Zeitschrift für Interkulturellen Fremdsprachenunterricht
Didaktik und Methodik im Bereich Deutsch als Fremdsprache
ISSN 1205-6545 Jahrgang 21, Nummer 2 (Oktober 2016)

Die Vorhersage der Schwierigkeit deutscher C-Test-Texte: Untersuchungen am Beispiel des onDaF

Nicole Kaufmann
TestDaF-Institut
Universitätsstraße 134
44799 Bochum
E-Mail: nicole.kaufmann@rub.de

Abstract: C-Tests sind Lückentests zur schnellen und zuverlässigen Messung der allgemeinen Sprachkompetenz. Obwohl die Tests bereits hinsichtlich vieler verschiedener Aspekte untersucht wurden, konnte die Frage, warum manche C-Test-Texte schwieriger sind als andere, noch nicht hinreichend beantwortet werden. Im vorliegenden Beitrag wird ein Versuch unternommen, diese Frage im Hinblick auf zehn Texte zu beantworten, die im onDaF (Online-Einstufungstest Deutsch als Fremdsprache) Anwendung finden. Untersucht werden verschiedene Textcharakteristika sowie Fehler, die 202 Lerner[1] in einzelnen Lücken gemacht haben. Statistische und linguistische Analysen führen zu dem Schluss, dass keiner der untersuchten Parameter, und auch keine Kombination dieser, zuverlässig die Schwierigkeit deutscher C-Test-Texte voraussagen kann. Die Ergebnisse stützen das Argument, dass die erfolgreiche Bearbeitung eines C-Tests umfassende Kompetenzen in verschiedensten Bereichen der Zielsprache erfordert.

C-tests are gap-filling tests for the quick and reliable assessment of global language competence. Although C-tests have already been researched with regard to many different aspects, the question why some C-test texts are more difficult than others remains largely unanswered. In the present paper, an approach is made to answer this question with regard to ten texts used in the onDaF (Online Placement Test of German as a Foreign Language). Different text characteristics are studied as well as mistakes in individual gaps made by 202 learners. Statistical and linguistic analyses lead to the conclusion that none of the parameters taken into account, or a combination of those, reliably predict the difficulty of German C-test texts. The findings support the argument that taking a C-test successfully requires a proficient knowledge of many different aspects of the target language.

Schlagwörter: C-Test, Sprachtesten, Schwierigkeitsgenerierende Parameter, Rasch-Modell, Schwierigkeitsanalyse, Online-Einstufungstest; C-test, language testing, difficulty generating parameters, Rasch model, difficulty analysis, online placement test

1. Einleitung

Gegenstand der Untersuchung sind C-Test-Texte und deren Schwierigkeit, die für den Online-Einstufungstest Deutsch als Fremdsprache (onDaF) verwendet werden. Beim onDaF handelt es sich um einen vollständig internetbasierten C-Test, bestehend aus acht kurzen Texten, die aufeinanderfolgend auf dem Bildschirm des Testteilnehmers erscheinen (vgl. Eckes 2010). Zur Bearbeitung jedes Textes haben die Teilnehmer maximal fünf Minuten Zeit. Unmittelbar nach Ende des Tests erhalten sie ihr Ergebnis in Form des erreichten Niveaus (A2-C1) analog zum GER.

Zur Weiterentwicklung des onDaF erstellt das TestDaF-Institut regelmäßig neue Texte, die in eine inzwischen mehr als 400 Texte umfassende Itembank aufgenommen werden, sofern sie den Qualitätsansprüchen genügen. Vor der Verwendung dieser Texte im Test finden in Deutschland und an ausländischen Institutionen Erprobungen statt. Mit Erprobung sind hier Testläufe gemeint, bei denen Gruppen von Probanden die Texte als Papierversion ausfüllen. Die Rahmenbedingungen (Bearbeitungszeit, Aufsicht durch Prüfer etc.) entsprechen dabei denen im Online-Test. Die Probanden weisen ähnliche Charakteristika (Alter, Sprachniveau, etc.) auf wie die Zielgruppe, die den onDaF ablegt.

Jeder Proband, der an einer Erprobung teilnimmt, erhält ein Set aus zehn Texten. Die Reihenfolge der Texte in einem Set, beginnend mit dem leichtesten, entscheidet sich anhand von Experteneinschätzungen. In der Auswertung der Erprobungen zeigt sich häufig, dass die intuitiv geschätzte Schwierigkeit eines Textes nicht mit dessen tatsächlicher Schwierigkeit übereinstimmt. Dies kann einen negativen Einfluss auf die Motivation und Konzentration der Probanden haben, da es möglich ist, dass sie bereits mit einem der ersten Texte überfordert sind. Um die Texte nach ihrer tatsächlichen Schwierigkeit

sortieren zu können, bedarf es Methoden, die Schwierigkeit zuverlässig vorhersagen zu können. Bis heute gibt es jedoch keinen methodischen Ansatz, mit dem dies möglich ist.

Das Wissen über schwierigkeitsgenerierende Parameter könnte die Arbeitsprozesse bei der Testerstellung beschleunigen, wenn bspw. Texte für ein bestimmtes Niveau benötigt werden. In Erprobungen könnte man zudem der Forderung nach steigender Schwierigkeit aufeinanderfolgender Texte besser nachkommen, um potenziell qualitativ hochwertigere Ergebnisse zu erzielen.

Ziel der vorliegenden Untersuchung ist es, Erkenntnisse darüber zu gewinnen, wie sich die Schwierigkeit deutscher C-Test-Texte vorhersagen lässt. Dazu wird zunächst ein Überblick über Studien gegeben, die sich direkt oder indirekt mit der Schwierigkeit von C-Test-Texten und ihrer Vorhersage befassen. Bevor überprüft wird, inwieweit die in der Literatur dargestellten Prädiktoren für die onDaF-Texte geeignet sind, werden die für die vorliegende Studie verwendeten Texte auf ihre Tauglichkeit für den Einsatz im onDaF hin untersucht. Dies stellt sicher, dass die Texte eine ausreichend hohe Qualität für den Online-Einstufungstest aufweisen. Im Anschluss folgen Untersuchungen auf Text- und Lückenebene, welche auf die einzelnen Prädiktoren Bezug nehmen. Die Studie schließt mit einer Zusammenfassung und der Diskussion der gewonnenen Erkenntnisse.

2. Der C-Test

Der C-Test ist ein aus mehreren kurzen Lückentexten bestehender Test zur Messung allgemeiner Sprachkompetenz in Fremd-, Zweit- oder Muttersprachen (vgl. Grotjahn, Klein-Braley & Raatz 2002). Die Aufgabe eines Testteilnehmers ist es, Lücken in vier bis acht Lückentexten zu füllen, d.h. die durch systematische Tilgung manipulierten Wörter korrekt zu rekonstruieren. Dabei stützt sich der C-Test auf das Prinzip der reduzierten Redundanz (vgl. u.a. Grotjahn 1992; Klein-Braley 1985a, 1997; Raatz & Klein-Braley 2002). Mit steigender Sprachkompetenz sollte es für Testteilnehmer einfacher werden, diese Lücken zu füllen, da sie lernen, von der Redundanz der natürlichen Sprache Gebrauch zu machen.

3. Aktueller Forschungsstand

Auf die Frage, wie die Schwierigkeit eines C-Test-Textes vorhergesagt werden kann, gibt es bislang keine zufriedenstellende Antwort. Ein möglicher und häufig genutzter Ansatz besteht in der intuitiven Vorhersage der Textschwierigkeit durch erfahrene Lehrpersonen oder Testleiter (vgl. Klein-Braley 1985b, 1994). Obwohl dieses Verfahren relativ unproblematisch zu realisieren ist, verweist Klein-Braley darauf, dass diese Personen die Texte häufig als schwieriger einstufen, als sie tatsächlich sind.

Ein alternativer Ansatz zu den Expertenurteilen ist die Verwendung von Regressionsanalysen. Klein-Braley (1994) führte Regressionsanalysen mit deutschen und englischen C-Test-Texten für verschiedene Zwecke durch: die Entscheidung über die Akzeptanz oder Ablehnung von Texten, die Sortierung von Texten innerhalb eines Tests sowie die Entwicklung paralleler Tests. Die Autorin ermittelte verschiedene Textcharakteristika wie die Anzahl an Sätzen und Silben sowie die durchschnittliche Anzahl an Wörtern im Satz und errechnete die Korrelation mit der Schwierigkeit der Texte. Sie fand die höchsten Korrelationen zwischen der Schwierigkeit der Texte und den Werten für das Type-Token-Ratio (TTR) und der Anzahl an Sätzen (vgl. Klein-Braley 1994: 273). Für die von ihr untersuchten Texte lieferten die Regressionsanalysen zufriedenstellende Ergebnisse. Bolten (1992) verwies in einer Untersuchung zur Schwierigkeit von C-Test-Texten für Deutsch als Fremdsprache darauf, dass die von Klein-Braley (1984) entwickelten Regressionsgleichungen wenig hilfreich bei der Bestimmung der Schwierigkeit von C-Test-Texten seien, da sie „die Spezifik von C-Tests nur wenig berücksichtigen“ (Bolten 1992: 198). In seinen Analysen mit C-Tests fand er nur marginale Übereinstimmungen von erwarteten und beobachteten Werten. Bolten argumentierte, dass z.B. semantisch-syntaktische Verbindungen, Ellipsen oder Partizipialkonstruktionen zur Schwierigkeit eines Textes beitragen können (1992: 200).

Mit der Vorhersage der Schwierigkeit von niederländischen C-Test-Texten bzw. deren intrinsischen Schwierigkeit haben sich Anckaert & Beeckmans (1992) auseinandergesetzt. Sie untersuchten, inwieweit sich die Schwierigkeit der Texte anhand mehrerer Indikatoren für eine bekannte Population vorhersagen ließ, um zu evaluieren, ob die Texte eine intrinsische Schwierigkeit aufwiesen. Sie haben Werte für das TTR und die durchschnittliche Satzlänge (DSL) für zwei Tests mit je vier Texten errechnet, entnommen aus verschiedenen Zeitschriften. Zudem untersuchten sie den Grad der Spezialisierung der Texte. Hierfür ermittelten sie die Anzahl der Leser der Zeitschriften sowie Textcharakteristika, für deren Erhebung sie die untersuchten acht Texte von verschiedenen Personen auf Ordinalskalen sortieren ließen. Jene schätzten die Texte danach ein, wie „théoretique“, „formel“ und „didactique“ (Anckaert & Beeckmans 1992: 151) sie waren, und sortierten sie jeweils im Hinblick auf diese Kriterien. Die Autoren zogen bei ihren Analysen auch die verschiedenen Kompetenzniveaus der Probanden in Betracht. Sie stellten abschließend fest, dass die verwendeten Prädiktoren nicht als Indikatoren für die Vorhersagbarkeit der Schwierigkeit von niederländischen C-Test-Texten geeignet waren (Anckaert & Beeckmans 1992: 159).

Kaufmann, Nicole (2016), Die Vorhersage der Schwierigkeit deutscher C-Test-Texte: Untersuchungen am Beispiel des onDaF. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 21: 2, 111-126. Abrufbar unter http://tujournals.ulb.tu-darmstadt.de/index.php/zif/.

Im Rahmen der *DESI-Studie* (Deutsch Englisch Schülerleistungen International) fand ein C-Test Anwendung. Er wurde genutzt, um den allgemeinen Sprachstand der Schülerinnen und Schüler zu messen (vgl. Harsch & Schröder 2007). Vorab untersuchten die Autoren die Texte in einem theoretischen Ansatz. Zunächst stuften sie die Texte im Hinblick auf drei Merkmale auf der Textebene ein. Diese waren das Thema, das Textniveau und Lösungsstrategien, wobei sich das letztgenannte Merkmal „auf die Komplexität des Lösungsprozesses und der dabei anzuwendenden Strategien [bezog]" (Harsch & Schröder 2007: 216). Beim Textniveau waren die Ausprägungen z.B. „leicht", „mittel" und „schwer". Auf der Ebene der Lücken betrachteten die Autoren die Merkmale Sprachniveau, Semantik und Itemfokus (Harsch & Schröder 2007: 218). Folgende Merkmale sind zu nennen, für die verschiedene Schwierigkeitsstufen vermutet wurden: Zugang zum Thema, Wortschatz, Parataxe und Hypotaxe, Satzverknüpfungen, Zeitformen, Passiv, Rekurs auf Weltwissen und Interpolationstechnik, grammatikalische und textuelle Formen der Lücken, Semantik der Lücke sowie Fokus der Lücken (vgl. Harsch & Schröder 2007: 215-219).

Im Rahmen ihrer Arbeit zu mentalen Prozessen bei der Bearbeitung von C-Test-Texten ermittelte Stemmer (1991) die Schwierigkeit französischsprachiger Texte und analysierte, welche Gründe für eine höhere oder niedrigere Schwierigkeit vorliegen können. Die Ergebnisse legten nahe, dass die Verteilung von referentiellen und lexikalischen Kohäsionsmitteln Auswirkung auf die Schwierigkeit haben kann (vgl. Stemmer 1991: 251). Die Autorin stellte fest, dass referentielle Kohäsionsmittel in den Texten vorwiegend Strukturwörter waren, während Inhaltswörter über die Lexik Kohäsion herstellten. Demnach müssten Kandidaten weniger Schwierigkeiten bei der Erstellung referentieller Kohäsion haben, weil Strukturwörter tendenziell einfacher zu erkennen und zu rekonstruieren sind als Inhaltswörter. Diese Analyse und deren Schlüsse bezogen sich jedoch auf die unveränderten Texte, in denen noch keine Wörter getilgt worden waren.

Des Weiteren stellte die Autorin fest, dass die Schwierigkeit der von ihr untersuchten Texte u.a. von der Länge der Bedeutungseinheiten in einem Text abhing (vgl. Stemmer 1991: 327). Die Tatsache, dass die Länge von Bedeutungsabschnitten Einfluss auf die Schwierigkeit haben kann, ergibt besonders im Hinblick auf die Abhängigkeiten der Lücken untereinander Sinn. Würden Analysen ergeben, dass C-Test-Texte mit längeren Bedeutungsabschnitten tatsächlich stets schwieriger sind als solche mit kürzeren, könnte dies als Prädiktor für die Schwierigkeit gesehen werden.

Hastings (2002) analysierte Fehler von Teilnehmern eines englischen C-Tests. Sie suchte in Texten nach Mustern, die auf integrierte psycholinguistische Prozesse bei der Bearbeitung hinweisen können. Im Hinblick auf die Fehler stellte sie u.a. fest, dass hochfrequente Wörter leichter zu rekonstruieren sind und dass das Thema eines Textabschnitts die Rekonstruktion eines Wortes vereinfachen kann (vgl. Hastings 2002: 57). Zur Überprüfung dieser Hypothese verwies sie auf doppelt getilgte Wörter innerhalb eines Textes, die unterschiedlich leicht oder schwierig für die Probanden waren. Auch kommt sie zu dem Schluss, dass syntaktische Effekte die Wahl eines Wortes beeinflussen können. Demnach ist zu erwarten, dass fortgeschrittene Lerner den gegebenen Kontext einer Lücke anders nutzen als schwächere Lerner (vgl. Hastings 2002: 59; Sigott 2004: 189). Des Weiteren argumentierte sie, dass die Flexionsmorphologie die Schwierigkeit eines getilgten Wortes beeinflussen kann (vgl. Hastings 2002: 65; Grießhaber 1998).

Wie die vorangegangene Darstellung des aktuellen Forschungsstandes zeigt, sind noch keine geeigneten Instrumente für die Vorhersage der Schwierigkeit von C-Test-Texten vorhanden. Dass es Aspekte auf der Textebene sowie auf der Wortebene gibt, die die Schwierigkeit einer Lücke bestimmen können, erscheint jedoch plausibel. Betrachtet man die genannten schwierigkeitsbestimmenden Aspekte sowie weitere Merkmale der deutschen Sprache, die allgemein als potenziell schwierig für Deutschlerner gelten, ergeben sich folgende, potenzielle Schwierigkeitsprädiktoren für einzelne Wörter und Texte:

- Type-Token-Ratio & durchschnittliche Satzlänge,
- Frequenz von Wörtern im Sprachgebrauch,
- Ellipsen,
- Partizipialkonstruktionen,
- Kenntnis von getilgten Wörtern, Anzahl möglicher Ergänzungen,
- Der bestimmte Artikel,
- Flexion der Nomina,
- Bedeutungseinheiten im C-Test-Text und deren Länge (= Kontext),
- Gebrauch von Präpositionen,
- Nicht-kontinuierliche Abhängigkeiten zwischen Satzteilen, z.B. Partikelverben,
- Wortbildung durch Präfixe und Suffixe,
- Verwendung des Passiv,
- Nominalisierungen,
- Kompositabildung,
- Kohäsionsmittel in Form von Inhalts- oder Strukturwörtern.

4. Untersuchung deutscher C-Test-Texte

Die vorliegende Untersuchung befasst sich anhand einiger der oben genannten Text- bzw. Wortmerkmale mit der Frage, ob sich die Schwierigkeit von deutschen C-Test-Texten vorhersagen lässt und welche der Textmerkmale sich hierfür am besten eignen. Gegenstand der Untersuchung sind mehrere C-Test-Texte, die zuvor für den Einsatz im onDaF erstellt wurden.[2]

Kaufmann, Nicole (2016), Die Vorhersage der Schwierigkeit deutscher C-Test-Texte: Untersuchungen am Beispiel des onDaF. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 21: 2, 111-126. Abrufbar unter http://tujournals.ulb.tu-darmstadt.de/index.php/zif/.

Diese durchliefen zunächst die am TestDaF-Institut üblichen Qualitätskontrollen und darauffolgend verschiedene Analysen sowohl auf der Textebene als auch auf der Wortebene. Im Fokus standen auf der Satzebene die Thematik, das Type-Token-Ratio und die durchschnittliche Satzlänge sowie das Verhältnis von getilgten Inhalts- zu getilgten Strukturwörtern. Auf der Wortebene waren es die Wortfrequenz und Fehler von 202 Probanden in einzelne Lücken des schwierigsten Textes, des leichtesten Textes und des Textes mit mittlerer Schwierigkeit.

Für die Qualitätskontrolle fand zunächst eine Vorerprobung der Texte im TestDaF-Institut statt. Dies ermöglichte eine erste Revision hinsichtlich potenzieller Rechtschreibfehler, akzeptabler Varianten o.Ä. Auch konnten so Einschätzungen von Experten zur potenziellen Schwierigkeit der neuen Texte gesammelt werden. Im Anschluss daran füllten DaF-Lerner in Erprobungen an Testzentren in Deutschland und im Ausland die Lückentexte aus, die sie in gedruckter Form erhielten. Die Ergebnisse dieser Erprobung waren Gegenstand der Analysen.

4.1. Probanden

Insgesamt 225 Personen haben die Texte des untersuchten Erprobungssets bearbeitet, wobei 23 Probanden (Pbn) den Test abgebrochen haben. Somit standen die Daten von 202 Personen für die Analysen zur Verfügung. Die Gruppe unterteilte sich in 134 Frauen (66,34 %) und 68 Männer (33,66 %), von denen 156 (77,23 %) Personen zwischen 19 und 29 Jahre alt waren (M = 25,57, SD = 10,02). Sieben Probanden waren älter als 40 Jahre, zwei Probanden machten keine Angabe zu ihrem Alter.

Insgesamt gaben die Probanden 37 verschiedene Herkunftsländer an. Die am häufigsten genannten waren China und Indonesien (je 28 Pbn), gefolgt von Italien (19 Pbn), Dänemark und Aserbaidschan (je 15 Pbn) sowie der Slowakei (10 Pbn). Sechs Probanden machten keine Angabe zu ihrem Heimatland.

Drei der Probanden waren zum Zeitpunkt der Erprobung Schüler oder an einem Studienkolleg eingeschrieben, sieben arbeiteten in einem nichtakademischen Beruf. Zudem machten 29 Probanden keine Angabe zu ihrer beruflichen Situation. Die übrigen 163 Probanden (80,69 %) waren während der Erprobung als Studenten an einer Hochschule eingeschrieben, wobei sie insgesamt 31 verschiedene Fachrichtungen angaben. Das am häufigsten studierte Fach bildete die Germanistik bzw. Deutsch als Fremdsprache (49 Pbn; 24,26 %). Dem folgten Kommunikations- und Medienwissenschaften (13 Pbn; 6,44 %), Informatik (12 Pbn; 5,94 %) und Wirtschaftswissenschaften (10 Pbn; 4,95 %).

4.2. Statistische Analysen

Die Analyse der erprobten Texte erfolgte nach dem diskreten Ratingskalenmodell (vgl. Andrich 1978; Eckes 2006, 2007). Dieses gehört zur Klasse der Rasch-Modelle, in denen von einer lokalen stochastischen Unabhängigkeit ausgegangen wird (vgl. Bühner 2011: 485). Das bedeutet, dass bei konstanter Fähigkeit einer Person die Wahrscheinlichkeit, dass sie ein Item korrekt löst, unabhängig von der Wahrscheinlichkeit ist, dass sie ein anderes Item korrekt löst. Somit entspricht die Wahrscheinlichkeit, alle Items eines Tests korrekt zu beantworten, dem Produkt der Wahrscheinlichkeiten für die korrekte Lösung der einzelnen Items.

Bei C-Tests ist es aufgrund des Konstrukts naheliegend, dass keine lokale Unabhängigkeit besteht, sondern eine mehr oder weniger starke Abhängigkeit zwischen aufeinanderfolgenden Lücken bzw. den Lücken innerhalb eines Textes vorherrscht, sowohl auf inhaltlicher als auch auf sprachlicher Ebene (vgl. Eckes 2007; Eckes & Baghaei 2015). Bei der Auswertung der Antworten in C-Tests werden daher nicht die einzelnen Lücken als einzelne Items betrachtet, sondern jeweils ein Text als Item mit Werten von 0 bis 20 (bei zwanzig zu füllenden Lücken).

Tab. 1: Deskriptive Statistiken der Texte

Text	*N*	Min	Max	*M*	*SD*
1	202	0	20	8,76	4,55
2	202	1	20	11,19	4,29
3	202	0	20	7,79	4,60
4	202	0	20	6,75	4,33
5	202	1	20	11,19	4,61
6	202	0	19	7,90	3,95
7	202	0	20	7,31	4,32
8	202	1	20	9,38	4,45
9	202	0	20	10,30	4,19
10	202	0	20	8,38	4,46

Anmerkungen: N = Anzahl der Probanden; Min = erreichte Mindestpunktzahl; Max = erreichte Höchstpunktzahl; M = Mittelwert; SD = Standardabweichung.

Tab. 1 fasst verschiedene Kennwerte bezüglich der Antworten der Probanden zusammen. Im Mittel haben die Probanden 6,75 bis 11,19 Lücken korrekt ergänzt. Bei einigen Texten kam es vor, dass Probanden keine Lücke richtig füllen konnten, während bei neun von zehn Texten mindestens einmal die Höchstpunktzahl zwanzig erreicht wurde. Die Standardabweichung (*SD*) liegt zwischen 3,95 und 4,61.

In Tab. 2 sind weitere statistische Werte der Texte zusammengefasst. Der zentrale Wert für die empirischen Untersuchungen ist die Textschwierigkeit (Schwierigkeitsmaß). Wie die Rasch-Analyse zeigt, stimmen die intuitiv geschätzten Schwierigkeiten in keinem der zehn Fälle mit den errechneten überein. In der Tabelle sind die Daten nach der Schwierigkeit sortiert. So stellte Text 4 mit einem Wert von 0,53 für die Probanden den schwierigsten Text dar, während der darauffolgende Text 5 mit -0,55 der leichteste war. Der Standardfehler liegt bei 0,03 bis 0,04.

Die Werte für Infit und Outfit haben einen Erwartungswert von 1 und können zwischen 0 und +∞ liegen (vgl. u.a. Linacre 2002; Wright & Linacre 1994). Sie geben an, wie gut die Daten anhand des Modells vorhersagbar sind. Als akzeptable Werte wurden von Linacre (2002) 0,5 bis 1,5 angegeben. Strengere Werte finden sich bei Wright & Linacre (1994). Hier reichen die Spannen von 0,7 bis 1,3 sowie von 0,8 bis 1,2. Die untersuchten Texte entsprechen mit Infit-Statistiken von 0,79 bis 1,13 und Outfit-Statistiken von 0,80 bis 1,14 allen diesen Richtwerten. Diese Ergebnisse stützen die Annahme der Eindimensionalität der Texte im untersuchten Erprobungsset, d.h. die Annahme, dass „die Daten durch eine einzige latente Dimension zu erklären [sind]“ (Eckes 2010: 160).

Tab. 2**:** Probabilistische Itemkennwerte

Text	Schwierigkeitsmaß	Standardfehler	Infit	Outfit
4	0,53	0,04	0,88	0,87
7	0,39	0,04	1,13	1,14
3	0,26	0,04	0,92	0,94
6	0,24	0,04	0,83	0,85
10	0,12	0,03	0,99	1,01
1	0,02	0,03	1,10	1,07
8	-0,12	0,03	0,82	0,81
9	-0,34	0,03	0,79	0,80
2	-0,55	0,03	1,00	0,96
5	-0,55	0,03	0,97	0,97

Anmerkung: Texte nach sinkender Schwierigkeit sortiert.

Im Rahmen der Rasch-Analyse wurde die Personenreliabilität errechnet, die ähnlich zu Cronbachs Alpha ist. Die möglichen Werte liegen zwischen 0 und 1, wobei ein geringer Wert besagt, dass ein hoher Messfehler vorliegt, während bei einem Wert von 1 der Messwert identisch mit dem wahren Wert ist und somit kein Messfehler vorliegt (vgl. Bortz & Döring 2010: 196).

Für die untersuchten Texte liegt der Wert des Reliabilitätskoeffizienten bei ,97, was auf eine sehr hohe Messgenauigkeit hinweist.

In Abb. 1 sind die Schätzwerte der Itemparameter (Schwierigkeit der untersuchten Texte) und der Personenparameter (die Personenfähigkeit) als zwei Facetten auf einer gemeinsamen Logit-Skala dargestellt (vgl. Bühner 2011). Beide Arten von Parametern können Werte zwischen -∞ und +∞ annehmen. Je nach Übereinstimmung der Verteilung von Fähigkeit und Schwierigkeit lässt sich schlussfolgern, ob die verwendeten Items, in diesem Fall die C-Test-Texte, für die Lernergruppe angemessen schwierig sind. Betrachtet man die erprobten Texte, so lässt sich feststellen, dass einige Texte eher schwierig für die Probanden waren. Für 58 Personen (28,71 %) waren selbst die leichtesten Texte zu anspruchsvoll. Es ist jedoch auch zu sehen, dass die Fähigkeit von 34 Probanden (16,83 %) teils deutlich über der Schwierigkeit des Textes mit dem höchsten Logit-Wert liegt. Da die Mittelwerte nicht weit auseinanderliegen, war die Schwierigkeit des Tests insgesamt angemessen.

Abb. 1: Grafische Darstellung der Textschwierigkeit und der Personenfähigkeit

```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +
                     |
                     |
                     |
                     |
                     |
                  .  |
    3                +
                  .  |
                     |
                  .  |
                  #  |
                     |
                     |
    2             .  +
                  .  |
                  . T|
                 .#  |
                  #  |
                  #  |
                  .  |
    1         .####  +
                .##  |
                 ## S|T
                ###  |  TEXT4
                ###  |S TEXT7
                 ##  |  TEXT3   TEXT6
              #####  |  TEXT10
    0        .#####  +M TEXT1
            #######  |  TEXT8
           ######## M|  TEXT9
        ###########  |S
        .##########  |  TEXT2   TEXT5
               .###  |T
                 .#  |
   -1          .###  +
             .##### S|
             ######  |
                 .#  |
                 .#  |
                ###  |
                  .  |
   -2             #  +
                    T|
                     |
                     |
                  #  |
                     |
                  .  |
   -3                +
               <less>|<frequent>
```

Anmerkungen: Ein # entspricht zwei Probanden; ein Punkt entspricht einem Proband.

Des Weiteren wurden die Daten auf Fälle von differenziellen Itemfunktionen (DIF) hin untersucht. Wie Badia, Prieto & Linacre (2002) schreiben, liegt ein Fall von DIF vor, wenn ein Item bei gleicher Fähigkeit für eine bestimmte Personengruppe aufgrund von Merkmalen wie dem Geschlecht, der Nationalität o.Ä. schwieriger zu lösen ist als für eine andere Gruppe. Am TestDaF-Institut üblich ist die Überprüfung auf DIF im Bereich des Geschlechts der Probanden (vgl. Eckes 2010).

Die Werte in der Spalte „DIF-Kontrast" in Tab. 3 geben die Differenz zwischen den Schwierigkeiten für die beiden Personenklassen wieder. Das negative Vorzeichen zeigt an, dass die Texte 2, 5 und 7 etwas leichter für die männlichen Probanden waren als für die weiblichen. Für einen auffälligen Unterschied zwischen den Klassen müsste dieser Kontrast jedoch bei mindestens 0,5 liegen. Auffällig ist der *t*-Wert von Text 2, der mit -2,10 über dem allgemein als Grenzwert angenommenen Absolutbetrag von 2,0 liegt. Hier liegt ein Hinweis auf einen bedeutsamen Unterschied vor. Aufgrund der multiplen Vergleiche muss jedoch die Bonferroni-Korrektur vorgenommen werden. Das korrigierte Alpha-Niveau liegt bei ,005 und somit unter dem *p*-Wert von Text 2 (,0373). Im Hinblick darauf, dass keine DIF-Differenz größer oder gleich 0,50 war, muss kein Text von den weiteren Analysen ausgeschlossen werden. Insgesamt zeigen die statistischen Analysen, dass alle Texte für den Einsatz im onDaF geeignet sind.

Tab. 3: Ergebnisse der DIF-Analyse für das Geschlecht der Probanden

Text	DIF-Kontrast (w-m)	DIF-Standardfehler	*t*	*df*	*P*
1	0,00	0,07	0,00	154	,9999
2	-0,15	0,07	-2,10	156	,0373
3	0,09	0,08	1,21	154	,2293
4	0,03	0,08	0,33	154	,7449
5	-0,03	0,07	-0,38	156	,7071
6	0,00	0,08	0,00	154	,9999
7	-0,04	0,08	-0,56	153	,5784
8	0,00	0,07	0,00	154	,9999
9	0,03	0,07	0,45	155	,6558
10	0,04	0,07	0,53	154	,6001

Anmerkungen: t = Werte der t-Statistik; df = Freiheitsgrade; p = Signifikanzwert.

4.3. Untersuchung von potenziell schwierigkeitsbestimmenden Merkmalen

Anhand der vorangegangenen Untersuchungen zeigte sich, dass alle Texte für den Einsatz im onDaF geeignet waren. Im nächsten Schritt wurden verschiedene Merkmale der Texte betrachtet, die einen Hinweis auf deren Schwierigkeit bzw. die Vorhersage der Schwierigkeit geben können. Ein schnell zu ermittelndes Merkmal war die Thematik der Texte, die Gegenstand der ersten Untersuchung wurde. Darauf folgten weitere Merkmale, die jedoch zunächst ermittelt bzw. errechnet werden mussten, wie das Type-Token-Ratio und die durchschnittliche Satzlänge. Auf der niedrigsten Untersuchungsebene fand schließlich die Betrachtung einzelner Lücken statt.

4.3.1. Thematik

Texte, die der Itembank des TestDaF-Instituts hinzugefügt werden, erhalten eine Zuordnung zu einer von über 20 vorgegebenen thematischen Kategorien. Anhand dieser Kategorien ließ sich überprüfen, inwieweit das Thema eines onDaF-Textes mit dessen Schwierigkeit zusammenhängt. In diesem Fall hätte die Betrachtung der neu erprobten Texte eine viel zu kleine Stichprobe abgedeckt. Die Untersuchung umfasste daher die Themen sämtlicher, zum Zeitpunkt der Untersuchung in der Itembank vorhandenen 426 Texte. Es gab Themen, die alle Niveaustufen abdeckten, aber auch mehrere, die auf zwei oder drei Niveaus beschränkt waren. Für 16 von 24 Themen gab es in der Itembank mindestens einen Text pro Niveaustufe. Bei den übrigen acht ist jedoch fraglich, ob es an der Schwierigkeit des Themas lag, dass nur gewisse Stufen vorkamen, oder an der relativ geringen Zahl an Texten, die vorhanden waren. So standen bspw. 40 Texten zum Thema „Gesellschaft" nur vier Texte zum Thema „Interkulturelle Kommunikation" gegenüber. Einer Kategorie waren sogar nur drei Texte zugeordnet, so dass anhand dieser gar nicht alle Niveaustufen abgedeckt werden konnten.

Betrachtet man nur die Texte, die alle Niveaustufen abdeckten und eine mittlere Anzahl an Texten aufwiesen (15 bis 28 Texte), lässt sich tendenziell erkennen, welche thematischen Kategorien eine höhere Schwierigkeit aufweisen. Bei den Themengebieten „Medizin" und „Bildung allgemein" fand sich eine relativ ausgewogene Verteilung mit durchschnittlich 4,75 bzw. 6,5 Texten pro Niveaustufe. Bei anderen Gebieten war die Verteilung hingegen sehr ungleichmäßig. So war beim

Thema „Geschichte“ die größte Anzahl an Texten auf dem Niveau B2 (zehn Texte im Vergleich zu zwei Texten auf A2) zu finden, während es sich bei der Kategorie „Familie“ umgekehrt verhielt (zehn Texte auf A2 im Vergleich zu einem Text auf B2). Eine mögliche Erklärung hierfür liegt darin, dass Lerner bereits früh mit dem Thema Familie konfrontiert werden, was es ihnen ermöglicht, relativ problemlos Texte zu diesem Thema zu bearbeiten. Im Gegensatz dazu wird das Thema Geschichte erst später im Unterricht behandelt.

Zusammenfassend lässt sich sagen, dass es Themen gibt, die anhand der Daten als tendenziell schwieriger eingestuft werden können, da mehr der Texte auf die Niveaus B2 und C1 entfallen (bspw. „Psychologie“ und „Gesellschaft“), während in anderen Bereichen die unteren Niveaus A2 und B1 stärker abgedeckt werden (bspw. „Studium“ und „Familie“). Anhand der vorliegenden Texte kann jedoch nicht argumentiert werden, dass Items, die auf bestimmte Themenkategorien entfallen, immer schwierig oder leicht für die Probanden zu lösen sind.

4.3.2. Type-Token-Ratio und durchschnittliche Satzlänge

Wie zuvor erwähnt wurden von Anckaert & Beeckmans (1992) und Klein-Braley (1984) Werte für das TTR und die DSL mehrerer Texte errechnet. Der TTR-Wert gibt Auskunft über das Verhältnis von unterschiedlichen Wörtern zur Gesamtheit der Wörter eines Textes und kann Werte zwischen 0 und 1 annehmen. Im Deutschen spricht man zuweilen auch von einem *Diversifikationsquotienten* (vgl. bspw. Wirtz 2013), der Auskunft über die Komplexität von Texten gibt. Je höher dieser Quotient ist, desto mehr unterschiedliche Wörter finden sich in einem Text. Es ist zu erwarten, dass ein Text mit hohem TTR-Wert schwieriger ist als einer mit einem niedrigen. Ähnliches gilt für die DSL. Diese errechnet sich durch die Division der Gesamtzahl an Wörtern eines Textes durch dessen Anzahl an Sätzen und gibt Auskunft über die syntaktische Komplexität (vgl. Anckaert & Beeckmans 1992).

Tab. 4: Type-Token-Ratio und durchschnittliche Satzlänge der untersuchten Texte

Text	TTR	DSL	Schwierigkeitsmaß
5	0,86	10,83	-0,55
2	0,85	12,17	-0,55
9	0,79	14,20	-0,34
8	0,87	14,75	-0,12
1	0,94	13,25	0,02
10	0,87	11,20	0,12
6	0,92	17,75	0,24
3	0,85	15,40	0,26
7	0,87	12,00	0,39
4	0,91	16,00	0,53

Anmerkungen: Texte nach steigender Schwierigkeit sortiert. TTR = Type-Token-Ratio; DSL = Durchschnittliche Satzlänge.

Tab. 4 fasst beide Werte für die Texte der vorliegenden Untersuchung zusammen. Um weitere Informationen darüber zu gewinnen, inwieweit die Indikatoren TTR und DSL zur Vorhersage der Schwierigkeit der deutschen C-Tests herangezogen werden können, wurde die bivariate Korrelation (vgl. Bortz & Döring 2010) mit der in den vorbereitenden Analysen errechneten Schwierigkeit berechnet. Der Pearson-Korrelationskoeffizient kann Werte zwischen -1 und +1 annehmen und stellt den linearen Zusammenhang zweier Variablen dar. Das Vorzeichen zeigt dabei die Richtung des Zusammenhangs an, die Größe des Wertes dessen Stärke (vgl. Bortz & Döring 2010).

Die Korrelation von DSL und Schwierigkeit liegt bei ,444 (p = ,198), die von TTR und Schwierigkeit bei ,496 (p = ,145), ohne dass die Korrelationen signifikant sind. Dass kein eindeutiger Zusammenhang festgestellt werden kann, zeigt sich auch im Streudiagramm. Abb. 2 zeigt den Zusammenhang zwischen Schwierigkeit und TTR-Werten. Es lässt sich ein tendenziell positiver Zusammenhang zwischen den beiden Prädiktoren erkennen, wobei dieser auch auf Zufall zurückzuführen sein kann. Problematisch ist auch, dass zwei Werte aus der erwarteten elliptischen Form ausbrechen. Zudem liegt die Irrtumswahrscheinlichkeit mit ,145 deutlich über dem üblicherweise zugrunde gelegten Wert von ,05. Das bedeutet, dass kein statistisch signifikanter Wert vorliegt. Für die Schwierigkeit und die DSL-Werte ist kein Zusammenhang zwischen den Variablen erkennbar (s. Abb. 3).

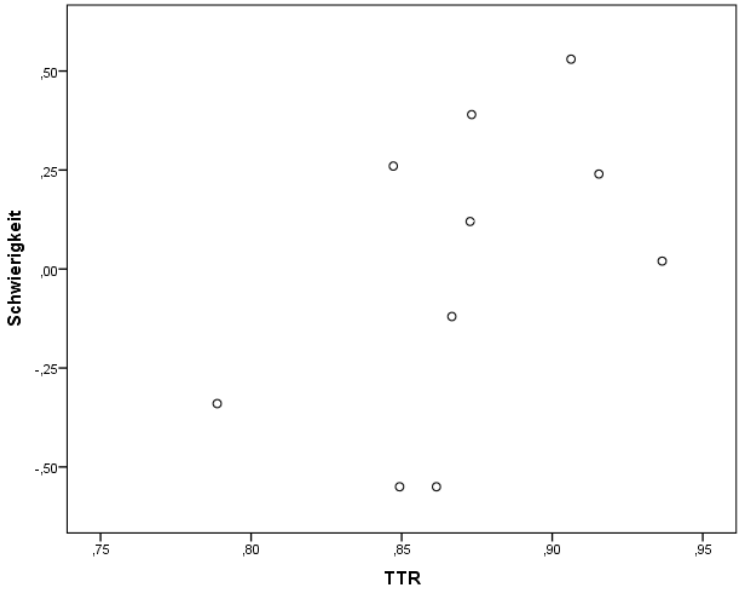

Abb. 2: Streudiagramm zur Verdeutlichung des Zusammenhangs zwischen Schwierigkeit und TTR-Werten

Anmerkungen: TTR = Type-Token-Ratio; Schwierigkeit in Logit.

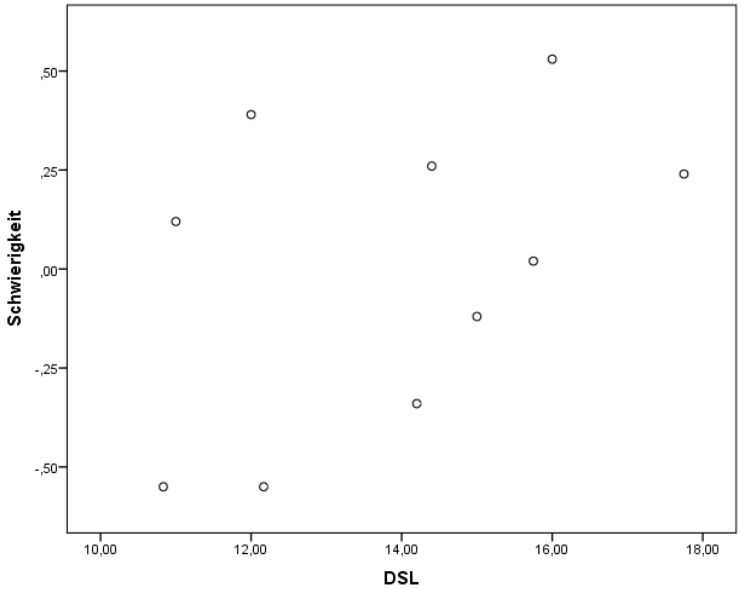

Abb. 3: Streudiagramm zur Verdeutlichung des Zusammenhangs zwischen Schwierigkeit und DSL-Werten

Anmerkungen: DSL = Durchschnittliche Satzlänge; Schwierigkeit in Logit.

4.3.3. Getilgte Inhaltswörter und Strukturwörter

Klein-Braley (1985b) argumentierte, dass Texte mit einem hohen Anteil an Strukturwörtern potenziell einfacher für Lernende sind als solche mit einem hohen Anteil an Inhaltswörtern, da letztere schwieriger zu rekonstruieren sind. Dies konnte von Stemmer (1991: 328) bestätigt werden. Da davon ausgegangen werden kann, dass bei einem Text mit einer hohen Anzahl an Inhaltswörtern entsprechend viele getilgt sind, wurden nachfolgend nur die gelöschten Inhalts- und Strukturwörter untersucht und deren Verhältnis zueinander (IWSW) ermittelt.

Tab. 5: Verhältnis getilgter Inhalts- und Strukturwörter

Text	IWSW	Schwierigkeitsmaß
2	0,33	-0,55
5	0,54	-0,55
9	0,67	-0,34
8	1,50	-0,12
1	1,50	0,02
10	1,50	0,12
6	1,00	0,24
3	1,22	0,26
7	1,00	0,39
4	1,86	0,53

Anmerkungen: Texte nach steigender Schwierigkeit sortiert. IWSW = Verhältnis getilgter Inhaltswörter zu getilgten Strukturwörtern.

Betrachtet man die Werte in Tab. 5, so fällt auf, dass die Ergebnisse zum Teil mit denen von Klein-Braley (1985b, 1994) übereinstimmen. Die leichtesten Texte (Text 2, Text 5) haben den niedrigsten IWSW-Wert. Auch der Hypothese entsprechend verhält es sich bei Text 4, der als schwierigster Text im Set mit einem Wert von 1,86 den geringsten Anteil an getilgten Strukturwörtern aufweist. Anhand der Streuung der Datenpunkte in Abb. 4 kann von einem engen Zusammenhang zwischen der Schwierigkeit eines Textes und dem Verhältnis getilgter Inhalts- zu Strukturwörtern ausgegangen werden. Die Korrelation liegt bei ,775 mit p = ,008.

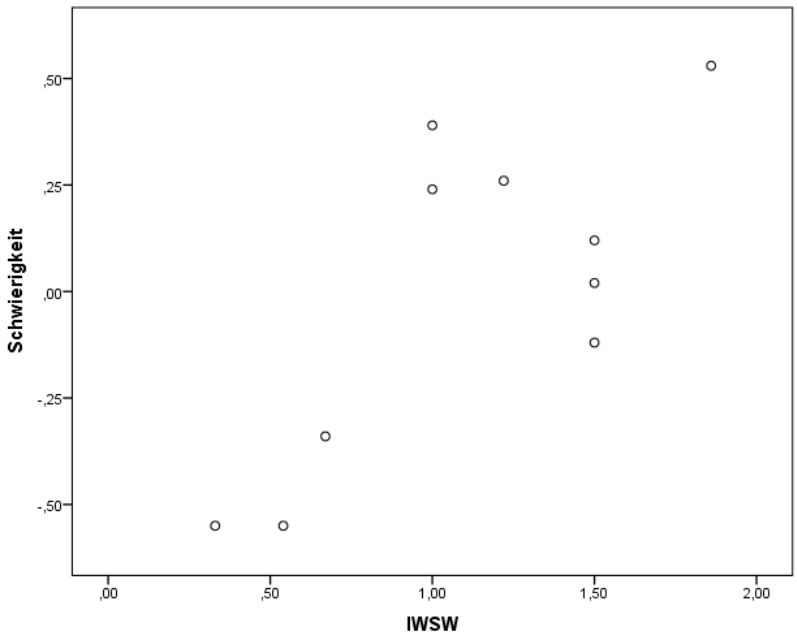

Abb. 4: Streudiagramm zur Verdeutlichung des Zusammenhangs zwischen Schwierigkeit und IWSW-Werten

Anmerkungen: IWSW = Verhältnis getilgter Inhaltswörter zu getilgten Strukturwörtern; Schwierigkeit in Logit.

4.3.4. Lineare Regression

Wie bereits erwähnt hat Klein-Braley (1984) zufriedenstellende Ergebnisse bei der Vorhersage der Schwierigkeit mittels Regressionsanalysen erzielt. Die Ergebnisse der vorliegenden Studie decken sich jedoch nicht mit denen von Klein-Braley. Die Prädiktoren in den hier verwendeten Gleichungen waren das Type-Token-Ratio, das Verhältnis von Strukturwörtern zu Inhaltswörtern sowie die durchschnittlichen Satzlängen, in unterschiedlichen Kombinationen. Diese Variablen beziehen sich zum einen auf die Bandbreite des Vokabulars, zum anderen auf die syntaktische Komplexität der Texte (vgl. z.B. Harsch &

Schröder 2007; Klein-Braley 1994). Die Korrelationen zwischen vorhergesagter und errechneter Schwierigkeit waren in keinem Fall zufriedenstellend. Die höchste Korrelation fand sich noch unter Verwendung aller drei genannten Variablen und lag bei ,661 (p = ,05). Zur zuverlässigen Vorhersage der Schwierigkeit waren die Regressionsgleichungen daher nicht geeignet.

4.3.5. Wortfrequenz und Wortschwierigkeit

Die Argumentation legt nahe, dass ein Wort, welches im Sprachgebrauch häufig vorkommt und somit als hochfrequent gilt, potenziell leichter zu erkennen bzw. zu rekonstruieren sein müsste als eines, welches seltener im sprachlichen Alltag Verwendung findet. Zur Überprüfung dieser These lassen sich Häufigkeitslisten heranziehen, die auf großen Korpora basieren. Problematisch ist hierbei, dass die Häufigkeit eines Wortes stark von der Thematik des Textes abhängt (vgl. Klein-Braley 1994: 276-277). Besonders bei kurzen Texten zu einem Thema, wie den C-Test-Texten, ist es fraglich, ob sich bei einem Vergleich zufriedenstellende Ergebnisse ergeben. Texte für den onDaF dürfen jedoch kein zu spezifisches Vokabular umfassen, so dass untersucht werden konnte, ob die Position in zwei Wortlisten in Zusammenhang mit der Schwierigkeit für die Kandidaten stand.

Der Vergleich umfasste jeweils die 10 Prozent der leichtesten und schwierigsten getilgten Wörter in dem untersuchten Set aus zehn Texten sowie deren Position in zwei Worthäufigkeitslisten. Bei diesen handelte es sich zum einen um den *Frequency Index* (FI) von Jones & Tschirner (2006), zum anderen um die *DeReWo – Korpusbasierte Grundformliste* (DeReWo 2013). Beim Vergleich der Daten war zu beachten, dass die Listen unterschiedliche Strukturen aufweisen. Die Häufigkeitsliste von Jones & Tschirner besteht aus den 4034 häufigsten Wörtern in einem Korpus von 4,2 Millionen gesprochenen und geschriebenen Wörtern, nach ihrer tatsächlichen Häufigkeit gelistet. Bei der korpusbasierten Wortgrundformenliste DeReWo hingegen handelt es sich um eine Lemmaliste, die die Wörter in verschiedene Häufigkeitskategorien unterteilt. Alle Wörter bzw. Kategorien werden hierbei relativ zum Aufkommen des bestimmten Artikels betrachtet, welcher am häufigsten im Sprachgebrauch verwendet wird.

Die in den Texten vielfach korrekt ergänzten Wörter sind in beiden Listen an hohen Positionen vorhanden. Im Gegensatz dazu sind für sechs der zwanzig schwierigsten Wörter keine Einträge im FI zu finden. Dies würde dafür sprechen, dass die Wörter, die häufig im Sprachgebrauch vorkommen, leichter für die Probanden zu rekonstruieren waren. Auch das Verhältnis getilgter Inhalts- und Strukturwörter entspricht der Erwartung. So ist unter den leichtesten nur ein Inhaltswort zu finden, während diese bei den schwierigsten deutlich überwiegen. Ausnahmen bilden die Wörter „auf“, „ein“ und „das“. Obwohl sie in den Häufigkeitslisten an sehr hohen Positionen zu finden sind, stellten sie sich im C-Test als schwierig heraus.

Klein-Braley (1985b) argumentiert, dass die Häufigkeit eines Wortes nicht allein für dessen Schwierigkeit in C-Tests ausschlaggebend ist. Ihr zufolge ist die Einbettung des Wortes in den spezifischen Kontext ebenfalls relevant (Klein-Braley 1985b: 37). Dies lässt sich anhand der vorliegenden Studie bestätigen. Betrachtet man Wörter, die in mehr als einem Text getilgt wurden, fällt auf, dass sie für die Probanden in den einzelnen Texten unterschiedlich schwierig zu rekonstruieren waren. So konnte bspw. die Präposition „in“ in Text 1 von 124 Probanden (61,39 %) korrekt ergänzt werden, während sie in Text 4 von nur 61 Probanden (30,20 %) als Lösung eingetragen wurde. Als Merkmal zur Vorhersage der Schwierigkeit der untersuchten Texte ist die Wortfrequenz daher nur bedingt geeignet.

4.3.6. Fehlerkategorien

Die detaillierte Betrachtung der einzelnen Lücken erfolgte für drei Texte des Sets: Text 5 mit der geringsten Schwierigkeit (-0,55), Text 4 mit der höchsten Schwierigkeit (0,53) sowie Text 1 mit mittlerer Schwierigkeit (0,02). Die Fragestellung lautete, welche der zu ergänzenden Wörter schwierig für die Probanden waren und weshalb sie sich als schwierig erwiesen hatten. Für die Analyse fand eine Einteilung der fehlerhaften Antworten der Probanden in fünf unterschiedliche Kategorien statt:

Fehlerkategorie 1: Unausgefüllte Lücke
Fehlerkategorie 2: Originallösung oder akzeptable Variante mit Rechtschreibfehler
Fehlerkategorie 3: Morphologiefehler
Fehlerkategorie 4: Lexikalischer Fehler
Fehlerkategorie 5: Nicht kategorisierbare Antwort

Fehlerkategorie 1: Wurde eine Lücke nicht gefüllt, kann dies an deren Schwierigkeit liegen. Wenn zum Ende eines Textes mehrere unausgefüllte Lücken aufeinander folgen, kann es jedoch auch auf ein Zeitproblem hindeuten. Des Weiteren kann ein Proband die Motivation verloren oder eine Lücke „‚aus Versehen’ nicht bearbeitet [haben]“ (Krauth 1995: 48). Bei einer ersten Betrachtung der drei Texte fand sich dieses Muster für verschiedene Probanden. Die drei letzten Lücken (Lücke 18-20) oder mehr haben im ersten Text 15 Probanden, in Text 5 insgesamt 22 Probanden und in Text 4 sogar 53 Probanden nicht ausgefüllt. Um erfassen zu können, welcher der oben genannten Gründe in Einzelfällen vorlag, müsste jedoch auf andere Verfahren zurückgegriffen werden.

Fehlerkategorie 2: Da es für die Bewertung von C-Tests unerheblich ist, ob ein Proband eine akzeptable Variante oder das Originalwort rekonstruiert hat, konnte darauf verzichtet werden, zwei voneinander getrennte Kategorien einzuführen. Der

Grad der orthographischen Abweichung von der richtigen Lösung diente als Maßstab dafür, ob ein Fehler als Orthographiefehler zählte oder in die Kategorie nicht kategorisierbarer Antworten fiel.

Fehlerkategorie 3: Unter Fehlerkategorie 3 fallen morphologische, also Wortbildungsfehler (z.B. Flexionsfehler).

Fehlerkategorie 4: Wenn ein Proband ein Wort in die Lücke geschrieben hat, das nicht mit der Originallösung oder potenziell akzeptablen Varianten übereinstimmte, zählte dieser Fehler zu Kategorie vier. Die Interpretation von lexikalischen Fehlern erwies sich als schwierig, weil kaum gesagt werden kann, ob es sich hierbei um das Wort handelt, von dem der Proband dachte, dass es zu rekonstruieren galt, oder ob er es eingetragen hat, weil ihm kein anderes Wort eingefallen ist, das mit den gegebenen Buchstaben beginnt. Dennoch war die Fehlerkategorie von Wert für die Auswertung, um zu prüfen, ob diese Art von Fehlern bei bestimmten Wörtern oder Wortarten vermehrt zu finden ist.

Fehlerkategorie 5: Bei den Datensätzen fanden sich einzelne Antworten, die sich nur mit viel Interpretation einer der obigen vier Kategorien zuordnen ließen. Die Ergänzung dieser Kategorie sollte daher den Grad der Spekulation reduzieren. In die Kategorie 5 entfielen Antworten wie *Wanderausen (=„Wanderausstellung“), die kein Wort darstellen, aber auch Antworten wie *ungenehme (=„ungenutzte“), bei denen zwar erkennbar ist, was gemeint ist – in diesem Fall „unangenehme“ – die Antwort jedoch so verfremdet ist, dass man nicht mehr von einem lexikalischen Fehler sprechen kann. Zudem sind Verletzungen des Tilgungsprinzips, d.h. das Eintragen von mehr als einem Wort in eine Lücke wie bei „in den“ (=„in“) dieser Kategorie zugeordnet.

In Tab. 6 ist die Verteilung der Fehler auf die einzelnen Kategorien zu sehen. Die prozentualen Angaben geben einen ersten Hinweis darauf, was für die Probanden schwierig war. Die anschließende einzelne Betrachtung der Texte bzw. der gegebenen Antworten erlaubte Detailanalysen der Fehler sowie den Vergleich der Antworten für eine Lücke und lückenübergreifend.

Tab. 6: Verteilung der Fehler auf die fünf Fehlerkategorien

	Text 1		**Text 4**		**Text 5**	
	f_k	%	f_k	%	f_k	%
FK 1	837	36,81	1080	40,43	740	41,83
FK 2	65	2,86	55	2,06	37	2,09
FK 3	628	27,62	362	13,55	199	11,25
FK 4	497	21,86	880	32,95	693	39,17
FK 5	247	10,86	295	11,04	100	5,61
Summe	2274	100,00	2671	100,00	1769	100,00

Anmerkung: f_k = Absolutbetrag der gefundenen Fehler; % = prozentualer Anteil an der Gesamtfehlersumme für den entsprechenden Text.

Die Anzahl an Fehlern steigt von Text 5 zu Text 1 um 28,55 %, von Text 1 zu Text 4 um weitere 17,46 %. Insgesamt findet sich von Text 5 zu Text 4 eine Steigerung um 50,99 %. Betrachtet man die Verteilung auf die Fehlerkategorien, so fällt auf, dass jeweils der größte Anteil der Fehler in den drei Texten auf unausgefüllte Lücken (FK 1) entfällt (Text 1: 36,81 %, Text 2: 40,43 %, Text 5: 41,83 %). Parallelen finden sich auch bei der Fehlerkategorie 2, die in allen Texten zwischen 2,04 % und 2,86 % der Gesamtfehlersumme ausmacht. Unterschiede gibt es besonders bei den Kategorien 3 und 4, also auf der morphologischen und lexikalischen Ebene. Beim leichtesten Text des Sets entfallen 11,25 % auf Morphologiefehler und 39,17 % auf lexikalische Fehler. Für den Text mit mittlerer Schwierigkeit ergeben sich Werte von 27,62 % und 21,86 % während sich das Verhältnis für den schwierigen Text wiederum umkehrt mit 13,55 % zu 32,95 %. Es ist demnach nicht möglich zu argumentieren, dass die Anzahl der Fehler einer bestimmten Kategorie direkte Hinweise auf die Schwierigkeit eines Textes gibt.

In den Analysen zu Fehlertypen und -häufigkeiten in den drei ausgewählten Texten zeigte sich, welche Lücken im Hinblick auf die jeweilige Fehlerkategorie auffällig waren. Dabei fanden sich unterschiedliche Schwierigkeiten, welche die Kandidaten bei der Bearbeitung hatten.Zur Zusammenfassung der Ergebnisse lassen sich die schwierigsten Lücken pro Text in Anlehnung an Little & Singleton (1992) in Tabellenform darstellen. Dies gibt Aufschluss darüber, ob für die verschiedenen Schwierigkeitsgrade der Texte Muster bei den Fehlern zu finden sind, die sich zur Vorhersage eignen. Unter Muster werden in diesem Zusammenhang die Parameter der Wortart, der Fehlersumme und deren Verteilung auf die Niveaustufen verstanden.

Die Tabellen 7a bis 7c zeigen für Text 1 und 5, welche getilgten Wörter von weniger als 50 % der Probanden korrekt zu rekonstruieren waren, und für Text 4, aufgrund der insgesamt deutlich höheren Fehleranzahl, diejenigen, die weniger als 25 % der Probanden korrekt ergänzten.

Tab. 7a: Lücken in Text 1, die weniger als 50 % der Pbn korrekt ergänzt haben

		Σ	FK 1	FK 2	FK 3	FK 4	FK 5
Adjektive	berü**hmtes**	133	20,30	6,77	42,86	0,75	29,32
	weit**eren**	139	36,69	0,00	17,27	41,01	5,04
Adverbien	ba**ld**	176	31,25	0,00	0,00	67,05	1,70
Pronomina	w**as**	125	86,40	1,12	0,00	8,00	4,00
Artikel	ei**nem**	179	11,73	0,00	83,24	5,03	0,00
Nomina	Jah**ren**	121	0,83	9,09	87,60	0,00	2,48
	Publ**ikum**	115	33,04	6,09	1,74	35,65	23,48
	Wanderaus**stellung**	188	62,23	1,60	0,00	0,00	36,17
	Stä**dten**	122	18,03	3,28	50,82	4,10	23,77
Präpositionen	i**m**	154	42,21	0,00	0,00	53,25	4,55

Tab. 7b: Lücken in Text 4, die weniger als 25 % der Pbn korrekt ergänzt haben

		Σ	FK 1	FK 2	FK 3	FK 4	FK 5
Adjektive	al**ten**	171	26,90	0,00	5,26	67,84	0,00
	ungen**utzten**	189	56,61	1,59	2,12	12,70	26,98
	gra**uen**	169	68,64	0,00	9,47	8,28	13,61
Nomina	Gärt**nern**	196	5,61	4,59	0,00	86,73	3,06
	Däc**her**	157	33,76	0,64	0,64	8,28	56,69
Präpositionen	inmi**tten**	178	63,48	15,17	0,00	1,12	20,22
	a**uf**	184	32,07	0,00	0,00	66,85	1,09

Tab. 7c: Lücken in Text 5, die weniger als 50 % der Pbn korrekt ergänzt haben

		Σ	FK 1	FK 2	FK 3	FK 4	FK 5
Artikel	d**ie**	122	13,93	0,00	68,85	16,39	0,82
Adjektiv	versch**iedene**	127	48,82	6,30	8,66	20,47	15,75
Konjunktion	a**ls**	145	36,55	0,00	0,00	63,45	0,00
Nomina	Ver**eine** / Ver**bände**	186	59,14	1,61	1,61	32,80	4,84
Präposition	v**on**	124	32,26	0,00	3,23	62,10	2,42
Verb	han**delt**	104	51,92	7,69	8,65	6,73	25,00

Anmerkungen zu Tabellen 7a bis 7c: Σ = Summe der fehlerhaften Antworten. In Spalte 2 bis 6 steht der prozentuale Anteil der Fehler in den fünf Fehlerkategorien. Die fettgedruckten Buchstaben waren im Test getilgt.

Der Vergleich der Tabellen zeigt, dass nur bei drei der 21 Lücken Wörter zu finden sind, die bei der Übersicht zu den Besonderheiten der deutschen Sprache nicht vorkamen. Die übrigen Lücken entfallen auf Nomina, Verben und Adjektive, ergänzt durch einen Fall des bestimmten Artikels sowie zwei Präpositionen.

Es zeigt sich daher, dass die schwierigen Lücken in den Texten für den onDaF mit dem übereinstimmen, was für Deutschlerner als schwierig gilt. Relativiert werden muss diese Aussage jedoch dahingehend, dass die häufigsten Fehlerarten bei den Lücken nicht immer mit den erwarteten Schwierigkeitswerten übereinstimmen. So findet sich z.B. der größte Anteil an Fehlern für die Nomina in Text 4 im Bereich der Lexik und undefinierbaren Antworten, der in Text 5 bei den unausgefüllten Lücken. Demnach kann zwar von der Wortart tendenziell auf die Schwierigkeit einer Lücke geschlossen

werden. Ob es aber möglich ist, von der Anzahl und der grammatikalischen Form dieser Wörter Schlüsse im Hinblick auf die Schwierigkeit des gesamten Textes zu ziehen, ist nicht zu sagen.

Die Hypothesen, die sich anhand der Fehleranalyse aufstellen ließen, lauten wie folgt:

- Die gegenseitige Abhängigkeit der Lücken innerhalb eines Satzes (oder Bedeutungsabschnitts) kann zu einer hohen Fehleranzahl führen.
- Die Flexion für Kasus und Genus (und zum Teil für Numerus) kann große Schwierigkeiten bereiten.
- Die Wahl der richtigen Wortart stellt für die Probanden teilweise eine Herausforderung dar.
- Die Nichtbeachtung des (weiteren) Kontexts führt zu Fehlern.
- Die Tilgung von Artikel und Bezugsnomen erschwert die Rekonstruktion beider Wörter.
- Präpositionen am Satzanfang, die eine kohäsionsstiftende Funktion haben, sind schwierig zu rekonstruieren.
- Die Zugehörigkeit zum Vokabular der gehobenen Sprache kann ein Indikator für die Schwierigkeit eines Wortes darstellen.
- Die Komplexität deutscher Komposita kann sich auf die Schwierigkeit einzelner Lücken auswirken.

5. Zusammenfassung und Diskussion

Die Schwierigkeit neu erstellter C-Test-Texte bereits vor der Erprobung und psychometrischen Analyse mit hoher Präzision vorhersagen zu können, könnte zur Qualität der Erprobung und der daraus gewonnenen Erkenntnisse beitragen. Auch wäre eine solche Vorhersage hilfreich, wenn Texte für ein spezielles Schwierigkeitsniveau entwickelt werden sollen. In der vorliegenden Studie wurden Textcharakteristika auf ihre Tauglichkeit für die Vorhersage der Schwierigkeit untersucht. Die Ergebnisse der Untersuchungen lassen sich wie folgt zusammenfassen:

- Die Schwierigkeit von C-Test-Texten hängt nicht wesentlich von der Thematik ab.
- Die durchschnittliche Satzlänge und das Type-Token-Ratio sind als Prädiktoren zur Vorhersage der Schwierigkeit der vorliegenden Texte eher ungeeignet.
- Die Schwierigkeit eines Wortes kann im Zusammenhang mit dessen Aufkommen im Sprachgebrauch stehen, wobei jedoch weitere Einflussfaktoren, insbesondere der Kontext, von Bedeutung sind.
- Das Verhältnis von getilgten Inhalts- zu Strukturwörtern ist tendenziell als Prädiktor geeignet.
- Anhand der Detailanalysen sind Schlussfolgerungen über die Schwierigkeit einzelner Lücken und den Zusammenhängen zwischen benachbarten Lücken möglich. Es ist jedoch nicht möglich, auf dieser Grundlage Aussagen über die Schwierigkeit eines Textes zu treffen.
- Die Merkmale der deutschen Sprache, die potenziell schwierig für Deutschlerner sind, können in einem C-Test zu Fehlern führen.
- Die Abhängigkeit der Lücken eines Textes voneinander und der damit zusammenhängende Kontext sind wichtig für die Rekonstruktion der Wörter.

Die Anzahl der in der Studie untersuchten Texte und Teilnehmerantworten lässt jedoch keine zuverlässigen, allgemeingültigen Aussagen zu. Hierfür würde es einer größeren Datenmenge bedürfen. Unabhängig von der Stichprobengröße gibt es weitere Gründe, weshalb einige der untersuchten Prädiktoren als solche ungeeignet sind.

So gibt die Zugehörigkeit eines Textes zu einer Themenkategorie nur vage Auskunft dafür, welche Inhalte der Text aufweist. Die Themenkategorien sind von großer Bedeutung für eine faire Testdurchführung beim onDaF, da sie verhindern, dass ein Teilnehmer Texte zu ähnlichen Themen mit ähnlichen Inhalten erhält und sich dies auf sein Ergebnis auswirkt, wenn er mit der Thematik wenig (oder sehr) vertraut ist. Für genauere Vorhersagen, ob ein Text für einen Teilnehmer eher schwierig oder leicht zu bearbeiten ist, bedarf es jedoch der Kenntnis des genauen Inhaltes und des Wissens darüber, ob der Teilnehmer über das themenspezifische Vokabular verfügt.

In der betrachteten Literatur wurde darauf verwiesen, dass die durchschnittliche Satzlänge als grober Kennwert für die syntaktische Komplexität eines Satzes geeignet ist, da längere Sätze tendenziell komplexer und somit schwieriger sind (vgl. Klein-Braley 1985b). Ein potenzieller Grund für die abweichenden Ergebnisse in dieser Studie lässt sich in Anlehnung an die Forschung zur Schwierigkeit von Aufgaben zum Leseverstehen finden. So argumentiert Grotjahn (2001: 84), dass die durchschnittliche Satzlänge nicht als Grund für die Schwierigkeit eines Textes herangezogen werden kann, da die Satzlänge mit dem Aufkommen von Konjunktionen wie „weil“, „da“ und „oder“ korreliert. Wenn man davon ausgeht, dass diese Aussage auch auf C-Test-Texte zutrifft, ist die durchschnittliche Satzlänge für die Vorhersage nicht geeignet.

Trotz der geringen Anzahl an Texten ließ sich eine hohe Korrelation zwischen der Schwierigkeit und dem Verhältnis getilgter Inhalts- zu getilgten Strukturwörtern finden. Dies spricht für einen Zusammenhang zwischen den beiden Faktoren. Es ist jedoch möglich, dass eine wiederholte Messung mit anderen Texten zu abweichenden Ergebnissen führen würde. Daher kann anhand dieser einen Analyse nicht hinreichend geklärt werden, inwieweit sich das Verhältnis von Inhalts- zu Strukturwörtern für die Vorhersage eignet. Bei der Erstellung neuer Erprobungssets für den onDaF hat es sich bislang als wenig hilfreich erwiesen.

Die Hypothesen, die sich aus der Fehleranalyse ergaben, entsprechen zum Teil dem aktuellen Forschungsstand zur Schwierigkeit von C-Tests und spiegeln die schwierigen Besonderheiten der deutschen Sprache wieder. Es lässt sich an diesem Punkt jedoch keine Aussage darüber treffen, wie sich diese Beobachtungen auf C-Test-Texte als Ganzes auswirken, da der Ausgangspunkt stets die Schwierigkeit einer einzelnen Lücke oder Bedeutungseinheit war. Natürlich liegt die Schlussfolgerung nahe, dass eine Kombination der gefundenen schwierigen Charakteristika zu einem schwierigen Text führt, aber welche Rolle einzelne sprachliche Elemente tatsächlich spielen und inwieweit sie für die Vorhersage der Schwierigkeit von C-Test-Texten und somit von C-Tests geeignet sind, kann anhand der Ergebnisse dieser Studie nicht festgestellt werden.

Die Ergebnisse stützen jedoch die Argumentation, dass die erfolgreiche Bearbeitung von C-Tests voraussetzt, dass Teilnehmer „über ein strukturiertes und differenziertes Sprachwissen verfügen und auf unterschiedliche Komponenten dieses Wissens zugreifen können“ (Eckes 2010: 126). Lerner benötigen eine fundierte Ausbildung in allen linguistischen Bereichen, u.a. in der Morphologie, der Orthographie und der Lexik, um im onDaF eine hohe Punktzahl zu erzielen. Wie sich gezeigt hat, führt fehlendes Wissen in einzelnen Bereichen sowohl bei leichten als auch bei schwierigen Texten zu Fehlern. Zudem müssen die Teilnehmer dazu in der Lage sein, Sprache auf allen Ebenen zu verarbeiten, vom einzelnen Buchstaben hin zum ganzen Text (vgl. Sigott 2004: 200). Die Nichtbeachtung des Kontextes stellt in einzelnen Fällen kein Problem dar, so dass Lücken unabhängig vom restlichen Text gelöst werden können. Andere Lücken können nur mit Hilfe der vorangegangenen oder darauffolgenden Wörter gelöst werden. Fehlen einem Teilnehmer Kenntnisse der Orthographie wird er es ebenso schwierig haben, auch wenn er weiß, welche Wörter gesucht sind. Das Zusammenspiel all dieser Aspekte erlaubt es nicht, einen einzelnen Faktor zu benennen, der darüber entscheidet, ob ein C-Test-Text für einen Lerner schwierig zu bearbeiten ist oder nicht.

Literaturverzeichnis

Anckaert, Philippe & Beeckmans, Renaud (1992), Le C-Test. Difficulté intrinsèque, pouvoir discriminant et validité de contenu. In: Grotjahn, Rüdiger (Hrsg.), 145-172.

Andrich, David (1978), A rating formulation for ordered response categories. *Psychometrika* 43: 4, 561-573.

Badia, Xavier; Prieto, Luis & Linacre, John M. (2002), Differential Item and Test Functioning (DIF & DTF). *Rasch Measurement Transactions* 16: 3, 889.

Bolten, Jürgen (1992), Wie schwierig ist ein C-Test? Erfahrungen mit dem C-Test als Einstufungstest in Hochschulkursen Deutsch als Fremdsprache. In: Grotjahn, Rüdiger (Hrsg.), 193-203.

Bortz, Jürgen & Döring, Nicola (2010), *Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler.* Heidelberg: Springer-Medizin-Verlag.

Bühner, Markus (2011), *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München, Don Mills: Pearson Studium.

DeReWo (2013), *Korpusbasierte Wortlisten DeReWo.* Allgemeine Anmerkungen, Technical Report IDS-KL-2013-01 [Online unter http://www.ids-mannheim.de/derewo. 22.11.2013].

Eckes, Thomas (2006), Rasch-Modelle zur C-Test-Skalierung. In: Grotjahn, Rüdiger (Hrsg.), *Der C-Test. Theorie, Empirie, Anwendungen/The C-Test: Theory, Empirical Research, Applications*. Frankfurt am Main: Lang, 1-44.

Eckes, Thomas (2007), Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen. *Diagnostica* 53: 2, 68-82.

Eckes, Thomas (2010), Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung. In: Grotjahn, Rüdiger (Hrsg.), *Der C-Test: Beiträge aus der aktuellen Forschung.* Frankfurt am Main: Lang, 125-192.

Eckes, Thomas & Baghaei, Purya (2015), Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education* 28, 85-98.

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

Grotjahn, Rüdiger (1992), Der C-Test. Einleitende Bemerkungen. In: Grotjahn, Rüdiger (Hrsg.), 1-18.

Grotjahn, Rüdiger (Hrsg.) (1992a), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer.

Grotjahn, Rüdiger (2001), Determinants of the difficulty of foreign language reading and listening comprehension tasks: predicting task difficulty in language tests. In: Pürschel, Heiner & Raatz, Ulrich (Hrsg.), *Tests and translation. Papers in memory of Christine Klein-Braley*. Bochum: AKS-Verlag, 79-102.

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

Harsch, Claudia & Schröder, Konrad (2007), Textrekonstruktion: C-Test. In: Beck, Bärbel & Klieme, Eckhard (Hrsg.), *c)*. Weinheim, Basel: Beltz, 212-225.

Hastings, Ashley J. (2002), Error analysis of an English C-test: Evidence for integrated processing. In: Grotjahn, Rüdiger (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: AKS-Verlag, 53-66.

Jones, Randall L. & Tschirner, Erwin P. (2006), *A frequency dictionary of German. Core vocabulary for learners*. London, New York: Routledge.

Klein-Braley, Christine (1984), Advance prediction of difficulty with C-Tests. In: Culhane, Terry; Klein-Braley, Christine & Stevenson, Douglas K. (Hrsg.), *Practice and problems in language testing 7*. Colchester: University of Essex, 97-114.

Klein-Braley, Christine (1985a), Reduced redundancy as an approach to language testing. In: Klein-Braley, Christine & Raatz, Ulrich (Hrsg.), *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis*. Bochum: AKS-Verlag, 1-19.

Klein-Braley, Christine (1985b), Advance prediction of test difficulty. In: Klein-Braley, Christine & Raatz, Ulrich (Hrsg.), *Fremdsprachen und Hochschule 13/14: Thematischer Teil: C-Tests in der Praxis*. Bochum: AKS-Verlag, 23-41.

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

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

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

Linacre, John M. (2002), What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions* 16: 2, 878.

Little, David & Singleton, David (1992), The C-Test as an elicitation instrument in second language research. In: Grotjahn, Rüdiger (Hrsg.), 173-192.

Raatz, Ulrich & Klein-Braley, Christine (2002), Introduction to language testing and to C-Tests. In: Coleman, James A.; Grotjahn, Rüdiger & Raatz, Ulrich (Hrsg.), *University language testing and the C-test*. Bochum: AKS-Verlag, 75-93.

Sigott, Günther (2004), *Towards identifying the c-test construct*. Frankfurt am Main: Lang.

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

Wirtz, Markus A. (Hrsg.) (2013), *Dorsch - Lexikon der Psychologie* (16. Aufl.). Bern: Verlag Hans Huber [Online unter https://portal.hogrefe.com/dorsch/. 23.11.2013].

Wright, Benjamin D. & Linacre, John M. (1994), Reasonable mean-square fit values. *Rasch Measurement Transactions* 8: 3, 370.

Anmerkungen

[1] Aus Gründen der sprachlichen Vereinfachung werden Ausdrücke wie „Lerner“ oder „Proband“ im generischen Sinne verwendet.

[2] Aus Gründen der Testsicherheit können die C-Test-Texte nicht abgedruckt werden. Unter www.ondaf.de [Link „Beispieltest“] sind Texte einsehbar, die dasselbe Format haben wie die untersuchten Texte.