

# Screening depressiver Störungen mittels Allgemeiner Depressions- Skala (ADS-K) und State-Trait Depressions Scales (STDS-T)

## Eine vergleichende Evaluation von Cut-Off-Werten

Dirk Lehr, Andreas Hillert, Edgar Schmitz und Nadia Sosnowsky

**Zusammenfassung.** Cut-Off-Werte dienen in Screening-Tests einer ökonomischen Identifikation von Erkrankungen bzw. Fällen. In zwei Fall-Kontroll-Studien wurden jeweils Patienten mit depressiven Störungen und klinisch unauffällige Kontrollgruppen untersucht. Zum Screening depressiver Störungen wurde in der ersten Studie ( $N = 296$ ) die *Allgemeine Depressions-Skala* in ihrer Kurzversion (ADS-K), in der zweiten ( $N = 154$ ) die *State-Trait Depression Scales* in ihrer Traitvariante (STDS-T) eingesetzt. Der Anwendungsbereich der Instrumente wurde zunächst in Rasch-Analysen mittels Item-Person-Map untersucht. Hiernach eignen sich beide Instrumente für Populationen mit leicht überdurchschnittlicher bis hin zu klinisch relevanter Depressivität. Receiver-Operating-Characteristics (ROC)-Kurven zeigten hohe diskriminatorische Fähigkeiten für ADS-K und STDS-T an. Unterschiedliche Cut-Off-Werte wurden hinsichtlich ihrer Sensitivität, Spezifität sowie ihres Youden-Indexes evaluiert. Nach dem Youden-Index lag der optimale Cut-Off-Wert für die Diagnostik einer depressiven Störung bei ADS-K  $\geq 18$ , für die STDS-T bei  $\geq 25$  Punkten. Die Analyse liefert eine rationale Grundlage für kategoriale diagnostische Entscheidungen.

Schlüsselwörter: Depression, Item-Response-Theorie, Receiver-Operating-Characteristic (ROC)-Kurve, Sensitivität, Spezifität, Cut-Off-Werte

Assessing depressive disorders using the *Center for Epidemiologic Studies-Depression Scale* (CES-D) and *State-Trait Depression Scales* (STDS-T): A comparative analysis of cut-off scores

**Abstract.** Using cut-off scores in screening tests is an economical way to identify disorders in individuals. Two case-control studies were conducted among healthy and inpatient psychotherapy samples. Patients were suffering from depressive disorders. In the first study ( $N = 296$ ) we used a German short form of the *Center for Epidemiologic Studies-Depression Scale* (CES-D) for screening purposes. In the second study ( $N = 154$ ) we used the trait scale of the *State-Trait Depression Scales* (STDS-T). Based on Rasch model analyses, the item-person maps showed that both instruments are suitable to assess depressive mood in healthy and patient samples. Analyses of the areas under the receiver-operating characteristic curves revealed a high diagnostic power of both instruments. Several cut-off scores were evaluated regarding sensitivity, specificity, and Youden-criterion. Following the Youden-criterion, the optimal cut-off scores for diagnosing depressive disorders were  $\geq 18$  for the CES-D and  $\geq 25$  for the STDS-T. The analyses provide empirical evidence for clinical decision-making.

Key words: depression, item-response-theory, receiver-operating characteristic (ROC) curve, sensitivity, specificity, cut-off scores

## Problemlage

Depressive Störungen gehen mit starkem individuellem Leiden einher und sind gesundheitsökonomisch von hoher Relevanz (Wittchen, Müller, Pfister, Winter & Schmidt, 1999). Die „Global Burden of Disease“ Studie (Mathers, Lopez & Murray, 2006) identifiziert depressive Störungen als eine zentrale Ursache für „verlorene“

Lebensjahre („Disability Adjusted Life Years“), was nicht zuletzt vor dem Hintergrund der hohen Komorbiditätsraten zu einer Vielzahl weiterer Erkrankungen zu erklären ist (Blazer, Kessler, McGonagle & Swartz, 1994). Depressivität stellt einen Negativprädiktor für den Verlauf verschiedener Erkrankungen (z. B. Faller & Schmidt, 2004) oder operativer Eingriffe dar (z. B. Basler & Zimmer, 1997). Die gesundheitsökonomische Bedeutung wird u. a. in langen, durch Depressionen bedingten Arbeitsunfähigkeitszeiten deutlich (z. B. Weber, Hörmann & Köllner, 2006).

Zur Diagnostik depressiver Symptome liegen verschiedene international erprobte Instrumente in deutsch-

Die Autoren danken Heinz-Dieter Basler und Stefan Koch für hilfreiche Kommentare und Anregungen sowie Hans-Helge Müller für seine methodische Unterstützung.

sprachiger Adaptation vor. Diese wurden jedoch für unterschiedliche Zielgruppen entwickelt. Während die *Hospital Anxiety and Depression Scale* (HADS; Herrmann, Buss & Snaith, 1995) auf die Erfassung depressiver Störungen im Zusammenhang mit körperlichen Erkrankungen ausgerichtet ist, findet das *Beck-Depressions-Inventar* (BDI) im Bereich psychisch erkrankter, klinischer Populationen häufige Anwendung (Hautzinger, Bailer, Worrall & Keller, 2000). Eine vereinfachte Version des *Beck-Depressions-Inventar* (BDI-V) wurde für nicht-klinische Populationen entwickelt und auch bei Patienten mit unterschiedlichen psychischen Störungen erfolgreich eingesetzt (Schmitt, Altstötter-Gleich, Hinz, Maes & Brähler, 2006).

Epidemiologische Studien sowie Interventionsstudien in der Schnittmenge von nicht-klinischen, subklinischen und klinischen Populationen benötigen ökonomische und valide Instrumente. Für diese Zielpopulation empfehlen Hautzinger und Bailer (2002) die *Allgemeine Depressions-Skala* (ADS; Hautzinger & Bailer, 1993), eine Adaptation der *Center for Epidemiologic Studies-Depression Scale* (CES-D; Radloff, 1979). Ein weiteres ökonomisches Instrument stellen die *State-Trait Depression Scales* (STDS) in der Trait-Variante (Spaderna, Schmukle & Krohne, 2002) dar. Krohne, Schmukle, Spaderna und Spielberger (2002) berichten hohe Übereinstimmungen zwischen der ADS bzw. CES-D und den STDS ( $r = .62$  bis  $.79$ ). Während die STDS in der Trait-Variante (STDS-T) per definitionem ein stabiles Merkmal messen soll, erfasst die ADS ebenfalls eine dispositionelle Neigung zur Depression, ohne jedoch ihre Fähigkeit einzubüßen, sensitiv auf situationale Veränderungen zu reagieren (Mohiyedini, Hautzinger & Bauer, 2002). Im Sinne der Internationalen Klassifikation psychischer Störungen (ICD-10; Dilling, Mombour & Schmidt, 2000) für den Bereich depressiver Episoden erfragen beide Verfahren sowohl das Vorliegen von negativen Symptomen als auch die Abwesenheit positiver Affekte. Die Ebenen der Depression betreffend (Hautzinger & de Jong-Meyer, 2003) fokussiert die STDS-T ausschließlich affektive Symptome, die ADS-K erfragt darüber hinaus motivationale, kognitive und soziale Symptome.

Spaderna et al. (2002) postulieren, die STDS messe Depressivität bei nicht-klinischen Stichproben, während der Nutzen der ADS bei dieser Population zweifelhaft sei. Im Gegensatz zu anderen Instrumenten könnten die STDS in klinisch unauffälligen Stichproben Depressivität differenziert erfassen (Krohne et al., 2002). Empirische Evidenz für den vermuteten Anwendungsbereich wird jedoch nicht angeführt. Neben dem empfohlenen Einsatzbereich unterscheiden sich beide Verfahren auch in ihrer theoretischen Konzeption. Während die *Allgemeine Depressions-Skala* psychische Beeinträchtigungen im Sinne der symptomatisch-deskriptiven Diagnostik erfasst, basieren die *State-Trait Depression Scales* hingegen auf einer spezifischen Persönlichkeitstheorie, die Annahmen zur Ätiologie depressiver Symptome macht (Krohne et al., 2002).

In epidemiologischen Studien, z. B. zum Zusammenhang zwischen Arbeitsbelastungen und Depression (z. B. Larisch, Joksimovic, vom dem Knesebeck, Starke & Siegrist, 2003), stellt sich regelmäßig die Frage, welche Ausprägung in einer Depressivitätsskala auf eine klinisch relevante Symptomatik hinweist (vgl. Schmitt et al., 2006).

Höfler (2004) betont, dass besonders anwendungsorientierte Forschung vor der Aufgabe stehe, quantitative Testergebnisse auf der Grundlage von Cut-Off-Werten in kategoriale oder binäre Urteile und in anschließende Entscheidungen zu überführen. Eine differenzierte Beschreibung der Gütekriterien von Cut-Off-Werten, z. B. ihre Sensitivität und Spezifität, findet sich jedoch selten für Depressivitätsscreenings. Für die ADS geben Hautzinger und Bailer (1993) wenige Cut-Off-Werte mit der zugehörigen Sensitivität, jedoch ohne die korrespondierende Spezifität an, während Spaderna et al. (2002) für die STDS weder Cut-Off-Werte noch deren Gütekriterien thematisieren. Neben dem Problem, dass Angaben zu Sensitivität und Spezifität nicht standardmäßig zu finden sind, werden sie regelmäßig wie „wahre Werte“ behandelt. Dabei wird außer Acht gelassen, dass Sensitivität und Spezifität nur Schätzwerte darstellen und damit Schwankungen unterliegen. Die routinemäßige Angabe entsprechender Konfidenzintervalle (KI) wird empfohlen (Cohen, 1994) und erlaubt eine Abschätzung der Zuverlässigkeit der Gütekriterien.

Gegenwärtig stellen die wenig evaluierten Cut-Off-Werte der Screenings eine kaum belastbare Datenbasis zur Diagnostik von klinisch relevanten Ausprägungen der Depressivitätswerte dar. Das Fehlen von evaluierten Cut-Off-Werten ist verwunderlich, weil letztlich die Validität zahlreicher Studien und deren versorgungsrelevante sowie gesundheitsökonomische Implikationen davon abhängen.

Zusammenfassend stellt sich folgende Problemlage dar: Mit der Kurzform der *Allgemeinen Depressions-Skala* sowie der Traitversion der *State-Trait Depression Scales* liegen zwei ökonomische Depressivitätsscreenings vor. Unklarheiten bestehen über die Breite des Anwendungsbereiches der Instrumente. Valide Kriterien für die Identifizierung eines Falls liegen derzeit nicht vor. Anwendungsorientierte Forschung benötigt aber eine empirische Basis für kategoriale Diagnostik als Grundlage für nachfolgende Entscheidungsprozesse. Daraus leiten sich die beiden Ziele der vorliegenden Untersuchung ab:

1. Auf der Grundlage der Item-Response-Theorie (IRT) soll der Anwendungsbereich beider Instrumente empirisch untersucht werden. Dazu wird eine gemeinsame Metrik der Personen- und Itemparameter erstellt. Dies wird in der Form einer Item-Person-Map dargestellt.

2. Primäre Zielsetzung der Studie ist die Konstruktion und Evaluation von Cut-Off-Werten. Sofern die Instrumente den Übergang von nicht-klinischer zu klinischer Depressivität messgenau erfassen, sollen verschiedene Cut-Off-Werte im Hinblick auf ihre Gütekriterien evaluiert werden.

## Methode

### Untersuchungsplan und testtheoretische Konzeption

Im Rahmen gemeinsamer Forschungsprojekte der Medizinisch-Psychosomatischen Klinik Roseneck, der Technischen Universität München und der Philipps-Universität Marburg zu psychosozialen Risikofaktoren für psychische Erkrankungen im Beruf wurden die *Allgemeine Depressions-Skala* in ihrer Kurzversion (ADS-K) sowie die *State-Trait Depression Scales* in ihrer Traitvariante (STDS-T) untersucht. Unter Verwendung jeweils eines Verfahrens wurden zwei Fall-Kontroll-Studien (ADS-K-Studie und STDS-T-Studie) durchgeführt, in denen psychisch manifest erkrankte Personen als Fälle definiert wurden.

In einem ersten Schritt wurde die Reliabilität der Verfahren überprüft. Es folgten statistische Analysen, die der Frage nachgingen, ob die Verfahren im Schnittfeld von klinischen und nicht-klinischen Stichproben sinnvoll eingesetzt werden könnten. Die Item-Response-Theorie bietet zur empirischen Spezifizierung der Zielpopulation eines Tests eine geeignete Methodik an. Ein zentraler Vorteil der IRT besteht in der Möglichkeit, eine gemeinsame Metrik („logits“ oder „log odds units“) zu erstellen, auf der sich die Eigenschaften der Items (Itemschwierigkeit) und Personen (Personenparameter oder Fähigkeit) gleichermaßen abbilden lassen (Bond & Fox, 2001). Auf diese Weise kann bestimmt werden, in welchem Bereich der Personenparameter (z. B. Ausprägung von Depressivität) die Items eines Fragebogens den höchsten Informationsgehalt aufweisen bzw. ob Stichprobe und Items zueinander passen (vgl. Gauggel, Böcker, Zimmermann, Privou & Lutz, 2004; Stansbury, Ried & Velozo, 2006).

Depressivitätsscreenings sollten Items enthalten, die den Bereich zwischen den mittleren Personenparametern der klinischen und nicht-klinischen Stichproben messen. Eine hohe Messgenauigkeit in diesem Depressivitätsbereich sollte damit einhergehen, dass sich der gesundheitliche Unterschied zwischen Patienten mit depressiven Störungen und nicht manifest erkrankten Personen valide abbilden lässt. Entsprechend wären sehr große Effektstärken (Cohen, 1988) für die Mittelwertsunterschiede zu erwarten.

Eine Anwendbarkeit der Verfahren in diesem Spektrum der Depressivität vorausgesetzt, können gute diskriminatorische Eigenschaften erwartet werden. Die im Rahmen der „Signal Detection Theory“ (Lusted, 1971) entwickelte „Receiver-Operating-Characteristic“ (ROC)-Kurve bietet eine bewährte Grundlage zur Konstruktion von Cut-Off-Werten (Kraemer, Kazdin, Offord, Kessler, Jensen & Kupfer, 1999; Kramer, 1988; Schäfer, 1989). ROC-Kurven verdeutlichen anschaulich die gegenläufige Beziehung der Testgütekriterien Sensitivität und Spezifität. Bei der Konstruktion von Cut-Off-Werten geht es unter Berücksichtigung von ökonomischen und ethischen Konsequenzen

oder Zielen (Weiß, 2005) um eine wertende Abwägung zwischen:

- hoher Sensitivität, d. h. der Wahrscheinlichkeit, eine tatsächlich depressive Person durch ein positives Testergebnis als Fall zu identifizieren und
- hoher Spezifität, d. h. der Wahrscheinlichkeit, eine nicht depressive Person mittels negativem Testergebnis als Nichtfall zu erkennen (vgl. Höfler, 2004).

Entsprechend sollten Untersuchungen zu Cut-Off-Werten nicht einzelne „optimale“ Testtrennwerte fokussieren, sondern eine Vielzahl von Cut-Off-Werten hinsichtlich ihrer Gütekriterien evaluieren (z. B. Schmitt et al., 2006; Geisser, Roth & Robinson, 1997). Dadurch kann eine rationale Grundlage zur Festsetzung eines Cut-Off-Wertes bereitgestellt werden, der unter Abwägung der Konsequenzen oder Ziele im Einzelfall „optimal“ ist.

Gute diskriminatorische Eigenschaften werden nach Wehberg, Sauerbrei und Schuhmacher (2002) durch die Höhe von Sensitivität und Spezifität sowie die Größe der „Area under the Curve“ (ROC-Analyse) angezeigt. Summarische Validitätsmaße betrachten Sensitivität und Spezifität parallel und eignen sich als heuristisches Optimalitätskriterium zur Bestimmung von Cut-Off-Werten (Abel, 1993). Werden beide Kriterien als gleichwertig betrachtet, stellt das Maximum des Youden-Index ( $Y = \text{Sensitivität} + \text{Spezifität} - 100$ ) ein geeignetes Maß zur Bestimmung von Cut-Off-Werten dar. Bei sämtlichen Validitätskriterien macht die Angabe der entsprechenden Konfidenzintervalle die Zuverlässigkeit der Schätzungen deutlich.

Neben dem Statistikpaket SPSS 12.0 kamen Winsteps 3.61 (Linacre & Wright, 1999) für die Rasch-Analysen und BiAS 8.2 (Ackermann, 2006) sowie MedCalc 8.2.1 (Schoonjans, 2006) für die ROC-Analysen zum Einsatz.

## Instrumente

Kurzversion der *Allgemeinen Depressions-Skala* (ADS-K). Die ADS-K umfasst 15 Aussagen zur Depressivität. Anhand von vierstufigen Ratings (von „selten oder überhaupt nicht“ bis „meistens, die ganze Zeit“) kann die Auftretenshäufigkeit der depressiven Symptome innerhalb der letzten Woche angegeben werden. Durch Addition wird ein Gesamtwert gebildet, der zwischen 0 und 45 liegen kann.

Traitversion der *State-Trait Depression Scales* (STDS-T). Die STDS-T beinhaltet fünf Aussagen, die vorhandene depressive Affekte beschreiben (Dysthymia). Weitere fünf Items beschreiben positive Affekte (Euthymia). Auf einer vierstufigen Häufigkeitsskala (von „fast nie“ bis „fast immer“) wird eingeschätzt, wie man sich „im Allgemeinen“ fühlt. Als abwesende positive Affekte gehen die Items der Euthymia-Komponente mit umgekehrter Codierung in den Gesamtwert ein. Dabei können minimal 10 bis maximal 40 Punkte erreicht werden.

## Stichproben

In die ADS-K-Studie wurden 296 berufstätige Lehrkräfte eingeschlossen. Das durchschnittliche Alter betrug 48.44 Jahre ( $SD = 8.5$ ), 70.2% waren weiblich. Davon befanden sich 97 Personen zum Zeitpunkt der Befragung in stationärer psychotherapeutischer Behandlung. In die Studie wurden konsekutiv Patienten eingeschlossen, die eine depressive Symptomatik aufwiesen. Einschlusskriterien waren die folgenden ICD-10 Diagnosen: F31.30; F32; F33; F34; F43.2(0-2). Die Diagnostik wurde durch den jeweiligen Bezugstherapeuten durchgeführt. Die übrigen 199 Personen nahmen freiwillig an einer Untersuchung zu Arbeit und Gesundheit im Lehrerberuf teil. Sie wiesen zum Zeitpunkt der Befragung keine klinisch relevante Einschränkung des Funktionsniveaus auf, d. h. sie waren arbeitsfähig und werden als nicht-klinische Kontrollgruppe bezeichnet. Die Kontrollgruppe wies in der ADS-K mit 9.40 Punkten ( $SD = 7.83$ ) ein Niveau an Depressivität auf, welches dem Durchschnitt der klinisch unauffälligen Normierungsstichprobe (Hautzinger & Bailer, 1993) entspricht ( $M = 10.72$ ,  $SD = 8.03$ ,  $d = .16$ ). Die Verteilung der Depressivitätswerte ist in Abbildung 1 nach Gruppen getrennt dargestellt.

Die STDS-T-Studie basierte auf den Daten von 154 berufstätigen Lehrkräften. Die Rekrutierung sowie die Ein- und Ausschlusskriterien entsprachen denen der ADS-K-Studie. In diese Studie wurden 50 Patienten mit depressiver Störung und 104 klinisch unauffällige Personen aufgenommen. Das durchschnittliche Alter der überwiegend weiblichen Teilnehmer (68.2%) lag bei 51.83 Jahren ( $SD = 5.94$ ). Die Depressivitätswerte (vgl. Abb. 1) der Kontrollgruppe ( $M = 17.75$ ,  $SD = 4.60$ ) waren unauffällig und lagen im Bereich der von Spaderna et al. (2002) berichteten Werte für nicht-klinische Stichproben ( $M = 19.69$ ,  $SD = 5.25$ ,  $d = .38$ ). In beiden Stichproben war das Vorliegen mindes-

tens einer weiteren komorbiden Störung die Regel (83.5% für ADS-K-Studie, 80% für STDS-T-Studie).

## Ergebnisse

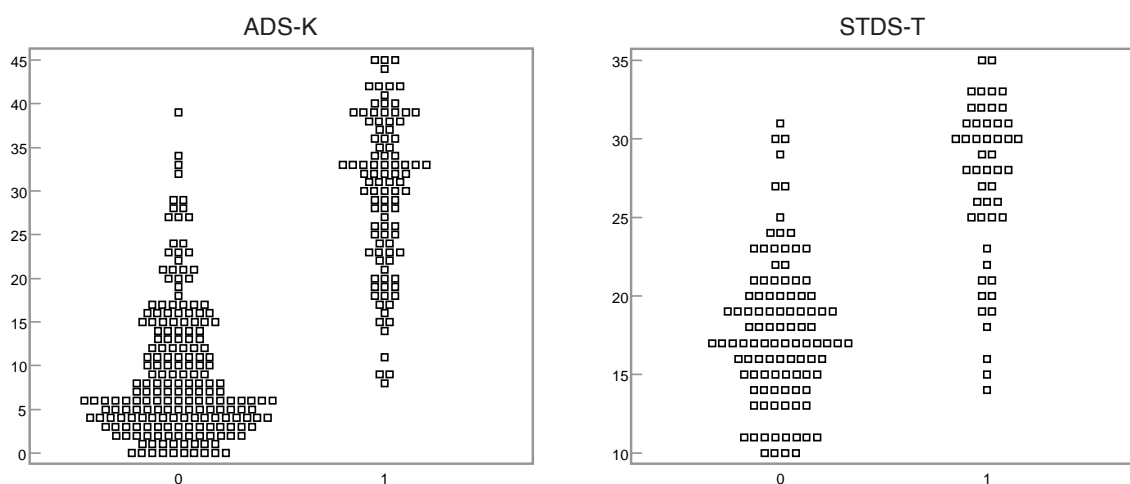
### Reliabilität

Die Reliabilität nach Cronbach-alpha betrug für die ADS-K Alpha = .95, für die STDS-T Alpha = .94. Beide Skalen erwiesen sich als hoch reliabel, wobei jedoch die Werte für Cronbach-alpha an der oberen Grenze des Wünschenswerten liegen und auf mögliche Redundanzen hinweisen (Streiner, 2003). Die Rasch-Reliabilität gibt Auskunft über die Reproduzierbarkeit der Personenkennwerte. Nach Linacre (1997) führt dies zu konservativeren, aber angemessenen Schätzungen der durch den Test aufgeklärten Varianz. Die Rasch-Reliabilität betrug für die ADS-K = .91, für die STDS-T = .92. Die hohe Übereinstimmung bezüglich der Reliabilitätsmaße zeigte sich auch für die Korrelation der Summenwerte mit den Personenparametern (ADS-K:  $r = .96$ , STDS-T:  $r = .99$ ).

### Person-Item-Korrespondenz

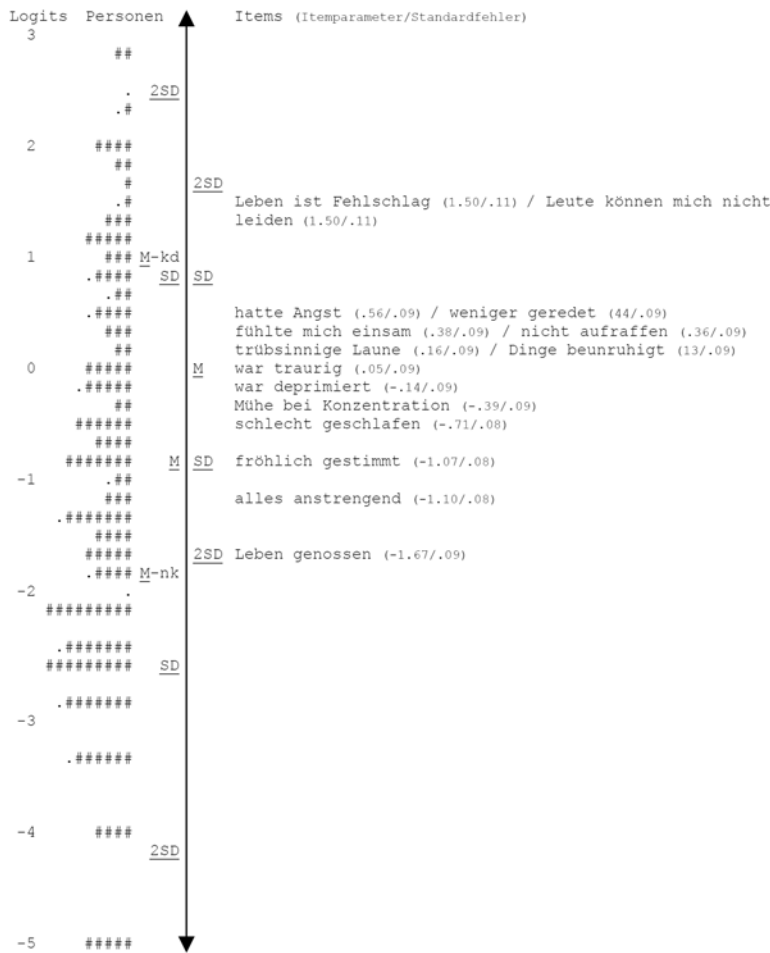
Die Analyse der Daten im Rahmen der IRT basierte auf dem Rating-Skalen-Modell. Dieses Modell geht davon aus, dass die Abstände zwischen aufeinander folgenden Antwortschwellen (z. B. Abstände der Schwellenparameter zwischen den Antwortkategorien *selten/manchmal* oder *manchmal/öfters* oder *öfters/meistens*) innerhalb eines Items unterschiedlich groß sein können. Die Schwellenabstände sind jedoch über die Items hinweg gleich.

Die Abbildungen 2 und 3 zeigen die gemeinsame Metrik (Logit-Skala) der Fähigkeitswerte der Personen (links) und der Itemschwierigkeiten (rechts). Bei der Kalibrierung



ADS-K = Kurzform der Allgemeinen Depressionsskala. STDS-T = Traitvariante der State-Trait Depression Scales. 0 = nicht-klinische Stichprobe, 1 = klinische Stichprobe. Jedes Quadrat repräsentiert eine Person.

Abbildung 1. Depressivitätswerte der klinischen und nicht-klinischen Stichproben.



Anmerkungen: jedes „#“ entspricht 2 Personen, jedes „.“ einer Person. *M* = Mittelwert; *SD* = eine Standardabweichung, *2SD* = zwei Standardabweichungen; *M-nk* = Mittelwert der nicht-klinischen Stichprobe, *M-kd* = Mittelwert der klinisch-depressiven Stichprobe; in Klammern sind Itemparameter mit Standardfehler angegeben.

Abbildung 2. Item-Person-Map für die ADS-K.

wurde der mittleren Itemschwierigkeit der Wert Null zugewiesen. Ein höherer Itemparameter zeigt ein schwierigeres Item an, ein größerer Personenkennwert eine stärkere Depressivität.

Die Infit-Statistik (infit mean square criteria, *Infit-MNSQ*) wird zur Kontrolle der Modellpassung vorgeschlagen (Wright & Linacre, 1994). Für Ratingsskalen sollten die Items einen *Infit-MNSQ*-Wert von 0.6 bis 1.4 aufweisen (Bond & Fox, 2001). Einen optimalen Wert stellt *Infit-MNSQ* = 1.0 dar. Niedrigere Werte weisen auf redundante Informationen hin. Eine hoher Wert von *Infit-MNSQ* = 1.4 bedeutet, dass 40 Prozent mehr Rauschen vorliegt als dies unter Annahme der Modellgeltung zu erwarten wäre. Hohe Werte stellen die Eindimensionalität des Konstrukts in Frage.

In der ADS-K wurde ein Item („habe ich schlecht geschlafen“) mit schlechter Infit-Statistik (*Infit-MNSQ* = 1.80) identifiziert. Ein zweites Item („hatte ich das Gefühl,

dass mich die Leute nicht leiden können“) zeigte einen grenzwertigen Infit-Wert (*Infit-MNSQ* = 1.33). Die übrigen Items lagen im empfohlenen Bereich (*Infit-MNSQ* = .60 bis 1.23). Die Items der STDS-T wiesen insgesamt gute Infit Werte auf (*Infit-MNSQ* = .70 bis 1.21). Für beide Verfahren konnte ein monotoner Anstieg der Schwellenparameter über die jeweils vier Antwortkategorien beobachtet werden. Nach Linacre (1999) gibt die Outfit-Statistik (outfit mean square criteria, *Outfit-MNSQ*) darüber Auskunft, ob einzelne Antwortkategorien modelldiskonforme Informationen liefern. In beiden Verfahren wurde der kritische Wert von *Outfit-MNSQ* = 2.0 für jede Antwortkategorie deutlich unterschritten. Dies zeigt eine adäquate Nutzung des vierstufigen Antwortformats an.

Für beide Verfahren galt, dass die Mehrzahl der Items ihren optimalen Trennbereich zwischen dem mittleren Personenparameter der nicht-klinischen und der depressiven Gruppe aufwies. Einzelne Items differenzierten in Bereichen, die über den mittleren Personenparameter der depressiven Gruppen lagen und damit eine starke Depressivität anzeigten. Sehr leichte Items, die den Bereich schwacher Depressivität erfassten, waren in beiden Verfahren nicht vorhanden. Items, die die Abwesenheit positiver Affektivität indizierten, wiesen in beiden Verfahren konsistent niedrige Itemparameter auf.

### Trennung der Gruppenmittelwerte

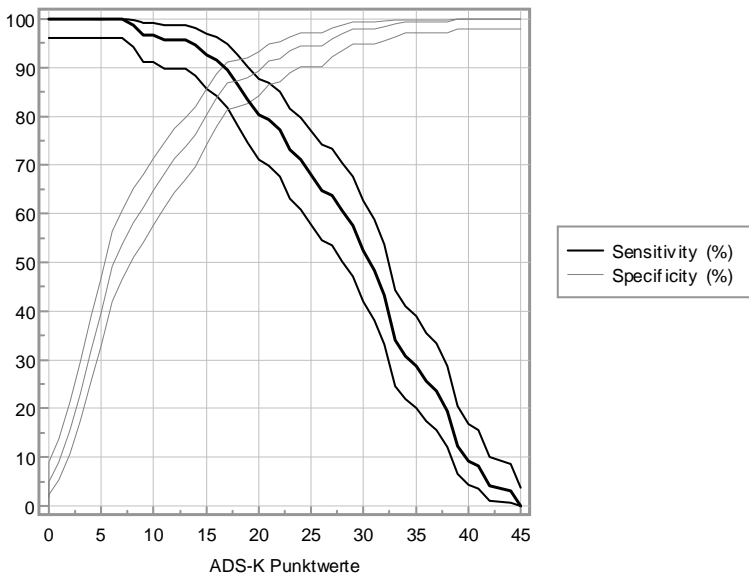
Der ADS-K-Wert lag für die nicht-depressive Vergleichsgruppe bei 9.40 (*SD* = 7.83), während die Patientengruppe einen mittleren Wert von 29.65 (*SD* = 9.00) aufwies. Einen nahezu identischen Wert berichten Hautzinger und Bailer (1993) für 86 depressive Patienten. Der Gruppenunterschied war signifikant ( $T = -19.88, df = 294, p < .001$ ) und entspricht nach Cohen (1988) einem starken Effekt von  $d = 2.45$ . Zur Berechnung der Effektstärke wurde die gepoolte Varianz bzw. Standardabweichung verwendet (Bortz, 1999).

In Bezug auf die STDS-T wies die Patientengruppe einen signifikant höheren Gesamtwert ( $M = 27.06, SD = 5.35$ ) gegenüber der Vergleichsgruppe ( $M = 17.75; SD = 4.60$ ) auf ( $T = -11.14, df = 152, p < .001$ ). Der Mittelwertsunterschied entsprach ebenfalls einem starken Effekt ( $d = 1.90$ ).

### Receiver-Operating-Characteristics

In die ROC-Analysen ging das Vorliegen einer depressiven Störung als Gruppenvariable (klinische Gruppe vs. Kontrollgruppe) und der ADS-K-Summenwert bzw. STDS-T-Gesamtwert als diagnostische Variablen ein. Die





Anmerkungen: ADS-K = Kurzform der Allgemeinen Depressionsskala. Die mittleren Linien stellen den Verlauf von Sensitivität und Spezifität dar. Durch die äußeren Linien wird die Breite des Konfidenzintervalls (95%) verdeutlicht.

Abbildung 5. Verlauf der Sensitivität und Spezifität über verschiedene Cut-Off-Werte der ADS-K.

zogene Person. Die Kurve näherte sich der linken oberen Ecke und zeigte die hohe diskriminatorische Güte der ADS-K an.

Im Vergleich dazu verlief die Kurve der STDS-T weniger prototypisch (vgl. Abb. 2), nach steilem Anstieg knickte sie vor dem Erreichen der linken oberen Ecke ab und verlief danach in Form einer Geraden. Dies ist auf eine „lokale Konstanz“ der Sensitivität im Bereich STDS-T = 22 bis 25 zurückzuführen. Der AUC-Wert betrug  $AUC = .89$  und unterschied sich signifikant von  $AUC = .50$  ( $Z = 7.85, p < .001, KI95\% = .83-.95$ ). Unter Berücksichtigung des Konfidenzintervalls wies die STDS-T nach Swets (1988) eine mittlere bis hohe diskriminatorische Fähigkeit auf.

### Cut-Off-Werte

Abbildung 5 macht für die ADS-K exemplarisch deutlich, wie mit einem höheren Trennwert ein Ansteigen der Spezifität und ein Abfallen der Sensitivität einhergehen. In ähnlicher Weise geht aus Abbildung 1 hervor, dass ein höherer Trennwert die Personen der nicht-klinischen

Tabelle 1. Gütekriterien der Cut-Off-Werte für ADS-K und STDS-T

	Cut-Off	Sensitivität	KI-u	KI-o	Spezifität	KI-u	KI-o	Youden	KI-u	KI-o
ADS-K	≥ 13	95.9	89.8	98.9	71.4	64.5	77.5	67.2	59.2	75.2
	≥ 14	95.9	89.8	98.9	73.9	67.2	79.8	69.6	62.0	77.5
	≥ 15	94.8	88.4	98.3	76.4	69.9	82.1	71.2	63.6	78.8
	≥ 16	92.8	85.7	97.1	80.4	74.2	85.7	73.2	65.7	80.7
	≥ 17	91.8	84.4	96.4	83.9	78.1	88.7	75.7	68.4	83.0
	≥ 18	<b>89.7</b>	<b>81.9</b>	<b>94.9</b>	<b>86.9</b>	<b>81.5</b>	<b>91.3</b>	<b>76.6</b>	<b>69.2</b>	<b>84.0</b>
	≥ 19	86.6	78.2	92.7	87.4	82.0	91.7	74.0	66.0	82.0
	≥ 20	83.5	74.6	90.3	87.9	82.6	92.1	71.4	63.0	79.9
	≥ 21	80.4	71.1	87.8	89.4	84.3	93.4	69.9	60.9	78.8
	≥ 22	79.4	70.0	86.9	91.5	86.7	94.9	70.8	61.8	79.9
≥ 23	77.3	67.7	85.2	92.0	87.3	95.3	69.3	59.9	79.7	
STDS-T	≥ 20	88.0	75.7	95.5	70.2	60.4	78.8	58.2	45.7	70.7
	≥ 21	84.0	70.9	92.8	76.9	67.6	84.6	60.9	48.4	73.4
	≥ 22	80.0	66.3	90.0	82.7	74.0	89.4	62.7	49.9	75.5
	≥ 23	78.0	64.0	88.5	84.6	76.2	90.9	62.6	49.6	75.7
	≥ 24	76.0	61.8	86.9	90.4	83.0	95.3	66.4	53.0	79.7
	≥ 25	<b>76.0</b>	<b>61.8</b>	<b>86.9</b>	<b>93.3</b>	<b>86.6</b>	<b>97.3</b>	<b>69.3</b>	<b>55.9</b>	<b>82.6</b>
	≥ 26	68.0	53.3	80.5	94.2	87.9	97.9	62.2	47.2	77.3
	≥ 27	62.0	47.2	75.4	94.2	87.9	97.9	56.2	40.0	70.4
≥ 28	58.0	43.2	71.8	96.2	90.4	98.9	54.2	36.9	71.3	

Anmerkungen: ADS-K = Kurzform der Allgemeinen Depressionsskala; STDS-T = Traitvariante der State-Trait Depression Scales. Cut-Off ≥ X bedeutet, dass eine depressive Störung bei einem Wert von X oder einen höheren Wert vorliegt. KI-u, KI-o = untere bzw. obere Grenze des 95% Konfidenzintervalls. Youden-Index = Sensitivität + Spezifität - 100.

Stichprobe spezifischer korrekt erkennt, aber immer weniger sensitiv für die Identifikation depressiver Personen wird.

Wurden beide Gütekriterien simultan maximiert, d. h. Sensitivität und Spezifität als gleichwertig betrachtet, erreichte der Youden-Index bei ADS-K  $\geq 18$  Punkten sein Maximum. Dieser Wert kann als optimaler Cut-Off-Wert empfohlen werden. Die Effizienz, d. h. der Anteil insgesamt richtiger Klassifikationsentscheidungen, lag an diesem Punkt bei 87.8 %.

Für die STDS-T erreichte der Youden-Index bei einem Trennwert von  $\geq 25$  Punkten sein Maximum. Die Effizienz betrug dabei 87.6 %. Während sich Sensitivität und Spezifität am optimalen Punkt für die ADS-K symmetrisch verhielten, überstieg die Spezifität ihre korrespondierende Sensitivität in der STDS-T deutlich. Die geschätzte Sensitivität und Spezifität für verschiedene Cut-Off-Werte sind in Tabelle 1 unter Angabe der Konfidenzintervalle aufgeführt.

## Diskussion

Im Mittelpunkt der vorliegenden Untersuchung standen zwei Instrumente zur Selbsteinschätzung der Depressivität, die Kurzform der *Allgemeinen Depressions-Skala* und die *State-Trait Depression Scales* in ihrer Trait Variante. Ziel der Studie war die Evaluation von Cut-Off-Werten zur Identifizierung eines klinisch relevanten Ausprägungsgrades an Depressivität.

## Anwendungsbereich der Instrumente

In einem ersten Schritt wurde der Anwendungsbereich der Instrumente auf der Grundlage der Item-Response-Theorie exploriert. Die Berücksichtigung der Korrespondenz von Personen und Items auf der gemeinsamen Metrik (Logit-Skala, Item-Person-Map) stellt eine intuitive sowie nützliche Perspektive auf psychologische Testverfahren dar. Die gemeinsame Metrik erlaubt empirisch begründete Empfehlungen zum Anwendungsbereich eines Tests zu geben.

In beiden Instrumenten stiegen die Items in ihrer Schwierigkeit monoton über etwa zwei Standardabweichungen des Personenparameters an und deckten damit ein weites Spektrum der Depressivität ab. In Übereinstimmung mit den Befunden von Stansbury et al. (2006) demonstrierte die Abwesenheit positiver Erlebnisinhalte konsistent ein geringeres Maß an Depressivität als das Vorliegen von Negativmerkmalen.

Das leichteste Item lag jeweils über dem Mittelwert der nicht-klinischen Kontrollgruppe, während die schwersten über dem mittleren Depressivitätswert der klinischen Gruppe lagen. Im optimalen Messbereich der Verfahren wurde ein leicht überdurchschnittliches bis klinisch bedeutsames Ausmaß an Depressivität erfasst. Dies bestä-

tigt Hautzingers und Bailers (2002) Empfehlungen zum Anwendungsbereich der ADS-K. Entgegen der Annahme von Spaderna et al. (2002) und Krohne et al. (2002) differenzierten STDS-T und ADS-K im nicht-klinischen Bereich gleichermaßen. Unterdurchschnittliche Depressivität konnte durch beide Instrumente allerdings nur suboptimal gemessen werden. Eine Beschränkung des Anwendungsbereichs der STDS-T auf klinisch unauffällige Populationen legt die Item-Person-Map nicht nahe.

Innerhalb des optimalen Messbereiches wies die STDS-T eine hohe Ökonomie auf. Die Items stiegen monoton an und wiesen dabei weder größere Messlücken noch Redundanzen auf. Die ADS-K zeigte innerhalb ihres Messbereiches ebenfalls keine Lücken. Jedoch wurden einige Stellen des Kontinuums durch mehrere Items redundant erfasst. Die Redundanzen decken sich mit den Untersuchungen von Stansbury et al. (2006) und dienen Cole, Rabin, Smith und Kaufman (2004) zur Konstruktion einer raschbasierten, ökonomischen CES-D Kurzform.

Entsprechend des identifizierten optimalen Anwendungsbereichs beider Verfahren kann der gesundheitliche Unterschied zwischen Patienten mit depressiven Störungen und der entsprechenden Kontrollgruppe valide abgebildet werden. Dies wird daran deutlich, dass sich die Gruppen durch etwa zwei Einheiten der Effektgröße Cohens  $d$  unterscheiden.

## Evaluation der Cut-Off-Werte

Die „Signal Detection Theory“ bildete den theoretischen Rahmen, um in einem zweiten Schritt die diagnostischen Eigenschaften der ADS-K und STDS-T zu untersuchen. Cut-Off-Werte für die ADS-K und die STDS-T wurden einer umfangreichen Evaluation unterzogen. Dies ist v. a. für anwendungs- und patientenorientierte Forschung von hoher Bedeutung und validiert die Instrumente an „echten“ (Amelang & Schmidt-Atzert, 2006), nicht auf reiner Selbsteinschätzung beruhenden, Kriterien.

Sowohl die ADS-K als auch die STDS-T bieten die Möglichkeit, Cut-Off-Werte zu nutzen, bei denen die Sensitivität und Spezifität gleichzeitig über 80 Prozent liegen. Werden Sensitivität und Spezifität als gleichwertig betrachtet und in ein summarisches Gütekriterium, den Youden-Index, überführt, resultierte für die ADS-K ein optimaler Cut-Off-Wert von  $\geq 18$  Punkten und von  $\geq 25$  Punkten für die STDS-T. Auf der Grundlage der vorgelegten Evaluation können diese Cut-Off-Werte je nach Intention, z. B. zur Identifikation depressiver oder „gesunder“ Personen, modifiziert werden (Weiß, 2005).

Cut-Off-Werte lenken die Aufmerksamkeit verstärkt auf die in den Tests erreichten Absolutwerte. Mit den Cut-Off-Werten werden valide Bezugspunkte vorgestellt, die eine Bewertung der gesundheitlichen Relevanz erlauben. Eine rein relationale Perspektive, z. B. Bezug auf verteilungsbasierte Cut-Off-Werte, Analyse von Korrelationen oder Mittelwertsdifferenzen, kann dies nicht sicherstellen.



Eine wichtige Frage stellt die Validität bzw. Generalisierbarkeit der ermittelten Cut-Off-Werte und ihrer Gütekriterien dar. Sensitivität und Spezifität werden häufig als statische Eigenschaften des Diagnostikums interpretiert. Allerdings hängen die Gütekriterien von den Charakteristika der Stichproben ab (Guggenmoos-Holzmann & van Houwelingen, 2000; Kramer, 1988). Die Sensitivität wird durch Charakteristika der klinischen, die Spezifität durch Merkmale der nicht-klinischen Stichprobe beeinflusst (Spektrum-Bias). Für beide Stichproben der ADS-K-Studie sowie die klinisch unauffällige STDS-Stichprobe konnten keine bedeutsamen Unterschiede zu den Referenzwerten (Hautzinger & Bailer, 1993; Spaderna et al., 2002) festgestellt werden. Entsprechend liegen keine Hinweise auf einen Spektrum-Bias vor. Ein möglicher Spektrum-Bias für die Sensitivität der STDS-T konnte nicht überprüft werden, da das Instrument erstmalig in einer klinischen Stichprobe eingesetzt wurde.

Die allgemein hohe Komorbiditätsrate bei depressiven Störungen (Blazer et al., 1994) war auch in den vorliegenden Studien zu beobachten. Die Sensitivität an den verschiedenen Cut-Off-Werten weist damit eine hohe ökologische Validität auf und eignet sich für ein breites Screening depressiver Störungen.

In ROC-Analysen wurde die wechselseitige Beziehung von Sensitivität und Spezifität untersucht. Die ROC-Kurve der ADS-K verlief idealtypisch, der Wert für die AUC war groß (vgl. Swets, 1988) und zeigte eine sehr gute klassifikatorisch-diagnostische Eigenschaft an. Die Konfidenzbänder um die ROC-Kurve waren eng, sodass die geschätzten Gütekriterien als zuverlässig angesehen werden können. Die ADS-K bietet mit ihren 15 vierstufigen Items eine vergleichsweise hohe Anzahl (45) von Abstufungsmöglichkeiten. Dies schlug sich auch darin nieder, dass Sensitivität und Spezifität in feineren Schritten adjustiert werden konnten. Im Vergleich dazu waren die Veränderungen in den Gütekriterien der STDS-T beim Wechsel zu benachbarten Cut-Off-Werten z. T. erheblich und führten zu einer gröberen wechselseitigen Adjustierung.

Die AUC der STDS-T kann als mittel bis groß bezeichnet werden, gute bis sehr gute diskriminatorischen Fähigkeiten anzeigend. Bedingt durch den geringeren Stichprobenumfang in der STDS-T-Studie waren die Konfidenzbänder um die ROC-Kurve breiter. Wurde die ROC-Kurve lokal für den Bereich der evaluierten Cut-Off-Werte betrachtet (vgl. Zweig & Campbell, 1993), so war die untere Grenze des Konfidenzintervalls jedoch deutlich von der Winkelhalbierenden entfernt. Bezüglich der Größe der AUC und der Höhe des Youden-Index zeigte die ADS-K deskriptiv die besseren Screeningfähigkeiten. Vor dem Hintergrund des Studiendesigns kann dies als weiter zu explorierende Tendenz, jedoch nicht als belastbarer Instrumentenvergleich bewertet werden.

Mit der *Allgemeinen Depressions-Skala* in ihrer Kurzform und den *State-Trait Depression Scales* in ihrer Trait Variante stehen zwei Screeningverfahren zur Verfügung, die hoch reliabel und ökonomisch ein überdurchschnitt-

liches bis klinisch hoch bedeutsames Ausmaß an Depressivität erfassen. Ihre Anwendung in epidemiologischen Studien zur Identifikation depressiver Störungen kann empfohlen werden. Die Instrumente sind besonders für Differenzierungen im Bereich leichter, subklinischer bis klinisch relevanter Depressivität geeignet. In Interventionsstudien könnte z. B. der Anteil von Personen bestimmt werden, der nach einem klinischen Ausmaß an Depressivität wieder unauffällige Werte erreicht. Beide Instrumente eignen sich vorwiegend zu Erfassung der affektiven Ebene der Depressivität. Motivationale, kognitive und soziale Symptome werden hingegen nur durch die ADS-K erfasst. Ungeachtet ihrer theoretisch-konzeptionellen Unterschiede erfüllen beide Verfahren ihre diskriminatorische Funktion gleichermaßen gut. Tendenziell zeigt die ADS-K eine höhere Diskriminationsstärke und bietet die Möglichkeit, Sensitivität und Spezifität genauer aufeinander abzustimmen. Vorteile der STDS-T liegen in ihrer höheren Ökonomie. Die vorliegende Studie zeigt, dass beide Instrumente valide das Vorliegen oder aber die Abwesenheit einer depressiven Störung anzeigen können. In Ergänzung zu der häufig eingenommenen relationalen Perspektive sollten sorgfältig evaluierte Cut-Off-Werte, z. B. in Studien zum Zusammenhang von Arbeitswelt und Gesundheit, eine stärkere Berücksichtigung bei der Bewertung von Studienergebnissen finden.

## Literatur

- Abel, U. (1993). *Die Bewertung diagnostischer Tests*. Stuttgart: Hippokrates-Verlag.
- Ackermann, H. (2006). *BIAS für Windows*. Hochheim Darmstadt: Epsilon Verlag.
- Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention*. Berlin: Springer.
- Basler, H. D. & Zimmer, C. (1997). Does dysphoric mood really predict the outcome of lumbar surgery? Methodological pitfalls in psychological research. *European Journal of Pain*, 1, 197–205.
- Blazer, D. G., Kessler, R. C., McGonagle, K. A. & Swartz, M. S. (1994). The prevalence and distribution of major depression in a national community sample: the National Comorbidity Survey. *American Journal of Psychiatry*, 151, 979–986.
- Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- Bortz, J. (1999). *Statistik* (5. Aufl.). Berlin: Springer.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997–1003.
- Cole, J. C., Rabin, A. S., Smith, T. L. & Kaufman, A. S. (2004). Development and validation of a rasch-derived CES-D short form. *Psychological Assessment*, 16, 360–372.
- Dilling, H., Mombour, W. & Schmidt, M. H. (2000). *Internationale Klassifikation psychischer Störungen. ICD-10* (4. Aufl.). Bern: Huber.
- Faller, H. & Schmidt, M. (2004). Prognostic value of depressive coping and depression in survival of lung cancer patients. *Psycho-Oncology*, 13, 359–363.
- Gaugel, S., Böcker, M., Zimmermann, P., Privou, C. & Lutz, D. (2004). Item-response-Theorie und deren Anwendung in der Neurologie: Messung von Aktivitätseinschränkungen neurologischer Patienten. *Nervenarzt*, 75, 1179–1186.

- Geisser, M. E., Roth, R. S. & Robinson, M. E. (1997). Assessing depression among persons with chronic pain using the Center for Epidemiological Studies-Depression Scale and the Beck Depression Inventory: A comparative analysis. *The Clinical Journal of Pain, 13*, 163–170.
- Guggenmoos-Holzmann, I. & van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity. *Statistics in medicine, 19*, 1783–1792.
- Hautzinger, M. & Bailer, M. (1993). *Allgemeine Depressions Skala – ADS*. Weinheim: Beltz.
- Hautzinger, M., Bailer, M., Worrall, H. & Keller, F. (2000). *Beck-Depressions-Inventar (BDI)*. Testhandbuch (3. Aufl.). Bern: Huber.
- Hautzinger, M. & Bailer, M. (2002). ADS – Allgemeine Depressions Skala. In E. Brähler, J. Schumacher & B. Strauß (Hrsg.), *Diagnostische Verfahren in der Psychotherapie* (S. 25–27). Göttingen: Hogrefe.
- Hautzinger, M. & de Jong-Meyer, R. (2003). Depression. In H. Reinecker (Hrsg.), *Lehrbuch der Klinischen Psychologie und Psychotherapie. Modelle psychischer Störungen* (S. 215–257). Göttingen: Hogrefe.
- Herrmann, C., Buss, U. & Snaith, R. P. (1995). *Hospital Anxiety and Depression Scale – Deutsche Version (HADS-D)*. Manual. Bern: Huber.
- Hilgers, R. A. (1991). Distribution-free confidence bounds for ROC curves. *Methods of Information in Medicine, 30*, 96–101.
- Höfler, M. (2004). *Statistik in der Epidemiologie psychischer Störungen*. Berlin: Springer.
- Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S. & Kupfer, D. J. (1999). Measuring the potency of a risk factors for clinical or policy significance. *Psychological Methods, 4*, 257–271.
- Kramer, M. S. (1988). *Clinical epidemiology and biostatistics: a primer for clinical investigators and decision makers*. New York: Springer.
- Krohne, H. W., Schmukle, S. C., Spaderna, H. & Spielberger, C. D. (2002). The State-Trait Depression Scales: An international comparison. *Anxiety, Stress, and Coping, 15*, 105–122.
- Larisch, M., Joksimovic, L., Knesebeck, O., Starke, D. & Siegrist, J. (2003). Berufliche Gratifikationskrisen und depressive Symptome: Eine Querschnittstudie bei Erwerbstätigen im mittleren Erwachsenenalter. *Psychotherapie, Psychosomatik und Medizinische Psychologie, 53*, 223–228.
- Linacre, J. M. (1997) KR-20 / Cronbach Alpha or Rasch Reliability: Which Tells the „Truth“? *Rasch Measurement Transactions, 11*, 580–581.
- Linacre, J. M. (1999). Understanding Rasch measurement: estimation methods for Rasch measures. *Journal of outcome measurement, 3*, 382–405.
- Linacre, J. M. & Wright, B. D. (1999). *WINSTEPS for PC compatibles*. Chicago: Mesa Press.
- Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science, 171*, 1217–1219.
- Mathers, C. D., Murray, C. J. L. & Lopez, A. D. (2006). The Burden of Disease and Mortality by Condition: Data, Methods and Results for the Year 2001. In A. D. Lopez, C. D. Mathers, M. Ezzati, D. T. Jamison & C. J. L. Murray (Eds.), *Global Burden of Disease and Risk Factors* (pp. 45–240). New York: Oxford University Press.
- Mohiyeddini, C., Hautzinger, M. & Bauer, S. (2002). Eine Latent State-Trait Analyse zur Bestimmung der dispositionellen und zustandsbedingten Anteile dreier Instrumente zur Erfassung von Depressionen: ADS, BDI und SDS. *Diagnostica, 48*, 12–18.
- Radloff, L. S. (1979). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.
- Schäfer, H. (1989). Constructing a cut-off point for a quantitative diagnostic test. *Statistics in Medicine, 8*, 1381–1391.
- Schmitt, M., Altstötter-Gleich, C., Hinz, A., Maes, J. & Brähler, E. (2006). Normwerte für das Vereinfachte Beck-Depressions-Inventar (BDI-V) in der Allgemeinbevölkerung. *Diagnostica, 52*, 51–59.
- Schoonjans, F. (2006). *MedCalc Software (Version 8.2.1)*. Mariakerke, Belgium.
- Spaderna, H., Schmukle, S. C. & Krohne, H. W. (2002). Bericht über die deutsche Adaption der State-Trait Depression Scales (STDS). *Diagnostica, 48*, 80–89.
- Stansbury, J. P., Ried, L. D. & Velozo, C. (2006). Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES-D) scale. *Journal of Personality Assessment, 86*, 10–22.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality assessment, 80*, 99–103.
- Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293.
- Weber, A., Hörmann, G. & Köllner, V. (2006). Psychische und Verhaltensstörungen – Die Epidemie des 21. Jahrhunderts? Der Erhalt oder die Wiederherstellung der seelischen Gesundheit müssen als gesamtgesellschaftliche Aufgabe begriffen werden. *Deutsches Ärzteblatt, 103*, 834–841.
- Wehberg, S., Sauerbrei, W. & Schumacher, M. (2002). Diagnosestudien. In M. Schumacher & G. Schulgen (Hrsg.), *Methodik klinischer Studien* (S. 291–311). Berlin: Springer.
- Weiß, C. (2005). *Basiswissen Medizinische Statistik* (3. Aufl.). Berlin: Springer.
- Wittchen, H. U., Müller, N., Pfister, H., Winter, S. & Schmidt-kunz, B. (1999). Affektive, somatoforme und Angststörungen in Deutschland – Erste Ergebnisse des bundesweiten Zusatzsurveys „Psychische Störungen“. *Gesundheitswesen, 61*, 216–222.
- Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.
- Zweig, M. H. & Campbell, G. (1993). Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561–577.

---

Dipl. Psych. Dirk Lehr

---

Philipps-Universität Marburg  
Zentrum für Methodenwissenschaften und  
Gesundheitsforschung  
Institut für Medizinische Psychologie  
Bunsenstr. 3  
35037 Marburg  
E-Mail: Dirk.Lehr@med.uni-marburg.de

---

PD Dr. Dr. Andreas Hillert

---

Medizinisch-Psychosomatische Klinik Roseneck  
Am Roseneck 6  
83209 Priem am Chiemsee

---

em. Prof. Dr. Edgar Schmitz

---

Technische Universität  
L. f. Psychologie  
Lothstraße 17  
80335 München

---

Dipl.-Psych. Nadia Sosnowsky

---

Institut für Pädagogische Psychologie und Soziologie  
Pädagogische Hochschule Ludwigsburg  
Reuteallee 46  
71634 Ludwigsburg