Grotjahn, Rüdiger. (Hrsg.). (2002). Der C-Test. Theoretische Grundlagen und praktische Anwendungen (Bd. 4, S. 211-225). Bochum: AKS-Verlag

Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis

Rüdiger Grotjahn*

The article, which is intended for the non-specialist, presents an overview of research into the C-Test. Topics dealt with include: (language-specific) modifications to the canonical C-Test format; development, analysis and scoring of C-Tests; reliability, objectivity, economy and validity; computer software; interpretation of C-Test results. The focus is on aspects that have proved important in the development and use of C-Tests. In the appendix the most important results are summarized in the form of a table.

1. Einleitung

Der C-Test ist ein ökonomisches und reliables Verfahren zur globalen Feststellung der allgemeinen Kompetenz in Fremd-, Zweit- und Erstsprachen. Er gehört zu den am gründlichsten untersuchten neueren Sprachtests (vgl. die C-Test-Bibliographie von Grotjahn, 1996) und ist insbesondere dann erfolgreich einsetzbar, wenn das Ziel eine vom vorangehenden Unterricht bzw. von der individuellen Lerngeschichte weitgehend unabhängige globale Sprachstandsfeststellung ist. Entsprechend werden C-Tests u.a. als Einstufungstest an Universitäten, Studienkollegs und Sprachenschulen eingesetzt (vgl. Grotjahn, Klein-Braley & Raatz, 1992; Klein-Braley & Grotjahn, 1998; sowie http://www.C-Test.de/).

Der vorliegende Beitrag richtet sich in erster Linie an den Nichtspezialisten: Im Vordergrund stehen Aspekte, die sich im Hinblick auf die Konstruktion und dem Einsatz von C-Tests in der Praxis als bedeutsam erwiesen haben. Zur Erleichterung der Orientierung sind die wichtigsten Gesichtspunkte im Anhang in einem „Leitfaden für die Praxis“ tabellarisch zusammengestellt. Weitere praxisbezogene Hinweise finden sich z.B. in Raatz & Klein-Braley (1995, 2002), Grotjahn (1987) und Grotjahn, Klein-Braley & Raatz (1992).

2. Das kanonische C-Test-Prinzip

C-Tests beruhen auf einer Variante des Cloze-Prinzips und auf dem theoretischen Prinzip der reduzierten Redundanz (vgl. Klein-Braley, 1997). Im Gegensatz zum klassischen Cloze Test werden mehrere kurze (aus etwa 60 bis 80

* Prof. Dr. R. Grotjahn, Seminar für Sprachlehrforschung, Ruhr-Universität Bochum, 44780 Bochum. E-mail: ruediger.grotjahn@ruhr-uni-bochum.de.

Der vorliegende Beitrag beruht in Teilen auf Grotjahn (1995).

Wörtern bestehende) Texte unterschiedlicher Thematik gewählt. Beginnend mit dem zweiten Wort des zweiten Satzes wird in jedem Text bei jedem zweiten Wort die zweite Hälfte getilgt. Wörter mit einem einzigen Buchstaben und Eigennamen bleiben unberücksichtigt. Bei Wörtern mit einer ungeraden Anzahl von Buchstaben ist die Zahl der getilgten Buchstaben um eins höher als die Zahl der nicht getilgten Buchstaben. In jedem Text sollte die gleiche Zahl von Tilgungen vorgenommen werden. Am Textende sollte ein kurzes unversehrtes Textstück als Kontext für die Lösungsfindung stehen bleiben.

Bei der Markierung der Lücken existieren mehrere Möglichkeiten: a) gleich lange durchgehende Striche für jede Lücke (ursprüngliches Prinzip); b) Strichlänge bei jeder Lücke in Abhängigkeit von der Zahl der getilgten Buchstaben; c) gestrichelte Linie mit einem Strich pro getilgtem Buchstaben. Die Varianten b) und c) machen den C-Test zwar zumeist leichter und reliabler, können jedoch einen negativen Effekt auf die Validität haben (Induzierung von Buchstabenzählen).

Die ausgewählten Texte werden in aufsteigender Schwierigkeit angeordnet. Es wird ein Punkt für jede exakte (oder bisweilen auch für jede akzeptable) Rekonstruktion des Originalworts gegeben (vgl. die detaillierten Hinweise in Grotjahn 1987a sowie Abschnitt 10). Entsprechend dem beschriebenen – als klassisch oder auch als kanonisch bezeichneten – Konstruktionsprinzip besteht ein C-Test meist aus vier bis sechs Texten mit jeweils 20 bis 25 Items (Lückenwörter), wobei häufig vier Texte mit 25 Items oder 5 Texte mit 20 Items gewählt werden.

3. Sprachspezifische Probleme mit dem kanonischen Konstruktionsprinzip

Im Französischen werden Zusammensetzungen mit Apostroph oder Bindestrich wie z.B. *d'autres* oder *celui-ci* bei der Tilgung zumeist jeweils als ein einziges Wort behandelt. Diese Regel führt jedoch z.B. im Italienischen in zahlreichen Fällen zu kaum rekonstruierbaren Wortverbindungen. So müsste in *dell'anno* oder *quell'epoca*, wenn diese als ein einziges Wort behandelt werden, *anno* bzw. *epoca* vollständig gelöscht werden. Es ist deshalb u.a. vorgeschlagen worden, jeweils nur die zweite Hälfte von entsprechender Verbindungen zu tilgen.

Im Deutschen oder auch im Italienischen ergibt sich z.B. bei Wortzusammensetzungen das Problem, dass das kanonische Konstruktionsprinzip zur Tilgung ganzer Wörter führt, die dann nicht mehr oder nur noch mit großen Problemen eindeutig rekonstruierbar sind. Es wird deshalb zuweilen der erste Buchstabe des letzten Wortes der Zusammensetzung unversehrt gelassen.

Bei den enklitischen Personalpronomina z.B. im Italienischen oder Portugiesischen sind nach der kanonischen Tilgungsregel z.B. in italienisch *regalarglielo* oder portugiesisch *dei-lho* die enklitischen Pronomina vollständig zu tilgen, was zu erheblichen Rekonstruktionsproblemen führen kann, insbesondere, wenn die Referenten der Pronomina ebenfalls zu den im Text beschädigten Wörtern gehören. Es ist deshalb vorschlagen worden, lediglich die zweite Hälfte der Enklitika zu tilgen.

Ein weiteres Problem können Graphemkombinationen darstellen, die einen Einzellaut repräsentieren. Tilgt man z.B. im Italienischen in *luglio*, *ogni* oder *viscere* die zweite Hälfte (d.h. *lio*, *ni* und *cere*), dann gehört der letzte nicht getilgte Buchstabe jeweils zu einer Graphemkombination, die keiner möglichen Aussprache dieses Buchstabens als Einzelgraphem entspricht. Da die Suche nach der Lösung häufig über innere Phonation und lautes Lesen erfolgt, kann der Getestete in solchen Fällen leicht irregeleitet werden. Dies kann vermieden werden, wenn man entsprechende Polygraphen entweder komplett tilgt (der Test wird schwieriger) oder komplett erhält (der Test wird leichter).

Besondere Probleme mit dem orthodoxen Tilgungsprinzip ergeben sich bei Sprachen wie Türkisch, Japanisch oder Chinesisch, die sich typologisch deutlich von den bisher zumeist in der C-Test-Forschung untersuchten Sprachen unterscheiden. Weiterführende Hinweise zur Sprachspezifik gibt Grotjahn (1995, 1997).

4. Sprachunspezifische Varianten des kanonischen C-Test-Formats

In einer Reihe von Arbeiten sind sprachunspezifische Modifikationen des kanonischen C-Test-Formats und deren Einfluss insbesondere auf die Testschwierigkeit untersucht worden (vgl. die Belege in Grotjahn, 1995, 1997). Entsprechende Studien sind in Bezug auf den praktischen Einsatz von C-Tests von erheblicher Bedeutung, da ein C-Test über eine Modifikation der Tilgungsregel gegebenenfalls leichter oder schwieriger gemacht werden kann und so eine bessere Passung zwischen Testschwierigkeit und Sprachstand möglich wird. Allerdings stellt sich das Problem, dass ein modifizierter C-Test möglicherweise partiell andere Fähigkeiten misst als ein entsprechender kanonischer C-Test.

Köberl & Sigott (1994) und Sigott & Köberl (1996) z.B. analysieren anhand von deutschen und englischen C-Tests drei verschiedene Modifikationsmöglichkeiten: Löschung a) von zwei Dritteln jedes zweiten Wortes; b) jedes zweiten Wortes mit Ausnahme des ersten Buchstabens; und c) der ersten Hälfte jedes zweiten Wortes. Die Varianten (a) und (b) reduzieren die Textredundanz und

führen zu signifikant schwierigeren C-Test-Texten. Als am reliabelsten im Deutschen, allerdings nicht im Englischen, erwies sich die Version (c).

Interessante Möglichkeiten eröffnen die Arbeiten von Kamimoto (1993) und Jafarpur (1999). Die Autoren untersuchen, ob sich mit Hilfe einer klassischen Itemanalyse auf der Basis der einzelnen Lücken die Reliabilität verbessern lässt, indem man hinsichtlich Schwierigkeit und Trennschärfe wenig zufriedenstellende Items von der Tilgung ausschließt. Da auf diese Weise die Zahl der Lücken reduziert und damit die Durchführungs- und Auswertungsökonomie erhöht werden kann, könnten verstärkt auch C-Tests eingesetzt werden, die aus mehreren längeren Texten bestehen. Allerdings dürfte ein itemanalysierter C-Test-Text aufgrund des resultierenden Tilgungsmusters partiell etwas anderes messen als ein kanonischer C-Test-Text. Auch wäre der Entwicklungsaufwand höher als bei einem kanonischen C-Test. Zudem kommt zumindest Jafarpur (1999) zu einer eher negativen Einschätzung hinsichtlich des Potentials der klassischen Itemanalyse zur Verbesserung von C-Tests.

5. Reliabilität, Objektivität und Ökonomie

C-Tests erweisen sich zumeist als hoch reliabel im Sinne einer Messkonsistenz auf der Ebene der einzelnen Texte (in der Regel lag Cronbachs Alpha zwischen 0.80 und 0.90). Dies ist um so erstaunlicher, als viele der untersuchten C-Test-Versionen zum ersten Mal eingesetzt wurden.

Auch die ermittelten Test-Retest-Reliabilitätskoeffizienten sind mit Werten zwischen 0.70 und 0.85 in Anbetracht der Tatsache, dass der Abstand zwischen den Testzeitpunkten bis zu einem Jahr betrug, erstaunlich hoch (vgl. Grotjahn, Klein-Braley & Raatz, 2002).

Neben einer vergleichsweise hohen Reliabilität besitzt der C-Test auch eine hohe Durchführungsobjektivität und erlaubt – zumindest dann, wenn lediglich exakte Lösungen oder vorher festgelegte akzeptable Lösungen als „korrekt“ gewertet werden – eine absolut objektive Auswertung.

Schließlich handelt es sich verglichen mit vielen anderen Testverfahren um ein relativ ökonomisches Testinstrument. Der Entwicklungsaufwand eines C-Tests ist weit geringer als z.B. der eines zufrieden stellenden Multiple-Choice-Tests. Die Durchführung eines C-Tests üblicher Länge dauert weniger als eine halbe Stunde, und die Auswertungszeit beträgt 1-2 Minuten pro Text und Proband (je nach Auswertungsmethode).

6. Validität

Die Ansicht, dass der C-Test ebenso wie der Cloze Test ein integratives Messinstrument zur Erfassung globaler Sprachkompetenz in Erst-, Zweit- und Fremdsprachen ist, stützt sich u.a. auf die zum Teil erstaunlich hohen Korrelationen von C-Tests mit verschiedenen Außenkriterien (z.B. Schulnoten, Lehrerurteilen über den Sprachstand der Schüler, Ergebnisse in anderen Sprachtests wie z.B. den TOEFL), sowie auf empirische Untersuchungen zur Konstruktvalidität unter Einschluss von introspektiven und experimentellen Verfahren. Zudem gibt es eine Reihe neuerer Belege, dass die Lerner bei der Rekonstruktion der C-Test-Lücken falls nötig auch den weiteren Kontext berücksichtigen und dass der C-Test damit nicht nur – wie zuweilen behauptet – auf der Mikroebene, sondern auch auf der Makroebene misst (vgl. z.B. Grotjahn, 2002; Grotjahn, Klein-Braley & Raatz, 2002; Grotjahn & Stemmer, 2002; Hastings, 2002; Klein-Braley, 1994, 1996; Sigott, 2002; Stemmer, 1991).

7. Textauswahl

Bei der Entwicklung von C-Tests geht man am besten von einer Auswahl von etwa 10 Texten aus. Sieht man einmal von dem Sonderfall diskursspezifischer C-Tests ab, dann sollten die gewählten Texte inhaltsneutral sein, kein Spezialvokabular enthalten, kein Spezialwissen verlangen, soweit wie möglich authentisch sein, eine Sinneinheit bilden und zudem möglichst zielgruppenadäquat sein. Die Auswahl erfolgt am besten durch Lehrende, die den Sprachstand und die Lerngeschichte der Probanden kennen. In Bezug auf den C-Test unerfahrene Lehrende überschätzen allerdings nicht selten die Schwierigkeit eines Textes.

In einer Reihe von Untersuchungen ist der Versuch unternommen worden, die Textschwierigkeit von C-Test-Texten anhand von Textmerkmalen zu bestimmen. Klein-Braley (1994) konnte dabei zeigen, dass es insbesondere anhand der *type-token-ratio* möglich ist, die empirische Schwierigkeit von englischen C-Test-Texten für bestimmte Lernergruppen regressionsanalytisch vorherzusagen.

Weiterhin ist untersucht worden, ob C-Tests zur Messung diskursspezifischer – und zwar vor allem fachsprachlicher Kompetenz – eingesetzt werden können (vgl. den Überblick bei Grotjahn, 1995 sowie Connelly, 1997; Daller, 1999; Daller & Grotjahn, 1999). Insgesamt gesehen spricht die Forschungslage zu Gunsten diskursspezifischer C-Tests.

8. Testinstruktion

Die Aufgabenstellung kann in der C-Test-Instruktion z.B. folgendermaßen beschrieben werden: „In den folgenden Texten fehlt bei einer Reihe von Wörtern ein Teil. Ergänzen Sie den fehlenden Teil in sinnvoller Weise.“ Genauere Hinweise zum Tilgungsprinzip sind nur im Hinblick auf einig sprachspezifische Konventionen nötig, wie z.B. „Wörter mit Bindestrich, wie z.B. *celui-ci*, zählen als ein einziges Wort.“ Weiterhin hat es sich als sinnvoll erwiesen, in der Testinstruktion ein (relativ großzügig bemessenes) Zeitlimit für jeden C-Test-Text anzugeben (ca. 4-5 Min. bei 20-25 Lücken pro Text). Bei der Administration ist die Einhaltung der Zeitvorgaben durch Nennung einer Restzeit bei jedem Text zu gewährleisten (z.B.: „Sie haben noch 2 Minuten Zeit zur Bearbeitung des Textes!“; nach 2 Minuten: „Sie sollten jetzt mit dem nächsten Text beginnen!“). Hierdurch wird sicher gestellt, dass die Kandidaten alle Texte bearbeiten, wodurch sich auch die Reliabilität erhöht.

9. Voruntersuchungen

Die Texte sollten möglichst anhand von Muttersprachlern und Lernern vorgetestet werden. Nicht selten weisen allerdings auch erstmals eingesetzte C-Tests erstaunlich gute Testkennwerte auf.

Beträgt die Lösungshäufigkeit bei Muttersprachlern weniger als 90%, ist der entsprechende Text auszuscheiden. Pilotuntersuchungen mit Muttersprachlern haben neben der Identifikation von zu schwierigen oder idiosynkratischen Texten den Vorteil, dass akzeptable Lösungsvarianten ermittelt werden können (vgl. Abschnitt 10).

Danach sollten zur Abschätzung der Schwierigkeit und Trennschärfe der einzelnen Texte sowie der Schwierigkeit und Reliabilität des Gesamttests Voruntersuchungen bei der Lernergruppe stattfinden, für die der Test bestimmt ist. Es ist zu beachten, dass die einzelnen Items (Lückenwörter) innerhalb eines Textes abhängig voneinander sind (sog. lokale stochastische Abhängigkeit) und dass deshalb bei der Berechnung der Reliabilität von den Summenwerten jedes Textes auszugehen ist. Zu schwierige oder zu leichte Texte sowie Texte mit einer zu geringen Trennschärfe und negativem oder nur geringem Beitrag zur Reliabilität werden ausgesondert (siehe die detaillierten Ausführungen in Grotjahn, 1987a; 1992 sowie die Auflistung der Schritte und Formeln bei Raatz & Klein-Braley, 1985). Die entsprechenden Analysen lassen sich leicht z.B. mit den Programm-Paketen SPSS für Windows oder ITAMIS-PC durchführen (vgl. Abschnitt 11).

Komplexer ist eine Überprüfung der Dimensionalität der eingesetzten C-Test-Texte mit Hilfe des klassischen latent-additiven Testmodells von Moosbrugger & Müller (1982) oder mit Hilfe probabilistischer Modelle für Ratingskalen (vgl. z.B. Müller, 1999; Arras, Eckes & Grotjahn, 2002). Der Praktiker sollte hier auf jeden Fall die Kooperation mit einem Spezialisten suchen. Ziel entsprechender Analysen ist, Texte zu identifizieren, die nicht in der gleichen Weise wie die übrigen Texte die zu messende Eigenschaft erfassen. Hinweise zu einem relativ einfachen Verfahren der Dimensionalitätsprüfung von C-Tests mit Hilfe des klassischen latent-additiven Testmodells auf der Basis von Standardprozeduren in SPSS finden sich in Grotjahn (1987a, 1992) und Raatz (1985).

10. Akzeptable Varianten und Orthographiefehler

Im Gegensatz zum Cloze Test sind bei C-Tests akzeptable Varianten, d.h. Lösungen, die zwar nicht dem Original entsprechen, jedoch bezogen auf den Gesamttext semantisch und grammatisch korrekt sind, zumeist sehr selten. Zudem finden sie sich in erster Linie bei weiter fortgeschrittenen Lernern. Wertet man eine größere Zahl von akzeptablen Varianten als „korrekt“, erhöht sich in der Regel die Reliabilität des entsprechenden Tests geringfügig, ohne dass sich jedoch die Punktwerte der Lerner wesentlich ändern.

Entscheidet man sich für die Berücksichtigung akzeptabler Varianten, sollte man möglichst mit vorher festgelegten Listen akzeptabler Lösungen arbeiten. Will man akzeptable Varianten insgesamt vermeiden, so kann man, um Eindeutigkeit zu erzielen, bei den betreffenden Wörtern weniger Buchstaben tilgen, wodurch allerdings die Schwierigkeit des Tests potenziell abnimmt.

Bei der Beurteilung von Orthographiefehlern stellt sich das Problem, dass die Kategorie „Orthographiefehler“ häufig mit einer erheblichen Unschärfe behaftet ist und dass als Folge beträchtliche Inkonsistenzen bei der Beurteilung auftreten können. Zudem führt die Wertung von Orthographiefehlern als korrekt zu einer Reduzierung der Auswertungsökonomie. Aus diesen und weiteren Gründen ist es zumeist angezeigt, orthographische Abweichungen zur Erhöhung der Auswertungsobjektivität und -ökonomie als Fehler zu werten (vgl. Grotjahn, 1995). Dies gilt allerdings nur sehr eingeschränkt für den Fall nicht sehr weit fortgeschrittener Lerner mit einer hohen Zahl von Orthographiefehlern (vgl. Arras, Eckes & Grotjahn, 2002).

11. Software

In Koller & Zahn (1996) wird das am Sprachenzentrum der Universität Erlangen entwickeltes Softwarepaket „Erltest“ vorgestellt, das eine bequeme und flexible

Entwicklung, benutzerfreundliche Administration und schnelle statistische Auswertung von C-Tests und anderen Lückentestformaten erlaubt (weitere Informationen unter http://www.phil.uni-erlangen.de/erltest/).

Einen einfachen web-basierten C-Test auf der Basis von HTML und JavaScript stellt Röver (2002) vor.

Das von Germann (1996) für *Word für Windows* 6.0 entwickelte Programm zur automatischen Erstellung von C-Tests ist unter neueren Versionen von *Word* nicht mehr lauffähig.

Für die statistische Analyse von Testdaten existiert eine Vielzahl von speziellen Programmen, die jedoch z.T. wenig benutzerfreundlich sind. An jedem deutschen universitären Rechenzentrum frei oder gegen eine geringe Lizenzgebühr zugänglich ist das "Statistical Package for the Social Sciences" (SPSS), das eine komfortable Berechnung der Itemschwierigkeiten, Trennschärfen und Reliabilität erlaubt (Prozedur „Analysieren → Skalieren → Reliabilitätsanalyse" in SPSS 10.0 für Windows). Ein ausführliches Beispiel wird in Diehl & Staufenbiel (2001, S. 539ff.) besprochen.

Dem Buch von Diehl & Staufenbiel (2001) liegt auch eine CD mit dem leistungsfähigen Programm ITAMIS-PC zur Item- und Skalenanalyse bei. Hinweise zum Leistungsumfang und zur Installation finden sich bei Diehl & Staufenbiel (2001, S. 701f.).

12. Interpretation der Ergebnisse aus C-Tests

Die Ergebnisse aus (kanonischen), nicht-diskursspezifischen C-Tests sind als Maß allgemeiner (fremd)sprachlicher Kompetenz zu interpretieren – und nicht etwa als Maß der Lesekompetenz (zur Begründung vgl. Grotjahn, 1987b; Grotjahn & Tönshoff, 1992).

Weiterhin werden C-Tests üblicherweise als **normorientierte** Messverfahren eingesetzt. Bei normorientierten Tests werden die individuellen Ergebnisse relativ zu den Ergebnissen einer Referenzgruppe, wie z.B. den Mitlernern in einem Sprachkurs, interpretiert. Man spricht deshalb auch von **bezugsgruppenorientierten** Tests. Die Ergebnisse werden häufig z.B. folgendermaßen formuliert: Der Kandidat gehört zu den oberen 10% der Gruppe. Ziel ist, die Kandidaten in eine Rangordnung zu bringen und möglichst verlässlich zwischen ihnen zu differenzieren.

Mit Hilfe von **kriteriumsorientierten** Tests will man dagegen ermitteln, ob und eventuell auch in welchem Ausmaß ein Lernender ein im Detail beschriebenes Kriterium, wie z.B. Kommunikationsfähigkeit im akademischen Kontext oder auch die in einem Kurs über einen bestimmten Zeitraum vermittelte

Grammatik, erreicht hat (vgl. Ingenkamp, 1997; S. 117-130; Klauer, 1987; Lynch & Davidson, 1997).

Am unproblematischsten ist eine solche kriteriumsorientierte Leistungsfeststellung, wenn die Prüfung die geforderten Leistungsmerkmale möglichst weitgehend widerspiegelt (vgl. Grotjahn, 2000; McNamara, 2000). Aus diesem Grund ist auch eine Interpretation von C-Test-Ergebnissen z.B. im Hinblick auf die Europarats-Stufen (vgl. Council of Europe, 2001; Europarat, 2001) oder die UNICERT-Stufen (vgl. Eggensperger & Fischer, 1998) ohne gründliche Validierungsuntersuchungen nicht möglich.

Literaturverzeichnis

Arras, Ulrike, Eckes, Thomas & Grotjahn, Rüdiger. (2002). C-Tests im Rahmen des „Test Deutsch als Fremdsprache“ (TestDaF): Erste Forschungsergebnisse. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4). Bochum: AKS-Verlag.

Connelly, Michael. (1997). Using C-Tests in English with post-graduate students. *English for Specific Purposes, 16*(2), 139-150.

Council of Europe. (2001). *A Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.
[auch abrufbar unter http://www.goethe.de/z/50/commeuro/]

Daller, Helmut. (1999). *Migration und Mehrsprachigkeit: Der Sprachstand türkischer Rückkehrer aus Deutschland*. Frankfurt am Main: Lang.

Daller, Helmut & Grotjahn, Rüdiger. (1999). The language proficiency of Turkish returnees from Germany: An empirical investigation of academic and everyday language proficiency. *Language, Culture and Curriculum, 12*(2), 156-172.

Diehl, Joerg M. & Staufenbiel, Thomas. (2001). *Statistik mit SPSS Version 10.0*. Eschborn: Klotz.

Eggensperger, Karl-Heinz & Fischer, Johann. (Hrsg.). (1998). *Handbuch UNICERT*. Bochum: AKS-Verlag.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin: Langenscheidt.
[ausführliche Information: http://www.goethe.de/z/50/commeuro/]

Germann, Ulrich. (1996). C-Tests automatisch erstellen – mit Word für Windows 6.0. In Grotjahn Rüdiger (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 279-304). Bochum: Brockmeyer.

Grotjahn, Rüdiger. (1987a). How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses. In Rüdiger Grotjahn, Christine Klein-Braley & Douglas K. Stevenson (Hrsg.), *Taking their measure: The validity and validation of language* tests (S. 219-253). Bochum: Brockmeyer.

Grotjahn, Rüdiger. (1987b). Ist der C-Test ein Lesetest? In Anthony Addison & Klaus Vogel (Hrsg.), *Lehren und Lernen von Fremdsprachen im Studium* (S. 230-248). Bochum: AKS-Verlag.

Grotjahn, Rüdiger. (1992). Der C-Test im Französischen. Quantitative Analysen. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 1, S. 205-255). Bochum: Brockmeyer.

Grotjahn, Rüdiger. (1995). Der C-Test: State of the Art. *Zeitschrift für Fremdsprachenforschung, 6*(2), 37-60.

Grotjahn, Rüdiger. (1996). The C-Test bibliography: version December 1995. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 435-457). Bochum: Brockmeyer. [auch abrufbar unter http://www.C-Test.de/]

Grotjahn, Rüdiger. (1997). Der C-Test: Neuere Entwicklungen. In Monica Gardenghi & Mary O'Connell (Hrsg.), *Prüfen, Testen, Bewerten im modernen Fremdsprachenunterricht* (S. 117-128). Frankfurt am Main: Lang.

Grotjahn, Rüdiger. (2000). Testtheorie: Grundzüge und Anwendungen in der Praxis. In Armin Wolff & Harald Tanzer (Hrsg.), *Sprache – Kultur – Politik: Beiträge der 27. Jahrestagung Deutsch als Fremdsprache vom 3.-5. Juni 1999 an der Universität Regensburg* (S. 304-341). Regensburg: Fachverband Deutsch als Fremdsprache.

Grotjahn, Rüdiger. (2002).'Scrambled' C-Tests: Eine Folgeuntersuchung. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4). Bochum: AKS-Verlag.

Grotjahn, Rüdiger, Klein-Braley, Christine & Raatz, Ulrich. (1992). C-Tests in der praktischen Anwendung. Erfahrungen beim Bundeswettbewerb Fremdsprachen. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 1, S. 263-296). Bochum: Brockmeyer.

Grotjahn, Rüdiger, Klein-Braley, Christine & Raatz, Ulrich. (2002). C-Tests: an overview. In James A. Coleman, Rüdiger Grotjahn & Ulrich Raatz (Hrsg.), *University language testing and the C-Test* (S. 93-114). Bochum: AKS-Verlag.

Grotjahn, Rüdiger & Stemmer, Brigitte. (2002). C-Tests and language processing. In James A. Coleman, Rüdiger Grotjahn & Ulrich Raatz (Hrsg.), *University language testing and the C-Test* (S. 115-130). Bochum: AKS-Verlag.

Grotjahn, Rüdiger & Tönshoff, Wolfgang. (1992). Textverständnis bei der C-Test-Bearbeitung. Pilotstudien mit Französisch- und Italienischlernern. In Grotjahn Rüdiger (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 1, S. 19-95). Bochum: Brockmeyer.

Hastings, Ashley J. (2002). Error analysis of an English C-Test: Evidence for integrated processing. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4). Bochum: AKS-Verlag.

Ingenkamp, Karlheinz. (1997). *Lehrbuch der Pädagogischen Diagnostik* (Studienausgabe, 4. Aufl.). Weinheim & Basel: Beltz.

Jafarpur, Abdoljavad. (1999). Can the C-Test be improved with classical item analysis? *System, 27*(1), 79-89.

Kamimoto, Tadamitsu. (1993, November). Tailoring the test to fit the students: improvement of the C-Test through classical item analysis. *Language Laboratory, 30*, 47-61.

Klauer, Karl J. (1987). *Kriteriumsorientierte Tests.* Göttingen: Hogrefe.

Klein-Braley, Christine. (1994). *Language testing with the C-Test: A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty.* Habilitationsschrift Universität Duisburg.

Klein-Braley, Christine. (1996). Towards a theory of C-Test processing. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 343-366). Bochum: Brockmeyer.

Klein-Braley, Christine. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing, 14*(1), 47-84.

Klein-Braley, Christine & Grotjahn, Rüdiger. (1998). C-Tests in der Schule? *Praxis des neusprachlichen Unterrichts, 45*(4), 411-417.

Köberl, Johann & Sigott, Günther. (1994). Adjusting C-Test difficulty in German. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 2, S. 179-192). Bochum: Brockmeyer.

Koller, Gerhard & Zahn, Rosemary. (1996). Computer based construction and evaluation of C-Tests. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 401-418). Bochum: Brockmeyer.

Lynch, Brian K. & Davidson, Fred. (1997). Criterion referenced testing. In Caroline Clapham & David Corson (Hrsg.), *Encyclopedia of language and education. Vol. 7: Language testing and assessment* (S. 263-273). Dordrecht: Kluwer.

McNamara, Tim F. (2000). *Language testing*. Oxford: Oxford University Press.

Moosbrugger, Helfried & Müller, Hans. (1982). A classical latent additive test model (CLA model). *The German Journal of Psychology, 6*, 145-149.

Müller, Hans. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen: Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items*. Bern: Huber.

Raatz, Ulrich. (1985). Better theory for better tests? *Language Testing, 2*, 60-75.

Raatz, Ulrich & Klein-Braley, Christine. (1985). How to develop a C-Test. *Fremdsprachen und Hochschule, 13/14*, 20-22.

Raatz, Ulrich & Klein-Braley, Christine. (2002). Introduction to language testing and to C-Tests. In James A. Coleman, Rüdiger Grotjahn & Ulrich Raatz (Hrsg.), *University language testing and the C-Test* (S. 75-91). Bochum: AKS-Verlag.

Röver, Carsten. (2002). Web-based C-Tests. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4). Bochum: AKS-Verlag.

Sigott, Günther. (2002). High-level processes in C-Test taking? In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 4). Bochum: AKS-Verlag.

Sigott, Günther & Köberl, Johann. (1996). Deletion patterns and C-Test difficulty across languages. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 159-172). Bochum: Brockmeyer.

Stemmer, Brigitte. (1991). *What's on a C-Test taker's mind: Mental processes in C-Test taking*. Bochum: Brockmeyer.

Anhang: Leitfaden zur Konstruktion und zum Einsatz von C-Tests

Schritt	**Hinweis**
1. Textauswahl	**inhaltlich geschlossene, kohärente und kohäsive Abschnitte aus mehreren längeren Texten** **Textmerkmale** (kanonischer C-Test): authentische Texte; keine Fiktion, kein verbaler Humor, keine direkte Rede, kein fach- und kulturspezifischer Inhalt; unmarkierte Syntax und Lexik; angemessener Schwierigkeitsgrad; Berücksichtigung der Lerngeschichte der Gruppe … **Bearbeitung der Texte**: so wenig wie möglich **Verwendung der Vorhersageformel** zur Bestimmung der Textschwierigkeit (Englisch) **Zahl der Texte:** Auswahl von ca. doppelt so vielen Texten, wie der C-Test enthalten soll
2. Testkonstruktion	**Testlänge**: 3 – 8 Texte in Abhängigkeit von der gewünschten Messgenauigkeit; zumeist: 5 Texte à 20 Lücken oder 4 Texte à 25 Lücken (max. Punktwert = 100) **Löschungen**: kanonisches Prinzip (2. Hälfte jedes 2. Wortes eines Textes – beginnend mit dem 2. Wort des 2. Satzes) **Zahl der Lücken im Text**: mindestens 20; gleiche Zahl von Lücken pro Text **Alternativen bei der Markierung der Lücken**: a) gleich lange, durchgehende Striche für jede Lücke; b) Strichlänge bei jeder Lücke in Abhängigkeit von der Zahl der getilgten Buchstaben; c) gestrichelte Linie mit einem Strich pro getilgtem Buchstaben. Die Varianten b) und c) machen den C-Test zwar zumeist leichter und reliabler, können jedoch einen negativen Effekt auf die Validität haben (Induzierung von Buchstabenzählen). **Sprachspezifik**: sprachspezifische Löschungsregeln (z.B. bei Polygraphen und Enklitika im Spanischen, Italienischen oder Portugiesischen, Komposita im Deutschen oder Italienischen)

	Anordnung der Texte im Test: nach aufsteigender Schwierigkeit **Berücksichtigung von potenziellen Einflussfaktoren**: Testerfahrung; Vertrautheit mit dem C-Test-Format; C-Test-Training; kulturelle Differenzen im Antwortverhalten; Testangst; Neigung zum Raten („hohe vs. niedrige Risikobereitschaft“); Feldabhängigkeit vs. Feldunabhängigkeit … **Variationen des C-Test-Formats**: deutliche Abweichungen vom kanonischen C-Test-Prinzip, wie z.B. Löschung der 1. Hälfte jedes zweiten Wortes, sind vor dem Einsatz in der Praxis empirisch zu überprüfen.
3. Testinstruktion	**Klarheit und Einfachheit der Instruktion**: z.B. „In den folgenden Texten fehlt bei einer Reihe von Wörtern ein Teil. Ergänzen Sie den fehlenden Teil in sinnvoller Weise.“ Verwendung von weitgehend sprachfreien Instruktionen bei multilingualen Adressaten; C-Test-Beispiel zur Übung … **Timing**: Angabe des Zeitlimits (z.B. „Sie haben 5 Minuten Zeit zur Bearbeitung jedes Texts“) **Angabe sprachspezifischer Konventionen**: z.B. im Französischen: „Wörter mit Bindestrich, wie *celui-ci*, zählen als *ein* Wort.“ **Ermutigung**: Hinweis, dass von den Lernern kein perfektes Ergebnis erwartet wird; Aufforderung an die Lerner, auch dann Lücken aufzufüllen, wenn sie sich nicht ganz sicher ist, dass die Lösung korrekt ist. …
4. Testadministration	**Überwachung**: Sicher stellen, dass alle Kandidaten verstehen, was sie tun sollen; Gewährleistung angemessener und vergleichbarer Bedingungen für alle Kandidaten … **Timing**: Einhaltung der Zeitvorgaben durch Nennung einer Restzeit bei jedem Text (z.B.: „Sie haben noch 2 Minuten Zeit zur Bearbeitung des Textes!“; „Sie sollten jetzt mit dem nächsten Text beginnen!“)

5. Testauswertung	**Kanonische Regel**: Jede dem Original entsprechende Lösung erhält einen Punkt. Die Summe der korrekten Lösungen eines Textes ergibt den Punktwert für den Text (= Item) und die Summe der Punktwerte der Einzeltexte ergibt den Gesamtpunktwert für einen C-Test. **Orthographie**: Falls gewünscht (z.B. im Fall von Anfängern), können Lösungen, die lediglich orthographisch falsch sind, als korrekt gewertet werden. **Alternative Lösungen**: Morpho-syntaktisch und semantisch angemessene Alternativen können als korrekt gewertet werden – und zwar, soweit möglich, anhand von vorher erstellten Lösungslisten.
6. Testanalyse	**Lokale stochastische Unabhängigkeit**: Die zu rekonstruierenden Elemente innerhalb eines Textes sind nicht unabhängig voneinander (die korrekte oder inkorrekte Rekonstruktion eines Wortes kann die Wahrscheinlichkeit, andere Wörter desselben Textes korrekt zu rekonstruieren, erhöhen oder erniedrigen). Aus diesem Grund sind alle statistischen Analysen auf der Basis der Summenwerte für die Texte durchzuführen, d.h. jeder Text gilt als Item (nicht die Lücken!) **Schwierigkeit**: P_{Item} = Prozentsatz korrekter Lösungen in einem Text (Item). Texte, mit einem unangemessenen Schwierigkeitsgrad sind auszuscheiden (zumeist: Lösungsrate von weniger als 20% oder mehr als 80%). **Itemtrennschärfe**: Items (d.h. Texte), die nicht hinreichend zwischen guten und schlechten Kandidaten differenzieren, sind auszuscheiden. **Eindimensionalität**: Die Eindimensionalität ist z.B. mit Hilfe des Klassischen Latent-Additiven Testmodells oder mit Hilfe probabilistischer Modelle für ordinale Itemanworten sicher zu stellen (Aussonderung von Texten, die das Kriterium der Eindimensionalität verletzen und evtl. auch Ausschluss von Kandidaten mit ‚extremen' Werten).

	Reliabilität: Berechnung der Reliabilität z.B. mit Hilfe von Cronbachs Alpha (die Verwendung von Alpha setzt Eindimensionalität und gleiche Zahl von Lücken pro Text voraus). Texte, die nicht hinreichend zur Reliabilität beitragen, sind auszusondern.
7. Interpretation	**Allgemeine Sprachkompetenz**: Kanonische C-Tests mit nicht-fachsprachlicher Textbasis sind als Instrumente zur Messung allgemeiner (fremd)sprachlicher Kompetenz zu interpretieren (und nicht z.B. als Leseverstehenstests). **Normorientierung**: C-Tests werden üblicherweise als normorientierte (d.h. bezugsgruppenorientierte) und nicht als kriteriumsorientierte Verfahren verwendet. **Rangordnung**: Die Kandidaten werden entsprechend ihren Testergebnissen in eine Rangordnung gebracht. Diese kann für Entscheidungen (z.B. Selektion, Einstufung) benutzt werden.