

C. Semmler · P. L. Klumb

Geschlechtsunterschiede in der Prävalenz depressiver Symptomatik: Ein Resultat differentieller Validität der Erhebungsinstrumente?

Eingegangen: 24 Juli 2003 / Angenommen: 12 Januar 2004 / Online veröffentlicht: 7 April 2004
© Springer-Verlag 2004

Zusammenfassung In dieser Studie wurde die prädiktive Validität der 20 Items der deutschen Version der Center for Epidemiological Studies-Depression Scale (CES-D, ADS) geschlechterbezogen untersucht. Mit den Antworten von 516 Teilnehmerinnen und Teilnehmern der Berliner Altersstudie (70- bis 103-jährige Frauen und Männer) wurden Itemanalysen durchgeführt. Für Männer waren die meisten Items schwieriger und weniger trennscharf als für Frauen. Darüber hinaus wurden die Itemwerte zu psychiatrischen Depressionsdiagnosen nach DSM-III-R in Beziehung gesetzt, um deren Sensitivität und Spezifität für beide Geschlechter zu ermitteln. Die Itemsensitivitäten fielen eher niedrig aus und unterschieden sich für Frauen und Männer sehr stark. Diese Unterschiede spiegeln sich in der Sensitivität des Gesamtwerts (Cut-Off: 16) wider, die Konfidenzintervalle der Schätzungen überlappten jedoch. Mit Regressionsanalysen wurden die Itemantworten durch das Geschlecht der Person, die Depressionsdiagnose und die Interaktion dieser Prädiktoren vorhergesagt. Die Items 8 und 15 standen in keinem Zusammenhang zur psychiatrischen Depressionsdiagnose. Die Werte von sieben Items wurden signifikant durch die psychiatrische Diagnose und das Geschlecht bzw. die Interaktion von Diagnose und Geschlecht vorhergesagt. Auch wenn auf der Ebene des Gesamtwerts keine statistisch signifikanten Unterschiede in der Validität gefunden wurden, kann es von praktischer Bedeutung sein, dass einzelne CES-D Items nicht die gleiche diagnostische Qualität für Frauen und Männer aufwiesen.

Schlüsselwörter Geschlechtsunterschiede · Depressivität · Differentielle Validität · Erhebungsinstrumente

Gender differences in the prevalence of depressive symptoms: A consequence of differentially valid measurement instruments?

Abstract In the sample of the Berlin Aging Study (516 men and women aged 70–103 years), we examined gender differences in the predictive validity of the German version of the Center for Epidemiological Studies-Depression scale (CES-D). Separate item analyses demonstrated that most items were more difficult and had lower corrected item-total correlations for men than for women. To determine sensitivity and specificity, we calculated the agreement between item scores and psychiatric diagnoses of depression according to DSM-III-R for both sexes. For most items, sensitivity was low and differed between men and women. Although these differences cumulated in the sensitivity of the total score (cut-off: 16), they could not be confirmed in ROC analyses. Regressing individual item scores on DSM-III-R diagnoses, the participants' sex and their interaction revealed that two items (8 and 15) did not correlate with the diagnoses. Seven items were predicted by psychiatric diagnoses of depression, but also by sex or by the diagnosis/sex interaction. Although there were no statistically significant validity differences for the total score, it may be of practical significance that not all of the CES-D items had the same measurement properties for men and women.

Keywords Gender differences · Depressivity · Differential validity · Measurement instruments

Dr. C. Semmler (✉)
IFT Institut für Therapieforchung,
Parzivalstraße 25, 80804 München
E-Mail: semmler@ift.de
Tel.: 089-36080435
Fax: 089-36080449

P. L. Klumb
Institut für Gesundheitswissenschaften,
Technische Universität Berlin,
BH 8, Ernst-Reuter-Platz 1, 10587 Berlin

Hinsichtlich der Prävalenz unipolarer depressiver Störungen werden in der Literatur immer wieder Geschlechtsunterschiede berichtet (z. B. Aneshensel et al. 1981; Radloff 1975; Rosenfield 1980) und auf biologische, psychologische und sozioökonomische Einflüsse zurückgeführt.

Um den Geschlechtsunterschied zu etablieren, reichen allerdings bloße Replikationen des Befundes nicht aus. Dazu ist es auch notwendig, die Angemessenheit der verwendeten Konzepte und Methoden geschlechterbezogen zu überprüfen (Jahn 2002). In allen Phasen des Forschungsprozesses kann die nicht angemessene Berücksichtigung der Kategorie Geschlecht zu Verzerrungen der Ergebnisse führen und damit zu artifiziellen Geschlechtsunterschieden (Eichler et al. 2000; Ruiz u. Verbrugge 1997). Mit dieser Studie soll der Einfluss eines methodischen Faktors untersucht werden: die Angemessenheit eines Depressivitäts-Screenings für beide Geschlechter. In der Allgemeinbevölkerung kann das Auftreten depressiver Symptomatik mit der vom Center for Epidemiological Studies des amerikanischen National Institute of Mental Health entwickelten Depressionsskala CES-D erfasst werden (CES-D, Radloff 1977; deutsch: Allgemeine Depressionsskala, ADS, Hautzinger 1988; Hautzinger u. Bailer 1992). Die Skala hat damit einen breiteren Einsatzbereich als diagnostische Instrumente, die für den Einsatz in klinischen Settings konstruiert wurden. Häufig steht wegen der Größe der Stichproben bei Screening-Instrumenten die ökonomische Anwendbarkeit im Vordergrund. Zusätzlich müssen solche Instrumente aber auch die Kriterien der klassischen Testtheorie (Objektivität, Reliabilität und Validität) erfüllen. Da die Validität eines Instruments innerhalb von Subgruppen anders ausfallen kann als in der Gesamtpopulation, ist das Ziel dieser Arbeit, die Validität der deutschen Version der CES-D geschlechterbezogen zu überprüfen. Während die altersbezogene Untersuchung der diagnostischen Eigenschaften der CES-D sowohl im angelsächsischen als auch im deutschen Sprach- und Kulturraum nur geringe Altersunterschiede ergab (z. B. Hertzog et al. 1990; Riediger et al. 1998; Weyerer et al. 1992), fand die geschlechterbezogene Überprüfung der Validität der CES-D bislang fast ausschließlich im angelsächsischen Sprachraum statt. Dabei wurden Geschlechtsunterschiede hinsichtlich der faktoriellen Struktur der englischen Version gefunden, die wir im Folgenden kurz darstellen werden.

Clark et al. (1981) untersuchten die Faktorenstruktur der CES-D an 588 Frauen und 412 Männern des „Los Angeles Metropolitan Area Sample“ (18–92 Jahre). Eine Faktorenanalyse über die Gesamtstichprobe erbrachte die von Radloff (1977) ermittelten Faktoren. Es ergaben sich jedoch verschiedene Faktorenmuster für Frauen und Männer. Diese mangelnde faktorielle Invarianz für die Geschlechter ist ein Hinweis darauf, dass die Skala von Frauen und Männern anders verwendet wird und damit nicht die gleichen Eigenschaften für beide Geschlechter aufweist. In einer anderen Studie wurden 4 Substichproben mit insgesamt über 2000 Jugendlichen der Schulklassen 9–12 aus dem US-Bundesstaat Oregon untersucht (Roberts et al. 1990), und es zeigten sich ebenfalls Geschlechtsunterschiede in den Faktorladungen. Eine Analyse auf Itemebene führten Stommel et al. (1993) an 2 Stichproben durch: 1) 708 Krebspatientinnen und -patienten (361 Frauen und 347 Männer; 22–89 Jahre) und 2) 504 Krankenpflegerinnen und -pfleger (252 Frauen und 252 Männer; 18–88 Jahre).

Die Autorinnen stellten ein Grundmodell auf, bei dem die Faktorladungen, die Fehlervarianzen und die Interfaktor-Kovarianzen der Items für beide Geschlechter gleich sein sollten. Da dieses Grundmodell nicht mit den Stichprobendaten der Krebspatientinnen und -patienten übereinstimmte, wurden die Gleichheitsbeschränkungen des Grundmodells für 3 Items gelockert. Nach dem Ausschluss dieser 3 auffälligen Items, sowie 2 weiterer wegen extremer Schiefe, ergab sich eine Skala mit 15 Items. Obwohl Frauen immer noch einen signifikant höheren Gesamtwert als Männer erhielten, war der Geschlechtsunterschied bei der verkürzten Skala, im Vergleich zur Originalskala, signifikant geringer. Aufgrund der in diesen Studien berichteten Geschlechtsunterschiede in der faktoriellen Struktur kann man nicht davon ausgehen, dass die englische CES-D für beide Geschlechter in gleicher Weise anwendbar ist. Von größerer praktischer Bedeutung als Unterschiede in der Konstruktvalidität (faktoriellen Validität) sind Unterschiede in der kriterienbezogenen Validität, insbesondere hinsichtlich der Korrelation mit Außenkriterien. Bei unserer geschlechterbezogenen Überprüfung der deutschen Version galt unsere Aufmerksamkeit deshalb besonders der prädiktiven Validität. Dazu haben wir Moderatoreffekte in der Prädiktion von Itemwerten sowie Geschlechtsunterschiede in Sensitivität und Spezifität untersucht.

Methoden

Stichprobe

Diese Untersuchung basiert auf Daten der Ersterhebung der Berliner Altersstudie (BASE). In dieser Studie wurden 516 Männer und Frauen im Alter von 70 bis 103 Jahren (mittleres Alter: 84,9 Jahre) untersucht, die zwischen 1990 und 1992 aus dem amtlichen Melderegister von Berlin (West) gezogen und nach Alter und Geschlecht so geschichtet wurden, dass alte und sehr alte Männer und Frauen zu gleichen Anteilen in der Stichprobe vertreten waren (Baltes et al. 1996).

Untersuchungsvariablen

Es wurden 3 Variablenkomplexe der Berliner Altersstudie benutzt:

- 1) das biologische Geschlecht der Studienteilnehmer,
- 2) die Items der CES-D und
- 3) psychiatrische Diagnosen über das Vorliegen einer depressiven Erkrankung.

Die per Interview erhobenen 20 Items der CES-D erfassen die Häufigkeit depressiver Symptome in der vorangegangenen Woche. Die vierstufige Antwortskala weist folgende Antwortmöglichkeiten auf:

- kaum oder überhaupt nicht (weniger als 1 Tag)
- manchmal (1–2 Tage),

- öfters (3–4 Tage),
- meistens, die ganze Zeit (5–7 Tage).

Diesen Stufen sind Punktwerte von 0 bis 3 zugeordnet. Der Gesamtscore wird durch Addition der einzelnen Itemwerte gebildet. Er kann Werte zwischen 0 und 60 annehmen. Im Allgemeinen wird ein Cut-Off von 16 angesetzt, d. h., von diesem Wert an gelten die Symptome als klinisch relevante Störung (Clark et al. 1981; Orme et al. 1986; Weyerer et al. 1992).

Ob eine depressive Erkrankung vorlag, wurde in mehreren Schritten ermittelt. Zunächst entschieden 3 Forschungspsychiater anhand des halbstrukturierten klinischen Interviews „Geriatric Mental State Version A/History and Aetiology Schedule“ (GMS-A/HAS; Copeland et al. 1986; McWilliam et al. 1988), ob eine Studienteilnehmerin bzw. ein Studienteilnehmer ein depressives Syndrom aufwies oder nicht. Lag ein solches Syndrom nicht vor, wurde diese Person als nichtdepressiv diagnostiziert. Alle Personen mit einem depressiven Syndrom wurden anschließend danach differenziert, ob eine nach den Kriterien des DSM-III-R (American Psychiatric Association 1987; Wittchen et al. 1989) spezifizierte depressive Erkrankung vorlag. Bei Personen mit subdiagnostischer depressiver Symptomatik schätzte der untersuchende Psychiater den Krankheitswert der Symptome ein. Dieser wurde angenommen, wenn die depressive Erkrankung Wohlbefinden und Alltagsfunktionen beeinträchtigte und keine Reaktion auf ein kürzlich eingetretenes Lebensereignis darstellte. In diesem Fall erhielten Personen die Diagnose „Nicht näher bezeichnete depressive Störung“. Andernfalls wurde von „Symptomen ohne Krankheitswert“ gesprochen.

Die diagnostische Information haben wir auf 2 Kategorien reduziert:

- In der Kategorie „nichtdepressiv“ sind Personen ohne depressives Syndrom und Personen mit depressiven Symptomen ohne Krankheitswert zusammengefasst.
- Die Kategorie „depressiv“ schließt Personen mit nach DSM-III-R spezifizierten depressiven Störungen und Personen mit „nicht näher bezeichneten depressiven Störungen“ ein.

Diese Unterteilung ist sinnvoll, da aus der CES-D nur abgeleitet werden kann, ob eine depressive Störung vorliegt; es ist aber keine Spezifizierung nach verschiedenen Diagnosen möglich (Hautzinger 1988).

Analysestrategien

Geschlechtsunterschiede hinsichtlich der psychiatrischen Depressionsdiagnosen wurden mit dem asymptotischen Vierfelder- χ^2 -Test geprüft. Der Vergleich der durchschnittlichen Itemantworten von Männern und Frauen erfolgte über den U-Test von Mann-Whitney, da die Voraussetzungen des t-Tests [Normalverteilung (Kolmogoroff-Smirnov-Anpassungstest mit Lilliefors-Schranken) und Varianzhomogenität (Levene-Test)] verletzt waren.

Die interne Konsistenz der CES-D wurde für die Gesamtstichprobe sowie für Frauen und für Männer über Cronbachs α bestimmt. Anschließend wurden die Items Itemanalysen unterzogen.

Die Grundlage für die geschlechtsbezogene Ermittlung der Itemgüte bot der Vergleich von Itemwerten der CES-D mit den diagnostischen psychiatrischen Informationen. Je nach Übereinstimmung beider Informationen ergeben sich 4 Kombinationen aus Itemwert und Diagnose (richtig-positiv, falsch-positiv, richtig-negativ und falsch-negativ), anhand derer Sensitivität und Spezifität als Kennwerte der Itemgüte ermittelt wurden. Die Sensitivität repräsentiert den Anteil der Richtig-Positiven an allen Personen mit einer depressiven Erkrankung. Es handelt sich dabei um die Wahrscheinlichkeit, dass der Itemwert bei einer erkrankten Person positiv ist. Die Spezifität repräsentiert den Anteil der Richtig-Negativen an allen Gesunden. Da pro Item nur das Vorhandensein eines depressiven Symptoms erfasst wird, können aus einem einzelnen Item noch keine Aussagen über das Vorliegen einer depressiven Störung abgeleitet und demnach keine perfekten Kennwerte erwartet werden. Versagt ein Item jedoch in der Trennung von nichtdepressiven und depressiven Personen, bietet es auch bei der Bildung des Gesamtwerts keine nützlichen Informationen. Deswegen wurde von jedem Item ein Mindestmaß an diagnostischer Qualität erwartet, d. h. eine Trefferquote jenseits der Zufallswahrscheinlichkeit. Zur Ermittlung der geschlechtsspezifischen Sensitivität und Spezifität wurden die Itemantworten dichotomisiert, um anzugeben, welche Itemantworten negativ und welche positiv sind. Wegen der linkssteilen Verteilung der Itemantworten und aus Gründen der Einheitlichkeit wurde für alle Items und für Männer und Frauen folgender Cut-off festgelegt: Nur der Itemwert null ist negativ, während die anderen 3 Itemwerte (1, 2 und 3) positiv sind (Cut-off: 0 vs. 1, 2 und 3). Die Kennwerte wurden nur berechnet, wenn zwischen den dichotomisierten Itemantworten und der psychiatrischen Depressionsdiagnose ein signifikanter Zusammenhang, gemessen über den Phi-Koeffizienten (ϕ^2), vorlag. Da der Gesamtscore eine additive Verknüpfung der Informationen aller Items ist, schlagen sich differenzielle Prädiktionen der Items im Gesamtscore nieder. Für den Gesamtwert wurden Sensitivität und Spezifität auf der Grundlage des Standard-Cut-offs von 16 Punktwerten ermittelt. Zusätzlich haben wir Receiver-operator-characteristic(ROC)-Kurven erstellt und die Fläche unter der Kurve („area under the curve“, AUC; Murphy et al. 1987) berechnet.

Um die Ermittlung von Items mit einer geschlechtsspezifischen Prädiktion zufallskritisch abzusichern, wurde für jedes Item ein Regressionsmodell aufgestellt, nach dem die Items durch die Depressionsdiagnose, das Geschlecht und die Interaktion zwischen Diagnose und Geschlecht vorhergesagt wurden. Damit sollte überprüft werden, ob die Itembeantwortung signifikant vom Geschlecht der Probanden beeinflusst wurde. Idealerweise wird ein Item nur durch die Diagnose vorhergesagt. Leistet zusätzlich auch das Geschlecht einen signifikanten Vorhersagebeitrag, muss bereits von einer differentiellen Prädiktion ausgegangen werden. Eine zusätzlich fehlende

Korrelation zur Diagnose verstärkt diese Verzerrung. Dasselbe Regressionsmodell wurde auf den Gesamtscore angewendet.

Ergebnisse

Depressionsprävalenz und Itemkennwerte

Psychiatrische Depressionsdiagnosen lagen bei 31% der Frauen und 20% der Männer vor, wurden also häufiger an Frauen als an Männer vergeben ($\chi^2=8,52$, $z=2,92$; $p<0,01$). Außerdem unterschieden sich Frauen und Männer bei der Beantwortung von 12 Items sowie beim Gesamtwert signifikant in ihrer zentralen Tendenz (s. Tabelle 1). Cronbachs α über alle 20 Items der CES-D lag für die Gesamtstichprobe bei 0,87 (Frauen: 0,88, Männer: 0,83).

Die Mehrheit der Probandinnen und Probanden erlebte die angesprochenen Symptome „kaum oder überhaupt nicht“ (Antwortmöglichkeit Null), sodass sich für beide Geschlechter linkssteile Antwortverteilungen ergaben (s. Tabelle 1). Davon ausgenommen sind die positiv gepolten Items 8 (Hoffnung), 12 (Fröhliche Stimmung) und 16 (Leben genießen). Am stärksten wurden die Items 15 (unfreundliche Leute) und 19 (Nicht-leiden-können) von den Probanden abgelehnt, von Frauen noch mehr als von Männern. Entsprechend der Schiefe waren die meisten Itemschwierigkeiten hoch ($<0,20$; Fisseni 1990; s. Tabelle 1). Dies traf bei Männern doppelt so oft zu wie bei Frauen (Männer: 14 Items, Frauen: 7 Items). Alle Items waren für Frauen leichter, eine Ausnahme waren die Items 15 und 19, deren Itemschwierigkeiten bei beiden Geschlechtern extrem hoch waren (vgl. Schiefe). Bis auf die Items 11 (Schlafprobleme) und 20 (Schwunglosigkeit) hatten alle Items für Frauen höhere Trennschärfen als für Männer (s. Tabelle 1). Lediglich 5 Items korrelierten bei Männern hoch mit dem Gesamtwert, während es bei Frauen 11 Items waren (Trennschärfe $>0,5$; Bortz u. Döring 1995). Während für Männer 4 Items Trennschärfen $<0,3$ aufwiesen, hatte für Frauen nur 1 Item eine zu geringe Trennschärfe.

Validitätsindikatoren

Bei fast allen Items lag für Frauen eine höhere Sensitivität vor (s. Abb. 1).^{1 2} Die Items 3 (Trübsinn), 10 (Angst), 14 (Einsamkeit), 17 (Weinen) und 18 (Traurigkeit) wiesen für Frauen eine um mindestens 10% höhere Sensitivität auf als für Männer. Demnach wurden Männer durch diese

¹ Für die Items 8 und 15 wurden Sensitivität und Spezifität nicht ermittelt, da zwischen den dichotomisierten Itemantworten (Cut-off: 0 vs. 1, 2 und 3) und der psychiatrischen Depressionsdiagnose bei Item 8 für beide Geschlechter und bei Item 15 für Männer keine signifikanten Zusammenhänge vorlagen.

² Die Verwendung anderer Cut-offs auf Itemebene (0, 1 vs. 2, 3; 0, 1, 2 vs. 3 oder eines optimalen Cut-offs, d. h. auf Basis der Itemdichotomisierung mit dem größten Zusammenhang zur psychiatrischen Depressionsdiagnose) erbrachte ähnliche Ergebnisse.

Items häufiger als Frauen falsch-negativ klassifiziert. Während für Männer 7 Items (2, 3, 9, 10, 13, 17, 19) eine Sensitivität unter 50% aufwiesen, waren es für Frauen nur 3 (Items 9, 17, 19). Für die meisten Items waren die Spezifitätswerte für Männer und Frauen höher als die Sensitivitätswerte (s. Abb. 2) und wiesen geringere Geschlechtsunterschiede als letztere auf. Nur die Items 11 (Schlafprobleme) und 20 (Schwunglosigkeit) wiesen für Männer eine um mindestens 10% höhere Spezifität als für Frauen auf, d. h., es gab hier bei Frauen eine wesentlich höhere Rate falsch-positiver Zuordnungen. Die Kennwerte des Gesamtscores auf Basis des Standard-Cut-offs (s. Abb. 1 und 2) differierten für beide Geschlechter stark. Die Sensitivität lag für Frauen bei 80% und für Männer bei 67%. Wenn Frauen einen positiven Gesamtscore aufweisen (mindestens 16 Punktwerte), liegt mit 80%iger Wahrscheinlichkeit eine depressive Störung vor. Bei positiv klassifizierten Männern liegt hingegen nur mit einer Wahrscheinlichkeit von 67% eine depressive Störung vor. Die Spezifität fiel bei beiden Geschlechtern höher aus und erzielte Werte von 78% (Frauen) und 83% (Männer). Negativ klassifizierte Frauen sind also mit einer Wahrscheinlichkeit von 78%, negativ klassifizierte Männer mit einer Wahrscheinlichkeit von 83% nichtdepressiv. Mit ROC-Kurven ließen sich diese Unterschiede jedoch nicht zufallskritisch absichern. Die 95%igen Konfidenzintervalle der Fläche unter der Kurve (AUC) waren für beide Geschlechter groß und überlappend (Frauen: 0,81–0,91, Männer: 0,80–0,91).

Der multiple Regressionskoeffizient R als Maß für den Zusammenhang zwischen den 3 Prädiktorvariablen psychiatrische Diagnose, Geschlecht sowie deren Interaktion und den Kriterien war mit Ausnahme von Item 8 (Hoffnung) bei allen Items signifikant ($p<0,001$).³ Die Varianzaufklärung lag zwischen 3,6% (Item 15) und 20,6% (Item 6). Item 8 konnte durch keine der Prädiktorvariablen vorhersagt werden. Bei 11 Items (1, 2, 4, 5, 6, 7, 9, 13, 14, 16, 19) hatte nur die psychiatrische Diagnose einen signifikanten Einfluss auf das jeweilige Item. Die Items 11 (Schlafprobleme), 12 (fröhliche Stimmung) und 20 (Schwunglosigkeit) wurden sowohl durch die Diagnose als auch durch das Geschlecht vorhergesagt. Bei 4 Items (3-Trübsinn, 10-Angst, 17-Weinen, 18-Traurigkeit) trugen sowohl die Diagnose als auch die Interaktion zwischen Diagnose und Geschlecht signifikant zur Aufklärung der jeweiligen Itemvarianz bei. Signifikante Prädiktoren zur Vorhersage des Items 15 (unfreundliche Leute) waren das Geschlecht und die Interaktion zwischen Diagnose und Geschlecht.

Zwischen dem Gesamtwert und den 3 Prädiktoren Diagnose, Geschlecht und Interaktion lag ein multipler Zusammenhang von 0,61 vor, was einer aufgeklärten Kriteriumsvarianz von 37% entspricht. Sowohl Depressionsdiagnose als auch biologisches Geschlecht waren signifikante Prädiktoren.

³ Die genauen Ergebnisse der Regressionsanalysen für alle Items und den Gesamtwert können bei den Autorinnen angefordert werden.

Tabelle 1 CES-D: Mittelwerte, Mediane, Varianzen, Schiefen, Schwierigkeiten und Trennschärfe sowie U-Test von Items und Gesamtwert

Item	Mittelwert		Median		U-Test		Varianz		Schiefe		Schwierigkeit		Trennschärfe	
	M	F	M	F	p	M	F	M	F	M	F	M	F	
1 Beunruhigung	0,65	0,68	0	0	0,87	0,75	0,94	1,22	1,28	0,22	0,23	0,43	0,50	
2 Appetitmangel	0,42	0,63	0	0	0,02	0,75	1,09	2,08	1,45	0,14	0,21	0,37	0,53	
3 Trübsinn	0,35	0,51	0	0	0,05	0,56	0,81	2,27	1,76	0,12	0,17	0,56	0,67	
4 So gut wie andere	0,84	0,96	0	0	0,28	1,26	1,48	0,83	0,71	0,28	0,32	0,39	0,39	
5 Konzentrationsprobleme	0,46	0,59	0	0	0,07	0,57	0,76	1,72	1,50	0,15	0,20	0,47	0,53	
6 Niedergeschlagenheit	0,49	0,64	0	0	0,08	0,66	0,87	1,75	1,36	0,16	0,21	0,68	0,78	
7 Anstrengung	0,63	0,87	0	0	0,02	0,88	1,21	1,38	0,96	0,21	0,29	0,48	0,53	
8 Hoffnung	2,02	2,17	2	3	0,04	1,06	1,13	-0,67	-0,99	0,67	0,72	0,01	0,01	
9 Leben als Fehlschlag	0,25	0,31	0	0	0,37	0,41	0,53	2,85	2,50	0,08	0,10	0,44	0,50	
10 Angst	0,24	0,45	0	0	0,00	0,38	0,72	2,95	1,87	0,08	0,15	0,50	0,72	
11 Schlafprobleme	0,56	0,92	0	1	0,00	0,72	1,21	1,41	0,85	0,19	0,31	0,48	0,37	
12 Fröhliche Stimmung	1,44	1,71	2	2	0,00	1,07	1,1	-0,07	-0,46	0,48	0,57	0,48	0,51	
13 Weniger Reden	0,39	0,55	0	0	0,05	0,47	0,76	1,78	1,62	0,13	0,18	0,29	0,41	
14 Einsamkeit	0,52	0,77	0	0	0,00	0,81	1,05	1,65	1,11	0,17	0,26	0,56	0,63	
15 Unfreundliche Leute	0,10	0,09	0	0	0,19	0,11	0,17	3,64	5,85	0,03	0,03	0,12	0,35	
16 Leben genießen	1,78	2,01	2	2	0,01	1,33	1,26	-0,45	-0,77	0,59	0,67	0,39	0,41	
17 Weinen	0,14	0,33	0	0	0,00	0,17	0,43	3,53	2,12	0,05	0,11	0,33	0,56	
18 Traurigkeit	0,4	0,71	0	0	0,00	0,45	0,86	1,81	1,29	0,13	0,24	0,54	0,75	
19 Nicht-leiden-Können	0,14	0,12	0	0	0,45	0,26	0,23	4,07	4,55	0,05	0,04	0,27	0,33	
20 Schwunglosigkeit	0,52	0,83	0	0,5	0,00	0,7	1,04	1,70	1,02	0,17	0,28	0,61	0,49	
Gesamtwerte	12,3	15,8	10	13	0,00	63,43	103,8	1,22	1,33	-	-	-	-	

M=Männer, F=Frauen. Items mit signifikanten Geschlechtsunterschieden in der zentralen Tendenz (U-Test) wurden fettgedruckt.

Abb. 1 Geschlechtsspezifische Sensitivität von Items und Gesamtwert der CES-D

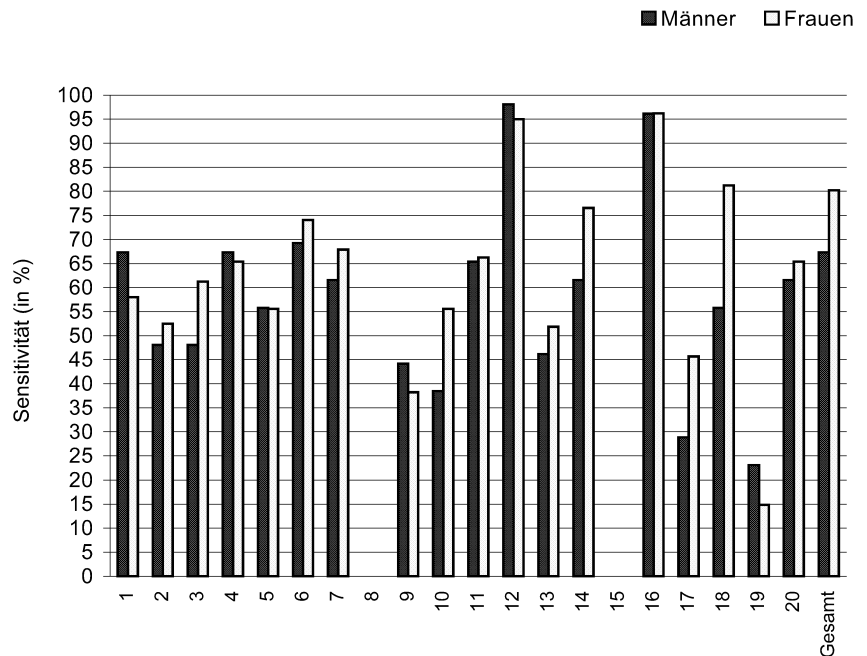
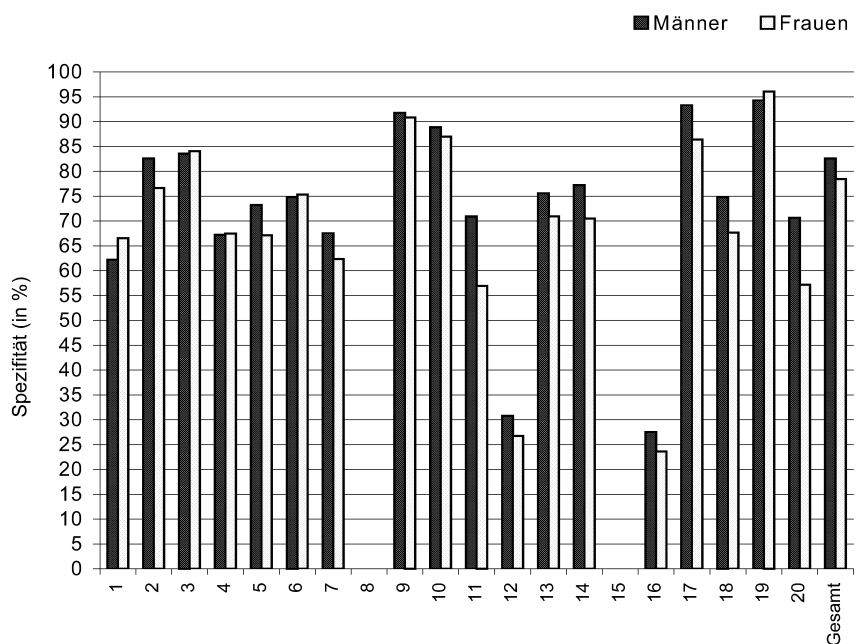


Abb. 2 Geschlechtsspezifische Spezifität von Items und Gesamtwert der CES-D



Diskussion

In dieser Studie haben wir die prädiktive Validität der deutschen Version der CES-D geschlechterbezogen überprüft. Übereinstimmend mit früheren Untersuchungen (z. B. Clark et al. 1981; Hautzinger 1988; Riediger et al. 1998; Roberts et al. 1990; Weyerer et al. 1992) wies die Skala für beide Geschlechter hohe Konsistenzwerte auf. Zentrale Tendenz, Schiefe, Schwierigkeit und Trennschärfe der Items sowie Ausprägung des Gesamtwerts unterschieden sich hingegen für Frauen und Männer erheblich. Im Gegensatz dazu konnte Hautzinger (1988) in

den von ihm zur Validierung eingesetzten Stichproben jüngeren und mittleren Alters (Schülerinnen und Schüler sowie Studentinnen und Studenten, $n=192$; Studentinnen und Studenten, $n=210$; depressive Patientinnen und Patienten, $n=53$) keine signifikanten Geschlechtsunterschiede in Mittelwerten, Standardabweichungen und Spannweiten des Gesamtwerts feststellen. Mögliche Erklärungen dafür sind zum einen die im Vergleich zur örtlich repräsentativen Stichprobe der Berliner Altersstudie eingeschränkte Heterogenität der Stichproben, zum anderen Unterschiede zwischen den Kohorten. Sowohl soziale Normen als auch Geschlechtsrollenselbstkonzepte von Frauen und

Männern haben sich im Verlauf des vergangenen Jahrhunderts stark verändert. Um diese Kohorteneffekte zu untersuchen, müssten zusätzlich zu den Informationen über das biologische Geschlecht solche über dessen soziokulturelle Dimension herangezogen werden.

Die diagnostische Qualität der hier geprüften Items (ermittelt über Sensitivität und Spezifität) lag in einigen Fällen unter dem geforderten Mindestmaß und unterschied sich bei Frauen und Männern. Für Frauen lag insgesamt eine höhere Sensitivität vor, und 5 Items wiesen bedeutsame Geschlechtsunterschiede auf. Im Vergleich zu Frauen wird bei Männern der wahre Anteil von Depressiven eher unter- und der von Nichtdepressiven eher überschätzt. Tendenziell lag für Männer eine höhere Spezifität vor, und 2 Items zeigten Geschlechtsunterschiede in beträchtlichem Ausmaß. Nichtdepressive Männer werden im Vergleich zu nichtdepressiven Frauen eher als solche erkannt. Folglich wird der wahre Anteil nichtdepressiver Frauen im Vergleich zu Männern eher unterschätzt, während derjenige depressiver Frauen eher überschätzt wird. Obwohl sich diese auf Itemebene ermittelten Unterschiede in der Sensitivität und Spezifität des Gesamtwerts (Cut-off: 16) niederschlugen, ließen sie sich jedoch mit ROC-Kurven nicht inferenzstatistisch absichern. Die Konfidenzintervalle der Fläche unter der Kurve (AUC) waren für beide Geschlechter groß und überlappend. Da die CES-D ein vielfach erprobtes Instrument ist, waren grobe Auffälligkeiten auf der Gesamtscore-Ebene nicht zu erwarten.

In anderen Studien zur deutschen Version der CES-D werden lediglich Kennwerte für die Gesamtstichprobe, jedoch keine geschlechtsspezifischen Werte angegeben. Die von Hautzinger (1988) und Hautzinger u. Bailer (1992) berichteten Sensitivitäten lagen weit höher als die der Männer unserer Studie und geringfügig höher als die der Frauen. Die hohen Sensitivitäten könnten durch die Zusammensetzung der Stichproben aus depressiven Patientinnen und Patienten bedingt sein. Die Basisrate hat einen Einfluss auf die Prädiktionskraft eines Instruments: Je höher die Prävalenz in einer Bevölkerungsgruppe, desto geringer ist der Anteil von falsch-positiven unter den positiv Getesteten (vgl. Gigerenzer et al. 1998). Außerdem wurde ein anderer Cut-off-Wert verwendet. Während wir einen Cut-off von 16 Punktwerten benutzten, lag der Cut-off von Hautzinger (1988) bei 18 und von Hautzinger u. Bailer (1992) bei 23.

Regressionsanalysen ergaben, dass die psychiatrischen Depressionsdiagnosen, mit Ausnahme der Items 8 und 15, signifikant zur Vorhersage der Itemwerte beitrugen. Diese beiden Items stehen also in keinerlei Zusammenhang zum Konstrukt der depressiven Beeinträchtigung und sind damit invalide. 7 Items (11-Schlafprobleme, 12-Fröhliche Stimmung, 20-Schwunglosigkeit, 3-Trübsinn, 10-Angst, 17-Weinen, 18-Traurigkeit) wurden über die Diagnose hinaus signifikant durch das biologische Geschlecht oder die Interaktion zwischen Diagnose und Geschlecht vorhergesagt. Die prädiktive Validität dieser Items ist damit nicht für beide Geschlechter identisch.

Methodische und praktische Implikationen

Die geschlechtsspezifische diagnostische Qualität einzelner Items ist von theoretischer und praktischer Bedeutung. Die Frage nach der praktischen Bedeutsamkeit der gefundenen Unterschiede konnte mit dieser Studie jedoch nicht ausreichend beantwortet werden. Für ihre Beantwortung sind weitere Studien erforderlich. Es erscheint uns vorläufig nicht empfehlenswert, die CES-D zur Bestimmung von Geschlechtsunterschieden in depressiver Symptomatik einzusetzen. Es gibt jedoch eine Reihe von Modifikationsmöglichkeiten. Als wichtigste erachten wir eine komplexere und geschlechtsspezifische Bildung des Gesamtwerts sowie den Gebrauch von geschlechtsspezifischen Cut-off-Werten (vgl. Fuhrer u. Rouillon 1989). Vor einer Umformulierung, aber auch einem Ersetzen bzw. Entfernen einzelner Items (wie z. B. Items 8 und 15), ist eine Absicherung durch weitere Befunde notwendig.

Einschränkungen

Die Validierung der Items der deutschen Version der CES-D erfolgte an psychiatrischen Depressionsdiagnosen. Psychiatrische Diagnosen werden oftmals als „Goldstandard“ angesehen, weil in der Psychiatrie keine natürlichen externen Validierungskriterien existieren (Häfner 1978). Dabei ist zu beachten, dass die psychiatrische Depressionsdiagnostik selbst vielerlei Einflüssen ausgesetzt ist. Die Auswahl von Diagnosekriterien ist ein solcher Einflussfaktor: Bei der Verwendung von Feighner-Kriterien (Feighner et al. 1972) hätten sich möglicherweise andere Zusammenhangsmuster mit den Itemwerten ergeben (vgl. Weyerer et al. 1992), da diese etwas besser mit der durch die CES-D erfassten depressiven Symptomatik übereinstimmen als die DSM-III-R-Kriterien.

Eine weitere Einschränkung ist in der Zusammensetzung der Stichprobe zu sehen. Die Ergebnisse dieser Arbeit basieren auf Daten der Berliner Altersstudie, in der aufgrund der Schichtung nach Alter und Geschlecht sehr alte Männer überrepräsentiert waren. Möglicherweise zeichnen sie sich durch eine besonders positive Selektion in verschiedenen Funktionsbereichen wie z. B. der Gesundheit aus. Nach statistischer Kontrolle der physischen Morbidität blieben die berichteten Ergebnisse allerdings unverändert. Dennoch sind die anhand dieser Stichprobe gewonnenen Erkenntnisse nicht ohne weiteres generalisierbar. Außerdem schränken Kohorten-, Alters- und Zeiteffekte die Generalisierbarkeit ein. Daher wäre eine Replikation dieser Studie mit anderen Altersgruppen wünschenswert.

Literatur

- American Psychiatric Association (1987) Diagnostic and Statistical Manual of Mental Disorders DSM-III-R (3rd ed. revised). American Psychiatric Association, Washington, DC

- Aneshensel CS, Frerichs RR, Clark VA (1981) Family roles and sex differences in depression. *J Health Soc Behav* 22:379–393
- Baltes PB, Mayer KU, Helmchen H, Steinhagen-Thiessen E (1996) Die Berliner Altersstudie (BASE): Überblick und Einführung. In: Mayer KU, Baltes PB (Hrsg) Die Berliner Altersstudie. Akademie Verlag, Berlin, S 21–54
- Bortz J, Döhning N (1995) Forschungsmethoden und Evaluation für Sozialwissenschaftler, 2., vollständig überarbeitete und aktualisierte Aufl. Springer, Berlin
- Clark VA, Aneshensel CS, Frerichs RR, Morgan TM (1981) Analysis of sex and age in response to items on the CES-D Scale. *Psychiatry Res* 5:171–181
- Copeland JRM, Dewey ME, Griffiths-Jones HM (1986) A computerized psychiatric diagnostic system and case nomenclature for elderly subjects: GMS and AGE-CAT. *Psychol Med* 16:89–99
- Eichler M, Fuchs J, Maschewsky-Schneider U (2000) Richtlinien zur Vermeidung von Gender-Bias in der Gesundheitsforschung. *Z Gesundheitswiss* 8:293–310
- Feighner JP, Robins E, Guze SB, Woodruff RA, Winoku G, Munoz R (1972) Diagnostic criteria for use in psychiatric research. *Arch Gen Psychiatry* 26:57–63
- Fisseni H (1990) Lehrbuch der psychologischen Diagnostik. Hogrefe, Göttingen
- Fuhrer R, Rouillon F (1989) La version française de l'échelle CES-D (Center for Epidemiologic Studies—Depression Scale). Description et traduction de l'échelle d'autoévaluation. *Psychiatr Psychobiol* 4:163–166
- Gigerenzer G, Hoffrage U, Ebert A (1998) AIDS Counselling for low-risk patients. *AIDS Care* 10:197–211
- Häfner H (1978) Psychiatrische Epidemiologie. Springer, Berlin
- Hautzinger M (1988) Die CES-D Skala. Ein Depressionsmeßinstrument für Untersuchungen in der Allgemeinbevölkerung. *Diagnostica* 34:167–173
- Hautzinger M, Bailer M (1992) Allgemeine Depressions Skala. Beltz Test GmbH, Weinheim
- Hertzog C, Van Alstine J, Usala PD, Hultsch DF (1990) Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. *Psychol Assess* 2:64–72
- Jahn I (2002) Methodische Probleme einer geschlechtergerechten Gesundheitsforschung. In: Hurrelmann K, Kolip P (Hrsg) Geschlecht, Gesundheit und Krankheit. Huber, Bern, S 142–155
- Kessler RC, Brown RC, Broman CL (1981) Sex differences in psychiatric help seeking: Evidence from four large surveys. *J Health Soc Behav* 22:49–64
- McWilliam C, Copeland JRM, Dewey ME, Wood M (1988) The Geriatric Mental State Examination: A case-finding instrument in the community. *Br J Psychiatry* 152:205–208
- Murphy JM, Berwick DM, Weinstein MC, Borus JF, Budman SH, Klerman GL (1987) Performance of screening and diagnostic tests. *Arch Gen Psychiatry* 44:550–555
- Orme JG, Reis J, Herz EJ (1986) Factorial and discriminant validity of the Center for Epidemiological Studies-Depression (CES-D) Scale. *J Clin Psychol* 42:28–33
- Radloff LS (1975) Sex differences in depression: The effects of occupation and marital status. *Sex Roles* 1:249–265
- Radloff LS (1977) The CES-D scale: A self-report depression scale for research in the general population. *Appl Psych Meas* 1:385–401
- Riediger M, Linden M, Wilms HU (1998) Die deutsche Version der CES-D als Instrument der gerontologischen Forschung. *Z Klin Psychol Psych* 46:344–364
- Roberts RE, Andrews JA, Lewinsohn PM, Hops H (1990) Assessment of depression in adolescents using the Center for Epidemiologic Studies Depression Scale. *Psychol Assess* 2:122–128
- Rosenfield (1980) Sex differences in depression: Do women always have higher rates? *J Health Soc Behav* 21:33–42
- Ruiz TM, Verbrugge LM (1997) A two way of gender bias in medicine. *J Epidemiol Community Health* 51:106–109
- Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R (1993) Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Res* 49:239–250
- Verbrugge LM (1989) The twain meet: Empirical explanations of sex differences in health and mortality. *J Health Soc Behav* 30:282–304
- Weyerer S, Geiger-Kabisch C, Denzinger R, Pfeifer-Kurda M (1992) Die deutsche Version der CES-D Skala. Ein geeignetes Meßinstrument zur Erfassung von Depressionen bei älteren Menschen? *Diagnostica* 38:354–365
- Wittchen HU, Saß H, Zaudig M, Koehler K (1991) Diagnostisches und statistisches Manual psychischer Störungen DSM-III-R (deutsche Bearbeitung und Einführung), 3., korrigierte Aufl. Beltz, Basel