Kindheit und Entwicklung 11 (2), 90–97 © Hogrefe-Verlag, Göttingen 2002

ADHS im Lehrerurteil: Ein Vergleich von Klinik- und Normstichprobe anhand der Conners-Skalen

Michael Huss[1], Christina Stadler[2], Harriet Salbach[1], Patrick Mayer[3], Marlies Ahle[1] und Ulrike Lehmkuhl[1]

[1] Klinik für Psychiatrie, Psychosomatik und Psychotherapie des Kindes- und Jugendalters, Charité, Virchow-Klinikum der Humboldt-Universität zu Berlin
[2] Klinik für Psychiatrie und Psychotherapie des Kindes- und Jugendalters der Johann Wolfgang Goethe-Universität Frankfurt am Main
[3] Psychologisches Institut der Freien Universität Berlin

Zusammenfassung. Die Conners-Rating-Skalen (CRS) stellen das in der ADHS-Diagnostik am häufigsten eingesetzte Fragebogenverfahren dar. In deutlichem Kontrast zu der Verbreitung der CRS steht der empirische Kenntnisstand über deren Gütekriterien. Angesichts fehlender Normdaten wurden im deutschen Sprachraum bislang nur US-amerikanische Normen eingesetzt. Ob damit möglicherweise systematische Klassifikationsfehler begangen wurden, ist nicht bekannt. Die Studie analysiert die Verhaltenseinschätzung der Lehrer für 994 kinderpsychiatrische Patienten sowie für 424 gesunde Probanden und setzt diese mit den US-amerikanischen Werten in Beziehung. Dabei zeigen sich in Abhängigkeit von der jeweiligen Skala zum Teil deutliche Abweichungen zwischen deutschen und US-amerikanischen Normen. Die stärksten Abweichungen finden sich in der Betragensskala. Die Skala für Hyperaktivität/Impulsivität weicht mäßig ab. Die Aufmerksamkeitsskala ist hingegen über die Altersgruppen hinweg nahezu identisch. Bei einem Vergleich von Klinik- und Normstichprobe ergeben sich in der ROC-Analyse die schlechtesten Trennwerte für die Betragensskala. Mit den Skalen für Hyperaktivität/Impulsivität sowie für Aufmerksamkeit lassen sich die Stichproben deutlich besser trennen.
Schlüsselwörter: Aufmerksamkeitsdefizit-/Hyperaktivitätsstörungen, Conners-Lehrer-Fragebogen, interkultureller Vergleich, Normierung

Teacher-rated ADHD – A comparison of a clinical and a field sample with the Conners Rating Scale

Abstract. The Conners Rating Scale (CRS) is the most often used questionnaire in ADHD assessment. However, there is a discrepancy between the widespread use of the CRS and the amount of empirically based knowledge about the test characteristics. Lacking national norm data, German-speaking countries usually use the US-American norms. Whether this may result in incorrect classifications is not yet known. The study analyzes behavioral ratings of teachers for 994 patients treated in a child and adolescent psychiatric clinic and for 424 healthy controls. German and US raw scores are compared. Discrepancies between the norm data depended on the scale. While conduct problems differed strongly, the impulsivity/hyperactivity scale showed only moderate discrepancies and the attention problem scale was almost identical. Using ROC analyses to discriminate the German normative sample from the clinical sample, the worst discrimination results were found for the conduct scale. The impulsivity/hyperactivity scale and especially the attention problem scale showed much better results in the ROC analyses.
Key words: Attention Deficit/Hyperactivity Disorder, ADHD, Conners Teacher Rating Scale, cross-cultural comparison, norm

Im Rahmen der Diagnostik einer Aufmerksamkeits-Defizit-Hyperaktivitäts-Störung (ADHS) bzw. eines Hyperkinetischen Syndroms (HKS) nimmt das Lehrerurteil eine zentrale Stellung ein. Nach den diagnostischen Kriterien gemäß ICD-10 kann die Diagnose eines HKS nur dann vergeben werden, wenn hinreichende Informationen über das Verhalten des Kindes in der Schule oder einer vergleichbaren Institution vorliegen. In der Regel verfügt der diagnostizierende Arzt oder Psychologe nicht über die zeitlichen Ressourcen, eine Verhaltensbeobachtung in der Schule vorzunehmen. Auch muß davon ausgegangen werden, daß nur in wenigen Fällen ein persönliches Gespräch mit dem Lehrer geführt wird. Meist erfolgt die Verhaltenseinschätzung schriftlich mittels standardisierter Fragebogen. Eine gute Normierung und Validierung solcher Verfahren hat damit weitreichende praktische Konsequenzen.

Die am häufigsten eingesetzten Fragebogen – gewissermaßen die ‚Klassiker' der ADHS-Forschung – sind die sog. Conners-Skalen (CRS). Sie wurden in den 60er Jahren entwickelt und erfuhren eine rasche, weltweite Verbreitung. Bis heute führen sie die Liste der in der ADHS-Forschung verwendeten Verfahren an (Deimel et al., 1997).

Die Verbreitung der CRS steht – insbesondere im deutschen Sprachraum – in deutlichem Kontrast zu dem empirischen Kenntnisstand über das Instrument. Unseres

DOI: 10.1026//0942-5403.11.2.90

Wissens liegt nur eine Studie zu Reliabilitäts- und Validitätsaspekten aus dem deutschen Sprachraum vor (Brocke et al., 1986), deren Daten bereits vor 18 Jahren erhoben wurden. Deutsche Normwerte fehlen ganz.

Entscheidet sich der deutschsprachige Praktiker trotz unzureichender Forschungslage für die Conners-Skalen, so stellt sich als nächste Frage, welche Form er wählen soll.

Innerhalb der vergangenen 30 Jahren wurden die CRS sowohl in der Eltern- als auch in der Lehrerversion mehreren Änderungen unterzogen, die sich hinsichtlich Umfang, Inhalt und Testgüte unterscheiden. Die Lehrerversion, auf die wir uns in der vorliegenden Arbeit beschränken, wurde von Conners erstmals 1969 mit 39 Items veröffentlicht. Bis zum Erscheinen der aktuellen Revision (Conners et al., 1998) berichten Deimel et al. (1997) im Rahmen einer Übersichtsarbeit von sechs weiteren Fassungen der Lehrerversion, deren Umfang zwischen zehn und 39 Items variiert. Im wesentlichen lassen sich vier Grundformen unterscheiden, aus deren Variation die übrigen Versionen abgeleitet werden können: 1. die ursprüngliche Langversion mit 39 Items (Conners, 1969), 2. die revidierte und gekürzte Fassung mit 28 Items (Goyette et al., 1978), 3. die Kurzform mit zehn Items (Conners, 1989) und 4. die aktuelle Überarbeitung mit 59 Items als Lang- und mit 28 Items als Kurzversion (Conners et al., 1998). Über die letztgenannte Fassung (59 bzw. 28 Items) liegen bislang nur wenige internationale Daten vor.

Hinsichtlich der Einschätzung, welche Version sich für die ADHS-Diagnostik am besten eignet, ist ein Trend zugunsten der 28-Item-Version von 1978 auszumachen. So sprechen sich Barkley (1987) als auch Blondis et al. (1991) für den Einsatz der revidierten und gekürzten Fassung mit 28 Items aus. Conners (1989, S. 3) empfiehlt diese ebenfalls für den klinischen Gebrauch, da sie aufgrund der Beschränkung auf die wichtigsten Items schneller zu bearbeiten und mit den überarbeiteten Formulierungen einfacher zu verstehen sei. Auch biete die an einer Stichprobe durchgeführte Parallelnormierung der Lehrer- und Elterneinschätzung (Goyette et al., 1978) den Vorteil der direkten Vergleichbarkeit von Daten (Conners, 1989). Trotz der genannten Vorteile bezieht sich die Mehrzahl der Reliabilitäts- und Validitätsstudien auf die Langversion mit 39 Items von 1969.

Im deutschen Sprachraum wurde unseres Wissens bislang nur eine empirische Arbeit über die Gütekriterien der Conners Lehrerversion veröffentlicht. Es handelt sich um die Studie von Brocke und Mitarbeitern (1986), bei der eine übersetzte 39-Item-Version an 576 Berliner Grundschülern im Alter von durchschnittlich 9,9 Jahren evaluiert wurde. Auf der Grundlage umfassender Reliabilitäts- und Validitätsanalysen kommen Brocke et al. (1986) zu dem Schluß, daß es sich – trotz gewisser Schwächen in den Gütekriterien – bei den Lehrer-Conners-Skalen um ein brauchbares Instrument handelt, das bislang in der internationalen ADHS-Forschung durch kein vergleichbares Verfahren zu ersetzen ist.

Fragestellung

Die vorliegende Studie erhebt nicht den Anspruch einer vollständigen deutschen Normierung der CTRS. Sie soll anhand eines ausgewählten Vergleichs von großen Klinik- und Normstichproben dem Praktiker Anhaltspunkte dafür geben, welche ‚Fehler' zu erwarten sind, wenn – wie bisher – die alten US-Normen unkritisch bei der aktuellen Diagnostik eingesetzt werden. Von besonderem Interesse ist dabei, ob deutsche Lehrer in Relation zu den US-Normen in ihrem Urteil abweichen. Sollte dies der Fall sein, so ist von weiterem Interesse, ob sich spezifische Abweichungsmuster über die verschiedenen Skalen der CTRS ergeben. Wie in der dimensionalen Diagnostik üblich, sind dabei Alters- und Geschlechtseffekte zu berücksichtigen. Dem Screening-Charakter der CTRS entsprechend soll ferner geprüft werden, wie gut die CTRS zwischen einer klinischen Inanspruchnahmepopulation und einer nicht-klinischen Normstichprobe trennt.

Methodik

Stichproben

Die Normstichprobe wurde in den Jahren 1998 und 2000 an zwölf Schulen im Würzburger und Berliner Raum sowie in Krumbach, einer Kleinstadt in Bayern, erhoben. Vertreten waren Grundschulen, Hauptschulen, Gesamtschulen und Gymnasien. Insgesamt schätzten die Lehrer das Verhalten von 424 Schülern (231 Mädchen und 193 Jungen) ein. Der Altersmittelwert beträgt 9 Jahre und 3 Monate (Standardabweichung: 1 Jahr und 4 Monate) mit einer Altersspanne von 6 bis 13 Jahren.

Die Klinikstichprobe umfaßt 1345 Patienten, die in dem Zeitraum zwischen 1997 und 2000 in unserer Klinik vorgestellt wurden und von denen eine Verhaltenseinschätzung durch den Lehrer mittels der CTRS erhoben werden konnte. Dem Screening-Charakter der CTRS entsprechend, erfolgte die Lehrerbefragung unabhängig vom jeweiligen Störungsbild. Die Klinikstichprobe ist damit im Sinne einer Inanspruchnahmepopulation zu verstehen. Die Lehrer erhielten die Fragebogen durch die Eltern, die zuvor ihr Einverständnis zu der Befragung gegeben hatten. Die ausgefüllten Fragebogen wurden dann auf dem Postweg zurückgesandt oder von den Eltern zum nächsten Untersuchungstermin mitgebracht. Acht Fragebogen mußten bereits bei der Dateneingabe ausgesondert werden, da sie widersprüchlich ausgefüllt

waren. Aufgrund der Beschränkung auf den Altersbereich von sechs bis 13 Jahren wurden 351 klinische Fälle ausgeschlossen. Damit resultiert eine klinische Stichprobe von 994 Patienten (270 Mädchen, 724 Jungen). Der Altersdurchschnitt beträgt 8 Jahre und 11 Monate und liegt damit etwas unterhalb der Normstichprobe. Die Standardabweichung fällt mit 1 Jahr und 6 Monaten (Altersspanne sechs bis 13 Jahre) ähnlich wie die der Normstichprobe aus.

Statistische Analysen

Zunächst wurde eine deskriptive Datenanalyse auf Skalenebene und bezogen auf den sog. Total-Problem-Score (Summe aller Items) gemäß der üblichen Alters- und Geschlechtsstratifizierungen vorgenommen. Anhand internationaler Vergleichsdaten wird anschließend eine Beurteilung vorgenommen, ob und in welchem Umfang beim Einsatz von US-Normen an deutschen Stichproben Verzerrungen zu erwarten sind. Schließlich soll anhand von ROC-Analysen ermittelt werden, wie gut die CTRS klinische Stichproben von Normstichproben trennt.

Umgang mit fehlenden Werten

Vor dem Normvergleich wurden die Fragebogen zunächst einer Missing-Analyse unterzogen. 417 (98,3%) der Fragebogen aus der Normstichprobe waren vollständig ausgefüllt. Bei sieben Fragebogen fehlte jeweils die Angabe zu einem Item. In der Klinikstichprobe war der Anteil von Fragebogen mit nicht angekreuzten Items höher. Nur 849 (85,4%) der Fragebogen lagen vollständig beantwortet vor, in 145 (14,6%) Fällen war jeweils ein Item nicht angekreuzt. Im Sinne einer konservativen Missing-Analyse waren bei beiden Stichproben Fragebogen mit mehr als einer fehlenden Antwort bereits im Vorfeld ausgeschlossen worden. In einem Imputationsverfahren nach PRELIS 2.30 (Jöreskog & Sörbom, 2000) wurden die fehlenden Werte mit den Daten der jeweils ähnlichsten Fragebogen ersetzt. Die Imputation ist dem Ersetzen fehlender Werte durch Null (Conners, 1989; Achenbach, 1991) oder durch den jeweiligen Mittelwert überlegen, da sie gemäß des sonstigen Antwortprofils die wahrscheinlichste Angabe einfügt. Angesichts der relativ großen Fallzahl hätte auch ein fallweiser Ausschluß erfolgen können. In einer früheren Arbeit über die Conners-Eltern-Skalen (Huss et al., 2001) konnten wir jedoch zeigen, daß Kinder, deren Eltern mindestens eine Frage nicht beantwortet hatten, insgesamt deutlich höhere Symptomausprägungen aufwiesen. Dieses Ergebnis ließ sich auch an den aktuellen Stichproben zeigen (Summenwert: $F = 12{,}8$; $p < 0.000$**). Ein fallweiser Ausschluß der Fragebogen mit fehlenden Angaben hätte somit zu einer Verzerrung der Stichprobe geführt, wobei die schwerer betroffenen Kinder von den Analysen ausgeklammert worden wären.

Ergebnisse

Zunächst wurde eine deskriptive Analyse der Geschlechts- und Stichprobeneffekte bezogen auf den Summenwert und die Skalen der CTRS durchgeführt. Wie Abbildung 1 zu entnehmen ist, liegen die Werte der Klinikstichprobe erwartungskonform deutlich über denen der Normstichprobe (Mittelwert des Summenwerts: 29,2 vs. Norm: 15,6; $F = 186{,}5$; $df = 1$; $p < .000$**). Darüber hinaus zeichnen sich in beiden Gruppen ausgeprägte Geschlechtseffekte ab (Mittelwert Mädchen: 16,9 vs. Jungen: 29,6; $F = 179{,}6$; $df = 1$; $p < .000$**). In einem gemeinsamen univariaten Modell ergeben sich erwartungskonform sowohl für den Einfluß des Geschlechts als auch der Stichprobenzugehörigkeit hoch signifikante Haupteffekte (Geschlecht: $F = 89{,}6$; $df = 1$; $p < .000$ **; Stichprobe: $F = 112{,}4$; $df = 1$; $p < .000$**). Zusätzlich findet sich ein deutlich schwächer ausgeprägter Interaktionseffekt zwischen den beiden genannten Hauptfaktoren, der darauf zurückzuführen ist, daß in der Klinikstichprobe die Jungen überproportional hohe Merkmalsausprägungen haben (Geschlecht * Stichprobe $F = 4{,}5$; $p = .04$*). Das Alter wurde in den genannten Varianzmodellen als Kovariate berücksichtigt (Abb. 1).

Die drei Skalen der CTRS – Betragensprobleme, Impulsivität/Hyperaktivität und Unaufmerksamkeit – verhalten sich ähnlich wie der Summenwert. Die Erklärungskraft der Geschlechts- und Stichprobeneffekte weicht jedoch pro Skala etwas ab. So überwiegt in der

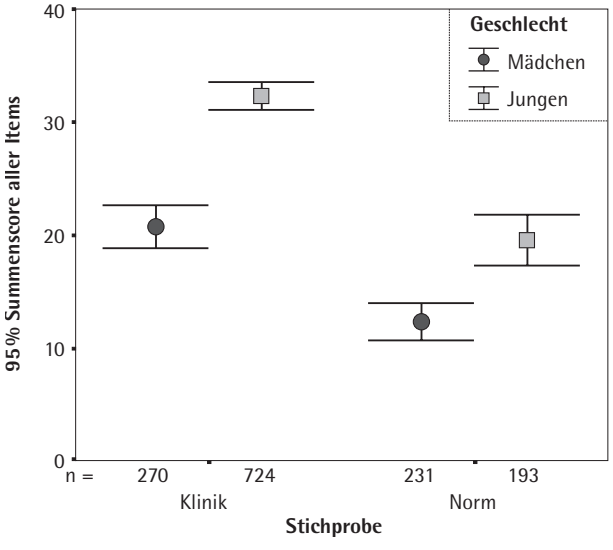

Abbildung 1. Geschlechts- und Stichprobeneffekte bezogen auf den Summenwert aller Items

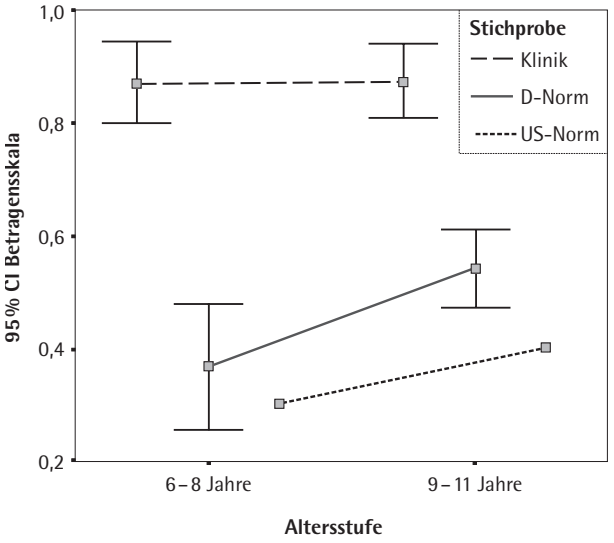

Abbildung 2. Rohdatenvergleich der drei Stichproben für die Betragensskala

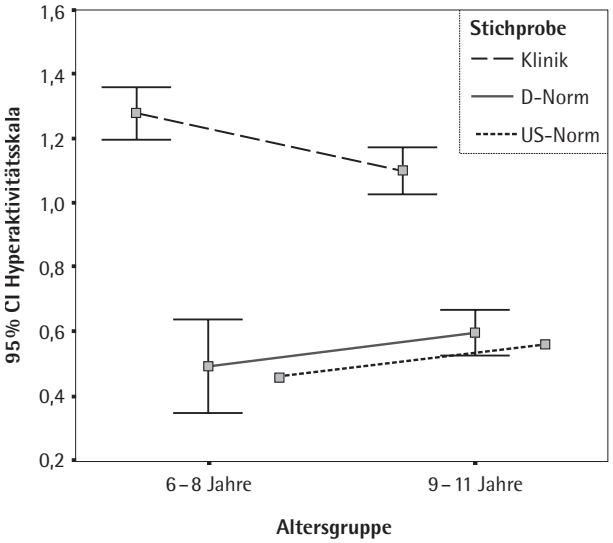

Abbildung 3. Rohdatenvergleich der drei Stichproben für die Skala ‚Impulsivität/Hyperaktivität'

Betragens- und der Impulsivitäts/Hyperaktivitäts-Skala jeweils der Geschlechtseffekt (Geschlechtseffekt für Betragen: F = 52,4; Geschlechtseffekt für Impulsivität/Hyperaktivität: F = 120,1; Stichprobeneffekt für Betragen: F = 37,4; Stichprobeneffekt für Impulsivität/Hyperaktivität: F = 77,5). Bei der Aufmerksamkeitsskala überwiegen dagegen die Stichprobeneffekte (Geschlechtseffekt: F = 44,4 versus Stichprobeneffekt: F = 154,3). Alle genannten F-Werte sind inferenzstatistisch in einem univariaten Varianzmodell mit $p < 0.000$** hoch signifikant.

Die oben beschriebenen Analysen machen deutlich, daß die Beurteilung von CTRS-Skalenwerten immer geschlechtsstratifiziert erfolgen sollte. Hinsichtlich der zu berechnenden Altersstrata gibt Conners (1989) fünf Altersstufen vor (3 – 5 Jahre; 6 – 8 Jahre; 9 – 11 Jahre; 12 – 14 Jahre; 15 – 17 Jahre). In der vorliegenden Studie beschränken wir uns auf eine Analyse der beiden unteren Altersgruppen, die im Rahmen der ADHS-Diagnostik von besonderem Interesse sind.

Um den Fehler einschätzen zu können, der begangen wird, wenn für deutsche Stichproben US-Normen verwendet werden, ist ein interkultureller Vergleich der Rohdaten erforderlich. Die Abbildungen 2, 3, und 4 geben die Skalen-Rohwerte gegliedert nach Altersstufe und Stichprobe wieder.

Dabei wird deutlich, daß sowohl in der deutschen als auch in der US-amerikanischen Normstichprobe alle Skalenwerte mit steigendem Alter zunehmen. Die Klinikstichprobe erweist sich als altersindifferent (Betragensskala), den Normstichproben gegenläufig (Impulsivitätsskala) bzw. den Normstichproben gleichsinnig (Aufmerksamkeitsskala).

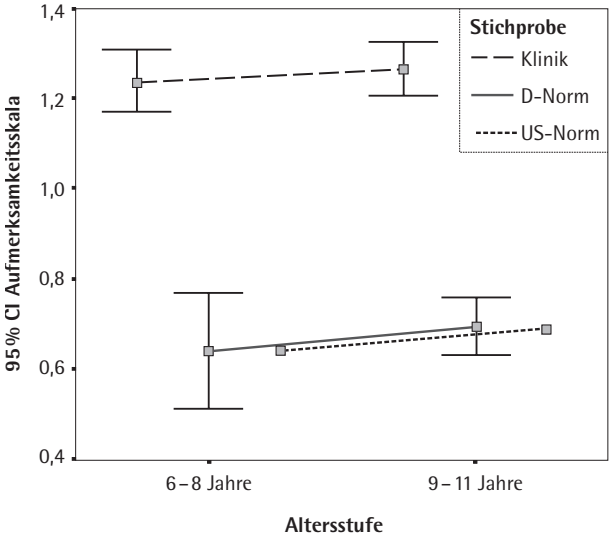

Abbildung 4. Rohdatenvergleich der drei Stichproben für die Aufmerksamkeitsstörung

Je nach Skala unterscheiden sich die Abweichungen zwischen US- und deutscher Norm. Die stärksten Abweichungen finden sich in der Betragensskala. Die deutschen Lehrer kreuzen auf dieser Skala deutlich höhere Werte an. Mäßig erhöhte Werte findet man bei der Impulsivitätsskala. Die Skala für Aufmerksamkeitsstörungen erweist sich zwischen den USA und Deutschland als nahezu identisch.

Neben Normierungsaspekten der CTRS stellt sich in der Praxis auch die Frage, wie gut das Instrument zwischen klinischen und nicht-klinischen Stichproben differenzieren kann. Methodisch bietet sich in diesem Zusam-

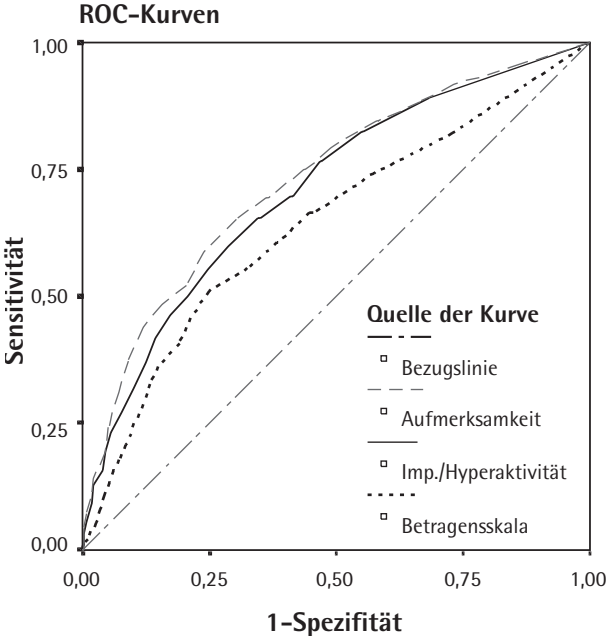

Abbildung 5. ROC-Kurven für den Vergleich von Klinik- und Normstichprobe für die drei Skalen

menhang der Einsatz sogenannter Receiver Operating Characteristic Curve (ROC) an. Die ROC zeigt für eine Verteilung das Verhältnis zwischen Sensitivität und Spezifität eines Tests für alle möglichen Trennpunkte an. Anhand der Fläche unter der Kurve kann die Güte des jeweiligen Tests eingeschätzt werden. Je größer die Fläche, desto besser der Test. Bei einem optimalen Test würde sowohl eine Sensitivität als auch eine Spezifität nahe 1 zu erwarten sein. Die Fläche unter der ROC-Kurve wäre dann maximiert. Ein nutzloser Test zeichnet sich dagegen durch einen Flächenanteil nahe 0,5 aus (siehe Diagonale in Abb. 5). Die Vorhersagegüte des Tests liegt dann im Zufallsbereich.

Wie Tabelle 1 zu entnehmen ist, weisen die Skalen unterschiedliche Kennwerte der Testgüte auf. Die schlechtesten Flächenwerte ergeben sich für die Betragensskala. Mit einer Fläche von 0,642 liegt das Ergebnis mäßig über der Zufallsquote. Eine treffsichere Trennung zwischen Klinik- und Normstichprobe ist damit anhand der Betragensskala nicht möglich. Bessere Werte ergeben sich mit .710 für die Skala Impulsivität/Hyperaktivität' und insbesondere mit .730 für die Aufmerksamkeitsskala.

Diskussion

Die vorgelegten Analysen machen deutlich, daß die US-Normen in dem Altersbereich zwischen sechs und elf Jahren nicht unkritisch auf deutsche Stichproben übertragen werden sollten. Bemerkenswert erscheint uns, daß keine globalen Abweichungen über alle Skalen und Stratifizierungsmerkmale wie Geschlecht und Alter hinweg zu verzeichnen sind. Als möglicher Globalfaktor hätte sich beispielsweise der Kultureffekt im Sinne einer systematischen Abweichung in den Lehrerurteilen niederschlagen können. Als weiterer potentieller Globalfaktor ist die Zeitdifferenz von etwa 15 Jahren zwischen den Erhebungen der beiden Normstichproben in Betracht zu ziehen.

In der vorgelegten Studie hat die jeweils analysierte Skala einen erheblichen Einfluß auf die Abweichungen zwischen US- und deutschen Skalenwerten. Die deutlichsten Unterschiede zwischen US- und deutscher Normstichprobe finden sich in der Betragensskala. Sie erweist sich auch als die Skala mit den schlechtesten Testkennwerten in den ROC-Analysen. Neben möglichen Kultureffekten sind in diesem Zusammenhang auch Änderungen im Urteilsanker innerhalb der vergangenen 15 Jahre in Betracht zu ziehen. Für diese Einschätzung spricht auch die bereits 1982 veröffentlichte Arbeit von Glow et al. (1982), bei der sich die Betragensskala bei der Überprüfung der Ein-Jahres-Stabilität der CTRS als besonders schlecht erwies.

Um darüber hinaus mögliche Kultureffekte für die Betragensskala besser einschätzen zu können, sei zunächst deren inhaltliche Definition auf Itemebene genauer betrachtet. Die Betragensskala wird im Lehrerurteil der CTRS aus acht Items berechnet. Von den Lehrern sollen u. a. oppositionelle Verhaltensweisen der Schüler eingeschätzt werden: „verhält sich unverschämt und frech" (Item 4), „mault und schmollt" (Item 10); „ist streitsüchtig" (Item 12); „verleugnet Fehler oder beschuldigt ande-

Tabelle 1. Flächenanteile, Standardfehler und Konfidenzintervall der ROC-Kurven

	ROC-Fläche	Standardfehler	95% Konfidenzintervall	
			Untergrenze	Obergrenze
Betragensprobleme	0,642	0,15	0,611	0,672
Impulsivität/Hyperaktivität	0,710	0,15	0,681	0,738
Aufmerksamkeitsprobleme	0,730	0,14	0,702	0,757

re“ (Item 23); „ist unkooperativ mit Lehrern“ (Item 27). In die Skala gehen aber auch emotionale Komponenten ein: „zeigt Wutausbrüche und unvorhergesehenes Verhalten“ (Item 5); „zeigt schnelle und ausgeprägte Stimmungswechsel“ (Item 11). Item 10 „ist übermäßig empfindlich gegenüber Kritik“ fällt inhaltlich etwas aus dem Rahmen der Betragensprobleme. Interessanterweise liegt dieses Item auch in der Originalarbeit (Goyette et al., 1978) mit .63 vergleichsweise gering auf der Betragensskala.

Für die Einschätzung, wie es zu den Abweichungen in den Betragensskalen zwischen US- und deutschen Skalenwerten kommt, lassen sich eine Reihe von Argumentationslinien verfolgen. In unserer Diskussion potentieller Effekte werden wir drei Aspekte aufgreifen: 1. Die Prävalenz der Betragensprobleme ist in den beiden Kulturen unterschiedlich ausgeprägt; 2. Der Urteilsanker der Lehrer unterscheidet sich; 3. Es handelt sich um nicht kulturbezogene Stichprobeneffekte.

Bei dem ersten Argument, demzufolge die Betragensprobleme in den untersuchten Kulturen unterschiedlich stark ausgeprägt sind, würden die gefundenen Unterschiede im Lehrerurteil nur die tatsächliche Prävalenz der Störung spiegeln. Bei deutschen Schülern müßten demnach häufiger und intensiver Betragensprobleme auftreten. Dieser Annahme widersprechen jedoch die umfassenden interkulturellen Analysen von Crijnen et al. (1999) an rund 12 000 Kindern und Jugendlichen aus zwölf Kulturen. Dabei lag die deutsche Normstichprobe auf der Aggressionsskala der Child Behavior Checklist 1,4 Wertpunkte unter dem internationalen Gruppenmittelwert. Verglichen mit den übrigen elf nationalen Normstichproben hatte die deutsche Stichprobe auf der Aggressionsskala die geringste Ausprägung. Auf der Skala Betragensprobleme ergaben sich keine Unterschiede gegenüber den anderen Nationen. Die Iteminhalte der CBCL-Aggressions- und Betragensskala zeigt hohe Ähnlichkeit mit der Betragensskala der CTRS, wobei einschränkend darauf hinzuweisen ist, daß es sich bei der CBCL-Beurteilung um ein Eltern-Rating der Kinder und Jugendlichen handelt.

Der zweite Erklärungsansatz geht von unterschiedlichen Urteilsankern zwischen den Lehrern aus. Demnach würde das gleiche Ausmaß an Betragensstörung in verschiedenen Kulturen unterschiedlich schwer eingeschätzt. Da bei interkulturellen Vergleichen kein ‚Archimedischer Punkt‘ definierbar ist, sind für Abweichungen in den Urteilsankern der Lehrer grundsätzlich zwei Betrachtungsweisen möglich. Einerseits könnten die US-Lehrer ihre Schüler hinsichtlich der Betragensprobleme besonders milde einschätzen. Zum anderen könnten die deutschen Lehrer das Auftreten von Betragensproblemen als besonders schwerwiegend einschätzen und damit zu einer Verschiebung der Basisrate beitragen.

In diesem Zusammenhang ergeben sich wichtige Hinweise aus anderen kulturvergleichenden Studien mittels der 39-Items-Version der CTRS. So konnten beispielsweise Taylor et al. (1984) zeigen, daß die Rohwerte der Betragensskala in einer englischen Feldstichprobe von 437 Kindern und Jugendlichen mit einem Mittelwert von 0,39 (Jungen) bzw. 0,20 (Mädchen) deutlich unter der US-amerikanischen Stichprobe (n = 291) von Werry (1975) liegen, die mit 0,21 für die Jungen und 0,08 für die Mädchen angegeben werden. Der Trend zu höheren Rohwerten für die Betragensskala läßt sich auch in der Arbeit von Werry et al. (1976) für eine Stichprobe von 418 Kindern und Jugendlichen aus Neuseeland sowie bei einer Studie aus Hong Kong (Luk et al., 1988) mit 914 Kindern und Jugendlichen nachweisen. In allen internationalen Vergleichen lag die US-Stichprobe in den Rohwerten der Betragensskala unter denen anderer Nationen. Da sich alle genannten Analysen auf die alte 39-Item-Version beziehen, lassen sich keine direkten Mittelwertsvergleiche mit unserer Stichprobe vornehmen. Im Rahmen des hier verfolgten Erklärungsansatzes über kulturbedingte Unterschiede in den Urteilsankern der Lehrer, sprechen die zitierten Studienergebnisse jedoch deutlich für die Annahmen einer systematischen Urteilsverschiebung bei den Lehrern im Sinne einer milden Verhaltensbeurteilung bei US-Lehrern.

Die zweite Skala der CTRS, die mit „Hyperaktivität/Impulsivität“ umschrieben wird, setzt sich aus sieben Items zusammen. Wie schon der Name vermuten läßt, finden sich in dieser Skala einerseits Items, die sich auf die motorische Unruhe der Kinder beziehen wie zum Beispiel „ist unruhig im Sinne von zappelig“ (Item 1); „ist unruhig, immer auf dem Sprung“ (Item 14). Zum anderen werden impulsive Verhaltensweisen beschrieben: „ist erregbar, impulsiv“ (Item 15); „Forderungen muß sofort entsprochen werden“ (Item 3). Weitere Items umfassen zusätzlich auch Aspekte von Betragensproblemen: „stört andere Kinder/Jugendliche“ (Item 8); „fordert die Aufmerksamkeit des Lehrers übermäßig ein“ (Item 16). Des weiteren findet sich auch ein Item ohne direkt nachvollziehbare inhaltliche Zuordnung (Item 2: „macht unangebrachte Geräusche“). In unserer Studie ergaben sich weniger ausgeprägte Unterschiede zwischen US- und deutscher Normierung. Auch die Trennschärfe zwischen Klinik- und Normstichprobe ist in dem ROC-Ansatz bei der Hyperaktivitäts-/Impulsivitätsskala besser als bei den Betragensproblemen. Die interkulturellen Einflüsse scheinen sich bei den genannten Merkmalen weniger stark auszuwirken.

Als nahezu identisch können die beiden Normstichproben für die Skala „Aufmerksamkeitsprobleme“ eingeschätzt werden. Auch hier soll eine inhaltliche Beschreibung der Items zu einem besseren Verständnis der Befunde beitragen. Die Skala setzt sich aus sieben Items zusammen. In erster Linie beschreiben die Items – der Skalenbezeichnung entsprechend – Aufmerksamkeitsproble-

me: „Ablenkbarkeit und Aufmerksamkeitsspanne sind problematisch“ (Item 7), „tagträumt“ (Item 9), „bringt angefangene Dinge nicht zu Ende“ (Item 21).

Die erstaunliche Übereinstimmung der Aufmerksamkeitsprobleme über Kulturen und größere Zeiträume hinweg ist möglicherweise darauf zurückzuführen, daß mit dieser Skala in erster Linie Merkmale erfragt werden, bei denen eine starke neurobiologische Beeinflussung vermutet wird. Sozial-interaktive Komponenten treten bei der Aufmerksamkeitsskala in den Hintergrund.

Bei einem systematischen Vergleich mit den von Conners (1989) und Goyette et al. (1978) mitgeteilten sozioökonomischen Daten lassen sich Stichprobeneffekte nicht ausschließen, erscheinen jedoch in vielfacher Hinsicht als unwahrscheinlich. Sowohl bei der US- als auch bei der deutschen Normstichprobe handelt es sich um Feldstichproben, die aus unterschiedlichen Schulsystemen rekrutiert wurden. Die Rekrutierung der US-Stichprobe erfolgte randomisiert anhand des Einwohnerregisters von Pittsburgh. Über Verweigerer-Raten liegen keine Angaben vor. Die deutsche Stichprobe wurde über den Schulkontakt nach Verfügbarkeit ausgewählt.

Die Altersmittelwerte der Gesamtstichprobe unterscheiden sich nur unwesentlich (Deutsche Normstichprobe: 9,4 Jahre mit SD = 1 Jahr; US: 9,9 Jahre ohne Angaben über die Streuung). Das Geschlechtsverhältnis weicht zwischen den jeweiligen Gesamtstichproben ab (D: 55% Mädchen zu 45% Jungen; US: 55% Jungen zu 45% Mädchen). Da die Analysen jedoch alters- und geschlechtsstratifiziert erfolgen, sind keine diesbezüglichen Effekte zu erwarten. Der Anteil der Geschwisterkinder ist in der US-Stichprobe aufgrund des familienbasierten Rekrutierungsmodus überproportional hoch. Die 570 in Pittsburgh untersuchten Kinder stammen aus 277 Familien. Damit hat durchschnittlich fast jedes Kind ein Geschwister in der Stichprobe. Der Geschwisteranteil in der deutschen Normstichprobe ist nicht bekannt, muß aber aufgrund der demographisch zu erwartenden Kinderzahl pro Familie etwas niedriger angesetzt werden. Die ethnische Zugehörigkeit wird in der US-Stichprobe zu 98% mit kaukasisch (weiß) angegeben. In der deutschen Stichprobe liegt dieser Anteil bei 94%.

An praktischer Relevanz läßt sich aus der vorgelegten Studie ableiten, daß die US-Normen der Conners-Skalen für Lehrer nicht unkritisch bei deutschen Stichproben verwendet werden sollten. Insbesondere für sozial auffälliges Verhalten sind erhebliche Verzerrungen zu erwarten. Werden deutsche Normwerte nicht berücksichtigt, so resultiert daraus eine fälschlich hohe Rate an Schülern mit vermeintlichen Hinweisen auf eine ‚Sozialstörung‘. Für den Bereich der Aufmerksamkeitsstörungen lassen sich dagegen mit US-Normen sehr präzise Aussagen treffen. Die Trennschärfe zwischen Norm- und Klinikstichprobe ist akzeptabel, reicht jedoch nicht aus, um auf der Grundlage der Conners-Skalen für Lehrer eine hinreichend genaue Zuordnung vornehmen zu können.

Literatur

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist and 1991 Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Barkley, R. A. (1987). The assessment of attention deficit-hyperactivity disorder. *Behavioral Assessment, 9,* 207–233.

Blondis, T. A., Snow, J. H., Stein, M. & Roizen, N. J. (1991). Appropriate use of measures of attention and activity for the diagnosis and management of attention deficit hyperactivity disorder. In P. J. Accardo, T. A. Blondis & B. G. Whitman (Eds.), *ADDH in Children (pp. 2–9).* New York: Dekker.

Brocke, B., Schuck, P. & Bruns, I. (1986). Testtheoretische Analyse der Conners-Skala zur Erfassung des Hyperkinese-Syndroms. *Zeitschrift für Klinische Psychologie, 15,* 177–200.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *American Journal of Psychiatry, 126,* 884–888.

Conners, C. K. (1989). *Conners Rating Scales – Manual.* New York: Multi Health Systems.

Conners, C. K., Siatrenios, G., Parker, J. D. & Epstein, J. N. (1998). Revision and restandardization of the Conners Teacher Rating Scale (CTRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology, 26,* 279–291.

Crijnen, A. A., Achenbach, T. M. & Verhulst, F. C. (1999). Problems reported by parents of children in multiple cultures: The Child Behavior Checklist Syndrom Constructs. *American Journal of Psychiatry, 156,* 569–574.

Deimel, W., Schulte-Körne, G. & Remschmidt, H. (1997). Welchen Nutzen haben die Conners-Lehrer-Fragebögen für die klinische Forschung und Praxis? *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 25,* 174–186.

Glow, R. A., Glow, P. H. & Rump, E. E. (1982). The stability of child behavior disorders: A one year test-retest study of Adelaide Versions of the Conners Teacher and Parent Rating Scales. *Journal of Abnormal Child Psychology, 10,* 33–60.

Goyette, C. H., Conners, C. K. & Ulrich, R. F. (1978). Normative data on Revised Conners Parent and Teacher Rating Scales. *Journal of Abnormal Child Psychology, 6,* 221–236.

Huss, M., Iseler, A. & Lehmkuhl, U. (2001). Interkultureller Vergleich der Conners-Skalen: Läßt sich die US-amerikanische Faktorenstruktur an einer deutschen Stichprobe replizieren? *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 29,* 16–24.

Jöreskog, K. & Sörbom, D. (2000). Softwarepaket LISREL 8.30 und PRELIS 2.30. Scientific Software International, Inc.

Luk, S. L., Leung, P. W. & Lee, P. L. (1988). Conners' Teacher Rating Scale in Chinese children in Hong Kong. *Journal of Child Psychology and Psychiatry, 29,* 165–174.

Taylor, E. A. & Sandberg, S. (1984). Hyperactive behavior in English schoolchildren: A questionnaire survey. *Journal of Abnormal Child Psychology, 12,* 143–156.

Werry, J. S., Sprague, R. L. & Cohen, M. (1975). Conners' Teacher Rating Scale for use in drug studies with children – An empirical study. *Journal of Abnormal Child Psychology and Psychiatry, 3,* 217–229.

Werry, J. S. & Hawthorne, D. (1976). Conners Teacher Questionnaire – Norms and validity. Australia and New Zealand. *Journal of Psychiatry, 10,* 257–262.

Dr. Michael Huss
Dipl.-Psych. Harriet Salbach
Dipl.-Psych. Marlies Ahle
Prof. Dr. Ulrike Lehmkuhl

Klinik für Psychiatrie, Psychosomatik und Psychotherapie
des Kindes- und Jugendalters, Charité
Virchow-Klinikum der Humboldt-Universität zu Berlin
Augustenburger Platz 1
D-13353 Berlin

Dr. Christina Stadler

Klinik für Psychiatrie und Psychotherapie
des Kindes- und Jugendalters der
Johann Wolfgang Goethe-Universität
Deutschordenstraße 50
D-60590 Frankfurt am Main

Dipl.-Psych. Patrick Mayer

Psychologisches Institut
der Freien Universität Berlin
Habelschwerdter Allee 45
D-14195 Berlin