

Hans-Ludwig Freese

Was mißt und was leistet 'Leistungsmessung mittels C-Tests'?*

Der vorliegende Beitrag setzt sich kritisch mit dem C-Test auseinander. Es werden folgende Hauptkritikpunkte genannt: (1) Hohe Korrelationen zwischen C-Tests und Außenkriterien (z.B. Werte aus anderen Sprachtests oder Lehrerurteile) dürfen nicht – wie weithin üblich – dahingehend interpretiert werden, daß C-Test und Außenkriterien das gleiche messen. (2) Da man nicht weiß, welche Fähigkeiten die Außenkriterien genau erfassen, erfährt man über eine Korrelation mit Außenkriterien auch nichts darüber, was mit dem C-Test genau gemessen wird. (3) Ohne Informationen über die differentielle Benachteiligung bestimmter Personengruppen ist der Einsatz von C-Tests in höchstem Maße fragwürdig. Es wird argumentiert, daß nur Analysen der bei der Lösung von C-Tests beteiligten Prozesse, etwa im Stil der Intelligenz- und Leseprozessforschung, darüber Aufschluß geben können, was der C-Test genau mißt.

1. Einleitung

Es wäre zu schön, um wahr zu sein: ein Test, der gleichermaßen im Mutter-, Fremd- und Zweitsprachenunterricht einsetzbar ist, der zur Einschätzung aller möglichen Beherrschungsgrade einer Sprache taugt, der systemlinguistische und kommunikative Kompetenz erfaßt, der dazu noch reliabel und valide, in seiner Konstruktion narrensicher und in der Anwendung praktikabel und ökonomisch ist. Alle diese Vorzüge versammeln nach Ansicht ihrer Verfechter Sprachtests, die nach dem C-Test-Prinzip erstellt sind.

C-Tests beruhen auf einer Modifikation des Cloze-Prinzips. Die von U. Raatz und C. Klein-Braley eingeführte Abwandlung des Cloze-Verfahrens

* Bei dem vorliegenden Beitrag handelt es sich – bis auf einige geringfügige vom Herausgeber dieses Bandes vorgenommene Kürzungen und Änderungen – um ein unveröffentlichtes (und auch ursprünglich nicht zur Veröffentlichung bestimmtes) Manuskript aus dem Jahre 1985. Für die verspätete Publikation sprechen u.a. folgende Gründe: 1. In einer Reihe von Publikationen zum C-Test wird auf das Manuskript von 1985 Bezug genommen; 2. Das Manuskript enthält eine Reihe von Kritikpunkten, die auch heute noch wichtig erscheinen. Der Herausgeber dankt dem Autor für seine Bereitschaft, eine nur noch bedingt aktuelle Arbeit der kritischen Diskussion zugänglich zu machen. (Anmerkung des Herausgebers)

dient vor allem dazu, derartige Tests leichter zu machen, ihnen den Charakter von Leseverständnisprüfungen zu nehmen und nicht zuletzt dazu, "objektiver" entscheiden zu können, was als richtige und falsche Antwort gewertet werden soll.

Insbesondere Oller (vgl. z.B. Oller, 1979) sah im Cloze-Verfahren (und im Diktat) eine ideale Testprozedur, da sie den ganzheitlichen und kommunikativen Charakter von Sprache berücksichtigt, der bei den sog. Discrete-Point-Tests ausgeblendet wird. Die schon von Carroll (1961) vorgebrachte Kritik an der elementaristischen Testpraxis und seine Forderung nach Tests, die sich auf den gesamten kommunikativen Effekt einer Äußerung richten, wurde für Oller zum Ausgang seiner Bemühungen nachzuweisen, daß Sprachbeherrschung ein unteilbares, nicht in spezifische Teilfertigkeiten und Wissens Elemente zerlegbares Phänomen darstellt, was sich statistisch in der Existenz eines allgemeinen Faktors, des sog. "language proficiency factors", spiegelt. Dabei wird Sprachbeherrschung als eine hypothetische Größe behandelt, deren Extension aus dem Umfang der Ausnutzung der in allen sprachlichen Äußerungen enthaltenen Redundanz erschlossen wird. Oller bezeichnet das System der internalisierten linguistischen Wissensbestände, mit deren Hilfe sich Übergangswahrscheinlichkeiten für die Produktion und Rezeption von sprachlichen Äußerungen nutzen lassen, als Erwartungsgrammatik (vgl. Oller, 1976). Der Ausdruck "Erwartungsgrammatik" darf jedoch nicht darüber hinwegtäuschen, daß Cloze- und C-Tests nicht eine explizite Grammatik als Theorie der Sprache zugrunde liegt und daß damit gegen das Postulat, daß jeder Meßoperation eine Theorie des zu messenden Gegenstandsbereichs zugrunde liegen muß, verstoßen wird.

2. Validierung mit Hilfe von Außenkriterien

In Ermangelung einer expliziten Theorie der Sprache und Sprachbearbeitung müssen die Konstrukteure von Cloze- und C-Tests zum Nachweis der Gültigkeit ihrer Messungen Zuflucht zur sog. Außenvalidierung nehmen, d.h., sie versuchen nachzuweisen, daß das, was C-Tests messen, etwas mit dem zu tun hat, was sich auf andere Weise über die zu messenden Fähigkeiten ermitteln läßt. Aus der Korrelation von C-Test-Werten mit dem Lehrerurteil und Testwerten, die mit Hilfe anderer Sprachtests gewonnen wurde, schließen sie auf das, was C-Tests messen, zurück.

Aus der Korrelation von C-Test-Werten mit Lehrereinschätzungen und Schulnoten bzw. mit Einstufungstestwerten in der Größenordnung von etwa

.40 bis .80 schließen sie (fälschlicherweise), daß die untersuchten C-Tests im wesentlichen das messen, was die Lehrer mit ihren globalen Schulnoten beurteilen.

Diese Interpretation geht von der Annahme aus, daß zwei hoch miteinander korrelierende Prüfverfahren quasi dieselben Fähigkeiten erfassen. Der Korrelationskoeffizient erlaubt aber nur Aussagen über das statistische "Zusammenvariieren" zweier Meßreihen. Angenommen, zwei Tests korrelieren mit .80 miteinander. Diese hohe Korrelation ist zwar zurückzuführen auf ihre hohen Ladungen auf den ihnen gemeinsamen Faktoren. Die Stärke dieser Faktoren kann allerdings für die beiden Tests sehr unterschiedlich sein. Farhady (1983, S. 317) zieht hieraus den folgenden Schluß:

"This implies that they [the tests] do not test the same underlying factors in the same way, though they correlate highly with each other and will probably provide overall similar results. Thus, a high correlation between two given tests should not be interpreted as though they tested the same thing."

Außerdem ist zu beachten, daß die Korrelation zwischen den verschiedenen Verfahren zur Gewinnung von Informationen über Testpersonen, über die Selektions- bzw. Plazierungsentscheidungen getroffen werden müssen, zwar hoch, aber nicht so hoch ist, daß die Entscheidungen auf der Grundlage unterschiedlicher Testinformationen gleich ausfallen. Es ist gänzlich offen, ob im Einzelfall mit Hilfe des Lehrerurteils, Einstufungstests oder C-Tests "bessere" oder "schlechtere" Entscheidungen getroffen werden.

Farhady gibt zu bedenken, daß, obwohl Discrete-Point-Einstufungstests und integrative Tests wie der C-Test im allgemeinen quasi identische Meßergebnisse zeitigen können, es sehr wohl möglich ist, daß bestimmte Testpersonengruppen je unterschiedliche Testresultate erzielen, je nachdem ob sie mit dem einen oder anderen Testformat getestet wurden. So wird z.B. ein thailändischer Schüler, der in der Zweitsprache große Probleme mit der Lautung und der Phonem-Graphem-Zuordnung hat, vermutlich bei einem C-Test völlig versagen, obwohl er über sonst gut entwickelte zweitsprachliche Fähigkeiten verfügt. Dasselbe gilt für Legastheniker, Probanden, die nur eine geringe formale Schulbildung aufweisen, und möglicherweise für viele andere mehr. Bevor also nicht die differentielle Benachteiligung bestimmter Probandengruppen ausgeschlossen werden kann, ist der Einsatz von C-Tests in höchstem Maße fragwürdig.

3. Zur Konstruktvalidität von C-Tests

Da man nicht weiß, welche Fähigkeiten im einzelnen mit dem Lehrerurteil bzw. mit Einstufungstests erfaßt werden, erfährt man aus ihren Korrelationen mit C-Tests auch nichts darüber, welche Fähigkeiten mit diesen erfaßt werden. Um das Konstrukt "globale Sprachbeherrschung" aufzuklären, haben Raatz und Klein-Braley C-Test-Werte mit einer Reihe von Variablen wie Schulnoten in Deutsch, Englisch, Mathematik, einer Einschätzung von Grammatikkenntnissen, dem diagnostischen Test Deutsch DTD 4-6, dem Rechtschreibtest DRT 4-5, dem sprachfreien Intelligenztest "Frankfurter Denkaufgaben FDA 3-6" und einem Konzentrationstest (test d2) korreliert und die Korrelationsmatrix faktorisiert (vgl. z.B. Raatz, 1985). Es zeigt sich, daß der C-Test mit allen diesen Variablen, deren Auswahl nicht begründet wird, mit Ausnahme des Konzentrationstests, signifikant positiv korreliert ist. Der C-Test besitzt die höchsten Korrelationen mit dem Rechtschreibtest sowie mit dem Untertest Textstrukturierung, hohe Korrelationen mit Deutsch und Grammatik, mittlere Korrelationen mit Mathematik und allgemeiner Intelligenz (vgl. Raatz, 1985, S. 49).

Diese Untersuchung wurde an 75 deutschen Schülern des 5. Schuljahres mit einem muttersprachlichen C-Test durchgeführt. Es ist nicht auszuschließen, daß die entsprechenden Korrelationen bei älteren Schülern, in der Fremd- oder Zweitsprache, bei verschiedenen Nationalitäten usw. sehr unterschiedlich ausfallen, was zu dem Schluß nötigen würde, daß C-Tests bei verschiedenen Probandengruppen und Textschwierigkeiten (s. weiter unten) verschiedene Fähigkeiten anzupfen.

Was aber an diesem Ergebnis verräterisch erscheint, ist die Tatsache, daß muttersprachliche C-Tests höher mit Rechtschreibtests als mit der Deutschnote korrelieren. Kaum jemand dürfte sich dazu bereit finden, Rechtschreibfähigkeiten als den besten Indikator für muttersprachliche Kompetenz auszugeben. Es bleibt unerfindlich, warum man in diese Analyse nur einen sprachfreien Intelligenztest einbezogen hat. Wäre ein verbaler Intelligenztest verwendet worden, so wäre m.E. der Charakter von C-Tests als faktoriell hochgesättigt mit verbaler Intelligenz, d.h. der Fähigkeit, komplexes sprachliches Material zu verstehen und zu verarbeiten, noch deutlicher geworden. Nicht umsonst lassen sich Werte aus Tests zur allgemeinen Intelligenz am besten durch Wortschatz- und Leseverständnisleistungen vorher-sagen.

Außerdem scheinen mir die verwendeten Tests zur Überprüfung der faktoriellen Struktur von C-Tests zu unspezifisch ausgewählt worden zu sein. Textspezifische Tests, die die für die jeweiligen Texte relevanten lexikalischen Kenntnisse, Rechtschreibkenntnisse und Leseverständnisleistungen abprüfen, würden sehr viel höher mit den globalen C-Test-Werten korrelieren und damit mehr über das, was C-Tests erfassen, aussagen.

Daraus folgt, daß wegen der mangelnden Explikation des theoretischen Konstrukts "allgemeine Sprachkompetenz", von der behauptet wird, sie werde von C-Tests erfaßt, eine Interpretation von C-Test-Ergebnissen fast der Beliebigkeit anheimgegeben ist, da nicht bekannt ist, was eigentlich damit gemessen wird. Was im Falle der Noten gemessen wird, wird erst aus der Kenntnis des Curriculums und der Kenntnis eines je besonderen Unterrichts erschließbar, ähnliches gilt für traditionelle Fremdsprachentests. Da aber C-Tests gar keinen Bezug auf die Bedingungen nehmen, unter denen sprachliche Fähigkeiten erworben wurden, sind die damit gewonnenen Informationen nicht für didaktische und unterrichtsorganisatorische Entscheidungen relevant.

Wie problematisch der Einsatz von C-Tests ist, mag folgendes Beispiel illustrieren: In einem deutschsprachigen C-Test erreicht ein ausländischer Schüler denselben Testwert wie ein deutscher Schüler. Heißt das, daß beide die gleiche globale Sprachkompetenz im Deutschen besitzen und daß beide in gleicher Weise von demselben Sprachunterricht profitieren? Das ist zumindest unwahrscheinlich. Der muttersprachliche Sprecher wird aller Wahrscheinlichkeit nach über andere Konfigurationen von Fähigkeiten verfügen als der Sprecher, für den die Sprache des Tests Fremd- oder Zweitsprache ist. Ein in der Zweitsprache literarisch gebildeter Sprecher, der über eine geringe interpersonale Kommunikationsfähigkeit verfügt, wird wahrscheinlich weitaus besser mit C-Tests zurechtkommen als ein Muttersprachler mit einer geringen Schulbildung, dem man aber eine globale Sprachkompetenz von hohen Graden bescheinigen kann.

Dieser Sachverhalt läßt sich auch am Problem der Auswahl von Texten mit einem geeigneten Schwierigkeitsgrad für C-Tests illustrieren. Dem C-Test liegt die Annahme zugrunde, daß mit "schwierigen" Texten je höhere Grade der allgemeinen Sprachkompetenz erfaßt werden. Aus der linearen Beziehung zwischen Lebensalter und Textverständnis und aus der Annahme, daß der Lernprozeß in der Muttersprache bzw. Umgangssprache frühestens in der Pubertät abgeschlossen ist, wird geschlossen, daß Sprach-

kompetenz eine Größe sei, die rein quantitativen Veränderungen unterliegt. Es sprechen aber gute Gründe für die Annahme, daß die Sprachentwicklung bereits im Alter von 6-7 Jahren, was die Beherrschung des formalen Regelsystems anlangt, weitgehend abgeschlossen ist und daß alle übrigen Aspekte der Sprachbeherrschung in ihrer Entwicklung prinzipiell niemals gänzlich zum Stillstand kommen. Das würde bedeuten, daß ein L1-Sprecher von etwa 7 Jahren einen Grad der (formalen) Sprachbeherrschung erreicht hat, den ein noch so fortgeschrittener L2-Sprecher nur in den seltensten Fällen erreicht, obwohl dieser u.U. höhere C-Testwerte aufzuweisen hat.

Die Leseprozeßforschung legt darüber hinaus die Annahme nahe, daß sich leichtere Texte mit Rückgriff auf andere Fähigkeiten bewältigen lassen als schwierigere Texte. Schwierigere Texte sind u.a. eben dadurch schwierig, daß sie seltenere und längere Wörter und längere Sätze enthalten, die an die Verarbeitungskapazität, das Weltwissen, die Fähigkeit zu schlussfolgerndem Denken usw. andere Ansprüche stellen als leichtere Texte. C-Tests, die für siebenjährige L1-Sprecher "leicht" sind, können für erwachsene L2-Sprecher, die mit C-Tests fertig werden, die wiederum für die siebenjährigen L1-Sprecher "schwierig" sind, zu schwer sein. Ist aber die Frage nicht abwegig, ob siebenjährige L1-Sprecher, die sich problemlos in alltäglichen Kommunikationssituationen verständigen können, über eine größere oder geringere allgemeine Sprachkompetenz verfügen als L2-Benutzer, die mit anspruchsvollen Texten fertig werden, aber keine elementaren umgangssprachlichen Fertigkeiten erworben haben? Oder um ein anderes Beispiel zu geben, kann man geistig Behinderten, die in C-Tests ganz und gar scheitern, eine allgemeine Sprachkompetenz absprechen, obwohl sie erstaunliche sprachliche Fähigkeiten in der alltäglichen Kommunikation unter Beweis stellen?

4. Abschließende Wertung

M.E. sind sprachliche Leistungen viel zu komplex und voraussetzungsreich, als daß sie mit Hilfe eines so einfachen Verfahrens wie C-Tests dingfest gemacht werden könnten. Zwar messen C-Tests irgendetwas, was sie aber jeweils messen, darüber können nur Analysen der beteiligten Aufgabenlösungsprozesse, etwa im Stil der Intelligenz- und Leseprozeßforschung, Aufschluß geben. Solange nicht bekannt ist, welche Sprachfähigkeiten jeweils gemessen werden, ist ihr Einsatz zu Selektions- und Klassifikationsentschei-

dungen äußerst fragwürdig, auch schon deshalb, weil sie nicht auf mögliche unterschiedliche Zielsetzungen von Unterricht Bezug nehmen.

Schlimmer wäre noch, wenn derartige Tests normative Wirkungen zeitigten, eine Gefahr, die bei jeder Testverwendung gegeben ist, d.h., wenn Lehrer zur Vorbereitung ihrer Schüler auf derartige Prüfungen verstärkt Fertigkeiten wie Rechtschreibsicherheit trainierten oder exzessiv viele Einsetzübungen in der Art von C-Tests durchführten, was doch wohl nicht erwünscht sein kann, oder wenn Lehrer den Umkehrschluß zögen und zur Entwicklung und Förderung der allgemeinen Sprachkompetenz sich auf Einsetzübungen kaprizierten. Die Fähigkeit, Lücken in Lückentests zu füllen, läßt sich sicherlich durch Training steigern, ob aber damit auch die allgemeine Sprachkompetenz angehoben wird, ist nicht nur eine Frage der Definition, sondern in erster Linie ein empirisches Problem, das weiterer Aufklärung bedarf.

Literaturverzeichnis

- Carroll, John B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In Center for Applied Linguistics (Hrsg.), *Testing the English proficiency of foreign students* (S. 30-40). Washington, DC: Center for Applied Linguistics. (reprinted in: Allen, Harold B. & Campbell, Russell N. (Hrsg.). (1972). *Teaching English as a second language: a book of readings* (S. 313-320). New York: McGraw Hill.)
- Farhady, Hossein. (1983). The disjunctive fallacy between discrete-point and integrative tests. In John W. Oller, Jr. (Hrsg.), *Issues in language testing research* (S. 311-321). Rowley, Mass.: Newbury House. (zuerst in: *TESOL Quarterly*, 13, 1979, 347-357.)
- Oller, John W., Jr. (1976). Evidence of a general language proficiency factor: an expectancy grammar. *Die Neueren Sprachen*, 75, 165-174.
- Oller, John W., Jr. (1979). *Language tests at school. A pragmatic approach*. London: Longman.
- Raatz, Ulrich. (1985). The factorial validity of C-Tests. In Christine Klein-Braley & Ulrich Raatz (Hrsg.), *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis* (S. 42-54). Bochum: AKS.

Adressen der Autoren

- Ulrike Arras, c/o Rüdiger Grotjahn, Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, D-44780 Bochum.
- Rupprecht S. Baur, Universität GH Essen, Fachbereich 3, Literatur- und Sprachwissenschaften, Universitätsstraße 12, D-45117 Essen.
- Brian Bresnihan, Kobe Shoka Daigaku, 8-2-1 Gakuen Nishi-machi, Niskiku, Kobe-shi 651-21, Japan.
- James A. Coleman, School of Languages and Area Studies, University of Portsmouth, Wiltshire Building, Hampshire Terrace, Portsmouth PO 1 2BU, England.
- Hans-Ludwig Freese, Freie Universität Berlin, Fachbereich 12, Erziehungs- und Unterrichtswissenschaften, Institut für Allgemeine und Vergleichende Erziehungswissenschaft, Arnimallee 10, D-14195 Berlin.
- Ulrich Germann, c/o Rüdiger Grotjahn, Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, D-44780 Bochum.
- Rüdiger Grotjahn, Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, D-44780 Bochum.
- Heike Hohenbleicher, c/o Rüdiger Grotjahn, Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, D-44780 Bochum.
- Gerhard Jakschik, Arbeitsamt Stuttgart, Psychologischer Dienst, Postfach, D-70049 Stuttgart.
- Christine Klein-Braley, Universität-GH-Duisburg, Fachbereich 3, Sprach- und Literaturwissenschaften, D-47048 Duisburg.
- Johann Köberl, Universität Klagenfurt, Institut für Anglistik und Amerikanistik, Universitätsstraße 65-67, A-9022 Klagenfurt, Österreich.
- Gregor Meder, Universität GH Essen, Fachbereich 3, Literatur- und Sprachwissenschaften, Universitätsstraße 12, D-45117 Essen.
- Ulrich Raatz, Universität-GH-Duisburg, Fachbereich 2, Psychologie, D-47048 Duisburg.
- Stratton Ray, Box 66, Teachers College, Columbia University, New York, NY 10021, USA.
- Undine Roos, Kürenbergweg 22, D-45279 Essen.
- Günther Sigott, Universität Klagenfurt, Institut für Anglistik und Amerikanistik, Universitätsstraße 65-67, A-9022 Klagenfurt, Österreich.
- Wolfgang Tönshoff, Goethe-Institut PKIN, Plac Defilad 1, PL-00-901 Warschau, Polen.