentielle Unidimensionalität für einen Fragebogen hinreichend erfüllt werden können [52]. Diese beinhalten:

- einen hohen Eigenwert des 1. Faktors, der mind. 20% der Variabilität erklärt,
- eine Ratio von erstem zu zweitem Eigenwert >4,
- eine Explained Common Variance of the general factor mit

$$ECV = \frac{(\sum \lambda_{gen}^2)}{(\sum \lambda_{gen}^2)+(\sum \lambda_{grp1}^2)+(\sum \lambda_{grp2}^2)} \geq 0{,}6, \quad (15)$$

- einen Koeffizient ω_H mit

$$\omega_H = \frac{(\sum \lambda_{gen})^2}{(\sum \lambda_{gen})^2(\sum \lambda_{grp1})^2(\sum \lambda_{grp2})^2+\sum(1-h^2)} \geq 0{,}8, \quad (16)$$

- Faktor-Ladungen auf den Generalfaktor >0,3 und signifikant höher als auf die Gruppenfaktoren, sowie
- Fit-Indices der Einfaktorlösung:

  SRMR ≤ 0,06, RMSEA ≤ 0,08, TLI ≥ 0,95, CFI ≥ 0,95

  (für die Berechnungsmöglichkeiten siehe Hu und Bentler [58]) [22, 60].

Können diese Kriterien ausreichend erfüllt werden, ist davon auszugehen, dass durch eine eventuell vorliegende Multidimensionalität die Parameterschätzungen bei Nutzung eines unidimenisonalen IRT-Modells nicht signifikant beeinflusst werden und folglich die Anwendung eines unidimensionalen Modells gerechtfertigt werden kann [60].
Zusammenfassend dient die Überprüfung der in diesem Abschnitt aufgeführten Kriterien folglich dazu, entscheiden zu können, ob es für den BDI-II zulässig ist, die weiteren Analysen mittels unidimensionaler IRT-Modelle durchzuführen.

Ist die Entscheidung zur Dimensionalität der IRT-Analyse getroffen, steht als nächster Schritt die Auswahl eines geeigneten IRT-Modells an. Wie in Abschnitt 1.3.3 bereits ausgeführt, existieren hierfür mittlerweile mehr als 100 verschiedene Modelle [22]. Für Fragebögen wie den BDI-II mit polytomen, geordneten Antwortkategorien in Form von Likert-Skalen bieten sich insbesondere die drei bereits in Abschnitt 1.3.3 näher

beschriebenen parametrischen IRT-Modelle an [25]: das Partial Credit Modell [26], das Generalized Partial Credit Modell [27] und das Graded Response Modell [28].
Werden parametrische IRT-Modelle zur Beschreibung eines Datensatzes genutzt, stellt die Anpassungsgüte, auch bezeichnet als Goodness-of-Fit (GOF), allerdings ein häufiges Problem dar [18, 61]. Als Anpassungsgüte wird hierbei bezeichnet, wie gut ein gewähltes Modell den zugrundeliegenden Datensatz abbilden kann [62]. Die Anpassungsgüte ist von zentraler Bedeutung, da es nur dann zulässig ist, auf Grundlage des gewählten Modells allgemeingültige Aussagen über das BDI-II, sowie die Depressivität der Probanden abzuleiten, wenn das entsprechende IRT-Modell den Datensatz treffsicher abbilden kann [23].
Zur Überprüfung der Anpassungsgüte von IRT-Modellen existiert eine Vielzahl von Möglichkeiten. In der vorliegenden Arbeit wurden hierfür Überprüfungen der Anpassungsgüte auf Modell-, Item- und Personen-Ebene durchgeführt.

Als erstes wurde die Anpassung des Modells als Gesamtes, also auf Modell-Ebene, überprüft. Dieses Verfahren wird als globaler Goodness-of-Fit bezeichnet und soll im folgenden Abschnitt näher erläutert werden [23, 62].

(3) Goodness-of-Fit auf Modell-Ebene

Eine Möglichkeit, die Anpassung eines Modells für einen vorliegenden Datensatz zu überprüfen, stellt die globale GOF-Statistik dar. Hierbei wird überprüft, ob sich die unter dem gewählten Modell erwarteten Antworthäufigkeiten signifikant von den tatsächlich im vorliegenden Datensatz beobachteten Häufigkeiten unterscheiden [62].

Prinzipiell unterscheidet man hierbei Ansätze, bei denen alle Antworthäufigkeiten evaluiert und die daher als Full-Information-GOF-Statistik bezeichnet werden, von Ansätzen, bei denen nur bestimmte Antworthäufigkeiten berücksichtigt und die daher als Limited-Information-GOF-Statistik bezeichnet werden [23].

Verdeutlichen lässt sich dies, wenn der vorliegende Datensatz – bestehend aus den Antworten von N Probanden auf n Items mit je m_i Antwortkategorien – in eine Kontingenztafel eingetragen wird. Analog wird anhand der vom gewählten IRT-Modell vorhergesagten Antworthäufigkeiten eine separate zweite Kontingenztafel erstellt [62].

Um zu überprüfen, ob sich die Häufigkeiten in den beiden Kontingenztafeln signifikant unterscheiden, kann zum einen die Full-Information-GOF-Statistik genutzt werden:

$$X^2 = N \sum_{c=1}^{C} \frac{(p_c - \hat{\pi}_c)^2}{\hat{\pi}_c} \tag{17}$$

Hierbei repräsentiert p_c die tatsächlich beobachtete Häufigkeit und $\hat{\pi}_c$ die unter dem gewählten Modell erwartete Häufigkeit einer Zelle c [62, 63]. Da die Kontingenztafeln

jeweils aus $C = \prod_i^n m_i$ Zellen bestehen, werden diese Kontingenztafeln insbesondere bei Fragebögen mit mehreren Antwortoptionen pro Item allerdings sehr umfangreich.

Auf Grund dessen entwickelten Maydeu-Olivares und Joe [63] eine Weiterentwicklung dieses Verfahrens, die Limited-Information GOF-Statistik, deren bekanntester Vertreter die M_2-Statistik ist. Diese Weiterentwicklung beruht darauf, dass sich die Häufigkeiten der oben beschriebenen Kontingenztafel auch als Marginalhäufigkeiten darstellen lassen. Bei der M_2-Statistik werden dann, wie der Name schon nahelegt, nur die Marginalhäufigkeiten erster und zweiter Ordnung berücksichtigt, während Informationen, die in höherrangigen Marginalhäufigkeiten enthalten sich, nicht berücksichtigt werden. Cai und Hansen [64] entwickelten diese Statistik zur M_2*-Statistik weiter, die eine wietere Zusammenfassung der Marginalhäufigkeiten erster und zweiter Ordnung beinhaltet und daher auch für umfangreiche polytome Fragebögen eine geringe Typ-I-Fehlerrate aufweist. Aufgrund der Komplexität wird für die genaue Formel der M_2*-Statistik, sowie deren Herleitung auf die Originalarbeit von Cai und Hansen verwiesen [64].

Die Auswertung der M_2*-Statistik erfolgt dabei durch den Vergleich mit einer X^2-Verteilung mit $\left(\left[w + \frac{w(w-1)}{2}\right] - q\right)$ Freiheitsgraden, wobei q für die Anzahl von Modellparametern, w für die Anzahl univariater und $\left(\frac{w(w-1)}{2}\right)$ für die Anzahl bivariater Residuen steht [65]. Überträgt man diese Formel auf das Beck Depressions-Inventar-II, entstehen bei dem aus 21 Items mit je vier Antwortkategorien bestehenden Fragebogen somit $w = 21$ univariate und $\left(\frac{w(w-1)}{2}\right) = 210$ bivariate Residuen. Das Partial Credit Modell [26] berücksichtigt insgesamt 64 Modellparameter, die sich aus einem fixierten Steigungsparameter für alle Items, sowie pro Item drei Schwellenparametern zusammensetzt, sodass die M_2*-Statistik bei ausreichendem Modellfit einer X^2-Verteilung mit 167 Freiheitsgraden folgen müsste. Für das Generalized Partial Credit Modell [27] und das Graded Response Modell [28], die pro Item einen eigenen Steigungsparameter sowie ebenfalls pro Item drei Schwellenparameter berücksichtigen und somit 84 Modellparameter aufweisen, müsste die M_2*-Statistik folglich einer X^2-Verteilung mit 147 Freiheitsgraden folgen.
Kommt es hierbei zu einer signifikanten Abweichung von dieser X^2-Verteilung ($p<,05$), weichen die vom Modell vorhergesagten Häufigkeiten signifikant von den tatsächlich in der Stichprobe beobachteten Häufigkeiten ab und es muss somit von einer eingeschränkten Anpassungsgüte des untersuchten Modells ausgegangen werden [65].

Eine weitere Möglichkeit zur Evaluation der Anpassungsgüte eines Modells auf Modell-Ebene ist die Nutzung von GOF-Indices. Diese können, ergänzend zur M_2*-Statistik,

bei polytomen Fragebögen eine Aussage darüber machen, wie gut die in Frage kommenden Modelle den zugrunde liegenden Datensatz beschreiben können [58].
Man unterscheidet hierbei relative Fit-Indices, die die proportionale Verbesserung des Fits bei Vergleich des gewählten Modells (M_T) mit einem restriktiveren genesteten Basismodell (M_0) beschreiben, von absoluten Fit Indices, die - ohne Referenzmodell – anzeigen, wie gut das gewählte Modell die vorliegenden Daten widerspiegelt [58].
Zur Evaluation der Anpassungsgüte eines Modells eignen sich dabei die bereits zur Überprüfung der Faktorenstruktur vorgestellten Indices: der TLI und CFI, die beide zu den relativen Fit-Indices gehören, sowie der SRMR und RMSEA, die zu den absoluten Fit-Indices zählen [58]. Die verwendeten Grenzwerte gelten dabei analog zu den oben dargestellten auch für die Evaluation der Anpassungsgüte eines Modells [22, 59].

Um einen direkten Vergleich zweier in Frage kommender Modelle auch dann durchführen zu können, wenn es sich bei den Modellen um nicht genestete Modelle handelt, bieten sich das Akaike's Information Criterion (AIC) und das Schwarz Bayesian Information Criterion (BIC) an, die ebenfalls zu den absoluten Fit-Indices zählen [62, 66].
Hierbei wird mithilfe einer Signifikanztestung überprüft, ob die Differenz der AIC und BIC-Indices zwischen den zwei Modellen groß genug ist, um die Überlegenheit eines der beiden Modelle ableiten zu können. Umschließt das entsprechende 95%-Konfidenzintervall dabei den Nullpunkt, wird angenommen, dass der Fit der beiden verglichenen Modelle gleichwertig ist. Ist dies nicht der Fall, wird dasjenige Modell gewählt, das den signifikant niedrigeren AIC- bzw. BIC-Index aufweist. Die Herleitung der Formel und Grundlagen zur Berechnung sind ausführlich in der Arbeit von Merkle, You und Preacher dargestellt [66].

Die vorhergehenden theoretischen Ausführungen legten für die praktische Durchführung der Analysen am BDI-II das nachfolgend beschriebene Vorgehen nahe. Aufgrund des polytomen Antwortformates mit 21 Items wurde die M_2*-Statistik zur Evaluation des globalen Modell-Fits gewählt. Ergänzend wurden die absoluten und relativen Fit-Indices TLI, CFI, RMSEA und SRMR für die drei in Frage kommenden Modelle errechnet, um so dasjenige Modell zu finden, das den vorliegenden Datensatz als Gesamtes am präzisesten abbilden kann. Um bei annähernd vergleichbarem Fit zweier Modelle zu überprüfen, ob sich anhand der Fit-Indices die Überlegenheit eines der Modelle ableiten lässt, sollen ΔBIC und ΔAIC zum Einsatz kommen.

Neben der Modell-Ebene kann die Anpassungsgüte auch auf Item-Ebene näher evaluiert werden [23]. Dies ermöglicht es, für jedes Item einzeln zu überprüfen, ob das gewählte Modell das tatsächliche Antwortverhalten in der zugrunde liegenden Stich-

probe korrekt vorhersagen kann [67]. Dieses Verfahren wird als lokaler Goodness-of-Fit bezeichnet [23] und im nachfolgenden Abschnitt näher erläutert.

(4) Goodness-of-Fit auf Item-Ebene

Um den Fit eines Modells auf Item-Ebene zu evaluieren, wird die Kongruenz zwischen der mit dem IRT-Modell vorhergesagten und der tatsächlich beobachteten Häufigkeiten der Antwortkategorien für jedes Item einzeln überprüft [67]. Prinzipiell lässt sich das Vorgehen in fünf Schritte untergliedern:

1. Schätzung der Item- und Personenparameter ($\hat{\theta}_v$) anhand des IRT- Modells,
2. Unterteilung der Probanden anhand von $\hat{\theta}_v$ in Z homogene Gruppen,
3. Evaluation der beobachteten Antworthäufigkeiten in jeder Gruppe,
4. Berechnung der vom Modell vorhergesagten Antworthäufigkeiten für jede einzelne Gruppe und anschließende
5. Berechnung der X^2-basierten Statistiken zum Vergleich der beobachteten und der vom Modell vorhergesagten Antworthäufigkeiten [67, 68].

Orlando und Thissen stellten mit dem $S\text{-}X^2$-Test [67] eine Item-Fit-Statistik für dichotome IRT-Modelle vor, die aufgrund der zuverlässigeren Ergebnisse die traditionellen Item-Fit-Statistiken, wie Yen's Q_1 und McKinley und Mill's G^2, weitestgehend ablöste. Der $S\text{-}X^2$-Test wurde von Kang und Chen [68] zur Nutzung für polytome Items erweitert und lässt sich folgendermaßen berechnen:

$$S - X^2 = \sum_{z=(m_i-1)}^{F-(m_i-1)} \sum_{k=0}^{m_i-1} N_z \frac{(O_{ikz} - E_{ikz})^2}{E_{ikz}} \tag{18}$$

Hierbei stehen O_{ikz} und E_{ikz} für die beobachtete bzw. vom Modell vorhergesagte Häufigkeit von Antwortkategorie k_i in Probandengruppe z, sowie F für den maximalen Test-Score. Die erwarteten Häufigkeiten lassen sich mithilfe des von Thissen, Pommerich, Billeaud und Williams beschriebenen rekursiven Algorithmus berechnen [67–69].

Da für diese Analysen multiple statistische Test durchgeführt werden, sollte eine Korrektur der p-Werte für multiple Tests, bspw. nach Benjamini-Hochberg [70], durchgeführt werden. Als ausreichend guter Fit wird in der Regel gewertet, wenn der p-Wert des $S\text{-}\chi^2$-Tests nach Benjamini-Hochberg-Korrektur größer als 0,01 ist [71].

Zusammenfassend wurde in den beiden vorausgehenden Abschnitten sowohl auf Modell- als auch auf Item-Ebene überprüft, wie gut die drei für die vorliegende Arbeit in Frage kommenden IRT-Modelle den vorliegenden Datensatz abbilden können, oder im Umkehrschluss, wie zuverlässig auf Basis des jeweiligen Modells allgemeingültige Aussagen über das BDI-II, sowie die Depressivität der Probanden abgeleitet werden können.

(5) Goodness-of-Fit auf Personen-Ebene

'Mängel eines Tests hinsichtlich der Modellkonformität können aber nicht nur auf die Items oder die Modelleigenschaften zurückzuführen sein, sondern auch darauf, dass einzelne Personen auf die Testitems nicht in angemessener Weise reagieren' [11, Seite 398], sondern aberrantes Antwortverhalten zeigen. Aberrantes Antwortverhalten bedeutet, dass das gezeigte Antwortmuster unter der gewählten Modell-Annahme nur mit sehr geringer Wahrscheinlichkeit auftreten würde. Liegt bei einem Probanden ein derartiges Antwortmuster vor, kann angenommen werden, dass die Wahl der Antworten durch einen anderen Mechanismus gesteuert wurde, als den vom Modell konstruierten [65]. Problematisch ist dies insofern, als der IRT-Score dieser Probanden folglich keine valide Aussage zur latenten Variablen des Probanden erlaubt [72].

Beim untersuchten Fragebogen könnte zum Beispiel ein wenig motivierter Proband die Antwortkategorien rein zufällig auswählen, um den Fragebogen schneller zu beenden. Der Auswahlmechanismus wäre entsprechend nicht, wie vom Modell angenommen, durch sein Depressions-Level beeinflusst, sondern fußt dabei rein auf dem Zufall, sodass der IRT-Score entsprechend keine valide Aussage zur Depressivität dieses Probanden erlaubt. Als weitere Gründe für aberrantes Antwortverhalten sind neben fehlender Motivation auch mangelnde Sprachkenntnisse, Ablenkung oder Über- bzw. Untertreibung vorstellbar [11, 65].

Eine Möglichkeit aberrante Antwortmuster rechnerisch zu detektieren stellt Levine und Rubin's Likelihood-basierter Index l_0, bzw. dessen Standardisierung in Form von Drasgow's Z_h-Statistik dar, die sich folgendermaßen berechnen lassen:

$$l_0 = P(X_i|\theta_i) = \sum_{i=1}^{n}\sum_{k=1}^{m_i} \delta_k(v_i) log P_{ik}(\theta), \;\; bzw. \;\; Z_h = \left[\frac{l_0(\theta) - E\big(l_0(\theta)\big)}{SD\big(l_0(\theta)\big)}\right] \qquad (19)$$

Hierbei entspricht $\delta_k(v_i)$ einem zufälligen Vektor von Item-Antworten, $SD\big(l_0(\theta)\big)$ der Standardabweichung und $E\big(l_0(\theta)\big)$ dem Mittelwert aller l_0-Werte der Stichprobe [65]. Durch die Nutzung des standardisierten Z_h-Index anstelle des originalen l_0-Index kann das Confounding durch den Theta-Schätzer deutlich reduziert werden [72].
Der in der normalverteilten z-Statistik häufig genutzte Cut-off-Wert von $-1{,}96$ kann nicht ohne Weiteres auf den Z_h-Index übertragen werden, da die Verteilung zwar einer Normalverteilung ähnlich ist, jedoch einer empirischen Verteilung folgt [65, 72]. Als Cut-off zur Detektion von aberrantem Antwortverhalten wurde daher in der vorliegenden Stichprobe die minus-zweifache Standardabweichung von Z_h genutzt [24].

Die Antwortmuster von Personen mit aberrantem Antwortverhalten sollten zwar nicht zur Einschätzung der latenten Variablen verwendet werden, können aber unter

Umständen trotzdem relevante Informationen enthalten [11]. Um bspw. nähere Informationen zu den Gründen aberranten Antwortverhaltens herausarbeiten zu können, erfolgte die graphische Darstellung des Z_h-Index in Abhängigkeit von der latenten Variablen, sowie die Gegenüberstellung der mittleren Itemscores der Probanden mit typischem und der Probanden mit atypischem Antwortverhalten [41].
Da bei Probanden mit aberrantem Antwortverhalten keine valide Einschätzung der latenten Variablen auf Basis des Personenparameters möglich ist [65, 72], wurden diese Probanden für die weiteren Analysen ausgeschlossen. Um den Einfluss aberranten Antwortverhaltens auf die globalen und lokalen GOF-Statistiken eruieren zu können, wurden diese nach Entfernung der betreffenden Probanden erneut errechnet.

Die Evaluation der Anpassungsgüte von IRT-Modellen sollte immer auch die Überprüfung der zwei zentralen Annahmen parametrischer IRT-Modelle – die lokale Unabhängigkeit aller Itempaare und die Monotonie aller Item Response Funktionen – beinhalten [61], die daher im Folgenden näher erläutert werden sollen.

(6) Überprüfung der lokalen Unabhängigkeit

Eine mögliche Ursache für den mangelnden Fit eines unidimensionalen IRT-Modells kann das Vorliegen von lokalen Abhängigkeiten zwischen einzelnen Itempaaren sein. Lokale Abhängigkeit bedeutet hierbei, dass nach Kontrolle der latenten Variablen eine signifikante Korrelation zwischen Items bestehen bleibt, die Items also nicht voneinander unabhängig sind [73–75].
Um diesen Sachverhalt konkret am BDI-II zu verdeutlichen, kann man sich eine Gruppe von Probanden mit exakt derselben Ausprägung der latenten Variablen, also in diesem Fall der depressiven Symptomatik, vorstellen. Betrachtet man sich das Antwortverhalten dieser Probanden auf zwei beliebige Items des BDI-II, die die Depressivität der Probanden messen sollen, müssten bei gleicher Ausprägung dieser theoretisch alle Probanden dieselbe Antwort auf die beiden Items wählen. Praktisch kommt es durch unsystematische Einflüsse in der Regel zu einer symmetrischen Streuung der tatsächlichen Antworten um diese theoretische Antwort [11].
Zeigt sich hier jedoch eine systematische Abweichung der manifesten Antworten in eine Richtung, muss davon ausgegangen werden, dass die Beantwortung der Items neben der beabsichtigten Variable ‚Depressivität' noch durch eine weitere, unbeabsichtigte Variable systematisch beeinflusst wird [11].
Prinzipiell lassen sich hierbei zwei Formen lokaler Abhängigkeiten unterscheiden:

- die ‚underlying local dependence', die die Folge von zusätzlichen, nicht berücksichtigten latenten Variablen und daher eng mit der Dimensionalität des Fragebogens verknüpft ist und

- die ‚surface local dependence', die durch eine Ähnlichkeit mehrerer Items (bezüglich Inhalt oder Position innerhalb des Fragebogens) entsteht [73, 76].

In der Literatur wurden diverse Möglichkeiten beschrieben, um die lokale Abhängigkeit von Items rechnerisch zu evaluieren. Eine der bekanntesten stellt Yen's $Q3$-Statistik [76] dar, die die Korrelation zwischen der Performance in zwei Testitems i und j nach Berücksichtigung der Performance im gesamten Test nutzt. Hierfür wird zunächst die Abweichung (d_{iv}) zwischen der durch das gewählte IRT-Modell vorhergesagten Antwort eines Probanden v auf ein Item i ($E_{iv} = \hat{P}_i\left(\hat{\theta}_v\right)$) – basierend auf dem $\hat{\theta}_v$-Schätzer des Probanden v nach Beantwortung aller Items – mit dem tatsächlich vorliegenden Antwortverhalten (x_{iv}) errechnet:

$$d_{iv} = x_{iv} - E_{iv} \tag{20}$$

Um nun den Index $Q3_{ij}$ von zwei Items i und j zu erhalten, muss die Pearson Product-Moment-Korrelation zwischen den Abweichungen (berechnet mithilfe von Formel 20) von Item i und j berechnet werden [74, 76]:

$$Q3_{ij} = r_{d_i d_j} \tag{21}$$

Zur Bewertung des so entstandenen $Q3$-Index differieren die empfohlenen Richtwerte in der Literatur. Ein häufig verwendeter statischer Cut-off-Wert für Lokale Abhängigkeit liegt bei einem Wert der $Q3$-Statistik über 0,2, wie er von Chen und Thissen [73] empfohlen wurde. Da der Einfluss der Anzahl von Items und Antwortkategorien, sowie der Stichprobengröße in den statischen Cut-off-Werten nicht berücksichtigt wird, wurde von Christensen [77] empfohlen, die $Q3$-Statistik eines Items relativ zur mittleren Item-Residualkorrelation zu sehen. Als hinweisend für das Vorliegen von Lokaler Abhängigkeit eines Itempaars wurde in der vorliegenden Arbeit den Empfehlungen von Christensen [77] folgend daher eine Korrelation von 0,2 über der durchschnittlichen Item-Residualkorrelation gewertet.

(6) Evaluation der Form der Item Response Funktionen (IRFs)

Eine weitere Ursache für eine mangelnde Modell-Anpassung kann sich durch die Anforderung parametrischer IRT-Modelle an die graphische Form der IRFs ergeben. Hierbei unterscheidet man zum einen die Forderung nach Monotonie und zum anderen die nur bei parametrischen Modellen bestehende Forderung nach logistisch oder normal-ogive geformten IRFs [61]. Bevor weiter auf die graphische Überprüfung dieser Kriterien mittels non-parametrischer IRT-Modelle eingegangen wird, sollen diese beiden Anforderungen zunächst näher erläutert werden.

Prinzipiell bedeutet Monotonie in diesem Zusammenhang, dass die Wahrscheinlichkeit, eine höhere Antwortkategorie zu wählen, zunimmt, je stärker die latente Variable

ausgeprägt ist. Die IRFs sind entsprechend nichtfallende Funktionen der latenten Variablen θ [22].
Konkret auf den BDI-II bezogen bedeutet eine Verletzung der Monotonie eines Items, dass die Wahrscheinlichkeit, eine höhere Antwortkategorie bei diesem Item zu wählen – trotz steigendem Depressionslevel des Probanden – dennoch fällt. Wäre dies der Fall, könnte somit nicht automatisch geschlossen werden, dass ein Proband, der in diesem Item eine höhere Kategorie gewählt hat, auch depressiver ist, als ein Proband, der eine niedrigere Kategorie gewählt hat. Entsprechend wäre dieses Item in der vorliegenden Form nicht, oder nur eingeschränkt dafür geeignet, die Depressionsschwere eines Probanden einzuschätzen, sodass die Monotonie der Items eine zentrale Grundlage für jede Form der IRT-Analyse darstellt [22].

Die zweite Anforderung beschäftigt sich mit der Form der IRFs. Der Unterschied zwischen parametrischen und non-parametrischen IRT-Modellen definiert sich dabei grundlegend durch die Form der Beziehung zwischen der Auswahlwahrscheinlichkeit einer Item-Antwort – $P(\theta_v)$ – und der Ausprägung der latenten Variablen des Probanden (θ_v) [61].
Während die non-parametrischen Modelle keine speziellen Anforderungen an diese Beziehung stellen und die Form der IRF folglich eine von den Daten getriggerte Repräsentation dieser Beziehung ist [78], wird bei den parametrischen Modellen für diese Beziehung a priori eine vorgegebene – in der Regel logistische – Form angenommen [49]. Stimmt die tatsächlich vorliegende Form dieser Beziehung nicht mit der bei parametrischen IRT-Modellen angenommenen, logistischen Form überein, führt dies zu einer mangelnden Anpassungsgüte des parametrischen Modells. Dies führt dazu, dass Item- und Personenparameter, die mithilfe des Modells errechnet wurden, keine validen Aussagen zu den Items auf der einen und der Einschätzung der latenten Variablen der Probanden auf der anderen Seite erlauben würden [79].

Um nun Items identifizieren zu können, für die sich die genannten Vorgaben an die Form der IRFs parametrischer IRT-Modelle als zu unflexibel erweisen, stellt die graphische Evaluation der non-parametrischen IRFs eine Möglichkeit dar [61]. Diese non-parametrischen IRFs können hierfür mittels Kernel Smoothing Regression geschätzt werden, auf die daher im Weiteren näher eingegangen werden soll [49].

Die Kernel Smoothing Regression basiert auf der Generierung lokaler, gewichteter Durchschnittswerte, die die Mittelwerte der Antwortvariablen in der Nähe eines festgelegten Punktes als Referenzwert nutzen, um hieraus sinnvolle Schätzungen für die Regressionskurve zu generieren [79]. Die Kernel-Schätzungen der IRFs an den Evaluationspunkten q werden hierbei durch nachfolgende Formel dargestellt [78]:

$$P_{ik}(\theta_q) = \sum_{v=1}^{N} \omega_{vq} X_{vik} = \sum_{v=1}^{N} \frac{K\left(\frac{\theta_v - \theta_q}{h}\right)}{\sum_{r=1}^{N} K\left(\frac{\theta_r - \theta_q}{h}\right)} \; X_{vik}, \tag{22}$$

Hierbei entspricht X_{vik} der Wahrscheinlichkeit, dass Person v bei Item i Kategorie k wählt. K steht für die Kernel-Funktion, eine nichtnegative, kontinuierliche symmetrische Funktion, die die Form der Verteilung der Kernel-Gewichte definiert und h für den Smoothingparameter, der die Gewichtung definiert und somit den Grad der Glättung der Kurve beschreibt [49, 78].

Um Vergleiche zwischen parametrischen und non-parametrischen IRT-Modellen zu ermöglichen, müssen beide Modelle auf die gleiche Verteilung zurückgreifen [61]. Da die in Frage kommenden parametrischen Modelle eine Gauß-Verteilung $(K(x)=e(-x^2/2))$ annehmen, wurde diese auch als Grundlage der non-parametrischen IRT-Berechnungen gewählt. Bei Verwendung der Gauß-Verteilung lässt sich der optimale Smoothingparameter anhand der Silverman Formel

$$h = 1{,}06\sigma_\theta n^{-1/5} \tag{23}$$

berechnen, wobei σ_θ für die Standardabweichung der latenten Variablen steht [49].

Da es sich bei den IRFs der non-parametrischen Modelle um eine direkt von den Daten getriggerte Repräsentation der Beziehung zwischen der Auswahlwahrscheinlichkeit einer Item-Antwort und der Ausprägung der latenten Variablen handelt [78], kann direkt von der IRF des Items abgelesen werden, ob das Kriterium der Monotonie erfüllt wird. Diese liegt vor, wenn die non-parametrische IRF bei ansteigender latenter Variable an keiner Stelle fällt.

Durch den optischen Vergleich der Item Response Funktionen des parametrischen und non-parametrischen Modells kann der Einfluss der Abweichung von der geforderten IRF-Form parametrischer Modelle auf den Item-Misfit abgeschätzt werden [61]. Um parametrische und non-parametrische IRFs leichter vergleichen zu können, bietet sich die Berechnung von punktweisen Konfidenzintervallen an, die Aussagen darüber ermöglichen, wie zuverlässig die IRFs in den verschiedenen Bereichen der latenten Variablen definiert sind [61]. Für die Möglichkeit der Berechnung dieser Konfidenzintervalle wird auf die Arbeit von Ramsay verwiesen [80].

Die bisherigen Analysen dienten einem gemeinsamen Ziel: ein IRT-Modell zu finden, das den vorliegenden Datensatz adäquat abbilden kann.
Dies ist die grundlegende Voraussetzung dafür, im Folgenden anhand des gewählten IRT-Modells allgemeingültige Aussagen über das BDI-II und die Depressivität der Probanden ableiten zu können und soll nun schrittweise erarbeitet werden.

3.4.3 IRT-Analyse

Nach der Festlegung auf ein geeignetes IRT-Modell kann mit der Auswertung der Ergebnisse der IRT-Analyse begonnen werden. Zur Bestimmung der Itemparameter wurde das MML-Schätzverfahren – unter Verwendung des von Bock und Aitkin entwickelten EM-Algorithmus [30] – gewählt.

Die in Frage kommenden parametrischen Modelle beinhalten für jedes Item $m_i - 1$ Schwellenparameter [14] – für den BDI-II also entsprechend drei ($\delta_{i_{1-3}}$). Der Mittelwert dieser Schwellenparameter, auch bezeichnet als Locationparameter (LP), dient in der IRT dabei stichprobenunabhängig als Maß für die Itemschwierigkeit [15].

Während das PCM einen einheitlichen Steigungs- bzw. Diskriminationsparameter α für alle Items annimmt, differenzieren das GPCM und das GRM pro Item einen separaten Diskriminationsparameter α_i [14]. Je höher dieser ausfällt, desto besser kann ein Item zwischen Probanden mit hoher und Probanden mit niedriger Ausprägung der latenten Variablen diskriminieren [15]. Nach Baker [81] wird die Diskriminiationsfähigkeit eines Items bei einem α zwischen 0,01 und 0,34 hierbei als sehr niedrig, zwischen 0,35 und 0,64 als niedrig, zwischen 0,65 und 1,34 als moderat, zwischen 1,35 und 1,69 als hoch und über 1,7 als sehr hoch bewertet.

Anhand dieser Itemparameter lässt sich für jeden Probanden itemunabhängig ein Personenparameterschätzer $\hat{\theta}_v$, auch bezeichnet als Traitscore, bestimmen, der gegenüber dem klassischen Summenscore den Vorteil bietet, dass durch die Gewichtung mittels der Itemparameter hier das individuelle Antwortmuster mit berücksichtigt wird [6, 15]. Wird das BDI-II in der Praxis mittels IRT ausgewertet, wird die Depressionsschwere eines Probanden – analog zum klassischen Summenscore – anhand des Personenparameterschätzers $\hat{\theta}_v$ eingeschätzt. Wie bereits in Abschnitt 1.3.3 ausführlich erläutert, stehen zur Bestimmung hierfür verschiedene Schätzverfahren zur Verfügung. Aufgrund der Limitation der ML-basierten Verfahren in Bezug auf Extremwerte wurde die Entscheidung zur Verwendung eines Bayes'schen Schätzers getroffen. Abhängig von den Ergebnissen der Dimensionalitätsanalyse soll hierfür bei Verwendung eines unidimensionalen IRT-Modells der EAP-, sowie bei Verwendung eines multidimensionalen Modells der MAP-Schätzer angewendet werden [82].

Die Reliabilität bzw. Präzision, mit der diese modellbasierten Personenparameterschätzer $\hat{\theta}_v$ bestimmt werden können, stellt ein zentrales Gütekriterium eines Tests dar [11]. Anders als in der klassischen Testtheorie kann hierbei für jede Merkmalsausprägung ein individueller Standardmessfehler errechnet werden, der von der Testinformations-

funktion $T(\theta_v)$, die der Summe aller Iteminformationsfunktionen $I_i(\theta_v)$ eines Tests entspricht, abhängig ist:

$$SEE(\theta) = \sqrt{T(\theta)^{-1}} = \sqrt{\sum_{i=1}^{n} I_i(\theta)^{-1}} \quad (24)$$

Graphisch lassen sich die Iteminformationsfunktionen eines Tests mithilfe von Iteminformationskurven darstellen [25] und sollen für alle 21 BDI-II Items errechnet werden. Die Testinformationsfunktion kann mittels Testinformationskurve graphsich umgesetzt werden. Um Vergleiche mit der klassischen Reliabilität in der KTT zu ermöglichen, kann der Standardmessfehler und damit automatisch auch die Testinformation mittels folgender Formel in die klassische Reliabilität umgerechnet werden [83]:

$$Rel = 1 - SEE(\theta)^2 = 1 - T(\theta)^{-1} \quad (25)$$

Die Testinformationsfunktion ermöglicht eine Einschätzung, in welchen Bereichen der latenten Variable das Instrument eine besonders hohe Aussagekraft aufweist und somit gut zwischen Abstufungen der Merkmalsausprägung diskriminieren kann, und in welchen Bereichen das Instrument Schwächen aufweist. Die gleiche Aussage lässt sich anhand der Iteminformationsfunktionen auch für jedes einzelne Item ableiten.

Für das BDI-II klinisch relevant ist die Item- bzw. Testinformationsfunktion insbesondere in den Grenzbereichen behandlungsbedürftiger Depressivität. Basierend auf der klassischen Testtheorie wird häufig der in Abschnitt 3.2 beschriebene Bewertungsmaßstab genutzt [2, 19]: als klinisch unauffällig gelten Probanden mit einem Summenscore zwischen 0 und 13 Punkten. Ein Summenscore zwischen 14 und 19 weist auf ein mildes, zwischen 20 und 29 auf ein moderates und zwischen 30 und 63 auf ein schweres depressives Syndrom hin.
Für die Personenparameterschätzer $\hat{\theta}_v$ existieren bisher keine derartigen klinisch validierten Cut-off-Werte. Um dennoch abschätzen zu können, welche Items in den Übergangsbereichen eine besonders hohe Aussagekraft aufweisen, wurden die gerade genannten, klinisch validierten Cut-off-Scores der klassischen Testtheorie mittels linearer Transformation in IRT-Traitscores konvertiert [84] und als Anhaltspunkte für die Einschätzung der Iteminformation in den Übergangsbereichen genutzt.

Mit Hilfe der IRT kann zudem überprüft werden, ob ein Instrument die latente Variable für alle Probanden gleichartig misst, unabhängig von verschiedenen demographischen Eigenschaften, wie beispielsweise dem Alter oder dem Geschlecht der Probanden [85]. Sie kann für den BDI-II folglich die Frage beantworten, ob es zulässig ist, die depressive Symptomatik von Männern und Frauen anhand ihres BDI-II-Testscores – auf die IRT bezogen anhand ihres Personenparameterschätzers $\hat{\theta}_v$ – zu vergleichen.

Allgemeiner formuliert ermöglicht die IRT somit die Überprüfung, ob es zulässig ist, die latente Variable zwischen Subgruppen, die durch demographische Eigenschaften definiert werden, anhand ihres Traitscores zu vergleichen [7]. Diese Vergleiche sind allerdings nur dann zulässig, wenn die Wahrscheinlichkeit einer Item-Antwort nach Kontrolle der latenten Variablen keinen signifikanten Unterschied zwischen den Subgruppen aufweist, die Itemparameter also folglich zwischen den Subgruppen konstant sind [7, 22]. Wird diese Annahme verletzt, spricht man von Differential Item Functioning (DIF). Prinzipiell unterscheidet man hierbei uniformes von non-uniformen DIF [85]. Uniformes DIF liegt vor, wenn die Wahrscheinlichkeit über die komplette Bandbreite der latenten Variable für eine Subgruppe konstant höher bzw. niedriger ist, sich also bei gleichem Steigungsparameter α_i nur die Schwellenparameter δ_{ik} zwischen den Subgruppen unterscheiden. Non-uniformes DIF dagegen liegt vor, wenn sich die Wahrscheinlichkeit zwischen den Subgruppen, je nach Bereich der latenten Variablen, in dem der Vergleich stattfindet, unterscheidet, entsprechend also Unterschiede im Steigungsparameter α_i zwischen den Subgruppen bestehen [7, 85, 86].

Zur rechnerischen Detektion von DIF existieren multiple Ansätze. Einen der bekanntesten stellt hierbei die ordinal logistische Regression dar, die die Wahrscheinlichkeit, eine bestimmte Antwortkategorie zu wählen, als Funktion des Testscores (Summenscore oder Personenparameterschätzer), der Gruppenzugehörigkeit und der Interaktion zwischen Gruppenzugehörigkeit und Testscore vorhersagt [85–87].

Hierfür werden für jedes einzelne Item neben einem Null-Modell drei weitere genestete, hierarchische Modelle formuliert. Modell 1 berücksichtigt im Vergleich zum Null-Modell zusätzlich einen Term für den Trait-Score des Probanden $(\delta_{i1} * \theta_v)$, Modell 2 zusätzlich zu Modell 1 einen Term für die Gruppenzugehörigkeit $(\delta_{i2} * g)$ und Modell 3 zusätzlich zu Modell 2 einen Term für die Interaktion zwischen Gruppenzugehörigkeit und Trait-Score $(\delta_{i3} * (\theta_l * g))$ [7, 85]:

$$Modell\ 0: logit\ (p_{ik}|\theta_v, g) = log\left(\frac{p_{ik}}{1-p_{ik}}|\theta_v, g\right) = \alpha_{ik} \quad (26)$$

$$Modell\ 1: logit\ (p_{ik}|\theta_v, g) = log\left(\frac{p_{ik}}{1-p_{ik}}|\theta_v, g\right) = \alpha_{ik} + \delta_{i1} * \theta_v \quad (27)$$

$$Modell\ 2: logit\ (p_{ik}|\theta_v, g) = log\left(\frac{p_{ik}}{1-p_{ik}}|\theta_l, g\right) = \alpha_{ik} + \delta_{i1} * \theta_v + \delta_{i2} * g \quad (28)$$

$$Modell\ 3: logit\ (p_{ik}|\theta_v, g) = log\left(\frac{p_{ik}}{1-p_{ik}}|\theta_l, g\right) = \alpha_{ik} + \delta_{i1} * \theta_v + \delta_{i2} * g + \delta_{i3} * (\theta_v * g) \quad (29)$$

Das Grundprinzip der ordinal logistischen Regression basiert nun auf dem Vergleich dieser logistischen Modelle mittels Likelihood-Ratio-X^2-Test [87]:

$$LR = -2\ln(\Delta LR) = -2\ln L_{Modell\ 1} - (-2\ln L_{Modell\ 2}) \quad (30)$$

Die Teststatistik sollte, wenn keine signifikante Abweichung zwischen den Modellen besteht, einer X^2-Verteilung mit Freiheitsgraden gleich der Differenz von Parameterschätzern zwischen den beiden Modellen folgen [86]. Da multiple Vergleiche durchgeführt werden müssen, die bei Verwendung des Signifikanzniveaus $\alpha=0,05$ zu Verzerrungen führen können, sollte für das Signifikanzniveau α eine Korrektur nach Bonferroni durchgeführt werden [85]. Bei 21 Items im BDI-II wurde daher ein Bonferroni-korrigiertes Signifikanzniveau von $\alpha = \frac{0,05}{21} = 0,0024$ verwendet.

Soll nun überprüft werden, ob ein Item Hinweise auf DIF bezüglich einer Variablen aufweist, werden die Modelle 1 und 3 mithilfe des Likelihood-Ratio-Tests miteinander verglichen. Die Teststatistik sollte, sofern kein DIF vorliegt, einer X^2-Verteilung mit zwei Freiheitsgraden folgen, da Modell 3 zwei Parameter mehr berücksichtigt als Modell 1. Ein signifikantes Ergebnis deutet auf das Vorliegen von DIF hin und sollte mittels weiterer Tests näher auf die Form des DIFs hin untersucht werden. Uniformes DIF manifestiert sich hierbei im Vergleich zwischen Modell 1 und 2, wohingegen sich non-uniformes DIF im Vergleich von Modell 2 und 3 zeigt. Für beide Vergleiche wird als Bewertungsmaßstab die X^2-Verteilung mit einem Freiheitsgrad genutzt. Um falsch-positive und falsch-negative Ergebnisse der DIF-Analyse aufzudecken, wurde eine iterative Purifikation durchgeführt. Hierfür werden die LR-Analysen mittels erneuerter IRT-Schätzer so lange wiederholt, bis in zwei aufeinander folgenden Analysen dieselben Items als DIF-Items identifiziert werden [7, 52, 85].

Um das Ausmaß des detektierten DIFs zu quantifizieren, werden verschiedene Effektsize-Maße verwendet. Eines der bekanntesten stellt der Leitfaden von Jodoin und Gierl [88] dar, der ein DIF mit einem pseudo-R^2_Δ von kleiner $0,035$ als vernachlässigbar, zwischen $0,035$ und $0,07$ als moderat und größer als $0,07$ als massiv einstufte [89]. Eine weitere Empfehlung, der auch das PROMIS-Netzwerk folgt, nennt als Kriterium für signifikantes DIF dagegen einen Unterschied von $0,02$ in McFaddens R^2 [7, 22].

In der vorliegenden Arbeit wurde eine Analyse auf Differential Item Functioning bezüglich Alter und Geschlecht der Probanden durchgeführt. Während das Geschlecht bereits in dichotomer Form vorliegt, musste das Alter der Probanden zunächst dichotomisiert werden. Als Cut-off-Wert zwischen den beiden Subgruppen wurde ein Alter von 30 Jahren gewählt und die Probanden somit in eine Subgruppe ‘jüngerer Probanden’ (<30 Jahre) und eine Subgruppe ‘älterer Probanden’ (≥30 Jahre) unterteilt.

4. Ergebnisse

4.1 Klassische Item Analyse

In Anhang 1 sind die Ergebnisse der klassischen Item Analyse dargestellt. Die absolute und prozentuale Verteilung der Antworten ergab, dass Kategorie 3 von Item 18 (‘Appetitveränderungen’ – ‘Ich habe überhaupt keinen Appetit’ bzw. ‘Ich habe ständig großen Hunger’) die am seltensten und Kategorie 1 von Item 9 (‘Suizidalität’ – ‘Ich denke nie daran, mich umzubringen’) die am häufigsten gewählte war. Die große Spannbreite der beobachteten Testscores (0 bis 58) verdeutlicht das breite Spektrum der depressiven Symptomatik in der untersuchten Stichprobe, das Histogramm in Abbildung 3 (links) die Verteilung mit einem klaren Maximum im Bereich niedriger Depressivität.

Die Inter-Item-Korrelationen, die aufgrund mangelnder Übersichtlichkeit bei 21 Items nicht einzeln dargestellt werden können, lagen zwischen 0,14 bei Itempaar 16/21 (‘Appetitveränderungen’/ ‘Suizidalität’) und 0,65 bei Itempaar 7/14 (‘Selbstablehnung’/ ‘Wertlosigkeit’). Der Mittelwert aller Inter-Item-Korrelationen betrug 0,36.

Um die Reliabilität überprüfen zu können, wurde wie in Abschnitt 3.4.1 beschrieben, zunächst die Annahme der essentiellen τ-Äquivalenz überprüft, die aufgrund des signifikanten Ergebnisses der robusten F-Statistik (F=5,742, p<0,01) jedoch abgelehnt werden muss. Die Skewness (Schiefe) von 1,39 (p<0,01) spricht für das Vorliegen einer rechtsschieferen und die Kurtosis (Wölbung) von 2,26 (p<0,01) für das Vorliegen einer steileren Verteilung der Testscores in der vorliegenden Stichprobe, als dies bei einer Normalverteilung anzunehmen wäre. Dies bestätigt sich auch bei der optischen Auswertung von Histogramm und QQ-Plot (Abbildung 3). Zur Einschätzung des Ausmaßes der Abweichung von der Normalverteilung wurde im Anschluss der Shapiro-Wilk-Test berechnet, der aufgrund des Ergebnisses (W=0,89, p<0,01) eine signifikante Abweichung von der Normalverteilung nahelegt.

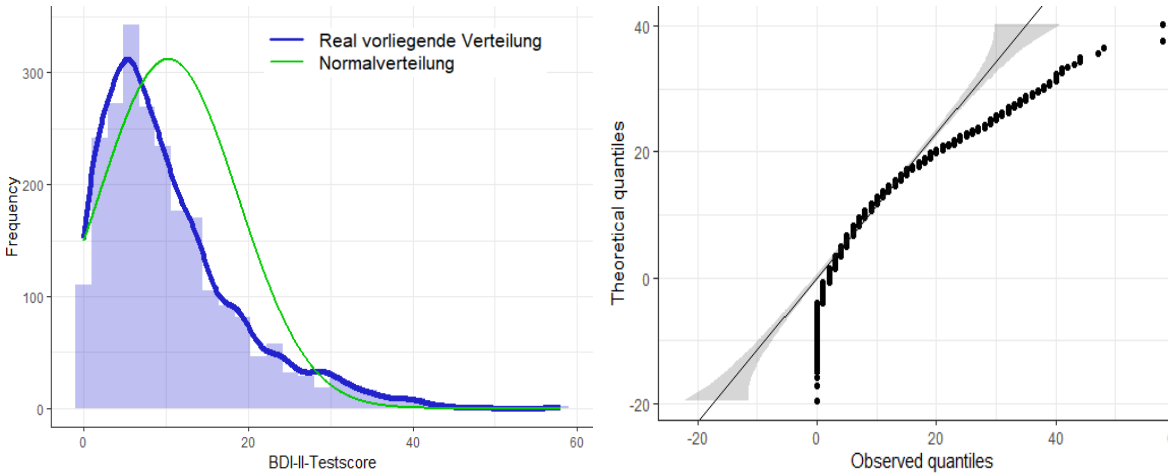


Abbildung 3: Test auf Normalverteilung der Daten

Bei nicht essentiell τ-äquivalenten und nicht-normalverteilten Daten könnte die Reliabilität bei Verwendung von Cronbach α tendenziell unterschätzt werden, sodass zusätzlich zu Cronbach α die Reliabilitätskoeffizienten Omega ω_t und GLB errechnet wurden, die insgesamt eine gute Übereinstimmung mit Cronbach α zeigten (Tabelle 3).

Tabelle 3: Reliabilitätskoeffizienten

Cronbach α	Omega ω_t	Greatest Lower Bound (GLB)
0,91 [0,91; 0,92]	0,91 [0,91; 0,92]	0,91

4.2 IRT-Modellauswahl

4.2.1 Evaluation der Dimensionalität des BDI-II

Die Überprüfung der Faktorstruktur des BDI-II mittels konfirmatorischer Item Faktor Analyse erfolgte für die sechs in Tabelle 2 (Seite 24) dargestellten Strukturen, sowie für die fünf korrespondierenden Bifaktor-Modelle. Insgesamt zeigten die fünf Bifaktor-Modelle gegenüber den klassischen Faktormodellen einen besseren Fit für den vorliegenden Datensatz (Tabelle 4).

Zusammenfassend den besten Fit zeigte hierbei das Bifaktor-Modell mit den drei Gruppenfaktoren ‚kognitiv', ‚somatisch' und ‚affektiv' nach Beck et al. [2].

Tabelle 4: Ergebnisse der Item Faktor Analyse

Modell	SRMR	RMSEA	TLI	CFI
Faktor-I_{Steer}	0,051	0,044	0,980	0,982
Faktoren-II_{Beck}	0,226	0,074	0,942	0,949
Faktoren-II_{Dozois}	0,226	0,072	0,945	0,952
Faktoren-$II_{Gorenstein}$	0,220	0,074	0,943	0,950
Faktoren-III_{Osman}	0,250	0,071	0,947	0,955
Faktoren-III_{Beck}	0,280	0,132	0,818	0,844
Bifaktor-II_{Beck}	0,037	0,036	0,986	0,990
Bifaktor-II_{Dozois}	0,035	0,035	0,987	0,991
Bifaktor-$II_{Gorenstein}$	**0,033**	0,035	0,987	0,990
Bifaktor-III_{Osman}	0,035	0,035	0,987	0,990
Bifaktor-III_{Beck}	0,036	**0,032**	**0,989**	**0,992**

4.2.2 Festlegung der Dimensionalität der geplanten IRT-Analyse

Um die Frage zu klären, ob die in Abschnitt 4.2.1 nachgewiesene Bifaktorstruktur des BDI-II im Folgenden die Verwendung eines unidimensionalen IRT-Modells erlaubt, oder ob sie die Verwendung der komplexeren multidimensionalen IRT-Modelle erfordert, erfolgte als nächster Schritt die schrittweise Prüfung der in Abschnitt 3.4.2 beschriebenen Kriterien für essentielle Unidimensionalität. Die Ergebnisse wurden kurz zusammengefasst in Tabelle 5 dargestellt. Bei nachgewiesener essentieller Unidimensionalität wurden im Folgenden die weiteren Analysen den Empfehlungen des PROMIS-Netzwerks [52] folgend anhand unidimensionaler IRT-Modelle durchgeführt.

Tabelle 5: Evaluation der Kriterien für essentielle Unidimensionalität

Kriterium für essentielle Unidimensionalität	*Ergebnis*	*Erfüllt?*
Hoher Eigenwert des ersten Faktors, der mindestens 20% der Variabilität erklärt	*1. Eigenwert: 8,79* *Variabilität: 41,9%*	*Ja*
Quotient von erstem zu zweitem Eigenwert >4	***3,82***	***Nein***
Coeffizient omega H >0,8	*Omega H: 0,81*	*Ja*
ECV >0,6	*ECV: 0,68*	*Ja*
Faktor-Ladungen auf den Generalfaktor >0,3	*alle ≥0,4*	*Ja*
Unidimensionales Modell: *TLI >0,95/ CFI >0,95/ RMSEA <0,06/ SRMR <0,08*	*0,980/ 0,982/ 0,044/ 0,051*	*Ja*

4.2.3 Goodness-of-Fit auf Modell-Ebene

Um eine fundierte Wahl für eines der drei in Frage kommenden unidimensionalen Modelle treffen zu können, wurde zunächst die Limited-Information-GOF-Statistik M_2* nach Cai und Hansen [64], die absoluten Fit-Indices RMSEA, SRMR, AIC und BIC, sowie die relativen Fit-Indices TLI und CFI [58] für jedes der in Frage kommenden Modelle berechnet und die Ergebnisse in Tabelle 6 zusammengefasst.

Tabelle 6: Limited GOF-Statistik M_2* und ausgewählte GOF-Indices

	M_2*	df	p	RMSEA	SRMR	AIC	BIC	TLI	CFI
PCM	1662,5	167	<0,01	0,061	0,081	73350,7	73721,4	0,961	0,962
GPCM	954,2	147	<0,01	0,048	0,058	**72848,0**	**73334,4**	0,976	0,979
GRM	**836,1**	147	<0,01	**0,044**	**0,051**	72862,0	73348,4	**0,980**	**0,982**
$GRM_{Personfit}$	686,7	147	<0,01	0,040	0,046	67127,2	67610,3	0,984	0,986

Insgesamt zeichnete sich ab, dass die Anpassungsgüte beim Partial Credit Modell durch die Restriktion der Diskriminationsparameter auf einen konstanten Wert deutlich schlechter ist, als bei den beiden weniger restriktiven Modellen GPCM und GRM, sodass das PCM aus den weiteren Überlegungen ausgeschlossen wurde.

Um zu überprüfen, ob die Differenz der AIC- und BIC-Werte zwischen GPCM und GRM groß genug ist, um die Überlegenheit des GPCM ableiten zu können, wurde eine Signifikanztestung mit einem α-Level von 0,05 durchgeführt. Da die 95%-Konfidenzintervalle der Differenz zwischen GRM und GPCM mit

$$-74{,}5 < \Delta AIC < 102{,}5$$
$$-74{,}5 < \Delta BIC < 102{,}5$$

den Nullpunkt umschließen, muss allerdings die Gleichwertigkeit der beiden Modelle in Bezug auf die Anpassungsgüte an den vorliegenden Datensatz angenommen werden.

Insgesamt zeigen sowohl das GRM, als auch das GPCM eine gute Anpassung an den vorliegenden Datensatz. Das signifikante Ergebnis der M_2*-Statistik lässt jedoch eine

gewisse Abweichung der Daten von den Modellannahmen erkennen, die im Weiteren auf Item-Ebene näher untersucht wurde.

4.2.4 Goodness-of-Fit auf Item-Ebene

Zur Evaluation des GOF auf Item-Ebene erfolgte für das GPCM und das GRM die Berechnung des $S\text{-}X^2$-Tests nach Kang und Chen [67, 68]. Die Ergebnisse mit entsprechender Signifikanztestung und Benjamini-Hochberg-Korrektur (*cor.p*) sind in Tabelle 7 (Block 1 und 2) dargestellt. Im GPCM zeigten die Items ‘Selbstablehnung’, ‘Selbstvorwürfe’ und ‘Unruhe’ eine unzureichende Modellanpassung, im GRM die Items ‘Selbstvorwürfe’, ‘Unruhe’, ‘Entscheidungsunfähigkeit’ und ‘Ermüdung’.

Tabelle 7: Ergebnisse des $S\text{-}X^2$-Test nach Kang und Chen für das GPCM und das GRM

	GPCM			GRM			$GRM_{Personfit}$		
	$S\text{-}X^2$	**df**	**cor.p**	**$S\text{-}X^2$**	**df**	**cor.p**	**$S\text{-}X^2$**	**df**	**cor.p**
BDI1	41,01	54	0,90	37,14	54	0,96	43,48	47	0,72
BDI2	64,44	61	0,48	70,12	63	0,35	64,13	58	0,38
BDI3	70,51	73	0,65	77,82	72	0,37	73,43	63	0,30
BDI4	61,68	58	0,48	61,35	56	0,37	61,73	51	0,28
BDI5	87,91	79	0,40	105,73	82	0,08	89,84	70	0,12
BDI6	75,60	72	0,48	87,57	75	0,27	59,74	62	0,69
BDI7	127,76	81	**<0,01***	121,28	83	0,02	86,30	77	0,33
BDI8	135,13	76	**<0,01***	127,29	72	**<0,01***	106,56	59	**<0,01***
BDI9	41,36	34	0,34	42,12	35	0,28	35,54	33	0,46
BDI10	81,30	88	0,75	92,15	96	0,66	81,50	91	0,77
BDI11	153,61	82	**<0,01***	169,63	90	**<0,01***	132,77	72	**<0,01***
BDI12	67,04	68	0,63	74,31	64	0,28	65,24	57	0,33
BDI13	132,44	92	0,02	150,29	98	**<0,01***	123,07	89	0,03
BDI14	83,35	64	0,11	60,54	63	0,66	46,21	54	0,77
BDI15	87,44	58	0,02	87,04	59	0,03	76,57	49	0,03
BDI16	51,56	63	0,89	52,86	61	0,80	49,00	57	0,77
BDI17	111,17	88	0,11	128,75	94	0,03	104,52	78	0,07
BDI18	95,28	63	0,02	96,88	66	0,03	91,54	62	0,03
BDI19	91,89	62	0,02	81,76	61	0,08	73,97	55	0,10
BDI20	112,67	82	0,04	135,54	87	**<0,01***	101,97	76	0,07
BDI21	85,27	81	0,48	99,39	78	0,10	101,29	70	0,03

** p-Wert signifikant nach Korrektur nach Benjamini-Hochberg für multiple statistische Tests*

Zusammenfassend zeigt sich keines der beiden überprüften Modelle dem anderen gegenüber eindeutig überlegen. Für die weiteren Analysen wurde daher aufgrund der besseren Interpretierbarkeit der Ergebnisse und den Empfehlungen des PROMIS-Netzwerks [52] folgend fortan das GRM genutzt.

4.2.5 Goodness-of-Fit auf Personen-Ebene

Um ein mögliches Problem bei der Anpassung von IRT-Modellen, nämlich aberrantes Antwortverhalten einzelner Probanden, aufzudecken, wurde als nächstes der Person-

Fit mittels Drasgow's Z_h-Index [72] errechnet. Tabelle 8 stellt die Charakteristika und Abbildung 4 die Häufigkeitsverteilung der Z_h-Statistik dar. Als hinweisend für ein atypisches Antwortmuster wurde ein Z_h-Index kleiner als die minus-zweifache Standardabweichung des Z_h-Index gewertet ($Z_h < -2{,}04$) [24]. Insgesamt zeigten somit 93 der insgesamt 2419 Probanden aberrantes Antwortverhalten.

Tabelle 8: Charakteristika der Z_h-Verteilung

Mittelwert	Standardabweichung	Median	Minimum	Maximum	Skewness	Kurtosis
0,14	1,02	0,37	-5,84	2,53	-1,57	3,94

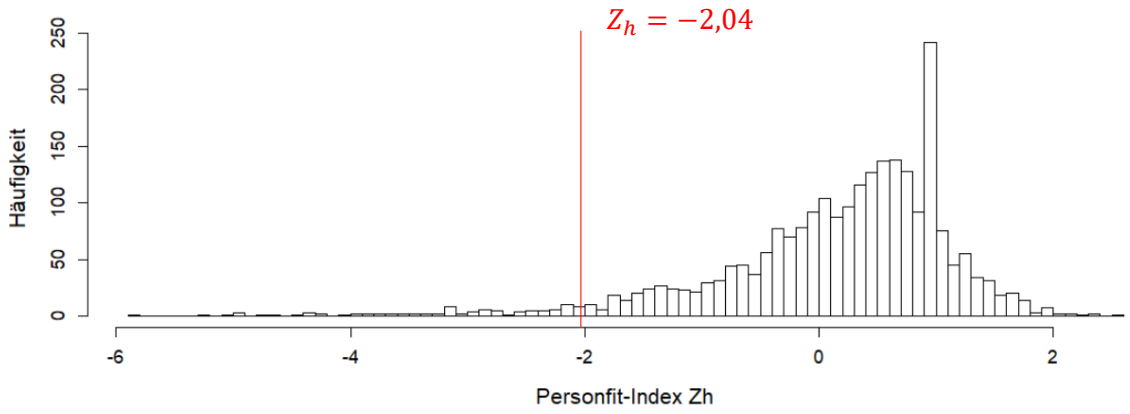


Abbildung 4: Häufigkeitsverteilung der Z_h-Statistik

Die Verteilung der Z_h-Indices in Abhängigkeit von der latenten Variablen θ ist graphisch in Abbildung 5 dargestellt. Sie deutet auf eine inverse Beziehung hin, die zwischen dem Z_h-Index, der einen Indikator für aberrantes Antwortverhalten darstellt, und dem Trait-Score des Probanden, der auf die Anzahl bzw. Schwere der depressiven Symptomatik zurückgeht, besteht.

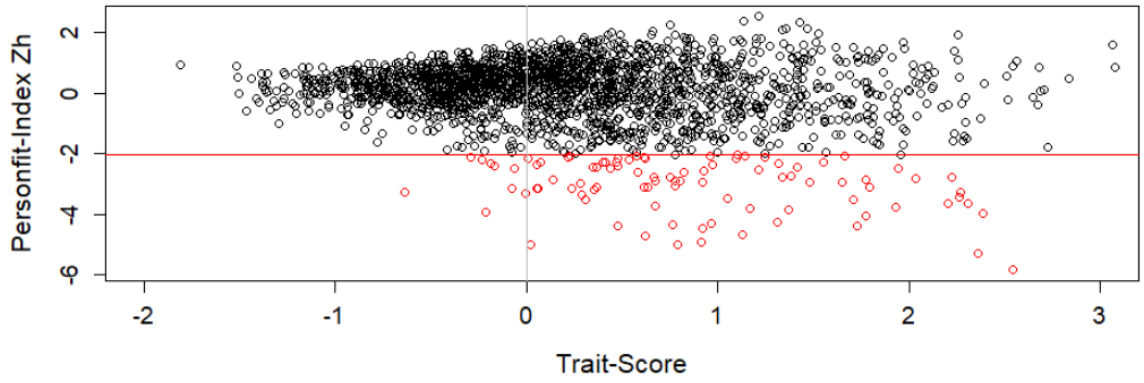


Abbildung 5: Z_h-Index abhängig von der latenten Variablen

Der Vergleich der Merkmalsverteilung in der Gruppe mit aberrantem im Vergleich zur Gruppe mit typischem Antwortverhalten ist in Tabelle 9 dargestellt. Hier zeigt sich kein signifikanter Unterschied bezüglich der Merkmale Alter und Geschlecht, während Probanden mit aberrantem Antwortverhalten signifikant höhere BDI-II-Trait-Scores aufwiesen, als die Probanden mit typischem Antwortverhalten.

Tabelle 9: Merkmalsverteilung zwischen Probanden mit typischem und atypischem Antwortmuster

Merkmal	Atyp. Antwortmuster	Typ. Antwortmuster	Teststatistik	p-Wert
Weibl. Geschlecht	68,82%	68,19%	X^2 = 0,01	0,92
Alter	*mean* = 22,66	*mean* = 22,59	t = 0,17	0,87
BDI-II-Trait-Score	*mean* = 0,87	*mean* = -0,03	t = 11,56	**<0,01***

Als Teststatistik wurde für das Geschlecht der Pearson X^2-Test angewendet, für die Variablen Alter und Trait-Score der Welch-t-Test

Die mittleren Item-Scores der Probanden mit typischen und der Probanden mit atypischem Antwortverhalten sind in Abbildung 6 gegenübergestellt. Aufgrund der gerade beschriebenen inversen Beziehung wurde eine Zentrierung der mittleren Item-Scores der Probanden mit atypischem Antwortmuster auf die Gruppe mit typischem Antwortverhalten durchgeführt, um so die Differenzen in den Antwortmustern deutlicher hervorzuheben [41]. Zusammenfassend lässt sich feststellen, dass Probanden mit aberrantem Antwortmuster in Relation höhere Itemscores bei den Items ‘Versagensgefühle’, ‘Selbstablehnung’, ‘Entscheidungsunfähigkeit’ und ‘Wertlosigkeit’ aufwiesen, sowie in Relation niedrigere Itemscore bei den Items ‘Traurigkeit’, ‘Verlust an Freude’, ‘Unruhe’, ‘Interessenlosigkeit’, ‘Verlust an Energie’, ‘Schlafgewohnheiten’, ‘Reizbarkeit’, ‘Appetitveränderungen’ und ‘Ermüdung’.

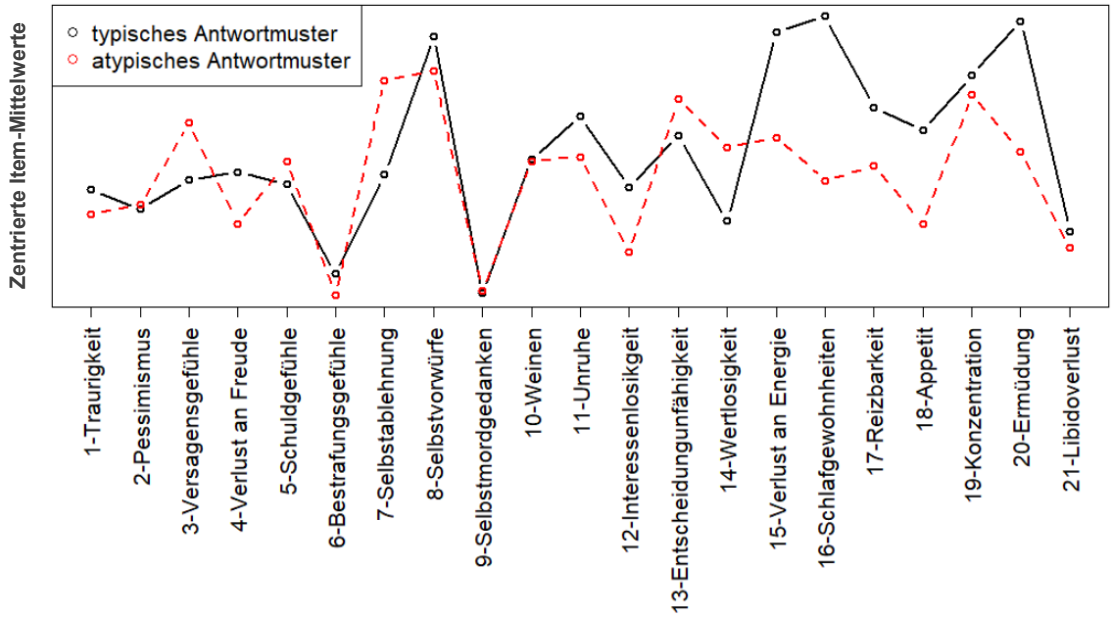


Abbildung 6: Zentrierte Item-Mittelwerte bei typischem und atypischem Antwortmuster

Bei Probanden, die aberrantes Antwortverhalten aufweisen, kann die latente Variable, also im Fall des BDI-II die Depressivität eines Probanden, nicht anhand des Personenparameters θ_v abgeschätzt werden. Umgekehrt sind keine Rückschlüsse von diesem Antwortverhalten auf das BDI-II möglich, sodass diese Probanden für die weiteren Analysen ausgeschlossen wurden. So entstand die finale Stichprobe, die sich aus 2326 jungen Menschen in der Ausbildungsphase zusammensetzte.

Mit dieser Stichprobenzusammensetzung wurde die Evaluation des GOF auf Modell- und Item-Ebene für das GRM nochmals erneuert und die Ergebnisse in Tabelle 6 (Seite 41) und Tabelle 7 (Seite 42) mit $GRM_{Personfit}$ kenntlich gemacht. Insgesamt zeigte sich eine deutliche Verbesserung des Fits auf Modell- und auch auf Item-Ebene. Allerdings zeigten die Items ‘Selbstvorwürfe’ und ‘Unruhe’ weiterhin eine eingeschränkte Anpassung an das GRM.

Um die Evaluation der Anpassungsgüte des GRM zu vervollständigen, erfolgte im Folgenden die Evaluation der lokalen Unabhängigkeit und der Form der Item Response Funktionen.

4.2.6 Überprüfung der lokalen Unabhängigkeit

Zur Überprüfung der lokalen Unabhängigkeit der Itempaare wurde Yen’s [74] Q_3-Statistik berechnet. Verwendet man, wie von Christensen [77] empfohlenen, eine Korrelation von 0,2 über der durchschnittlichen Item-Residualkorrelation als Grenzwert für das Vorliegen von lokaler Abhängigkeit, entsteht ein Bereich von $-0{,}242<Q_3<0{,}242$, in dem lokale Unabhängigkeit der Itempaare angenommen werden kann. Keines der untersuchten Items lag außerhalb dieses Bereiches (Anhang 2), sodass von lokaler Unabhängigkeit aller Itempaare ausgegangen werden kann.

4.2.7 Evaluation der Form der Item Response Funktionen (IRFs)

Die Anpassung von non-parametrischen IRT-Modellen bietet die Möglichkeit, zwei weitere Ursachen einer mangelnden Anpassung an parametrische IRT-Modelle zu evaluieren: die Verletzung der Monotonie und die Abweichung von der bei parametrischen Modellen geforderten logistischen Form der IRF.

Für die Kernel-Smoothing-Regression wurde, wie im Abschnitt 3.4.2 beschrieben, die Gauß-Verteilung genutzt. Der anhand der Silverman-Formel errechnete optimale Smoothingparameter lag bei 0,224.

Die non-parametrische IRF des Items ‘Unruhe’, das im $S\text{-}X^2$-Test nach Kang und Chen [67, 68] eine mangelnde Modellanpassung an das GRM aufwies, zeigt im Bereich sehr hoher Depressivität, einem Trait-Score von 2,9 bis 3,1 entsprechend, minimale Hinweise auf eine Verletzung der Monotonie (Abbildung 7). Das Item ‘Selbstvorwürfe’ hingegen, das im $S\text{-}X^2$-Test [67, 68] ebenfalls nur eine eingeschränkte Modellanpassung aufwies, zeigt im gesamten Kurvenverlauf keine Hinweise auf eine Verletzung der Monotonie (Anhang 3). Für beide Items zeigten sich die parametrischen und non-parametrischen IRFs über die gesamte Breite der latenten Variablen nahezu deckungsgleich.

Die Items ‘Weinen’ und ‘Ermüdung’ weisen, ähnlich wie das Item ‘Unruhe’, im Bereich hoher Depressivität *(Trait-Score > 3)* Hinweise auf eine Verletzung der Monotonie auf (Anhang 3 und Anhang 4). Für alle drei Items liegen diese Hinweise jedoch in einem

Bereich der latenten Variable, in dem die Konfidenzintervalle aufgrund der Stichprobenzusammensetzung extrem weit gefasst und die Genauigkeit der Schätzung der Regressionskurve daher eingeschränkt ist.

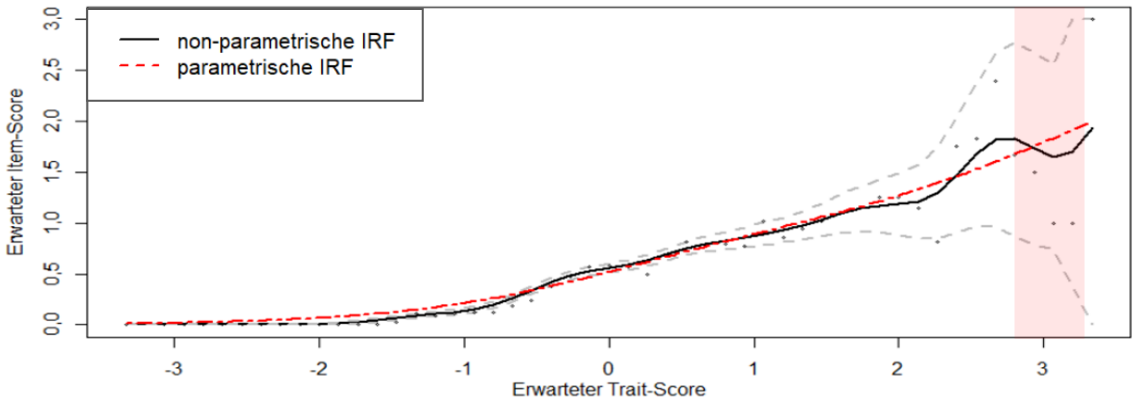


Abbildung 7: Non-parametrische Item Response Funktion von Item 11

Betrachtet man sich das Kriterium der logistisch geformten IRFs, weisen alle Items in den mittels Kernel-Smoothing-Regression errechneten non-parametrischen IRFs eine annähernd logistische Form auf und zeigen graphisch somit eine gute Anpassung an die Modellannahme parametrischer IRT-Modelle (Anhang 3 und Anhang 4).

Fasst man die Ergebnisse des ersten Teils dieser Arbeit zusammen, konnte mit dem unidimensionalen GRM ein IRT-Modell gefunden werden, das den vorliegenden Datensatz, wenn auch mit kleinen Schwächen, insgesamt zuverlässig abbilden kann.
Dies ermöglicht es für die weiteren Analysen, anhand des GRM Aussagen über die Informationsstruktur des BDI-II und seiner Items, sowie über die Depressivität der Probanden ableiten zu können.

4.3 IRT-Analyse

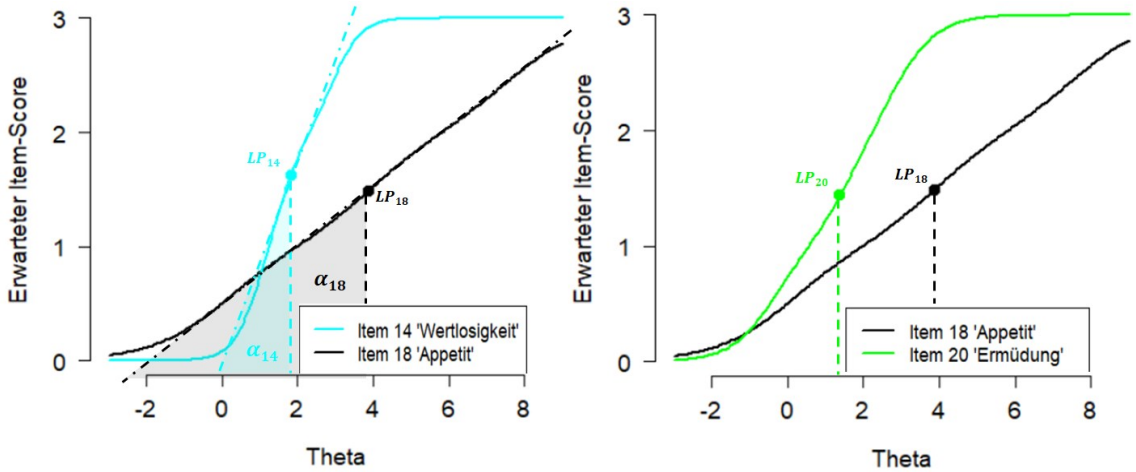
4.3.1 Schätzung der Itemparameter

Anhang 5 stellt die mittels MML-Schätzer bestimmten Itemparameter – α_i als Diskriminations- und $\delta_{i_{1-3}}$ als Schwellenparameter – mit entsprechender Standardabweichung dar.

Betrachtet man sich nun die Diskriminationsparameter α_i aller Items, zeigen diese nach dem Bewertungsmaßstab von Baker [81] eine mindestens moderate, neun Items sogar eine sehr hohe Diskriminationsfähigkeit. Errechnet man die Pearson-Korrelation zwischen den Diskriminationsparametern der IRT (α_i) und der KTT (r_{pbis}), ergibt sich eine positive signifikante $(p<0{,}05)$ Korrelation von $0{,}89$ $[0{,}75;\ 0{,}96]$. Die positive Korrelation signalisiert, dass hohe Werte bei beiden Parametern eine hohe Diskriminationsfähigkeit anzeigen [35].

Die beste Diskriminationsfähigkeit wies hierbei – sowohl in der IRT, als auch in der KTT – das Item 'Wertlosigkeit' (α=2,787, r_{pbis}= 0,725) auf, während die Items 'Schlafgewohnheiten' (α=1,030, r_{pbis}=0,434), 'Appetitveränderung' (α=0,995, r_{pbis}=0,467) und 'Libidoverlust' (α=1,017, r_{pbis}=0,459) als die am schlechtesten diskriminierenden Items identifiziert werden konnten (Anhang 5 und Abbidung 8 li.).

Betrachtet man sich die Schwellenparameter $\delta_{i_{1-3}}$ umspannen diese einen sehr weiten Bereich der latenten Variablen (-1,05 bis 7,66), konzentrieren sich mit den Locationparametern zwischen 1,36 bis 3,88 allerdings eher im höheren Bereich der Merkmalsausprägung. Errechnet man die Korrelation der Schwierigkeitsparameter der IRT (LP) und KTT (m_{Item}), ergibt sich eine negative signifikante (p<0,05) Korrelation von -0,43 [-0,73; -0,01]. Die negative Korrelation erklärt sich dadurch, dass in der KTT ein hoher mittlerer Item-Score für ein einfaches Item, ein hoher Locationparameter in der IRT dagegen für ein schwieriges Item spricht. Wie die deutlich niedrigere Korrelation der Schwierigkeitsparameter bereits vermuten ließ, kommen IRT und KTT bezüglich der schwierigsten bzw. einfachsten Items zu unterschiedlichen Einschätzungen. So stellten sich in der IRT die Items 'Appetitveränderung' (LP=3,876) und 'Libidoverlust' (LP=3,285) als die schwierigsten und die Items 'Ermüdung' (LP=1,359) und 'Verlust an Energie' (LP=1,462) als die leichtesten Items dar (Abbidung 8 re.). In der KTT dagegen werden die Items 'Suizidgedanken' (m_{Item}=0,153) und 'Bestrafungsgefühle' (m_{Item}=0,196) als die schwierigsten und die Items 'Schlafveränderung' (m_{Item}=0,775) und 'Ermüdung' (m_{Item}=0,764) als die leichtesten Items eingeschätzt (Anhang 5).



Abbidung 8: Höchste und niedrigste Diskriminationsfähigkeit (li.) bzw. Schwierigkeit (re.) aller Items

4.3.2 Item- und Testinformation

Anhand der Itemparameter lässt sich für jedes Item eine Iteminformationskurve, die auf die Iteminformationsfunktion $I(\theta_v)$ zurückgeht, berechnen. Diese wurden für alle 21 Items des BDI-II in Abbildung 9 zusammenfassend dargestellt. Trotz der Unübersichtlichkeit dieser Grafik sticht eine entscheidende Information hier deutlich hervor: Die

Frage nach dem Symptom der ‘Wertlosigkeit’ (Item 14 - cyanblau) weist über ein breites Spektrum der latenten Variablen einen besonders hohen Informationsgehalt auf.

Um abschätzen zu können, welche Items im Bereich der klinischen Cut-off-Scores, die bisher nur aus der KTT bekannt sind, den höchsten Informationsgehalt aufweisen und somit die beste Differenzierung der Depressivität ermöglichen, wurden diese Cut-off-Werte zunächst mittels linearer Transformation in IRT-Trait-Scores konvertiert. Auf diese Weise entstand als Grenzwert zwischen klinisch unauffälligen Probanden und Probanden mit Hinweis auf ein leichtes depressives Syndrom ein Trait-Score von 0,37, sowie zwischen leichter und moderater Depressivität von 1,09. Von einem schweren depressiven Syndrom wurde ab einem Trait-Score von 2,29 ausgegangen.

Überträgt man diese Werte auf Abbildung 9 und hebt diejenigen Iteminformationskurven optisch hervor, die an den Schnittpunkten den höchsten Informationsgehalt aufweisen, entsteht auf diese Weise Abbildung 10.

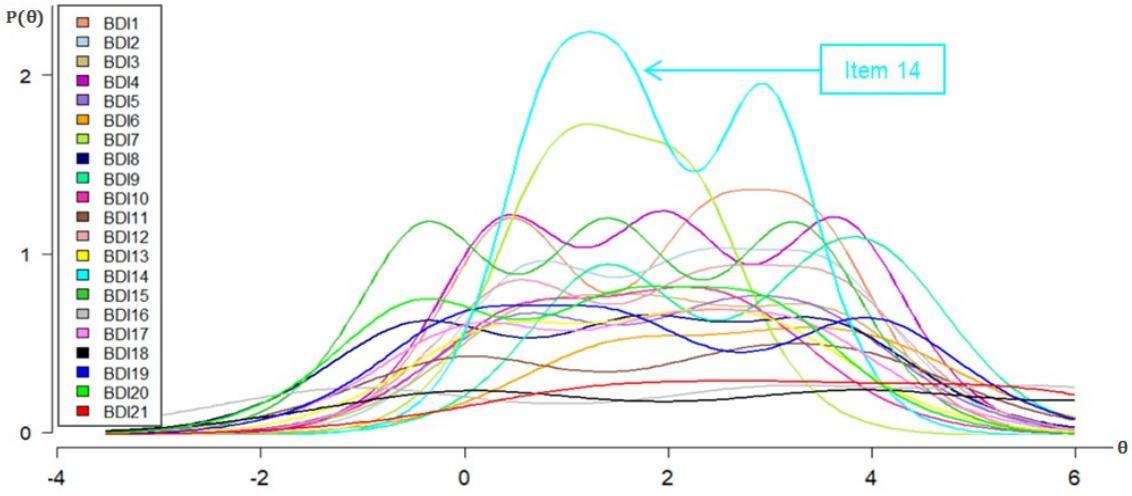


Abbildung 9: Iteminformationskurven aller 21 BDI-II-Items

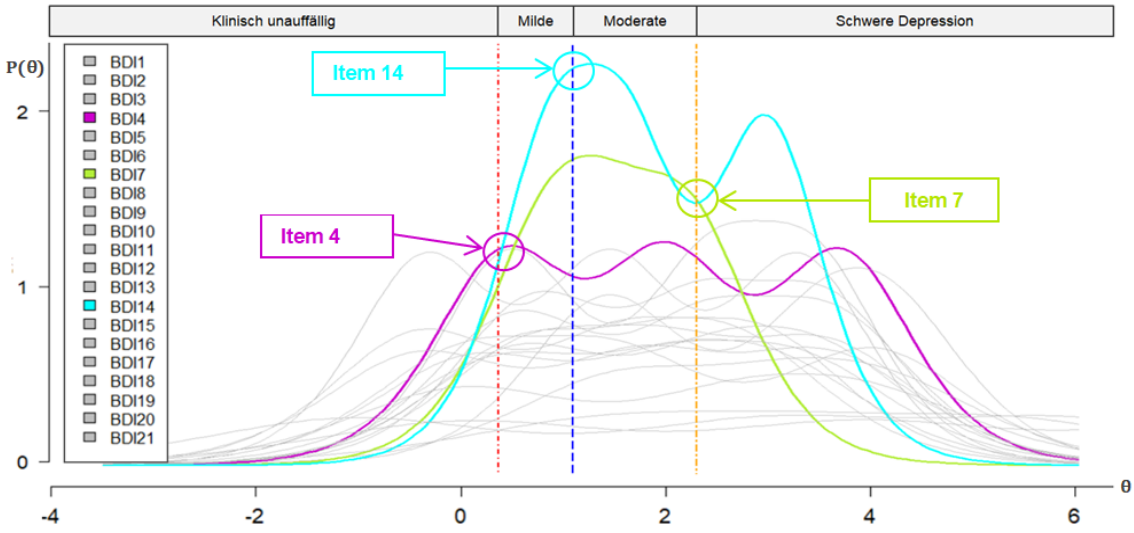


Abbildung 10: Hervorhebung einzelner Iteminformationskurven

Betrachtet man sich zunächst den Informationsgehalt am Grenzwert zwischen einerseits ‚klinisch unauffällig' und andererseits ‚milder Depression' (rote gestrichelte Linie) weist Item 4 ('Verlust an Freude', violetter Graph), den höchsten Informationsgehalt auf, knapp gefolgt von Item 14 ('Wertlosigkeit', cyanblauer Graph). Am Cut-off-Score zwischen milder und moderater Depression (gestrichelte blaue Linie) weist Item 14 ('Wertlosigkeit', cyanblauer Graph) den höchsten Informationsgehalt auf, während beim Cut-off-Score zwischen moderater und schwerer Depression (gestrichelte orange Line) Item 7 ('Selbstablehnung', hellgrüner Graph), die höchste Aussagekraft besitzt. Auch hier knapp gefolgt vom Item 'Wertlosigkeit', das somit an allen drei Grenzwerten einen sehr hohen Anteil an der Gesamtinformation aufweist.

Möchte man eine Aussage zum Test als Gesamtes machen, ist die Testinformation die entscheidende Größe. Diese setzt sich aus der Summe der Iteminformationen aller im Test enthaltenen Items zusammen. Für den BDI-II wurde die Testinformation – mit zugehörigem Standardmessfehler – in Abbildung 11 dargestellt.

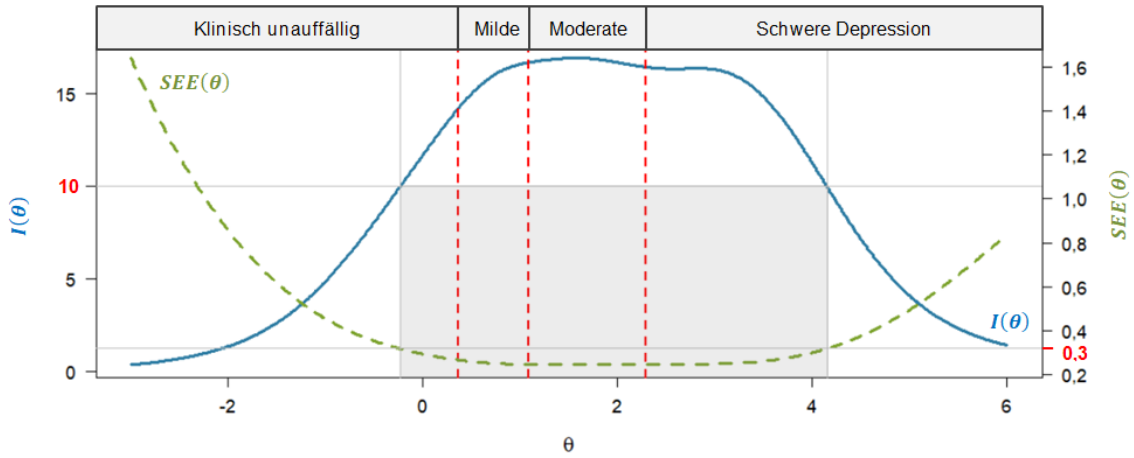


Abbildung 11: Testinformationskurve mit zugehörigem Standardmessfehler

Hier zeigt sich, dass die höchste Aussagekraft des BDI-II im Bereich der durchschnittlichen bis höheren Ausprägung der Depressivität liegt. In den Randbereichen nimmt die Testinformation dagegen ab, entsprechend steigt hier der Standardmessfehler. Die Depressivität der Probanden kann in diesen Bereichen folglich nur mit einer geringeren Sicherheit eingeschätzt werden.

Klinisch relevant ist vor allem der Bereich, in dem das BDI-II die Depressivität mit hinreichender Sicherheit einschätzten kann. Als ausreichend wird in der klassischen Testtheorie in der Regel eine Reliabilität über 90% gewertet. In der IRT entspricht dies einem Standardmessfehler $SEE(\theta_v)$ von circa $0{,}32$, oder anders ausgedrückt einer Testinformation $I(\theta_v)$ von 10. Dieser Bereich wurde daher in Abbildung 11 optisch hervorgehoben und umfasst den Bereich der latenten Variablen θ_v von $-0{,}23$ bis $4{,}15$, also einer durchschnittlichen bis ausgeprägten Depressivität. Betrachtet man sich die von

der KTT übertragenen klinischen Cut-off-Scores, zeigt sich an allen Übergangsbereichen (rote gestrichelte Linien) somit eine Reliabilität über 90%.

4.3.3 Schätzung der Personenparameter

Nach Bestimmung der Itemparameter lässt sich für jeden einzelnen Probanden ein Personenparameterschätzer $\hat{\theta}_v$ – aufgrund der Wahl eines unidimensionalen Modells bestimmt mittels EAP-Schätzer – errechnen, der durch Gewichtung anhand der Itemparameter auch das Antwortmuster der Probanden mit berücksichtigt. Abbildung 12 stellt die Trait-Scores der Probanden den klassischen Summenscores gegenüber.
Nimmt man sich einen beliebigen Summenscore vor – bspw. 5 Punkte (gestrichelte Markierung) – und vergleicht die korrespondierenden Trait-Scores (roter Bereich), zeigt sich in der vorliegenden Stichprobe ein Wertebereich zwischen $-0{,}9$ und $-0{,}2$ und verdeutlicht die im Antwortmuster der Probanden zusätzlich enthaltene Information zur Differenzierung der Depressivität, die mittels klassischem Summenscore verloren geht.

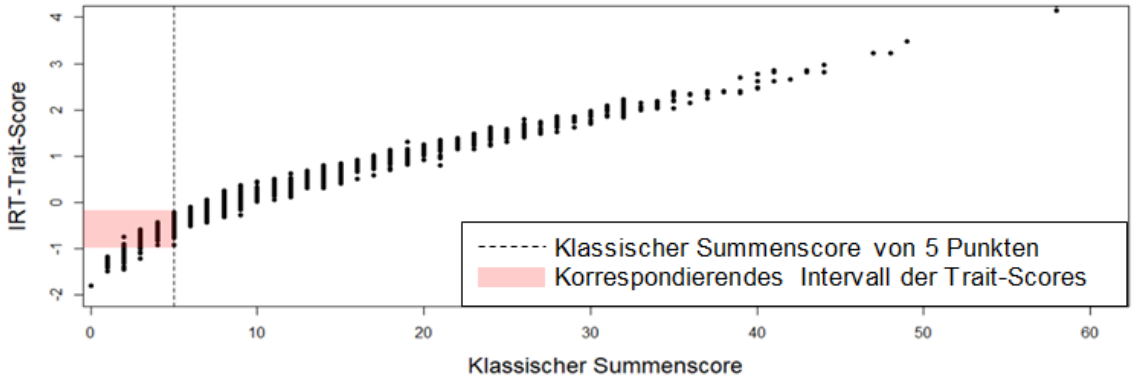


Abbildung 12: Vergleich der IRT-basierten Trait-Scores mit den KTT-basierten Summenscores

4.3.4 Evaluation auf Differential Item Functioning (DIF)

Als nächster Schritt erfolgte die Analyse auf DIF bezüglich Alter und Geschlecht.

(1) Alter

Um zu überprüfen, ob das Alter nach Kontrolle der latenten Variablen Einfluss auf das Antwortverhalten der Probanden ausübt, wurde dieses zunächst dichotomisiert:

1. Jüngere Probanden mit einem Alter < 30 Jahren $(n=2155)$
2. Ältere Probanden mit einem Alter ≥ 30 Jahren $(n=163)$

Abbildung 13 stellt graphisch die Ausprägung der Depressivität der jüngeren Probanden $(\theta_{mean} = -0{,}01)$ und älteren Probanden $(\theta_{mean} = -0{,}04)$ gegenüber.
Um DIF – uni- und nonuniformes – für ein Item nachzuweisen, müssen zunächst Modell 1 und Modell 3 miteinander verglichen werden. Wendet man hierfür den definierten Grenzwert von $0{,}0024$, der durch die Bonferroni-Korrektur für multiples Testen bei einem $\alpha=0{,}05$ entsteht, auf die Ergebnisse des globalen G^2-Tests ($Pr(X_{13}^2,2)$) an, wurden

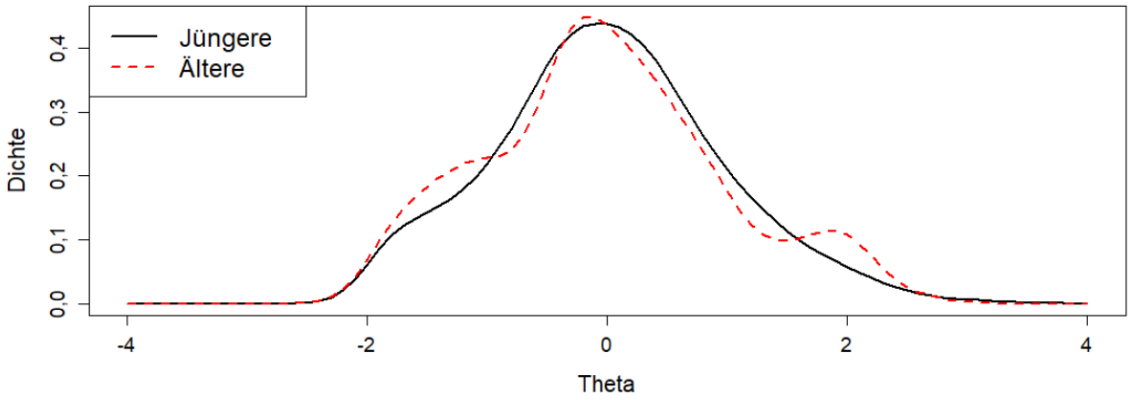


Abbildung 13: Merkmalsverteilung getrennt nach Altersgruppen

die Items ‘Selbstvorwürfe’ und ‘Libidoverlust’ als hinweisend auf das Vorliegen von DIF bezüglich des Alters markiert (Anhang 6). Anschließend wurden die Ergebnisse der Vergleiche von Modell 1 und Modell 2 $(\Pr(X_{12}^2,1))$ und der Vergleiche von Modell 2 und Modell 3 $(\Pr(X_{23}^2,1))$ berechnet. Hierbei zeigen beide markierten Items Hinweise auf uniformes DIF.

Als nächstes erfolgte die graphische Analyse der markierten Items (Abbildung 14 und Abbildung 15). Hierfür wurden die Item Response Funktionen und die dazugehörigen Option Characteristic Curves getrennt nach Alterskategorien dargestellt. Sofern kein DIF vorliegt, müssten die Kurven deckungsgleich sein. Für das Item ‘Selbstvorwürfe’ zeigt sich das Bild eines uniformen DIFs, bei dem sich bei annähernd gleichem Steigungsparameter $\boldsymbol{\alpha_i}$ die Schwellenparameter $\boldsymbol{\delta_{ik}}$ zwischen den untersuchten Gruppen unterscheiden. Das Item ‘Libidoverlust’ hingegen zeigt das Bild eines gemischten DIFs mit Unterschieden in den Parametern $\boldsymbol{\alpha_i}$ und $\boldsymbol{\delta_{ik}}$. Die Itemparameter sind mittig neben den OCCs getrennt nach Alterskategorie – oben in schwarz für die jüngeren und unten in rot für die älteren Probanden – dargestellt.

Item 8 (‘Selbstvorwürfe’)

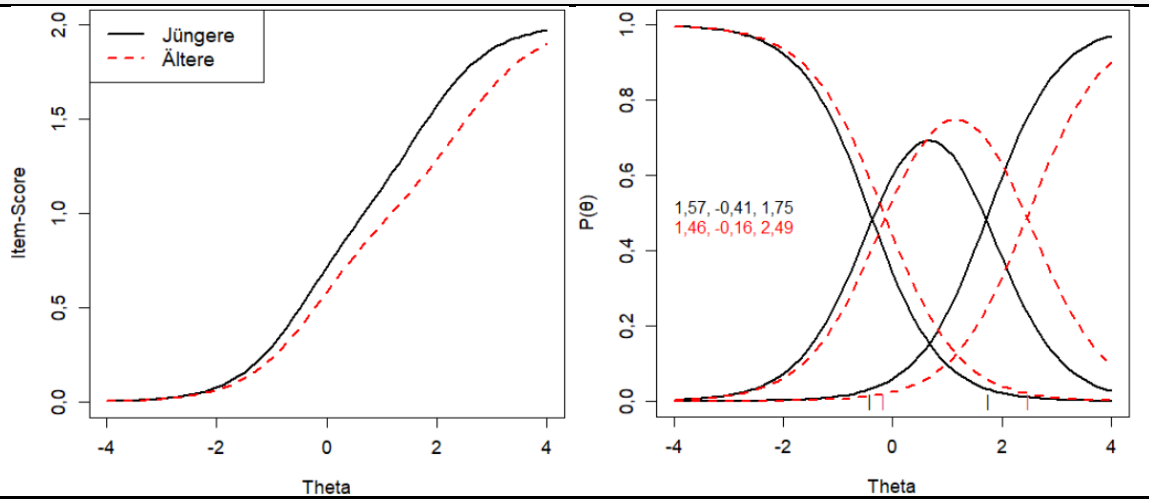


Abbildung 14: Graphische Auswertung der DIF-Analyse bezüglich des Alters – Item 8

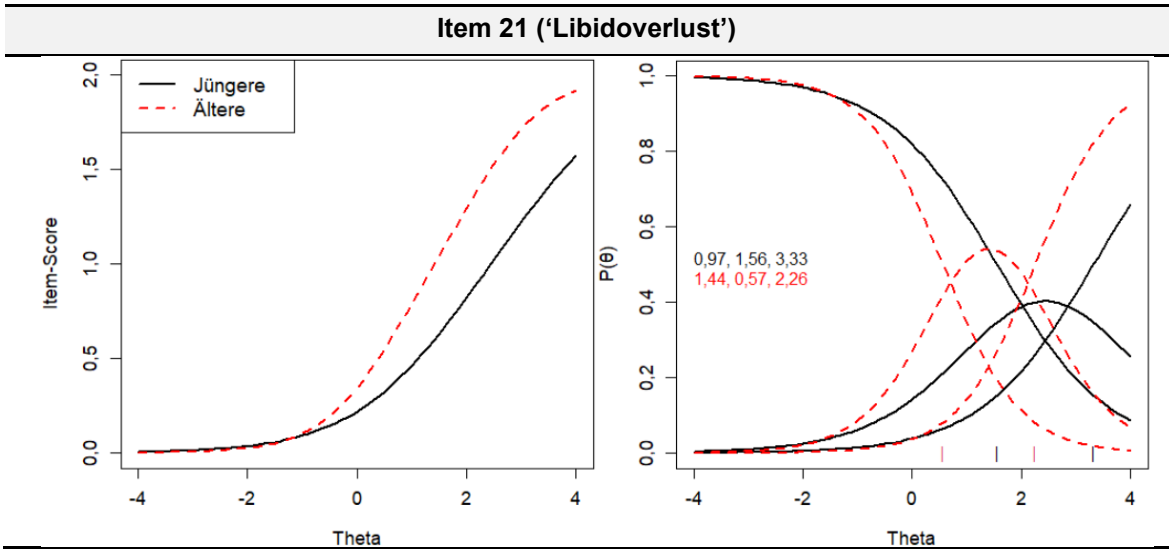


Abbildung 15: Graphische Auswertung der DIF Analyse bezüglich des Alters – Item 21

Um das Ausmaß des detektierten DIFs zu quantifizieren, erfolgte die Berechnung von ΔR^2 nach McFadden, welches für keines der Items den Grenzwert für klinisch relevantes DIF von 2% überschritt (Anhang 6).

(2) Geschlecht

Als nächstes erfolgte die Untersuchung auf DIF bezüglich des Geschlechtes, die in analoger Art und Weise wie oben für das Alter beschrieben durchgeführt wurde. Abbildung 16 stellt vergleichend den Ausprägungsgrad der Depressivität zwischen den biologischen Geschlechtern dar und zeigt eine im Schnitt etwas höhere Depressivität bei den Frauen ($\theta_{mean} = 0{,}09$) als bei den Männern ($\theta_{mean} = -0{,}19$).

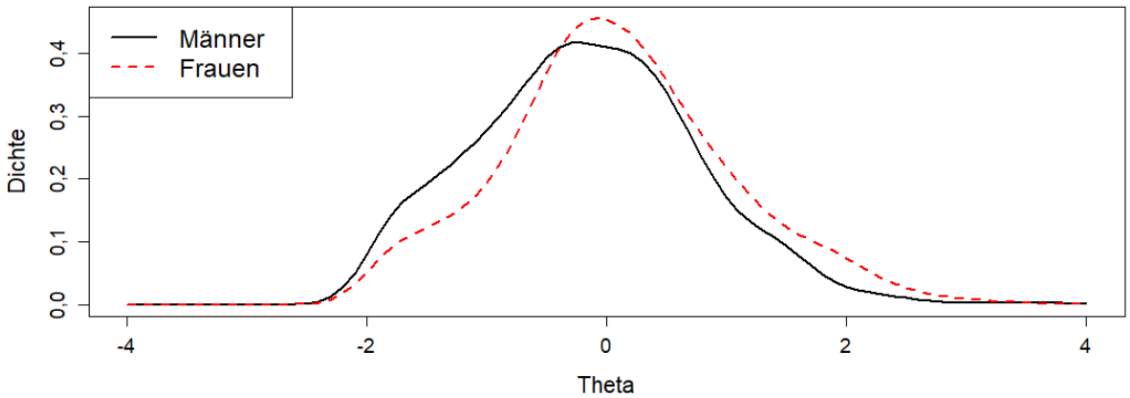


Abbildung 16: Merkmalsverteilung getrennt nach biologischem Geschlecht

Der globale G^2-Test ($Pr(X_{13}^2,2)$) ergab für insgesamt 10 Items – Item 1, 3, 4, 6, 8, 9, 10, 12, 13 und 20 – Hinweise auf das Vorliegen von DIF bezüglich des Geschlechtes (Anhang 6).

Um wiederum das Ausmaß des so detektierten DIFs zu quantifizieren, erfolgte die Auswertung von ΔR^2 nach McFadden, welches für das Item 'Weinen' mit einem ΔR_{13}^2 von 3,52%, einem ΔR_{12}^2 von 3,28% und einem ΔR_{23}^2 von 0,24% über dem Grenzwert

von 2% lag und damit auf das Vorliegen von klinisch signifikantem uniformen DIF bezüglich des Geschlechtes hinwies. Legt man den Bewertungsmaßstab von Jodoin und Gierl [88] zugrunde, so ist das DIF für ‘Weinen’ bezüglich des Geschlechts sehr knapp an der Grenze zwischen vernachlässigbarem und moderatem DIF (Anhang 6).

Abbildung 17 zeigt die Item Response Funktion und die dazugehörige Option Characteristic Curve – getrennt nach Geschlecht – für das Item ‘Weinen’. Optisch zeigt sich, neben dem uniformen DIF, auch eine non-uniforme Komponente, sodass von einem gemischten DIF bezüglich des Geschlechts ausgegangen werden muss. Im niedrigeren Ausprägungsgrad der latenten Variable wird ‘Weinen’ von weiblichen Probanden bereits bei einem in Relation zu den männlichen Probanden niedrigeren Level beschrieben, während sich dieses Verhältnis im Bereich der höheren Merkmalsausprägung umkehrt.

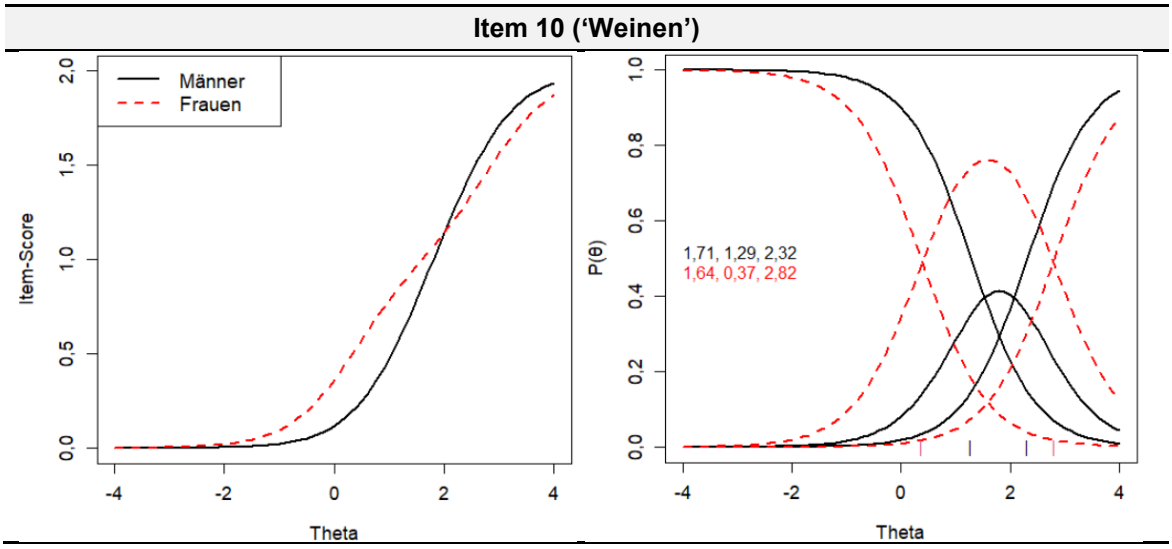


Abbildung 17: Graphische Auswertung der DIF-Analyse bezüglich des Geschlechts – Item 10

Um nun die Auswirkung des DIFs beim Item ‚Weinen’ auf den Testscore zu evaluieren, erfolgte die graphische Auswertung der Test Characteristic Curves, die für die beiden Geschlechter getrennt berechnet wurden. Deckungsgleiche Kurven, wie in Abbildung 18, signalisieren, dass kein signifikanter Effekt des DIFs auf den Testscore vorliegt.

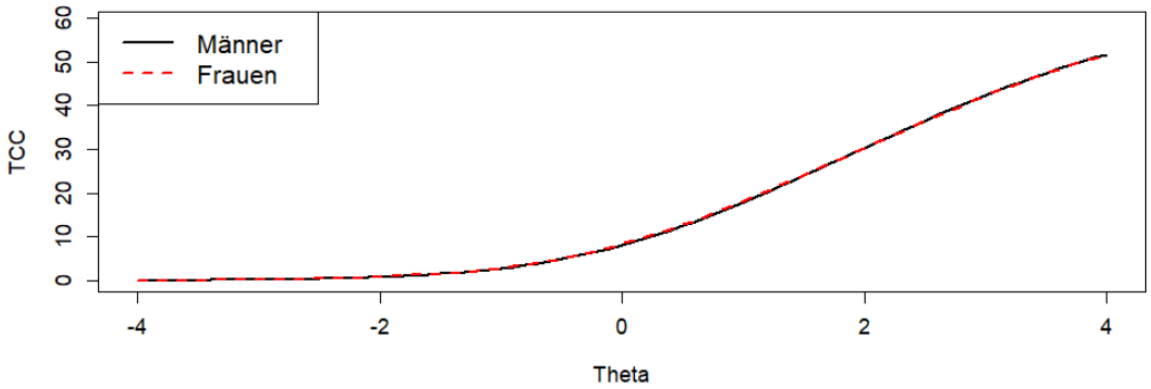


Abbildung 18: Auswirkungen des DIFs von Item 10 (‘Weinen’) auf den Testscore

5. Diskussion

Für die Früherkennung depressiver Störungen spielt die medizinische Primärversorgung, die in Deutschland vor allem durch die Hausärzte gewährleistet wird, eine entscheidende Rolle [90]. Circa 55% aller Diagnosen einer Depression gehen auf den Hausarzt zurück, sodass die unipolare Depression folglich einen häufigen Grund der Vorstellung beim Hausarzt darstellt [91]. Problematisch ist in diesem Zusammenhang allerdings, dass nur circa 5% der Patienten eine Depression als primären Konsultationsgrund nennen, während 57% sich aufgrund somatischer Beschwerden, wie Müdigkeit oder Abgeschlagenheit, in der Praxis vorstellen und die Depression als Ursache entsprechend erst im Verlauf exploriert werden muss [92]. Beesdo-Baum et al. [91, Seite 52] führten 2017 in Deutschland eine epidemiologische Querschnittsstudie durch, die untersuchte, wie *'häufig […] Patienten mit depressiven Störungen in der hausärztlichen Praxis erkannt*' werden. Hierbei zeigte sich eine Erkennungsrate psychischer Erkrankungen von 72,1%, allerdings wurde nur bei 33,7% aller Patienten, die am Stichtag die ICD-10-Kriterien erfüllten, auch die Diagnose einer Depression gestellt, sodass hier für die Zukunft noch weiterer Optimierungsbedarf besteht [91].

Ein möglicher Grund für diese niedrige Erkennungsrate könnte zum einen Zeitmangel in der hausärztlichen Routineversorgung sein, der zu einer unzureichenden Exploration psychischer Symptome führt, und zum anderen der Einsatz wenig geeigneter psychometrischer Screeningverfahren [90]. Entsprechend '*besteht in der medizinischen Primärversorgung ein Bedarf an Screening-Instrumenten, die neben einer hohen, möglichst fehlerfreien Detektionsrate unkompliziert und ökonomisch durchgeführt und ausgewertet werden können*' [90, Seite 8].

Zur Einschätzung des Schweregrads der depressiven Symptomatik eines Patienten werden in der Praxis häufig Fragebögen eingesetzt [3]. Eines der in Deutschland für diesen Zweck am häufigsten genutzten Instrumente stellt das Beck Depressions-Inventar-II [1] dar, dessen Bewertung in der Regel mittels des klassischen Summenscores erfolgt [2]. Eine noch differenziertere Einschätzung der Depressivität verspricht die Anwendung der Item Response Theorie, die die Itemparameter, bspw. die Item-Schwierigkeit, in der Score-Bildung berücksichtigt und somit die im Antwortmuster enthaltenen Informationen einfließen lässt [6].

Grundvoraussetzung hierfür ist allerdings die Eignung des BDI-II für unidimensionale IRT-Analysen. Diese Fragestellung stellt einen zentralen Punkt der vorliegenden Arbeit dar, deren Ergebnisse im folgenden Abschnitt eingehend betrachtet und kritisch hinterfragt werden sollen. In diesem Rahmen soll eine Einordnung in den bisherigen

Forschungsstand erfolgen, sowie Perspektiven für weitergehende Forschung und mögliche Limitationen der vorliegenden Arbeit diskutiert werden.

5.1 Ergebnisdiskussion und Einordnung in den Forschungskontext

5.1.1 Überprüfung der grundlegenden Eignung des BDI-II für IRT-Analysen

Einen zentralen Schritt, um die Eignung der deutschen Fassung des BDI-II [1, 2] für IRT-Analysen zu evaluieren, stellt die Feststellung der Dimensionalität des Fragebogens dar. Die Dimensionalität erfüllt dabei zwei wichtige Funktionen: zum einen stellt sie eine essentielle Grundlage für die Wahl eines geeigneten IRT-Modells dar und zum anderen begründet sich auf der Dimensionalität des Fragebogens die Bewertungsmethode eines Instrumentes [57].

Bei einem rein unidimensionalen Fragebogen, bei dem ausschließlich eine latente Variable Einfluss auf das Antwortverhalten ausübt, erscheint es ohne Diskussion sinnvoll, den Fragebogen mittels eines einzigen Scores – dem klassischen Summenscore oder einem unidimensionalen IRT-Trait-Score – zu bewerten [57].
Bei einem multidimensionalen Fragebogen, bei dem in der Faktorenanalyse eindeutig abgrenzbare Faktoren mit konstanter Itemzuordnung nachweisbar sind, ist dagegen die Bewertung mittels pro Faktor separatem Subscore ratsam [57]. Ein bekanntes Beispiel hierfür stellt die deutsche Version des Big Five Inventory-2 (BFI-2) [93] dar, der die fünf Persönlichkeitsdomänen – Extraversion, Verträglichkeit, Gewissenhaftigkeit, Negative Emotionalität und Offenheit – behandelt, in der konfirmatorischen Faktorenanalyse fünf klar abgrenzbare Faktoren mit konstanter Itemzuordnung aufweist und daher idealerweise anhand von fünf separaten Subscores bewertet wird [93].
Schwieriger gestaltet es sich, wenn die Ergebnisse der Faktorenanalyse, wie im Fall des BDI-II, keine konsistenten Ergebnisse zeigen [57]. Für den BDI-II umfassten die beschriebenen klassischen Faktorstrukturen hierbei Ein- bis Dreifaktorenstrukturen, wie bereits ausführlich in Abschnitt 3.4.2 dargestellt. Die Itemzuordnung zu den einzelnen Faktoren unterschied sich jedoch studien- bzw. stichprobenabhängig [37, 38].
Diese Inkonsistenz stellt für die Festlegung der Bewertungsmethode ein großes Problem dar [57], da beispielsweise die nachgewiesene Zweifaktorenstruktur einerseits eine Bewertung mittels zweier Subscores nahelegt, andererseits durch die stichprobenabhängige Zuordnung der Items zu den beiden Faktoren die Frage unbeantwortet bleibt, welche Items zu welchem Subscore zusammengerechnet werden sollen. Dies führte zu einer über Jahre hinweg andauernden Diskussion über die sinnvollste Art der Bewertung des Beck Depressions-Inventars-II [57].

Eine neuere Entwicklung, die auf Grundlage dieser Diskussion entstand, ist die zunehmende Anwendung von sog. Bifaktor-Modellen, bei denen die gemeinsame Item-Varianz durch die Varianz eines Generalfaktors (unidimensionaler Testscore) und multiplen Gruppenfaktoren (multidimensionale Subscores) widergespiegelt wird [37, 57]. In verschiedenen Studien, unter anderem von Subica et al. [37] und Faro et al. [38], konnte für das BDI-II ein überlegener Fit der Bifaktor-Struktur gegenüber der klassischen Faktorstruktur nachgewiesen werden. Dies bestätigte sich auch bei der hier untersuchten Stichprobe, bei der sich der beste Fit für die Bifaktor-Lösung mit dem durch alle Items repräsentieren Generalfaktor ‚Depressivität' und den drei Gruppenfaktoren ‚kognitiv', ‚somatisch' und ‚affektiv' (nach Beck et al. [2]) zeigte.

Für die Festlegung der Bewertungsmethode eines Fragebogens bietet die Evaluation von Bifaktormodellen einen großen Vorteil [57]. Zeigt sich bei einem Instrument für das Bifaktor-Modell der beste Fit, kann mittels eines einzigen, unidimensionalen Scores der Generalfaktor des Fragebogens abgebildet werden. Übertragen auf das BDI-II bedeutet es folglich, dass durch den wiederholt nachgewiesenen überlegenen Fit der Bifaktor-Modelle der Generalfaktor ‚Depressivität' am besten durch einen einzigen, unidimensionalen Score – dem klassischen Summenscore oder einem unidimensionalen Trait-Score – repräsentiert wird. Durch die stichprobenabhängig wechselnde Zuordnung der Items zu den Gruppenfaktoren spielen die Subscores für die Bewertung dagegen nur eine sehr untergeordnete Rolle [34]. Zusammenfassend lässt sich festhalten, dass die in der Praxis bewährte Bewertungsmethode des BDI-II mittels eines einzigen Scores durch die in der vorliegenden Arbeit nachgewiesene Bifaktorstruktur weiter gestützt wird.

Für die Wahl eines IRT-Modells ermöglichte die nachgewiesene Bifaktor-Struktur des BDI-II für die vorliegende Arbeit prinzipiell zwei mögliche Herangehensweisen:
Da sich der Generalfaktor ‚Depressivität' am besten durch einen einzigen Score darstellen lässt, erscheint einerseits die Anwendung eines unidimensionalen IRT-Modells, insbesondere nach dem Nachweis von essentieller Unidimensionalität des BDI-II, gerechtfertigt [22, 57]. Andererseits wäre die Anwendung eines Bifaktor-Modells möglich, wie dies beispielsweise von Williams et al. [10] zur psychometrischen Validierung des BDI-II in einer Stichprobe Erwachsener mit Störungen aus dem autistischen Formenkreis durchgeführt wurde. Dies stellt allerdings einen eher neuen, rechnerisch deutlich komplexeren und bisher nur selten für Studien zum BDI gewählten Weg dar. Durch den Nachweis essentieller Unidimensionalität war nicht davon auszugehen, dass sich die IRT-Trait-Scores und somit die Bewertung der Depressivität der Probanden signifikant zwischen dem unidimensionalen und dem Bifaktor-Modell unterscheiden [57]. Da bei

Anwendung dieser neuen IRT-Modelle trotzdem die Vergleichbarkeit der Ergebnisse mit bisherigen, internationalen Studien zum BDI eingeschränkt wäre, wurde letztendlich auf Basis dieser vorgenannten Überlegungen die Entscheidung zur Verwendung eines unidimensionalen IRT-Modells für die vorliegende Arbeit getroffen.

Um anhand des gewählten IRT-Modells sowohl Aussagen über die Informationsstruktur des BDI-II, also auch über die Depressivität von Probanden ableiten zu können, ist die Anpassungsgüte des gewählten Modells entscheidend. Streng genommen stellt kein real vorliegendes Modell je eine perfekte Repräsentation der Testdaten dar und kann bei ausreichend großer Stichprobe stets falsifiziert werden. Die entscheidende Frage ist somit nicht, ob das gewählte Modell perfekt passt, sondern ob es eine ausreichend gute Repräsentation des zugrunde liegenden Datensatzes erlaubt, um allgemeingültige Aussagen ableiten zu können [18].
Zusammenfassend zeigte das unidimensionale Graded Response Modell nach Samejima [28] – trotz der in Abschnitt 4.2 dargestellten Schwächen – insgesamt eine gute Modell-Anpassung in der untersuchten Stichprobe, sodass die Grundlage dafür vorlag, anhand des Antwortverhaltens der Probanden sowohl Rückschlüsse auf deren Depressivitätsgrad, als auch über die Informationsstruktur des BDI-II ableiten zu dürfen [23].

5.1.2 Nachweis der Messinvarianz bezüglich Alter und Geschlecht

Um die Depressivität von Probanden anhand ihres BDI-II-Traitscores vergleichen zu dürfen, ist es Voraussetzung, dass der Fragebogen die latente Variable, also die depressive Symptomatik, für alle Probanden gleichartig misst, unabhängig von verschiedenen demographischen Eigenschaften, wie beispielsweise Alter, Geschlecht oder Herkunft [7].

Der Vergleich zwischen den Geschlechtern ergab in der untersuchten Stichprobe, dass die weiblichen Probanden einen im Schnitt etwas höheren BDI-II-Traitscore aufwiesen, als die männlichen Probanden. Um herauszuarbeiten, ob dieser Unterschied durch eine real vorliegende erhöhte Depressivität der Frauen hervorgerufen wurde oder das Resultat einer unterschiedlichen Beschreibung der depressiven Symptomatik zwischen den Geschlechtern war, wurde eine Analyse auf Differential Item Functioning bezüglich des Geschlechts durchgeführt. Es wurde folglich überprüft, ob sich die psychometrischen Charakteristika der Items bzw. des Fragebogens systematisch zwischen den beiden Geschlechtern unterscheiden.
Die Analyse ergab für das BDI-II, dass zehn der 21 Items Hinweise auf ein geschlechtsspezifisch unterschiedliches Antwortverhalten zeigten. Ein klinisch relevantes Ausmaß nach Jodoin und Gierl [88] erreichte jedoch ausschließlich das Item ‘Weinen’.

Hier beschrieben Frauen im Bereich niedriger Depressivität frühzeitiger eine Intensivierung des Symptoms ‘Weinen’, als die männlichen Probanden, während sich dieses Verhältnis im Bereich höherer Depressivität zugunsten der männlichen Probanden umkehrte. Werden zur Berechnung der Trait-Scores geschlechtsspezifisch separate IRT-Parameter verwendet, zeigte sich allerdings kein signifikanter Einfluss auf den Trait-Score. Folglich war das verzerrende Potential des DIFs nicht ausgeprägt genug, um geschlechtsabhängig zu einer unterschiedlichen Einschätzung des Schweregrads der depressiven Symptomatik der Probanden zu gelangen. Einschränkend muss hier allerdings beachtet werden, dass diese Aussage nur dann zutreffend ist, wenn auch der gesamte BDI-II genutzt wird. Im Rahmen von computerbasiertem adaptiven Testen (CAT) kann das beobachtete DIF hingegen eine deutliche klinische Relevanz erlangen. Grundidee des adaptiven Testens ist es, nur diejenigen Items für einen Probanden einzusetzen, die an dessen individueller Merkmalsausprägung den höchsten Informationsgehalt aufweisen, während auf andere Items verzichtet wird. Auf diese Weise kann eine Einschätzung der Depressivität bereits mit weniger Items, somit geringerem Zeitaufwand und folglich einer geringeren Belastung des Probanden erfolgen [11]. Allerdings kann es bereits bei gering ausgeprägten geschlechtsspezifischen Unterschieden im Antwortverhalten, das bei Anwendung des gesamten Fragebogens keinen signifikanten Einfluss auf die Einschätzung der latenten Variablen ausübt, bei Anwendung von nur wenigen Items im Rahmen von adaptivem Testen zu einem signifikanten verzerrenden Einfluss auf die Einschätzung der Merkmalsausprägung kommen [22]. Auf die Messung der Depressivität übertragen könnte es in einem CAT somit durch die Anwendung einer Frage, die sich mit dem Symptom ‘Weinen’ befasst, trotz exakt gleichem Level der Depressivität geschlechtsabhängig zu unterschiedlichen Ergebnissen des adaptiven Tests und damit zu geschlechtsabhängig unterschiedlichen Einschätzungen der Depressionsschwere kommen.

Bei Leistungstest stellt das Entfernen bzw. Ersetzen eines Items, das Hinweise auf DIF zeigt, eine sinnvolle Lösungsstrategie dar. Bei psychologischen Tests, bei denen die Items jedoch verschiedene Symptombereiche abfragen, würde dies unter Umständen zum Verlust relevanter Informationen führen und sollte nicht unkritisch durchgeführt werden. Eine Möglichkeit dem verzerrenden Einfluss ohne Informationsverluste zu begegnen, stellt hier die Berechnung von gruppenspezifischen IRT-Parametern dar [22, 94] und sollte im Rahmen eines adaptiven Tests vor Anwendung eines Items, das sich mit dem Symptom ‘Weinen’ befasst, erwogen werden.

Vergleicht man die Ergebnisse der vorliegenden Arbeit mit bisherigen Erkenntnissen zu geschlechtsspezifischem Antwortverhalten, zeigen sich einige Übereinstimmungen. So

wurde bei Frauen in der Vergangenheit deutlich häufiger eine Depression ärztlich diagnostiziert, als bei Männern [3]. Zudem konnten zahlreiche Studien zeigen, dass Frauen, unabhängig vom genutzten Instrument zur Einschätzung der Depressivität im Vergleich zu Männern höhere Scores aufweisen [4, 40]. Dies stellte die Grundlage dar, die bekanntesten Depressions-Inventare auf systematische geschlechtsabhängige Antwortverzerrungen hin zu untersuchen. Hierbei kamen Williams et al. [10] in ihrer IRT-basierten Studie am BDI-II, die an einer Stichprobe aus dem autistischen Formenkreis durchgeführt wurde, zu ähnlichen Ergebnissen. Auch dort zeigte sich ein geschlechtsabhängig signifikant unterschiedliches Antwortverhalten auf das Item ‘Weinen’. Dies bestätigte sich auch bei der Studie von de Sá Junior et al. [36], die für Frauen mittels der portugiesischen Version des BDI-II durch das Item ‘Weinen’ ebenfalls eine tendenzielle Überschätzung der Depressivität im niedrigen Depressionsbereich, sowie eine Unterschätzung im Bereich hoher Depressivität nachweisen konnten.

Vergleicht man diese Ergebnisse mit Untersuchungen an anderen Depressions-Inventaren, beispielsweise Santor et al. [95] am BDI-I, Cole et al. [96] am CES-D sowie Teresi et al. [97] an der PROMIS-Itembank, zeigten sich in all diesen Studien Hinweise, dass das jeweilige Item ‘Weinen’ geschlechtsabhängig unterschiedlich beantwortet wurde. Da sich diese geschlechtsabhängigen Antwortverzerrungen bei verschiedenen Instrumenten und damit einhergehend bei unterschiedlich formulierten Items zum Symptom ‘Weinen’ zeigte, kann gefolgert werden, dass das unterschiedliche Antwortverhalten nicht auf die Formulierung der Items zurückgeführt werden kann. Das unterschiedliche Antworverhalten wäre möglicherweise dadurch zu erklären, dass auch nicht-depressive Frauen im Vergleich zu Männern häufiger und intensiver weinen, sodass in den niedrigen Bereichen der latenten Variable durch die Frage nach vermehrtem ‘Weinen’ eher die weibliche Neigung zu Tränen als Coping-Stategie, als tatsächlich eine erhöhte Depressivität erfasst wird [36]. Dies würde bei Frauen die tendenzielle Überschätzung der Depressivität im Bereich niedriger Merkmalsausprägung erklären.

Obwohl ‘Weinen’ häufig zu den Symptomen einer Depression gezählt wird, ist diese Verbindung wissenschaftlich bisher kaum belegt. So konnten Vingerhoets et al. [36, 98] in ihrem systematischen Literatur-Review keine solide empirische Grundlage für die Annahme finden, dass eine Depression zu einer erhöhten Häufigkeit von ‘Weinen’, bzw. dem ‘Unmöglichwerden von Weinen’ bei schwerster Depression führt. Trotzdem weisen nahezu alle Messinstrumente zur Einschätzung der Depressivität Items auf, die sich mit ‘Weinen’ befassen. Im Mittel wird hierbei jedoch nur eine moderate Korrelation mit der gemessenen Depressivität erreicht. Bisher gibt es kein einheitliches Vorgehen,

wie mit dem Symptom ‘Weinen’ in den Depressions-Inventaren umgegangen werden sollte, sodass hier noch weiterer Forschungsbedarf besteht [36, 98].

Betrachtet man sich als zweite mögliche Einflussvariable das Alter der Probanden, zeigte sich in der untersuchten Stichprobe eine nahezu gleichgestaltete Verteilung der Merkmalsausprägung in den beiden Teilstichproben (jünger als 30 vs. älter als 30).
Um nachzuweisen, ob es bei Nutzung des BDI-II zu altersspezifischen Antwortverzerrungen kommt, wurde eine Analyse auf Differential Item Functioning bezüglich des Alters angeschlossen. Hierbei ergaben sich in der vorliegenden Stichprobe Hinweise darauf, dass Probanden im jüngeren Alter (<30 Jahre) bereits bei geringerer Depressivität eine Zunahme an ‚Selbstvorwürfen‘ berichten, als ältere Probanden (≥30 Jahre). Das Symptom ‚Libidoverlust‘ hingegen wurde von jüngeren Probanden erst bei einer höheren Depressivität angegeben, als bei älteren Probanden. Beides überschritt allerdings nicht die Schwelle für klinisch relevantes DIF nach Jodoin und Gierl [88].

Vergleicht man die Ergebnisse der vorliegenden Arbeit mit denen internationaler Studien, ergeben sich in einigen Punkten Überschneidungen, jedoch auch vereinzelt Differenzen. So konnten Williams et al. [10] in einer Stichprobe von Probanden aus dem autistischen Formenkreis, die die englischsprachige Version des BDI-II bearbeiteten, ebenfalls Hinweise darauf finden, dass die Items ‘Selbstvorwürfe’ und ‘Libidoverlust’ je nach Altersstufe (jünger als 30 vs. älter als 30) signifikant unterschiedlich beantwortet wurden. Dies bestätigte sich auch bei de Sá Junior et al. [36], die für die portugiesische Version des BDI-II nachweisen konnten, dass bei jüngeren Probanden (<30 Jahre) die Depressivität durch das Item ‘Libidoverlust’ tendenziell unterschätzt wird, während sie bei älteren Probanden (>30 Jahre) eher überschätzt wird.
Schließt man in die Betrachtungen auch Analysen an Vorversionen des BDI-II mit ein, lassen sich anhand der Arbeit von Kim et al. [99] einige interessante Parallelen ziehen. So konnten Kim et al. für den BDI-I zeigen, dass sich das Antwortverhalten der Probanden auf das Item ‘Selbstvorwürfe’ altersabhängig (jünger als 60 vs. älter als 60) signifikant unterschied. Wie in der vorliegenden Arbeit zeigten auch dort jüngere Probanden im Vergleich zu älteren Probanden bei niedrigerer Depressivität eine vermehrte Angabe von ‘Selbstvorwürfen’. Interessant ist hierbei insbesondere, dass die Beobachtung trotz der deutlich unterschiedlichen Altersgrenze – 30 Jahre vs. 60 Jahre als Cut-off-Alter – identisch ausfiel. Hieraus ließe sich schließen, dass die Wahrnehmung von ‘Selbstvorwürfen’ mit steigendem Alter kontinuierlich abnimmt.
Für das Item ‘Schlafveränderungen’ zeigten sich in der Analyse von Kim et al. [99] altersabhängig sehr deutliche Unterschiede im Antwortverhalten, während sich in der vorliegenden Arbeit, sowie in den Arbeiten von Williams et al. [10] und de Sá Junior et

al. [36] mit einer der vorliegenden Arbeit ähnlichen Altersverteilung, hierfür kein Anhalt finden ließ. Folglich scheinen die Unterschiede im Antwortverhalten auf das Item ‘Schlafveränderungen’ erst im höheren Alter klinisch relevant zu werden und konnten daher anhand der in der vorliegenden Arbeit untersuchten Altersstruktur nicht nachgewiesen werden.

Bezüglich des Items ‘Libidoverlust’ zeigten sich in der Studie von Kim et al. [99] deutlich andere Ergebnisse als in der vorliegenden Arbeit. So zeigte sich dort, dass jüngere Probanden (<60 Jahre) bereits bei niedrigerer Depressivität von einem ‘Libidoverlust’ berichten, als ältere Probanden (>60 Jahre), also genau gegensätzlich zu der in der vorliegenden Arbeit beobachteten Verteilung, sowie der Verteilung bei de Sá Junior et al. [36] und Williams et al. [10].

Betrachtet man sich mögliche Gründe für das zwischen jüngeren und älteren Probanden nachgewiesene unterschiedliche Antwortverhalten bei der Frage nach einem ‘Verlust des Interesses an Sex’, muss offenbar zwischen den verschiedenen Altersbereichen unterschieden werden.

Im hohen Alter wird laut Kim et al. [99] die Depressivität anhand des Items ‘Libidoverlust’ im Vergleich zu mittelalten Probanden tendenziell unterschätzt. Das bedeutet, dass ältere Probanden bei gleicher Depressivitätsausprägung eine geringere Ausprägung des Symptoms ‘Libidoverlust’ berichten, als mittelalte Probanden. Dies könnte sich dadurch erklären, dass Menschen im höheren Alter unabhängig von einer Depression von einer natürlicherweise mit dem Alter einhergehenden, sowie durch somatische Erkrankungen mitbedingten, abnehmenden Libido betroffen sind. Bei Auftreten einer Depression im Alter könnte somit ein durch die depressive Störung bedingter ‘Libidoverlust’ von den Probanden als normale Abnahme der Libido fehlinterpretiert und somit im Fragebogen nicht erwähnt werden. Dies würde zu einer verminderten Angabe des Symptoms ‘Libidoverlust’ beitragen und wäre somit eine mögliche Erklärung für das in der Studie von Kim et al. [99] beobachtete altersabhängig unterschiedliche Antwortverhalten auf das Item ‘Libidoverlust’.

Eine mögliche Erklärung für das beobachtete, unterschiedliche Antwortverhalten im Bereich junger bis mittelalter Personen sowohl in der vorliegenden Arbeit, als auch bei de Sá Junior et al. [36] und Williams et al. [10], könnte durch ein bei jungen Menschen ausgeprägter vorhandenes Schamgefühl bedingt sein. Dieses häufig mit der Entwicklung der eigenen Persönlichkeit einhergehende Schamgefühl könnte dazu beigetragen haben, dass jüngere Probanden bei der Frage nach einem Libidoverlust diesen erst bei deutlich ausgeprägterer Symptomatik zugaben, als bspw. ältere Erwachsene. Ein weiteres Indiz, das diese Theorie unterstützt, ist, dass in der hier unter-

suchten Stichprobe bei den Probanden unter 20 Jahren die Frage nach einem ‘Verlust des Interesses an Sex’ annähernd doppelt so häufig unbeantwortet blieb (0,02%), wie in der Gruppe der Probanden über 30 Jahren (0,01%).

Einschränkend muss für all diese Schlüsse allerdings bedacht werden, dass sich die Altersstruktur der vorliegenden Arbeit deutlich von der der letztgenannten Studie unterscheidet. Während der Schwerpunkt der Arbeit von Kim et al. [99] im höheren Altersbereich lag, wurde in der vorliegenden Arbeit eine Stichprobe aus Personen in der Ausbildungsphase untersucht und somit insbesondere jüngere Menschen in die Studie eingeschlossen. Die in der vorliegenden Arbeit getroffenen Aussagen zu Differential Item Functioning bezüglich des Alters sind folglich auch nur für diesen begrenzten Altersbereich – 18 bis 56 Jahre – zulässig, während die von Kim et al. [99] gefolgerten Schlüsse nur für den höheren Altersbereich gelten. Um weitergehende Aussagen zu altersabhängig verschiedenem Antwortverhalten treffen und somit die hier diskutierten Schlüsse, die studienübergreifend getroffen wurden, überprüfen zu können, wäre eine umfangreichere Stichprobe und insbesondere eine homogenere Altersverteilung von sehr jungen bis sehr alten Probanden erforderlich.
Da die Gruppe der älteren Probanden mit nur 164 Personen für eine DIF-Analyse grenzwertig klein war, ist die Aussagekraft der Auswertung jedoch beschränkt. Auf eine Verschiebung des Cut-off-Wertes, der zu einer homogeneren Gruppengröße geführt hätte, wurde jedoch zugunsten der besseren Vergleichbarkeit mit den Studien von Williams et al. [10] und de Sá Junior et al. [36] verzichtet, die bei ähnlicher Altersstruktur beide ebenfalls als Cut-off 30 Jahre gewählt haben. Für Folgestudien, die das Alters-DIF näher untersuchen sollen, wäre – wie bereits oben erwähnt – eine homogenere Altersverteilung mit damit einhergehend größeren Teilstichproben notwendig.
Anhand der dargestellten Ergebnisse dieser Arbeit lässt sich zusammenfassen, dass die Depressivität von Probanden unterschiedlichen Alters und Geschlechts anhand des BDI-II-Trait-Scores verglichen werden darf und sich die psychometrischen Charakteristika des BDI-II insgesamt als relativ robust gegen Einflüsse durch konstruktferne Merkmale erwiesen haben. Einschränkend müssen hier allerdings die nur begrenzt miterfassten Merkmale der Probanden beachtet werden, die die DIF-Analysen auf die beiden Charakteristika ‚Alter‘ und ‚Geschlecht‘ limitierte.

5.1.3 Informationsstruktur des Beck Depressions-Inventars-II

Das Beck Depressions-Inventar wurde primär zur Evaluation depressiver Symptomatik im klinischen Setting konzipiert [9]. Seit seiner Veröffentlichung wurde das BDI-II jedoch in diversen Studien sowohl in klinischen, als auch in nicht-klinischen Stichproben eingesetzt und wies in beiden Populationen solide psychometrische Charakteristika auf

[34], sodass das BDI-II mittlerweile auch als Screening-Instrument für depressive Störungen empfohlen wird.

Mithilfe der IRT können im Vergleich zur KTT weitergehende Aussagen zur Informationsstruktur eines Instruments abgeleitet werden, die unter anderem eine Einschätzung zu sinnvollen Einsatzgebieten eines Fragebogens erlauben. Die entscheidende Größe stellt hierbei die Testinformation dar, die direkt mit dem Standardmessfehler eines Instrumentes in Verbindung steht [83].

Während in der KTT die Reliabilität nur als konstante Größe über die komplette Bandbreite der Depressivität angegeben werden kann, beispielsweise mittels Cronbach α [17], ermöglicht es die IRT für jeden einzelnen Bereich der Depressivität einzuschätzen, wie zuverlässig das BDI-II den Schweregrad der depressiven Symptomatik beschreiben kann. Für ein Screening-Instrument ist insbesondere der Grenzbereich zwischen klinisch unauffälligen und leichtgradig depressiven Personen relevant und legt für letztgenannte weitere diagnostische Schritte nahe, wohingegen der Schwerpunkt im klinischen Setting insbesondere in den Grenzbereichen zwischen leichter, moderater und schwerer Depressivität liegt und wiederum therapeutische Konsequenzen nach sich zieht.

Klassischerweise wird diese Unterteilung anhand von Cut-Off-Werten getroffen, die durch die Korrelation mit dem Goldstandard der Diagnosestellung definiert werden, für das BDI-II bspw. durch die Korrelation mit einem standardisierten klinischen Interview wie dem CIDI. Für den klassischen Summenscore konnten auf diese Weise in diversen Studien Cut-off-Werte detektiert werden [19]. Eine sehr häufig genutzte Einteilung hierfür ist: Probanden mit einem BDI-II-Summenscore zwischen 0 und 13 Punkten gelten als klinisch unauffällig. Ein Summenscore zwischen 14 und 19 weist auf eine mildes, zwischen 20 und 29 auf eine moderates und zwischen 30 und 63 Punkten auf ein schweres depressives Syndrom hin [2].

Aufgrund der bisher noch sehr eingeschränkten Anzahl IRT-basierter Studien zum BDI-II existieren aktuell noch keine IRT-basierten Cut-off-Werte zur Einteilung der Depressionsschwere anhand des Trait-Scores. Durch das Studiendesign der vorliegenden Arbeit als sekundäre Datenauswertung war bei fehlenden parallel erhobenen standardisierten klinischen Interviews auch in dieser Arbeit keine Berechnung von IRT-basierten Grenzwerten möglich. Um dennoch eine Einschätzung der Depressivität anhand der Trait-Scores zu ermöglichen, wurden daher in der vorliegenden Arbeit die mittels klassischem Summenscore erhobenen Cut-off-Werte in IRT-Trait-Scores konvertiert und als IRT-Grenzwerte genutzt. Um jedoch das volle Potential der besseren Diskriminationsfähigkeit der IRT-Trait-Scores ausschöpfen zu können, sind weitere IRT-basierte Studien erforderlich, die insbesondere eine parallele Erhebung von BDI-II

und strukturierten klinischen Interviews beinhalten und somit die Errechnung von IRT-basierten Cut-off-Werten erlauben.

Betrachtet man sich nun die Informationsstruktur des BDI-II und wählt – analog zu einer Reliabilität von 90% in der KTT – eine Testinformation größer 10 als Grenzwert [83], erlaubt das BDI-II somit im Bereich durchschnittlicher bis hoher Depressivität eine zuverlässige Einschätzung der Schwere der depressiven Symptomatik. Alle drei von der KTT übertragenen Grenzwerte liegen im Bereich maximaler Aussagekraft des BDI-II und legen damit eine hohe Aussagekraft in den Grenzbereichen nahe. Folglich bestätigt die Auswertung der Testinformation, dass das BDI-II durch die hohe Aussagekraft am Cut-off-Wert zwischen klinisch unauffälligen Personen und Personen mit leichter depressiver Störung als Screening-Instrument in der Allgemeinbevölkerung eingesetzt werden kann, aber durch die hohe Aussagekraft an den Cut-off-Werten zwischen leichter, moderater und schwerer depressiver Störung auch zur Einschätzung des Schweregrads der depressiven Symptomatik im klinischen Setting verwendet werden kann.

Im Bereich unterdurchschnittlicher Depressivität hingegen zeigt das BDI-II nur eine geringere Testinformation, damit einhergehend eine geringe Aussagekraft und einen hohen Standardmessfehler. Die Testinformation hebt somit eine Schwäche des BDI-II hervor, die bei der Evaluation mittels KTT verborgen bleibt, nämlich die nur geringe Aussagekraft im Bereich unterdurchschnittlicher Depressivität. Für die Praxis bedeutet dies, dass das BDI-II bei klinisch unauffälligen Probanden nicht zur präzisen Einschätzung depressiver Symptomatik eingesetzt werden sollte. Um entscheiden zu können, welcher Fragebogen hierfür geeigneter erscheint, errechneten Zhao et al. [32] eine gemeinsame Metrik für fünf in China gängige Depressionsinventare – das Beck Depressions-Inventar-II (BDI-II), den Center of Epidemiologic Studies Depression Scale (CES-D), den Patient Health Questionnaire-9 (PHQ-9), den Depression Anxiety and Stress Scale (DASS) und den Hospital Anxiety and Depression Scale (HADS) – und konnten nachweisen, dass der CES-D im Bereich niedriger Depressivität dem BDI-II überlegen ist, während das BDI-II im Bereich hoher Depressivität wiederum besser diskriminieren konnte, als alle anderen untersuchten Instrumente. Wahl et al. [33] führten eine ähnlich angelegte Studie im deutschsprachigen Raum durch, konnten aber von den drei neben dem BDI-II untersuchten Instrumenten, dem PHQ-2, dem PHQ-9 und dem Mental Health Inventory-5 (MHI-5), für keines eine deutliche Überlegenheit im Bereich der unterdurchschnittlichen Depressivität nachweisen.

Entsprechend lässt sich zusammenfassen, dass das BDI-II sowohl als Screening-Instrument in der Allgemeinbevölkerung eingesetzt werden kann, als auch zur Evalua-

tion der depressiven Symptomatik im klinischen Setting. Soll jedoch eine Einschätzung im Bereich unterdurchschnittlicher Depressivität erfolgen, ist das BDI-II nur bedingt geeignet. Als Alternative mit höherer Aussagekraft in diesem Bereich bietet sich hierfür beispielsweise der CES-D an.

Analog zur Auswertung der Testinformation lässt sich im Rahmen der IRT auch für jedes einzelne Item des BDI-II die zugehörige Iteminformation berechnen, die eine Aussage darüber ermöglicht, in welchem Bereich der Depressivität das Item eine hohe Aussagekraft besitzt. Interessant ist in diesem Zusammenhang der Vergleich mit der Kurzform des BDI-II, dem Beck-Depressions-Inventar Fast-Screen for Medical Patients (BDI-FS) [90]. Dieser Fragebogen, der als Screening-Tool für die medizinische Grundversorgung entwickelt wurde, klammert bewusst die somatischen Items des BDI-II aus und setzt sich aus sieben Items des BDI-II zusammen. Die Auswahl der Items erfolgte dabei zum einen auf Basis theoretischer Überlegungen (‘Traurigkeit’, ‘Verlust an Freude’ und ‘Suizidgedanken’) und zum anderen auf den höchsten Ladungen auf die kognitive Komponente in der Faktorenanalyse (‘Pessimismus’, ‘Versagensgefühle’, ‘Selbstablehnung’ und ‘Selbstvorwürfe’) [90].
Für ein Screening-Instrument ist insbesondere eine hohe Aussagekraft im Grenzbereich zwischen klinisch unauffälligen Probanden und Probanden mit leichter Depressivität entscheidend. Vergleicht man die anhand der vorliegenden Arbeit errechnete Informationsstruktur derjenigen Items, die Teil des BDI-FS sind, mit denjenigen, die im BDI-FS nicht berücksichtigt werden, so weisen die im BDI-FS berücksichtigten Items ‘Traurigkeit’, ‘Pessimismus’, ‘Verlust an Freude’ und ‘Selbstablehnung’ in der vorliegenden Stichprobe im entsprechenden Grenzbereich die höchste Aussagekraft auf (siehe Abbildung 9, Seite 48). Während das Item ‘Suizidgedanken’, das in der untersuchten Stichprobe nur eine begrenzte Aussagekraft im Grenzbereich aufwies, auf Basis theoretischer Überlegungen in den BDI-FS aufgenommen wurde, wurden die Items ‘Selbstvorwürfe’ und ‘Versagensgefühle’, die ebenfalls in der hier vorliegenden Arbeit nur eine geringe Aussagekraft im Grenzbereich aufwiesen, aufgrund der Ergebnisse der Faktorenanalyse aufgenommen. Betrachtet man sich die Informationskurven der BDI-II-Items, erscheint es sinnvoll, das Item ‘Wertlosigkeit’ im BDI-FS zu berücksichtigen, das in der vorliegenden Stichprobe am Übergang zwischen klinisch unauffälligen Probanden und Probanden mit leichter Depression ein deutlich höheres Diskriminationspotential aufwies und zudem in den Bereichen behandlungsbedürftiger Depressivität im Vergleich zu den abgefragten Symptomen eine noch deutlich bessere Diskriminationsfähigkeit zeigte. Durch zusätzliche Berücksichtigung des Symptoms ‘Wertlosigkeit’, ggf. auch im Austausch gegen eines der Items ‘Versagensgefühle’ oder

‘Selbstvorwürfe’ erscheint es wahrscheinlich, dass die Sensivität und Spezifität des BDI-FS weiter verbessert werden könnte.

Ähnliches gilt auch für den von der S3-Leitlinie [4] vorgeschlagenen Patient-Health-Questionnaire-2 (PHQ-2), einem 2-Fragen-Test, der eine Sensivität von 96% und eine Spezifität von 57% zur Erfassung einer unipolaren Depression aufweist und sich aus folgenden beiden Fragen zusammensetzt:

1. ‘*Fühlten Sie sich im letzten Monat häufig niedergeschlagen, traurig, bedrückt oder hoffnungslos?’*
2. ‘*Hatten Sie im letzten Monat deutlich weniger Lust und Freude an Dingen, die Sie sonst gerne tun?’* [4, Seite 37]

Werden beide Fragen mit ‘Nein’ beantwortet, liegt eine Depression nur mit einer Likelihood-Ratio von 0,07 vor. Entsprechend ist das Risiko, einen depressiven Patienten zu übersehen, extrem gering. Allerdings kommt es durch die geringere Spezifität des Tests zu einer höheren Rate falsch-positiver Befunde, die dann einen erhöhten diagnostischen Aufwand mit sich bringen [100]. Betrachtet man sich die Gebiete, die thematisch durch den PHQ-2 abgefragt werden, berücksichtigt das Instrument zwei der drei Major-Kriterien der Depression, die niedergeschlagene Stimmung und die Anhedonie [4]. Diese Symptombereiche werden beim BDI-II mittels der Items ‘Traurigkeit’ (Items 1), ‘Pessimismus’ (Item 2), ‘Verlust der Freude’ (Item 4) und ‘Verlust des Interesses’ (Item 12) abgefragt und decken sich teilweise mit der Auswahl im BDI-FS. Auch hier wäre entsprechend denkbar, dass durch die Hinzunahme des Symptoms ‘Wertlosigkeit’ im PHQ-2 die Sensivität und Spezifität des Zwei-Fragen-Test verbessert und somit das Auftreten von falsch-positiven und insbesondere von falsch-negativen Ergebnissen noch weiter reduziert werden könnte.

5.1.4 Evaluation der im Antwortmuster enthaltenen Information

Die IRT ermöglicht neben der genaueren Evaluation der Informationsstruktur eines Instrumentes zudem auch das Herausarbeiten der im Antwortmuster enthaltenen Informationen, auf die im nachfolgenden Abschnitt näher eingegangen werden soll. Hierbei unterscheidet man zum einen die auf dem Antwortmuster basierende noch feinere Differenzierungsmöglichkeit der depressiven Symptomatik mittels IRT-Trait-Score und zum anderen das Herausfiltern von Probanden, die ein auffälliges Antwortmuster im BDI-II zeigen und deren BDI-II-Traitscore entsprechend nicht unkritisch zur Einschätzung der Depressivität des Probanden eingesetzt werden sollte.

Wird das BDI-II der klassischen Testtheorie folgend ausgewertet, wird zur Einschätzung der Depressivität der einfache Summenscore genutzt [6]. Das heißt, dass die Punktzahl, die pro Item erreicht wird, zu einem Testscore aufsummiert wird. Jede

Frage geht aufgrund der Annahme paralleler Tests dabei zu einem gleichen Anteil in den Testscore ein [18], unabhängig davon, ob das Item ein Symptom abfragt, das bereits bei geringerer Ausprägung der Depressivität vermehrt höher bewertet wird (z.B. ‘Ermüdbarkeit’), oder ein Symptom abfragt, das erst bei hohem Depressionslevel bejaht wird, wie z.B. ‘Appetitveränderungen’ oder die ‘Suizidalität’ eines Probanden.
In den Trait-Score, der das IRT-Korrelat zum Summenscore in der KTT darstellt, geht die Bewertung eines Items dagegen gewichtet ein [6]. Das bedeutet, dass der Trait-Score höher ausfällt – der Proband also als depressiver eingeschätzt wird – wenn bei einem Item mit hohem Schwierigkeitsparameter, z.B. ‘Appetitveränderung’, Kategorie 3 gewählt wird, als wenn die gleiche Kategorie bei einem Item mit niedriger Schwierigkeit (z.B. ‘Ermüdbarkeit’) gewählt wird. Hierdurch wird das Antwortmuster in der Bewertung der Depressivität mit berücksichtigt und erlaubt so eine noch genauere Diskrimination der Depressivität, als dies mittels einfachem Summenscore in der KTT möglich ist [18].
Entscheidend ist diese zusätzliche Diskriminationsfähigkeit insbesondere dann, wenn anhand des Testscores eine Einteilung in Diagnosegruppen erfolgen soll. Hierbei unterscheidet man klinisch unauffällige Probanden von Probanden mit leichter, moderater und schwerer Depressivität. Von dieser Eingruppierung kann dann die Notwendigkeit weiterer Diagnostik beziehungsweise die Therapiebedürftigkeit eines Patienten abgeleitet werden. Wie bereits im vorausgehenden Abschnitt eingehend erklärt, existieren jedoch aktuell noch keine IRT-basierten Cut-off-Werte für diese Unterteilung, die es ermöglichen würden, das volle Potential dieser zusätzlichen Diskriminationsfähigkeit auszuspielen, sodass hier noch weiterer Forschungsbedarf besteht.

Neben der gerade detailliert erläuterten, im Antwortmuster der Probanden enthaltenen Information zur Einschätzung der Depressivität erlaubt die IRT zudem, Probanden mit aberrantem Antwortmuster zu detektieren. Ein Antwortmuster gilt dann als aberrant, wenn es unter der gewählten Modellannahme nur extrem unwahrscheinlich auftreten würde [11]. Eine wichtige Ursache für aberrantes Antwortverhalten ist, neben sprachlichen Problemen bei der Bearbeitung, der Versuch der Manipulation. So zeigen Probanden, die versuchen, einen Fragebogen in möglichst kurzer Zeit zu beantworten, also ohne die Fragen konzentriert zu lesen und ehrlich zu beantworten, häufig ein atypisches Antwortmuster. Auch der Versuch, die tatsächlich vorliegende Symptomatik verstärkt darzustellen, also zu simulieren, bzw. die Symptomatik zu verharmlosen (Dissimulation), führt häufig zu einem auffälligen Ergebnis der Person-Fit-Statistik.

Eine in den letzten Jahren vermehrt untersuchte Ursache für atypisches Antwortmuster stellt die bei manchen Probanden vorliegende, von der Mehrheit der Personen abweichende Ausprägung der Symptomkomplexe dar [101].

Bevor näher auf klinische Einsatzmöglichkeiten der Person-Fit-Statistik eingegangen wird, soll zunächst die genauere Betrachtung der Antwortmuster von Probanden mit atypischem Antwortverhalten in der vorliegenden Stichprobe vorangestellt werden. Insgesamt zeigte sich eine Häufung von aberrantem Antwortverhalten im Bereich höherer Depressivität. Betrachtet man sich die Items, denen von Probanden mit atypischem Antwortmuster in Relation vermehrt zugestimmt werden konnte, fiel auf, dass diese ausschließlich das Spektrum der kognitiven Symptome umfassten, während die in Relation weniger hoch bewerteten Items der somatisch-affektiven Komponente zuzurechnen waren. Eine mögliche Interpretation dieser Ergebnisse stellt die Annahme der Existenz von unterschiedlich häufig ausgeprägten Depressions-Typologien dar. Hierbei scheinen Probanden, die vermehrt kognitive Symptome schildern, während somatisch-affektive Symptome im Verhältnis weniger angegeben werden, vom vorliegenden Modell nicht adäquat mit abgebildet zu werden.

Zur Überprüfung der Hypothese, dass ein Zusammenhang zwischen atypischem Antwortverhalten und bekannten Depressions-Typologien besteht, führten Conijn et al. [101] eine Studie mittels Inventory of Depressive Symptomatology (IDS) durch. Hier zeigte sich neben einer signifikanten Korrelation von atypischem Antwortverhalten mit melancholischer Depression, atypischer Somatisierung und atypischer Selbstmord-Ideation eine deutlich signifikante Korrelation mit einer mittels Markerfragen evaluierten, mangelnden Datenqualität. Hier wären für die Zukunft Studien wünschenswert, die, insbesondere für das Beck Depressions-Inventar-II, eine weitere Überprüfung des Zusammenhangs zwischen atypischem Antwortverhalten und Ausprägungsform der Depressivität auf der einen und einer mangelnden Datenqualität auf der anderen Seite ermöglichen. Sollte sich hier der Zusammenhang zwischen atypischem Antwortverhalten und Depressions-Typologie bestätigen, wäre dies eine wichtige Grundlage dafür, die IRT-basierte Person-Fit-Statistik gewinnbringend im Rahmen von Therapiestudien einsetzen zu können. In verschiedenen Studien über die Wirksamkeit von Arzneimitteln zur Therapie einer Depression, wie bspw. Imipramin, konnte eine unterschiedlich stark ausgeprägte Wirksamkeit in Abhängigkeit von der Depressions-Typologie nachgewiesen werden. Eine hierfür bisher häufig vorgenommene Einteilung stellte die Unterteilung in ‚endogene' und ‚reaktive' Depression dar, wobei Personen mit einer ‚reaktiven' Depression eine schlechtere Medikamentenansprechrate aufweisen [102]. Interessant wäre in diesem Zusammenhang, ob sich die Ansprechraten von Arzneimitteln bzw. allgemeiner Therapieangeboten auch signifikant zwischen

Gruppen mit modellkonformen und Gruppen mit atypischem Antwortverhalten im BDI-II unterscheiden und somit eine Vorhersage des Therapieansprechens auf diese Weise ermöglicht werden könnte.

Einen weiteren, interessanten Ansatz verfolgten Wardenaar et al. [41], die versuchten, mittels der Person-Fit-Statistik in einer Stichprobe von Myokardinfarktpatienten herauszufinden, ob das BDI-I in dieser Stichprobe tatsächlich die Depressivität erfassen kann oder aufgrund der somatischen Erkrankung die Depressivität mittels des BDI-I in dieser Population tendenziell überschätzt wird. Hier zeigte sich bei Probanden mit atypischem Antwortverhalten, das durch die Angabe von vermehrt kognitiven Symptomen definiert war, eine höhere Korrelation des BDI-I-Testscores mit den Ergebnissen des parallel durchgeführten strukturierten Interviews (CIDI), als bei Probanden mit typischem Antwortmuster mit vorherrschend somatischen Symptomen. Hieraus wurde gefolgert, dass das BDI-I bei Patienten nach einem Myokardinfarkt nicht ausschließlich die Depressivität misst, sodass der Einsatz des BDI-I zur Einschätzung der Depressivität in dieser Patientengruppe nicht unkritisch erfolgen sollte. Dieser Ansatz bietet auch für viele weitere somatische Erkrankungen die Möglichkeit, den Einfluss der somatischen Erkrankung auf die Evaluation der Depressivität mittels BDI abzuschätzen, um so möglicherweise weitere Populationen zu identifizieren, bei denen der Einsatz von Fragebögen und speziell vom BDI-II zur Einschätzung der Depressivität nur eingeschränkt geeignet erscheint.

Wanders et al. [103] konnten in einer praktisch angelegten Studie zudem zeigen, dass der behandelnde Psychiater durch eine Warnung auf das Vorliegen von atypischem Antwortverhalten (‘Autofeedback’) in einem Depressionsinventar in bis zu 60% der Fälle neue Einblicke in den Fall gewinnen konnte und sogar in bis zu 75% hierdurch Anlass für eine erweiterte Patientenanamnese bestand. So können Patienten unter Umständen ein Interesse daran haben, bei der Evaluation der Depressionsschwere ein höheres Level vorzutäuschen (Simulation), während Patienten in anderen Situationen dagegen versuchen können, ihre depressive Symptomatik herunterzuspielen (Dissimulation). Diese Versuche spiegeln sich häufig in einem ungewöhnlichen Antwortmuster wieder und können mittels der IRT-basierten Person-Fit-Statistik identifiziert werden. Wanders stellte daraufhin in den Raum, dass der behandelnde Arzt durch ein automatisches Feedback über atypisches Antwortverhalten direkt nach dem Ausfüllen eines Depressionsinventars hilfreiche Informationen und Unterstützung in der klinischen Entscheidungsfindung erhalten könne [103]. Allerdings fehlen bisher noch klinische Folgestudien, die Wanders Hypothese stützen würden. An diesem Punkt setzt

die im Rahmen dieser Promotionsarbeit erstellte Online-Applikation an, auf die im nachfolgenden Abschnitt näher eingegangen werden soll.

5.2 Vorstellung der Online-Applikation

Um zukünftige Forschungsvorhaben zum Thema IRT-Analyse des Beck Depressions-Inventars-II zu unterstützen und somit die Datenlage zu diesem Thema zu verbessern, wurde im Rahmen dieser Promotionsarbeit eine frei verfügbare Online-Applikation erstellt, die grob zwei Teilbereiche umfasst.

Zum einen enthält die Applikation ein Online-Tool, mit dem es möglich ist, sich anhand der Antworten einer Person auf das Beck Depressions-Inventar-II neben dem klassischen Summenscore zusätzlich einen IRT-Trait-Score zu errechnen, der auf den IRT-Parametern der vorliegenden Arbeit basiert. Zudem wird, dem Vorbild von Wanders [103] folgend, ein automatisches Feedback zu atypischem Antwortverhalten auf Basis der Personfit-Statistik ausgegeben.

Zum anderen kann für einen Datensatz, der sich aus Antworten auf den BDI-II zusammensetzt, eine strukturierte IRT-Analyse nach dem Vorbild dieser Arbeit erstellt werden. Die Applikation ist unter der Internetadresse

https://bdi-ii.shinyapps.io/BDI-II-Rechner/

zu finden und führt den Anwender durch ein Menü, in dem schrittweise die hier ausführlich dargestellten Analysen durchlaufen werden. Die jeweiligen Ergebnisse können per Download gespeichert und für Veröffentlichungen genutzt werden. Um den Funktionsumfang der Applikation auch ohne entsprechenden Datensatz testen zu können, besteht zudem die Möglichkeit, sich einen Beispieldatensatz zu generieren, der sich aus jeweils zufällig ausgewählten Antwortmustern der in der vorliegenden Arbeit genutzten Stichprobe zusammensetzt. Durch Anonymisierung und mehrmalige zufallsbasierte Änderung der Reihenfolge ist eine Rückverfolgung von einem Antwortmuster auf den ursprünglichen Probanden nicht möglich.

Durch den kostenlosen Zugang und die Gestaltung der Applikation, die die Durchführung von IRT-Analysen für einen BDI-II-Datensatz auch ohne technische Vorkenntnisse in der Bedienung von R erlaubt, soll die Durchführung zukünftiger Studien erleichtert und somit die Voraussetzungen für die breitere Anwendung von IRT-Trait-Scores auch in der Praxis, beispielsweise im Rahmen von hausärztlichen Evaluationen der Depressivität, für die Zukunft verbessert werden. Allerdings muss beachtet werden, dass die Äquivalenz der Ergebnisse mittels der Online-Applikation bisher nicht durch ein Validierungsverfahren bestätigt werden konnte und die Ergebnisse somit nicht zur medizinischen Entscheidungsfindung genutzt werden sollten.

5.3 Limitationen der Studie

Im Folgenden sollen nun die Limitationen der vorliegenden Studie, die sich aus der Rekrutierung der Stichprobe, dem Studiendesign und dem verwendeten Messinstrument ergeben könnten, beleuchtet werden.

Die erste Limitation der vorliegenden Studie ergibt sich durch die Zusammensetzung der Stichprobe. Ein großer Vorteil der IRT gegenüber der KTT liegt zwar in der Stichprobenunabhängigkeit der IRT, die zur Auswertung der Item-Charakteristika keine repräsentative Stichprobe benötigt, allerdings wird eine große, heterogene Stichprobe gefordert, um eine zuverlässige Parameter-Schätzung zu gewährleisten [18]. Diese Heterogenität war in der vorliegenden Arbeit durch die Wahl einer primär nichtklinischen Stichprobe junger Menschen nur eingeschränkt gegeben. Trotz der relativ umfangreichen Gesamtstichprobe von 2376 Probanden wurden manche Antwortkategorien von nur sehr wenigen Probanden gewählt. Dies betraf insbesondere diejenigen Kategorien, die eher bei einem hohen Depressionslevel ausgewählt werden. Beispielsweise wurde in der vorliegenden Stichprobe Kategorie 3 von Item 18 (‘Appetitveränderungen’ – ‘Ich habe überhaupt keinen Appetit’ bzw. ‘Ich habe ständig großen Hunger’) insgesamt von nur 4 Probanden gewählt. Durch die nur geringe Probandenzahl im hohen Depressionsbereich ist somit die geforderte Heterogenität der Stichprobe nur eingeschränkt vorhanden und die Parameterschätzungen können folglich im Bereich hoher Depressivität unpräzise werden.
Durch Hinzunahme einer zusätzlichen klinischen Stichprobe, deren Depressionslevel eher im höheren Ausprägungsbereich zu erwarten ist, wäre eine homogenere Verteilung der Depressivität über das gesamte Spektrum hinweg und somit eine präzisere Schätzung der latenten Variable mittels IRT möglich gewesen.

Aus der vorausgehenden Limitation, die auf die Stichprobenzusammensetzung zurückgeht, fußt eine weitere Limitation der vorliegenden Arbeit: die fehlende Kreuzvalidierung. Um eine Kreuzvalidierung durchzuführen, müsste die vorliegende Stichprobe in zwei Teilstichproben unterteilt werden. Die anhand der ersten Teilstichprobe evaluierten Kennwerte könnten dann anhand der zweiten Teilstichprobe validiert werden.
Allerdings wäre bei einer Durchführung in der vorliegenden Stichprobe mit einer deutlichen Verzerrung durch die Stichprobenzusammensetzung zu rechnen gewesen, da sich die Problematik der nur geringen Probandenzahl im höheren Bereich der Depressivität deutlich verschärfen würde. Da bei Durchführung einer Kreuzvalidierung die Parameterschätzungen in der ersten Teilstichprobe an Präzision verloren hätten und zum anderen die Aussagekraft der Kreuzvalidierung anhand der zweiten Teilstichprobe

nur begrenzte Aussagekraft aufgewiesen hätte, wurde daher in der vorliegenden Arbeit auf die Durchführung einer Kreuzvalidierung verzichtet.

Eine weitere Limitation der Arbeit ergibt sich durch die Rekrutierung der Stichprobe ausschließlich an Universitäten, Fachhoch- und Berufsschulen. Die Analyse ergab in der vorliegenden Stichprobe keinen Hinweis auf klinisch signifikantes DIF bezüglich der besuchten Bildungseinrichtung. Aufgrund der beschränkten Anzahl von nur 75 Fachhochschülern – also deutlich unter der für DIF-Analysen empfohlenen Mindestzahl von 200 Probanden – und der damit einhergehenden nur beschränkten Aussagekraft dieses Ergebnisses, wurde auf eine ausführliche Darstellung in der Arbeit verzichtet. Durch die Wahl dieser Rekrutierungsmethode enthielt die Stichprobe zudem nur Probanden, die sich zum Erhebungszeitpunkt in der Ausbildungsphase befanden, sodass eine generelle Aussage zum Einfluss des Bildungsniveaus auf die psychometrischen Charakteristika des BDI-II nicht möglich war. Auch die Analyse auf DIF bezüglich des Alters konnte bei einer Spannbreite von 18 bis 58 Jahren und einer deutlich linksschiefen Altersverteilung nur in einem beschränkten Umfang erfolgen. Eine Generalisierbarkeit der Erkenntnisse auf die Gesamtbevölkerung, die auch Personen aus der bildungsfernen Schicht oder Personen im hohen Alter umfasst, kann entsprechend nicht ohne weiteres erfolgen. Zudem wurden einige Charakteristika, wie beispielsweise die ethnische Herkunft der Probanden oder etwaige somatische Erkrankungen, nicht miterfasst und standen somit einer DIF-Analyse nicht zur Verfügung.

Eine weitere Limitation der Studie ergibt sich aus der Durchführung der Studie als sekundäre Datenanalyse. Da die Primärstudien zur Evaluation der Depressivität in den Stichproben ausschließlich das BDI-II nutzten, standen keine Daten zur Korrelation mit anderen Möglichkeiten der Depressions-Einschätzung zur Verfügung. Aufgrund des großen Stichprobenumfangs erscheint die Durchführung strukturierter klinischer Interviews, die den Goldstandard der Diagnosestellung einer Depression darstellen, an der gesamten Stichprobe allerdings auch schwer umsetzbar. Prinzipiell wäre allerdings denkbar, eine Substichprobe zur Evaluation der Diskriminationsfähigkeit des BDI-II parallel mittels strukturierter klinischer Interviews auf das Vorliegen einer Depression hin zu untersuchen, um anschließend die Sensivität und Spezifität des BDI-II im Gesamten und im Speziellen auch für jedes Item separat errechnen zu können, sowie die Definition IRT-basierter Cut-off-Werte zu ermöglichen. Dies würde die Vorteile einer differenzierteren Einschätzung der Depressivität auch auf die Cut-off-Werte ausweiten und somit eine noch präzisere Einschätzung der Depressivität erlauben.

6. Zusammenfassung

Am Ende der Arbeit lässt sich zusammenfassend festhalten, dass sich das BDI-II ausreichend gut für eine unidimensionale IRT-Analyse mit dem GRM eignet, um anhand dieses Modells allgemeingültige Aussagen über die Informationsstruktur des BDI-II ableiten zu können. Die Evaluation der Informationsstruktur sowohl auf Test- als auch auf Item-Ebene zählt zu den zentralen Elementen der IRT. Während die KTT eine konstante Reliabilität über die komplette Bandbreite der Merkmalsausprägung annimmt und für das BDI-II somit lediglich die Aussage erlaubt, dass die Depressivität der Probanden mittels BDI-II ausreichend zuverlässig eingeschätzt werden kann, können in der IRT Messfehlerstatistiken errechnet werden. Diese ermöglichen es, für jeden Bereich der Depressivität eine separate Einschätzung darüber abzugeben, wie hoch die Aussagekraft und respektive wie hoch der Messfehler in diesem Bereich ist. Für das BDI-II ergab die Auswertung, dass der Fragebogen den Schweregrad der depressiven Symptomatik im Bereich durchschnittlicher bis hoher Depressivität zuverlässig einschätzen kann. Für die Praxis lässt sich folglich ableiten, dass das BDI-II sowohl als Screening-Instrument in der Allgemeinbevölkerung, als auch zur Differenzierung des Ausprägungsgrades depressiver Symptomatik in klinischen Stichproben geeignet ist. Während die vorgenannte Aussage mit der Einschätzung anhand der KTT übereinstimmt, die eine hohe Reliabilität über die komplette Bandbreite der Depressivität annimmt, kommt die IRT im Bereich unterdurchschnittlicher Depressivität zu einem deutlich anderen Ergebnis. Hier zeigt das BDI-II in der Testinformation eindeutige Defizite, sodass das BDI-II aufgrund des damit einhergehenden hohen Messfehlers folglich nicht zur Einschätzung milder depressiver Symptomatik bei klinisch unauffälligen Probanden eingesetzt werden sollte. Dieser ‘blinde Fleck’ des BDI-II bleibt bei der Evaluation der Reliabilität in der KTT hingegen verborgen.

Während die KTT den Fragebogen primär als Gesamtes evaluiert, liegt der Fokus der IRT, wie der Name bereits nahelegt, auf den einzelnen Items eines Fragebogens. Die Analyse der Informationsstruktur der einzelnen Items des BDI-II zeigt, dass insbesondere das Item ‘Wertlosigkeit’ ein besonders ausgeprägtes Diskriminationspotential aufwies. Während die Zustimmung eines Probanden zu starker ‘Veränderung des Appetits’ mit einer sehr ausgeprägten depressiven Symptomatik einherging, war die Zustimmung zu starker ‘Ermüdung’ nur mit einer niedrigen Depressivität verknüpft.

Die Informationsstruktur der Items stellt in der IRT einen ganz zentralen Aspekt dar, da – anders als in der KTT, bei der alle Items zu gleichem Anteil in den Summenscore eingehen – die Items in der IRT nach Itemparametern gewichtet in den Trait-Score eingehen. Somit wird in der IRT die im Antwortmuster enthaltene Information in der

Scorebildung berücksichtigt und erlaubt somit eine noch differenziertere Einschätzung der depressiven Symptomatik, als dies mit dem klassischen Summenscore möglich ist.
Um das BDI-II sinnvoll in der Praxis einsetzen zu können, muss es möglich sein, anhand des BDI-II-Testscores die Schwere der depressiven Symptomatik zwischen Probanden vergleichen zu dürfen. Dies ist jedoch nur dann zulässig, wenn das Antwortverhalten der Probanden auf die BDI-II-Items unabhängig von jeglichen demographischen Eigenschaften ist. In der vorliegenden Stichprobe wurde dies für die beiden Eigenschaften Alter und Geschlecht untersucht. Zusammenfassend zeigte sich die Bewertung der Depressivität der Probanden dabei unabhängig von den untersuchten Variablen und erlaubt es somit in der Praxis, die Depressivität von Probanden anhand ihres BDI-II-Testscores zu vergleichen, auch wenn die Probanden ein unterschiedliches Geschlecht und Alter aufweisen. Einschränkend muss hier jedoch beachtet werden, dass das Symptom ‘Weinen’ zwischen den Geschlechtern abhängig von der Depressivität unterschiedlich wahrgenommen wurde. Der verzerrende Effekt war jedoch nicht stark genug, um bei Anwendung des gesamten BDI-II zu einer signifikant unterschiedlichen Einschätzung der Depressivität der Probanden zu gelangen.

Fazit und Ausblick:

Die IRT kann einen wichtigen Beitrag dazu leisten, das Diskriminationspotential der Fragebogendiagnostik weiter zu verbessern, sowie zusätzlich Informationen bezüglich aberrantem Antwortverhalten, sowie DIF für die Praxis nutzbar zu machen.
Konkret für das BDI-II bedeutet dies, dass, ohne zusätzliche Belastung des Probanden, durch Anwendung der IRT die Diskrimination des Schweregrads depressiver Symptomatik mittels BDI-II verbessert werden kann. Ein automatisches Feedback-System kann zudem dazu beitragen, durch die Detektion aberranten Antwortverhaltens im BDI-II Personen zu identifizieren, deren BDI-II-Score keine zuverlässige Einschätzung der depressiven Symptomatik erlaubt und kann dem behandelnden Arzt so hilfreiche Informationen und Unterstützung in der klinischen Entscheidungsfindung bieten. Die Analyse auf DIF ist insbesondere im Hinblick auf die zunehmende Anwendung von CATs ein wichtiger Baustein, um eine sichere Diagnostik mittels Fragebögen zu ermöglichen. Hier zeigte insbesondere das Item ‚Weinen‘ Verzerrungspotential zwischen den Geschlechtern und sollte entsprechend für CATs nicht unkritisch eingesetzt werden.
Bevor die IRT gewinnbringend auch für die klinische Praxis eingesetzt werden kann, besteht jedoch noch weiterer Forschungsbedarf, der sich insbesondere auf den Nachweis der klinischen Relevanz des mittels IRT zusätzlich ermöglichten Diskriminationspotentials des BDI-II, sowie auf die Generierung IRT-basierter Grenzwerte zur Einschätzung der Depressionsschwere konzentriert.

7. Literaturverzeichnis

1. Hautzinger M, Keller F, Kühner C. BDI-II. Beck-Depressions-Inventar Revision. Frankfurt: Harcourt Test Services; 2006.
2. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation;1996.
3. Wittchen H-U. Depressive Erkrankungen. Berlin: Robert-Koch-Inst; 2010.
4. Deutsche Gesellschaft Für Psychiatrie, Psychotherapie Und Nervenheilkunde, Ärztliches Zentrum Für Qualität In Der Medizin (ÄZQ). S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression - Langfassung, 2. Auflage: Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN); Bundesärztekammer (BÄK); Kassenärztliche Bundesvereinigung (KBV); Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF); 2015.
5. Nowotny M, Kern D, Breyer E, Bengough T, Griebler R, (Hg.). Depressionsbericht Österreich: Eine interdisziplinäre und multiperspektivische Bestandsaufnahme. Bundesministerium für Arbeit, Soziales, Gesundheit und Konsumentenschutz. Wien, 2019.
6. Rusch T, Lowry PB, Mair P, Treiblmaier H. Breaking Free from the Limitations of Classical Test Theory: Developing and Measuring Information Systems Scales Using Item Response Theory. Inf Manag. 2017;54:189–203.
7. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. J Stat Softw. 2011;39:1–30.
8. Alexandrowicz RW, Fritzsche S, Keller F. Die Anwendbarkeit des BDI-II in klinischen und nicht-klinischen Populationen aus psychometrischer Sicht: Eine vergleichende Analyse mit dem Rasch-Modell. Neuropsychiatr. 2014;28:63–73.
9. Beck AT, Ward CH, Mendelson M, Mock M, Erbaugh J. An Inventory for Measuring Depession. Arch Gen Psychiatry. 1961:561–71.
10. Williams ZJ, Everaert J, Gotham KO. Measuring Depression in Autistic Adults: Psychometric Validation of the Beck Depression Inventory-II. Assessment. 2020:1-19.
11. Moosbrugger H, Kelava A. Testtheorie und Fragebogenkonstruktion. 3rd ed. Berlin, Heidelberg: Springer-Verlag; 2020.
12. Lienert GA, Raatz U. Testaufbau und Testanalyse. 6th ed. Weinheim: Beltz; 1998.
13. McDonald RP. Modern Test Theory. In: Little TD, editor. The Oxford Handbook of Quantitative Methods, Volume 1: Foundations. Oxford: Oxford University Press USA; 2013. p. 118–143.
14. Paek I, Cole K. Using R for item response theory model applications. Abingdon, Oxon, New York, NY: Routledge; 2020.
15. de Ayala RJ. The IRT Tradition and its Applications. In: Little TD, editor. The Oxford Handbook of Quantitative Methods, Volume 1: Foundations. Oxford: Oxford University Press USA; 2013. p. 144–169.
16. Gulliksen H. Theory of Mental Tests. New York: Wiley. 1950.
17. Lord FM, Novick MR, editors. Statistical theories of mental test scores. Reading,MA: Addison-Wesley; 1968.
18. Hambleton RK, Jones RW. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educ Meas. 1993;12:38–47.
19. Glischinski M von, Brachel R von, Hirschfeld G. How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). Qual Life Res. 2019;28:1111–8.
20. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.

21. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR, editors. Statistical theories of mental test scores. Reading,MA: Addison-Wesley; 1968. p. 397–479.
22. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45:22-31.
23. Ranger J, Much S. Analyzing the Fit of IRT Models With the Hausman Test. Front Psychol. 2020;11:Article 149.
24. Embretson SE, Reise SP. Item response theory. New Jersey: Lawrence Erlbaum Associaties; 2000.
25. Reckase MD. Multidimensional Item Response Theory. New York, NY: Springer New York; 2009.
26. Masters GN. A rasch model for partial credit scoring. Psychometrika. 1982;47:149–74.
27. Muraki E. A generalized partial credit model: Application of an EM algorithm. Appl Psychol Meas. 1992;14:59–71.
28. Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometrika Monographs. 1969;34.
29. García-Pérez MA. An Analysis of (Dis)Ordered Categories, Thresholds, and Crossings in Difference and Divide-by-Total IRT Models for Ordered Responses. Span J Psychol. 2017;20:1-27.
30. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. 1981;46:443–59.
31. Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess. 2014;26:513–27.
32. Zhao Y, Chan W, Lo BCY. Comparing five depression measures in depressed Chinese patients using item response theory: an examination of item properties, measurement precision and score comparability. Health Qual Life Outcomes. 2017;15:60.
33. Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014;67:73–86.
34. Wang Y-P, Gorenstein C. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. Braz J Psychiatry. 2013;35:416–31.
35. de Sá Junior AR, de Andrade AG, Andrade LH, Gorenstein C, Wang Y-P. Response pattern of depressive symptoms among college students: What lies behind items of the Beck Depression Inventory-II? J Affect Disord. 2018;234:124–30.
36. de Sá Junior AR, Liebel G, de Andrade AG, Andrade LH, Gorenstein C, Wang Y-P. Can Gender and Age Impact on Response Pattern of Depressive Symptoms Among College Students? A Differential Item Functioning Analysis. Front Psychiatry. 2019;10:1–11.
37. Subica AM, Fowler JC, Elhai JD, Frueh BC, Sharp C, Kelly EL, Allen JG. Factor structure and diagnostic validity of the Beck Depression Inventory-II with adult clinical inpatients: comparison to a gold-standard diagnostic interview. Psychol Assess. 2014;26:1106–15.
38. Faro A, Pereira CR. Factor structure and gender invariance of the Beck Depression Inventory – second edition (BDI-II) in a community-dwelling sample of adults. Health Psychol Behav Med. 2020;8:16–31.
39. Nuevo R, Dunn G, Dowrick C, Vázquez-Barquero JL, Casey P, Dalgard OS, et al. Cross-cultural equivalence of the Beck Depression Inventory: a five-country analysis from the ODIN study. J Affect Disord. 2009;114:156–62.
40. Kuehner C. Gender differences in unipolar depression: an update of epidemiological findings and possible explanations. Acta Psychiatr Scand. 2003;108:163–74.

41. Wardenaar KJ, Wanders RBK, Roest AM, Meijer RR, de Jonge P. What does the beck depression inventory measure in myocardial infarction patients? a psychometric approach using item response theory and person-fit. Int J Methods Psychiatr Res. 2015;24:130–42.
42. Kowarik A, Templ M. Imputation with the R Package VIM. J Stat Softw. 2016;74:1–16.
43. Kindt T, Rabkow N, Pukas L, Keuch L, Sapalidis A, Piloty-Leskien A, et al. A Comparison of Depressive Symptoms in Medical and Psychology Students in Germany – Associations with Potential Risk and Resilience Factors. JMP. 2021:1–13.
44. Ehring E, Frese T, Fuchs S, Dudo K, Pukas L, Stoevesandt D, Watzke S. Asking future doctors: what support options do medical students want to cope with medical school? J Public Health (Berl.). 2021.
45. Rabkow N, Pukas L, Sapalidis A, Ehring E, Keuch L, Rehnisch C, et al. Facing the truth - A report on the mental health situation of German law students. Int J Law Psychiatry. 2020;71.
46. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. https://www.R-project.org/.
47. Revelle W. psych: Procedures for Personality and Psychological Research (Version 1.9.12). 2019. https://CRAN.R-project.org/package=psych. Accessed 4 Apr 2021.
48. Chalmers RP. mirt : A Multidimensional Item Response Theory Package for the R Environment. J Stat Softw. 2012;48:1–29.
49. Mazza A, Punzo A, McGuire B. KernSmoothIRT: An R Package for Kernel Smoothing in Item Response Theory. J Stat Softw. 2014;58:1–34.
50. Peters G-JY. Userfriendlyscience (UFS). OSF. (20.08.2019). https://doi.org/10.17605/OSF.IO/TXEQU. Accessed 4 Apr 2021.
51. Zhang Z, Yuan K-H. Robust Coefficients Alpha and Omega and Confidence Intervals With Outlying Observations and Missing Data: Methods and Software. Educ Psychol Meas. 2016;76:387–411.
52. HealthMeasures. PROMIS® Instrument Development and Validation: Scientific Standards Version 2.0. 2013. https://www.healthmeasures.net/images/PROMIS/PROMISStandards_Vers2.0_Final.pdf. Accessed 4 Apr 2021.
53. Hedderich J, Sachs L. Angewandte Statistik: Methodensammlung mit R. 16th ed.: Springer Spektrum; 2018.
54. Trizano-Hermosilla I, Alvarado JM. Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. Front Psychol. 2016;7:769.
55. Bock RD, Gibbons R, Muraki E. Full-Information Item Factor Analysis. Appl Psychol Meas. 1988;12:261-280.
56. Mair P. Modern Psychometrics with R. Cham: Springer International Publishing; 2018.
57. Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. J Pers Assess. 2010;92:544–59.
58. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Edu Modeling. 1999;6:1–55.
59. Cangur S, Ercan I. Comparison of Model Fit Indices Used in Structural Equation Modeling Under Multivariate Normality. J Mod Stat Methods. 2015;14:152–67.
60. Rodriguez A, Reise SP, Haviland MG. Evaluating bifactor models: Calculating and interpreting statistical indices. Psychol Methods. 2016;21:137–50.
61. Douglas J, Cohen A. Nonparametric Item Response Function Estimation for Assessing Parametric Model Fit. Appl Psychol Meas. 2001;25:234–43.
62. Maydeu-Olivares A, García-Forero C. Goodness-of-Fit Testing. International Encyclopedia of Education. 2010;7:190–6.
63. Maydeu-Olivares A, Joe H. Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. Psychometrika. 2006;71:713–32.

64. Cai L, Hansen M. Limited-information Goodness-of-fit Testing of Hierarchical Item Factor Models. Br J Math Stat Psychol. 2013;66:245–76.
65. Felt JM, Castaneda R, Tiemensma J, Depaoli S. Using Person Fit Statistics to Detect Outliers in Survey Research. Front Psychol. 2017;8:1–9.
66. Merkle EC, You D, Preacher KJ. Testing nonnested structural equation models. Psychol Methods. 2016;21:151–63.
67. Orlando M, Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. Appl Psychol Meas. 2000;24:50–64.
68. Kang T, Chen TT. An Investigation of the Performance of the Generalized S-X2 Item-Fit Index for Polytomous IRT Models. J Educ Meas. 2008;45:391–406.
69. Thissen D, Pommerich M, Billeaud K, Williams VSL. Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses. Appl Psychol Meas. 1995;19:39–49.
70. Benjamini Y, Hochberg Y. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. J Educ Behav Stat. 2000;25:60–83.
71. Brady KJS, Ni P, Sheldrick RC, Trockel MT, Shanafelt TD, Rowe SG, et al. Describing the emotional exhaustion, depersonalization, and low personal accomplishment symptoms associated with Maslach Burnout Inventory subscale scores in US physicians: an item response theory analysis. J Patient Rep Outcomes. 2020;4:42.
72. Drasgow F, Levine MV, Williams EA. Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices. Br J Math Stat Psychol. 1985;38.
73. Chen W-H, Thissen D. Local Dependence Indexes for Item Pairs Using Item Response Theory. J Educ Behav Stat. 1997;22:265–89.
74. Yen WM. Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. Appl Psychol Meas. 1984;8:125–45.
75. Edwards MC, Houts CR, Cai L. A diagnostic procedure to detect departures from local independence in item response theory models. Psychol Methods. 2018;23:138–49.
76. Yen WM. Scaling Performance Assessments: Strategies for Managing Local Item Dependence. J Educ Meas. 1993;30:187–213.
77. Christensen KB, Makransky G, Horton M. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. Appl Psychol Meas. 2017;41:178–94.
78. Rajlic G. Visualizing Items and Measures: An Overview and Demonstration of the Kernel Smoothing Item Response Theory Technique. Quant Method Psychol. 2020;16:363–75.
79. Lee Y-S, Wollack JA, Douglas J. On the Use of Nonparametric Item Characteristic Curve Estimation Techniques for Checking Parametric Model Fit. Educ Psychol Meas. 2009;69:181–97.
80. Ramsay JO. Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika. 1991;56:611–30.
81. Baker FB. The basics of item response theory. 2nd ed. College Park Md.: ERIC Clearinghouse on Assessment and Evaluation; 2001.
82. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Appl Psychol Meas. 1982;6:431–44.
83. Thissen D. Reliability and measurement precision. In: Wainer H, editor. Computerized adaptive testing: A primer. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2000. p. 159–183.
84. Castro SMJ, Trentini C, Riboldi J. Item response theory applied to the Beck Depression Inventory. Rev Bras Epidemiol. 2010;13:1–13.
85. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques: DIFdetect and difwithpar. Med Care. 2006.

86. Rogers JH, Swaminathan H. A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning. Appl Psychol Meas. 1993;17:105–16.
87. Bulut O, Suh Y. Detecting Multidimensional Differential Item Functioning with the Multiple Indicators Multiple Causes Model, the Item Response Theory Likelihood Ratio Test, and Logistic Regression. Front. Educ. 2017;2.
88. Jodoin MG, Gierl MJ. Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. Appl Meas Educ. 2001;14:329–49.
89. Desjardins CD, Bulut O. Handbook of educational measurement and psychometrics using R. Boca Raton, London, New York: CRC Press; 2018.
90. Beck AT, Brown GK, Steer RA. Beck-Depressions-Inventar-FS (BDI-FS). Manual. Deutsche Bearbeitung von Sören Kliem & Elmar Bräher. Frankfurt am Main: Pearson Assessment; 2013.
91. Beesdo-Baum K, Knappe S, Einsle F, Knothe L, Wieder G, Venz J, et al. Wie häufig werden Patienten mit depressiven Störungen in der hausärztlichen Praxis erkannt? : Eine epidemiologische Querschnittsstudie. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2018;61:52–64.
92. Becker N, Abholz H-H. Prävalenz und Erkennen von depressiven Störungen in deutschen Allgemeinarztpraxen - eine systematische Literaturübersicht. Z Allg Med. 2005;81:474–81.
93. Danner D, Rammstedt B, Bluemke M, Treiber L, Berres S, Soto C, John O. Die deutsche Version des Big Five Inventory 2 (BFI-2): ZIS - GESIS Leibniz Institute for the Social Sciences; 2016.
94. Cho S-J, Suh Y, Lee W-Y. After Differential Item Functioning Is Detected: IRT Item Calibration and Scoring in the Presence of DIF. Appl Psychol Meas. 2016;40:573–91. doi:10.1177/0146621616664304.
95. Santor DA, Ramsay JO, Zuroff DC. Nonparametric Item Analyses of the Beck Depression Inventory: Evaluating Gender Item Bias and Response Option Weights. Psychol Assess. 1994;6:255–70.
96. Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale. J Clin Epidemiol. 2000;53:285–9.
97. Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychol Sci Q. 2009;51:148–80.
98. Vingerhoets AJJM, Rottenberg J, Cevaal A, Nelson JK. Is there a relationship between depression and crying? A review. Acta Psychiatr Scand. 2007;115:340–51.
99. Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF. Differential Functioning of the Beck Depression Inventory in Late-Life Patients: Use of Item Response Theory. Psychol Aging. 2002;17:379–91.
100. Whooley MA, Avins AL, Miranda J, Browner WS. Case-Finding Instruments for Depression: Two Questions Are as Good as Many. J Gen Intern Med. 12;1997:439–45.
101. Conijn JM, Spinhoven P, Meijer RR, Lamers F. Person misfit on the Inventory of Depressive Symptomatology: Low quality self-report or true atypical symptom profile? Int J Methods Psychiatr Res. 2017;26.
102. Paykel ES. Depressive typologies and response to amitriptyline. Br J Psychiatry. 1972;120:147–56.
103. Wanders RBK, Meijer RR, Ruhé HG, Sytema S, Wardenaar KJ, de Jonge P. Person-fit feedback on inconsistent symptom reports in clinical depression care. Psychol Med. 2018;48:1844–52.

8. Thesen der Dissertation

1. Die Faktorenstruktur der deutschen Version des Beck Depressions-Inventars-II lässt sich am besten durch das Bifaktor-Modell mit dem Generalfaktor ‚Depressivität' und den drei Gruppenfaktoren ‚kognitiv', ‚somatisch' und ‚affektiv' beschreiben. Die Depressivität eines Probanden lässt sich beim BDI-II folglich am treffendsten durch einen einzigen Score – den klassischen Summenscore oder einen unidimensionalen Trait-Score – ausdrücken.
2. Unidimensionale parametrische IRT-Modelle, insbesondere das Graded Response Modell nach Samejima, sind gut dafür geeignet, den vorliegenden Datensatz, der sich aus den Antworten einer Stichprobe junger Menschen in der Ausbildungsphase auf den BDI-II zusammensetzt, abzubilden. Diese hohe Anpassungsgüte ermöglicht es, auf Grundlage der IRT inferenzstatistische Aussagen über die zugrunde liegende Depressivität von Probanden ableiten zu können.
3. Mittels IRT können Personen, die – beispielsweise durch Manipulationsversuche in Form von Simulation/ Dissimulation – ein aberrantes Antwortmuster im BDI-II aufweisen und deren Schweregrad der depressiven Symptomatik entsprechend nicht anhand des BDI-II-Ergebnisses eingeschätzt werden sollte, identifiziert werden.
4. Das BDI-II eignet sich sowohl als Screening-Instrument in der Allgemeinbevölkerung, als auch als Instrument zur Einschätzung des Schweregrads depressiver Symptomatik in klinischen Populationen.
5. Die Frage nach dem Symptom der ‚Wertlosigkeit' weist an allen drei klinischen Cut-off-Scores einen sehr hohen Anteil an der Testinformation auf und kann in diesen Bereichen am zuverlässigsten diskriminieren.
6. Probanden im Altersbereich von 18 bis 56 Jahren weisen abhängig von ihrem Lebensalter kein klinisch relevant unterschiedliches Antwortverhalten im BDI-II auf. Folglich kann die Depressivität von Probanden in diesem Altersbereich mittels BDI-II-Testscore miteinander verglichen werden.
7. Das Symptom ‚Weinen' zeigt zwischen den Geschlechtern ein systematisch unterschiedliches Antwortverhalten auf. Für männliche Probanden wird die Depressivität anhand der Frage nach dem Symptom ‚Weinen' im niedrigen Ausprägungsgrad tendenziell unterschätzt, während sich dieses Verhältnis in höheren Ausprägungsbereich umkehrt, führt bei Anwendung des gesamten BDI-II jedoch zu keiner signifikant unterschiedlichen Einschätzung der Depressivität der Probanden.
8. Die aus der IRT resultierenden Erkenntnisse erlauben es, den Goldstandard BDI-II noch informativer einzusetzen, die Ausprägung der Depressivität genauer einzuschätzen, aberrantes Antwortverhalten zu detektieren und für DIF anzupassen.

ANHANG

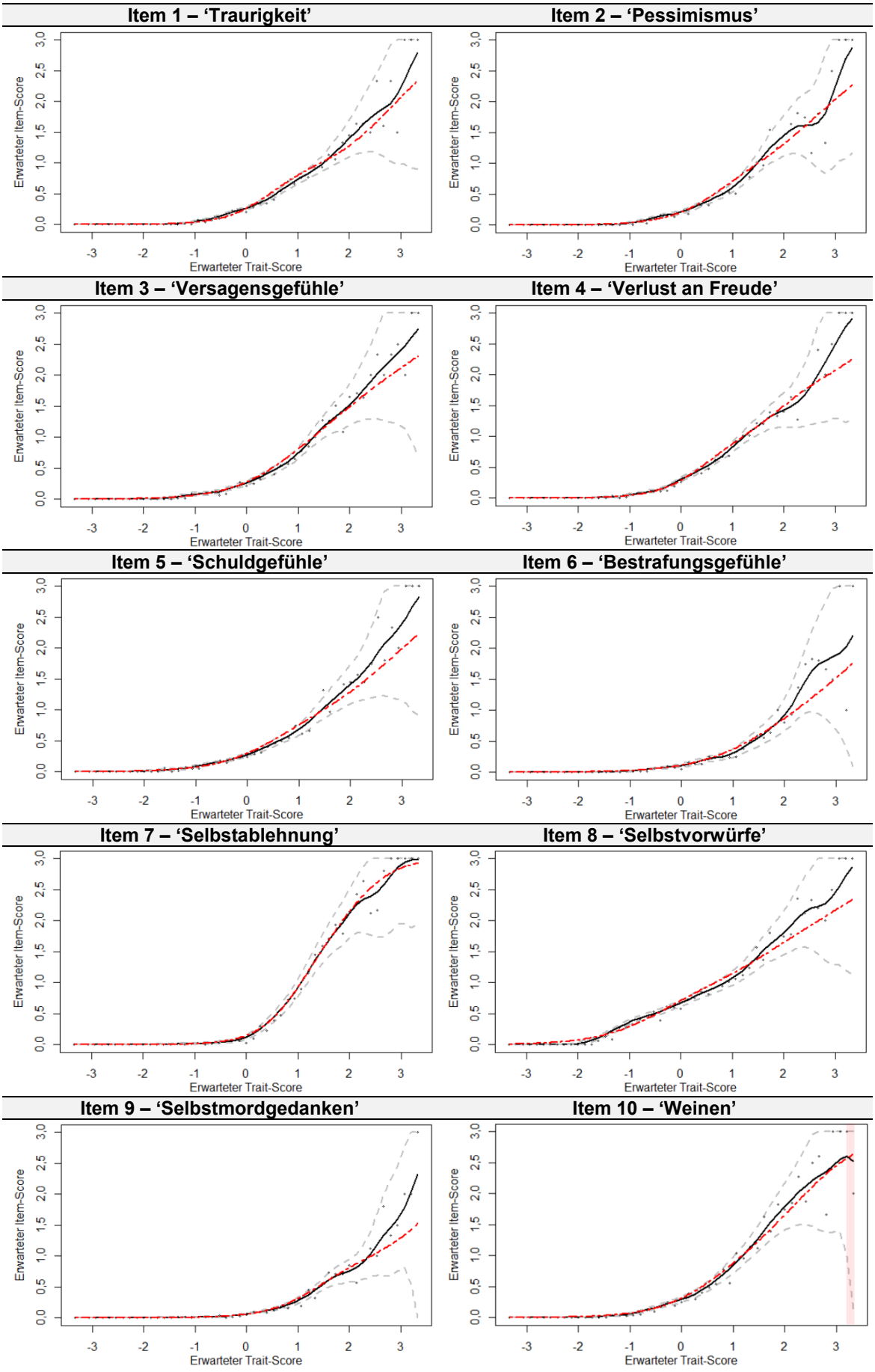
Anhang 1: Deskriptive Charakteristika des BDI-II

Item	Symptom	Kategorie 0		Kategorie 1		Kategorie 2		Kategorie 3		M	SD	r_{pbis}	α	Omega	GLB
		[%]	[abs.]	[%]	[abs.]	[%]	[abs.]	[%]	[abs.]						
BDI 1	Traurigkeit	64,3	1556	31,9	771	3,0	73	0,8	19	0,40	0,59	0,64	0,91	0,91	0,90
BDI 2	Pessimismus	70,1	1696	24,6	596	4,2	101	1,1	26	0,36	0,62	0,60	0,91	0,91	0,90
BDI 3	Versagensgefühle	68,1	1648	21,9	530	8,6	208	1,4	33	0,43	0,71	0,56	0,91	0,91	0,90
BDI 4	Verlust an Freude	63,4	1534	29,7	719	6,2	150	0,7	16	0,44	0,64	0,64	0,91	0,91	0,90
BDI 5	Schuldgefühle	65,5	1585	28,9	699	3,8	91	1,8	44	0,42	0,65	0,56	0,91	0,91	0,90
BDI 6	Bestrafungsgefühle	83,3	2015	13,3	322	2,1	50	1,3	32	0,21	0,54	0,46	0,91	0,91	0,91
BDI 7	Selbstablehnung	75,4	1825	9,4	227	10,2	247	5,0	120	0,45	0,87	0,63	0,91	0,91	0,92
BDI 8	Selbstvorwürfe	39,5	956	48,0	1161	10,7	258	1,8	44	0,75	0,72	0,56	0,91	0,91	0,92
BDI 9	Selbstmordgedanken	**84,2**	**2036**	14,8	359	0,5	12	0,5	12	0,17	0,43	0,51	0,91	0,91	0,91
BDI 10	Weinen	66,3	1603	24,3	587	5,3	128	4,2	101	0,47	0,78	0,57	0,91	0,91	0,91
BDI 11	Unruhe	51,1	1235	43,2	1044	3,7	90	2,1	50	0,57	0,67	0,47	0,91	0,91	0,91
BDI 12	Interessenlosigkeit	65,6	1586	29,5	714	3,8	91	1,2	28	0,41	0,62	0,60	0,91	0,91	0,89
BDI 13	Entscheidungsunfähigkeit	60,3	1458	30,1	727	6,2	149	3,5	85	0,53	0,76	0,55	0,91	0,91	0,92
BDI 14	Wertlosigkeit	77,3	1870	12,4	301	9,0	218	1,2	30	0,34	0,69	0,67	0,91	0,91	0,92
BDI 15	Verlust an Energie	39,5	956	47,3	1143	11,9	289	1,3	31	0,75	0,71	0,64	0,91	0,91	0,92
BDI 16	Schlafgewohnheiten	29,2	707	64,0	1547	6,3	152	0,5	13	0,78	0,57	0,37	0,91	0,91	0,91
BDI 17	Reizbarkeit	53,2	1287	37,5	906	6,9	168	2,4	58	0,59	0,73	0,55	0,91	0,91	0,91
BDI 18	Veränderung des Appetits	51,0	1234	45,1	1092	3,7	89	**0,2**	**4**	0,53	0,58	0,42	0,91	0,91	0,91
BDI 19	Konzentrationsschwierigkeiten	51,6	1247	32,0	773	15,4	373	1,1	26	0,66	0,77	0,56	0,91	0,91	0,91
BDI 20	Ermüdung	39,0	943	47,7	1154	10,3	250	3,0	72	0,77	0,75	0,58	0,91	0,91	0,92
BDI 21	Libidoverlust	76,7	1856	16,7	405	5,3	128	1,2	30	0,31	0,63	0,40	0,91	0,91	0,92
BDI-II										10,35	8,53	-	0,91	0,91	0,91

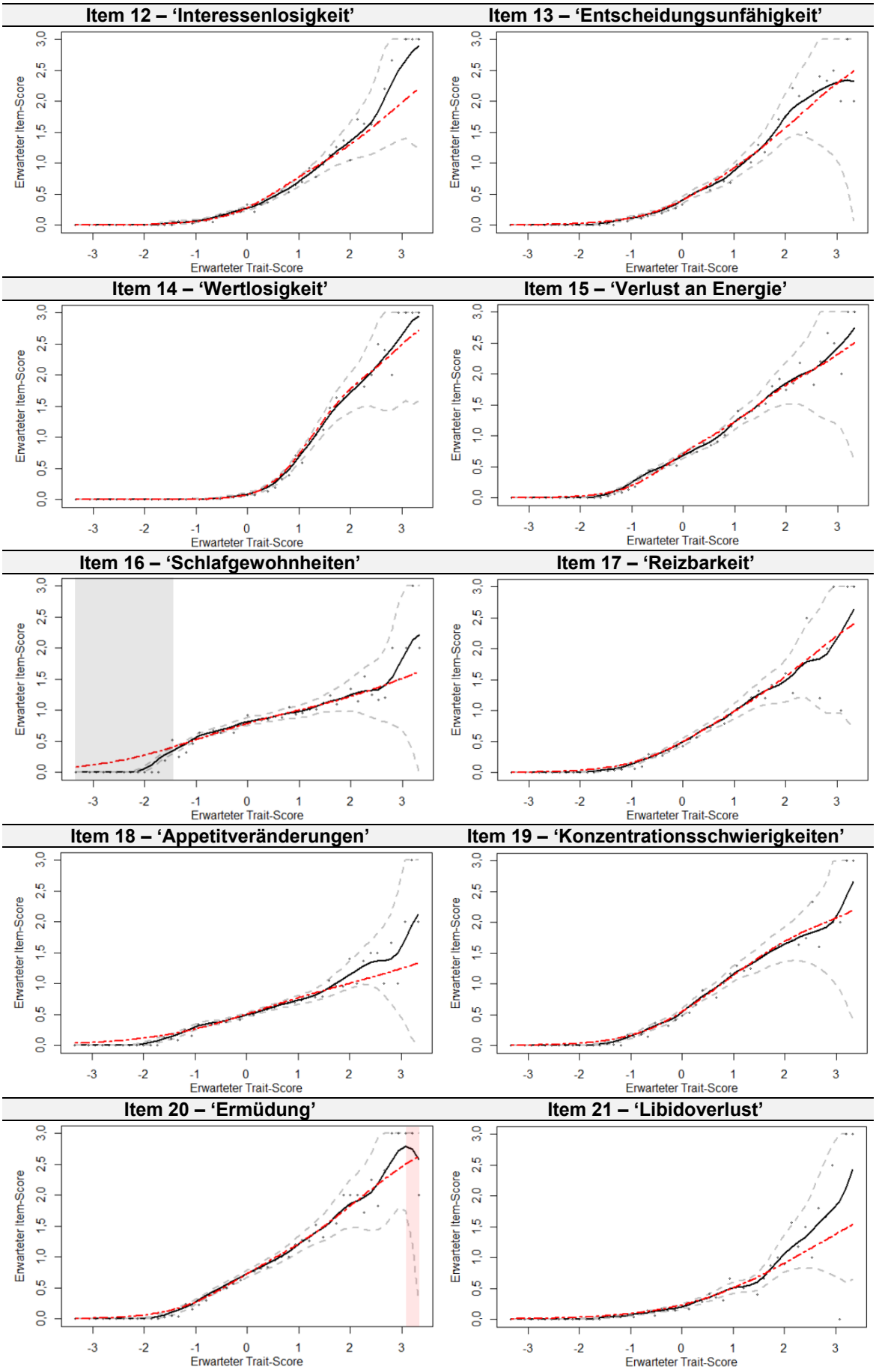
M= Mittelwert, SD= Standardabweichung, r_{pbis} = Point biserial correlation,α/ Omega/ GLB= Cronbach α/ Omega/ GLB, wenn das jeweilige Item entfernt wird

Anhang 2: Q3-Test auf lokale Unabhängigkeit

	BDI1	BDI2	BDI3	BDI4	BDI5	BDI6	BDI7	BDI8	BDI9	BDI10	BDI11	BDI12	BDI13	BDI14	BDI15	BDI16	BDI17	BDI18	BDI19	BDI20	BDI21	
BDI1		-0,02	-0,06	-0,09	-0,05	0,02	-0,06	-0,11	0,05	0,14	-0,07	-0,15	-0,04	-0,02	-0,10	-0,06	-0,03	-0,06	-0,07	-0,14	-0,08	**BDI1**
BDI2			0,04	-0,11	-0,04	-0,06	0,01	-0,02	-0,02	-0,08	-0,04	-0,06	0,00	-0,06	-0,06	-0,04	-0,11	-0,09	-0,04	-0,12	-0,06	**BDI2**
BDI3				-0,08	0,08	0,03	0,10	-0,02	0,00	-0,11	-0,12	-0,10	-0,09	0,07	-0,13	-0,05	-0,13	-0,04	-0,04	-0,13	-0,07	**BDI3**
BDI4					-0,08	-0,05	-0,11	-0,07	-0,04	-0,05	-0,09	0,24	-0,10	-0,16	-0,05	-0,07	-0,03	-0,05	-0,07	-0,08	0,05	**BDI4**
BDI5						0,09	0,02	0,06	0,01	-0,12	-0,10	-0,11	0,01	-0,01	-0,14	-0,08	-0,08	-0,03	-0,05	-0,13	-0,03	**BDI5**
BDI6							-0,04	0,00	0,00	-0,07	-0,01	-0,09	-0,08	-0,01	-0,11	-0,02	0,00	-0,03	-0,02	-0,04	-0,05	**BDI6**
BDI7								0,10	0,06	-0,05	-0,15	-0,14	-0,10	0,20	-0,11	-0,05	-0,21	-0,09	-0,15	-0,19	-0,11	**BDI7**
BDI8									-0,01	-0,04	-0,03	-0,07	-0,01	0,02	-0,15	-0,06	-0,07	-0,05	-0,16	-0,15	-0,05	**BDI8**
BDI9										0,00	-0,10	-0,05	-0,06	0,08	-0,11	-0,04	-0,09	-0,07	-0,13	-0,15	-0,09	**BDI9**
BDI10											-0,02	-0,11	-0,02	-0,06	-0,09	-0,01	0,00	0,01	-0,08	-0,05	0,00	**BDI10**
BDI11												-0,02	0,00	-0,11	-0,05	0,04	0,14	0,01	-0,02	0,02	0,01	**BDI11**
BDI12													-0,04	-0,12	-0,03	-0,06	-0,01	-0,04	-0,05	-0,01	0,09	**BDI12**
BDI13														-0,08	-0,04	-0,05	-0,06	-0,04	0,03	-0,06	-0,03	**BDI13**
BDI14															-0,18	-0,11	-0,16	-0,05	-0,12	-0,16	-0,09	**BDI14**
BDI15																0,02	-0,03	-0,04	0,06	0,16	-0,03	**BDI15**
BDI16																	-0,01	0,13	-0,02	0,10	-0,03	**BDI16**
BDI17																		-0,02	-0,05	0,06	0,06	**BDI17**
BDI18																			0,00	0,05	0,03	**BDI18**
BDI19																				0,09	-0,02	**BDI19**
BDI20																					0,03	**BDI20**
BDI21																						**BDI21**



Anhang 3: Vergleich parametrischer und non-parametrischer IRFs – Teil 1



Anhang 4: Vergleich parametrischer und non-parametrischer IRFs – Teil 2

Anhang 5: Vergleich zwischen Kenngrößen der IRT und der KTT

Item	Symptom	IRT-Analyse									KTT-Analyse	
		α_i	SE	δ_{i1}	SE	δ_{i2}	SE	δ_{i3}	SE	LP	r_{pbis}	m_{Item}
BDI 1	Traurigkeit	2,198	0,102	0,479	0,035	2,447	0,088	3,385	0,155	2,104	0,684	0,384
BDI 2	Pessimismus	1,940	0,091	0,715	0,040	2,330	0,088	3,527	0,171	2,191	0,653	0,341
BDI 3	Versagensgefühle	1,658	0,079	0,698	0,042	1,937	0,078	3,530	0,170	2,055	0,629	0,407
BDI 4	Verlust an Freude	2,190	0,098	0,429	0,034	1,990	0,070	3,699	0,196	2,039	0,694	0,425
BDI 5	Schuldgefühle	1,624	0,078	0,582	0,041	2,526	0,103	3,430	0,159	2,180	0,624	0,396
BDI 6	Bestrafungsgefühle	1,417	0,084	1,548	0,075	3,153	0,160	4,009	0,231	2,904	0,509	0,196
BDI 7	Selbstablehnung	2,397	0,112	0,881	0,038	1,346	0,048	2,186	0,076	1,471	0,705	0,418
BDI 8	Selbstvorwürfe	1,583	0,069	-0,398	0,040	1,765	0,071	3,535	0,161	1,634	0,632	0,730
BDI 9	Selbstmordgedanken	1,948	0,110	1,416	0,058	3,594	0,182	4,120	0,250	3,043	0,568	0,153
BDI 10	Weinen	1,663	0,077	0,603	0,041	1,985	0,079	2,681	0,110	1,756	0,639	0,454
BDI 11	Unruhe	1,319	0,066	0,029	0,042	2,792	0,129	3,841	0,200	2,221	0,532	0,551
BDI 12	Interessenlosigkeit	1,841	0,086	0,536	0,038	2,412	0,093	3,602	0,174	2,183	0,648	0,390
BDI 13	Entscheidungsunfähigkeit	1,540	0,071	0,385	0,040	2,082	0,085	2,983	0,128	1,817	0,618	0,506
BDI 14	Wertlosigkeit	2,787	0,135	0,911	0,037	1,574	0,052	2,976	0,121	1,821	0,725	0,315
BDI 15	Verlust an Energie	2,172	0,090	-0,339	0,034	1,445	0,052	3,281	0,144	1,462	0,704	0,739
BDI 16	Schlafgewohnheiten	1,030	0,058	-1,050	0,069	3,079	0,163	5,837	0,417	2,622	0,434	0,775
BDI 17	Reizbarkeit	1,555	0,071	0,110	0,038	2,043	0,083	3,246	0,148	1,800	0,61	0,570
BDI 18	Veränderung des Appetits	0,995	0,059	0,055	0,050	3,912	0,225	7,659	0,817	3,876	0,467	0,518
BDI 19	Konzentrationsschwierigkeiten	1,617	0,071	0,075	0,037	1,471	0,062	4,002	0,213	1,849	0,626	0,641
BDI 20	Ermüdung	1,726	0,073	-0,402	0,038	1,587	0,062	2,891	0,117	1,359	0,653	0,764
BDI 21	Libidoverlust	1,017	0,066	1,419	0,088	3,147	0,187	5,290	0,371	3,285	0,459	0,291

α_i = Diskriminationsparameter der IRT, $\delta_{1\text{-}3}$ = Schwellenparameter, LP= Locationparameter, r_{pbis} = Point biserial correlation/ Diskriminationsparameter KTT, m_{Item} = item mean score/ Schwierigkeitsparameter KTT

Anhang 6: DIF-Analyse bezüglich Alter und Geschlecht

	DIF bezüglich Alter						DIF bezüglich Geschlecht					
	$Pr(X_{13}^2,2)$	$Pr(X_{12}^2,1)$	$Pr(X_{23}^2,1)$	$\Delta R_{13}^2{}_{McFadden}$	$\Delta R_{12}^2{}_{McFadden}$	$\Delta R_{23}^2{}_{McFadden}$	$Pr(X_{13}^2,2)$	$Pr(X_{12}^2,1)$	$Pr(X_{23}^2,1)$	$\Delta R_{13}^2{}_{McFadden}$	$\Delta R_{12}^2{}_{McFadden}$	$\Delta R_{23}^2{}_{McFadden}$
BDI 1	0,0032	0,0137	0,0197	0,0039	0,0021	0,0019	**0,0001**	**0,0000**	0,6851	0,0055	0,0055	0,0000
BDI 2	0,1630	0,0667	0,6071	0,0011	0,0010	0,0004	0,2449	0,1042	0,6761	0,0008	0,0008	0,0001
BDI 3	0,5816	0,3299	0,7136	0,0003	0,0003	0,0000	**0,0000**	**0,0000**	0,6432	0,0057	0,0057	0,0001
BDI 4	0,1261	0,0546	0,5040	0,0011	0,0010	0,0001	0,0025	**0,0008**	0,3840	0,0032	0,0030	0,0004
BDI 5	0,5161	0,2698	0,7459	0,0005	0,0004	0,0000	0,7182	0,9677	0,4164	0,0002	0,0000	0,0002
BDI 6	0,3539	0,1540	0,8323	0,0011	0,0010	0,0000	**0,0020**	0,0113	0,0141	0,0051	0,0026	0,0025
BDI 7	0,1106	0,0432	0,5754	0,0014	0,0013	0,0004	0,9308	0,7806	0,7974	0,0000	0,0000	0,0000
BDI 8	**0,0024**	**0,0017**	0,1310	0,0028	0,0022	0,0005	**0,0012**	0,0532	**0,0018**	0,0029	0,0008	0,0021
BDI 9	0,0279	0,0075	0,9758	0,0039	0,0039	0,0000	**0,0008**	**0,0002**	0,3528	0,0075	0,0071	0,0005
BDI 10	0,0087	0,0754	0,0118	0,0026	0,0009	0,0017	**0,0000**	**0,0000**	**0,0039**	**0,0346**	**0,0322**	**0,0024**
BDI 11	0,0984	0,1106	0,1480	0,0012	0,0007	0,0005	0,5183	0,7429	0,2720	0,0003	0,0000	0,0003
BDI 12	0,3731	0,1631	0,8704	0,0006	0,0006	0,0000	**0,0000**	**0,0000**	0,6677	0,0056	0,0056	0,0001
BDI 13	0,2947	0,1566	0,5084	0,0006	0,0005	0,0001	**0,0004**	**0,0001**	0,7275	0,0037	0,0036	0,000
BDI 14	0,2729	0,1070	0,9959	0,0009	0,0009	0,0000	0,1834	0,1327	0,2874	0,0011	0,0007	0,0004
BDI 15	0,0286	0,1039	0,0346	0,0016	0,0006	0,0010	0,1120	0,0507	0,4542	0,0010	0,0008	0,0001
BDI 16	0,6905	0,4591	0,6609	0,0002	0,0001	0,0001	0,3422	0,2528	0,3602	0,0006	0,0003	0,0002
BDI 17	0,1149	0,1272	0,1571	0,0010	0,0005	0,0005	0,2778	0,4170	0,1677	0,0006	0,0001	0,0004
BDI 18	0,4429	0,3223	0,4204	0,0005	0,0003	0,0002	0,2391	0,1102	0,5778	0,0008	0,0007	0,0001
BDI 19	0,0692	0,0647	0,1650	0,0012	0,0008	0,0004	0,9638	0,8260	0,8732	0,0000	0,0000	0,0000
BDI 20	0,1946	0,2742	0,1494	0,0007	0,0003	0,0005	**0,0001**	**0,0000**	0,8927	0,0039	0,0039	0,0000
BDI 21	**0,0001**	**0,0001**	0,1286	0,0061	0,0053	0,0008	0,5414	0,8949	0,2714	0,0004	0,0000	0,0004

Die hier dargestellte Tabelle basiert auf den finalen Regressionsanalysen mittels erneuerter IRT-Schätzer

ERKLÄRUNGEN

(1) Ich erkläre, dass ich mich an keiner anderen Hochschule einem Promotionsverfahren unterzogen bzw. eine Promotion begonnen habe.

(2) Ich erkläre, die Angaben wahrheitsgemäß gemacht und die wissenschaftliche Arbeit an keiner anderen wissenschaftlichen Einrichtung zur Erlangung eines akademischen Grades eingereicht zu haben.

(3) Ich erkläre an Eides statt, dass ich die Arbeit selbstständig und ohne fremde Hilfe verfasst habe. Alle Regeln der guten wissenschaftlichen Praxis wurden eingehalten; es wurden keine anderen als die von mir angegebenen Quellen und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht.

Nürnberg, den 20.08.2021

Johanna Wittmann

DANKSAGUNG

Mein größter Dank gilt meinem Betreuer, Herrn apl. Prof. Dr. Stefan Watzke, der durch sein Engagement und seine Flexibilität die Rahmenbedingungen dafür geschaffen hat, dass ich meine Promotion trotz der größeren örtlichen Distanz vollständig berufsbegleitend erstellen konnte. Die Art und Weise der Zusammenarbeit habe ich allzeit als unkompliziert, wertschätzend und motivierend empfunden. Ich möchte mich daher herzlich für die in mich investierte Zeit, die konstruktiven und hilfreichen Anmerkungen, sowie die ständige Erreichbarkeit bedanken.

Auch bei Herrn Dr. Jochen Ranger, der mich durch seine fachlichen Ratschläge und Anmerkungen bei meiner Arbeit unterstützt hat, möchte ich mich auf diesem Wege herzlich bedanken.

Nicht unerwähnt lassen möchte ich auch die Kommilitonen, die durch die Probandenbefragungen und die damit verbundenen Vorarbeiten die Durchführung meiner Doktorarbeit erst ermöglicht haben.

Zudem möchte ich mich an dieser Stelle bei meiner Familie für die Motivation und die – nicht nur im Rahmen dieser Arbeit – unermüdliche Unterstützung bedanken.