Grotjahn, Rüdiger. (Hrsg.). (2002). Der C-Test. Theoretische Grundlagen und praktische Anwendungen (Bd. 4, S. 175-209). Bochum: AKS-Verlag

C-Tests im Rahmen des „Test Deutsch als Fremdsprache“ (TestDaF): Erste Forschungsergebnisse

Ulrike Arras, Thomas Eckes & Rüdiger Grotjahn*

After a brief description of TestDaF (Test Deutsch als Fremdsprache – Test of German as a Foreign Language) an overview of the studies carried out in the present article is given. Aspects dealt with include the use of C-Tests for calibration of items, the selection of texts for a German C-Test, necessary modifications of the classical deletion principle, and problems encountered with the detailed scoring system used. In the next section some results of the statistical analyses are presented with a focus on how to deal with spelling errors. On the basis of a detailed analysis of data from two groups of learners of German ($N = 93$ and $N = 187$) it is concluded that with more advanced learners and carefully piloted C-Tests, it makes only a relatively small difference whether spelling errors are counted as correct or incorrect and that therefore for the sake of economy and reliability spelling errors should rather be counted as incorrect. Subsequently it is shown that a C-Test – consisting of four texts each with 20 gaps and developed for the calibration of TestDaF's reading and listening comprehension items – was highly reliable (Cronbach's alpha = .84) and could be successfully scaled on the basis of Müller's (1999) Continuous Rating Scale Model. Furthermore, the C-Test correlated substantially with the reading, listening, writing and speaking parts of TestDaF (Spearman's $r > .64$). Particularly important in this context is the high correlation between the C-Test and oral proficiency measured by a Simulated Oral Proficiency Interview (SOPI) – a new result in the C-Test literature. In the last part of the article it is argued that the C-Test can also be used as a very efficient and, compared to TestDaF's writing subtest, much more economical means for screening subjects with low writing ability.

1. Der Test Deutsch als Fremdsprache

Wer in Deutschland studieren möchte, muss ausreichende Sprachkenntnisse im Deutschen nachweisen, um zu einem Fachstudium zugelassen zu werden. Dieser Nachweis kann durch die neue standardisierte Prüfung „Test Deutsch als Fremdsprache“ (TestDaF) erbracht werden. TestDaF wird weltweit zu festgelegten Terminen durchgeführt und zentral im TestDaF-Institut in Hagen erstellt und

* Ulrike Arras, TestDaF-Institut, Elberfelder Str. 103, D-58084 Hagen. E-mail: ulrike.arras@testdaf.de.

PD Dr. Thomas Eckes, TestDaF-Institut, Elberfelder Str. 103, D-58084 Hagen. E-mail: thomas.eckes@testdaf.de.

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

Wir danken Herrn Prof. Dr. Ulrich Raatz für die kritische Lektüre einer ersten Fassung des vorliegenden Artikels.

ausgewertet. Es handelt sich um eine Prüfung auf fortgeschrittenem Niveau, die sprachliche Fähigkeiten und Fertigkeiten in hochschulbezogenen Kontexten erfasst. Die vier Bereiche „Leseverstehen“, „Hörverstehen“, „Schriftlicher Ausdruck“ und „Mündlicher Ausdruck“ werden getrennt geprüft, und es wird ein differenziertes sprachliches Leistungsprofil der Studienbewerber/innen ausgewiesen. Die Charakterisierung der Leistungen der Prüfungsteilnehmer/innen erfolgt dabei nicht über Punktwerte, sondern auf der Basis der TestDaF-Niveaustufen (TDN) 3 bis 5, die in etwa den Stufen 3 (*Independent User*) bis 5 (*Good User*) der *Association of Language Testers in Europe* (ALTE) bzw. den Stufen B2.1 (*Lower Vantage Level*) bis C1.2/C2.1 (*Higher Effective Proficiency*) des Europarats entsprechen (vgl. Association of Language Testers in Europe, 1998, Kap. 2; Europarat, 2001, Kap. 3; TestDaF-Institut, 2001, S. 96ff.). Der Testteil „Leseverstehen“ besteht aus drei Texten mit insgesamt 30 Items (Zuordnungs-, Mehrfachwahl- und Auswahl-Items), der Testteil „Hörverstehen“ aus 25 Items (gesteuerte Notizen, Alternativformen). Die Überprüfung der schriftlichen Ausdrucksfähigkeit erfolgt anhand einer Texterstellungsaufgabe. Zur Bewertung der mündlichen Ausdrucksfähigkeit wird ein “Simulated Oral Proficiency Interview” (SOPI) eingesetzt, das aus 10 situativ eingebetteten Situationen besteht, in denen die Prüfungsteilnehmenden sprachlich reagieren müssen (vgl. Kniffka & Üstünsöz-Beurer, 2001). Die Leistungen in den produktiven Testteilen „Schriftlicher Ausdruck“ und „Mündlicher Ausdruck“ werden von mindestens zwei Korrektorinnen und Korrektoren unabhängig voneinander anhand eines Kriterienkatalogs bewertet.

Wie in standardisierten internationalen Tests üblich, werden alle für TestDaF neu entwickelten Aufgaben vor ihrem Einsatz erprobt. Ziel der Erprobungsphase ist u.a. auch eine Kalibrierung der Schwierigkeiten der Aufgaben in den Subtests „Leseverstehen“ und „Hörverstehen“ und eine Zuordnung der Punktscores zu den TestDaF-Niveaustufen 3, 4 oder 5. Hierzu werden Ankeritems, d.h. Items mit einem feststehenden Schwierigkeitsgrad, sowie das Rasch-Modell der probabilistischen Testtheorie eingesetzt. Bislang wurden zur Verankerung vom *University of Cambridge Local Examinations Syndicate* (UCLES) entwickelte Grammatik- und Lexik-Aufgaben im Format kurzer Lückentexte verwendet. Die insgesamt 15 Ankeritems stammen aus der deutschen Itembank des computeradaptiven Tests “Linguaskill”. Detailliertere Hinweise zum TestDaF finden sich unter http://www.testdaf.de sowie z.B. in Grotjahn (2001), Grotjahn & Kleppin (2001) und Projektgruppe TestDaF (2000).

2. Überblick über die durchgeführten Untersuchungen

2.1. Zielsetzung und methodisches Vorgehen

Es wurde überprüft, ob nicht anstelle der bereits mehrfach verwendeten UCLES-Ankeritems C-Tests zur Verankerung neuer TestDaF-Aufgaben eingesetzt werden könnten. Für den Einsatz von C-Tests sprechen u.a. folgende Argumente: (1) Der C-Test eignet sich gut zur Erfassung globaler Sprachkompetenz. (2) Er korreliert in der Regel (relativ) hoch mit dem Lese- und dem Hörverstehen. (3) Er ist sowohl hinsichtlich der Testerstellung als auch hinsichtlich der Testauswertung hoch praktikabel und ökonomisch (vgl. zu den Argumenten 1-3 Grotjahn, 1995; Raatz & Klein-Braley, 1998). (4) Eine computerbasierte Testerstellung, Testdurchführung und Testauswertung ist leicht möglich (vgl. Grießhaber, 1998; Koller & Zahn, 1996; Röver, 2002; sowie auch http://www.phil.uni-erlangen.de/erltest/). Dieses Argument ist auch im Hinblick auf die in Entwicklung befindliche computer-basierte TestDaF-Version zu sehen. (5) Der C-Test ist als Einstufungstest für Deutsch als Fremdsprache an deutschen Hochschulen seit langem etabliert und gilt in diesem Kontext als adäquates Messinstrument (vgl. z.B. Bolten, 1992; Grießhaber, 1998; Grotjahn & Allner, 1996).[1]

Bei der Erprobung neuer TestDaF-Aufgaben wurden die von UCLES zur Verfügung gestellten Aufgaben, deren Schwierigkeitsgrad bekannt ist, und die eigens entwickelten C-Test-Items zusammen eingesetzt. Den Versuchspersonen war nicht bekannt, dass die UCLES-Aufgaben und der C-Test rein testmethodischen Zielen dienten. Ziel war es, sowohl die neuen Lese- und Hörverstehensaufgaben als auch den C-Test mittels der UCLES-Items zu verankern. Durch die Verankerung des C-Tests sollten die Voraussetzungen geschaffen werden, diesen später selbst als Anker für die Erprobung neuer TestDaF-Aufgaben einsetzen zu können.

Vor Durchführung des Tests wurden die Teilnehmenden gebeten, Angaben zu Geschlecht, Alter, Herkunftsland, Erstsprache, Fremdsprachenkenntnissen sowie zur Dauer des bisherigen Deutschunterrichts zu machen.

2.2. Versuchspersonen

Die Versuchspersonen entsprachen mit Ausnahme einer Voruntersuchung mit Muttersprachlerinnen und Muttersprachlern des Deutschen alle der Zielgruppe

[1] Inwieweit mit Hilfe von C-Tests akademische Sprachfähigkeit im Sinne der Cognitive-Academic Language Proficiency von Cummins (1984) erfasst werden kann, hat vor allem Daller (1999) untersucht (vgl. auch Daller & Grotjahn, 1999).

von TestDaF: Es handelte sich um Studierende, die entweder bereits in Deutschland studienvorbereitende Deutschkurse besuchten oder – noch im Heimatland – einen Studienaufenthalt in Deutschland planten.

2.3. Textauswahl

In den verschiedenen Versuchsgruppen wurden unterschiedlich viele C-Test-Texte mit jeweils 20 Lücken eingesetzt. Die Texte wurden zunächst – der vermuteten Schwierigkeit nach – aufsteigend angeordnet. Die endgültige Auswahl und Anordnung der Texte erfolgte auf der Basis der Ergebnisse der testmethodischen Auswertungen. Zur Bearbeitung standen pro Text fünf Minuten zur Verfügung.

Es wurden Texte aus dem akademischen Kontext ausgewählt, die hinsichtlich Thematik und Textsorte dem TestDaF-Format entsprachen und ein hinreichendes Schwierigkeitsspektrum abdeckten. Geeignete Quellen waren populärwissenschaftliche Zeitschriften, Info-Materialien zum Thema Studium, Buchbesprechungen u.Ä. Alle ausgewählten Texte wurden gekürzt und falls notwendig hinsichtlich Lexik und Syntax geändert. Ein – allerdings erst zu einem späteren Zeitpunkt aufgenommener Text – wurde selbst verfasst (Uni-Café). Um genügend C-Test-Texte für die endgültige Auswahl zur Verfügung zu haben, wurden für die ersten Erprobungen mehr Texte als für einen C-Test üblich eingesetzt.

2.4. Testkonstruktion

Die Texte wurden nach dem klassischen Tilgungsprinzip eingerichtet, d.h. beginnend mit dem zweiten Satz wurde die zweite Hälfte jedes zweiten Wortes getilgt. Am Ende des Textes verblieb ein unversehrter (Teil)-Satz. Von der Tilgung ausgenommen blieben Eigennamen und Abkürzungen.

Teils von Beginn an, teils bei der Revision der Items wurde in Ausnahmefällen aus sprachspezifischen Überlegungen abweichend vom Prinzip getilgt. Das betraf vor allem Komposita, deren zweite Komponente nach dem Prinzip n/2 bzw. (n + 1)/2 entweder ganz oder bis auf einen Buchstaben getilgt worden wäre. Hier war es im Hinblick auf die Lösbarkeit und zur Vermeidung von Varianten notwendig, von der zweiten Komponente genügend Text unversehrt zu lassen (vgl. auch Abschnitt 2.7 sowie die Hinweise zu sprachspezifischen Varianten des Tilgungsprinzips in Grotjahn, 1995).

Das Tilgungsprinzip wurde bei der Revision auch aufgegeben, um die Schwierigkeit von Lücken bzw. Textstellen zu erhöhen, und zwar insbesondere bei Lücken, die bei Deutschlernenden eine sehr hohe Lösungsrate und somit eine geringe Trennschärfe aufwiesen. Dies betraf z.B. Konjunktionen wie „oder“

(Lücke 18 im Text „Sprachkurse“): Hier wurde statt ursprünglich od_____ nunmehr o_____ getilgt.

Weiterhin wurden die Texte auch hinsichtlich der Lexik revidiert, wenn dies zur Vermeidung von akzeptablen Varianten oder zu einer gewünschten Veränderung der Schwierigkeit führte. So wurde im Text „Die menschliche Haut“ die erste Lücke statt „Die Haut“ (getilgt Ha_____) zu „Dieses Organ“ (getilgt Or_____), weil im Einleitungssatz bereits die Vokabel „Haut“ vorkam, wodurch die Lösungsrate bei der ersten Lücke relativ hoch war. Als Folge wird allerdings vermutlich an dieser Textstelle nicht nur Sprachbeherrschung, sondern auch Weltwissen geprüft: Die Versuchspersonen müssen nicht nur die Vokabeln Haut und Organ kennen, sondern auch über das Wissen verfügen, dass es sich bei der Haut um ein Organ handelt. Durch das Demonstrativpronomen „dieses“ wird der Bezug zum Subjekt des vorhergehenden Satz, nämlich die Haut, hergestellt, so dass sich hier die Sprachbeherrschung auch darin zeigt, ob die Funktion des Demonstrativpronomens erkannt und gegebenenfalls eine entsprechende Lesestrategie verwendet wird.

Ein Beispiel für abgeänderte Lexik zur Reduzierung der Schwierigkeit sei aus dem Text „Sprachkurse“ angeführt. In der ursprünglichen Fassung war das Wort „alteingesessenen“ von der Tilgung getroffen: alteinge_____ . Diese Lücke erwies sich als besonders schwierig, vermutlich wegen der niedrigen Frequenz der Vokabel, und wurde daher zu „vorhandenen“ (getilgt: vorha_____) geändert.

Der Text „Stellenausschreibung“ (es handelt sich um eine Stellenausschreibung für eine Professur an einer deutschen Hochschule) hatte sich bereits in der Voruntersuchung mit Muttersprachlerinnen und Muttersprachlern als zu schwierig erwiesen.

Als ungeeignet erwies sich außerdem der Text „Demokratie“ nach Auswertung der Daten aus der Voruntersuchung mit Muttersprachlerinnen und Muttersprachlern (Lösungsraten < 90% bei einigen Lücken; Probleme entstanden vor allem im Bereich von Morphologie und Lexik).

In die Erprobung wurden zunächst 12 Texte à 20 Lücken aufgenommen.

2.5. Untersuchungen

Die 12 C-Test-Texte wurden zuerst in einer Voruntersuchung sowohl mit 11 deutschen Studierenden (am Seminar für Sprachlehrforschung der Ruhr-Universität Bochum) als auch mit zwei deutschen Hochschullehrkräften erprobt. Ziel der Analysen war zum einen die Feststellung der Lösungsraten. Betrug die Lösungsrate einer einzelnen Lücke weniger als 90%, wurde die Lücke revidiert (z.B. durch Verringerung der Zahl der getilgten Buchstaben) oder auch der Text

verworfen. Zum anderen diente die Voruntersuchung der Feststellung von akzeptablen Varianten. C-Test-Texte, die insgesamt eine zu niedrige Lösungsrate und zu viele Varianten aufwiesen, wurden verworfen.

Nach Revision bzw. Auswahl anhand der Ergebnisse aus der Voruntersuchung wurden acht Texte in einer ersten Erprobung mit Deutschlernenden eingesetzt. Es handelt sich um zwei Lerngruppen am Studienkolleg Bochum. Die erste Gruppe bestand aus 18 Personen (12 Männer und 6 Frauen), die einen Mittelstufen-Kurs besuchten. Bei der zweiten Gruppe handelte es sich um 17 Personen (10 Frauen und 7 Männer), die ca. vier Wochen vor der „Deutschen Sprachprüfung für den Hochschulzugang ausländischer Studienbewerber“ (DSH) standen und damit der Zielgruppe voll entsprachen.

Beide Gruppen erwiesen sich als sehr heterogen hinsichtlich Herkunftskultur, Muttersprache, Alter (21 bis 38 Jahre in der Bochumer Mittelstufen-Gruppe; 20 bis 66 Jahre in der Bochumer Oberstufen-Gruppe) und Dauer des DaF-Unterrichts (in der Oberstufen-Gruppe reichten die Angaben von vier Monaten bis vier Jahren, in der Mittelstufen-Gruppe von sieben Monaten bis drei Jahren Deutsch in der Schule plus zwei Monaten DaF in Deutschland).

Nach Auswertung der Ergebnisse und erneuter Revision problematischer Lücken wurden die C-Test-Texte erneut erprobt. Es handelte sich um eine Gruppe von 58 Deutschlernenden (32 Männer und 26 Frauen) unterschiedlichen Niveaus (Mittelstufe bis Oberstufe) am Studienkolleg Münster. Auch diese Gruppe erwies sich als äußerst heterogen, vor allem auch hinsichtlich der Dauer des Deutschunterrichts (die Angaben reichten von zwei Monaten bis zehn Jahren).

Nach der Auswertung wurde ein Text verworfen, andere Texte wurden wiederum revidiert. Außerdem wurden drei neue Texte (nach Vorerprobung mit sechs Muttersprachlerinnen und Muttersprachlern und anschließender Revision) aufgenommen, so dass in der nächsten Erprobung am Fachsprachenzentrum der Universität Hannover insgesamt 10 Texte zum Einsatz kamen. Die Gruppe bestand aus 113 Personen (63 Frauen, 50 Männer), vorwiegend aus der VR China, außerdem ein größerer Teil aus Russland bzw. der GUS sowie aus Osteuropa. Die Angaben zu den Vorkenntnissen reichten von „seit Schulzeit“ bis „seit drei Monaten“, so dass auch diese Gruppe hinsichtlich Deutschkenntnisse, Muttersprache etc. als heterogen zu bezeichnen ist.

Nach Auswertung der Daten der Hannover-Gruppe wurden die neuen C-Test-Texte revidiert. Auf der Grundlage von Itemanalysen wurde eine Auswahl von vier Texten vorgenommen. Diese Texte wurden schließlich im Rahmen der Erprobung zweier neuer TestDaF-Prüfungssätze zusammen mit den von UCLES entwickelten Ankeritems überprüft: Zunächst jeweils im Rahmen von Vorerprobungen mit 20 bzw. 12 Personen mit Deutsch als Muttersprache, anschließend

weltweit mit 188 bzw. 144 Deutschlernenden. Im Frühjahr 2002 wurden die C-Tests, wiederum zusammen mit den UCLES-Ankeritems, erneut im Rahmen dreier Erprobungssätze eingesetzt. Daten aus diesen Erprobungen liegen bislang nicht vor. Der zeitliche Verlauf der C-Test-Erstellung ist in Tabelle 1 dargestellt.

Tabelle 1: Zeitlicher Verlauf der C-Test-Erstellung

Datum	**Gruppe**	***N***	**Texte**	**Bemerkung**
01/01	Muttersprachler/innen	11 + 2	12	danach Auswahl von 8 Items
02/01	Studienkolleg Bochum	18 + 17	8	danach Revision
03/01	Studienkolleg Münster	58	8	danach Revision bzw. Verwerfen von Items und Erstellung neuer Items
03/01	Muttersprachler/innen	6	3	Voruntersuchung neue Items
04/01	Fachsprachenzentrum Hannover	113	10	danach definitive Auswahl und Revision der TestDaF-tauglichen Items
05/01	Bonn und Köln: Muttersprachler/innen Im Rahmen der Vorerprobung neuer TestDaF-Aufgaben (V004)	20	4	Items identisch mit vorhergehender Gruppe.
06/01	weltweit im Rahmen TestDaF-Erprobungsprüfung (E004)	188	4	Items identisch mit vorhergehender Gruppe, jedoch Revision der Lücke 13 (Text 2)
08/01	Muttersprachler/innen Im Rahmen der Vorerprobung neuer TestDaF-Aufgaben (V005)	12	4	Items identisch mit vorhergehender Gruppe
09/01	weltweit im Rahmen TestDaF-Erprobungsprüfung (E005)	144	4	Items identisch mit vorhergehender Gruppe
04/02	weltweit im Rahmen von drei TestDaF-Erprobungsprüfungen (E006, E007, E008)	705	4	keine Veränderungen

2.6. Auswertungskategorien

Da nicht ausgeschlossen werden konnte, dass bei der TestDaF-Zielgruppe unterschiedliche Modalitäten bei der C-Test-Auswertung zu Unterschieden in der Rangordnung der Probanden führen, erschien es notwendig, bei der Entwicklung der Ankeritems zunächst möglichst detaillierte Auswertungskategorien zu verwenden. Zudem erlaubt eine Auswertung anhand detaillierter Kategorien ein besseres Verständnis der von den Probanden jeweils erbrachten Leistung. Allerdings wird dieser Vorteil mit einer gewissen Unschärfe bei der Zuordnung zu den Kategorien und zugleich mit einem erheblichen Verlust an Auswertungsökonomie erkauft.

Im Rahmen der Vorerprobung an Muttersprachlerinnen und Muttersprachlern und im Rahmen der Erprobung an den DaF-Gruppen in Bochum und Münster sowie bei der Erprobung neuer TestDaF-Aufgaben (E004) wurde zwischen folgenden sechs Auswertungskategorien unterschieden:[2]

Kategorie 1: unausgefüllt
Kategorie 2: orthografisch richtiges Original
Kategorie 3: grammatisch und/oder inhaltlich nicht akzeptabel
Kategorie 4: orthografisch richtige Variante
Kategorie 5: orthografisch falsches Original
Kategorie 6: orthografisch falsche Variante

Die Kategorie „grammatisch und/oder inhaltlich nicht akzeptabel“ erweist sich zwar als ökonomisch bei der raschen Testauswertung, doch verhindert sie eine genauere Fehlerzuordnung und damit auch eine genauere Fehleranalyse sowie eine exaktere Einschätzung der Fähigkeit, da sie zwei unterschiedliche Bereiche umfasst: Inhalt und Grammatik, wobei Grammatik selbst unterschiedliche Teilaspekte beinhaltet (Syntax, Morphologie). Dies kann Auswirkungen auf die Validität haben, denn die Kategorie erlaubt keine Unterscheidung zwischen solchen Versuchspersonen, die das fehlende Wort erkennen, dieses aber grammatisch falsch ergänzen (zumeist morphologisch falsch), und denjenigen Versuchspersonen, die das erforderliche Wort gar nicht erkennen und daher die Lücke entweder unausgefüllt lassen oder inhaltlich und möglicherweise zugleich auch grammatisch falsch ergänzen.

---

[2] Andere, ökonomischere Kodierungen sehen lediglich die Kategorien „nicht ausgefüllt“, „falsch ausgefüllt“, „richtig ausgefüllt“ vor, wobei zur Kategorie „falsch ausgefüllt“ sowohl grammatische, lexikalische als auch orthografische Fehler gezählt werden.

Um Fehler präziser zuordnen zu können, wurde daher die Auswertungskategorie „grammatisch und/oder inhaltlich nicht akzeptabel“ folgendermaßen aufgelöst:

Kategorie 3 (neu): inhaltlich nicht akzeptabel
Kategorie 7: inhaltlich akzeptabel, aber grammatisch falsch

Auf diese Weise war es möglich, zwischen Versuchspersonen zu unterscheiden, die auf der Makroebene vermutlich über die erforderliche Textverstehenskompetenz verfügen und deshalb lexikalisch richtig, jedoch z.B. morphologisch falsch ergänzen, und solchen Versuchspersonen, die lexikalisch kontextuell unangemessen ergänzen. Ab der Erprobung E005 wurden bei der Auswertung der C-Tests daher sieben Kategorien angelegt.

Einige Beispiele sollen den Vorteil dieser Differenzierungsmöglichkeit verdeutlichen.[3]

Eine Versuchsperson ergänzt im Text „Museumsmanagement“ (Vp 1 – Text 2; Studienkolleg Münster) folgendermaßen: „Es tr**ägt** dem sich wande**rung** Charakter v**on** Museen h**ier** zu besucheror**dnung** Dienstleistungszentren Rech**t** und h**at** bereits i**n** Vorfeld interna**tionalen** Aufmerksamkeit a**n** sich gez**eigt**.“ Die Ergänzung der Lücken zeigt, dass die Versuchsperson den Text nicht versteht und die Kollokationen nicht beherrscht, auch wenn vereinzelt korrekt ergänzt wurde. Die korrekte Ergänzung einzelner Lücken kann auf Grund der Kenntnis sprachspezifischer Strukturen oder aber dank der Kenntnis des Tilgungsprinzips erfolgen, wie beispielsweise die Ergänzungen v**on** und h**at** vermuten lassen.

Die gleiche Versuchsperson zeigt im Text 8 „Hochschulpartnerschaften“ hingegen Textverstehenskompetenz, obwohl z.T. fehlerhaft ergänzt wird: „Diese Partner**schaft** beruhen i**n** der Re**gel** auf form**en** Vereinbarungen zw**ischen** zwei Hochs**schulen** bzw. ih**re** Leitungen u**nd** sehen langfr**ist** Maßnahmen d**es** Austausches u**nd** der Zusamme**narbeit** vor.“ Zwar weisen die Ergänzungen z.B. morphologische Fehler (Partner**schaft** statt Partner**schaften**, ih**re** statt ih**ren**) oder auch orthografische Fehler (Hochs**schulen** statt Hochs**chulen** – vermutlich auf Grund des unterbrochenen Schreibflusses verursacht) auf, doch kann festgehalten werden, dass die Ergänzungen hinsichtlich der Verstehenskompetenz qualitativ anders einzuordnen sind als unausgefüllte Lücken oder lexikalisch unangemessene Ergänzungen.

---

[3] Einige C-Test-Texte sind im Anhang aufgeführt. Aus Gründen der Testsicherheit können jedoch nur solche C-Tests veröffentlicht werden, die nicht im Rahmen des TestDaF zum Einsatz kommen.

Die Frage nach der korrekten Wortart bleibt jedoch auch hierbei problematisch. Wenn wie im Text „Menschliche Haut“ die Textstelle „faszin_____“ (Kontext: „Die Haut hat noch eine weitere faszinierende Eigenschaft“), zu „faszinierend“ ergänzt wird, so liegt durch die morphologisch falsche Ergänzung auch die falsche Wortart vor (Adverb statt Adjektiv), so dass auch hier eine Zuordnung sowohl zur Kategorie grammatisch falsch, als auch – je nach Interpretation – zur Kategorie lexikalisch falsch möglich ist.

Ein weiteres Problem besteht darin, dass bestimmte fehlerhafte Ergänzungen auf Grund ihrer Wortart nur bestimmten Fehlerkategorien zugeordnet werden können. Artikel beispielsweise können nur morphologisch falsch sein, nicht aber inhaltlich oder orthografisch falsch, ebenso sind keine Varianten denkbar.

Orthografische Varianten, die auf der alten Schreibweise beruhen blieben unberücksichtigt und wurden als „orthografisch richtiges Original“ gewertet.

2.7. Einige Auswertungsprobleme

Die für die Lösung von C-Tests notwendigen Fähigkeiten und Fertigkeiten variieren je nach Sprache. So erfordert die korrekte Ergänzung von Wörtern mit flektierten Suffixen nicht nur die Beherrschung der Lexik, sondern auch der Flexion (vgl. Grießhaber, 1998, S. 157). Aus diesem Grunde stellt vor allem die Wortbildung beim C-Test in deutscher Sprache ein (Auswertungs-)Problem dar: Wird ein Kompositum von der Tilgung getroffen, so kann u.U. der gelöschte Teil nicht mehr – eindeutig – rekonstruiert werden. Dies betrifft nicht nur Komposita, die aus zwei (oder mehreren) Substantiven zusammengesetzt sind. Auch solche Zusammensetzungen, die Präfixe und/oder Suffixe aufweisen, können problematisch sein. Die Folge ist, dass z.T. mehrere akzeptable Varianten möglich sind, was wiederum Auswirkungen auf den Auswertungsmodus und gegebenenfalls auf die Auswertungsökonomie hat.

Einige Beispiel sollen dies verdeutlichen. Das Wort „Einzelheiten“ im Text „Förderung von Hochschulpartnerschaften“ (Kontext: „Einzelheiten der akademischen Zusammenarbeit werden ... geregelt“) verbleibt nach der Tilgung als Einzel_____ . Bei der Vorerprobung mit Muttersprachlerinnen und Muttersprachlern wurden häufig keine Lösung oder aber Varianten (z.B. Einzelaspekte) angegeben. Deshalb wurde revidiert zu Einzelh_____ .

Als weiteres Beispiel sei die Lücke aus dem Text „Die menschliche Haut“ angeführt: Das betroffene Wort ist „Wärmeabgabe“, das nach dem Tilgungsprinzip als Wärmea_____ verbleibt. (Kontext: Die Haut ... reguliert ... die Wärmeabgabe an die Umwelt“). Eindeutig würde das zur rekonstruierende Wort nur, wenn die erste Komponente „Wärme“, das Präfix „ab“ und der erste Buchstabe der letzten Komponente „g“ verbliebe, also: „Wärmeabg_____“. Da-

mit würde die Textstelle vermutlich erheblich leichter, akzeptable Varianten wären jedoch nicht mehr möglich. Bei der Revision wurde „Wärmeab_____ “ belassen, was zur Folge hat, dass Varianten wie Wärmeabfuhr und Wärmeableitung zu akzeptieren sind.

Ebenso verhält es sich mit dem Wort „Konfliktbewältigung (Text „Buchtipp“ mit dem Kontext: „Zum heutigen Handwerkszeug der Führungskräfte zählen auch die sogenannten Kompetenzen wie Teamarbeit, Konfliktbewältigung oder konstruktive Kritik“). Getilgt ergibt sich Konfliktb_____ . Mehrere akzeptable Varianten wurden festgestellt: Konfliktbeseitigung, Konfliktberatung bis hin zu Konfliktbereitschaft. Auch hier kann nur mehr Kontext, also mindestens Konfliktbew_____ , zu einer höheren Lösungsrate bzw. zur Vermeidung von Varianten führen.

Fraglich bleibt, inwiefern durch die veränderte Tilgung die Textstelle nicht nur eindeutiger und leichter, sondern zugleich auch weniger trennscharf wird. Denn sehr gute Probanden lösen wahrscheinlich auch die ursprüngliche Fassung korrekt und unterscheiden sich somit deutlicher von schwächeren Versuchspersonen.

Bei dem folgenden Beleg aus dem Text „Buchtipp“ „D_____ Titel „Soz_____ Kompetenz“ v_____ Rudolf Donnert befa_____ sich ....“ (Der Titel „Soziale Kompetenz“ befasst sich mit ...) wird vermutlich fälschlicherweise ein Kompositum erkannt: Statt „Soziale Kompetenz“ wurde häufig „Sozial Kompetenz“ gelöst. Entweder handelt es sich um einen Interferenzfehler (z.B. aus dem Englischen), oder aber ein Kompositum wird erkannt: „Sozialkompetenz“. Eine Ursache hierfür mag auch sein, dass „Soziale“ als Teil des Buchtitels groß geschrieben erscheint, obwohl es sich um ein Adjektiv handelt. In der Revision wurde dieses Problem umgangen, indem der Buchtitel von der Tilgung ausgespart wurde: D_____ neue Ti_____ „Soziale Kompetenz“ v_____ ... . Da sich der Text „Buchtipp“ jedoch auch in späteren Erprobungen – vor allem wegen der Anzahl an Varianten bei einigen Lücken – als ungeeignet erwies, wurde er schließlich verworfen.

Gehäuft treten auch Folgefehler auf. Es handelt sich um fehlerhafte Ergänzungen, die sich über mehr als eine Lücke erstrecken und in sich stimmig sind, auch wenn sie dem Kontext nicht entsprechen. Als Beispiel diene der erste Satz im Text „Globale Erwärmung“: „Und d_____ Mensch kön_____ daran d_____ Hauptschuld tra_____ . (Und der Mensch könnte daran die Hauptschuld tragen)“. Überraschend häufig wurde die erste Lücke fehlerhaft, nämlich durch „die“ ergänzt. Vermutlich wurde der Singular übersehen, worauf auch hinweist, dass in der Folge die Verbform ebenfalls im Plural erscheint: „könnten“ statt „könnte“. Dass die Singularform übersehen wird und die

Pluralform, also „Menschen“ statt „Mensch“ wahrgenommen wird, könnte an der Leseerwartung liegen, dass hier tatsächlich die Menschen, also alle, gemeint sind, aus stilistischen Gründen jedoch „Mensch“ verwendet wird. Zu überlegen bleibt, ob Ergänzungen infolge fehlerhaft gelöster Lücken evtl. als korrekte bzw. akzeptable Variante gewertet werden sollten.

In diesem Zusammenhang sei auch darauf hingewiesen, dass Varianten eigentlich nur dann als akzeptabel kategorisiert werden können, wenn sie mit dem Kontext in Einklang stehen. So wird beispielsweise im Text „Globale Erwärmung“ an der Textstelle: „Das geht aus einem Bericht des zwischenstaatlichen Gremiums für Klimaveränderung hervor, ...“ häufig statt „Gremiums“ der Plural: „Gremien“ ergänzt, was zunächst einmal eine mögliche Variante darstellt. Wie verhält es sich aber, wenn im Kontext nicht analog auch der vorausgehende Artikel als Genitiv und als Plural markiert ist, also statt „der zwischenstaatlichen Gremien“, „des zwischenstaatlichen Gremien“ ergänzt wird. „Des“ für sich genommen entspricht dem Original „des zwischenstaatlichen Gremiums“. Im Plural müssten beide Lücken entsprechend folgendermaßen gelöst werden: „Der zwischenstaatlichen Gremien“. Sowohl „Gremien“ als auch „der“ wären also akzeptable Varianten. Doch zusammen mit „Gremien“ wäre „des“ keine akzeptable Variante. Nur eine Version, entweder singularisch oder pluralisch, ist also akzeptabel.

Vermutlich ergänzen gerade kreative Lernende zuweilen, indem sie zugleich nicht von der Tilgung betroffene Textstellen „korrigieren“. Ein Beispiel sei angeführt: Im Text „Betreuung von ausländischen Studierenden“ ist eine Textstelle in d_____ fremden La_____ zu ergänzen (Kontext: Da reicht die Beratung .... nicht aus, um sich in dem fremden Land wohl zu fühlen). Ergänzt wird zuweilen der Plural, also „in den Ländern“, wobei der Vokal „a“ zu Umlaut „ä“ korrigiert wird. Obwohl zwar im Einführungssatz ausdrücklich vom Aufenthalt an deutschen Hochschulen die Rede ist, muss sich die Integration ausländischer Studierender nicht unbedingt allein auf Deutschland beziehen. Da es sich damit um eine sowohl grammatisch als auch inhaltlich sinnvolle Ergänzung handelt, ist zu überlegen, entsprechende Lösungen zu akzeptieren.

Eine weitere Frage ist, wie mit „doppelten“ Fehlern umzugehen ist, also Fehlern, die sowohl der Kategorie „orthografisch falsch“ als auch der Kategorie „grammatisch falsch“ zuzuordnen sind. Beispielsweise ist im Text „Uni-Café“ die Lücke „Semi_____ “ im folgenden Kontext zu füllen: (die Studierenden) … unterhalten sich, bereiten Seminare vor. Wird nun statt Seminare „Semina“ ergänzt, so kann hier einerseits ein Orthografiefehler vorliegen, weil vermutlich vom Lautbild auf das Schriftbild geschlossen wird, andererseits aber auch ein Morphologiefehler, denn der Plural wurde nicht erkannt. Klare Ent-

scheidungshilfen bei der Fehlerzuordnung fehlen ebenso wie die Gewichtung der Fehlerkategorien. Ähnlich verhält es sich mit der Lücke „Hochs_____ “ im Text „Hochschulpartnerschaften (Kontext: „ … zwischen zwei Hochschulen bzw. ihren Leitungen …“). Wenn nun ergänzt wird: „Hochshul“, kann es sich erstens um einen Orthografiefehler handeln (bzw. zwei Fehler, denn es fehlt ein „c“ sowie die Endung „e“ im Falle des Singulars oder die Endung „en“ im Falle des Plurals), oder aber es handelt sich um einen Morphologie-Fehler (die korrekte Endung fehlt). Auch hier ist unklar, welcher Fehlerkategorie die Ergänzung zuzuordnen ist und wie gewichtet werden soll.

Häufig stellt sich bei der Auswertung zudem die Frage, wo die Grenze zwischen Orthografie- und Lexik-Fehler liegt. So gab es im Item „Weltkonferenz“ unterschiedliche Ergänzungen der Lücke We_____ (Kontext: „Sie ist eine politische Verpflichtung auf internationaler Ebene, die Welt nachhaltig zu entwickeln“). Kann „Wält“ noch als orthografisch falsch eingestuft werden, wie verhält es sich dann mit „Walt“? Liegt bei der zweiten Ergänzung ein Lexikfehler vor, ist also evtl. „Wald“ gemeint? Gleiches gilt für die Ergänzung „Wehlt“, denn hier kann auch „wahlen“, also „wählt“ gemeint sein.

Ähnlich verhält es sich mit der Ergänzung der Textstelle I_____ der Re_____ im Text „Förderung von Hochschulpartnerschaften“ (Kontext: „Diese Partnerschaften beruhen in der Regel auf ...“). Ergänzt wurde mitunter „In der Regal“, wobei die zweite Ergänzung als Lexikfehler (also Regal statt Regel) interpretiert werden könnte. Allerdings widerspricht dieser Fehlerinterpretation der Umstand, dass die feste Wendung „in der Regel“ zum Grundwortschatz gehört und falls tatsächlich „Regal“ gemeint ist, auch der Artikel nicht korrekt wäre. Zudem verweist nichts im Kontext auf die Vokabel „Regal“ bzw. auf Einrichtungsgegenstände.

3. Statistische Analysen

3.1. Einfluss der Auswertungsmethode

Die detaillierte Auswertung der C-Tests anhand der sieben Kategorien beeinträchtigt zwar die Auswertungsökonomie beträchtlich, ermöglicht jedoch eine genauere Analyse der individuellen Leistung. Es soll deshalb im Folgenden überprüft werden, inwieweit das Auswertungsverfahren – und zwar speziell die unterschiedliche Kategorisierung von orthografischen Varianten[4] – einen Einfluss auf die Punktwerte und die Rangfolge der Probanden sowie auf die Relia-

[4] Beim praktischen Einsatz von C-Tests stellt sich immer wieder die Frage, wie insbesondere Orthografiefehler bei der Auswertung zu behandeln sind.

bilität und Validität des C-Tests hat und ob letztendlich nicht eine ökonomischere Auswertung vertretbar ist. Da das erweiterte, siebenstufige Kategoriensystem erst ab der Erprobungsprüfung E005 angewendet wurde (vgl. Abschnitt 2.6), liegen für eine Evaluation des siebenstufigen Systems noch nicht genug Daten vor. Wir beschränken uns deshalb auf eine Analyse des ursprünglichen sechsstufigen Kategoriensystem anhand der Daten aus den Studienkollegs Bochum und Münster sowie aus der weltweiten TestDaF-Erprobungsprüfung E004.

Wie bereits erwähnt, handelt es sich bei den Probanden aus den Studienkollegs Bochum und Münster jeweils um sehr leistungsheterogene Gruppen, deren Niveau zudem z.T. unterhalb der Leistung der TestDaF-Zielgruppe liegen dürfte. Dies gilt insbesondere für die 18 Personen aus dem Bochumer Mittelstufenkurs. Erwartungsgemäß finden sich in den unteren Leistungsgruppen relativ viele Orthografiefehler und nur relativ wenige akzeptable Varianten. Es wurde deshalb lediglich zwischen folgenden zwei Auswertungsmethoden differenziert:

Methode A:
Orthografisch richtige Originale und orthografisch richtige Varianten werden als korrekt gewertet.

Methode B:
Sowohl orthografisch richtige als auch orthografisch falsche Originale und Varianten werden als korrekt gewertet.

Tabelle 2 zeigt die Mittelwerte und Standardabweichungen für alle acht Texte unterschieden nach den beiden Auswertungsmethoden. Wegen der geringen Stichprobengrößen haben wir im Folgenden die Daten für Bochum und Münster stets zusammengefasst.

Tabelle 2: Mittelwerte und Standardabweichungen unterschieden nach Auswertungsmethode A und B (Studienkolleg Bochum und Münster; $N = 93$)

	Methode A		Methode B	
	Mittelwert	Standardabw.	Mittelwert	Standardabw.
Text 1	12,92	2,58	15,00	2,53
Text 2	10,74	2,43	12,01	2,44
Text 3	10,55	2,90	12,91	2,52
Text 4	8,89	3,19	10,95	3,09
Text 5	8,71	3,34	11,08	3,13
Text 6	11,06	3,16	13,74	2,88
Text 7	10,94	3,76	12,52	3,80
Text 8	10,65	3,72	13,30	3,86
Gesamt	84,46	18,82	101,50	17,25

Wie die Tabelle 2 zeigt, führt die Bewertung von Orthografiefehlern als „korrekt“ bei den meisten Texten zu deutlich höheren Punktwerten. Bezogen auf den Gesamtpunktwert beträgt der Unterschied zwischen den beiden Auswertungsmethoden mehr als 17 Punkte und ist hochsignifikant (*t*-Test für gepaarte Beobachtungen). Anzumerken ist auch, dass die Methode B erwartungsgemäß zu einer geringfügig reduzierten Varianz führt. Der Unterschied ist allerdings nicht signifikant ($F(92, 92) = 1{,}19$).

Tabelle 3 zeigt die Pearson-Produkt-Moment-Korrelationen zwischen den Auswertungsmethoden A und B für alle acht Texte. Insgesamt besteht eine sehr hohe Korrelation zwischen den beiden Auswertungsmethoden. Zugleich weisen die Daten jedoch darauf hin, dass die Berücksichtigung von Orthografiefehlern zumindest bei einzelnen Texten zu nicht zu vernachlässigenden Differenzen in den Punktzahlen und in der Rangordnung der Probanden führen kann. Allerdings darf auch aufgrund der hohen Gesamtkorrelation von 0,953 keineswegs der Schluss gezogen werden, dass die beiden Auswertungsmethoden zu gleichen Rangordnungen führen.

Tabelle 3: Korrelation zwischen den Auswertungsmethoden A und B (Studienkolleg Bochum und Münster; $N = 93$)

Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Gesamt
,861	,888	,864	,934	,813	,869	,939	,915	,953

Anmerkung: Alle Korrelationen sind auf dem Niveau von 0,01 (2-seitig) signifikant.

Wir hatten bereits darauf hingewiesen, dass Orthografiefehler vor allem von leistungsschwächeren Probanden gemacht werden. Die Abbildung 1 zeigt den Zusammenhang zwischen der Zahl der orthografisch falschen und der Zahl der orthografisch richtigen Originale in den Studienkollegs Bochum und Münster in Form eines sog. Sonnenblumen-Streudiagramms.[5] Da die Zahl der orthografisch falschen Originale über ihren gesamten Wertebereich bei acht Texten mit jeweils 20 Lücken maximal den Wert „160 – Zahl der orthografisch korrekten Originale“ erreichen kann, ist im oberen Wertebereich der Variablen „Zahl der orthografisch richtigen Originale“ die Streuung der Variablen „Zahl der orthografisch falschen Originale“ notwendigerweise reduziert. Technisch gesprochen handelt es sich im vorliegenden Fall um eine sog. heteroskedastische (d.h. varianzinhomogene) Regression, die häufig mit Nicht-Linearität einhergeht. Aufgrund der hohen Streuung der Variablen „Zahl der orthografisch falschen Originale“ im unteren Wertebereich der Variablen „Zahl der orthografisch richtigen

[5] Die Zahl der Blätter zeigt die Zahl der jeweiligen Probanden an. Da nur eine sehr geringe Zahl von orthografisch falschen Varianten vorkam, haben wir lediglich Originale berücksichtigt.

Originale“ kann im unteren Leistungsspektrum nur sehr bedingt von der Zahl der orthografisch richtigen Originale auf die Zahl der orthografisch falschen Originale geschlossen werden.

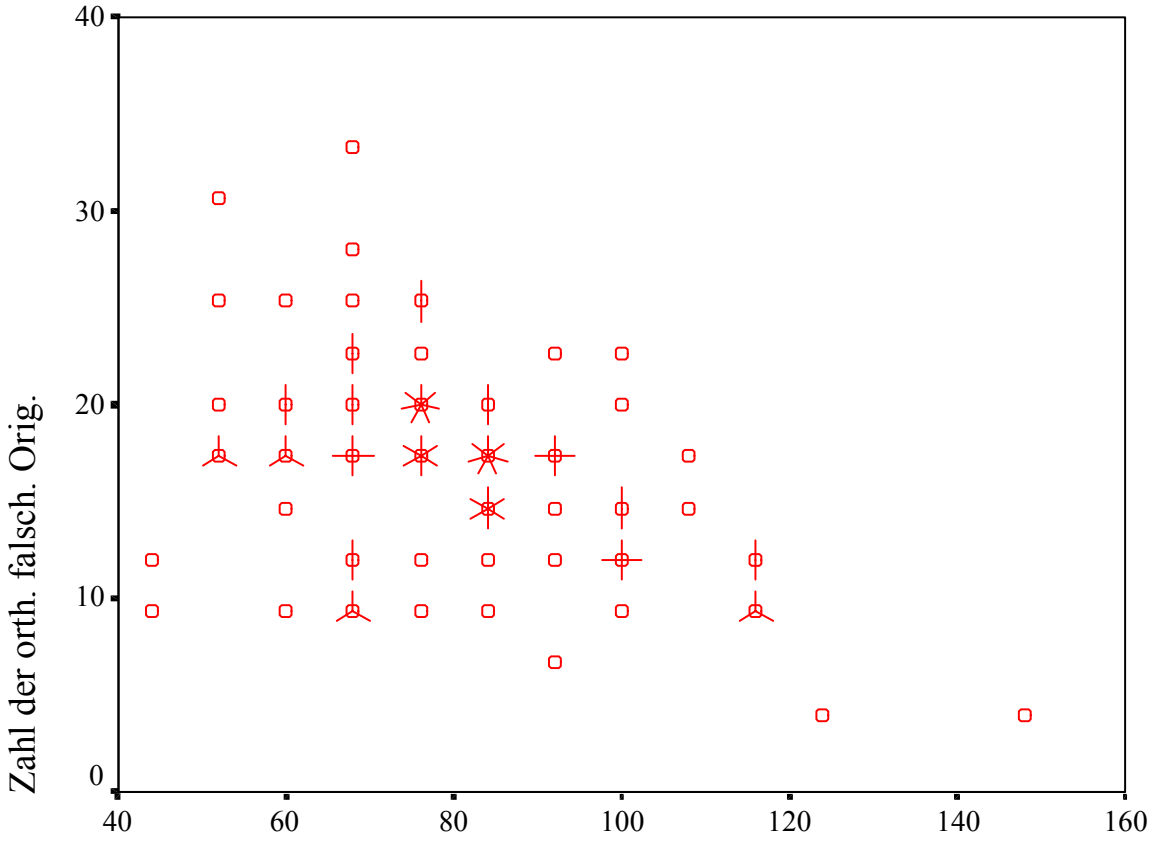

Abbildung 1: Streudiagramm der Zahl der orthografisch falschen und der orthografisch richtigen Originale (Studienkolleg Bochum und Münster; $N = 93$)

In Abbildung 2 ist für die gleiche Stichprobe die Regression der Zahl der orthografisch falschen Originale auf die Zahl der orthografisch richtigen Originale grafisch dargestellt – und zwar sowohl für ein lineares Modell als auch für ein Polynom zweiten Grades. Wie die Werte für den adjustierten Determinationskoeffizienten zeigen (linear: $R_a^2 = 0{,}176$; Polynom 2. Grades: $R_a^2 = 0{,}233$), ergibt sich für die quadratische Funktion zwar eine etwas bessere Anpassung als für das lineare Modell; die Anpassung ist jedoch auch hier relativ schwach.[6]

[6] Daneben wurden noch einige weitere theoretisch sinnvolle nicht-lineare Modelle mit Hilfe der Prozedur „Kurvenanpassung“ in SPSS 10.0 berechnet. (Die Prozedur erlaubt die Schätzung der Parameter von insgesamt 11 verschiedenen Modellen.) Es ergab sich jedoch in keinem Fall eine bessere Anpassung.

Die Abbildung 2 liefert zugleich auch eine mögliche Begründung für die leicht reduzierte Streuung der Punktwerte im Fall der Auswertungsmethode B in Tabelle 2. Die Streuung der Punktwerte der Methode B setzt sich additiv zusammen aus der Streuung der Zahl der orthografisch richtigen Originale/Varianten und der Streuung der Zahl der orthografisch falschen Originale/Varianten sowie der Kovarianz der beiden Variablen. Da mit zunehmender Zahl orthografisch richtiger Originale/Varianten die Zahl der orthografisch falschen Originale/Varianten notwendigerweise abnimmt, ist die Kovarianz negativ. Dies führt ab einer bestimmten Größenordnung zu einem Absinken der Varianz der Auswertungsmethode B und damit möglicherweise auch zu einer reduzierten Reliabilität der Methode B im Vergleich zur Methode A (vgl. die genaueren Ausführungen in Grotjahn, 1989, 1992).

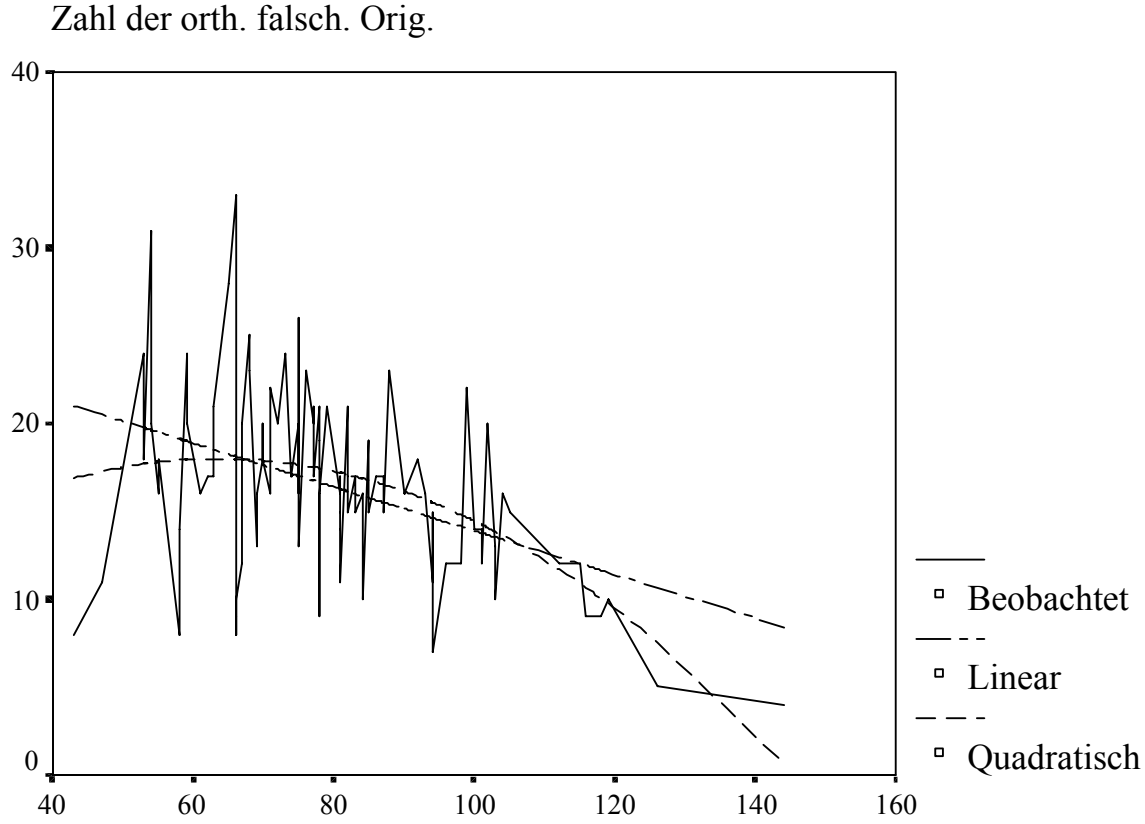

Abbildung 2: Regression der Zahl der orthografisch falschen Originale auf die Zahl der orthografisch richtigen Originale (Studienkolleg Bochum und Münster; $N = 93$)

Abbildung 3 zeigt die Regression des Gesamtpunktwertes entsprechend Auswertungsmethode B (GESAMT_B) auf den Gesamtpunktwert entsprechend Auswertungsmethode A (GESAMT_A). Der Zusammenhang ist zwar linear,

allerdings heteroskedastisch: Im unteren Leistungsbereich kann nur sehr eingeschränkt von der Variablen GESAMT_A auf die Variable GESAMT_B geschlossen werden. Insofern ist auch die hohe Gesamtkorrelation von 0,95 irreführend. Dies zeigt sich deutlich, wenn wir die Stichprobe am Median der Variablen GESAMT_A teilen: Für die schwächere Leistungsgruppe (Punktzahl < 82; $N = 44$) ergibt sich dann eine Korrelation von lediglich 0,86.

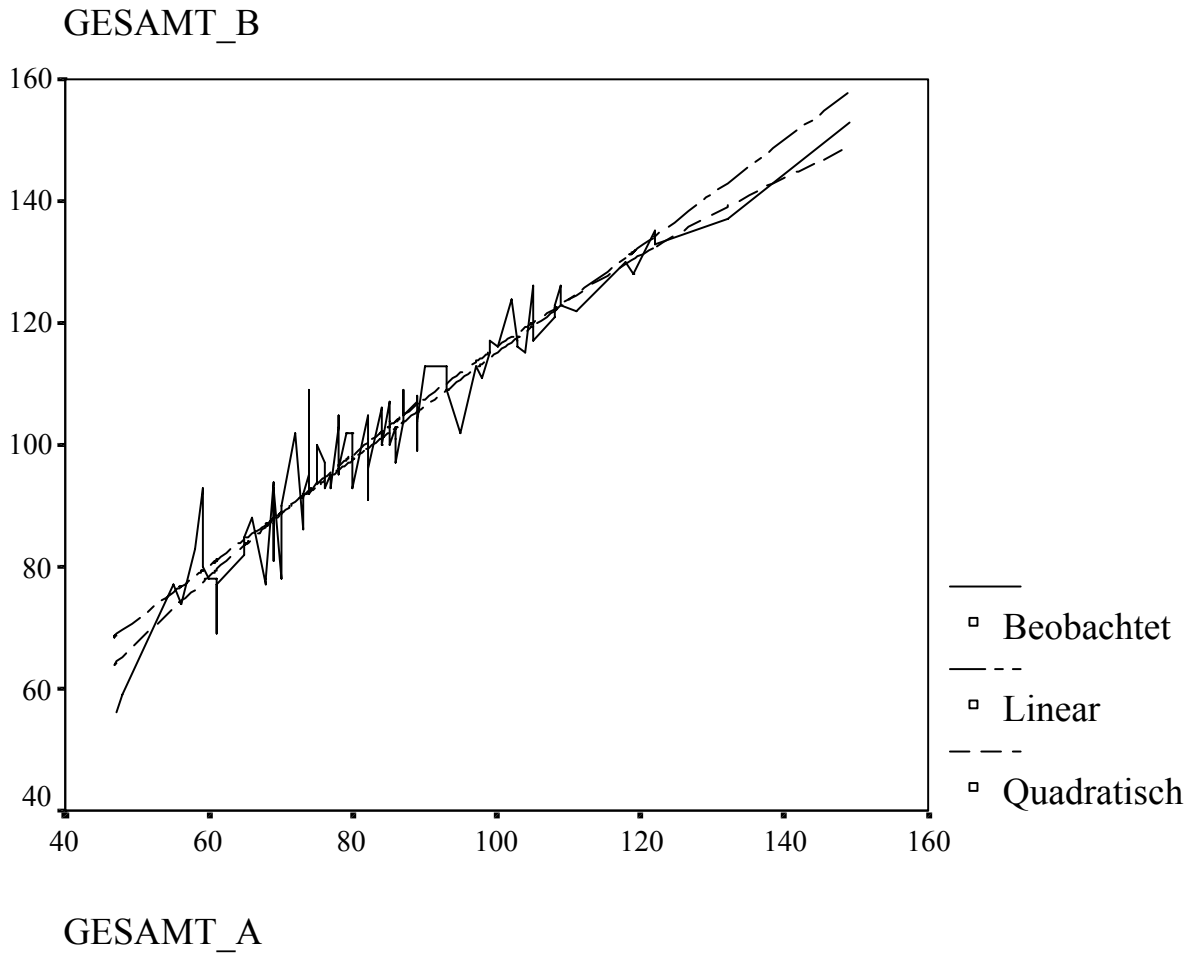

Abbildung 3: Regression des Gesamtpunktwertes entsprechend Auswertungsmethode B auf den Gesamtpunktwert entsprechend Auswertungsmethode A (Studienkolleg Bochum und Münster; $N = 93$)

Wir wollen nun die bisherigen Ergebnisse mit den Daten aus der weltweiten TestDaF-Erprobungsprüfung E004 vergleichen. Die Tabelle 4 zeigt analog zu Tabelle 2 die Mittelwerte und Standardabweichungen der Punktzahlen für die beiden Auswertungsmethoden A und B.

Tabelle 4: Mittelwerte und Standardabweichungen unterschieden nach Auswertungsmethode A und B (E004; *N* = 187)

	Methode A		Methode B	
	Mittelwert	Standardabw.	Mittelwert	Standardabw.
Text 1 (neu)	15,13	3,31	15,52	3,24
Text 5	11,95	4,17	12,14	4,12
Text 7	11,28	3,84	11,50	3,90
Text 8	10,02	3,45	10,12	3,44
Gesamt	48,39	12,23	49,28	12,10

Vergleicht man die Tabellen 2 und 4 fällt sofort auf, dass die unterschiedliche Wertung von Orthografiefehlern in der Gruppe E004 nur einen minimalen Effekt zu haben scheint: Wertet man die vergleichsweise wenigen Orthografiefehler als korrekt, werden unerheblich höhere Punktwerte erzielt, und die Varianz reduziert sich insgesamt gesehen erwartungsgemäß geringfügig. Entsprechend hoch ist auch die Korrelation zwischen den Auswertungsmethoden A und B (vgl. Tabelle 5). Auch hier darf jedoch die hohe Gesamtkorrelation von 0,997 nicht zu dem Schluss verleiten, die Rangfolgen seien identisch: In 20% der Fälle ergaben sich Unterschiede zwischen 5 und maximal 19 Rängen (bei einer Differenz von 0 bis 2 Gesamtpunktwerten in 93% der Fälle und 3 bzw. 4 Punktwerten in den restlichen Fällen).

Tabelle 5: Korrelation zwischen den Auswertungsmethoden A und B (*N* = 187; E004)

Text 1 (neu)	Text 5	Text 7	Text 8	Gesamt
,983	,995	,994	,994	,997

Anmerkung: Alle Korrelationen sind auf dem Niveau von 0,01 (2-seitig) signifikant.

Abbildung 4 zeigt wiederum die Regression des Gesamtpunktwertes entsprechend Auswertungsmethode B auf den Gesamtpunktwert entsprechend Auswertungsmethode A. Der Zusammenhang ist linear und homoskedastisch: Es kann bei der Stichprobe E004 über den gesamten Wertebereich der Variablen „GESAMT_A“ auf die Werte der Variablen „GESAMT_B“ geschlossen werden. Dies bedeutet konkret, dass es bei den vorliegenden vier Texten in der Gruppe E004 nur eine sehr geringe Auswirkung auf den C-Test-Punktwert hat, ob Orthografiefehler als korrekt oder als nicht korrekt gewertet werden.

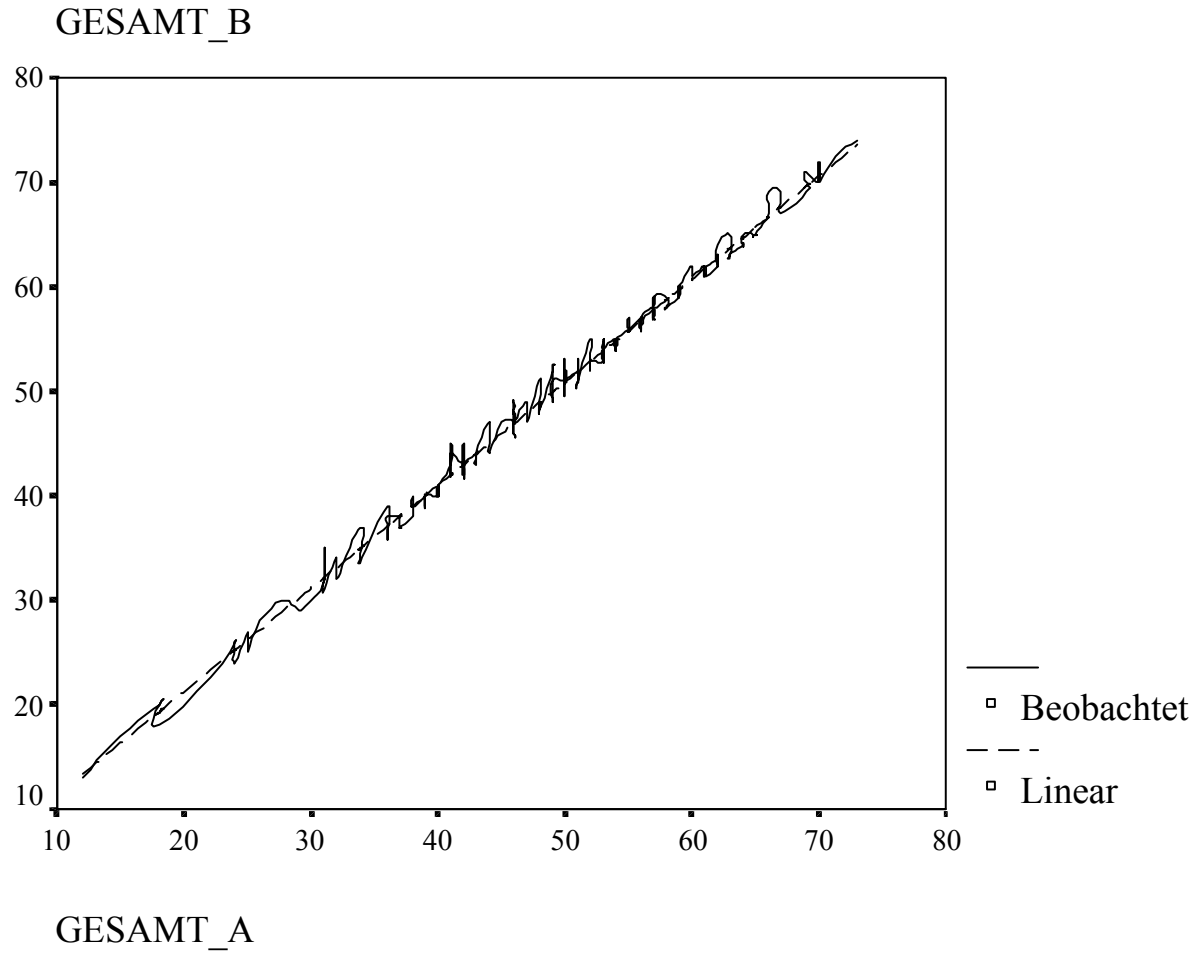

Abbildung 4: Regression des Gesamtpunktwertes entsprechend Auswertungsmethode B auf den Gesamtpunktwert entsprechend Auswertungsmethode A (E004; $N = 187$)

Für den geringen Effekt der Auswertungsmethode in E004 bieten sich zumindest zwei Erklärungsmöglichkeiten an:

1. Die Revision der Texte 5, 7 und 8 hat zu einer deutlich verminderten Zahl von Orthografiefehlern geführt.
2. Die Probanden der Gruppe E004 haben einen höheren Leistungsstand als die Probanden aus den Studienkollegs Bochum und Münster.

Die erstgenannte Begründung dürfte auf jeden Fall zutreffend sein. Hierfür spricht nicht nur die im Vergleich zu den Daten aus den Studienkollegs sehr geringe Zahl von Orthografiefehlern, sondern auch die Tatsache, dass der Anteil an orthografisch falschen Lösungen (Originale + Varianten) beim neuen Text 1 mit 1,91% deutlich größer ist als bei revidierten Texten (0,93%; 1,09%; 0,48%).

Für einen zusätzlichen Effekt der Variablen „Leistungsstand“ könnten die Ergebnisse eines Vergleichs der Leistungen der Teilnehmenden aus den Studien-

kollegs und der Gruppe E004 bei den gemeinsamen Texten 5, 7 und 8 sprechen. Die relevanten Daten finden sich in Tabelle 6. Die mittlere Leistung auf der Basis der Auswertungsmethode A ist in der Gruppe E004 signifikant höher als in den Studienkollegs ($t(278) = 2{,}48$; $p < 0{,}05$). Da wegen der vorgenommenen Revisionen allerdings nicht ausgeschlossen werden kann, dass die Texte leichter geworden sind, ist keine gesicherte Interpretation im Sinne von Leistungsunterschieden möglich.

Tabelle 6: Mittelwerte und Standardabweichungen für die Summe der Punktzahlen aus den Texten 5, 7 und 8 (Studienkollegs: $N = 93$; E004: $N = 187$)

	Methode A		Methode B	
	Mittelwert	Standardabw.	Mittelwert	Standardabw.
Studienkollegs	30,29	9,00	36,89	8,84
E004	33,26	9,67	33,76	9,65

3.2. Rasch-Skalierung der C-Test-Texte

Um einen genaueren Einblick in die Skaleneigenschaften des C-Tests zu erhalten, insbesondere um die Annahme der Eindimensionalität der vier C-Test-Texte bzw. -Items zu überprüfen, wurden die Items einer Rasch-Skalierung unterzogen. Wir beschränkten uns bei der Darstellung auf die im Gesamtzusammenhang relevantere Stichprobe E004 (Auswertungsmethode A).

Da C-Test-Items allgemein ein quasi-kontinuierliches Antwortformat aufweisen (im vorliegenden Fall mit Werten zwischen 0 und 20), verwendeten wir ein Testmodell, das gezielt zur Analyse kontinuierlicher Items entwickelt worden ist. Es handelte sich um das **kontinuierliche Ratingskalen-Modell** von Müller (1999). Der Einfachheit halber sei dieses Modell im Folgenden kurz „KRS-Modell“ oder „KRSM“ genannt.

Das KRS-Modell gehört zur Klasse der eindimensionalen Rasch-Modelle. Im Unterschied zum dichotomen Rasch-Modell wird im KRSM die Form der Antwortverteilung (oder genauer: die Dichtefunktion) nicht nur von der Differenz zwischen Personparameter (Fähigkeit) und Itemparameter (Schwierigkeit), sondern auch vom sog. Dispersionsparameter bestimmt. Der Dispersionsparameter gibt (vereinfacht gesprochen) an, inwieweit die Schwellenwerte entlang der kontinuierlichen Ratingskala monoton steigen (analog zum diskreten Ratingskalen-Modell mit äquidistanten Schwellen; vgl. Andrich, 1982). Anders ausgedrückt, benutzen Personen ein kontinuierliches Item tatsächlich als ein solches, dann nimmt der Dispersionsparameter Werte größer 0 an (**regulärer Fall**). Benutzen sie es dagegen eher als ein dichotomes Item, dann tendiert der Dispersionsparameter gegen 0 (**degenerierter Fall**). Negative Werte des Dispersionsparameters

schließlich zeigen eine klare Modellverletzung an (**irregulärer Fall**); eine Verletzung der Modellannahme könnte z.B. auf die Tendenz der Personen, extreme Antworten zu geben, zurückgehen.

Das KRSM lässt sich formal wie folgt skizzieren. Eine Person *v* bearbeitet Item (hier: Text) *i* mit Mittelpunkt *c* (hier: $c = 10$) und Länge *d* (hier: $d = 20$). Das Ergebnis sei x_{vi}. Wie in dieser allgemeinen Klasse von Modellen üblich, werden Eindimensionalität des Personmerkmals θ_v (hier: globale Sprachkompetenz) und lokale Unabhängigkeit der Item-Bearbeitungen (hier allerdings für jede einzelne Person) angenommen. Die Itemschwierigkeit sei (analog zum dichotomen Rasch-Modell) β_i, der (itemunabhängig konzipierte) Dispersionsparameter sei λ. Item *i* wird auf der latenten Merkmalsdimension durch ein sog. Schwellenintervall von $\beta_i - \lambda d$ bis $\beta_i + \lambda d$ repräsentiert. Die Dichtefunktion ist proportional zu (vgl. Müller, 1999, S. 97):

$$\exp[x_{vi}(\theta_v - \beta_i) - x_{vi}^2 \lambda].$$

Bei Modellgültigkeit, d.h., wenn die C-Test-Items im Sinne des KRSM skalierbar sind (vgl. zur Kontrolle der Modellgeltung die Ergebnisdarstellung weiter unten), folgt die Antwortverteilung einer doppelt gestutzten Normalverteilung („doppelt gestutzt“ wegen der endlichen Ausdehnung der kontinuierlichen Ratingskala). Anders als das KRS-Modell macht das (ebenfalls für kontinuierliche Ratingskalen konzipierte) klassische latent-additive Testmodell (KLA-Modell; Moosbrugger, 1992; Moosbrugger & Müller, 1982; vgl. auch die Ausführungen in Grotjahn, 1992, S. 235–239) keine Annahmen über den Verteilungstyp der manifesten Variablen. Aus dem KLA-Modell lässt sich daher nicht ableiten, welche Arten der Itembearbeitung zulässig oder sinnvoll sind. Das KLA-Modell ist damit im strengen Sinne kein probabilistisches Testmodell (vgl. Müller, 1999, S. 20f.).

Die KRSM-Analysen der C-Test-Daten aus der Prüfung E004 wurden anhand des Programms CRSM (Continuous Rating Scale Model, Version 1.3; vgl. Müller, 1999, S. 177) vorgenommen.[7] Dieses Programm erlaubt die Schätzung von Item-, Person- und Dispersionsparametern sowie die Berechnung verschiedener Kenngrößen für die Güte der Modellanpassung. Tabelle 7 fasst die Ergebnisse der KRSM-Analysen zusammen.

---

[7] Wir danken Herrn Dr. Hans Müller für die freundliche Überlassung des von ihm entwickelten CRSM-Programms und für seine Kooperationsbereitschaft bei der Implementierung des Programms.

Tabelle 7: Ergebnisse der Rasch-Skalierung nach dem kontinuierlichen Ratingskalen-Modell (E004)

C-Test-Text	Itemparameter	Dispersionsparameter (adjustiert)	*U*-Index	*Q*-Index
Text 1	–5,94	9,51	0,68	0,08
Text 2	0,48	6,81	0,77	0,07
Text 3	1,47	6,01	0,84	0,09
Text 4	3,99	9,28	0,71	0,09
Gesamt	0,00	7,90	0,75	0,08

Zunächst eine kurze Erläuterung zu den statistischen Schätz- bzw. Kenngrößen in den Spalten 2 bis 5 von Tabelle 7. Der Itemparameter (Spalte 2) gibt (wie in Rasch-Modellen üblich) die Schwierigkeit eines Items (d.h. C-Test-Textes) in Einheiten der Logit-Skala an (höhere Werte stehen für eine größere Itemschwierigkeit).

Die adjustierten Werte des Dispersionsparameters (Spalte 3) liefern eine grobe Schätzung der itemspezifischen Parameterwerte. Damit lässt sich die Modellannahme für alle Items konstanter Dispersionsparameter überprüfen. Der nichtadjustierte, itemunabhängige Parameterwert ist in der letzten Zeile (“Gesamt”) aufgeführt. Zugleich geben die adjustierten Werte des Dispersionsparameters Hinweise auf Items mit degenerierter Antwortverteilung (Werte um 0) oder mit irregulärer Antwortverteilung (negative Schätzwerte; vgl. Müller, 1999, S. 150).

Der *U*-Index (Spalte 4) dient der Kontrolle der Modellgeltung. Dieser im Rahmen des Partial-Credit-Modells entwickelte Index (vgl. z.B. Masters & Wright, 1997) basiert auf den standardisierten Residuen der Antworten von *N* Personen auf Item *i* (d.h. auf den standardisierten Abweichungen der beobachteten von den erwarteten Antworten). *U*-Werte nahe oder gleich 0 verweisen auf einen starken Overfit („Überanpassung“, d.h., das Antwortmuster der Personen hat stark deterministischen Charakter); Werte nahe 1 sprechen dagegen für eine hohen Grad der Modellgültigkeit (vgl. Göbel, Müller & Moosbrugger, 1999).

Ebenfalls zur Kontrolle der Modellgeltung wird der *Q*-Index eingesetzt (Spalte 5). Dieser Index, der Werte zwischen 0 und 1 annimmt, drückt aus, wie weit die Wahrscheinlichkeit des beobachteten Antwortmusters vom Minimum bzw. Maximum entfernt liegt. Anders als beim *U*-Index zeigen niedrigere *Q*-Werte eine größere Modellgültigkeit an, d.h. je kleiner der *Q*-Index, desto besser das Item (vgl. Rost, 1996, S. 366f.). Nach Göbel et al. (1999) sollte der *Q*-Index weniger als 0,5 betragen.

Wie Tabelle 7 zeigt, entsprechen die Schwierigkeiten der vier Texte der angestrebten Abstufung. Text 1 ist der mit Abstand leichteste, Text 4 der schwie-

rigste. Legt man die Standardfehler der paarweisen Differenzen von Itemschwierigkeiten zugrunde, sind die Schwierigkeitsunterschiede zwischen den Items alle statistisch signifikant. Der hohe (itemunabhängige) Wert des Dispersionsparameters ($\lambda = 7{,}90$) verweist darauf, dass der reguläre Fall des KRS-Modells angenommen werden kann. Darüber hinaus liegen die adjustierten Parameterschätzungen auf durchgängig hohem Niveau, sodass diese Schlussfolgerung auf alle C-Test-Texte in gleichem Maße zutrifft. Für den *U*-Index ergeben sich Werte zwischen 0,68 (Text 1) und 0,84 (Text 3). Diese Werte sprechen dafür, dass bei jedem einzelnen Text von vernachlässigbar geringen Modellabweichungen auszugehen ist. Schließlich lassen die sehr niedrigen Werte des *Q*-Index nicht den geringsten Zweifel daran, dass die vier Texte des hier betrachteten C-Tests durch das KRSM adäquat skalierbar sind.

Insgesamt kann also von einer sehr hohen Modellgültigkeit des beobachteten Antwortverhaltens bei der Bearbeitung der vier C-Test-Texte gesprochen werden. Der vorliegende C-Test erweist sich damit als **eindimensionales** Instrument zur Messung der globalen Sprachkompetenz. Weitere Erkenntnisse zur Reliabilität und Validität dieses C-Tests werden in den folgenden beiden Abschnitten berichtet.

3.3. Analysen zur Reliabilität

In Tabelle 8 sind Mittelwerte, Standardabweichungen, Schiefe, Minimum und Maximum für den C-Test und die vier Subtests des TestDaF aufgeführt. Für die Stufen „TDN 5“ bis „unter TDN 3“ wurden die Zahlen 5 bis 2 als Codierung verwendet. Wie die Tabelle zeigt, ist vor allem der Subtest Leseverstehen zu leicht und differenziert damit nicht hinreichend im oberen Leistungsspektrum.

Tabelle 8: Mittelwerte, Standardabweichung, Schiefe, Minimum und Maximum (E004)

	Mittelwert	Stand. Abw.	Schiefe	Min	Max	N
C-Test (A)	48,39	12,23	-0,472	12	73	187
C-Test (B)	49,28	12,10	-0,464	13	74	187
LV	24,55	4,08	-0,945	9	30	187
HV	14,73	4,23	-0,113	3	24	187
SA				2	5	155
MA				2	5	145

In Tabelle 9 finden sich die Reliabilitäten der vier Subtests des TestDaF und des C-Tests differenziert nach den Auswertungsmethoden A und B. Für den C-Test, das Leseverstehen (LV) und das Hörverstehen (HV) wurde Cronbachs

Alpha berechnet (jeweils $N = 187$), für die auf Schätzurteilen beruhenden Daten aus dem Schriftlichen Ausdruck (SA) und dem Mündlichen Ausdruck (MA) der gewichtete Kappa-Koeffizient von Cohen. Der gewichtete Kappa-Koeffizient ist ein Maß für die Interrater-Reliabilität, das sowohl berücksichtigt, wie oft übereinstimmend geurteilt wurde, als auch wie weit nicht-übereinstimmende Urteile auseinander liegen (vgl. Bortz, Lienert & Boehnke, 1990, S. 482ff.; Krauth, 1995, S. 54-62). Kappa sollte mindestens 0,70 betragen, um von einer „guten“ Übereinstimmung sprechen zu können (vgl. Bortz & Döring, 2002, S. 277).

Tabelle 9: Reliabilitäten (E004)

C-Test (A)*	C-Test (B)*	LV*	HV*	SA**	MA**
,842	,836	,78	,76	,35	,44

* Cronbachs Alpha

** gewichteter Kappa-Koeffizient nach Cohen

Die Werte für Cronbachs Alpha sind für einen C-Test mit insgesamt nur 80 Lücken, der maximal 20 Minuten in Anspruch nimmt, sehr zufriedenstellend. Erwartungsgemäß ist die Reliabilität im Fall der Auswertungsmethode B wegen des systematischen varianzmindernden Effekts der Wertung von Orthografiefehlern als korrekt reduziert, allerdings nur minimal. Der Wert für das Leseverstehen stellt möglicherweise eine leichte Überschätzung der tatsächlichen Reliabilität dar, da er auf der Basis von z.T. lokal stochastisch abhängiger Einzelitems berechnet wurde.[8]

Bei den in Tabelle 9 angegebenen Kappa-Werten für MA und SA handelt es sich um Mittelwerte (vgl. Krauth, 1995, S. 62). Der Wert für SA ist das Mittel der Kappa-Koeffizienten von vier Rater-Paaren (Spannweite: 0,10 bis 0,52), die insgesamt 165 Probanden beurteilt haben. Der Wert für MA beruht auf den Kappa-Koeffizienten von sechs Rater-Paaren (Spannweite: 0,17 bis 0,83), die insgesamt 145 Probanden beurteilt haben.[9]

---

[8] Beim Hörverstehen als online-Prozess spielt wegen der Begrenztheit des Arbeitsgedächtnisses die serielle Abhängigkeit der Items vermutlich eine deutlich geringere Rolle als beim Leseverstehen, da hier im Gegensatz zum Hörverstehen eine gezielte mehrfache Verarbeitung bestimmter Textinformationen möglich ist.

[9] Unterschiede in den Stichprobenumfängen sind dadurch zu erklären, dass nicht alle Probanden alle Testteile bearbeitet haben. Im Fall des SA wurden zudem einige Probanden von mehr als einem Rater-Paar beurteilt. Als Folge beruht der mittlere Kappa-Koeffizient nicht auf 165, sondern auf 180 Urteilen. Bei den weltweiten TestDaF-Prüfungen vom 26.04.2001 und 18.10.2001 ergaben sich im Übrigen nur geringfügig höhere Werte für das mittlere Kappa, so z.B. für die Prüfung vom 18.10.2001: 0,37 (SA; 400 Beurteilungen) und 0,50 (MA; 387 Beurteilungen). Auch die Streuungen und Spannweiten waren ähnlich.

Die Interrater-Reliabilitäten sind insb. im Fall des Schriftlichen Ausdrucks unbefriedigend.[10] Zur Zeit wird geprüft, ob mit Hilfe eines überarbeiteten Kriterienrasters und einer entsprechenden Schulung der Rater beim SA eine höhere Reliabilität erreicht werden kann (vgl. Arras & Grotjahn, 2002). Beim MA scheint es sich dagegen nicht um ein gleichermaßen generelles Reliabilitätsproblem zu handeln, das z.B. auf eine mangelnde Eindeutigkeit des Kategorienrasters zurückzuführen ist. Wie der beobachtete maximale Kappa-Wert von 0,83 andeutet, kann hier sicherlich über eine weitere Beurteilerschulung eine hinreichende Reliabilität erzielt werden.

3.4. Interkorrelationen zwischen C-Test, LV, HV, SA und MA

In Tabelle 10 sind die Korrelationen zwischen C-Test, LV, HV, SA und MA aufgeführt – und zwar in der oberen Dreiecksmatrix die Werte für den Spearman-Rang-Korrelationskoeffizienten und in der unteren Dreiecksmatrix die Werte für Kendalls Tau-b. Vor der Berechnung der Koeffizienten wurden jeweils Rangplatz-Transformationen vorgenommen. Da relativ viele Bindungen (gleiche Rangplätze) vorlagen, erfolgte die Berechnung der Spearman-Korrelation mit Hilfe der Formel für den Produkt-Moment-Korrelationskoeffizienten.

Tabelle 10: Spearman-Rang-Korrelationen (obere Dreiecksmatrix) und Kendalls Tau-b (untere Dreiecksmatrix) zwischen TestDaF-Subtests und C-Test (E004; C-Test-Auswertungsmethode A)

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

Anmerkung: Alle Korrelationen sind auf dem Niveau von 0,01 (2-seitig) signifikant.

[10] Da im Falle von divergierenden Beurteilungen (bei ca. 30-40%) eine Drittbeurteilung vorgenommen wurde, dürften die angegebenen Kappa-Werte die tatsächliche Reliabilität von SA und MA eher unterschätzen. Einen Hinweis darauf geben auch die relativ hohen Korrelationen zwischen C-Test und SA und MA (vgl. Tab. 10).

Erwartungsgemäß sind die Werte für Kendalls Tau-b durchgängig niedriger als die entsprechenden Werte für den Spearman-Koeffizienten. Die Muster entsprechen sich jedoch weitgehend. Der C-Test korreliert hoch mit allen vier TestDaF-Subtests und eignet sich damit gut als Ankertest.[11] 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 (zugrundeliegende) 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 (vgl. allerdings Wright, 1990), ist dies ein sehr wichtiges Resultat.

Berücksichtigt man zudem, dass MA und SA nur eine geringe Reliabilität aufweisen (vgl. Tabelle 9), dann dürften die Korrelationen des C-Tests mit MA und SA in Tabelle 10 die „wahren“ Korrelationen eher unterschätzen. Bei einer Verbesserung der Reliabilität von MA und SA ist deshalb zu erwarten, dass sich auch die entsprechenden Korrelationen mit dem C-Test noch erhöhen.[12]

3.5. Der C-Test als Screening-Test für TestDaF?

Angesichts der relativ hohen Korrelationen des C-Tests mit SA und MA stellt sich die Frage, ob der C-Test nicht als ökonomischer Screening-Test im Hinblick auf die wenig ökonomischen Testteile MA und SA eingesetzt werden kann. Könnte man z.B. hinlänglich verlässlich von einer niedrigen C-Test-Leistung auf eine niedrige Leistung im SA oder MA schließen, dann wäre u.a. folgendes Vorgehen denkbar: Man entwickelt einen weltweit frei zugänglichen, web-basierten C-Test (vgl. Röver, 2002) und empfiehlt Interessenten, sich nur dann zur TestDaF-Prüfung anzumelden, wenn ein bestimmter Punktwert im C-Test erreicht worden ist.

---

[11] Zusätzlich haben wir für den Zusammenhang zwischen C-Test und SA bzw. MA das Korrelationsverhältnis Eta mit SA und MA als kategorialer unabhängiger Variablen berechnet. Es ergab sich für MA 0,640 und für SA 0,693.

[12] Wegen der Verwendung von Cohens Kappa zur Abschätzung der Reliabilität von MA und SA ist eine Berechnung von messfehlerbereinigten Korrelationen auf der Basis der üblichen, auf dem Produkt-Moment-Korrelationskoeffizienten beruhenden Formeln für Minderungskorrekturen nicht möglich. Berechnet man trotz der Tatsache, dass Cohens Kappa keine Schätzung der Produkt-Moment-Korrelation ist und auch nicht im Sinne von Varianzanteilen interpretiert werden kann, minderungskorrigierte Korrelationen, so führen diese zu Werten größer als 1. Dies kann zugleich als Hinweis darauf gedeutet werden, dass die erhaltenen Kappa-Werte die tatsächliche Reliabilität der Beurteilung der schriftlichen und mündlichen Ausdrucksfähigkeit unterschätzen (vgl. auch Lord & Novick, 1968, S. 138).

Der Zusammenhang zwischen C-Test-Leistung und dem Schriftlichen bzw. Mündlichen Ausdruck ist in den Tabellen 11 und 12 dargestellt.

Tabelle 11: Zusammenhang zwischen C-Test und Schriftlichem Ausdruck (E004; C-Test-Auswertungsmethode A)

	SA				Gesamt
C-Test	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
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

Fortsetzung Tabelle 11

	SA				Gesamt
C-Test	unter TDN 3	TDN 3	TDN 4	TDN 5	
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

Tabelle 12: Zusammenhang zwischen C-Test und Mündlichem Ausdruck (E004; C-Test-Auswertungsmethode A)

	MA				Gesamt
C-Test	unter TDN 3	TDN 3	TDN 4	TDN 5	
12		1			1
15		1			1
18		3			3
23	1				1
24	1	1			2
25		2			2
26			1		1
28		1			1
29		1			1
30		1			1
31	1	2	1		4
32			1		1
33		1			1
34	1		1	1	3
36		2	1	1	4
37		3	2		5
38	1	2	1		4
39	1	6	1		8
40		3	1		4
41			4		4
42		4	3	1	8

Fortsetzung Tabelle 12

	MA				Gesamt
C-Test	unter TDN 3	TDN 3	TDN 4	TDN 5	
43			4		4
44		1	1		2
45	1				1
46		1	4	1	6
47			2		2
48		2	3	1	6
49		1	3	1	5
50		1	4	2	7
51		1	3	1	5
52		1	4		5
53		1	2	1	4
54			1	1	2
56		1	2	1	4
57			4	4	8
58			1		1
59			1	1	2
60			1		1
61			2	1	3
62			1		1
63				3	3
64			1	1	2
65			2		2
67				1	1
69				3	3
70				3	3
71			1		1
73				1	1
Gesamt	7	44	64	30	145

Sowohl beim Schriftlichen Ausdruck als auch beim Mündlichen Ausdruck zeigen die Daten in den Tabellen 11 und 12 ein charakteristische Muster: 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. Allerdings ist dieses Muster weit deutlicher im Fall des Schriftlichen Ausdrucks ausgeprägt. Hier findet sich z.B. unter den 27 Probanden mit einem C-Test-Wert kleiner oder gleich 37 – und damit bei einem substanziellen Anteil (25,5%) – kein einziges Mal das Urteil TDN 4 oder TDN 5. Beim Mündlichen Ausdruck lässt sich dagegen die Leistung von nur relativ wenige Probanden hinreichend verlässlich vorhersagen. In Einzelfällen gibt es hier sogar relativ drastische Ab-

weichungen: So kommt es z.B. vor, dass ein C-Test-Wert von 45 der Beurteilung „unter TDN 3“ entspricht, während ein C-Test-Wert von 34 sowohl der Beurteilung „unter TDN 3 als auch TDN 4 als auch TDN 5 entsprechen kann. Solche Beispiele relativieren deutlich die beobachtete Korrelation von 0,64 zwischen C-Test und MA.

Insgesamt liefert die vorliegende Studie erste Hinweise dafür, dass der C-Test als Screening-Test eingesetzt werden könnte – allerdings nur sehr eingeschränkt im Hinblick auf die mündliche Leistung. In Bezug auf den Schriftlichen Ausdruck könnte die Vorhersagekraft des C-Tests sicherlich noch verbessert werden – und zwar vermutlich deutlich durch eine Erhöhung der Interrater-Reliabilität bei der Beurteilung der schriftlichen Leistungen und zumindest geringfügig durch Verlängerung des C-Tests um einen weiteren Text.[13]

4. Zusammenfassung

Die vorliegende Untersuchung hat folgende Hauptergebnisse im Hinblick auf die Entwicklung von deutschen C-Tests und den Einsatz von C-Tests im Rahmen von TestDaF erbracht:

1. Es sollten möglichst sowohl Originallösungen als auch akzeptable Varianten als korrekt gewertet werden. Bei nicht sehr weit fortgeschrittenen Lernern sind akzeptable Varianten allerdings relativ selten und Bewertungsunterschiede haben deshalb nur einen geringen Effekt.
2. Gehört die orthografische Kompetenz mit zum messenden Konstrukt, sind orthografisch falsche Originale und Varianten als falsche Lösung zu werten.
3. Orthografiefehler kommen erwartungsgemäß vor allem bei weniger fortgeschrittenen Lernern vor. Hier kann es je nach Wertung von Orthografiefehlern zu deutlich unterschiedlichen Punktwerten und Rangfolgen der Probanden kommen.
4. Durch gründliche Vorerprobung und Revision lassen sich C-Tests entwickeln, in denen nur wenige Orthografiefehler vorkommen und die Wertung von Orthografiefehlern als korrekt kaum Auswirkungen auf die Gesamtpunktwerte hat. In einem solchen Fall kann es aus Ökonomiegründen gerechtfertigt sein, orthografisch falsche Lösungen als „nicht korrekt“ zu werten, auch wenn die orthografische Kompetenz nicht zum messenden Konstrukt gehört. Allerdings kann auch bei einer nur relativ kleinen Zahl von or-

[13] Wie allerdings die Tabelle 11 verdeutlicht, ist ein hinreichend verlässliches Screening nur im Hinblick auf das Erreichen der Stufen TDN 4 und TDN 5 möglich. Setzt man einen Fehler 2. Art von 5% an, dann zeigen die Tafeln von Taylor und Russell zudem, dass nur bei ca. 10% der Probanden auf die Teilnahme am SA verzichtet werden kann.

thografisch falschen Lösungen die Wertung von Orthografiefehlern als „nicht korrekt“ zu beträchtlichen Rangplatzdifferenzen führen.

5. Der im Kontext von TestDaF entwickelte, aus vier Texten mit jeweils 20 Lücken bestehende C-Test ist hoch reliabel. Die Analysen auf der Basis des kontinuierlichen Ratingskalen-Modells von Müller (1999) belegen zudem die gelungene Rasch-Skalierung der vier Texte.
6. Der C-Test korreliert relativ hoch mit allen vier Subtests des TestDaF (Spearmans *r* zwischen 0,64 und 0,68). Bemerkenswert ist die hohe Korrelation von 0,64 mit dem Mündlichen Ausdruck.
7. Der C-Test scheint sich sowohl für die Kalibrierung der Lese- und Hörverstehensitems als auch für ein ökonomisches Screening schwächerer Probanden zu eignen. Ein solches Screening bietet sich vor allem im Hinblick auf den wenig ökonomischen Testteil „Schriftlicher Ausdruck“ an – z.B. auf der Basis einer freiwilligen Selbstevaluation mit Hilfe eines web-basierten, weltweit einsetzbaren und automatisch auswertbaren C-Tests.

Literaturverzeichnis

Andrich, David. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105-113.

Arras, Ulrike & Grotjahn, Rüdiger. (2002). *TestDaF: Aktuelle Entwicklungen*. Vortrag auf der AKS-Tagung in Chemnitz, 28. Februar 2002 [erscheint in den Tagungsberichten].

Association of Language Testers in Europe (ALTE). (1998). *ALTE Handbuch europäischer Sprachprüfungen und Prüfungsverfahren*. Cambridge: The University of Cambridge Local Examinations Syndicate.

Bolten, Jürgen. (1992). Wie schwierig ist ein C-Test? Erfahrungen mit dem C-Test als Einstufungstest in Hochschulkursen Deutsch als Fremdsprache. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 1, S. 193-203). Bochum: Brockmeyer.

Bortz, Jürgen, Lienert, Gustav A. & Boehnke, Klaus. (1990). *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer-Verlag.

Bortz, Norbert & Döring, Nicola. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschafler* (3. Aufl.). Berlin: Springer-Verlag.

Cummins, Jim. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In Charlene Rivera (Hrsg.), *Language proficiency and academic achievement* (S. 2-19). Clevedon, England: Multilingual Matters.

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.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Göbel, Silke, Müller, Hans & Moosbrugger, Helfried. (1999). *Überprüfung einer NEO-FFI-Version mit kontinuierlichen Items*. Poster präsentiert auf der 4. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie, Universität Leipzig.

Grießhaber, Wilhelm. (1998). Der C-Test als Einstufungstest. In Karl-Heinz Eggensperger & Johann Fischer (Hrsg.), *Handbuch Unicert* (S. 153-167). Bochum: AKS-Verlag.

Grotjahn, Rüdiger. (1989). Der C-Test im Bundeswettbewerb Fremdsprachen: Eignung und Probleme (Französisch). In Thomas Finkenstaedt & Konrad Schröder (Hrsg.), *Zwischen Empirie und Machbarkeit: Erstes Symposium zum Bundeswettbewerb Fremdsprachen* (S. 41-56). Augsburg: Universität.

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. (2001). *TestDaF: Theoretical Basis and Empirical Research.* Paper presented at the ALTE European Year of Languages Conference, 5 to 7 July 2001, Barcelona. [erscheint 2002 in den Tagungsberichten].

Grotjahn, Rüdiger & Allner, Burkhardt. (1996). Der C-Test in der Sprachlichen Aufnahmeprüfung an Studienkollegs für ausländische Studierende an Universitäten in Nordrhein-Westfalen. In Rüdiger Grotjahn (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* (Bd. 3, S. 279-342). Bochum: Brockmeyer.

Grotjahn, Rüdiger & Kleppin, Karin. (2001). TestDaF: Stand der Entwicklung und einige Perspektiven für Forschung und Praxis. In Karin Aguado & Claudia Riemer (Hrsg.), *Wege und Ziele: Zur Theorie, Empirie und Praxis des Deutschen als Fremdsprache (und anderer Fremdsprachen). Festschrift für Gert Henrici zum 60. Geburtstag* (S. 419-433). Baltmannsweiler: Schneider Verlag Hohengehren.

Kniffka, Gabriele & Üstünsöz-Beurer, Dörthe. (2001). TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassettengesteuerten Testformats. *Fremdsprachen Lehren und Lernen, 30*, 127-149.

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.

Krauth, Joachim. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Psychologie Verlags Union.

Lord, Frederic M. & Novick, Melvin R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, Geofferey N. & Wright, Benjamin D. (1997). The partial credit model. In Wim J. van der Linden & Ronald K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 101-121). New York: Springer.

Moosbrugger, Helfried. (1992). Testtheorie: Klassische Ansätze. In Reinhold S. Jäger & Franz Petermann (Hrsg.*), Psychologische Diagnostik: Ein Lehrbuch* (S. 310-322). Weinheim: Psychologie Verlags Union.

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.

Projektgruppe TestDaF. (2000). TestDaF: Konzeption, Stand der Entwicklung, Perspektiven. *Zeitschrift für Fremdsprachenforschung, 11*(1), 63-82.

Raatz, Ulrich & Klein-Braley, Christine. (1998). Gleichwertige Zertifikate – überall und immer oder Wie man UNICERT-Zertifikate mittels C-Tests kalibrieren kann. In Karl-Heinz Eggensperger & Johann Fischer (Hrsg.), *Handbuch Unicert* (S. 107-114). Bochum: AKS-Verlag.

Rost, Jürgen. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.

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

Wright, Frank. (1990). *Testing foreign language ability: Two approaches in theory and practice.* Unpublished M. Litt. Thesis, Trinity College, University of Dublin, Ireland.

Anhang

Ausgewählte C-Test-Texte

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.

Globale Erwärmung wird langsam katastrophal

Die Erdatmosphäre erwärmt sich deutlich schneller, als Experten bislang vermutet haben - mit möglicherweise verheerenden Folgen. Und d_____ Mensch kön_____ daran d_____ Hauptschuld tra_____ . Das ge_____ aus ei_____ Bericht d_____ zwischenstaatlichen Grem_____ für Klimaverä_____ der UNO her_____ , der a_____ Montag i_____ Schanghai veröffe_____ wurde. D_____ Wissenschaftler erwa_____ eine schne_____ und poten_____ katastrophalere glo_____ Erwärmung i_____ diesem Jahrh_____ . Bislang war nur mit einem Temperaturanstieg von 1 bis 3,5 Grad zwischen 1990 und 2100 gerechnet worden.

Weltkonferenz

1992 beschlossen auf der UN-Weltkonferenz für Umwelt und Entwicklung in Rio de Janeiro 170 Staaten die Agenda 21. Sie i_____ eine politi_____ Verpflichtung a_____ internationaler Eb_____ , die We_____ künftig nachh_____ zu entwi_____ . Dies so_____ weder a_____ Kosten d_____ Natur, and_____ Regionen o_____ anderer Mens_____ noch z_____ Lasten zukün_____ Generationen gesc_____ . Mit d_____ Agenda 21 wu_____ ein ganzhei_____ Ansatz v_____ Chancengleichheit und Gerechtigkeit aller Menschen untereinander gewählt.

Verwüstungsprozesse

Über ein Drittel des Landes der Erde ist Wüste. Ein gro_____ Teil i_____ seit d_____ Verbreitung d_____ Zivilisation z_____ Wüste gewo_____ . Der Pro_____ der Verwü_____ hat si_____ in d_____ letzten Jah_____ dramatisch beschl_____ und bedr_____ die Zuk_____ von vie_____ Millionen Mens_____ und Tie_____ . Diese Entwi_____ ist m_____ einem sta_____ Produktivitätsrückgang verbunden, denn je trockener ein Gebiet ist, desto weniger kann geerntet werden.