

Der C-Test: Aktuelle Entwicklungen¹

1. Einleitung

Der C-Test gehört zu den am gründlichsten untersuchten neueren Sprachtests (vgl. die C-Test-Bibliographie von Grotjahn 1996 sowie <http://www.c-test.de>). Er gilt als ein ökonomisches und reliables Verfahren zur globalen Feststellung der allgemeinen Kompetenz in Fremd-, Zweit- und Erstsprachen. C-Tests sind insbesondere dann erfolgreich einsetzbar, wenn das Ziel eine vom vorangehenden Unterricht bzw. von der individuellen Lerngeschichte weitgehend unabhängige globale Sprachstandsfeststellung ist. C-Tests sind mittlerweile für mehr als 20 Sprachen entwickelt worden und in einer Vielzahl unterschiedlicher Kontexte eingesetzt worden (vgl. die Hinweise in Grotjahn 1995, 2002a; Coleman/Grotjahn/Raatz 2002).

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 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. Hat ein Wort eine ungerade Zahl von Buchstaben, werden entweder $(n+1)/2$ Buchstaben (der Test wird schwieriger) oder $(n-1)/2$ Buchstaben getilgt (der Text wird leichter). Wörter mit einem einzigen Buchstaben sowie zumeist auch Eigennamen bleiben unberücksichtigt. 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 (vgl. die detaillierteren Hinweise in Grotjahn 1995).

Bei der Markierung der Lücken existieren mehrere Möglichkeiten: a) gleich lange durchgehende Striche für jede Lücke (ursprüngliches Prinzip); b) Wahl der Strichlänge bei jeder Lücke in Übereinstimmung mit 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 reli-

¹ Der vorliegende Beitrag beruht in wesentlichen Teilen auf früheren Publikationen des Verfassers.

abler, sie können jedoch einen negativen Effekt auf die Validität haben, da sie zum Buchstabenzählen verleiten.

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 detaillierteren Hinweise in Grotjahn 1995 sowie Abschnitt 9). 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, d.h. Lückenwörtern, wobei häufig vier Texte mit 25 Items oder fünf Texte mit 20 Items gewählt werden. Dies hat den Vorteil, dass die Testergebnisse unmittelbar als Prozentzahlen interpretiert werden können.

Die Abbildung 1 zeigt ein Beispiel für einen deutschen C-Test-Text aus den Erprobungsuntersuchungen zum Test *Deutsch als Fremdsprache* (TestDaF) – dem deutschen Pendant zu internationalen (universitären) sprachlichen Zulassungstests wie TOEFL oder IELTS. In der Instruktion wird den Lernern lediglich mitgeteilt, dass in den Texten die Lücken zu ergänzen sind und dass fünf Minuten pro Text zur Verfügung stehen. Weitere Beispiele für C-Tests finden sich unter <http://www.c-test.de> oder <http://www.deutsch-als-fremdsprache.de/ctest/>

Uni-Café

Ganz in der Nähe der Universität gibt es ein gern besuchtes Café. Hier tref_____ sich d_____ Studierenden zwis_____ den Vorle_____, sitzen b_____ Tee, Kaf_____ und Bröt_____, lesen Zei_____, unterhalten si_____, bereiten Semi_____ vor. Man_____ sitzen all_____, um z_____ lesen, and_____ sitzen zusa_____, um si_____ zu unter_____ und u_____ zu disku_____. Jeden T_____ ist das Café geöffnet, auch spät am Abend kann man dort noch etwas trinken oder essen.

Abbildung 1: C-Test-Text aus Arras/Eckes/Grotjahn (2002: 208)

Es gibt eine Reihe von sprachspezifischen Problemen mit dem kanonischen Konstruktionsprinzip. Im Deutschen ergibt sich z.B. bei Wortzusammensetzungen das Problem, dass das kanonische Prinzip zur Tilgung ganzer Wörter führt, die dann nicht mehr oder nur noch mit großen Problemen eindeutig rekonstruierbar sind. Es wird deshalb zuweilen erst ab dem zweiten Buchstaben des letzten Wortes der Zusammensetzung getilgt.

Ein weiteres Problem können Graphemkombinationen darstellen, die einen Einzellaut repräsentieren, wie z.B. im Deutschen das <sch>. Tilgt man z.B. bei dem Wort *waschen* die zweite Hälfte, d.h. <chen>, dann entspricht die Aussprache des letzten nicht getilgten Buchstabens, d.h. des <s>, nicht der Aussprache der zu rekonstruierenden Graphemkombination <sch>. Da die Suche nach der Lösung häufig über innere Phonation oder 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.

3. Sprachenspezifische Varianten des kanonischen C-Test-Formats

In einer Reihe von Arbeiten sind sprachenspezifische Modifikationen des kanonischen C-Test-Formats und deren Einfluss insbesondere auf die Testschwierigkeit untersucht worden (vgl. die Belege in Grotjahn 1995). Entsprechende Studien sind in Bezug auf den praktischen Einsatz von C-Tests von erheblicher Bedeutung, da ein C-Test über eine Modifikation der Tilgungsregel 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. Die Varianten (a) und (b), d.h. die Löschung von zwei Dritteln jedes zweiten Wortes bzw. die Löschung jedes zweiten Wortes mit Ausnahme des ersten Buchstabens, führten aufgrund der Reduktion der Textredundanz zu signifikant schwierigeren C-Test-Texten. Als am reliabelsten im Deutschen, allerdings nicht im Englischen, erwies sich die Variante (c), d.h. die Löschung der **ersten** Hälfte jedes zweiten Wortes.

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. Man kann hier allerdings einwenden, dass ein itemanalytischer C-Test-Text aufgrund des resultierenden Tilgungsmusters vermutlich partiell etwas anderes misst als ein kanonischer C-Test-Text. Auch wäre der Entwicklungsaufwand höher als bei einem kanoni-

schen C-Test. Zudem kommt zumindest Jafarpur (1999) zu einer eher negativen Einschätzung des Potentials der klassischen Itemanalyse zur Verbesserung von C-Tests.

Eine andere Form einer Itemanalyse einzelner Lücken wird in Kontra/Kormos (2004) verwendet. Die Autorinnen administrieren einen herkömmlichen C-Test mit fünf Texten und 100 Lücken, berücksichtigen dann jedoch bei der Berechnung der Testergebnisse und Kennwerte nur insgesamt 25 Lücken mit einer Schwierigkeit zwischen 0.3 und 0.8 sowie einer Trennschärfe größer als 0.3. Nach Einschätzung der Autorinnen führt dieses Verfahren zu zufriedenstellenderen Testkennwerten. Das von Kontra und Kormos gewählte Verfahren hat zwar den Vorteil, dass es den Entwicklungsaufwand eines C-Tests nicht erhöht und beim Einsatz geeigneter Software auch den Auswertungsaufwand nicht vergrößert. Es sind jedoch m.E. u.a. folgende Probleme mit dem Verfahren verbunden: 1) Insbesondere bei kleineren Probandenstichproben ist die Selektion der Lücken aufgrund statistischer Kriterien mit einem relativ großen Messfehler behaftet. Als Folge können sich bei einer Replikation deutlich differierende Reliabilitäts- und Validitätskoeffizienten ergeben. 2) Die Inhaltsvalidität der itemanalysierten Form ist fraglich, da die selektierten Einzelitems wahrscheinlich keine repräsentative Stichprobe aus der jeweiligen Sprache darstellen. Bevor deshalb an einen Einsatz des Verfahrens in der Praxis gedacht werden sollte, bedarf es einer gründlichen Erprobung an einer Vielzahl von C-Tests und Probandengruppen.

4. 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 liegt 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). Weiterhin können C-Tests mit relativ geringem Aufwand als computer- oder auch web-basierte Tests implementiert werden (vgl. z.B. Koller/Zahn 1996; Röver 2002).

5. Validität

Die Ansicht, dass der C-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 introspektiver und experimenteller 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. Entsprechende Befunde finden sich z.B. in Grotjahn (2002b), Hastings (2002), Kontra/Kormos (2004) und Sigott (2002, 2004).

Ein weiterer wichtiger aktueller Befund stammt aus einer kürzlich abgeschlossenen Begleituntersuchung zum Test Deutsch als Fremdsprache – TestDaF (vgl. auch <http://www.testdaf.de>). Die Tabelle 1 zeigt die Korrelationen zwischen den vier Subtests des TestDaF und einem zur Verankerung der TestDaF-Teile „Leseverstehen“ und „Hörverstehen“ neu entwickelten C-Test für Deutsch als Fremdsprache (vgl. Arras/Eckes/Grotjahn 2002; Eckes/Grotjahn 2004).

Tabelle 1: Spearman-Rang-Korrelationen (obere Dreiecksmatrix) und Kendalls Tau-b (untere Dreiecksmatrix) zwischen TestDaF-Subtests und C-Test

	C-Test	LV	HV	SA	MA
C-Test		.647 <i>N</i> = 187	.636 <i>N</i> = 187	.682 <i>N</i> = 155	.640 <i>N</i> = 145
LV	.483 <i>N</i> = 187		.641 <i>N</i> = 187	.605 <i>N</i> = 155	.556 <i>N</i> = 145
HV	.473 <i>N</i> = 187	.487 <i>N</i> = 187		.600 <i>N</i> = 155	.625 <i>N</i> = 145
SA	.559 <i>N</i> = 155	.493 <i>N</i> = 155	.499 <i>N</i> = 155		.539 <i>N</i> = 113
MA	.521 <i>N</i> = 145	.450 <i>N</i> = 145	.522 <i>N</i> = 145	.490 <i>N</i> = 113	

Anmerkung: Alle Korrelationen sind auf dem Niveau von 0.01 (2-seitig) signifikant.

LV = Leseverstehen; HV = Hörverstehen; SA = Schriftlicher Ausdruck; MA = Mündlicher Ausdruck

(Quelle: Arras/Eckes/Grotjahn 2002: 200)

Der C-Test korreliert hoch mit allen vier TestDaF-Subtests. Auffallend ist vor allem die hohe Korrelation von 0.64 (Spearman) zwischen C-Test und MA. Da ein Testmethodeneffekt wegen der Unterschiedlichkeit der Aufgabenformate auszuschließen ist, scheinen C-Test und MA in nicht unbeträchtlichem Maße die gleichen zugrundeliegenden Fähigkeiten zu erfassen. Angesichts der Tatsache, dass in der C-Test-Literatur bisher kaum Befunde zum Zusammenhang zwischen C-Test-Leistung und mündlicher Kompetenz vorliegen, ist dies ein sehr wichtiges Resultat.

Am höchsten korreliert der C-Test mit den Ergebnissen aus dem Testteil „Schriftlicher Ausdruck“, in dem das Schreiben eines komplexen Textes anhand von Vorgaben gefordert ist. Auf diesen Zusammenhang werde ich noch genauer eingehen.

6. 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 au-

thentisch sein, eine Sinneinheit bilden und zudem möglichst zielgruppenadäquat sein. Um eine Benachteiligung einzelner Testteilnehmer möglichst weitgehend zu vermeiden, erfolgt die Auswahl 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* als Maß lexikalischer Redundanz, d.h. anhand des Quotienten aus der Häufigkeit von Worttypen und Wortformen, möglich ist, die empirische Schwierigkeit von englischen C-Test-Texten für bestimmte Lerngruppen regressionsanalytisch, d.h. statistisch, vorherzusagen.

Weiterhin ist untersucht worden, ob C-Tests zur Messung diskurspezifischer – 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 diskurspezifischer C-Tests.

7. Testinstruktion

Wichtig ist auch die Form der C-Test-Instruktion. 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 einige sprachspezifische Konventionen nötig, wie z.B. „Wörter mit Bindestrich, wie z.B. französisch *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!“; und 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.

8. Voruntersuchungen

Die C-Test-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

kompetenten 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.

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, d.h. Lückenwörter, innerhalb eines Textes voneinander abhängig sind 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 1987, 1992, 2002a sowie die Auflistung der Schritte und Formeln bei Raatz/Klein-Braley 1985). Die entsprechenden Analysen lassen sich leicht z.B. mit gängigen Statistik-Programmen wie SPSS für Windows durchführen.

Komplexer ist eine Überprüfung der Dimensionalität der eingesetzten C-Test-Texte z.B. mit Hilfe mit Hilfe probabilistischer Modelle für Ratingskalen. 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. Beispiele für eine C-Test-Skalierung mit Hilfe probabilistischer Modelle für Ratingskalen finden sich z.B. in Arras/Eckes/Grotjahn (2002), Eckes (2004) und Eckes/Grotjahn (2004).

9. 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).

10. Interpretation der Ergebnisse aus C-Tests

Bei der Interpretation von C-Test-Resultaten ist u.a. folgendes zu beachten: Die Ergebnisse aus kanonischen, nicht-diskursspezifischen C-Tests sind als Maß allgemeiner (fremd)sprachlicher Kompetenz zu interpretieren – und nicht etwa lediglich als Maß der Lesekompetenz (zur Begründung vgl. z.B. 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 und die Ergebnisse 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. Brown/Hudson 2002; Ingenkamp 1997: 117-130; Klauer 1987).

Am unproblematischsten ist eine solche kriteriumsorientierte Leistungsfeststellung, wenn die Prüfung die relevanten Merkmale der zielsprachlichen Verwendungssituation weitgehend widerspiegelt (vgl. Grotjahn 2000; McNamara 2000). Dies ist z.B. bei einem kommunikativen Performanztests wie dem ACTFL *Oral Proficiency Interview* weitgehend der Fall. Die Beziehung zwischen C-Tests und Konstrukten wie fremdsprachliche Lese- oder Schreibkompetenz ist dagegen weit indirekter. Eine kriterienbezogene Interpretation von C-Test-Ergebnissen auf der Basis eines Vergleichs der beim Lösen von C-Tests involvierten Fähigkeiten und Fertigkeiten und den für die Lösung be-

stimmter zielsprachlicher Probleme benötigten Fähigkeiten und Fertigkeiten ist deshalb sehr schwierig und auch nur eingeschränkt möglich.

Eine andere, theoretisch zwar weniger zufriedenstellende, jedoch praktikablere Möglichkeit einer kriterienbezogenen Interpretation von C-Test-Ergebnissen beruht auf der empirisch ermittelten Korrelation zwischen den C-Test-Ergebnissen und anderen Messwerten, die ihrerseits kriteriumsorientiert interpretiert werden können. Ich möchte dies anhand der Niveaustufenbeschreibungen des TestDaF und den immer wichtiger werdenden Kompetenzstufen des „Gemeinsamen europäischen Referenzrahmens für Sprachen“ des Europarats verdeutlichen.

Wie bereits erwähnt, korreliert der im Rahmen von TestDaF neu entwickelte C-Test am höchsten mit dem TestDaF-Subtest „Schriftlicher Ausdruck“. Beim Subtest „Schriftlicher Ausdruck“ handelt es sich um einen kriteriumsorientierten Test, bei dem die Schreibleistung der Kandidaten anhand der drei Hauptmerkmale „Gesamteindruck“, „Behandlung der Aufgabe“ und „Sprachliche Realisierung“ sowie einer Reihe von spezifischen Deskriptoren den TestDaF-Niveaustufen TDN 5, TDN 4, TDN 3 und unter TDN 3 zugeordnet wird. Die TestDaF-Niveaustufen für den Schriftlichen Ausdruck lauten folgendermaßen:

TDN 5 Kann sich in studienbezogenen Alltagssituationen (z.B. Bericht für Stipendiengeber) sowie im fächerübergreifenden wissenschaftlichen Kontext (z.B. Protokoll, Thesenpapier) zusammenhängend und strukturiert sowie der jeweiligen Textsorte angemessen äußern.

TDN 4 Kann sich in studienbezogenen Alltagssituationen (z.B. Bericht für Stipendiengeber) sowie im fächerübergreifenden wissenschaftlichen Kontext (z.B. Protokoll, Thesenpapier) weitgehend zusammenhängend und strukturiert sowie der jeweiligen Textsorte weitgehend angemessen äußern, allerdings mit sprachlichen Mängeln, die jedoch das Textverständnis nicht behindern.

TDN 3 Kann sich in studienbezogenen Alltagssituationen (z.B. Bericht für Stipendiengeber) nur eingeschränkt zusammenhängend und strukturiert schriftlich äußern; kann sich im fächerübergreifenden wissenschaftlichen Kontext (z.B. Protokoll, Thesenpapier) schriftlich nicht so äußern, dass das Textverständnis hinreichend gesichert ist.

abgerufen unter <http://www.testdaf.de> am 25.5.2003

Die Tabelle 2 zeigt den Zusammenhang zwischen den C-Test-Leistungen und den Leistungen im Schriftlichen Ausdruck.

Tabelle 2: Zusammenhang zwischen C-Test und Schriftlichem Ausdruck (weltweite TestDaF Erprobungsprüfung E004 Juni 2001)

C-Test	Schriftlicher Ausdruck				Gesamt
	unter TDN 3	TDN 3	TDN 4	TDN 5	
15	1				1
18	1	2			3
23	1				1
24	1				1
26		1			1
28		1			1
30	1				1
31	2	1			3
32		2			2
33		1			1
34	2	1			3
36		4			4
37	2	3			5
38		2	1		3
39		3	2		5
40		1	1		2
41	1	2	1		4
42		2	2		4
43		3			3
44		1	1		2
45		1			1
46		1	3	2	6
47		2			2
48		2	4	1	7
49		3			3
50	1	5			6
51		4	3		7
52			4	1	5
53		2	3		5
54		1	4	1	6
55			3		3

C-Test	Schriftlicher Ausdruck				Gesamt
	unter TDN 3	TDN 3	TDN 4	TDN 5	
56		1	2	2	5
57		3	5	1	9
58			3		3
59		1	1	2	4
60			2		2
61		1	4	1	6
62			2	2	4
63			2	1	3
64		1		2	3
65			2	1	3
66			1	1	2
67			2		2
69				3	3
70			1	2	3
71				1	1
73			1		1
Gesamt	13	58	60	24	155

Quelle: Arras/Eckes/Grotjahn (2002: 202f.)

Die Daten in Tabelle 2 weisen ein charakteristisches Muster auf: Niedrige C-Test-Werte gehen in der Tendenz mit einer Zuordnung zu einer niedrigen TDN und hohe C-Test-Werte mit einer Zuordnung zu einer hohen TDN einher. So findet sich z.B. unter den 27 Probanden mit einem C-Test-Wert kleiner oder gleich 37 – und damit bei einem substantziellen Anteil von 25.5% – kein einziges Mal das Urteil TDN 4 oder TDN 5. Eine genauere Zuordnung der C-Test-Ergebnisse zu den TestDaF-Niveaustufen ist allerdings problematisch und mit einem relativ hohen Mess- und Klassifikationsfehler behaftet. Dies gilt insbesondere in Bezug auf die Stufen TDN 4 und TDN 5. Ob diese Unschärfe als gravierender Mangel anzusehen ist, hängt jedoch von den Entscheidungen ab, die mit Hilfe der C-Test-Ergebnisse getroffen werden sollen.

Noch problematischer ist es, wenn man auf der Basis der Korrelation zwischen C-Test und TestDaF eine Einordnung der C-Test-Ergebnisse auf den Europaratsstufen vornehmen will. Die Abbildung 2 zeigt den Zusammenhang zwischen den TestDaF-Niveaustufen und den Referenzstufen des Europarats. Danach entspricht der von TestDaF angezielte Kompetenzbereich in etwa den Europaratsstufen “Lower Vantage” bis “Higher Effective Operational Proficiency”.

Kompetenzstufen

A Elementare Sprachverwendung		B Selbstständige Sprachverwendung				C Kompetente Sprachverwendung	
A1 Breakthrough	A2 Waystage	B1 Threshold		B2 Vantage		C1 Effective Operational Proficiency	C2 Mastery
		B1.1	B1.2	B2.1	B2.2	C1.1	C1.2

TestDaF:

TDN 3	TDN 4	TDN 5
-------	-------	-------

Abbildung 2: Zusammenhang zwischen den Niveaustufen des TestDaF und den Kompetenzstufen des Gemeinsamen europäischen Referenzrahmens für Sprachen (Europarat 2001: 33ff.; <http://www.testdaf.de>)

Bei der Zuordnung der TestDaF-Niveaustufen zu den Europaratstufen stellt sich nämlich das Problem, dass die dargestellten Beziehungen zwischen den TestDaF-Stufen und den Europaratsstufen in erster Linie die Urteile von Experten widerspiegeln; umfangreiche Validierungsuntersuchungen fehlen bisher. Dies schränkt die Möglichkeit der Zuordnung der C-Test-Ergebnisse zu den Europaratsstufen mit Hilfe des TestDaF weiter ein. Man sollte deshalb m.E. darauf verzichten, die Zuordnung von C-Test-Ergebnissen zu den Europaratsstufen über Tests vorzunehmen, die selbst nicht hinreichend präzise auf den Europaratsskalen verortet sind.

Wünscht man sich trotz der aufgezeigten Probleme z.B. im unterrichtlichen Kontext eine Verortung von C-Test-Ergebnissen auf den Skalen des Europarats, dann kann man z.B. folgendermaßen vorgehen: Der Lehrende stuft selbst die Schülerleistungen auf der Basis einer genügend umfangreichen und komplexen Sprachstichprobe z.B. auf der Globalskala des Europarats ein und setzt dann die C-Test-Ergebnisse zu den Einstufungen in Beziehung. Sind sowohl die Einstufungen der Schülerleistungen auf den Europaratsskalen als auch die C-Test-Ergebnisse hinlänglich reliabel und korrelieren beide zudem hoch miteinander, dann kann der entsprechende C-Test z.B. für eine schnelle, approximative Verortung einer vergleichbaren Lernergruppe auf der Globalskala des Europarats benutzt werden. Das geschilderte Verfahren dürfte allerdings für eine Zuordnung **einzelner** Lerner zu den Europaratsstufen nur sehr eingeschränkt verlässlich sein.

11. Literatur

- Arras, Ulrike/Eckes, Thomas/Grotjahn, Rüdiger: "C-Tests im Rahmen des ‚Test Deutsch als Fremdsprache‘ (TestDaF): Erste Forschungsergebnisse." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag 2002, 175-209.
- Brown, James D./Hudson, Thom: *Criterion-referenced language testing*. Cambridge: Cambridge University Press 2002.
- Coleman, James A./Grotjahn, Rüdiger/Raatz, Ulrich (Hrsg.): *University language testing and the C-Test*. Bochum: AKS-Verlag 2002.
- Connelly, Michael: "Using C-tests in English with post-graduate students." In: *English for Specific Purposes* 16, 2 (1997), 139-150.
- Daller, Helmut: *Migration und Mehrsprachigkeit: Der Sprachstand türkischer Rückkehrer aus Deutschland*. Frankfurt am Main: Lang 1999.
- Daller, Helmut/Grotjahn, Rüdiger: "The language proficiency of Turkish returnees from Germany: An empirical investigation of academic and everyday language proficiency." In: *Language, Culture and Curriculum* 12, 2 (1999), 156-172.
- Eckes, Thomas: "Rasch-Modelle zur C-Test-Skalierung." In: Grotjahn, Rüdiger (Hrsg.): *The C-Test: Theorie, empirical research, applications*. Frankfurt am Main: Lang 2004 (erscheint).
- Eckes, Thomas/Grotjahn, Rüdiger: "Der C-Test als Ankertest für TestDaF: Analysen auf der Basis eines probabilistischen Testmodells." In: Grotjahn, Rüdiger (Hrsg.): *The C-Test: Theorie, empirical research, applications*. Frankfurt am Main: Lang 2004 (erscheint).
- Europarat: *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin: Langenscheidt 2001.
- Grotjahn, Rüdiger: "How to construct and evaluate a C-Test: A discussion of some problems and some statistical analyses." In: Grotjahn, Rüdiger/Klein-Braleay, Christine/Stevenson, Douglas K. (Hrsg.): *Taking their measure: The validity and validation of language tests*. Bochum: Brockmeyer 1987, 219-253.
- Grotjahn, Rüdiger: "Der C-Test im Französischen. Quantitative Analysen." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bd. 1. Bochum: Brockmeyer 1992, 205-255.
- Grotjahn, Rüdiger: "Der C-Test: State of the Art." In: *Zeitschrift für Fremdsprachenforschung* 6, 2 (1995), 37-60.
- Grotjahn, Rüdiger: "The C-Test bibliography: version December 1995." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bd. 3. Bochum: Brockmeyer 1996, 435-457. [auch abrufbar unter <http://www.c-test.de>].
- Grotjahn, Rüdiger: "Testtheorie: Grundzüge und Anwendungen in der Praxis." In: Wolff, Armin/Tanzer, Harald (Hrsg.): *Sprache – Kultur – Politik: Beiträge der 27. Jahrestagung Deutsch als Fremdsprache vom 3.-5. Juni 1999 an der Universität Regensburg*. Regensburg: Fachverband Deutsch als Fremdsprache 2000, 304-341.
- Grotjahn, Rüdiger: "Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag 2002a, 211-225. [auch abrufbar unter <http://homepage.ruhr-uni-bochum.de/Ruediger.Grotjahn/>].

- Grotjahn, Rüdiger: "'Scrambled' C-Tests: Eine Folgeuntersuchung." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag 2002b, 83-121.
- Grotjahn, Rüdiger/Klein-Braley, Christine/Raatz, Ulrich: "C-Tests: An overview." In: Coleman, James A./Grotjahn, Rüdiger/Raatz, Ulrich (Hrsg.): *University language testing and the C-Test*. Bochum: AKS-Verlag 2002, 93-114.
- Grotjahn, Rüdiger/Tönshoff, Wolfgang: "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. Bochum: Brockmeyer 1992, 19-95.
- Hastings, Ashley J.: "Error analysis of an English C-Test: Evidence for integrated processing." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag 2002, 53-66.
- Ingenkamp, Karlheinz: *Lehrbuch der Pädagogischen Diagnostik: Studienausgabe* (4., neu ausgestattete Aufl.; 1. Aufl. 1985). Weinheim: Beltz 1997.
- Jafarpur, Abdoljavad: "Can the C-test be improved with classical item analysis?" In: *System* 27, 1 (1999), 79-89.
- Kamimoto, Tadimitsu: "Tailoring the test to fit the students: improvement of the C-Test through classical item analysis." In: *Language Laboratory* 30 (1993), 47-61.
- Klauer, Karl J.: *Kriteriumsorientierte Tests*. Göttingen: Hogrefe 1987.
- Klein-Braley, Christine: *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*. Unveröffentl. Habilitationsschrift, Universität Duisburg 1994.
- Klein-Braley, Christine: "C-Tests in the context of reduced redundancy testing: An appraisal." In: *Language Testing* 14, 1 (1997), 47-84.
- Köberl, Johann/Sigott, Günther: "Adjusting C-test difficulty in German." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bd. 2. Bochum: Brockmeyer 1994, 179-192.
- Koller, Gerhard/Zahn, Rosemary: "Computer based construction and evaluation of C-tests." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Bd. 3. Bochum: Brockmeyer 1996, 401-418.
- Kontra, Edit H./Kormos, Judit: "Strategy use and the construct of C-tests." In: Grotjahn, Rüdiger (Hrsg.): *The C-Test: Theorie, empirical research, applications*. Frankfurt am Main: Lang 2004 (erscheint).
- McNamara, Tim F.: *Language testing*. Oxford: Oxford University Press 2000.
- Raatz, Ulrich/Klein-Braley, Christine: "How to develop a C-Test." In: Klein-Braley, Christine/Raatz, Ulrich (Hrsg.): *Fremdsprachen und Hochschule, 13/14: Thematischer Teil: C-Tests in der Praxis*. Bochum: AKS-Verlag 1985, 20-22.
- Röver, Carsten: "Web-based C-tests." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag 2002, 123-130.
- Sigott, Günther: "High-level processes in C-Test taking?" In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag 2002, 67-82.
- Sigott, Günther: *Towards identifying the C-Test construct*. Frankfurt am Main: Lang 2004.

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

Materialien Deutsch als Fremdsprache

Heft 73

Integration durch Sprache

Herausgegeben

von

Armin Wolff; Torsten Ostermann; Christoph Chlosta

Fachverband Deutsch als Fremdsprache

Regensburg 2004

Materialien Deutsch als Fremdsprache

- Herausgeber:** Fachverband Deutsch als Fremdsprache (FaDaF)
- Redaktion:** Dr. Armin Wolff, Universität Regensburg
Reinhold Zollner, Universität Regensburg
- Heft 73** Armin Wolff; Torsten Ostermann; Christoph Chlosta
Integration durch Sprache
Beiträge der 31. Jahrestagung DaF 2003
- 1. Auflage** Regensburg 2004
- ISBN** 3-88246-265-5
- Redaktionsadresse
und Vertrieb:** Dr. Armin Wolff
Universitätsstr. 31
D-93053 Regensburg
Tel. 0941/943-2426/25
Tel. 0941/943-3008
Fax 0941/943-2410
e-mail: armin.wolff@sprachlit.uni-regensburg.de
- Preis incl. Versand:** € 24.00
(für FaDaF-Mitglieder € 20.00)
- Druck und Einband:** Dr. Joachim Kuns
Druck + Verlag
Charlottenstr. 14
52070 Aachen