Originalarbeit

Zur Messqualität des Beck-Depressionsinventars (BDI-II) in unterschiedlichen klinischen Stichproben

Eine Item-Response-Theorie Analyse

Ferdinand Keller[1], Christine Kühner[2], Rainer W. Alexandrowicz[3], Ulrich Voderholzer[4,5,6], Adrian Meule[4,5], Jörg M. Fegert[1], Tanja Legenbauer[7], Martin Holtmann[7], Anne-Kathrin Bräscher[8], Martin Cordes[9], Lydia Fehm[10], Anne-Katharina Fladung[11], Thomas Fydrich[10], Alfons Hamm[12], Jens Heider[13], Jürgen Hoyer[14], Tina In-Albon[15], Tania M. Lincoln[11], Wolfgang Lutz[16], Jürgen Margraf[17], Babette Renneberg[18], Angelika Schlarb[19], Henning Schöttke[20], Tobias Teismann[17], Julia Velten[17], Ulrike Willutzki[21], Michael Witthöft[8], Max Ziem[14] und Martin Hautzinger[22]

[1]Klinik für Kinder- und Jugendpsychiatrie / Psychotherapie, Universitätsklinikum Ulm, Deutschland
[2]AG Verlaufs- und Interventionsforschung, Zentralinstitut für Seelische Gesundheit, Mannheim, Medizinische Fakultät Mannheim, Universität Heidelberg, Deutschland
[3]Abteilung für Methodenlehre, Institut für Psychologie, Universität Klagenfurt, Österreich
[4]Klinik für Psychiatrie und Psychotherapie, Klinikum der LMU München, Deutschland
[5]Schön Klinik Roseneck, Prien am Chiemsee, Deutschland
[6]Klinik für Psychiatrie und Psychotherapie, Universitätsklinikum Freiburg, Deutschland
[7]LWL-Universitätsklinik für Kinder- und Jugendpsychiatrie und Psychotherapie, Ruhr-Universität Bochum, Hamm, Deutschland
[8]Psychologisches Institut, Johannes-Gutenberg-Universität Mainz, Deutschland
[9]Institut für Psychologie, Poliklinische Psychotherapieambulanzen, Universität Osnabrück, Deutschland
[10]Institut für Psychologie, Humboldt-Universität zu Berlin, Deutschland
[11]Institut für Psychologie, Arbeitsbereich Klinische Psychologie und Psychotherapie, Universität Hamburg, Deutschland
[12]Zentrum für Psychologische Psychotherapie, Universität Greifswald, Deutschland
[13]Psychotherapeutische Universitätsambulanz, Campus Landau, Universität Koblenz-Landau, Deutschland
[14]Institut für Klinische Psychologie und Psychotherapie, Technische Universität Dresden, Deutschland
[15]Klinische Psychologie und Psychotherapie des Kindes- und Jugendalters, Universität Koblenz-Landau, Deutschland
[16]Poliklinische Psychotherapieambulanz, Klinische Psychologie und Psychotherapie, Universität Trier, Deutschland
[17]Forschungs- und Behandlungszentrum für psychische Gesundheit, Fakultät für Psychologie, Ruhr-Universität Bochum, Deutschland
[18]Hochschulambulanz für Psychotherapie, Diagnostik und Gesundheitsförderung, Freie Universität Berlin, Deutschland
[19]Klinische Psychologie und Psychotherapie des Kindes- und Jugendalters, Abteilung für Psychologie, Universität Bielefeld, Deutschland
[20]Institut für Psychologie, Klinische Psychologie und Psychotherapie, Universität Osnabrück, Deutschland
[21]Department für Psychologie und Psychotherapie, Fakultät für Gesundheit, Universität Witten / Herdecke, Deutschland
[22]Klinische Psychologie und Psychotherapie, Fachbereich Psychologie, Eberhard-Karls-Universität Tübingen, Deutschland

Zeitschrift für Klinische Psychologie und Psychotherapie (2022)
https://doi.org/10.1026/1616-3443/a000676

Zusammenfassung: *Theoretischer Hintergrund:* Das BDI-II ist ein Selbstbeurteilungsinstrument zur Erfassung des Schweregrads einer Depression. Es liegen kaum Analysen mit Modellen aus der Item-Response-Theorie (IRT) vor. *Fragestellung:* Wie hoch ist die Messgenauigkeit des BDI-II über die unterschiedlichen Ausprägungen des latenten Traits (Depressivität) hinweg und sind die Kategorien der Items jeweils aufsteigend geordnet? *Methode:* Anhand von sechs großen Datensätzen aus verschiedenen klinischen Bereichen wurden psychometrische Analysen mit dem Graded Response Model durchgeführt. *Ergebnisse:* In allen Stichproben fand sich eine hohe interne Konsistenz. Die Schwellenwerte waren mit Ausnahme von Item 6 („Bestrafungsgefühle“) geordnet. Gemäß Testinformationsfunktion misst das BDI-II im mittleren bis hohen Depressionsbereich sehr gut (Reliabilität > .90) und im unteren Bereich gut. *Schlussfolgerung:* Für das BDI-II ergibt sich eine hohe und relativ gleichbleibende Messpräzision über einen weiten Bereich des latenten Traits, weshalb es insbesondere im klinischen, aber auch im nicht klinisch relevanten Wertebereich zur Erhebung des Schweregrades einer Depression gut geeignet ist.

Schlüsselwörter: Beck Depressionsinventar II, Item-Response-Theorie (IRT), Reliabilität, Messgenauigkeit, Depression

The Measurement Quality of the Beck Depression-Inventory (BDI-II) in Different Clinical Samples: An Item Response Theory Analysis

Abstract: *Background:* The Beck Depression Inventory (BDI-II) is a self-report instrument for assessing the severity of depression. To date, publications on psychometric properties based on item response theory (IRT) are largely missing. *Objective:* To determine how high the measurement precision is across the latent trait and whether the item categories are in ascending order. *Methods:* Using six large data sets from different clinical settings, we performed psychometric analyses using the graded response model. *Results:* We identified high internal consistencies in all samples. Apart from item 6 (“penalty feelings”), all categories were ordered. According to the Test Information Function, the BDI-II shows very good measurement precision (reliability > .90) in the moderate to high depression range, and good precision in the lower range. *Conclusions:* Our data revealed high and relatively stable measurement precision across a broad range of the depression construct. We consider the BDI-II to be well suited for assessing depression severity levels, particularly in clinical but also in nonclinical settings.

Keywords: Beck Depression-Inventory II, item response theory (IRT), reliability, measurement precision, depression

Das Beck Depressionsinventar (BDI-II) ist ein Selbstbeurteilungsinstrument zur Erfassung des Schweregrads einer Depression, das hauptsächlich im klinischen Bereich oder im Rahmen von Therapiestudien eingesetzt wird (vgl. Kuehner et al., 2022). Ausgehend vom Manual des revidierten BDI (Beck et al., 1996), in dem die Ergebnisse zu 500 erwachsenen Patient_innen analysiert werden, wurde das BDI-II in zahlreiche Sprachen übersetzt und weltweit eingesetzt. Psychometrische Studien zum BDI-II, die Trennschärfekoeffizienten und interne Konsistenz (Cronbachs Alpha) berichten, liegen in großer Zahl vor. Übersichtsarbeiten berichten über gute Itemtrennschärfen und sehr gute interne Konsistenzwerte (Huang & Chen, 2015; Wang & Gorenstein, 2013). Für die deutsche Version des BDI-II fanden sich vergleichbare Größenordnungen in verschiedenen Stichproben (Hautzinger et al., 2006; Kühner et al., 2007).

Ebenfalls breit untersucht wurde die faktorielle Struktur des BDI-II. Die meisten Studien identifizierten unter Verwendung exploratorischer Faktorenanalysen (meist anhand von Hauptkomponentenanalysen) zwei Faktoren, die als kognitiv und somatisch-affektiv bezeichnet wurden (Huang & Chen, 2015) und der ursprünglichen von Beck et al. (1996) vorgeschlagenen Faktorzusammensetzung sehr ähnlich sind. Für die deutsche Version des BDI-II fand sich ebenfalls eine zwei-faktorielle Lösung (Keller et al., 2008). Andere exploratorische Faktorenanalysen identifizierten jedoch auch drei Faktoren, mit einem eigenständigen affektiven Faktor (vgl. Huang & Chen, 2015; Wang & Gorenstein, 2013). Eine Aggregierung von publizierten Korrelationsmatrizen der Items aus 16 Studien erbrachte ebenfalls eine Lösung mit zwei Faktoren, aber auch die Annahme eines Generalfaktors war angesichts der guten Anpassung des Ein-Faktor-Modells vertretbar (Huang & Chen, 2015).

Konfirmatorische Faktorenanalysen zum BDI-II begannen hauptsächlich mit Ward (2006), der insgesamt sechs Datensätze von Beck et al. und weiteren Studien vergleichend untersuchte. Zusätzlich zu den bis dahin vorgeschlagenen Modellen mit zwei oder drei korrelierten Faktoren nahm er auch ein Bifaktor-Modell auf. Bifaktor-Modelle sind dadurch charakterisiert, dass alle Items auf einem Generalfaktor und zusätzlich noch auf einem von mehreren spezifischen Faktoren laden, wobei der Generalfaktor mit den spezifischen Faktoren unkorreliert ist. Dabei zeigte sich ein Bifaktor-Modell als zumindest gleichwertig und meist besser passend, und es konnte ein starker Generalfaktor angenommen werden (Ward, 2006). Weitere Studien fanden ebenfalls eine Überlegenheit eines Modells mit einem starken Generalfaktor und spezifischen Faktoren, die allerdings wenig zusätzlichen Erklärungswert aufwiesen (Faro & Pereira, 2020; McElroy et al., 2018; Subica et al., 2014). Diese Studien untersuchten eine Reihe unterschiedlicher Faktormodelle aus der Literatur hinsichtlich Goodness-of-fit und favorisierten jeweils ein Bifaktor-Modell. Die Überlegenheit des Bifaktor-Modells steht allerdings in Frage, da Bifaktor-Modelle allein schon wegen der höheren Anzahl zu schätzender Parameter tendenziell immer etwas besser passen sollten. Vor allem aber wurden konzeptuelle Gründe gegen die Anwendung des sogenannten symmetrischen Bifaktor-Modells, das in den genannten Studien verwendet

wurde, dargelegt (Eid et al., 2017; Heinrich et al., 2020). Insgesamt lassen die hohen Korrelationen der Faktoren in den zwei- und drei-Faktorlösungen (> .80, teilweise > .90), der starke Generalfaktor in den Bifaktor-Modellen und die gute Anpassung des Ein-Faktor-Modells von Huang und Chen (2015) eine im Wesentlichen eindimensionale Struktur des BDI-II aus empirischer Sicht plausibel erscheinen.

Bei Jugendlichen liegen sowohl insgesamt als auch bezüglich der faktoriellen Struktur wesentlich weniger Studien vor. Zu nennen sind hier initiale Studien von Steer et al. (1998) und Osman et al. (2004). Eine ausführliche Beschreibung zu Faktorenanalysen bei Jugendlichen findet sich in Keller et al. (2020). Für die in Keller et al. (2020) untersuchte Stichprobe von 835 Jugendlichen ergab sich, dass ein Bifaktor-Modell mit einem starken Generalfaktor und zwei ergänzenden spezifischen Faktoren, kognitiv und somatisch, zu favorisieren war. Aufgrund der geringen zusätzlich erklärten Varianzanteile der spezifischen Faktoren legt auch diese Untersuchung letztlich eine eindimensionale Skala nahe.

Alternative Analysemethoden zur Gruppierung von Items kamen ebenfalls zum Einsatz. Dazu zählt insbesondere die Anordnung der Items gemäß ihrer Ähnlichkeit im Rahmen einer Nonmetrischen Multidimensionalen Skalierung. Bühler et al. (2012) fanden einen inneren Bereich von depressiven Kern-Items und fünf weitere Facetten mit kognitiven und psychovegetativen Items; zusätzlich zeichnete sich noch eine Facette Aktivierung ab. Daraus wurde ein modifiziertes Bifaktor-Modell abgeleitet, bei dem Items auf mehr als nur einem der drei spezifischen Faktoren laden können. Dieses Modell konnte an einer unabhängigen Stichprobe von psychosomatischen Patient_innen repliziert werden (Bühler et al., 2014). Mit derselben Methode zeigten Voderholzer et al. (2019) für Patient_innen mit Majorer Depression und Essstörung, dass die Relation der Kern-Items der Depression bei beiden Störungsbildern große Ähnlichkeiten aufwies.

Zusammenfassend ist festzustellen, dass das BDI-II in zahlreichen Studien hinsichtlich seiner psychometrischen Eigenschaften gemäß Kennwerten der Klassischen Testtheorie (Trennschärfe, Cronbachs Alpha) und seiner faktoriellen Struktur untersucht wurde. Weniger Aufmerksamkeit wurde bislang auf Eigenschaften gelegt, die sich mittels Modellen der Item-Response-Theorie (IRT, Rost, 2004) untersuchen lassen. Hierzu zählt insbesondere die Frage, in welchen Bereichen des Wertebereiches das BDI-II gut bzw. weniger gut misst. Im Gegensatz zur klassischen Testtheorie (KTT), in der die Reliabilität eines Tests lediglich durch einen einzigen Koeffizienten (z. B. Cronbachs Alpha) für den gesamten Wertebereich des Tests ausgedrückt wird, lässt sich mit IRT-Modellen die Messgenauigkeit spezifisch für jede Stelle des latenten Kontinuums bestimmen (Samejima, 1994). Ermöglicht wird dies durch die sogenannte *Test Information Function* (TIF), die für jeden Punkt auf dem latenten Kontinuum (Trait) angibt, wie gut der Test hier misst und damit einen lokalen Index der Messpräzision darstellt. Die TIF setzt sich dabei zusammen aus der Summe der Informationen, die die einzelnen Items über den Trait hinweg beitragen (*Item Information Curves*, IIC). Da der Wert der TIF keine inhaltliche interpretierbare Information darstellt, wird sie oft in den *Standard Error of Measurement* oder die Reliabilität umgerechnet (O'Connor, 2018). Weiterhin erlauben bestimmte IRT-Modelle die empirische Überprüfung, ob die Anordnung der Schwellenparameter auf der latenten Dimension der Abfolge der Antwortkategorien entspricht, da Schwellenvertauschungen den Annahmen des Modells widersprechen (Rost, 2004). Mittels IRT lassen sich sowohl ein- wie auch mehrdimensionale Modelle schätzen (z. B. Reckase, 2009; Chalmers, 2012). Aufgrund der empirischen Befundlage wird in dieser Studie ein eindimensionaler Ansatz gewählt, wie er auch in der klinischen Praxis bei Berechnung eines Gesamtwertes angenommen wird.

Bisher vorliegende Studien zum BDI-II mittels IRT-Modellen beschäftigten sich vorwiegend mit Itemanalysen (s. Überblick in Wang & Gorenstein, 2013) oder dem so genannten Differential Item Functioning von BDI-II Items, d. h. der Frage ob Items in Subgruppen unterschiedlich messen bzw. verstanden werden (de Sá Junior et al., 2019). IRT-Modelle wurden auch angewendet, um verschiedene Erhebungsverfahren zur Depression in eine „common metric" zu standardisieren und damit vergleichbar zu machen (Wahl et al., 2014) oder zur Entwicklung computerisierter adaptiver Tests (Reise & Waller, 2009). Zur Frage der Messgenauigkeit über den Wertebereich des Konstrukts hinweg ist die Zahl der Studien bisher gering. In einer großen Schulstichprobe von Jugendlichen fanden Olino et al. (2012), dass das BDI (noch in der ersten Version) sehr gut im mittleren bis hohen Depressionsbereich misst. Im Vergleich zur ebenfalls erhobenen CES-D (deutsch: Allgemeine Depressionsskala – ADS, Hautzinger et al., 2012) zeigte sich die bereits vermutete Überlegenheit des BDI bei höheren Schweregraden von Depression, während die ADS bei niedrigeren Schweregraden etwas mehr Information lieferte. Dieses Ergebnis spricht daher für den Einsatz des BDI in klinischen Stichproben und der ADS in epidemiologischen Stichproben (Olino et al., 2012). Die Analyse einer gemischten Stichprobe von 188 erwachsenen Patient_innen mit der Diagnose einer unipolaren depressiven Störung und 957 Personen aus einer nicht klinischen Stichprobe (Park et al., 2020) ergab eine gute Messeigenschaft des BDI-II über einen Bereich von –1 SD bis +3 SD. Dabei war der Bereich zwischen 0 und 2.5 SD fast plateauartig, was für eine

gleichbleibend hohe Messpräzision im mittleren bis oberen Schweregradbereich spricht (Park et al., 2020).

Vergleichende Analysen zur Messgenauigkeit des BDI mit anderen Depressionsinstrumenten in der Selbstbeurteilung ergaben die schon erwähnten Unterschiede gegenüber der ADS bei Jugendlichen (Olino et al., 2012). Weiterhin verwendeten Zhao et al. (2017) den Patient Health Questionnaire (PHQ-9, Löwe et al., 2002) und die Subskala Depression aus der Depression, Anxiety and Stress Scale (DASS, Lovibond & Lovibond, 1995) bzw. der Hospital Anxiety and Depression Scale (HADS, Zigmond & Snaith, 1983). Alle fünf Skalen wurden an einer Stichprobe ambulanter Patient_innen, die wegen einer Depression Behandlung suchten (n = 207), erhoben. BDI-II und ADS wiesen eine hohe Messgenauigkeit zwischen ca. –1 und 2.5 SD auf, wobei die ADS sogar über einen weiten Bereich etwas besser war und nur im Bereich > 2 SD etwas schlechter abschnitt als das BDI-II. Auch der PHQ-9 wies gute Messeigenschaften im Bereich –1 SD bis 2 SD auf, ebenso die DASS-Subskala Depression, während die HADS-Subskala Depression über den ganzen Bereich keine gute Reliabilität erzielte (maximale Reliabilität = .78) (Zhao et al., 2017).

Im vorliegenden Artikel lag der Schwerpunkt auf psychometrischen Analysen zum BDI-II gemäß IRT. Die zentrale Fragestellung bezog sich dabei auf die Messgenauigkeit des BDI-II über den Wertebereich des latenten Merkmals hinweg. Dieser Aspekt ist wichtig sowohl für die reliable Erfassung der Symptomatik in Therapiestudien, z. B. im Vergleich von Baseline- und Entlasserhebung, als auch im Bereich der so genannten cut-off- Werte, die in der Praxis häufig zur kategorialen Einschätzung des Schweregrads einer Depression oder in Therapiestudien als Eingangskriterium verwendet werden (Kuehner et al., 2022). Von zusätzlichem Interesse war die Frage, ob die teilweise sehr symptomspezifischen Formulierungen der Itemkategorien (siehe ein Beispiel im Methodenteil) auch aufsteigend geordnet waren. Bereits im Manual verwiesen Beck et al. (1996) darauf, dass bei vier der 21 Items (6, 9, 11 und 21) die erwartete ansteigende Reihenfolge der Kategorien nicht vollständig erfüllt war. Anhand der Iteminformationskurven sollte zudem der Frage nachgegangen werden, ob manche Items eher im subklinischen oder im klinischen Wertebereich gut messen. Damit ergaben sich die folgenden Fragestellungen, für deren Beantwortung sich spezifisch die IRT-Analysemethodik eignete:

1. Weist das BDI-II über einen weiten Bereich des latenten Traits hinweg eine hohe und relativ konstante Reliabilität auf?
2. Sind die einzelnen Itemkategorien aufsteigend geordnet und haben alle Kategorien einen eigenständigen Bereich mit maximaler Auswahlwahrscheinlichkeit auf dem latenten Trait?
3. Wie groß ist der Informationsgehalt der einzelnen Items (Höhe und Bereich der Iteminformationskurven)?

Alle drei Fragestellungen wurden auch auf mögliche Stichprobenunterschiede hin betrachtet, insbesondere bezüglich ambulanter und stationärer Behandlung, Aufnahme- und Entlassungszeitpunkt sowie erwachsene und jugendliche Patient_innen.

Zudem liegen detaillierte Ergebnisse zum BDI-II auf der Basis der KTT für deutschsprachige Stichproben mit wenigen Ausnahmen seit über 15 Jahren nicht mehr vor. Deshalb werden im Folgenden auch KTT-Ergebnisse zum BDI-II berichtet, was auch den Vergleich mit älteren KTT-Studien (Hautzinger et al., 2006) ermöglicht.

Methode

Stichproben

Insgesamt standen die Daten aus sechs Stichproben zur Verfügung, deren deskriptive Beschreibungen in Tabelle 1 zusammengefasst sind. Im Folgenden werden weitere Angaben zur Erhebung, zur Behandlung fehlender Werte und zur Diagnosestellung gemacht.

a) Daten aus dem Projekt KODAP (Koordination der Datenerhebung und -auswertung an Forschungs-, Lehr- und Ausbildungsambulanzen für psychologische Psychotherapie; Velten et al., 2018). Die Daten wurden zu Beginn der psychotherapeutischen Behandlung in den teilnehmenden Ambulanzen erhoben, wobei in allen Fällen strukturierte oder standardisierte diagnostische Interviews eingesetzt wurden. Der Altersbereich erstreckte sich von 16 bis 82 Jahren. Um die Altersspanne zu begrenzen wurden Jugendliche unter 18 Jahren (n = 8) und Personen über 75 Jahre (n = 6) ausgeschlossen. Der resultierende Datensatz enthielt 2585 BDI-II, von denen n = 2450 (94.8 %) vollständig ausgefüllt waren und in die folgenden Analysen eingingen. Bezüglich der ICD-10 Diagnosen konnten in der Dokumentation insgesamt fünf Diagnosen eingegeben werden. Die Haupt- bzw. Indexdiagnose sollte dabei an erster Stelle stehen. Für die Substichprobe der Personen mit einer Diagnose aus dem Kapitel F3 des ICD-10 wurde daher das Auftreten einer F3-Kodierung als Indexdiagnose als Kriterium verwendet.

b) Daten von psychosomatischen Patient_innen, in stationärer Behandlung in der Schön Klinik Roseneck (Prien am Chiemsee) bzw. tagesklinischer Behandlung in der Schön Klinik Tagesklinik München. Für eine Beschreibung des Behandlungsansatzes in den Schön Kliniken vgl. Voderholzer et al. (2019). Zur Verfügung standen eine Aufnahme- und eine Entlassungsstichprobe, die im Fol-

genden separat (d. h. nicht längsschnittlich) betrachtet werden. Der Altersbereich lag zwischen 12 und 88 Jahren. Analog zum KODAP-Datensatz wurden Personen über 75 Jahren (n = 58) aus den Analysen ausgeschlossen, des Weiteren Jugendliche im Alter von 12 Jahren (n = 9). Hier wurde unterteilt in eine Jugendlichenstichprobe (13–18 Jahre) und eine Erwachsenenstichprobe (19–75 Jahre). Im resultierenden Gesamtdatensatz von 14499 BDI-II bei Aufnahme waren n = 13821 (95.3 %) vollständig ausgefüllt und gingen in die folgenden Analysen ein. In der Entlassungsstichprobe lagen nach Anlegen derselben Alterskriterien 10439 BDI-II vor, von denen n = 9888 (94.7 %) vollständig ausgefüllt waren. Als Kriterium für die Substichprobe der Personen mit einer Diagnose aus dem Kapitel F3 des ICD-10 wurde das Auftreten einer F3-Kodierung als Hauptdiagnose herangezogen.

In beiden Stichproben war ein kleiner Anteil von Patient_innen mit der Diagnose einer bipolaren Störung (F31.x) enthalten: 1.2 % in KODAP und 0.6 % in der Stichprobe Roseneck Erwachsene bei Aufnahme. Da es sich in fast allen Fällen um eine depressive Episode im Rahmen einer bipolaren Störung handelte, wurden diese Daten nicht ausgeschlossen.

c) Daten einer Stichprobe von Jugendlichen in ambulanter oder stationärer Behandlung in Ulm bzw. Hamm (n = 835). Diese Stichprobe wurde bereits hinsichtlich der faktoriellen Struktur des BDI-II sowie hinsichtlich Messinvarianz bezüglich des Geschlechts publiziert (Keller, Kirschbaum-Lesch & Straub, 2020). Eine Analyse mit IRT-Modellen wurde in diesem Rahmen nicht durchgeführt.

Erhebungsinstrument

Die revidierte Version des Beck Depressions-Inventars (BDI-II, Hautzinger et al., 2006) ist ein Selbstbeurteilungsinstrument mit 21 Fragen, das bei Jugendlichen ab 13 Jahren und Erwachsenen die Bestimmung der Schwere einer Depression ermöglicht. Das BDI-II besteht aus 21 Items, die jeweils auf einer Skala mit 0 bis 3 anzukreuzen sind. Die Texte dieser vier Antwortkategorien sind für jedes Item spezifisch ausformuliert; bei Item 1 (Traurigkeit) lautet beispielsweise die Kategorie 3 „ich bin so traurig oder unglücklich, dass ich es nicht aushalte“. Der Wertebereich der Skala geht von 0 bis 63 und die Schweregrade lassen sich charakterisieren als nicht / minimal depressiv (0–13), leicht depressiv (14–19), mäßig depressiv (20–28) und schwer depressiv (29–63).

Statistische Analysen

Deskriptive Analysen und die Auswertungen zur KTT wurden mit SAS v9.4 vorgenommen. Neben Cronbachs Alpha wurde auch McDonalds Omega als modellbasierte Reliabilitätsschätzung berechnet, da dieses weniger Vorannahmen als Alpha macht (vgl. Schermelleh-Engel & Gäde, 2020). Für die Berechnung von Omega wurde das SAS Makro v0.1 von Hayes und Coutts (2020) verwendet, basierend auf den Faktorladungen und Fehlervarianzen aus einer einfaktoriellen Faktorenanalyse mit Maximum-Likelihood-Schätzung. Ein häufig verwendeter einfacher Hinweis auf Eindimensionalität besteht im Vergleich der Eigenwerte im Rahmen einer Faktorenanalyse. Erwartet werden für eine eindimensionale Skala ein hoher erster Eigenwert und ein niedriger zweiter Eigenwert, das heißt der erste Faktor erklärt viel und zugleich deutlich mehr Varianz als der zweite (und die weiteren) Faktoren. Zusätzlich wird häufig das Verhältnis der beiden Eigenwerte berechnet, dass > 4 sein sollte (vgl. Zhao et al., 2017); eine vertiefte Analyse würde den Einsatz weiterer Verfahren erfordern (vgl. Diskussion).

Für die IRT-Analysen wurden aus Gründen der Vergleichbarkeit mit den bereits vorliegenden Studien (Olino et al., 2012; Zhao et al., 2017) das Graded Response Model (GRM; Samejima, 1969) verwendet. Geschätzt werden dabei pro Item drei Schwellenparameter, die für jedes benachbarte Paar von Antwortkategorien angeben, ab welchem Wert des latenten Merkmals (= geschätztem Grad an Depressivität) eher eine der darüber liegenden Kategorien ausgewählt wird. Dazu kommt pro Item ein Diskriminationsparameter, der modelliert, wie gut die Items zwischen Personen mit niedrigen vs. höheren Ausprägungen des Traits trennen können. Da das GRM durch die kumulative Modellierung der Kategorienwahrscheinlichkeiten („difference model“; vgl. Ostini & Nerig, S. 12) zwingend ansteigende Kategorien annimmt, ist es nicht zur Prüfung deren Geordnetheit geeignet. Daher wurde zur Überprüfung von Fragestellung 2 das Nominal Response Model von Bock (1972) eingesetzt. Dieses postuliert keinerlei Geordnetheit der Kategorien („nominales Modell“), eine allfällige Ordnung ist daher empirisch feststellbar (vgl. Ostini & Nerig, 2006, S. 19*f*). Die Modellschätzungen für beide Modelle erfolgten mit dem R-Paket *mirt* (Chalmers, 2012) in der Version 1.33.2. Die Abbildungen wurden aus *mirt* übernommen, lediglich die gleitende Reliabilität wurde gesondert mit R erstellt, wobei sie für jeden Punkt des Traits (θ) nach der Formel: Reliabilität(θ) = TIF(θ) / (TIF(θ) + 1) berechnet wurde (vgl. O’Connor, 2018, Formel 6).

Tabelle 1. Stichprobenübersicht, deskriptive Angaben (Mittelwert und Standardabweichung) und Reliabilitätsmaße.

Stichprobe	N	Alter	Geschlecht (% weiblich)	BDI-II- Summe	Cronbachs Alpha	Omega
Erwachsene						
KODAP						
Gesamt	2450	37.8 (13.2)	65.6%	22.1 (11.7)	.918	.916
Depression (F3-Diagnose)	1080	39.9 (13.1)	64.4%	25.1 (11.0)	.900	.898
Roseneck Aufnahme						
Gesamt	10720	40.7 (14.7)	68.0%	28.5 (11.6)	.914	.912
Depression (F3-Diagnose)	5240	47.0 (13.1)	57.4%	28.8 (10.9)	.908	.905
Roseneck Entlassung						
Gesamt	7590	40.9 (14.5)	67.7%	15.8 (12.3)	.946	.946
Depression (F3-Diagnose)	3737	47.0 (12.9)	56.8%	15.4 (11.9)	.948	.948
Jugendliche						
Ulm / Hamm KJPP						
Gesamt	835	15.8 (1.4)	58.7%	19.7 (13.6)	.938	.940
Depression (F32/3, F41.2, F92.0)	471	15.9 (1.3)	69.9%	24.9 (13.2)	.926	.927
Roseneck Aufnahme						
Gesamt	3101	16.3 (1.3)	88.9%	29.5 (12.4)	.923	.922
Depression (F3-Diagnose)	930	16.3 (1.2)	85.7%	33.7 (11.2)	.904	.903
Roseneck Entlassung						
Gesamt	2298	16.2 (1.3)	89.9%	19.3 (13.7)	.950	.951
Depression (F3-Diagnose)	664	16.3 (1.2)	87.2%	22.7 (14.3)	.953	.954

Anmerkungen: Bei Geschlecht taucht in keiner Stichprobe die Kategorie „divers“ auf, weshalb bei den Analysen durchgängig von einer dichotomen Variablen ausgegangen wird.

Ergebnisse

In Tabelle 1 zeigen sich wie erwartet höhere BDI-II-Summenwerte für die stationären als die ambulanten Aufnahmestichproben sowie höhere Werte zur Aufnahme als zur Entlassung. Auch die internen Konsistenzwerte (Cronbachs Alpha) lagen erwartungsgemäß im hohen Bereich (≥ .90), wobei die Koeffizienten in den Entlassstichproben am höchsten sind. In Tabelle S1 im Elektronischen Supplement (ESM 1) finden sich die deskriptiven Angaben zu den einzelnen Items sowie die Trennschärfekoeffizienten (part-whole korrigiert) für die Stichproben Erwachsener. Die Trennschärfen lagen alle mindestens bei .40 und insbesondere in der Entlassstichprobe in ca. der Hälfte der Items sogar bei .70 und mehr. Bei den Jugendlichen (Tabelle S2 im ESM 1) ergaben sich analoge Ergebnisse bezüglich interner Konsistenz und Trennschärfen, auffällig war jedoch die geringe Trennschärfe von Item 21, die nur knapp über dem häufig verwendeten Grenzwert von .30 liegt. Die Reliabilität berechnet mit Omega für die verschiedenen Stichproben findet sich in Tabelle 1. Alle Omega-Werte waren ≥ .90 und lagen insgesamt jeweils nahe bei den Werten von Cronbachs Alpha. Die einzelnen Faktorladungen der ein-faktoriellen Lösung finden sich beispielhaft für die KODAP-Stichprobe in Tabelle S4 im ESM 1, zusammen mit den Diskriminationsparametern, die sich im GRM ergaben. Außerdem sind in der Tabelle die Parameterwerte für die Substichprobe mit einer Indexdiagnose Depression (F3) aufgeführt; substantielle Unterschiede zur Gesamtstichprobe sind nicht zu erkennen. Die Trennschärfekoeffizienten für die drei Erwachsenenstichproben mit einer Depressionsdiagnose sind in der Tabelle S3 im ESM 1 dargestellt. Auch hier sind keine wesentlichen Unterschiede zu den Trennschärfekoeffizienten der jeweiligen Gesamtstichproben (Tabelle S1) ersichtlich.

Bei der Prüfung auf Eindimensionalität über die Eigenwerte ergab sich, dass in den sechs Stichproben die ersten Eigenwerte zwischen 8.99 und 12.42 liegen, während die zweiten Eigenwerte zwischen 1.15 und 1.82 schwankten. Das Verhältnis von erstem zu zweitem Eigenwert war zumindest 4.95 und spricht damit für weitgehende Eindimensionalität des BDI-II.

In der folgenden Darstellung der IRT-Analysen werden aus Platzgründen beispielhaft die Ergebnisse der Stichprobe KODAP dargestellt; auf die Ergebnisse der übrigen Stichproben wird im Text eingegangen und alle zugehörigen Abbildungen finden sich im ESM 1.

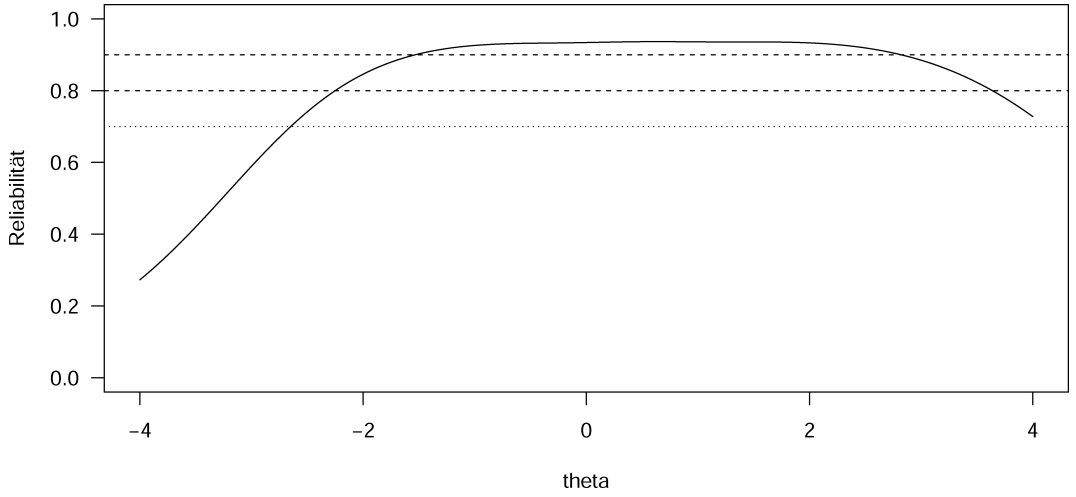

Abbildung 1. Reliabilität des BDI-II-Gesamtwerts in der Stichprobe KODAP über den Trait (theta) hinweg, geschätzt mit dem Graded Response Model.

Fragestellung 1: Messgenauigkeit des BDI-II über dessen Wertebereich hinweg

Bei der Inspektion der aus der TIF abgeleiteten Reliabilität über den Trait hinweg zeigte sich in der KODAP-Stichprobe (Abbildung 1) der erwartete plateauförmige Verlauf über einen weiten Bereich des Traits. Insbesondere im mittleren und hohen Bereich lag die Reliabilität bei einem sehr guten Wert von über .90 und bereits im unteren Bereich von −2 SD ergab sich eine Reliabilität von .85. Für den Bereich von −1 SD bis +2.3 SD lag die Reliabilität bei ≥ .93. In den weiteren Stichproben fand sich ebenfalls dieser plateauförmige Verlauf über einen weiten Bereich des Traits (siehe Abbildungen S7, S12, S17, S22 und S27 im ESM 1); analog zu KODAP lag die Reliabilität im mittleren und hohen Bereich bei über .90.

Fragestellung 2: Sind die einzelnen Itemkategorien aufsteigend geordnet, haben alle Kategorien einen eigenständigen Bereich auf dem latenten Trait?

Bei der Betrachtung der kategorienspezifischen Antwortwahrscheinlichkeiten (Abbildung 2) fällt auf, dass die Kategorie 2 (bezeichnet mit P3) von Item 6 (Bestrafungsgefühle) eine durchgehend niedrige Wahrscheinlichkeit aufweist; auch im nominalen Modell (Abbildung S31) zeigt sich, dass Kategorie 2 nicht wie erwartet zwischen den Kategorien 1 und 3 liegt. Leicht problematisch erscheint die Kategorie 2 bei Item 18 (Appetitveränderung), was sich im nominalen Modell aber nicht bestätigt, und Item 10 (Weinen). Positiv hervorzuheben sind die nahezu idealtypischen Kategorienkurven der Items 15 (Energieverlust), 20 (Ermüdung / Erschöpfung) und auch 4 (Verlust von Freude). Die Diagramme dieser Items lassen (auch im nominalen Modell) erkennen, dass alle Kategorien gut über den gesamten Wertebereich des latenten Merkmals Depressivität verteilt sind, in der Reihenfolge der Antwortkategorien sortiert sind und jede Kategorie ein spezifisches und hinreichend großes Intervall der Depressivität abzubilden vermag.

Die durchgehend niedrige Wahrscheinlichkeit von Kategorie 2 bei Item 6 zeigte sich in allen Stichproben in gleicher Weise (siehe Abbildungen S8, S13, S18 und S28 im ESM 1), abgesehen von den Jugendlichen in Roseneck bei Entlassung (Abbildung S23). Auch die Kategorie 2 bei Item 18 erreichte in allen Stichproben kein Maximum. „Grenzwertig" waren die Items 10 und 11 (Unruhe), deren Kategorie 2 in manchen Stichproben kein eigenständiges Maximum erreichte, während dies in anderen Stichproben gegeben war. Im nominalen Modell wurden diese Befunde bestätigt mit Ausnahme von Item 11, das in allen Stichproben eine geordnete Kategorienstruktur aufwies (siehe Abbildungen S31–S36 im ESM 1). Beim Item 21 (Verlust an sexuellem Interesse) hatten die beiden mittleren Kategorien nur eine flache Wahrscheinlichkeitskurve, so dass sich das Item als im Wesentlichen dichotom messend darstellt; diese Dichotomie ist in den Jugendlichenstichproben noch ausgeprägter. Item 21 war für Jugendliche zudem „schwierig", d. h. die Schwelle zu Kategorie 3 liegt in einem hohen Bereich des Traits (siehe Abbildungen S18, S23, S28 im ESM 1).

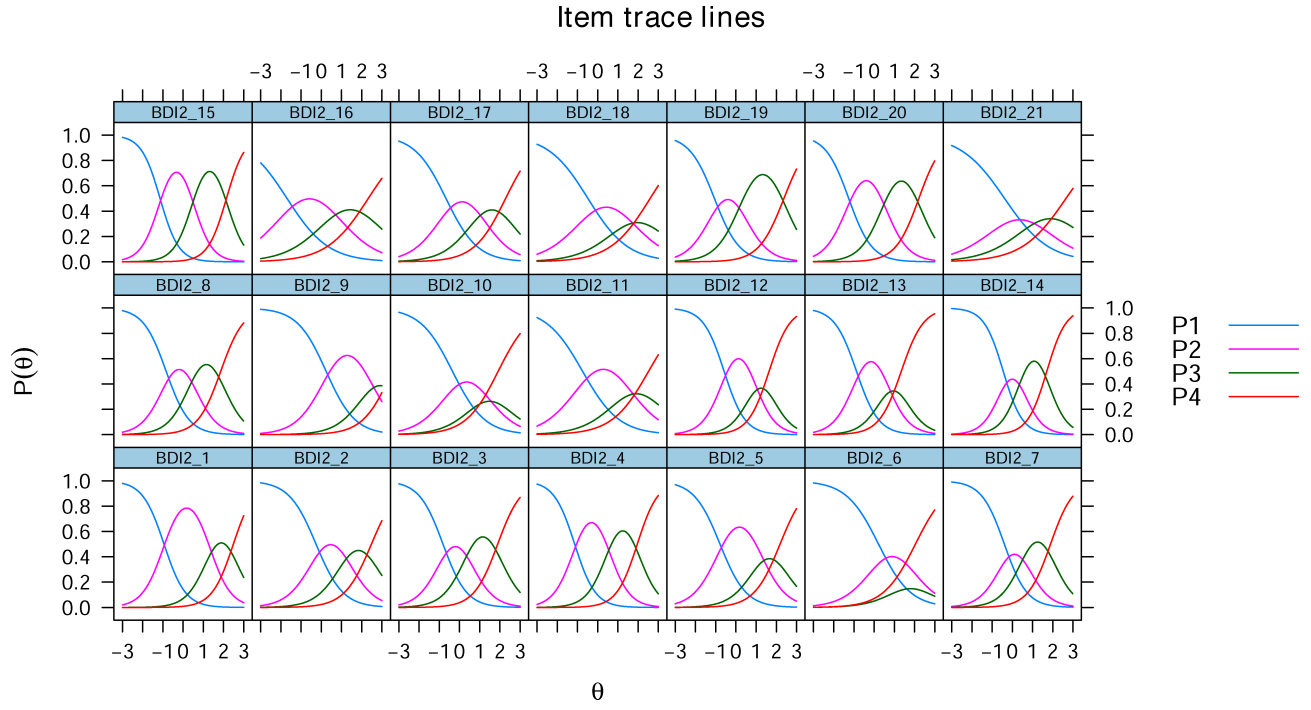

Abbildung 2. Kategorienspezifische Antwortwahrscheinlichkeiten in den einzelnen BDI-II-Items in der Stichprobe KODAP (P1 beschreibt die Wahrscheinlichkeit, die Kategorie 0 anzukreuzen, P2 die von Kategorie 1 usw.).

Fragestellung 3: welche Items haben einen hohen Informationsgehalt (und über einen weiten Bereich)

Aus der Höhe der Iteminformationskurven (IIC) für die KODAP-Stichprobe (Abbildung 3) lässt sich ableiten, dass das Item 14 (Wertlosigkeit) die höchste Iteminformation aufweist und sich diese auch über einen breiten Bereich des Traits erstreckt. Ähnlich hoch ist die IIC von Item 15 (Energieverlust), aber auch die Items 4 (Freudlosigkeit), 12 (Interesseverlust) und 13 (Entschlussunfähigkeit) weisen viel Informationsgehalt auf. Auf der negativen Seite, d. h. Items mit geringer und flacher IIC, liegen die Items 16 (Schlaf) und 18 (Appetitveränderung), und auch Item 21 (Verlust an sexuellem Interesse) liefert wenig Information.

Bezogen auf die anderen Stichproben wies das Item 14 ebenfalls konsistent die höchste Iteminformation auf (vgl. Abbildungen S9, S14, S19, S24 und S29 im ESM 1). Knapp dahinter lag das Item 7 (Selbstablehnung) in den beiden Jugendlichenstichproben, während bei den Erwachsenenstichproben Item 15 (KODAP) ähnlich hoch war wie Item 14 und bei den Aufnahmewerten der erwachsenen Roseneck-Patient_innen mehrere weitere Items ähnlich informativ waren. In der erwachsenen Entlassstichprobe aus Roseneck stach das Item 12 mit ähnlich hoher IIC wie Item 14 heraus (vgl. Abbildung S14). Auf der negativen Seite lagen weitgehend übereinstimmend in allen Stichproben die Items 16 und 18, und ebenso Item 21.

Diskussion

In einer psychometrischen Analyse des BDI-II an unterschiedlichen klinischen Stichproben zeigte sich gemäß klassischer Testtheorie jeweils eine hohe interne Konsistenz, die vergleichbar mit den Reliabilitätswerten aus anderen Studien ausfiel. Damit bestätigen sich die guten psychometrischen Werte des BDI-II für die deutsche Version (Hautzinger et al., 2006) auch in großen klinischen Stichproben. Die Bestimmung der Reliabilität über Omega erbrachte sehr ähnliche Reliabilitätswerte. In den beiden Entlassstichproben war die Reliabilität jeweils noch etwas höher als in den Aufnahmestichproben, und ebenso sind die meisten Trennschärfekoeffizienten höher. In den Entlassungsbögen scheinen die Items also (noch) homogener eingeschätzt zu werden. Für Erwachsene und Jugendliche ergaben sich nur kleine Unterschiede bezüglich der Items mit höchster Trennschärfe; ein analoges Ergebnis war in den Ausprägungen der Iteminformation aus den IRT-Analysen zu sehen.

Messpräzision des BDI-II über verschiedene Schweregradbereiche (IRT-Analyse)

Die Reliabilität erweist sich über einen weiten Bereich des Traits als sehr gut und dieses Ergebnis zeigt sich auch in allen Stichproben. Gute Messeigenschaften des BDI-II

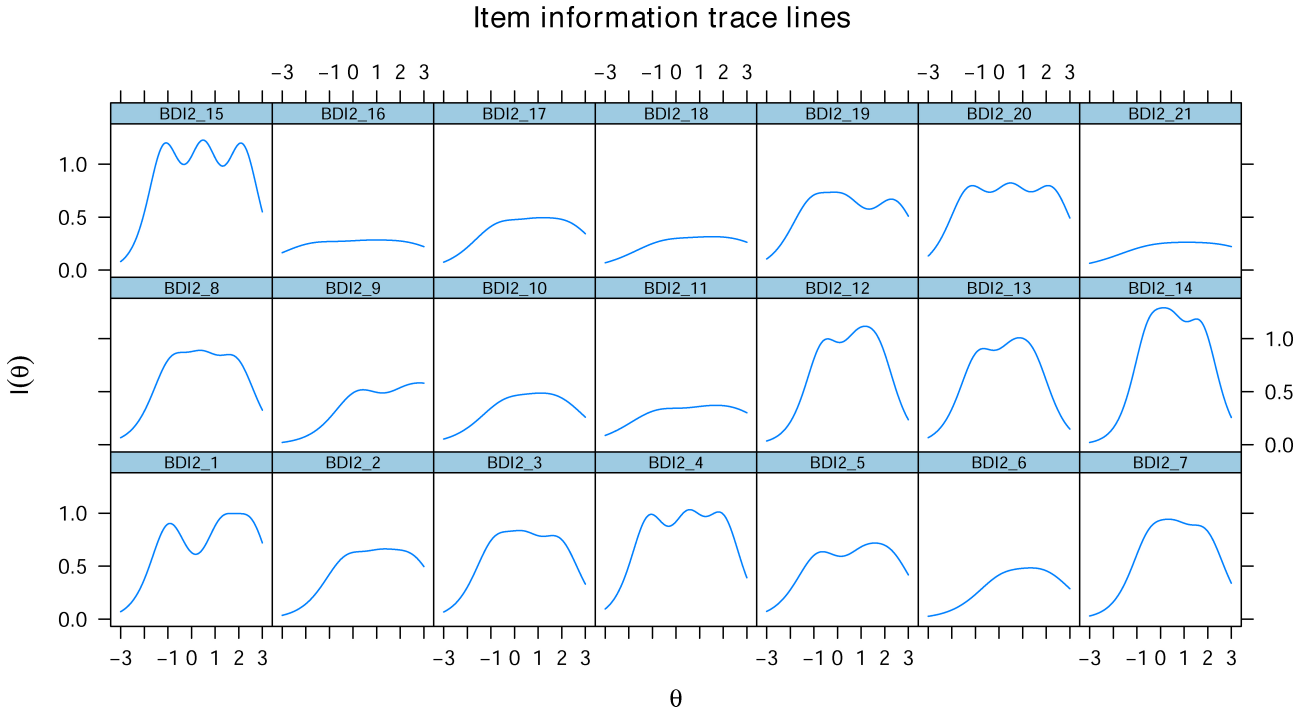

Abbildung 3. Iteminformationskurven der einzelnen BDI-II-Items in der Stichprobe KODAP.

über einen Bereich von -1 SD bis +3 SD fanden auch Park et al. (2020), und insbesondere der Bereich zwischen 0 und 2.5 SD war fast plateauartig, was für eine gleichbleibend hohe Messpräzision im mittleren bis oberen Schweregradbereich spricht (Park et al., 2020). Ein analoges Ergebnis berichten Zhao et al. (2017), bei deren Studie das BDI-II eine hohe Messgenauigkeit zwischen ca. -1 und 2.5 SD aufwies. Olino et al. (2012) kommen ebenfalls zu dem Schluss, dass das BDI sehr gut im mittleren bis hohen Depressionsbereich misst. Dort wurde allerdings das ursprüngliche BDI an einer wenig beeinträchtigten Jugendlichenstichprobe untersucht, für die der „mittlere" Bereich bereits vergleichsweise niedrig lag.

Das BDI-II misst somit sehr gut über einen weiten Symptombereich und bereits bei einem einstelligen Summenscore kann von einer Reliabilität von > .80 ausgegangen werden. Damit besteht eine gute Zuverlässigkeit für die Schweregradmessung in Studien und auch im Bereich der kritischen Schweregradeinteilungen (cut-offs: 14, 20, 29) liegt eine hohe Messpräzision vor. O'Connor (2018, p. 1000) empfiehlt für reliable Ja-Nein-Entscheidungen an den Cut-off Werten, die er als „diagnostische Reliabilität" bezeichnet, eine TIF ≥ 10; dies entspricht einer Reliabilität von ≥ .91 und ist in den untersuchten Stichproben gegeben.

In den Entlassstichproben ist die Reliabilität sowohl nach IRT wie KTT nochmals höher, das heißt das BDI-II wird noch homogener ausgefüllt. In einer Analyse des PHQ-9 über die Anzahl der therapeutischen Sitzungen hinweg nahm die Homogenität ebenfalls zu (Stochl et al., 2020) und auch Fried et al. (2016) fanden ein analoges Ergebnis und diskutieren eine Reihe von möglichen Gründen, die sich aber nicht empirisch bestätigen ließen.

Ordinale Anordnung der Kategorien

In fast allen Items hat jede Kategorie einen Bereich maximaler Auswahlwahrscheinlichkeit über den gesamten untersuchten Wertebereich des latenten Merkmals hinweg. Eine durchgängige Ausnahme stellt die Kategorie 2 bei Item 6 (Bestrafungsgefühl) dar, die an keiner Stelle des Traits maximal wird. Die Kategorie 2 wird auch in Item 18 (Appetitveränderung) knapp nicht maximal, und in Item 10 (Weinen) nur in einem sehr kleinen Bereich und dies auch nicht in allen Stichproben. Das Item 21 (Verlust an sexuellem Interesse) misst mehr oder weniger dichotom, zusätzlich ist es bei den Jugendlichen wenig trennscharf.

Ein Blick auf die von Park et al. (2020) publizierten kategorienspezifischen Antwortwahrscheinlichkeiten zeigt, dass auch in der koreanischen Version die Kategorie 2 des Items 6 kein Maximum aufweist und einen sehr ähnlichen Kurvenverlauf hat. Auch in Item 11 und Item 21 sind die Kurven für Kategorie 2 zu niedrig, wohingegen diejenigen für die Items 10 und 18, die in der eigenen Studie kritisch sind, zufriedenstellend ausfallen.

Bezüglich der Kategorienformulierungen lässt sich daraus schließen, dass beim Item 6 (Bestrafungsgefühle) die Kategorie 2 nicht richtig zur aufsteigenden Reihenfolge der übrigen Kategorien passt. Im deutschen BDI-II wird in der Kategorie 2 „Ich erwarte, bestraft zu werden" verwendet, während in den drei anderen Kategorien die Formulierung „... habe das Gefühl..." verwendet wird. Nachdem dieselbe Auffälligkeit auch in der koreanischen Version auftritt, liegt es nahe, dass sie ihren Ursprung bereits in der originalen englischen Version des BDI-II hat. In der Tat wird auch dort für Kategorie 2 „expect" benutzt, während die übrigen Kategorien mit „feel" formuliert sind. Durch die IRT-Analyse der Kategorienwahrscheinlichkeiten lassen sich solche sprachlichen Ungenauigkeiten aufdecken und sie repliziert den schon mittels einer nonparametrischen IRT-Analyse in Beck et al. (1996) erhaltenen Befund. Der praktische Effekt auf die Messqualität des Items dürfte aber gering sein und eine direkte Umformulierung der Kategorie 2 ist nicht erforderlich.

Informationsgehalt der einzelnen Items

Item 14 (Wertlosigkeit) ist in allen Stichproben dasjenige Item mit dem größten Informationsgehalt. In den Erwachsenenstichproben folgen die Items 4 (Verlust an Freude), 12 (Interessenverlust) und 13 (Entschlusslosigkeit), wobei das Item 12 insbesondere bei der Entlassungsstichprobe sehr gut misst. Bei den Jugendlichen hat das Item 7 (Selbstablehnung) in beiden Aufnahmestichproben eine ähnlich hohe Bedeutung wie Item 14 und ebenso in der Entlassungsstichprobe, in der zusätzlich wieder Item 12 hinzukommt. Bei den Erwachsenen scheint Item 7 zwar gut, aber weniger bedeutsam zu sein.

Bei Park et al. (2020) weist das Item 14 ebenfalls die höchste Information auf und nur geringfügig darunter liegen die Items 7 und 1 (Traurigkeit). Niedrige Beiträge liefern die Items 16 (Schlafveränderung), 18 und 21; auch dies deckt sich mit den eigenen Ergebnissen. Bei Zhao et al. (2017) ist Item 14 ebenfalls am informativsten und fast ebenso gut Item 4.

Einen stichprobenübergreifend hohen Informationsgehalt haben also vor allem einige Items aus dem kognitiv-affektiven Bereich, und sie messen auch relativ breit über den Trait hinweg. Nur begrenzt gut erscheinen die Items Schlaf und Appetit, obwohl sie zu den diagnostischen Kriterien gemäß DSM-IV bzw. DSM-5 gehören. Möglicherweise spielt hier die beidseitige Antwortmöglichkeit (mehr Schlaf, weniger Schlaf) eine Rolle. Item 21 liegt in den meisten Studien am unteren Ende bezüglich Informationsgehaltes und Trennschärfe, und insbesondere bei den Jugendlichen scheint es wenig geeignet. Da es sich aber auch nicht negativ auf die Bildung eines Summenwertes auswirkt, sollte es aus Konsistenzgründen nicht ausgeschlossen werden.

Limitationen

Die vorliegenden Analysen wurden an den Gesamtstichproben der jeweiligen Untersuchungszentren durchgeführt. Zwar zeigten die exploratorischen Analysen zur KTT keine bedeutsamen Abweichungen der depressiven Kerngruppe (F3-Hauptdiagnose), dennoch könnten / sollten in weiteren Analysen mögliche Unterschiede bezüglich Messpräzision und Itemqualität zwischen F3-Diagnosen und anderen Diagnosen untersucht werden, was den Rahmen der vorliegenden Arbeit gesprengt hätte. Zu bedenken ist freilich, dass viele Patient_innen eine F3-Nebendiagnose erhielten (bei KODAP 23 %, bei der Rosenеck-Stichprobe Erwachsener 32 %); Depressivität ist außerdem generell bei psychischen Störungen erhöht, was einige auf den ersten Blick geringe Mittelwertsunterschiede im BDI-II Gesamtwert zwischen depressiver Kerngruppe und Gesamtstichprobe miterklärt.

Methodisch wurde für die IRT-Analyse eine (weitgehende) Eindimensionalität angenommen, die sich anhand des Verhältnisses der ersten beiden Eigenwerte bestätigen ließ. Diese ist auch Voraussetzung für eine sinnvolle Interpretation der Summenbildung über die BDI-II-Items. Streng genommen ist die Summenbildung (als suffiziente Statistik) nur bei Gültigkeit des Rasch-Modells bzw. des Partial Credit Model (PCM) gerechtfertigt. Dennoch wurde in vorliegender Arbeit das GRM angewendet, um die Ergebnisse mit jenen von Olino et al. (2012) und Zhao et al. (2017) vergleichen zu können. Eine (hier nicht dargestellte) Kontrollauswertung mit dem Generalisierten PCM (Muraki, 1992) ergab zudem in allen Stichproben einen besseren Fit des GRM im Vergleich zum GPCM. Zwar fanden Alexandrowicz et al. (2014) beim Vergleich einer klinischen und einer nicht-klinischen (studentischen) Stichprobe, dass die Annahmen des PCM weitgehend erfüllt waren und den Personen aus beiden Gruppen in hohem Maße übereinstimmende Personenparameter pro Gesamtwert zugeordnet wurden. Dennoch weisen die hier gewonnenen Ergebnisse darauf hin, dass dem wegen der geschätzten Diskriminationsparameter „flexibleren" GRM gegenüber dem „strengeren" PCM (das konstante Diskrimination annimmt) der Vorzug zu geben ist. Dieser Frage soll in einem weiteren Artikel nachgegangen werden, in dem das PCM mit dem hier verwendeten GRM sowie dem GPCM verglichen und die Auswirkung der Modellwahl auf die Personenparameter im Detail untersucht werden.

Analysen zur Frage, ob eine einfache (ungewichtete) Summenbildung gerechtfertigt ist, sind auch deswegen

wichtig, weil verschiedentlich diskutiert wurde, ob ein einzelner Summenwert aus einem Depressionsmessinstrument das Konstrukt Depression adäquat abdecken kann. Fried et al. (2016) fanden bei vier unterschiedlichen Messinstrumenten, dass jeweils mehr als ein Faktor zur Beschreibung der Zusammenhänge in den Daten nötig war und daher die Annahme von Eindimensionalität zu verwerfen sei, freilich ohne auf die Höhe der Korrelationen zwischen den Faktoren einzugehen. Ein weiteres, in diesem Zusammenhang immer wieder geäußertes Argument lautet, dass ganz unterschiedliche Symptomkombinationen zu gleichen Summenwerten führen können (z. B. Heinrich et al., 2020). Die Suche nach qualitativ unterschiedlichen Subgruppen erbrachte jedoch keine überzeugenden Belege für symptomatisch abgrenzbare Subtypen von Depression (van Loo et al., 2012) und auch für die deutsche Version des BDI-II fanden sich keine Subgruppen, sondern eine weitgehende Gültigkeit der Annahmen des PCM (Keller, 2012). Eine kürzlich publizierte Studie von Stochl et al. (2020), in der unter anderem die Eindimensionalität des PHQ-9 untersucht wurde, nutzte insgesamt sechs Verfahren aus Faktorenanalyse und IRT und fand, dass gemäß PCM von Eindimensionalität ausgegangen werden konnte. Die faktorenanalytischen Verfahren verwiesen auf Multidimensionalität des PHQ-9, aber die Faktoren waren hoch korreliert und legten einen starken generellen Faktor nahe; die Verwendung eines Summenwertes zur Messung des Schweregrades wird daher als gerechtfertigt angesehen (Stochl et al., 2020). Mit der Wahl des GRM haben wir ein empirisch falsifizierbares Modell angewendet. Dessen Adäquatheit erscheint anhand unserer Ergebnisse plausibel, was jedenfalls als Hinweis für eindimensionale Abbildung von Depressivität zu werten ist. Noch differenziertere Befunde werden von der Anwendung des PCM zu erwarten sein.

Zusammenfassend kann geschlossen werden, dass das BDI-II gute Messeigenschaften gemäß der KTT aufweist. Analysen mit IRT erbrachten zusätzliche Aspekte und zeigten vor allem, dass das BDI-II über einen weiten Bereich sehr reliabel misst. Zwischen den verschiedenen Stichproben fanden sich keine ausgeprägten Unterschiede bezüglich der Messpräzision. Die Kategorien der Items waren bis auf die Ausnahme einiger weniger Kategorien, insbesondere Kategorie 2 bei Item 6 (Bestrafungsgefühle), geordnet, wobei bei diesem Item die abweichende Kategorienbeschreibung verantwortlich sein dürfte. Die hohe und relativ gleichbleibende Messpräzision des BDI-II über einen weiten Bereich des Traits weist darauf hin, dass das Verfahren insbesondere im klinischen, aber auch im nicht klinisch relevanten Wertebereich zur Erhebung des Schweregrades einer Depression gut geeignet ist.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/1616-3443/a000676

ESM 1. Statistische und psychometrische Kennzahlen, alle Abbildungen und zusätzliche Auswertungen.

Literatur

Alexandrowicz, R. W., Fritzsche, S. & Keller, F. (2014). Die Anwendbarkeit des BDI-II in klinischen und nicht-klinischen Populationen aus psychometrischer Sicht. Eine vergleichende Analyse mit dem Rasch-Modell. *Neuropsychiatrie, 28,* 63 – 73.

Beck, A. T., Steer, R. A. & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II. San Antonio.* TX: Psychological Corporation.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29 – 51.

Bühler, J., Keller, F. & Läge, D. (2012). Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten. *Zeitschrift für Klinische Psychologie und Psychotherapie, 41*, 231 – 242. https://doi.org/10.1026/1616-3443/a000170

Bühler, J., Keller, F. & Läge, D. (2014). Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression. *Psychological Assessment, 26,* 970 – 979. https://doi.org/10.1037/a0036755

Chalmers, R. P. (2012). mirt: A Multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1 – 29.

de Sá Junior, A. R., Liebel, G., Andrade, A.Gd., Andrade, L. H., Gorenstein, C. & Wang, Y. P. (2019). Can gender and age impact on response pattern of depressive symptoms among college students? A differential item functioning analysis. *Front. Psychiatry* 10:50. https://doi.org/10.3389/fpsyt.2019.00050

Eid, M., Geiser, C., Koch, T. & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22,* 541 – 562. https://doi.org/10.1037/met0000083

Faro, A. & Pereira, C. R. (2020). Factor structure and gender invariance of the Beck Depression Inventory – second edition (BDI-II) in a community-dwelling sample of adults. *Health Psychology and Behavioral Medicine, 8*, 16 – 31. https://doi.org/10.1080/21642850.2020.1715222

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F. & Borsboom, D. (2016). Measuring depression over time … Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28*(11), 1354 – 1367. https://doi.org/10.1037/pas0000275

Hautzinger, M., Bailer, M., Hofmeister, D. & Keller, F. (2012). *Allgemeine Depressionsskala (ADS)* (2., überarbeitete und neu normierte Auflage). Göttingen: Hogrefe.

Hautzinger, M., Keller, F. & Kühner, C. (2006). *BDI-II. Beck-Depressions-Inventar Revision – Manual.* Frankfurt: Harcourt Test Services

Hayes, A. F. & Coutts, J. J. (2020). Use omega rather than Cronbach's Alpha for estimating reliability. But… *Communication Methods and Measures, 14*, 1 – 24. https://doi.org/10.1080/19312458.2020.1718629

Heinrich, M., Zagorscak, P., Eid, M. & Knaevelsrud, C. (2020). Giving G a Meaning: An application of the bifactor-(S-1) Approach to realize a more symptom-oriented modeling of the Beck Depression Inventory-II. *Assessment, 27,* 1429 – 1447. https://doi.org/10.1177/1073191118803738

Huang, C. & Chen, J. H. (2015). Meta-analysis of the factor structures of the Beck Depression Inventory-II. *Assessment, 22(4),* 459 – 472. https://doi.org/10.1177/1073191114548873

Keller, F. (2012). Das Beck-Depressions-Inventar (BDI-II): Psychometrische Analysen mit probabilistischen Testmodellen. In W. Baros & J. Rost (Hrsg.), *Natur- und kulturwissenschaftliche Perspektiven in der Psychologie* (S. 120 – 132). Berlin: Verlag irena regener.

Keller, F., Hautzinger, M. & Kühner, C. (2008). Zur faktoriellen Struktur des deutschsprachigen BDI-II. *Zeitschrift für Klinische Psychologie und Psychotherapie, 37,* 245 – 254.

Keller, F., Kirschbaum-Lesch, I. & Straub, J. (2020). Factor structure and measurement invariance across gender of the Beck Depression Inventory-II in adolescent psychiatric patients. *Front. Psychiatry 11:*527559. https://doi.org/10.3389/fpsyt.2020.527559

Kühner, C., Bürger, C., Keller, F. & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck-Depressionsinventars (BDI-II): Befunde aus deutschsprachigen Stichproben. *Nervenarzt, 78,* 651 – 656.

Kuehner, C., Keller, F., Schricker, I. F., Beddig, T., Huffziger, S., Timm, C., Rachota-Ubl, B., Hautzinger, M. & Diener, C. (2022). Diagnostische Performanz und Validität des deutschsprachigen BDI-II: Eine Sekundäranalyse mit Daten aus klinischen und nichtklinischen Stichproben. *Psychiatrische Praxis.* https://doi.org/10.1055/a-1753-2298

Löwe, B., Spitzer, R. L., Zipfel, S. & Herzog, W. (2002). *Gesundheitsfragebogen für Patienten (PHQ D). Komplettversion und Kurzform.* Testmappe mit Manual, Fragebögen, Schablonen (2. Auflage). Karlsruhe: Pfizer.

Lovibond, S. H. & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales. (2nd. Ed.)* Sydney: Psychology Foundation.

McElroy, E., Casey, P., Adamson, G., Filippopoulos, P. & Shevlin, M. (2018). A comprehensive analysis of the factor structure of the Beck Depression Inventory-II in a sample of outpatients with adjustment disorder and depressive episode. *Irish Journal of Psychological Medicine, 35,* 53 – 61. https://doi.org/10.1017/ipm.2017.52

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159 – 176. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

O'Connor, B. P. (2018). An illustration of the effects of fluctuations in test information on measurement error, the attenuation of effect sizes, and diagnostic reliability. *Psychological Assessment, 30*(8), 991 – 1003. https://doi.org/10.1037/pas0000471

Olino, T. M., Yu , L., Klein, D. N., Rohde, P., Seeley, J. R., Pilkonis, P. A. et al. (2012). Measuring depression using item response theory: An examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research, 21,* 76 – 85. https://doi.org/10.1002/mpr.1348

Osman, A., Kopper, B. A., Barrios, F., Gutierrez, P. M. & Bagge, C. L. (2004). Reliability and Validity of the Beck Depression Inventory-II With Adolescent Psychiatric Inpatients. *Psychological Assessment, 16,* 120 – 132. https://doi.org/10.1037/1040-3590.16.2.120

Ostini, R. & Nerig, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks: SAGE.

Park, K., Jaekal, E., Yoon, S., Lee, S.-H. & Choi, K.-H. (2020). Diagnostic utility and psychometric properties of the Beck Depression Inventory-II among Korean adults. *Front. Psychol, 10,* 2934. https://doi.org/10.3389/fpsyg.2019.02934

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer-Verlag.

Reise, S. P. & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psycholology, 5,* 27 – 48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Rost, J. (2004). *Lehrbuch Testtheorie Testkonstruktion* (2. Aufl.). Bern: Hans Huber.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229 – 244.

Schermelleh-Engel, K. & Gäde, J. C. (2020). Modellbasierte Methoden der Reliabilitätsschätzung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. Aufl.) (S. 335 – 368). Berlin: Springer.

Steer, R. A., Kumar, G., Ranieri, W. F. & Beck, AT. (1998). Use of the beck depression inventory-II with adolescent psychiatric outpatients. *Journal of Psychopathology and Behavioral Assessment, 20*, 127 – 137. https://doi.org/10.1023/A:1023091529735

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., Jones, P. B. & Perez, J. (2020). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment* (Epub ahead of print). https://doi.org/10.1177/1073191120976863

Subica, A., Fowler, J., Elhai, J., Frueh, B., Sharp, C., Kelly, E. & Allen, J. G. (2014). Factor structure and diagnostic validity of the beck depression inventory ii with adult clinical inpatients: Comparison to a gold-standard diagnostic interview. *Psychological Assessment, 4,* 1106 – 1115. https://doi.org/10.1037/a0036998

van Loo, H. M., de Jonge, P., Romeijn, J.-W., Kessler, R. C. & Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: A systematic review. *BMC Medicine, 10*:156. https://doi.org/10.1186/1741-7015-10-156

Velten, J., Bräscher, A.-K., Fehm, L., Fladung, A.-K., Fydrich, T., Heider, J., Hentschel, S., Limberg-Thiesen, A., Lutz, W., Margraf, J., Schöttke, H., Witthöft, M. & Hoyer, J. (2018). Behandlungsdiagnosen in universitären Ambulanzen für psychologische Psychotherapie im Jahr 2016. *Zeitschrift für Klinische Psychologie und Psychotherapie, 47*, 175 – 185. https://doi.org/10.1026/1616-3443/a000490

Voderholzer, U., Hessler-Kaufmann, J. B., Lustig, L. & Läge, D. (2019). Comparing severity and qualitative facets of depression between eating disorders and depressive disorders: Analysis of routine data. *Journal of Affective Disorders, 257,* 758 – 764. https://doi.org/10.1016/j.jad.2019.06.029

Wahl, I., Löwe, B., Bjorner, J. B., Fischer, F., Langs, G., Voderholzer, U., Aita, S. A., Bergemann, N., Brähler, E. & Rose, M. (2014). Standardization of depression measurement: A common metric was developed for 11 self-report depression measures. *Journal of Clinical Epidemiology, 67*, 73 – 86. https://doi.org/10.1016/j.jclinepi.2013.04.019

Wang, Y.-P. & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II (BDI-II): A comprehensive review. *Brazilian Journal of Psychiatry, 35*, 416 – 431. https://doi.org/10.1590/1516-4446-2012-1048

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory–II. *Psychological Assessment, 18,* 81 – 88. https://doi.org/10.1037/1040-3590.18.1.81

Zhao, Y., Chan, W. & Lo, B. C.Y. (2017). Comparing five depression measures in depressed Chinese patients using item response theory: an examination of item properties, measurement precision and score comparability. *Health and Quality of Life Outcomes, 15,* 60. https://doi.org/10.1186/s12955-017-0631-y

Zigmond, A. S. & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67,* 361 – 370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

Historie
Onlineveröffentlichung: 17. 10. 2022

Interessenskonflikt
MH, FK und CK sind Herausgeber der deutschen Version des BDI-II und erhalten dafür Lizenzgebühren.

Förderung
Open Access-Veröffentlichung ermöglicht durch die Universität Ulm.

ORCID
Ferdinand Keller
https://orcid.org/0000-0002-6890-3869

Prof. Dr. Ferdinand Keller, Dipl.-Psych.
Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie
Universitätsklinikum Ulm
Steinhövelstr. 5
89075 Ulm
Deutschland
ferdinand.keller@uniklinik-ulm.de