KERSTIN ZIMMERMANN

Keine Zeit für den C-Test?

Eine empirische Untersuchung zum Einfluss einer Geschwindigkeitskomponente auf das Konstrukt des C-Tests

Universitätsverlag der TU Berlin

Technische Universität Berlin

Kerstin Zimmermann

Keine Zeit für den C-Test?

Eine empirische Untersuchung zum Einfluss einer Geschwindigkeitskomponente auf das Konstrukt des C-Tests

Kerstin Zimmermann

Keine Zeit für den C-Test?

Eine empirische Untersuchung zum Einfluss einer Geschwindigkeitskomponente auf das Konstrukt des C-Tests

Universitätsverlag der TU Berlin

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Universitätsverlag der TU Berlin, 2019
http://verlag.tu-berlin.de

Fasanenstr. 88, 10623 Berlin
Tel.: +49 (0)30 314 76131 / Fax: -76133
E-Mail: publikationen@ub.tu-berlin.de

Zugl.: Berlin, Techn. Univ., Diss., 2019
Gutachter: Prof. em. Dr. Ulrich Steinmüller
Gutachterin: Jun.-Prof. Dr. Anastasia Drackert (Ruhr-Universität Bochum)
Die Arbeit wurde am 18. Februar 2019 an der Fakultät I unter Vorsitz von Prof. Dr. Hans-Christian von Herrmann erfolgreich verteidigt.

Umschlagfoto: Aah-Yeah | https://www.flickr.com/photos/aah-yeah/35360272382/ | CC BY 2.0 | https://creativecommons.org/licenses/by/2.0/

Druck: docupoint GmbH
Satz/Layout: Georg Frein, Kerstin Zimmermann

ISBN 978-3-7983-3076-4 (print)
ISBN 978-3-7983-3077-1 (online)

Zugleich online veröffentlicht auf dem institutionellen Repositorium der Technischen Universität Berlin:
DOI 10.14279/depositonce-8288
http://dx.doi.org/10.14279/depositonce-8288

Danksagung

Eine Doktorarbeit schreibt man allein. Das dachte ich jedenfalls. Doch im Laufe des mehrjährigen Arbeitsprozesses wurde ich eines Besseren belehrt. Ohne die Unterstützung und Kooperationsbereitschaft zahlreicher Menschen wäre es mir nicht möglich gewesen, dieses Forschungsprojekt durchzuführen.

Mein Dank geht zunächst an *Dr. Thomas Eckes* und das *TestDaF-Institut* in Bochum, das mir zum Zwecke dieser Forschung freundlicherweise eine Papier-und-Bleistift-Version des onDaF zur Verfügung stellte.

Ich danke *Dr. Almut Schön*, Geschäftsführerin der *Zentraleinrichtung Moderne Sprachen* (ZEMS) an der Technischen Universität Berlin, für die Möglichkeit, in zahlreichen Deutschkursen Daten erheben zu dürfen. An dieser Stelle sei auch allen beteiligten Dozenten gedankt, die bereitwillig einen Teil ihrer Unterrichtszeit für mich geopfert haben: *Johanna Bräutigam, Ulla Kilias, Renate Klebe, Katrin Landesfeind, Dr. Claudia Müller* und *Wolfgang Zimmermann.*

Für die technische Unterstützung bei der Datenerhebung danke ich *Gesche Loft, Honorine Tchuatcheu* und *Wolfgang Zimmermann*.

Anna Fabian, *Dr. Dörte Grunzig* und *Sabine Prudent* sowie meinen Doktorschwestern *Dr. Simone Heine* und *Dr. Nicole Hartung* danke ich für ihre offenen Ohren, den dicken Geduldsfaden und unermüdliche moralische Unterstützung, deren Wert man gar nicht überschätzen kann.

Besonderer Dank gilt meinen Betreuern *Prof. Dr. Ulrich Steinmüller* und *Jun.-Prof. Dr. Anastasia Drackert*, die mir in fachlichen Fragen stets konstruktivem Feedback zur Seite standen. *Prof. Dr. Rüdiger Grotjahn* und *Philipp Puffer* danke ich für Hilfe in statistischen Fragen. Für ihr kritisches

Feedback zu diversen Kapiteln geht mein Dank an *Dr. Simone Heine, Heike Molnár* und *Dr. Almut Schön.*

Meinem Doktorvater *Prof. Dr. Ulrich Steinmüller* danke ich für die langjährige Begleitung durch alle Höhen und Tiefen, für seine Geduld und sein Vertrauen, für die Freiheit bei meiner Forschung und für seine immer offene Tür. Oft habe ich sein Büro frustriert betreten und voller Zuversicht wieder verlassen.

Zu guter Letzt möchte ich allen meinen anonymen *Probanden* danken, denn ohne deren Teilnahmebereitschaft hätte ich dieses Forschungsprojekt gar nicht durchführen können.[1]

1 Diese Dissertation ist am Fachgebiet Deutsch als Fremd- und Fachsprache des Instituts für Sprache und Kommunikation der Fakultät für Geistes- und Bildungswissenschaften an der Technischen Universität Berlin unter der Betreuung von Prof. em. Dr. Ulrich Steinmüller entstanden.

Abstracts

Deutsch

Der C-Test gilt als ein objektives, reliables und valides Instrument zum Erheben allgemeiner Sprachkompetenz. Zahlreiche Studien belegen den Zusammenhang des C-Tests mit verschiedenen sprachlichen Teilfertigkeiten: Leseverstehen, Hörverstehen, Sprechen, Schreiben, Wortschatz, Grammatik. Im Gegensatz dazu ist der S-C-Test – eine stark beschleunigte Variante des Testformats – noch kaum erforscht. Diesem Desiderat kommt die vorliegende Arbeit nach.

Die Studie beschäftigt sich mit der Frage, wie sich eine drastische Verkürzung der Bearbeitungszeit auf den C-Test auswirkt. Zugrunde liegt die Hypothese von GROTJAHN et al. (2010), dass eine Geschwindigkeitskomponente dazu beitrage, Sprachverwendung in Echtzeit zu simulieren und sich so der Zusammenhang zwischen der Leistung in Hörverstehens- und mündlichen Tests mit den C-Test-Ergebnissen erhöhen könnte.

Nach einer theoretischen Abhandlung von Entstehung, Konstruktion, Varianten und Einsatzmöglichkeiten des C-Tests werden ausgewählte Studien, die Korrelationsanalysen beinhalten oder einen S-C-Test verwenden, detailliert diskutiert.

Im empirischen Teil der Arbeit werden anhand von Daten erwachsener Deutschlerner im universitären Bereich C-Tests mit und ohne Geschwindigkeitskomponente miteinander verglichen. Es zeigt sich, dass der S-C-Test eine zufriedenstellende Reliabilität aufweist. Korrelationsanalysen ergeben, dass der S-C-Test einen stärkeren Zusammenhang mit Hörverstehen aufweist als der herkömmliche C-Test. Für die Fertigkeit Sprechen zeigen sich zwischen beiden Testverfahren kaum Unterschiede. Der Zu-

sammenhang des S-C-Tests mit den beiden genannten Fertigkeiten fällt bei leistungsstarken Probanden deutlich stärker aus als bei schwächeren Testteilnehmern.

Zusammenfassend weisen die Ergebnisse in die Richtung, dass der S-C-Test ein zuverlässiges Messinstrument darstellt, das insbesondere bei fortgeschrittenen Lernern dazu geeignet ist, in *low stakes*-Testsituationen Fremdsprachenkompetenz zu ermitteln.

Englisch

The C-test has come to be known as an objective, reliable and valid means of measuring general language proficiency. Many studies support its interrelation with several language sub-skills: reading comprehension, listening comprehension, speaking, writing, vocabulary, grammar. By contrast, the S-C-test – a substantially speeded up version of this testing format – has not yet been investigated in depth. The purpose of the study at hand is to help fill this gap.

This thesis deals with the question concerning the influence the drastic reduction of time allotted to complete the test has on the C-test. Underlying is an assumption made by GROTJAHN et al. (2010) hypothesizing that adding a speed factor to the C-test, thus simulating real time language use, might lead to higher correlations with listening comprehension and speaking tasks.

After a theoretical discussion of the origin, construction, variants and applications of C-tests, studies dealing with correlational analyses or using a speeded C-test will be examined.

In the second part of this thesis, C-tests and speeded C-tests will be compared, with data gathered from adult learners of German in a university context. As will be seen, S-C-tests reach a sufficient reliability. In fact, correlational analyses show that S-C-tests are more closely related to listening comprehension than commonly used C-tests. As far as speaking is concerned, both test versions yield similar results. Correlations with both language skills are considerably stronger for advanced learners than for less advanced ones.

To sum up, the S-C-test appears to be a reliable test instrument that is appropriate for measuring foreign language competence in low stakes situations, especially for advanced learners.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

The proof of the pudding is in the eating,
and if the test works then it is valid.
(RAATZ 1985: 62)

Wer fremdsprachliche Kompetenzen umfassend und zuverlässig testen möchte, braucht vor allem eins: Zeit. Standardisierte Testbatterien, die nicht nur die vier Fertigkeiten Hörverstehen, Leseverstehen, Sprechen und Schreiben separat messen, sondern auch Grammatik und zum Teil Wortschatz erheben, liefern zwar ein sehr differenziertes Kompetenzprofil, kosten aber sowohl die Testabnahmestelle als auch die Testteilnehmer[2] sehr viel Zeit. Für viele Situationen, in denen vom Testergebnis nicht allzu viel abhängt, sind derartige Testbatterien den Teilnehmern nicht zuzumuten und somit nicht geeignet.

Zeitökonomischer ist der C-Test. Er besteht aus mehreren, thematisch in sich geschlossenen kurzen Texten, bei denen, beginnend beim zweiten Wort des zweiten Satzes, jeweils die hintere Hälfte jedes zweiten Worts getilgt wird (vgl. KLEIN-BRALEY & RAATZ 1984: 136). Seit seiner Einführung vor über 30 Jahren hat sich der C-Test zu einem etablierten Testverfahren zur Messung globaler Sprachkompetenz entwickelt, das signifikant positive Korrelationen mit diversen sprachlichen Teilfertigkeiten aufweist

2 Unter Verweis auf die linguistischen Aspekte von EISENBERG (2017) sowie auf die von MARCHWACKA (2012: 11) genannten soziologischen Aspekte der geschlechtlichen Identität wird bewusst auf eine sogenannte geschlechtergerechte Sprache verzichtet. Alle biologischen und identifizierten Geschlechter sind stets intendiert, sofern nicht explizit ein Geschlecht ausdifferenziert wird.

(vgl. Übersicht in ECKES & GROTJAHN 2006: 295 ff.). Der C-Test wird international zu Zwecken des Screenings, der Einstufung und auch als Forschungsinstrument eingesetzt (vgl. GROTJAHN 1995: 44 f.; GROTJAHN 2011: 131 f.). Obwohl das Testformat ursprünglich für die flektierenden Sprachen Deutsch und Englisch entwickelt wurde, existieren inzwischen C-Test-Varianten für mehr als 25 verschiedene Sprachen (vgl. GROTJAHN 2011: 131).

RAATZ und KLEIN-BRALEY entwickelten den C-Test mit der Intention, ein Sprachtestformat zu etablieren, das einige Probleme des bis dahin stark verbreiteten Cloze-Tests löst. Zu den Kritikpunkten am Cloze-Test zählen u. a. die relative Testlänge, Schwierigkeiten bei der Auswertung, eine geringe Reliabilität sowie die Tatsache, dass Cloze-Tests teilweise auch für LI-Sprecher nur schwer lösbar sind (vgl. RAATZ & KLEIN-BRALEY 2002: 77 f.). Der neu zu entwickelnde Test sollte folglich deutlich kürzer sein als ein Cloze-Test. Er sollte über ein kanonisches Tilgungsprinzip verfügen und für Muttersprachler (nahezu) vollständig lösbar sein. Darüber hinaus sollte der neue Test den Testgütekriterien der Objektivität, Reliabilität und Validität genügen und leicht zu entwickeln sein (vgl. RAATZ & KLEIN-BRALEY 2002: 78).

Während es zum C-Test zahlreiche Forschungsarbeiten und Publikationen gibt (vgl. GROTJAHN 2014), die seine Validität belegen, stellt die Variante des sogenannten *speeded*-C-Tests ein weitgehend unerforschtes Gebiet dar. Unter einem *speeded*-C-Test wird ein C-Test mit stark beschränkter Bearbeitungszeit verstanden. Nur wenige Studien nutzen dieses Testformat als Forschungsinstrument, noch seltener ist der *speeded*-C-Test selbst Untersuchungsgegenstand (vgl. Kapitel 3.3).

Ziel der vorliegenden Arbeit ist es, einen Beitrag zur Erforschung des *speeded*-C-Tests zu leisten und zu ergründen, ob sich das Konstrukt des C-Tests durch das Hinzufügen einer Geschwindigkeitskomponente verän-

dert. Zentral ist hierbei die Beantwortung der von GROTJAHN et al. (2010) aufgestellten Hypothese, dass ein C-Test mit Geschwindigkeitsfaktor erhöhte Korrelationen mit den in Echtzeit ablaufenden Fertigkeiten Hörverstehen und Sprechen aufweisen könnte.

Zunächst wird jedoch das dem C-Test zugrunde liegende Prinzip der reduzierten Redundanz erläutert (Kapitel 2.1). Es folgt ein Überblick über diverse Testformate, die auf der Basis dieses Prinzips konzipiert wurden. Hierzu zählen neben dem klassischen Diktat auch der Cloze-Test, der direkte Vorläufer des C-Tests (Kapitel 2.2). Über die Kritik am Cloze-Test und die Forderung nach einem neuen Testformat (Kapitel 2.2.1) gelangt man zum Prinzip des C-Tests, dessen Aufbau anhand eines Beispiels erläutert wird (Kapitel 2.2.2). Im Anschluss wird der C-Test hinsichtlich der Erfüllung der klassischen Testgütekriterien Objektivität, Reliabilität und Validität untersucht (Kapitel 2.2.3) und sein Konstrukt umfassend diskutiert (Kapitel 2.2.4).

Der *speeded*-C-Test wird als die im Vordergrund dieser Untersuchung stehende Variante des C-Tests sodann in einem separaten Kapitel eingeführt, und es wird überprüft, wie er testtheoretisch klassifiziert werden kann und welche Aussagen man hinsichtlich der Erfüllung der klassischen Gütekriterien machen kann (Kapitel 2.2.5).

Neben dem *speeded*-C-Test existiert eine große Zahl weiterer Varianten des Testformats, die im folgenden Kapitel thematisiert werden. Hierzu zählen C-Test-Versionen in diversen flektierenden Sprachen sowie Variationen des Testprinzips für Sprachen typologisch anderer Bauart (Kapitel 2.3.1). Darüber hinaus werden auch fachsprachliche C-Tests konstruiert und eingesetzt (Kapitel 2.3.2). Weitere Varianten des C-Tests ergeben sich unter anderem aus den verschiedenen Darbietungsformen wie dem Ablegen des Tests am Computer oder mit Papier und Bleistift oder die Unter-

scheidung zwischen Einzel- und Gesamtdarbietung der C-Test-Texte (Kapitel 2.3.3 und 2.3.4).

Es schließt sich eine Betrachtung der mannigfaltigen Einsatzmöglichkeiten des C-Tests im mutter-, zweit- und fremdsprachlichen Bereich an. Dazu variiert der Zweck des Einsatzes von Einstufung und Screening bis hin zu Diagnostik und Sprachförderung im schulischen Bereich (Kapitel 2.4).

Nach dieser umfassenden Einführung zum C-Test folgt ein Überblick über die für die vorliegende Studie relevante Forschungsliteratur. Zunächst wird die Rolle der Fertigkeiten Hörverstehen und Sprechen im Fremdsprachenunterricht erörtert und die Bedeutung einer hohen Sprachverarbeitungsgeschwindigkeit diskutiert (Kapitel 3.1). Im Anschluss wird zwischen zweierlei Untersuchungstypen unterschieden: Zum einen werden korrelationsanalytische Studien herangezogen, die einen Hinweis über den Zusammenhang des C-Tests mit den Fertigkeiten Hörverstehen und Sprechen liefern (Kapitel 3.2). Zum anderen werden die wenigen vorhandenen Untersuchungen angeführt, die sich entweder eines S-C-Tests bedienen oder einen solchen ins Zentrum der eigenen Forschung stellen (Kapitel 3.3).

Im sich anschließenden empirischen Teil der Arbeit wird zunächst die hinsichtlich des *speeded*-C-Tests existierende Forschungslücke aufgezeigt sowie aus dieser heraus Forschungsfragen entwickelt (Kapitel 4.1). Nach einigen kurzen Hinweisen zu forschungsethischen Aspekten (Kapitel 4.2) werden die eingesetzten Forschungsinstrumente vorgestellt: ein C-Test, ein *speeded*-C-Test, ein Hörverstehenstest sowie ein Test zum Sprechen. Das Instrumentarium wird durch einen Fragebogen komplettiert, der neben persönlichen Angaben auch Erfahrungen mit und Einstellungen zu dem Testformat des C-Tests abfragt (Kapitel 4.3.1). Es schließt sich ein Bericht über die Pilotierung der für den *speeded*-C-Test angesetzten Bearbeitungszeit (Kapitel 4.3.2) sowie eine Beschreibung der Probanden anhand

des ausgewerteten Begleitfragebogens an (Kapitel 4.3.3). Das folgende Kapitel 4.3.4 legt den Aufbau der Studie dar. Der zeitliche Ablauf und das methodische Vorgehen bei der Durchführung der Datenerhebung werden ausführlich dargestellt und erläutert.

Nach einer deskriptiven Analyse der Testergebnisse werden zur Beantwortung der eingangs gestellten Forschungsfragen weitere statistische Analysen durchgeführt und interpretiert (Kapitel 4.4): Zunächst klärt eine Reliabilitätsanalyse die Frage nach der Zuverlässigkeit des *speeded*-C-Tests. Anschließend werden zahlreiche Korrelationen zwischen den beiden C-Test-Versionen und den Ergebnissen der Tests zum Hörverstehen und Sprechen errechnet.

Mithilfe von Gruppenvergleichen wird weiterhin überprüft, ob das Vertrauen in das Testformat oder das Schriftsystem der Alphabetisierung einen signifikanten Unterschied in den Testergebnissen des S-C-Tests bedingen.

Schließlich wird der Begleitfragebogen hinsichtlich des Vorgehens beim Lösen des C-Tests und des *speeded*-C-Tests ausgewertet. Eine Häufigkeitstabelle macht etwaige Unterschiede sichtbar.

Der empirische Teil endet mit einer kritischen Diskussion der Ergebnisse (Kapitel 4.4.4). In der sich anschließenden Reflexion der eigenen Forschung wird ein Ausblick auf potentielle Ansätze für künftige Studien zum *speeded*-C-Test gegeben (Kapitel 5).

Die Relevanz der Ergebnisse dieser Studie liegt auf der Hand: Liefert der *speeded*-C-Test ähnliche Ergebnisse wie der C-Test, ist er in *low-stakes*-Testsituationen wie etwa der Einstufung in Sprachkurse vorzuziehen. Erhöht sich durch die Geschwindigkeitskomponente zudem die Korrelation mit den Fertigkeiten Hörverstehen und Sprechen, würde der Test noch zuverlässiger (Kapitel 6).

2 Der C-Test

Der C-Test wurde 1981 von seinen Erfindern Ulrich Raatz und Christine Klein-Braley publik gemacht. Er ist ein mehrteiliger Lückentext, dessen Konstruktion auf dem Prinzip der reduzierten Redundanz (vgl. Kapitel 2.1) beruht. Seit seiner Einführung hat es zahlreiche Weiterentwicklungen des C-Tests gegeben. Diese betreffen sowohl die inhaltliche Ebene (z. B. fachsprachliche C-Tests) als auch die Darbietungsebene (z. B. Zeitbemessung pro Text). Dazu kommen zahlreiche Varianten für flektierende, agglutinierende und isolierende Sprachen (vgl. Kapitel 2.3.1). Auch die Einsatzgebiete des C-Tests haben sich seit 1981 deutlich erweitert, so dass der C-Test heute bei Erwachsenen und Kindern, Zweit- und Fremdsprachlern und zu diversen Zwecken (z. B. Einstufung oder Diagnostik) Verwendung findet.

Das folgende Kapitel liefert einen kurzen historischen Abriss zur Entstehung des C-Tests. Im Anschluss wird das kanonische Tilgungsprinzip des C-Tests erläutert und seine Tauglichkeit in Bezug auf die klassischen Testgütekriterien diskutiert. Danach wird das dem C-Test zu Grunde liegende Konstrukt genauer beleuchtet, ehe auf den hier zentralen *speeded*-C-Test detailliert eingegangen wird.

2.1 Tests auf der Basis reduzierter Redundanz

Der C-Test ist Teil einer Familie von Tests, die alle auf dem gleichen Prinzip basieren: der reduzierten Redundanz. Das sogenannte *Reduced Redundancy Testing* fußt sowohl auf Teilen der Informationstheorie als auch auf Teilen der Gestaltpsychologie (vgl. Klein-Braley 1983: 218).

Die Gestalttheorie ist ein 1912 zuerst von Max Wertheimer beschriebener Ansatz der Psychologie. Die auf Aristoteles' Metaphysik beruhende Formel „Das Ganze ist mehr als die Summe seiner Teile" (vgl. ARISTOTELES 1995: 168) umreißt nur knapp den Kern der Gestalttheorie. WERTHEIMER (1924) nennt das Wiedererkennen einer transponierten Melodie als Beispiel, um das Konzept der Gestalttheorie zu verdeutlichen:

> But what is it that enables us to recognize the melody when it is played in a new key? The sum of the elements is different, yet the melody is the same; indeed, one is often not even aware that a transposition has been made. […] There must be a something more than the sum of six tones, viz. a seventh something, which is the form-quality, the Gestaltqualität, of the original six. It is this seventh factor or element which enabled you to recognize the melody despite its transposition. (WERTHEIMER 1924: 3)

Es geht folglich darum, herauszustellen, welche Faktoren dazu beitragen, dass veränderte Informationen im Bewusstsein eines Rezipienten dennoch wiedererkannt werden. Die Gestalttheorie beinhaltet jedoch noch mehr. So kann das menschliche Gehirn auch aus unvollständigen Informationen etwas Sinnvolles, Bekanntes rekonstruieren. Ein visuelles Beispiel für dieses Prinzip zeigt Abbildung 1.

Obgleich Abbildung 1 lediglich gerade und gebogene Striche zeigt, fügt das Gehirn diese visuellen Informationen aufgrund ihrer Anordnung zueinander dergestalt zusammen, dass man auf der linken Seite einen Kreis und auf der rechten Seite ein Rechteck zu erkennen glaubt. Dieses anschauliche Beispiel kann auch den Zusammenhang von Gestalttheorie und C-Test verdeutlichen: Auch hier liegen einem Testteilnehmer unvollständige Informationen, d. h. ein lückenhafter Text, vor. Die Sprachkompetenz einer Person wird darüber ermittelt, inwiefern jemand dazu in der Lage ist, die unvollständigen Wörter und Sätze als „ganz" wahrzunehmen.

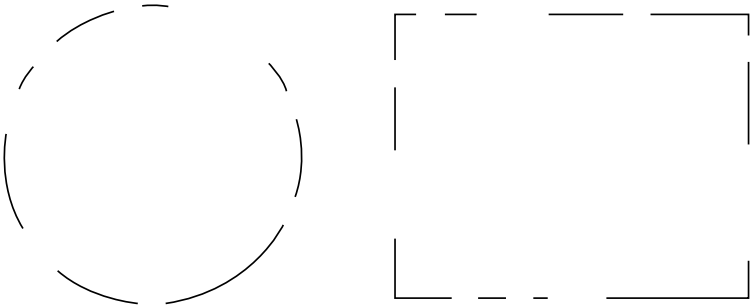


Abb. 1: Gesetz der Geschlossenheit (URL 8)

Ein Bezug zwischen *Reduced Redundancy Testing* und Informationstheorie besteht aufgrund der Tatsache, dass sowohl die Idee der Redundanz als auch ihre Reduktion durch *Noise* aus dieser Theorie stammen. Unter *Noise* versteht man jede Art von „unwanted signal that interferes with the communication, measurement or processing of an information-bearing signal" (VASEGHI 2000: 30). Das bedeutet, durch *Noise* kann der Kanal zwischen Sender und Empfänger einer Information gestört werden, wie Abbildung 2 veranschaulicht.

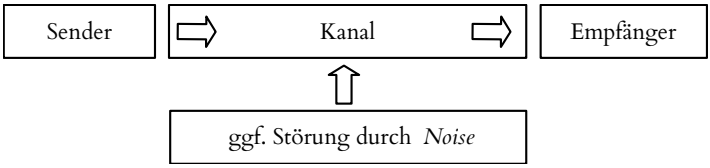


Abb. 2: Vereinfachtes Modell zur Informationsübertragung (nach SHANNON 1949: 3)

Unter Redundanz versteht man das mehrfache Vorhandensein von Informationseinheiten. Wie das folgende Beispiel illustriert, sind natürliche Sprachen redundant: In dem Satz „die Katze des kleinen Kindchens" sind

mehrere Informationen doppelt kodiert. So fällt zunächst auf, dass sowohl das Adjektiv „klein“ als auch das Diminutivsuffix „-chen“ Auskunft über die Größe bzw. das Alter des Kindes geben. „-chen“ beinhaltet zugleich die Information über das Genus des Wortes, welches ebenfalls durch den bestimmten Artikel „des“ eingegrenzt wird (feminin wäre unmöglich).

Des Weiteren liefert das „-e“ in „Katze“ einen Hinweis auf das Genus dieses Wortes, welches ebenfalls in dem vorgestellten bestimmten Artikel kodiert wird. Der Genitiv wird ebenfalls doppelt angezeigt, nämlich durch den bestimmten Artikel „des“ sowie durch die Endung „-s“ beim Wort „Kindchens“.

HABER (1974: 26) beschreibt das Verhältnis von Redundanz und *Noise* folgendermaßen:

> A few mistakes in ordinary English to [sic!] not render the whole message unintelligible. In fact the recipient can usually insert single missing symbols without too much guesswork. This is because of the redundancy. If the channel is perfect, the redundancy is unnecessary. If it is imperfect, it will be necessary to leave or insert some redundancy in the system in order to receive without error.

Wird also der Kanal mittels *Noise* gestört, wird die Redundanz natürlicher Sprachen reduziert. Dies ist auch in alltäglichen Situationen der Fall, sowohl im schriftsprachlichen Bereich, beispielsweise, wenn der Rand von einem Brief abgerissen oder verschmutzt ist oder ein Text in unleserlicher Schrift geschrieben wurde, als auch in der gesprochenen Sprache, wenn zum Beispiel bei Bahnhofsansagen ein einfahrender Zug das Hörverstehen beeinträchtigt (vgl. GRADMAN & SPOLSKY 1975: 67). Auf diese Weise ist der Leser bzw. Hörer darauf angewiesen, mit weniger Informationen und einem höheren Grad an potentieller Mehrdeutigkeit etwas Sinnvolles aus dem Gelesenen bzw. Gehörten zusammenzufügen. Muttersprachlern gelingt dies zumeist besser als Fremdsprachenlernern, wie SPOLSKY (1968: 81) schreibt:

> It has been observed that native speakers of a language can tolerate more interference in the channel than can foreign speakers of that language. Presumably this results from the native's mastery of his language, which makes it possible for him to utilize its redundancy to a higher degree […].

Diesen Umstand machen sich verschiedene Testformate zu Nutze, um die Sprachkompetenz der Testteilnehmer ermitteln zu können. Zugrunde liegt hier folglich immer die Frage: Wie gut kann ein Testteilnehmer mit der durch Störungen im Kanal hervorgerufenen reduzierten Redundanz umgehen?

Man kann generell zwischen Tests unterscheiden, deren Textgrundlage entweder gesprochene oder geschriebene Sprache ist. Zu den Testformaten mit gesprochenem Input wird unter anderem das **Diktat** (sowie das **Teildiktat**) gerechnet. Für gewöhnlich liest hierbei die Lehrkraft einen Text in Abschnitten vor, und die Lerner sollen das Gehörte niederschreiben. Beim Teildiktat liegt den Lernern ein Teil des gehörten Textes bereits vor (vgl. CAI 2012: 182). Das klassische Diktat ist bereits seit Beginn des 20. Jahrhunderts Bestandteil von Sprachtests (vgl. STANSFIELD 1985: 122) und hat seither vielfach Kritik erfahren. LADO (1961: 34) ist der Auffassung, dass das Diktat gänzlich ungeeignet sei, um Sprachkompetenz zu messen. Er argumentiert, dass das Diktat keinen Satzbau teste, da die Reihenfolge der Wörter bereits vorgegeben sei. Aus dem gleichen Grund teste es auch keinen Wortschatz. Als Hörverstehenstest eigne sich das Diktat ebenfalls nicht, da der Lehrer den zu diktierenden Text meist sehr langsam vorlese. Lediglich zum Testen von Rechtschreibung und Interpunktion könne man das Diktat einsetzen. SOMARATNE (1965: 48) schreibt, dass das Diktat hauptsächlich ein Rechtschreibtest sei. ANDERSON (1953: 43) klassifiziert das Diktat als unangemessenen Hörverstehenstest, da dieses hier nur ungenau gemessen werde und das Testergebnis zudem maßgeblich von der Rechtschreibkompetenz der Prüflinge beeinflusst werde. Darüber hinaus kritisiert er, dass

bedingt durch die Schreibgeschwindigkeit der Testteilnehmer insgesamt nur wenig Text die Testgrundlage bilden könne.

Im Gegensatz dazu findet jedoch OLLER (1971) bei Englischlernern (n = 100) signifikante Korrelationen auf dem Signifikanzniveau 0,001 zwischen einem Diktat und verschiedenen anderen Sprachtests (Vokabeltest, Grammatiktest, Schreibaufgabe, phonologische Unterscheidung) sowie der Gesamtheit dieser Tests.[3] Darüber hinaus ist die Korrelation mit dem Diktat die höchste in jeder der möglichen Konstellationen. OLLER (1971: 259) stellt heraus:

> The student is tested for his ability to (a) discriminate phonological units, (b) make decisions concerning word boundaries in order to discover sequences of words and phrases that make sense, i. e. that are grammatical and meaningful, and (c) translate this analysis into a graphemic representation.

OLLER und STREIFF (1975: 81) reanalysieren die Daten von Oller und erhalten noch höhere Korrelationswerte zwischen dem Diktat und den oben genannten weiteren Teilen der Testbatterie. Sie schlussfolgern, dass das Diktat dazu geeignet sei, die von OLLER (1973: 113) als Kernstück der allgemeinen Sprachkompetenz postulierte *Grammar of Expectancy* zu messen (OLLER & STREIFF 1975: 78).

Auch VALETTE (1977: 110) ist der Meinung, dass das Diktat ein geeignetes und genaues Instrument zur Erfassung des Hörverstehens sei, da die Testteilnehmer den diktierten Text sowohl auf der Wort- und Satz- als auch auf der Textebene verstehen müssen. Darüber hinaus liefere das Diktat auch Informationen zur allgemeinen Sprachkompetenz bei fortgeschritteneren Lernern (vgl. VALETTE 1977: 243).

Das Diktat verfügt über kein einheitliches Bewertungssystem. Es besteht die Möglichkeit, einzelne Wörter oder längere Phrasen zu bewerten, Inter-

3 Der verwendete Korrelationskoeffizient wird nicht angegeben.

punktionsfehler können anders gewichtet werden etc. Wie VALETTE (1977: 244) jedoch herausstellt, kommt es beim Diktat vor allem darauf an, dass man beim Bewerten konsistent vorgehe, das konkrete Bewertungssystem sei dann zweitrangig.

IRVINE et al. (1974) legten 159 iranischen Englischlernern ein Diktat, einen Cloze-Test und den TOEFL vor. Sie korrelierten die Ergebnisse des aus zwei unterschiedlichen Passagen bestehenden Diktats gegen den TOEFL und ermittelten einen Wert für Pearsons Produkt-Moment-Korrelationskoeffizienten von r = 0,69 (vgl. IRVINE et al. 1974: 249). Die Korrelation des Cloze-Tests mit dem TOEFL ist jedoch höher.

Im Gegensatz zu reineren Tests der Kategorie *Reduced Redundancy Testing* wird bei einem klassischen Diktat keine bewusste Störung vorgenommen, d. h., es wird kein systematischer *Noise* in den Kanal eingeführt. Dennoch zählt auch das Diktat zu dieser Gruppe von Tests, denn wie LADO (1961: 34) feststellt: „The words can in many cases be identified by context if the student does not hear the sound correctly."

Beim sogenannten ***Noise Test*** hören die Testteilnehmer mehrere kurze Sätze, die sie niederschreiben sollen, ähnlich wie bei einem Diktat. Die Sätze sind hierbei jedoch unabhängig voneinander (vgl. CAULFIELD & SMITH 1981: 57), so dass der *Noise Test* die Sprachkompetenz nur auf Satzebene und nicht auf Textebene testet. Beim Hören der Sätze wird der Grad der Kanalstörung meist in Form von unterschiedlichen Graden weißen Rauschens[4] variiert. Dabei wird mit dem stärksten Grad der Störung begonnen, um einen Gewöhnungseffekt zu vermeiden (vgl. KLEIN-BRALEY 1997: 55). SPOLSKY et al. (1968: 92) fanden in ihrer Untersuchung hingegen keinen Lerneffekt, und es gibt auch Studien, die eine konstante Störung von z. B. 3 dB (vgl. WHITESON et al. 1962: 17) oder 6 dB (vgl. CAULFIELD

4 Weißes Rauschen bezeichnet ein Signal, dessen Leistungsdichte für alle Frequenzen konstant ist (vgl. KIENCKE & EGER 2008: 222).

& SMITH 1981: 54) verwenden. Während der *Noise Test* sehr gut zwischen Lernern und muttersprachlichen Sprechern einer Sprache unterscheidet, war es hingegen nicht möglich, nachzuweisen, dass das Hinzufügen von *Noise* für diese Differenzierfähigkeit verantwortlich ist. Somit kann der *Noise Test* nicht als eine erfolgreiche Operationalisierung der reduzierten Redundanz betrachtet werden (KLEIN-BRALEY 1997: 56). Ein weiteres Problem dieses Testformats ist, dass es kein einheitliches Vorgehen bei der Punktevergabe gibt. So vergeben beispielsweise SPOLSKY et al. (1968: 84) einen Punkt für einen komplett richtig ausgeschriebenen Satz und null Punkte in allen anderen Fällen, während CAULFIELD und SMITH (1981: 56) fünf Punkte für einen vollständig korrekten Satz vergeben, vier Punkte, wenn der Satz einen phonemischen Fehler enthält, einen Punkt, wenn zumindest ein Wort des Satzes korrekt ist und keine Punkte, wenn gar nichts korrekt ist. Zwar ist letztere Variante der Punktevergabe offenbar differenzierter, jedoch sind die Abstände zwischen den Punkten völlig willkürlich gewählt. Des Weiteren existiert auch eine Multiple-Choice-Variante des *Noise Tests*, bei der die Testteilnehmer unter mehreren Antwortoptionen den gehörten Satz auswählen müssen (vgl. VALETTE 1977: 116). Diese Art von *Noise Test* lässt sich im Gegensatz zur offenen Version schnell und objektiv auswerten. JOHANSSON (1973) kommt in seiner Untersuchung zu der Schlussfolgerung, dass der *Noise Test* kein geeignetes Instrument sei, um die allgemeine Kompetenz in einer Fremdsprache zu ermitteln. Dies begründet er zum einen damit, dass die allgemeine Sprachkompetenz meist besser ausfalle als es die Ergebnisse des *Noise Test* vermuten lassen. Zum anderen spielen gemäß seinen Ergebnissen psychologische Faktoren (u. a. Intelligenz oder die Ergebnisse in einem *Gestalt Completion Test*) dahingehend eine Rolle, dass sie das Testergebnis der Teilnehmer beeinflussen können.

Der klassische **Cloze-Test** (auch *Fixed-Ratio-Cloze-Test* genannt) gilt als der direkte Vorgänger des C-Tests. Anders als beim C-Test werden hier

auch literarische Texte (z. B. Gulliver's Travels, vgl. TAYLOR 1953: 426) als Testgrundlage verwendet. Der Cloze-Test wurde ursprünglich von TAYLOR (1953) als ein Messinstrument für Lesekompetenz in der L1 vorgestellt. 1971 wurde das Testprinzip von Oller und Conrad auf den L2-Bereich übertragen.[5] Der wie das englische Wort „close" ausgesprochene Name des Tests bezieht sich auf den ebenfalls englischen Begriff „closure" (TAYLOR 1953: 415), was so viel wie „Schließung" bedeutet und darauf Bezug nimmt, dass es die Aufgabe des Testteilnehmers ist, den Cloze-Test zu „schließen", also zu vervollständigen (vgl. OLLER & CONRAD 1971: 183). Hier wird der Bezug der Testverfahren zur Gestaltpsychologie deutlich, denn ebenso, wie es dem Auge gelingt, eine unvollständige Figur innerlich zu vervollständigen, soll der Testteilnehmer den mit Lücken versehenen Text komplettieren. Dies ist jedoch nur möglich, wenn der redundante Text – wie die Figur – dem Testteilnehmer hinreichend bekannt ist. OLLER (1979: 343) schreibt dazu:

> When the material is almost completely redundant […] the task would seem to be somewhat like the process of filling in the gaps in imperfect patterns.

Beim Cloze-Test wird in einem längeren Textabschnitt jedes n-te Wort vollständig gelöscht und somit eine *pseudo-random deletion* erzeugt (vgl. ALDERSON 1983: 205). Üblich ist hierbei die Löschung jedes fünften, sechsten oder siebten Wortes (vgl. OLLER & JONZ 1994: 3). Untersuchungen zu unterschiedlich viel Kontext zwischen den einzelnen Lücken eines Cloze-Tests zeigen, dass mehr als zehn bis zwölf Wörter Kontext keinen gesteigerten Beitrag zum Lösen einer Lücke leisten. Jedoch vergleicht ALDERSON (1983: 211) Cloze-Tests mit Tilgungsraten von jedem sechsten, achten, zehnten und zwölften Wort miteinander und kommt zu dem Er-

5 Für einen Überblick über die Cloze-Test-Forschung bis 1970 siehe BICKLEY et al. (1970).

gebnis, dass ein Cloze-Test-Text in Abhängigkeit von der gewählten Tilgungsrate unterschiedliche Fähigkeiten misst.

Die gelöschten Wörter werden durch einheitlich lange Striche ersetzt. Dies soll verhindern, dass die Testteilnehmer sich von der Strichlänge bei der Lösungsfindung beeinflussen lassen (vgl. TAYLOR 1953: 416). Die Form des Tests setze die von Oller postulierte Fähigkeit der *Practical Expectancy Grammar* (vgl. OLLER 1973: 113) in Gang und teste deren Funktionieren. Der Cloze-Test ist somit mehr als ein Lesetest, denn wie OLLER (1973: 114) zusammenfasst, tritt der Testteilnehmer mit dem Cloze-Test in einen „aktiven" Prozess ein:

> […] the process of taking a cloze test involves more than "passive" reading. By sampling the information that is present the subject formulates hypotheses, or expectations, about information that is to follow. By sampling subsequent sequences, he either confirms or disconfirms these expectations. If the expectations are disconfirmed they must be revised and new hypotheses must be formed.

Ebenso wie der *Noise Test* verfügt auch der Cloze-Test über kein einheitliches Bewertungssystem.

Eine weitere Variante des Cloze-Tests stellt der **Multiple-Choice-Cloze-Test** dar. Dieser ist vom Testaufbau her mit dem gewöhnlichen Cloze-Test identisch, jedoch muss der Lerner hierbei nicht ganz selbstständig auf die richtige Lösung kommen, sondern zu jeder Lücke wird ihm eine Auswahl mehrerer Antwortoptionen präsentiert, aus denen die korrekte Antwort ausgewählt werden muss (vgl. CRANNEY 1972: 61). Die Reliabilität dieser Testvariante war mit Werten von KR-20 > 0,8 (vgl. CRANNEY 1972: 62) durchaus zufriedenstellend. HINOFOTIS und SNOW (1980) erstellen einen Multiple-Choice-Cloze-Test, indem sie den bereits von HINOFOTIS (1980) verwendeten Cloze-Test so modifizieren, dass je 25 Lücken frei auszufüllen und 25 Lücken im Multiple-Choice-Format über vier Antwortoptionen zu füllen sind. Die Distraktoren wurden dabei aus

häufigen falschen Antworten der offenen Testversion zusammengestellt. Der Multiple-Choice-Cloze-Test zeigte höhere Mittelwerte als der offene Cloze-Test, d. h., die Multiple-Choice-Variante ist gegenüber der offenen leichter. Die Autoren fanden Korrelationen von 0,74 (*acceptable scoring*), 0,71 (*exact scoring*) und 0,63 (Multiple-Choice) ($n = 66$). Über den verwendeten Koeffizienten machen die Autoren keine Angabe. Das Multiple-Choice-Format ist hier zwar beiden Auswertungsmethoden der offenen Testform unterlegen, jedoch sind die Unterschiede auf dem 0,05-Niveau nicht signifikant (vgl. HINOFOTIS & SNOW 1980: 131 f.).

Eine noch einfachere Variante des Cloze-Tests stellt das sogenannte **Reading-Input-Format** dar. Auch hier wird der Text nach einem mechanischen Tilgungsprinzip beschädigt. Den Testteilnehmern werden dabei lediglich die Originallösung sowie ein Distraktor präsentiert, wie im folgenden Beispiel:

> The great murder wave of the 1970s appears to have ebbed at last in big-city America. (First/Have) reports from police departments (las/in) twelve selected cities show (from/that) in nine of them, (the/at) number of homicides dropped (markedly/cities) – and in some cases (swiftly/nine) – last year. (VALETTE 1977: 213)

Neben der Tatsache, dass hier jede Lücke mit einer Ratewahrscheinlichkeit von 50 % richtig gelöst werden kann, genügt es bei diesem Testformat häufig, zu erkennen, welche Wortart eingesetzt werden muss, ohne dass die Bedeutung eines Wortes exakt verstanden werden muss. So lässt sich hier bei der Lösungssuche auch nach dem Ausschlussprinzip verfahren.

Unter dem Begriff Cloze-Test werden noch weitere Testverfahren, die ebenfalls auf dem Prinzip der reduzierten Redundanz natürlicher Sprachen basieren, verstanden. Beim **Rational-Deletion-Cloze-Test** gibt es keine fixe Tilgungsrate. Stattdessen hat der Testersteller die Möglichkeit, aus einem gegebenen Textstück genau die Wörter zu löschen, die er möchte. Somit kann je nach Auswahl der zu tilgenden Wörter (z. B. Inhaltswörter

vs. Funktionswörter) kontrolliert werden, welche sprachlichen Kenntnisse mit dem Test abgefragt werden (vgl. CHAPELLE & ABRAHAM 1990: 122). BROWN (2002) argumentiert für diese Variante des Cloze-Tests, da bei einer durch ein mechanisches Tilgungsprinzip erzeugten Auswahl von Löschungen im Text zu viele Lücken produziert würden, die einen *Item Facility Value* von null aufweisen. Somit müssten Cloze-Tests sehr lang konstruiert werden, damit genügend nützliche Items in einem Text vorhanden sind (vgl. BROWN 2002: 109). Je höher das *n* der Tilgungsrate gewählt wird, desto länger folglich der Test. Er fasst zusammen:

> […] the every nth word strategy is far too inefficient for responsible use in decision-making. Instead, we should probably use what we now know about the way cloze items discriminate […] to refine the strategies we use to tailor cloze tests that are efficient. We need to show the cloze test "who is boss" by shaping them to our language testing purposes. In short, we need to tailor our cloze. (BROWN 2002: 110)

Beim sogenannten **Cloze-Elide-Test** trifft das *Cloze*-Prinzip eher im umgekehrten Sinn zu. Denn hierbei müssen die Testteilnehmer in aus wenigen Abschnitten bestehenden Textpassagen einzelne Wörter durchstreichen, die nicht in den Satz gehören. Ein Beispiel für einen solchen Test im Englischen ist folgender Textabschnitt, in dem die Wörter *slip, pledge, leisure, decrease, dissent* und *clean* gestrichen werden müssen:

> Poor reading slip may be the result of a person's not having read enough. Such a reader leisure usually lacks vocabulary and has decrease bad reading habits or unconsciously resist chance because dissent they think slow readers clean get more out of what they read. (KLEIN-BRALEY 1997: 83)

Sämtliche Varianten des Tests unterscheiden sich insofern vom Diktat und vom *Noise Test*, als dass ersterer auf geschriebener Sprache basiert, während letztere den Testteilnehmern einen Audio-Input präsentieren. Dies gilt ebenso für den **L-Test**, einen weiteren auf dem Prinzip der reduzierten Redundanz basierenden Test (‚L' für *Letter-Deletion Procedure*). Ziel dieses

Testansatzes ist es, dass die Vorteile des *Rational Deletion Systems* und des C-Tests vereint werden. Abhängig von der Zielsprache werden hier flexible Tilgungsmuster verwendet:

> a certain number of letters may remain undeleted at the beginning (in inflected languages also at the end) of item words; this number varies from 0 (when the item is blank, as in a cloze-test) to about n/2, where *n* is the number of letters in the item word (in which case the item looks like a C-Test item); the distance between items varies according to the RD [rational deletion, Anmerkung K. Z.] selection system; number of items is determined by text-length, etc. (KOKKOTA 1988: 116)

Der L-Test hat im Gegensatz zu den weiteren vorgestellten Testformaten keine Verbreitung gefunden.[6]

Trotz seiner weiten Verbreitung und seiner verschiedenen Varianten hat das Cloze-Test-Format rege Kritik erfahren. Wie im Folgenden gezeigt wird, boten die Schwachstellen des Tests den Anreiz zur Optimierung des Testformats und somit zur Entwicklung des C-Tests.

2.2 Vom Cloze-Test zum C-Test

2.2.1 Kritik am Cloze-Test

Wenngleich OLLER (1973: 106) die Entwicklung des Cloze-Tests als „nothing less than a stroke of raw genius" bezeichnet, bietet dieses Testformat zahlreiche Kritikpunkte. RAATZ und KLEIN-BRALEY (2002: 77) merken an, dass zahlreiche Studien, die von der hohen Reliabilität und Validität des Tests berichten, aus den USA stammen und in äußerst heterogenen Lernergruppen durchgeführt wurden. Studien, die mit homogeneren Ler-

6 Ein konkretes Beispiel für einen L-Test findet sich in der Literatur nicht.

nergruppen arbeiten (vgl. ALDERSON 1979; ALDERSON 1983; KLEIN-BRALEY 1981) berichten indes von diversen Problemen mit dem Testformat. RAATZ und KLEIN-BRALEY (2002: 77) halten fest, dass ein Test relativ lang sein muss, um bei einer Tilgungsrate von meist jedem fünften bis jedem zehnten Wort über genügend Items zu verfügen.

Dadurch ergibt sich außerdem das Problem, dass es durch die Verwendung eines einzigen Textes zu einer Verzerrung kommen kann, etwa wenn Testteilnehmer eines Geschlechts besser abschneiden als andere. Ein weiterer Kritikpunkt ist die Tatsache, dass Tests über kein einheitliches Bewertungsverfahren verfügen (siehe oben). So ist es beispielsweise möglich, ausschließlich die Originallösung als korrekt zu bewerten (*exact scoring*) oder aber auch akzeptable Varianten als korrekt anzusehen (*acceptable scoring*). Beim *acceptable scoring* hat die Auswertung stets ein subjektives Moment. Dies macht die Auswertung nicht nur arbeitsintensiver, sondern, so könnte man meinen, auch unzuverlässiger. Jedoch vergleicht HINOFOTIS (1980: 127) die Ergebnisse beider Auswertungsmethoden bei einem Test mit 50 Lücken und kommt zu dem Ergebnis, dass das Akzeptieren von korrekten Alternativlösungen (KR-20 = 0,85) einer exakten Auswertung (KR-20 = 0,61) in puncto Reliabilität überlegen ist. Beim Korrelieren der Testergebnisse gegen den TOEFL zeigt sich außerdem, dass das *exact scoring* (0,71) signifikant niedrigere Werte liefert als das *acceptable scoring* (0,75). Der genutzte Korrelationskoeffizient wird nicht genannt.

Auch IRVINE et al. (1974) korrelieren einen Test mit beiden Auswertungsmethoden gegen den TOEFL und finden Werte von r = 0,78 (*exact scoring*) und r = 0,79 (*acceptable scoring*). Beide Auswertungsmethoden korrelieren mit r = 0,94 so hoch miteinander, dass es nur einen geringen Unterschied mache, welche Methoden man anwende (vgl. IRVINE et al. 1974: 250). Auch STUBBS & TUCKER (1974) korrelieren beide Auswertungsmethoden eines Cloze-Tests gegeneinander und finden mit 0,97 einen noch

höheren Wert. Diese Ergebnisse scheinen darauf hinzudeuten, dass die unterschiedlichen Auswertungsmethoden keinen problematischen Faktor beim Test darstellen. Es gibt jedoch noch weitere Aspekte, die hier eine Rolle spielen.

Die Auswertungsmethode ist auch deshalb ein erheblicher Faktor, da im Falle einer originalgetreuen Auswertung Tests häufig auch für Muttersprachler nur schwer lösbar sind. Dies stellt wiederum ein Problem dar, weil Sprachtests von Muttersprachlern (nahezu) perfekt lösbar sein sollten, da der Test sonst für Fremdsprachenlerner zu schwierig ist (vgl. BAUR et al. 2013: 4) oder etwas anderes als reine Sprachkompetenz misst. Wählt man hingegen ein Bewertungssystem, das auch Alternativlösungen zulässt, so kommt unweigerlich ein subjektives Moment ins Spiel. Darüber hinaus macht das *acceptable scoring* Testauswerter nötig, die über eine nahezu muttersprachliche Kompetenz in der Zielsprache verfügen (vgl. IRVINE et al. 1974: 250). Der Schwierigkeitsgrad von Tests variiert je nach gewählter Tilgungsrate. Dabei ist es jedoch nicht so, dass eine höhere Tilgungsrate von beispielsweise jedem sechsten Wort einen Cloze-Test-Text schwieriger macht als eine Tilgungsrate von beispielsweise jedem achten Wort (vgl. ALDERSON 1979: 113 f.). Vielmehr zeigte sich, dass der Schwierigkeitsgrad von Cloze-Tests davon abhängt, wie sich das Verhältnis zwischen den gelöschten Inhalts- und Funktionswörtern gestaltet. Dies ist ein Hinweis darauf, dass nicht alle Cloze-Tests Äquivalente sind, sondern, dass die gewählte Tilgungsrate den Schwierigkeitsgrad eines Textes durch die damit einhergehende Umverteilung der getilgten Funktions- und Inhaltswörter beeinflusst. Ebenso hängen die Faktoren Reliabilität und Validität sowohl vom gewählten Text als auch von der Tilgungsrate sowie vom Beginn der Tilgung ab. BROWN (2002: 79) stellt fest, dass die Datenlage bezüglich der Reliabilität und der Validität des Cloze-Tests extrem widersprüchlich ist, da die ermittelten Werte im Falle der Reliabilität je nach Studie zwischen 0,13

und 0,96 sowie bei der Validität zwischen 0,06 und 0,91 schwankten. Unklar bleibt, ob es sich hier um Werte für Cronbachs Alpha oder Kuder-Richardson handelt. Dies sei unter anderem darauf zurückzuführen, dass je nach Niveaustufe der getesteten Lerner unterschiedliche Items den Sprachstand gut messen. Hierzu führt er eine eigene Studie (BROWN 1994) an, bei der je nach Probandengruppe Reliabilitäten zwischen 0,31 und 0,95 sowie Validitäten zwischen 0,43 und 0,90 errechnet wurden.

2.2.2 Prinzip des C-Tests

Der C-Test wurde von Ulrich Raatz und Christine Klein-Braley erstmals 1982 auf einer internationalen Tagung in England vorgestellt (vgl. URL 23: C-Test: Der Sprachtest: Wie sieht ein C-Test aus?) und mit dem Ziel entwickelt, diverse mit dem Cloze-Test-Format verbundene Probleme zu lösen:

> Our aim in developing the C-test was to retain the underlying theory but to improve the sampling process in test development and therefore in subject performance. (KLEIN-BRALEY 1985a: 83)

Die Ähnlichkeit des C-Tests mit dem Test spiegelt sich auch im Namen des Tests wider, wobei das „C“ in C-Test für „Cloze“ steht (vgl. KLEIN-BRALEY 1997: 63). Beiden Testformaten ist gemein, dass die Testteilnehmer einen mit Lücken versehenen Text wieder herstellen bzw. die Lücken schließen müssen.

Der neu zu entwickelnde Test sollte, wie auch der Cloze-Test, über ein kanonisches Tilgungsprinzip verfügen. Im Gegensatz zum Cloze-Test sollte er jedoch kürzer sein und die Testgütekriterien zuverlässiger und eindeutiger erfüllen, als dies beim Cloze-Test der Fall war. Des Weiteren sollte der neue Test leicht zu erstellen und für Muttersprachler (nahezu) perfekt lösbar sein – auch dies ist ein Kritikpunkt am Cloze-Test, insbesondere bei einer Auswertung nach der Methode des *exact scoring*.

Ein C-Test besteht aus mehreren kurzen Texten von je 60 bis 100 Wörtern (vgl. GROTJAHN 2004: 535; SCHLAK et al. 2010: 15). Fiktionale Texte, verbaler Humor, direkte Rede sowie kulturspezifische Inhalte gilt es zu vermeiden (vgl. GROTJAHN 2002: 222). Jeder Text hat einen Titel. Zudem bleibt der erste Satz eines C-Test-Textes unversehrt, um als Kontext bzw. Einleitung in den Test zu fungieren. Beim zweiten Satz beginnt nun das kanonische Tilgungsprinzip: Bei jedem zweiten Wort wird jeweils die zweite Hälfte gelöscht. Dieses Prinzip bezeichnen KLEIN-BRALEY & RAATZ (1984: 136) als *rule of two*. Es wird angewandt, bis die gewünschte Anzahl von Lücken erreicht wird. Ein letzter Satz bleibt ebenfalls unversehrt als Kontext stehen. Der folgende Text (Abb. 3) liefert ein Beispiel für einen nach diesem kanonischen Tilgungsprinzip erstellten C-Test .

Fragen zur Berufswahl

Alte Berufe verschwinden, neue kommen hinzu: Bei d_____ Berufswahl ha_____ Schüler im_____ wieder Fra_____ oder Prob_____. Denn e_____ gibt ei_____ große Anz_____ sehr versch_____Berufe, u_____ es i_____ nicht ein_____, die rich_____ Wahl z_____ treffen. D_____ berufliche Zuk_____ sollte m_____ rechtzeitig pla_____. Dabei ka_____ es sinn_____ sein, sich beim Arbeitsamt beraten zu lassen. Manchmal hilft auch ein Test zu den persönlichen Berufsinteressen.

Abb. 3: Beispiel für einen C-Test (URL 24: onSET: Beispieltest)

Wie obiges Beispiel zeigt, sind die auszufüllenden Lücken üblicherweise von gleicher Länge. Üblich sind Testsets mit fünf Texten à 20 Lücken oder vier Texte à 25 Lücken (vgl. GROTJAHN 2002: 222), jedoch gibt es auch

Abweichungen davon. So besteht der in dieser Studie genutzte onDaF[7] (vgl. URL 3: onDaF) aus acht Texten à 20 Lücken. Obwohl ein C-Test aus mehreren Texten besteht, ist der Test insgesamt ökonomischer als der Cloze-Test, da bedingt durch das oben geschilderte Tilgungsverfahren mehr Lücken auf weniger Raum erzeugt werden können. Die Zusammensetzung eines C-Tests aus mehreren kurzen Texten hat zudem den Vorteil, dass sich die einzelnen Texte thematisch voneinander unterscheiden (sollen) und somit etwaige Vorteile einzelner Testteilnehmer, die zufällig auf dem einen oder anderen Gebiet besonders bewandert sind, ausgleichen können. Mit anderen Worten: Eine *test bias* kann hierdurch auf das ganze Testset bezogen vermieden werden.

Neben diesem auch von Computern durchführbaren Tilgungsprinzip (vgl. GERMANN 1996: 428 ff.; KOLLER & ZAHN 1996: 403 ff.) existiert eine Reihe weiterer Konventionen, die es bei der Erstellung von C-Tests zu beachten gilt. Hierzu zählen zum Beispiel das Vermeiden von Doppeltilgungen und der Umgang mit Zahlwörtern und Eigennamen sowie Komposita.[8] Die Textgrundlage von C-Tests bilden authentische Texte, wie etwa Zeitschriftenartikel. Nicht geeignet sind (im Gegensatz zum Vorgehen bei Cloze-Tests) literarische Texte (vgl. GROTJAHN 2002: 222) sowie Texte, die Tabu-Themen (Sex, Tod, Religion, Politik, Drogen etc.) behandeln (vgl. ALTE 2005: 54). Um beispielsweise Doppeltilgungen zu umge-

7 Der vom TestDaF entwickelte C-Test heißt zum jetzigen Zeitpunkt (2018) „onSET" und ist für die Sprachen Deutsch und Englisch verfügbar (vgl. URL 25: onSET: Über onSET). Da jedoch zum Zeitpunkt der Datenerhebung für die vorliegende Studie die offizielle Bezeichnung des Tests „onDaF" war, wird ausschließlich dieser Begriff im Folgenden verwendet. Hierdurch wird zugleich deutlich, dass es sich um den deutschsprachigen C-Test handelt.

8 Eine detaillierte Übersicht und praxisorientierte Darstellung der zu berücksichtigenden Regeln bei der C-Test-Erstellung findet sich in GROTJAHN (2002: 222 ff.) und SCHÖN et al. (2012: 66 ff.).

hen, müssen C-Tests häufig modifiziert werden. Hierbei gilt die Regel, dass möglichst wenig am Text verändert werden soll (vgl. GROTJAHN 2002: 222).

Items sollen voneinander unabhängige Testaufgaben sein. Das bedeutet, dass das erfolgreiche oder nicht erfolgreiche Lösen eines Items keinen Einfluss auf das Lösen der anderen Items im Test haben darf (vgl. KRAUTH 1995: 25 f.). Bei den Lücken des C-Tests ist dies offenkundig nicht der Fall, führt doch das (korrekte) Lösen einiger Lücken dazu, dass insgesamt mehr Kontext gegeben ist, der dabei helfen kann, die verbleibenden Lücken zu füllen. Andererseits kann das falsche Ausfüllen einer Lücke den Testteilnehmer in die Irre führen und das weitere Lösen des C-Test-Texts behindern. Die Lücken eines C-Tests sind also nicht voneinander unabhängig, sondern nur die einzelnen Texte, weshalb man einen C-Test-Text auch als Super-Item bezeichnet (vgl. RAATZ 1985: 64 f.).

2.2.3 Gütekriterien

Das Format des C-Tests ist als ein objektives, reliables und valides Verfahren zur Ermittlung allgemeiner Sprachkompetenz bekannt. Jedoch müssen einige Aspekte beachtet werden, um das Potential des C-Tests hinsichtlich seiner Testgüte voll auszuschöpfen.

2.2.3.1 Objektivität

Das Gütekriterium der Objektivität gibt den „Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher" sind, an (LIENERT & RAATZ 1998: 7). Um die **Durchführungsobjektivität** eines C-Tests gewährleisten zu können, ist es zwingend notwendig, die Durchführungsbedingungen exakt festzulegen. Hierzu zählt einerseits eine Festlegung auf das genutzte Medium (computerbasierter C-Test vs. Papier-und-Bleistift-Version). An-

dererseits muss entschieden werden, ob die Texte eines C-Test-Sets als Gesamtheit präsentiert werden, was das Vor- und Zurückblättern durch die Testteilnehmer erlaubt und dazu führt, dass die Teilnehmer die Bearbeitungszeit individuell auf alle Texte verteilen. Dies hat den Nachteil, dass einzelne Testteilnehmer es möglicherweise nicht schaffen, alle Texte in der vorgegebenen Zeit zu bearbeiten, da sie sich bei einem Text zu lange aufhalten. Des Weiteren besteht nach GROTJAHN (2010: 271) die Gefahr der sogenannten „Gesamtdarbietung mit unvollständiger Bearbeitung". Dieser Fall liegt vor, wenn Testteilnehmer die einzelnen Texte eines Testsets unterschiedlich sorgfältig bearbeiten und sich bei den späteren Texten weniger Mühe geben (vgl. GROTJAHN 2010: 271). Um die Ergebnisse nicht zu verfälschen, ist es daher ratsam, die C-Test-Texte einzeln darzubieten und mit einer – großzügigen – Zeitbemessung zu kombinieren, so dass alle Testteilnehmer gleich viel Zeit auf das Lösen eines Textes verwenden.

Die Auswertung des C-Tests ist im Gegensatz zum Cloze-Test eindeutig. Zwar gibt es auch hier die Möglichkeit, zwischen einem *exact scoring* und einem *acceptable scoring* zu unterscheiden, jedoch sind die Möglichkeiten zu akzeptablen Alternativlösungen durch den bereits vorgegebenen ersten Wortteil stark beschränkt. Im Rahmen der Erprobung eines C-Tests können diese wenigen Ausnahmen entdeckt und in eine Lösungsschablone eingetragen werden (vgl. KOLLER & ZAHN 1996: 405). Ein Beispiel für einen solchen Fall wäre das Wort *da[her]*, welches semantisch und syntaktisch gleichwertig mit dem Wort *da[rum]* ersetzt werden könnte. Das Zulassen von derartigen Alternativlösungen erhöht laut GROTJAHN (2004: 542) leicht die Reliabilität eines C-Tests. Da es sich um wenige Ausnahmen handelt, die zudem im Vorfeld des Testeinsatzes erkannt werden können, ist beim C-Test im Gegensatz zum Cloze-Test eine Auswertung durch den Computer möglich – die **Auswertungsobjektivität** des C-Tests ist also ausgesprochen hoch. Für jede korrekt ausgefüllte Lücke erhält ein Testteil-

nehmer einen Punkt. Die Gesamtzahl der im Test erreichten Punkte addiert sich folglich aus in allen Texten korrekt gelösten Lücken.[9] Vorausgesetzt, dass ein C-Test erprobt und kalibriert wurde, sollte somit auch eine hohe **Interpretationsobjektivität** gewährleistet sein. Das bedeutet, dass das Testergebnis von allen Prüfern bzw. Bewertern gleichermaßen interpretiert wird, beispielsweise, ob der erreichte Punktwert zum Bestehen ausreicht oder welcher Niveaustufe des Gemeinsamen Europäischen Referenzrahmens das Ergebnis zuzuordnen ist.

2.2.3.2 Reliabilität

Auch die Reliabilität des C-Tests ist ausgesprochen hoch. Sie ist ein Maß dafür, wie zuverlässig die Ergebnisse eines Tests sind. Die beobachtbaren Ergebnisse setzen sich aus dem wahren Testwert sowie dem Fehlerwert zusammen (vgl. BACHMAN 1990: 167). BACHMAN (1990: 171) weist darauf hin, dass, „since we can never know the true scores of individuals, we can never know what the reliability is, but we can estimate it from the observed scores". Er nennt drei potentielle Einflussquellen, die unabhängig von dem zu messenden Konstrukt das Testergebnis beeinflussen können: (1) testmethodische Faktoren wie beispielsweise die Art der Anleitung, (2) Eigenschaften der Testteilnehmer wie beispielsweise Lernstile oder das Geschlecht, (3) zufällige und vorübergehende Faktoren wie Müdigkeit oder eine bestimmte Gefühlslage (vgl. BACHMAN 1990: 164).

Frühe Autoren der C-Test-Literatur haben zur Berechnung der Reliabilität des C-Tests meist die Kuder-Richardson-Formel angewendet. Durch Analyse der Varianz der einzelnen Testteile im Verhältnis zur Varianz des Gesamttests gibt diese an, ob und inwieweit die Items eines Tests die glei-

9 Das C-Test-Format bietet jedoch die Möglichkeit einer differenzierteren Analyse der Lücken (vgl. Kapitel 2.4).

che latente Variable messen (vgl. BACHMAN 1990: 177). Mit dieser Berechnung wurden für den C-Test hohe Reliabilitätswerte bis zu KR-21 = 0,97 erzielt (vgl. z. B. BABAII & SHAHRI 2010: 46). Da die Lücken eines C-Tests jedoch nicht als voneinander unabhängige Items betrachtet werden können, führt die Berechnung der Reliabilität mittels der Kuder-Richardson-Formel dazu, dass die Reliabilität überschätzt wird (vgl. BACHMAN 1990: 177). Daher werden die Lücken eines C-Test-Texts als Super-Item behandelt und heutzutage Reliabilitätsanalysen, wie von RAATZ (1985: 64) vorgeschlagen, meist mittels einer Überprüfung auf innere Konsistenz mit Cronbachs Alpha durchgeführt. Im Allgemeinen wird bei Tests eine Reliabilität von mindestens $\alpha = 0{,}7$ angestrebt, um ein ausreichend messgenaues Instrument zu erhalten (vgl. MOOSBRUGGER & KELAVA 2007: 11). Da der C-Test jedoch generell ein sehr reliables Verfahren darstellt, wird für dieses Testformat eine Reliabilität von mindestens $\alpha = 0{,}8$ erwartet (vgl. GROTJAHN 2004: 538). Oftmals weisen C-Tests sogar einen Wert für Cronbachs Alpha von $\alpha = 0{,}9$ oder noch höher auf. Da auch Reliabilitätskoeffizienten zwischen 0,7 und 0,85 (vgl. GROTJAHN et al. 1996) ermittelt wurden, wertet GROTJAHN (1995: 45) dies als einen Beleg „sowohl für die Messgenauigkeit der eingesetzten C-Tests als auch für die Stabilität der gemessenen Eigenschaft".

2.2.3.3 Validität

Validität bezeichnet die Gültigkeit eines Tests, d. h., sie gibt Auskunft darüber, ob und zu welchem Grad ein Test das misst, was er messen soll (vgl. LIENERT & RAATZ 1998: 10). Objektivität und Reliabilität sind notwendige, jedoch nicht hinreichende Voraussetzungen für ein valides Testverfahren. Ob ein Testinstrument eine hinreichende Validität aufweist, lässt sich nur im Kontext seiner Nutzung und damit einhergehenden Fragestellungen beantworten (vgl. BACHMAN 1990: 238). Vor dem Hintergrund,

dass der C-Test in zahlreichen Kontexten Anwendung findet, unter anderem als Instrument zur Einstufung oder zur Sprachstandsdiagnose im schulischen Bereich (vgl. Kapitel 2.4), kann die Frage nach seiner Validität kaum vollumfänglich bearbeitet werden.

Während Validität klassischerweise in verschiedene Ausprägungen unterteilt wird, wie beispielsweise Inhalts-, Konstrukt- und Übereinstimmungsvalidität, stellt BACHMAN (1990: 243) heraus, dass „while these have typically been discussed as different kinds of validity […] they can be more appropriately viewed as complementary types of evidence that must be gathered in the process of validation". Im Folgenden werden einige hier relevante Formen der Validität herausgegriffen und auf den C-Test angewendet.

Bei der **pragmatischen Validität** komme es laut RAATZ (1985: 62) lediglich darauf an, dass ein Test in der Praxis seine Tauglichkeit unter Beweis stellt, unabhängig von seinem tatsächlichen Inhalt oder einem zugrundeliegenden theoretischen Konstrukt. Die Eignung des C-Tests zur Einstufung von Sprachkursteilnehmern ist durch diverse Publikationen belegt (vgl. u. a. SCHÖN et al. 2012), so dass eine pragmatische Validität des C-Test-Formats zumindest für diesen Kontext als gegeben angesehen werden kann.

Schwieriger ist bereits die Frage nach **Inhaltsvalidität**. Diese wird nicht empirisch und durch rechnerische Verfahren, sondern durch eine Beurteilung der Items durch Experten ermittelt (vgl. LIENERT & RAATZ 1998: 11). GROTJAHN (2000: 39) betont die Fehlbarkeit von Expertenurteilen, was die Frage nach der Inhaltsvalidität problematisch macht. RAATZ und KLEIN-BRALEY (2002: 81) geben beispielsweise an, dass die Inhaltsvalidität des C-Tests auf Lesen und Schreiben beschränkt sei. SIGOTT (2004: 58) hingegen konstatiert, dass „their content has to be representative of nothing less than language as a whole", wurden C-Tests doch als Messinstrument

globaler Sprachkompetenz entwickelt. Insgesamt werde die Frage nach der Inhaltsvalidität recht stiefmütterlich behandelt.

Gut erforscht ist hingegen die **kriterienbezogene Validität** des C-Tests. Diese ist ein Maß dafür, inwieweit eine latente, d. h. nicht beobachtbare, Variable mit einer manifesten korreliert (vgl. BORTZ & DÖRING 2002: 199 f.). Im Falle von Sprachtests sieht dies in der Praxis meist so aus, dass überprüft wird, ob ein Test mit einem bestehenden (etablierten) Außenkriterium zusammenhängt, d. h., wie groß der Zusammenhang der von beiden Tests gemessenen Merkmale ist. Man spricht daher auch von **Übereinstimmungsvalidität** (engl. *concurrent validity*). Diese zu ermitteln, stellt ein gängiges Verfahren zur Validierung von Tests dar. Neben den in Kapitel 3.1 diskutierten Korrelationen, die der C-Test mit verschiedenen Außenkriterien zum Hörverstehen und Sprechen aufweist, existieren zahlreiche Untersuchungen, die den C-Test auch mit anderen (in unterschiedlichem Maß etablierten und standardisierten) Sprachtests korrelieren. Die in Tabelle 1 zusammengefassten Werte belegen den Zusammenhang des integrativen C-Tests mit verschiedenen sprachlichen Teilkompetenzen. Die gefundenen Korrelationen variieren stark, was sich durch mehrere Faktoren erklären lässt, nämlich variierende Zielsprachen und damit einhergehende Außenkriterien, unterschiedlich große Probandengruppen und verschiedene Korrelationskoeffizienten. Nicht zuletzt unterscheiden sich natürlich auch die C-Tests von Studie zu Studie. Die kriterienbezogene Validität oder Übereinstimmungsvalidität lässt sich folglich nicht verallgemeinern, sondern es können lediglich Aussagen für sehr begrenzte Kontexte gemacht werden, beispielsweise, in welchem Maße ein konkretes C-Test-Set mit einem spezifischen Kriterium zusammenhängt.

	min. Wert	in Studie	max. Wert	in Studie
Leseverstehen	r = 0,29	Grotjahn (1992)	r = 0,72	Jafarpur (2002)
Hörverstehen	0,16	Huhta (1996)	r = 0,69	Eckes (2014)
Schreiben	0,37	Raatz & Klein-Braley (1994)	r = -0,78[10]	Coleman (1994)
Sprechen	−0,38	Jakschik (1996)	ρ= 0,640	Arras et al. (2002)
Grammatik	0,25	Dörnyei & Katona (1992)	r = 0,76	Bolten (1992)
Wortschatz	0,22	Huhta (1996)	0,94	Chapelle & Abraham (1990)
Diktat	0,28	Jakschik (1992)	r = 0,71	Klein-Braley (1996)

Tab. 1: Korrelationen des C-Tests mit verschiedenen sprachlichen Teilkompetenzen in der L2[11]

Mit der sogenannten ***predictive validity*** wird angegeben, inwieweit man ausgehend vom Punktwert eines Testteilnehmers im Test auf dessen voraussichtliche Leistung in einem anderen Test schließen kann (vgl. Bachman 1990: 250). Von der zuvor genannten kriterienbezogenen Validität unterscheidet sich die *predictive validity* somit in der zeitlichen Abfolge der Testabnahmen. Identisch ist hingegen, dass die Ergebnisse zweier Tests miteinander in Verbindung gebracht werden. In seiner Funktion als *Screening*-Test im Rahmen des TestDaF weist der C-Test (vgl. URL 3: onDaF) eine zufriedenstellende *predictive validity* auf (vgl. Eckes 2014). Das bedeutet, Teilnehmer, die beim onDaF einen entsprechenden Punktwert erreichen, haben realistische Chancen, auch den (ungleich teureren) TestDaF erfolgreich zu bestehen. Der C-Test hat hier folglich eine gewisse Vorhersagekraft.

10 Die negativen Koeffizienten sind der gegenläufigen Kodierung der Tests geschuldet.

11 Sofern nicht angegeben, ist der Korrelationskoeffizient nicht dokumentiert.

Eine Sonderstellung nimmt die sogenannte **Augenscheinvalidität** (engl. *face validity*) ein. Sie ist ein Maß für die Akzeptanz eines Tests bei den Prüflingen: „Augenscheinvalidität gibt an, inwieweit der Validitätsanspruch eines Tests vom bloßen Augenschein her einem Laien gerechtfertigt erscheint" (MOOSBRUGGER & KELAVA 2011: 15). Wenngleich Augenscheinvalidität folglich nichts mit der realen Gültigkeit eines Tests gemein haben muss, ist sie ein bedeutsames Kriterium, denn eine geringe Augenscheinvalidität kann dazu führen, dass sich Testteilnehmer durch mangelndes Vertrauen in das Testformat nicht bestmöglich anstrengen (vgl. GROTJAHN 2000: 45).

LEGENHAUSEN (1989) untersuchte Lehrer- und Schülerurteile in Bezug auf die Augenscheinvalidität und Schwierigkeit des C-Tests. Hierbei zeigte sich, dass bei $n = 58$ durch Lehrer ausgefüllte Kommentarbögen von 70,7 % der befragten Personen angegeben wurde, dass der Schwierigkeitsgrad des C-Tests zu hoch sei, während niemand das Niveau als zu niedrig erachtete (vgl. LEGENHAUSEN 1989: 71). Immerhin 31 % der Befragten schätzten die Validität des C-Tests negativ ein, die verbleibenden 69 % der Kommentarbögen enthielten hierzu keine Angaben.

MÜNZEL (1989: 104) berichtet gar, dass sich fast die Hälfte von 37 angemeldeten Schülern vom Bundeswettbewerb Fremdsprachen (Einzelwettbewerb) zurückgezogen hätten, nachdem ihnen bekannt wurde, dass im Rahmen des Wettbewerbs ein C-Test abzulegen sei – ein weiterer Hinweis auf die geringe Akzeptanz des Testformats.

In einer Untersuchung von JAFARPUR (1995) wurde die Einschätzung bzgl. des C-Tests sowohl von Lernern als auch von Lehrern über einen Fragebogen erhoben. Von fünf Lehrkräften, die das C-Test-Format beurteilen sollten, gaben drei an, dass es sich um einen guten Test handele und dass dieser Englischkompetenz messe. Die beiden Lehrkräfte, die nicht

diese Meinung teilten, gaben an, dass der C-Test ihrer Ansicht nach eher Intelligenz oder Rechtschreibung messe (vgl. JAFARPUR 1995: 207).

In seiner Validierungsstudie erhebt HUHTA (1996) sowohl quantitative als auch qualitative Daten zur Augenscheinvalidität des C-Tests. 54,4 % der Probanden gaben an, den C-Test nicht für ein geeignetes Instrument zur Sprachstandsmessung zu halten (vgl. HUHTA 1996: 211). Die ungarischen Studenten gaben als Gründe für ihr Misstrauen an, dass der C-Test ihrer Auffassung nach neben Sprachkompetenz auch Vorstellungsvermögen teste (vgl. HUHTA 1996: 213).

SIGOTT (2004: 58) fasst treffend zusammen, dass „[…] C-Tests do not fare very well as far as face validity is concerned". Anzumerken bleibt jedoch, dass die hier aufgeführten Studien nicht aus jüngerer Zeit stammen. Da, wie in Kapitel 2.4 dargelegt, der C-Test inzwischen eine weite Verbreitung in Schulen und Universitäten gefunden hat, sind durch den erhöhten Bekanntheitsgrad auch Veränderungen in der subjektiven Wahrnehmung von Testteilnehmern denkbar.

Zentral im Rahmen einer Validitätsprüfung ist zudem die Frage nach der hier bisher nicht diskutierten **Konstruktvalidität**. Hier gilt es zu ergründen, ob der C-Test mit dem ihm zugrunde gelegten theoretischen Konzept zusammenpasst (vgl. KLEIN-BRALEY 1985b: 55). Da die Konstruktvalidität somit untrennbar mit dem Konstrukt des C-Tests verbunden ist, wird dieser Fragestellung im Folgenden ein eigenes Kapitel gewidmet.

Exkurs: Nebengütekriterien

Neben den bereits diskutierten Hauptgütekriterien von Tests erfüllt der C-Test auch eine Reihe weiterer, sogenannter Nebengütekriterien. Während die Hauptgütekriterien einheitlich und eindeutig sind, existieren verschiedene Angaben dazu, welche Aspekte zu den Nebengütekriterien gerechnet werden (vgl. u. a. LIENERT & RAATZ 1998: 11 ff.; GROTJAHN 2000: 47 ff.; MOOSBRUGGER & KELAVA 2011: 18 ff.). Einige sollen hier dennoch Erwähnung finden:

Der Forderung nach **Authentizität** wird der C-Test in zweierlei Hinsicht gerecht: Zum einen werden bei der Entwicklung von C-Tests ausschließlich authentische Textgrundlagen genutzt (vgl. GROTJAHN 2002: 222). Dieser Aspekt betrifft die Authentizität der Vorlage (vgl. GROTJAHN 2000: 50). Zum anderen ist das dem C-Test zugrundeliegende Prinzip der reduzierten Redundanz ein im Alltag durchaus auftretendes Phänomen (beispielsweise, wenn durch einen einfahrenden Zug die Ansage am Bahnsteig nur bruchstückhaft zu hören ist). Somit wird der C-Test auch der Authentizität der Testsituation gerecht (vgl. ibid.). KOW YIP CHENG et al. (2009) verglichen Tests auf Basis von authentischen und konstruierten Textgrundlagen und fanden heraus, dass die Authentizität einen signifikanten Einfluss auf die Testperformanz hat.

In Hinblick auf **Ökonomie** schneidet der C-Test ebenfalls äußerst gut ab. Dies zeigt sich nicht nur durch seine vergleichsweise kurze Durchführungsdauer und die sehr hohe Auswertungsobjektivität, sondern auch bei der Erstellung der C-Test-Texte. Die kurze Durchführungsdauer sichert zudem die **Zumutbarkeit** für die Teilnehmer.

Die **Fairness** eines C-Tests kann durch eine zu spezifische Themenwahl der Texte gefährdet sein. Kommt es beispielsweise aufgrund des Geschlechts der Testteilnehmer zu einer Verzerrung (engl. *bias*) der Testergebnisse, liegt eine konstruktirrelevante Varianz vor (vgl. GROTJAHN 2000: 47 f.). Diese gilt es mittels einer Pilotierung der Testaufgaben zu vermeiden.

2.2.4 Das Konstrukt des C-Tests

Unter dem Konstrukt eines Tests versteht man die Merkmale, die mithilfe der Aufgaben und den dazugehörigen Bewertungskriterien gemessen werden (vgl. AMERICAN RESEARCH ASSOCIATION et al. 2014: 11). Der C-Test wird häufig als ein geeignetes Instrument zum Messen globaler Sprachkompetenz bezeichnet, was nicht verwunderlich ist, wenn man bedenkt, dass er von Klein-Braley und Raatz, den „Eltern" des C-Tests, als eben solcher entwickelt wurde:

> […] the C-test was explicitly developed as a test of general language competence […] (KLEIN-BRALEY & RAATZ 1984)

> C-Tests are construct valid tests, and […] the construct they are operationalizations of could be called ‚general language proficiency'. (KLEIN-BRALEY 1985b: 65)

> Ferner gilt der C-Test […] allgemein als integratives Messinstrument zur weitgehend globalen Erfassung der Kompetenz in Erst-, Zweit- und Fremdsprachen. (GROTJAHN 1995: 38)

> Der C-Test eignet sich gut zur Erfassung globaler Sprachkompetenz. (ARRAS et al. 2002: 177)

> C-Tests dienen der Messung der allgemeinen Sprachkompetenz in der Muttersprache oder in einer Fremdsprache. (ECKES 2007: 68)

> C-tests have proved to be objective, highly reliable and very economical means for measuring global language proficiency. (GROTJAHN 2012: 181)

Diese Sammlung von Zitaten verschiedener Autoren aus unterschiedlichen Dekaden erweckt den Anschein, als gäbe es keinerlei Unklarheit in Bezug auf das Konstrukt des C-Tests. Dieser in der Literatur allgegenwärtigen Auffassung stehen jedoch einige Aspekte entgegen, die es zu beachten gilt: Zum einen stellt sich die Frage, was globale Sprachkompetenz eigentlich

ist, und zum anderen bleibt zu diskutieren, ob der C-Test diesem Konstrukt auch tatsächlich entspricht.

Das Konzept der allgemeinen Sprachkompetenz findet sich zuerst bei SPOLSKY et al. (1968). Sie stellen fest, dass subjektive Testverfahren nicht hinreichend reliabel sind, da die Ergebnisse von Testleiter zu Testleiter stark variieren können. Sogenannte *discrete point*-Tests auf der anderen Seite, die einzelne linguistische Aspekte testen und objektiver ausgewertet werden können, liefern zwar Informationen über Sprachwissen, sagen jedoch nichts darüber aus, wie erfolgreich jemand sich in einer fremdsprachlichen Situation verhält (vgl. SPOLSKY et al. 1968: 80). Die Annahme Spolskys ist, dass Fremdsprachenlerner nicht viele Einzelkompetenzen, sondern eine „zentrale, integrative Sprachkompetenz, jene ‚over-all proficiency'" (VOLLMER 1982: 39) ausbilden, die „allen Sprachleistungen zugrunde liegt" (ibid.). Folglich fordern SPOLSKY et al. (1968) neue Messverfahren, die diese Lücke schließen. Ihre Arbeitshypothese lautet:

> THERE IS SUCH A FACTOR AS OVERALL LANGUAGE PROFICIENCY IN A SECOND LANGUAGE, AND IT MAY BE MEASURED BY TESTING A SUBJECT'S ABILITY TO SEND AND RECEIVE MESSAGES UNDER VARYING CONDITIONS OF DISTORTION OF THE CONDUCTING MEDIUM. (SPOLSKY et al. 1968: 81 [Versalschrift im Original]).

Wie aus dem Zitat ersichtlich wird, nehmen SPOLSKY et al. (1968) nicht nur an, dass es so etwas wie eine allgemeine Sprachkompetenz gibt, sondern sie machen zugleich einen Vorschlag, wie diese zu testen sei: Über eine Reduktion der sprachlichen Redundanz. Diesem Konzept folgen, wie bereits in den Kapiteln 2.1 und 2.2 diskutiert, sowohl der C-Test als auch sein Vorläufer, der Cloze-Test. Die Erwägungen SPOLSKYs et al. (1968) spiegeln Alltagserfahrungen wider:

> Underlying our proposal solution is the fact that the high degree of redundancy in a natural language system makes communication possible even

> when there is a considerable distortion of the conducting medium. Thus, we are able to understand radio and telephone messages even though the acoustic signal is reduced in band width, and even when there is much added noise; and we are able to read letters in what we call illegible handwriting. (SPOLSKY et al. 1968: 81).

Unterschiede in der Fähigkeit, mit reduzierter Redundanz umzugehen, so SPOLSKY et al. (1968: 81), resultierten aus einer unterschiedlich stark ausgeprägten Sprachkompetenz. Für die Autoren hat allgemeine Sprachkompetenz also quasi per definitionem einen sehr engen Bezug zur praktischen Sprachbeherrschung bzw. der Fähigkeit, mit einer Reduktion der Redundanz natürlicher Sprache umzugehen. Folgt man dieser Argumentation, scheint es einleuchtend zu sein, dass der C-Test ein Messinstrument globaler Sprachkompetenz ist.

OLLER (1983: 4 f.) argumentiert, dass jeder Art von Sprachverwendung – produktiver wie rezeptiver – eine einheitliche linguistische Kompetenz zugrunde liege, die er *Expectancy Grammar* nennt. Diese bezeichne die vom Lerner wie vom Muttersprachler „psychologically […] internalized grammar" (OLLER 1983: 4).

> A person speaking or writing is planning what to say next and monitoring the output to see whether or not it matches the intended meaning. A person listening or reading, on the other hand, is constantly generating hypotheses about what will come next in the sequence in terms of what the writer or speaker is intending to say. […] In both cases the planning ahead or the hypothesizing what will come next can be conceptualized in terms of grammar-based expectancies. (OLLER 1983: 4 f.)

Über eine korrelationsanalytische Untersuchung sowie durch eine Faktorenanalyse versucht OLLER (1983) die Existenz eines solchen allgemeinen Faktors zu belegen. Als einen solchen Beleg erwartet er, hohe Korrelationen zwischen verschiedenen Sprachtests (Schreibaufgabe, Wortschatz, Grammatik, Phonologie, Diktat) zu finden. Darüber hinaus sollte eine Faktorenanalyse einen gemeinsamen grundlegenden Faktor erkennen lassen.

Die Ergebnisse seiner Studie erfüllen diese Erwartung. Mit Korrelationen von 0,54 (Phonologie/Schreiben) bis 0,79 (Diktat/Schreiben) und Ladungen von 0,77 (Phonologie) bis 0,93 (Diktat) auf den *g-factor* sieht Oller seine Hypothese bestätigt ($n = 164$) (vgl. OLLER 1983: 7). Weitere von Oller herangezogenen Datensätze unter Verwendung anderer Sprachtests (Wortschatz, Grammatik, Leseverstehen, Diktat, Cloze-Test) lieferten ähnliche Ergebnisse (vgl. OLLER 1983: 8–10) und schienen seine Hypothese zu stützen.

Die Relevanz von Ollers *Expectancy Grammar* ergibt sich nun daraus, dass dies das Konstrukt sei, auf dem der C-Test basiere:

> […] the tests are based on a linguistic theory, namely Oller's construct of pragmatic expectancy grammar, we claim that C-Tests are content valid tests. (RAATZ 1985: 42 [Unterstreichung im Original])

Jedoch blieb OLLERs (1983) Idee der *Expectancy Grammar* nicht ohne Kritik. Diese richtet sich gegen sein methodisches Vorgehen, das fehlerbehaftet gewesen sei. So nimmt beispielsweise PANG (1984) Bezug auf eine Studie von OLLER & HINOFOTIS (1980), die versucht hatten darzulegen, dass Sprachtests einen einheitlich globalen Faktor der Sprachkompetenz messen. Hierzu hatten die Autoren eine Faktorenanalyse durchgeführt, die diese These zu bestätigen schien. PANG (1984: 207) überprüfte die Ergebnisse von OLLER & HINOFOTIS (1980) und kam zu dem Ergebnis, dass die Interpretation der Ergebnisse durch die beiden Autoren fehlerhaft war und sich die Existenz eines globalen Faktors in allen Sprachtests nicht belegen lasse.

Auch Bachman stellt in einem Interview aus jüngerer Zeit heraus, dass die Ergebnisse aus Ollers Untersuchungen durch die fehlerhafte Anwendung einer explorativen Faktorenanalyse methodologisch anfechtbar seien und sich die Annahme Ollers, dass allgemeine Sprachkompetenz einem einzigen zu Grunde liegenden einheitlichen Faktor zugeordnet werden könne, nicht halten lässt (vgl. CHEN 2011: 282).

Die zweite zentrale Frage in diesem Kontext ist, ob globale Sprachkompetenz dem Konstrukt des C-Tests tatsächlich entspricht. Damit steht zugleich zur Diskussion, ob und in wieweit die diversen Varianten des C-Tests (vgl. Kapitel 2.3) das gleiche Konstrukt repräsentieren. Nicht zuletzt hängt das ermittelte Konstrukt auch von der Zielgruppe ab. So stellt beispielsweise SIGOTT (2006) für die Zielsprache Englisch heraus, dass weiter fortgeschrittene Lerner die Lücken des C-Tests tendenziell durch Einbezug des Kontexts auf Satzebene lösen können, während vergleichsweise weniger weit fortgeschrittene Lerner mehr auf den Kontext auf der Textebene angewiesen sind (vgl. SIGOTT 2006: 144). REICHERT et al. (2014) untersuchten den Einfluss der L1 der Testteilnehmer auf die Validität und somit das Konstrukt des C-Tests. Sie legten 1262 luxemburgisch-deutschsprachigen und 281 französischsprachigen Achtklässlern sowohl einen deutschen als auch einen französischen zuvor erprobten C-Test vor. Ihre Ergebnisse zeigten, dass die L1 der Testteilnehmer bei der Validität des französischen C-Tests eine größere Rolle spielt als beim deutschen C-Test.
Die Frage dessen, was sich hinter dem Begriff „globale Sprachkompetenz" oder „general language proficiency" verbirgt, ist eine in der Literatur viel diskutierte. Eine klare Antwort auf diese Frage findet sich hingegen nicht. VOLLMER (1981: 96) stellt fest:

> […] language proficiency is what language proficiency tests measure. This circular statement is about all one can firmly say when asked to define the concept of language proficiency […]. This is even more so when it comes to the construct of overall language proficiency […].

Um herauszufinden, was der C-Test wirklich misst, finden sich in der Literatur sehr unterschiedliche methodische Ansätze:

In zahlreichen Studien wird das C-Test-Konstrukt über Korrelationen mit etablierten Sprachtests erforscht. Dies ist der bei weitem am häufigsten gewählte methodische Ansatz. Hierbei werden nicht nur die vier Teilfer-

tigkeiten Hörverstehen, Leseverstehen, Sprechen und Schreiben fokussiert, sondern ebenfalls Wortschatz- und Grammatiktests sowie Selbsteinschätzungen durch die Lerner, Lehrerurteile und Schulnoten herangezogen, um ein möglichst umfassendes Bild zu erhalten. Die Ergebnisse dieser Untersuchungen sind sehr unterschiedlich, da sie für verschiedene Sprachen und mit diversen Außenkriterien durchgeführt wurden, deren Qualität ebenfalls variiert. Wenngleich dieses Vorgehen weit verbreitet ist, merkt FREESE (1994: 306 f.) hierzu kritisch an, dass man über das Konstrukt des C-Tests auf diesem Wege keine Informationen erhalten könne, da bei den Außenkriterien ebenfalls häufig unklar sei, was diese genau messen.

STEMMER (1991) nutzte in ihrer Dissertationsschrift die Verfahren der Introspektion und Retrospektion, um die mentalen Prozesse beim Lösen eines C-Tests zu erforschen:

> We will gain insight into the subject's problem solving behavior, how the problem is approached, which type of knowledge is activated and how the subject attempts to activate it. This in turn may allow us to ultimately infer what the C-test measures. (STEMMER 1992: 100)

In der ersten Phase der Studie sollten 30 deutschsprachige Probanden beim Lösen eines französischen C-Tests laut denken, um so Informationen darüber zu erhalten, welche Prozesse beim Ablegen des Tests aktiv sind.

In der sich daran anschließenden zweiten Untersuchungsphase wurde den Teilnehmern die zuvor gemachte Aufnahme beim Lösen des C-Tests vorgespielt. Die Teilnehmer konnten nun das Gehörte kommentieren oder Fragen des Testleiters hierzu beantworten. Stemmer fand heraus, dass die von den Teilnehmern herangezogenen Lösungsstrategien sich meist auf den unmittelbaren Arbeitsbereich beziehen, d. h. 50 % der Strategien beziehen sich auf die konkrete auszufüllende Lücke und weitere 38 % auf die direkte textuelle Umgebung (vgl. STEMMER 1992: 120–121). Die Autorin gelangt so zu der Schlussfolgerung, dass das Textverständnis beim Lösen des

C-Tests keine Schlüsselrolle spiele und hält den C-Test daher nicht für ein Messinstrument globaler Sprachkompetenz:

> […] if general language proficiency is meant to include high level comprehension then our results suggest that the C-test cannot be regarded as a measure of general language proficiency. (STEMMER 1992: 126)

Auch KONTRA und KORMOS (2006) wählten das Verfahren des lauten Denkens, um Strategien beim Lösen von C-Tests aufzudecken und somit auch Informationen über das Konstrukt des C-Tests zu erhalten. Zehn ungarische Studenten der Anglistik lösten drei C-Tests für die Zielsprache Englisch mit jeweils 20 Lücken und dachten dabei laut. Bei der Analyse der so gewonnenen Daten identifizierten Kontra und Kormos sowohl allgemeine als auch itembezogene Strategien, die von den Probanden angewendet wurden. Bei letzteren werden unter anderem lexikalische, morphologische, syntaktische sowie textuelle Strategien differenziert und deren Vorkommen quantitativ ausgewertet. Anders als STEMMER (1991, 1992) konstatieren KONTRA und KORMOS (2006: 130), dass „the occurrence of higher level comprehension can be observed". Zudem kommen die Autorinnen zu der Schlussfolgerung, dass die kommunikative Kompetenz auf lexikalischer Ebene und auf der Phrasenebene ebenfalls mit einem C-Test erfasst werden könne, jedoch sei er für die morphologische Ebene weniger geeignet (vgl. KONTRA & KORMOS 2006: 135).

RAATZ (1985) wählt das Verfahren der Faktoranalyse, um dem Konstrukt des C-Tests näherzukommen. Bei 75 Schülern der fünften Klasse (Haupt- und Realschule sowie Gymnasium) führte er eine umfangreiche Testbatterie durch. Diese umfasste neben einem Diktat und einem mehrteiligen Diagnosetest (aktiver und passiver Wortschatz, Analogien finden, Textstrukturierung, Leseverstehen, Verstehen von Anweisungen) auch Tests zur nonverbalen Intelligenz und zur Konzentrationsfähigkeit. In die Analyse flossen darüber hinaus auch die Schulnoten in den Fächern

Deutsch (L1), Englisch (L2) und Mathematik mit ein, ebenso wie ein Lehrerurteil über die Grammatikkenntnisse der Schüler. RAATZ (1985: 51) identifiziert zwei Faktoren: Faktor I sei als „general language proficiency" zu bezeichnen, da hier auch das Diktat als ein integrativer Sprachtest dazu gehöre und auch die Schulnoten in den Fächern Deutsch und Englisch bereits als globale Einschätzungen der Sprachkompetenz angesehen werden könnten. Mit diesem Faktor weist der C-Test eine hohe Ladung von 0,71 auf. Faktor II besteht im Gegensatz dazu aus der nonverbalen Intelligenz, der Schulnote im Fach Mathematik, dem Konzentrationstest, dem Verstehen von Anweisungen und den Analogien. Zu diesem Faktor, den RAATZ (1985: 51) als „speeded factor of logical thinking" bezeichnet, zeigt der C-Test mit einem Wert von 0,33 keine hohe Ladung. Die Schlussfolgerung des Autors lautet:

> C-Tests measure chiefly general language proficiency, which is operationalised in the teacher ratings, the orthography test and the construct of successive extraction of information. (RAATZ 1985: 50)

Wenngleich RAATZ (1985: 53) selbst darauf hinweist, dass es ein generelles Problem ist, dass bei Faktorenanalysen immer nur diejenigen Faktoren eine Rolle spielen können, die man in Betracht zieht, ist dieses Ergebnis dennoch interessant, da es zumindest einen Hinweis darauf gibt, was der C-Test nicht misst, also beispielsweise scheinen nonverbale Intelligenz und Konzentrationsfähigkeit nicht Bestandteil seines Konstrukts zu sein.

ECKES und GROTJAHN (2006) bemängeln methodische Probleme bei zahlreichen Studien, die sich mit der Konstruktvalidität des C-Tests beschäftigen. Diese reichen von einer Nichtbeachtung der Abhängigkeit der Lücken voneinander bis hin zu unbekannten Reliabilitäten der gewählten Messinstrumente. Um derartige Schwierigkeiten zu umgehen, wählen sie in ihrer Untersuchung eine Raschanalyse und kombinieren diese mit dem Verfahren der Faktorenanalyse. Als Außenkriterium wählten die Autoren

den aus vier Subtests bestehenden TestDaF (vgl. URL 4: TestDaF): Hörverstehen, Leseverstehen, schriftlicher Ausdruck und mündlicher Ausdruck. Ihre Hypothese lautete, dass „the C-test and the four TestDaF sections would each relate to the same latent variable (i. e. general language proficiency)“ (ECKES & GROTJAHN 2006: 300). ECKES und GROTJAHN (2006) arbeiteten mit zwei unterschiedlichen Modellen: In Modell I wurden die Teile des TestDaF in rezeptive und produktive Fertigkeiten aufgeteilt. Der C-Test wurde einmal dem rezeptiven (Ia) und einmal dem produktiven Faktor (Ib) zugerechnet. In Modell II wurde im Gegensatz dazu eine Unterscheidung zwischen schriftlichen und mündlichen Fertigkeiten vorgenommen. Auch hier wurde der C-Test einmal zu dem schriftlichen (IIa) und einmal zu dem mündlichen Faktor (IIb) gerechnet. Hinzu kam ein Ein-Faktor-Modell, bei dem der C-Test allen vier Fertigkeiten nebengeordnet wurde. Abhängig von der Art der aus den verschiedenen Tests vorliegenden Daten, wurden geeignete Raschmodelle gewählt und mit WINSTEPS bzw. FACETS angewendet. Es stellte sich heraus, dass die Anpassungsgüte in allen Modelltypen sehr gut war. Jedoch waren die Zweifaktorenmodelle dem Ein-Faktor-Modell leicht überlegen. Die besten Ergebnisse wurden erzielt, wenn der C-Test den produktiven Teilen des TestDaF, also schriftlicher und mündlicher Ausdruck, zugeordnet wurde (Ib) (vgl. ECKES & GROTJAHN 2006: 315). Die Autoren interpretieren ihre Ergebnisse als einen Beleg dafür, dass der C-Test allgemeine Sprachkompetenz messe:

> In view of the consistently high factor correlations and the negligible differences in fit between the one-factor model and the two-factor models, it seems reasonable to prefer the more parsimonious one-factor model.“ (ECKES & GROTJAHN 2006: 315)

HASTINGS (2002) versucht das Konstrukt des C-Tests über eine explorative Analyse der beim Ausfüllen des Tests gemachten Fehler zu ergründen. Die

Datengrundlage bilden 200 C-Tests, die von internationalen Studenten in den USA bearbeitet wurden. Hastings unterzieht die Fehler einer qualitativen Analyse und berücksichtigt hierbei auch den sprachlichen Kontext der betreffenden Lücke im Text. Für das Textstück *Even a large piece of fish will cook i___ less than thirty minutes.* (HASTINGS 2002: 59) stellt er beispielsweise fest, dass der unmittelbare Kontext *i___ less than thirty minutes* zur Lösung *in* verleitet, während unter Hinzunahme des Wortes *cook* auch die Variante *it* als eine mögliche Lösung erscheint. Erst unter Beachtung des gesamten Satzes wird deutlich, dass *in* die korrekte Lösung ist. Anhand der Fehler, die beim Ausfüllen der Lücke von *sometimes* gemacht wurden, demonstriert Hastings, dass die falschen Varianten in ihrer Häufigkeit mit der Frequenz der Wörter im allgemeinen Sprachgebrauch zusammenpassen (vgl. HASTINGS 2002: 64). Er schlussfolgert, dass unbewusstes Wissen über die Häufigkeit von Wörtern beim Lösen von C-Tests offenbar eine Rolle spielt. Das Fazit seiner explorativen Fehleranalyse lautet:

> […] a C-test measures the ability to apply and integrate contextual, semantic, syntactic, morphological, lexical, and orthographic information and knowledge pertaining to a particular written language. Furthermore, the processing that is required for successful C-test performance seems comparable to natural language processing in both depth and complexity, and may in fact have much in common with natural language performance. (HASTINGS 2002: 66)

Dieses Fazit macht das Konstrukt des C-Tests nicht greifbar und erscheint in Anbetracht der punktuellen Analyse der fehlerhaft ausgefüllten Lücken recht kühn. Dennoch ist dieses ungewöhnliche Vorgehen insofern interessant und bereichernd, als dass es sich von den so häufig zu findenden korrelationsanalytischen Untersuchungen absetzt.

Auch BAGHAEI und GROTJAHN (2014) versuchen die mentalen Prozesse beim Lösen eines C-Tests zu ergründen. Sie zeigen mit ihrer Studie, dass das Genre der für einen C-Test gewählten Textgrundlage die mentalen Prozesse beim Ablegen eines C-Tests verändern kann. Dies habe wie-

derum Auswirkungen auf das Konstrukt des C-Tests, da je nach Texttypen ein anderer Bereich der Sprachkompetenz gemessen wird. Die Autoren kommen zu diesem Ergebnis nach einer Reanalyse der Daten von 200 türkischen Rückkehrern von DALLER und GROTJAHN (1999). Aus dieser Studie lagen C-Test-Texte auf Basis einer Zeitschrift für Deutschlerner sowie auf der Grundlage von germanistischen Fachtexten vor (vgl. DALLER & GROTJAHN 1999: 170–172). Während die Fachtexte eher die sogenannte *Cognitive Academic Language Proficiency (CALP)* widerspiegelten, entsprachen die Zeitschriftentexte mehr den *Basic Interpersonal Communication Skills (BICS)*[12] (vgl. BAGHAEI & GROTJAHN 2014: 169).

SIGOTT (2004) stellt heraus, dass eine Lücke im C-Test nicht immer die gleichen mentalen Prozesse triggern muss, abhängig davon, wie weit fortgeschritten der Testteilnehmer in der Zielsprache ist. Sigott spricht daher von einem *fluid construct* des C-Tests (SIGOTT 2006; vgl. auch SIGOTT 2004). Wenngleich es individuell verschieden sein kann, benötigen Anfänger tendenziell mehr Kontext zum Lösen einer C-Test-Lücke als weiter fortgeschrittene Lerner, da diese auf automatisiertes lexikalisches und morphosyntaktisches Wissen zurückgreifen können (SIGOTT 2004: 189–190). GROTJAHN (2011: 134) sieht in der Fluidität jedoch kein schwerwiegendes Problem in Bezug auf die Validität des C-Tests und betont, dass dieses Phänomen durchaus nicht unbekannt sei, denn auch in der Leseforschung habe man einen Zusammenhang zwischen dem Grad der Lesekompetenz und den Leseprozessen gefunden.

12 Das Begriffspaar *Basic Interpersonal Communication Skills (BICS)* und *Cognitive Academic Language Proficiency (CALP)* geht auf CUMMINS (1979) zurück. Während BICS eine alltagsbezogene kommunikative Kompetenz beschreibt, bezeichnet CALP eine bildungssprachliche L2-Kompetenz, in die neben sprachlichem Wissen auch kognitive Fähigkeiten mit einfließen.

Wie gezeigt wurde, gibt es sehr unterschiedliche methodische Ansätze, um das Konstrukt des C-Tests zu erforschen. GROTJAHN et al. (2002: 101) betonen die Tatsache, dass „[c]onstruct validation of tests is an ongoing process which is never completed".

2.2.5 Der *speeded*-C-Test

Der *speeded*-C-Test (S-C-Test) stellt eine Variante des C-Tests dar, bei der die Bearbeitungszeit pro Text gegenüber dem herkömmlichen C-Test drastisch reduziert ist. Dieses Kapitel ordnet den S-C-Test zunächst in das Kontinuum von Niveau- und Geschwindigkeitstests ein. Im Anschluss wird die Erfüllbarkeit der Testgütekriterien beim S-C-Test diskutiert sowie sein Konstrukt bzw. die Änderung seines Konstrukts behandelt.

2.2.5.1 Niveau- und Geschwindigkeitstests

In der psychologischen Testforschung wird innerhalb von Leistungstests zwischen sogenannten Niveautests (*Power-Tests*) und Geschwindigkeitstests (*Speed-Tests*) unterschieden (vgl. GROTJAHN 2010: 268). Diese Kategorien sind jedoch nicht absolut, sondern stellen vielmehr ein Kontinuum dar (vgl. LIENERT & RAATZ 1998: 35). Niveautests bestehen aus Aufgaben, deren Schwierigkeitsgrad kontinuierlich ansteigt. Insgesamt ist das Niveau so hoch, dass es nur sehr wenigen Testteilnehmern gelingt, alle Aufgaben korrekt zu lösen. Bei dieser Testform existiert entweder gar keine Zeitbegrenzung oder sie ist so großzügig bemessen, dass alle Testkandidaten unabhängig von ihrer persönlichen Arbeitsgeschwindigkeit dazu in der Lage sind, alle Aufgaben zu bearbeiten, wenngleich sie diese vielleicht nicht korrekt lösen können. ZIEGLER und BÜHNER (2012: 71) sprechen in diesem Zusammenhang von der „Ermittlung eines intellektuellen Leistungsniveaus".

Im Gegensatz dazu ist das Schwierigkeitsniveau der Aufgaben bei einem reinen Geschwindigkeitstest so gering, dass alle Testteilnehmer ohne die angesetzte Arbeitszeitbegrenzung dazu in der Lage wären, alle Aufgaben korrekt zu lösen. Die Frage, ob ein Testteilnehmer es schafft, alle Aufgaben zu lösen, hängt bei einem Geschwindigkeitstest also ausschließlich von seiner Arbeitsgeschwindigkeit ab (vgl. GROTJAHN 2010: 268). Durch die *speed*-Komponente kommt es bei Geschwindigkeitstests zu einem sogenannten *Speed-Accuracy-Trade-Off* (vgl. ZIEGLER & BÜHNER 2012: 70). Das bedeutet, dass es durch den Zeitdruck zu Fehlern kommt, die den Testteilnehmern ohne eine Begrenzung der Bearbeitungszeit nicht unterlaufen würden.

Neben den oben beschriebenen Reinformen von Niveau- und Geschwindigkeitstests, existieren auch Mischformen. Solche Tests bestehen einerseits aus Aufgaben von aufsteigendem Schwierigkeitsgrad (wie Niveautests), verfügen jedoch zugleich über eine Zeitbegrenzung (vgl. ZIEGLER & BÜHNER 2012: 71). Wenngleich nicht aus allen Studien zum C-Test hervorgeht, mit welcher Bearbeitungszeit pro Text dieser administriert wird, so ist doch eine Zeitbemessung von fünf Minuten pro C-Test-Text eine gängige Variante (vgl. u. a. RAATZ & KLEIN-BRALEY 2002: 75; GROTJAHN 2004: 541; ECKES 2010: 125). Obzwar Niveautests ohne eine Zeitbegrenzung durchgeführt werden sollen, ist der mit fünf Minuten angesetzte Zeitrahmen pro C-Test-Text so großzügig bemessen, dass alle Testteilnehmer unabhängig von ihrer individuellen Bearbeitungsgeschwindigkeit dazu in der Lage sind, den C-Test-Text vollständig zu lesen und zu bearbeiten. Sowohl LIENERT & RAATZ (1998: 35) als auch WILHELM & SCHULZE (2002: 538) sowie GROTJAHN (2010: 269) weisen darauf hin, dass es aus Gründen der Testökonomie in der Praxis durchaus üblich ist, auch Niveautests zeitlich zu begrenzen. MOOSBRUGGER & KELAVA (2012: 79) sprechen von einer „Zeitbegrenzung […], die von den Probanden nicht als

Zeitdruck empfunden wird". Die hier beschriebene C-Test-Variante mit einer Bearbeitungszeit von fünf Minuten pro Text kann also durchaus als (reiner) Niveautest klassifiziert werden.

Als Niveautest zeigt der C-Test in einer Vielzahl von Studien signifikant positive Korrelationen mit vielen sprachlichen Teilfertigkeiten (Leseverstehen, Hörverstehen, schriftlicher und mündlicher Ausdruck) und -kompetenzen (Wortschatz, Grammatikkenntnisse) in unterschiedlichem Maße (siehe Kapitel 3.2, vgl. z. B. die Übersicht in ECKES & GROTJAHN 2006: 295 ff.).

Unter einem *speeded*-C-Test versteht man einen C-Test, bei dem die den Testteilnehmern zur Verfügung stehende Bearbeitungszeit deutlich von den für gewöhnlich angesetzten fünf Minuten abweicht, so dass der Test unter einem gewissen Zeitdruck abgelegt werden muss (vgl. GROTJAHN 2010: 283). Es existieren einige Studien, in denen der C-Test mit einer *speed*-Komponente versehen wird. Hierbei handelt es sich meist um Untersuchungen mit Muttersprachlern (vgl. RAATZ 2002), (muttersprachlichen) Kindern (vgl. WOCKENFUSS 2009) oder hochkompetenten Fremdsprachenlernern (vgl. AGUADO et al. 2007), in denen die Bearbeitungszeit pro C-Test-Text deutlich herabgesetzt wird (vgl. Kapitel 3.3). Trotz des so aufgebauten Zeitdrucks handelt es sich laut GROTJAHN et al. (2010: 301) in diesen Fällen ebenfalls um einen Niveautest, wenngleich mit einer *speed*-Komponente. Es wird damit argumentiert, dass beim C-Test mit Geschwindigkeitskomponente relativ schwierige Texte bearbeitet werden, die auch unter Niveau-Bedingungen nicht von allen Teilnehmern vollständig korrekt gelöst werden können. Dies stimmt mit der Aussage von LIENERT & RAATZ (1998: 35) überein, dass das entscheidende Merkmal zur Unterscheidung zwischen Niveau- und Geschwindigkeitstests die Art bzw. die Schwierigkeit der Aufgaben sei.

WILHELM & SCHULZE (2002: 551) konstatieren, dass die Frage, ob ein Test mit oder ohne Geschwindigkeitsfaktor eingesetzt wird, in der Praxis davon abhängen sollte, zu welchem Zweck der Test verwendet wird:

> If predictive validity is of particular importance, speeded reasoning ability could be the better choice, especially if the criteria show similar time pressure on performance.

Der Vorhersagekraft des C-Tests kommt insbesondere in seiner Funktion als Screening- und Einstufungstest eine zentrale Rolle zu. Nicht zuletzt wäre ein Test, der verbesserte Korrelationen mit den Kompetenzen Hörverstehen und Sprechen aufweist, besonders geeignet, um Lerner in (universitäre) Sprachkurse mit kommunikativer Ausrichtung einzustufen (vgl. Kapitel 3.1).

2.2.5.2 Gütekriterien beim *speeded*-C-Test

Für einen sinnvollen Einsatz des *speeded*-C-Tests muss zunächst sichergestellt werden, dass er die Testgütekriterien erfüllt. Eine direkte Übertragung der zu den Gütekriterien des C-Tests gemachten Aussagen (vgl. Kapitel 2.2.3) auf den S-C-Test ist nicht möglich. Relativ unproblematisch scheint das Kriterium der Objektivität zu sein. Durchführungs- und Auswertungsobjektivität sind unter standardisierten Bedingungen für den S-C-Test ebenso zutreffend wie für den C-Test, und auch die Interpretationsobjektivität wird durch das Hinzufügen einer Geschwindigkeitskomponente nicht beeinträchtigt, sofern der Test hinreichend erprobt ist und *cut scores* ermittelt wurden.

Ob der S-C-Test über eine ausreichende Reliabilität verfügt, gilt es hingegen empirisch zu überprüfen (vgl. Kapitel 4.4.3.1). Der Aufbau von Zeitdruck beim Bearbeiten eines Tests kann dazu führen, dass nicht alle Teilnehmer es schaffen, die Texte vollständig zu bearbeiten. Dies kann wiederum der Reliabilität abträglich sein (vgl. GROTJAHN et al. 2010: 297).

Aus diesem Grund ist es im Falle des S-C-Tests auch von besonders großer Bedeutung, die Texte eines C-Tests in Form der Einzel- und nicht der Gesamtdarbietung zu präsentieren (vgl. Kapitel 2.3.4). GROTJAHN et al. (2010: 303) spekulieren jedoch, dass die Reliabilität des S-C-Tests höher ausfallen könnte als die des herkömmlichen C-Tests, denn:

> Geht man davon aus, dass die beiden latenten Variablen „deklaratives Wissen" und „Grad der Automatisierung" bei weit fortgeschrittenen Lernern weitgehend miteinander korrelieren und dass die wahre Varianz stärker steigt als die Fehlervarianz, ist bei einem zeitreduzierten C-Test eine Erhöhung der Varianz der beobachteten Werte und zugleich eine (deutlich) erhöhte Reliabilität zu erwarten.

Dies bedeutet im Umkehrschluss jedoch auch, dass bei weniger weit fortgeschrittenen Lernern, bei denen es noch eine größere Diskrepanz zwischen deklarativem Wissen und dem Grad der Automatisierung gibt, ein Anstieg der Reliabilität durch drastisches Begrenzen der Bearbeitungszeit nicht zu erwarten ist.

HEINE (2017) setzt ein identisches Textset als C-Test und S-C-Test bei monolingualen Muttersprachlern, bilingualen Muttersprachlern und Späterwerbern (Erwerbsbeginn > 16 Jahre) des Deutschen ein und ermittelte die Reliabilität der beiden Testversionen für alle Gruppen separat. Tabelle 2 fasst die Werte von Cronbachs Alpha zusammen (vgl. HEINE 2017: 132–133).

	C-Test	S-C-Test
Monolingual (n = 80)	0,851	0,924
Bilingual (n = 50)	0,585	0,950
Späterwerber (n = 89)	0,959	0,986
gesamt	0,971	0,986

Tab. 2: Reliabilitäten von C-Test und S-C-Test

Die Reliabilität beider Testversionen ist – mit Ausnahme des C-Tests bei den bilingualen Muttersprachlern[13] – sehr hoch. Insbesondere gilt es hier zu beachten, dass die Werte für Cronbachs Alpha für den S-C-Test in allen Gruppen sowie im Gesamtergebnis deutlich höher ausfallen als für den C-Test. Dies zeigt nicht nur, dass das Format des S-C-Tests eine mehr als zufriedenstellende Reliabilität aufweist, sondern dass sich die Reliabilität eines C-Test-Sets durch Hinzufügen einer Geschwindigkeitskomponente entsprechend der Hypothese von GROTJAHN et al. (2010: 303) tatsächlich erhöht, zumindest in der Gruppe von muttersprachlichen Sprechern und sehr erfolgreichen Späterwerbern des Deutschen.

Entscheidend ist letztendlich, ob der S-C-Test auch in puncto Validität bestehen kann. Die oben angeführten Belege zur Validität des C-Tests lassen sich nicht ohne weiteres auf den S-C-Test übertragen. Während sich die Frage nach der Inhaltsvalidität eines Tests generell als problematisch erwiesen hat, lässt sich die Inhaltsvalidität eines C-Tests jedoch auf ein identisches C-Test-Set mit hinzugefügter Geschwindigkeitskomponente übertragen. Da Inhaltsvalidität über Expertenurteile ermittelt wird, steht eine empirische Überprüfung hier ohnehin nicht zur Debatte. Andere Ausprägungen der Validität wie Übereinstimmungsvalidität, Konstruktvalidität oder Vorhersagevalidität müssen jedoch für den S-C-Test empirisch erforscht werden. Die beim C-Test ohnehin schwache Augenscheinvalidität könnte durch den aufgebauten Zeitdruck und die somit als stressiger zu bewertende Testsituation weiter sinken. Die Augenscheinvalidität des S-C-Tests wird im Rahmen dieser Studie erstmals untersucht und mit der des herkömmlichen C-Tests verglichen (vgl. Kapitel 4.4.3.7).

13 HEINE (2017: 133) selbst findet nach Überprüfung der vorliegenden Datensätze keine Erklärung für den für einen C-Test unerwartet niedrigen Wert von $\alpha = 0{,}585$.

2.2.5.3 Konstrukt des *speeded*-C-Tests

Während das Konstrukt des C-Tests Fokus zahlreicher Untersuchungen unterschiedlicher Methodologie ist, liegen zum Konstrukt des S-C-Tests so gut wie keine Daten vor. Die wenigen Studien, die einen S-C-Test verwenden, liefern kaum Anhaltspunkte zum Konstrukt des C-Tests. Auch thematisieren die Autoren ein möglicherweise vom herkömmlichen C-Test abweichendes Konstrukt nicht oder nur implizit (vgl. Kapitel 3.3).

AGUADO et al. (2007: 144) nutzen in ihrer Studie die *speed*-Komponente, um die Schwierigkeit der C-Test-Texte zu erhöhen. Sie stellen folgende Hypothese auf:

> Die kompetente Verwendung von Sprache unter Echtzeitbedingungen setzt sowohl eine breite deklarative (sprachliche) Wissensbasis als auch eine ausreichende Prozeduralisierung (Automatisierung) des Wissens […] voraus.

Durch eine deutliche Reduktion der Testzeit könnte sich also die Korrelation mit den in Echtzeit ablaufenden sprachlichen Teilfertigkeiten Hörverstehen und Sprechen erhöhen. Diese Vermutung wird in GROTJAHN et al. (2010: 315) geäußert. Die empirische Überprüfung dieser Hypothese steht im Zentrum dieses Dissertationsprojekts. Sollte sie sich als richtig erweisen, könnte die Durchführungsdauer des Tests deutlich reduziert werden, was das Testformat noch ökonomischer machen würde. Auch Versuche, beim Sitznachbarn abzuschreiben, würden deutlich erschwert (vgl. GROTJAHN 2010: 267).

GROTJAHN et al. (2010: 303) sehen im S-C-Test ein Instrument zur integrativen Messung deklarativer und automatisierter Wissensbestände (vgl. Kapitel 3.1). Im Vergleich zum gewöhnlichen C-Test rechnen sie bei sehr fortgeschrittenen Lernern mit einer Erhöhung der Reliabilität und der Varianz, d. h. mit einer verbesserten Differenzierbarkeit von hochkompetenten Fremdsprachensprechern. Diese Annahme begründen die Autoren damit, dass deklaratives Wissen und automatisiertes Wissen miteinander

korrelieren und mit zunehmender Fremdsprachenkompetenz die wahre Varianz schneller steige als die Fehlervarianz.

Einige weitere Spekulationen zum Konstrukt des S-C-Tests finden sich in GROTJAHN (2010). Zunächst stellt Grotjahn heraus, dass es sich beim S-C-Test nicht um einen typischen Geschwindigkeitstest handelt, bei dem die Items so einfach konstruiert sind, dass bei genügend Bearbeitungszeit jeder Teilnehmer dazu in der Lage wäre, alle Lücken korrekt auszufüllen (vgl. Kapitel 2.2.5.1). Er weist darauf hin, dass bei einem Niveautest die Varianz der Testwerte und die Reihenfolge, in die die Testteilnehmer gebracht werden, gleich bleiben, wenn man die Bearbeitungszeit erhöht (vgl. GROTJAHN 2010: 268). Da es sich beim herkömmlichen C-Test durch die sehr großzügig bemessene Bearbeitungszeit um einen Niveautest handelt, ist im Umkehrschluss die Frage relevant, ob aus dem S-C-Test eine ähnliche Varianz der Testwerte und Reihenfolge der Testteilnehmer resultiert wie aus dem C-Test.

Einen ersten Hinweis auf eine Veränderung des Konstrukts des C-Tests durch Hinzufügen einer Geschwindigkeitskomponente liefern FADAEIPOUR & ZOHOORIAN (2017). Die Autorinnen korrelieren einen *power*-C-Test und einen *speeded*-C-Test, bei dem die Bearbeitungszeit auf die Hälfte reduziert ist, mit den Ergebnissen eines Leseverstehenstests. Dabei fanden sie einen höheren Wert für den *power*-C-Test (r = 0,65) als für den *speeded*-C-Test (r = 0,55), was die Autorinnen als einen Hinweis auf eine Konstruktveränderung deuten (vgl. ibid. 47). Zugleich kann die mittlere Korrelation des *speeded*-C-Tests mit Leseverstehen als ein erster Schritt auf dem Weg zur Validierung des S-C-Tests gewertet werden.

Ob und wie sich das Konstrukt des C-Tests durch das Hinzufügen einer Geschwindigkeitskomponente verändert, ist der Fokus der vorliegenden Arbeit. Aufgrund der gewählten zentralen Forschungsfrage nach dem Verhältnis von C-Test und S-C-Test zu den Fertigkeiten Hörverstehen und

Sprechen, kann zwar auch hier kein vollständiges Bild des S-C-Test-Konstrukts gezeichnet, aber ein Beitrag zur Beantwortung der Frage nach der Konstruktvalidität und der kriterienbezogenen Validität des S-C-Tests geliefert werden (vgl. Kapitel 4.4.3).

2.3 Varianten des C-Tests

Seit der Einführung des C-Tests im Jahr 1981 hat es eine Vielzahl von Weiterentwicklungen gegeben, so dass man heutzutage nicht mehr von *dem* C-Test sprechen kann (vgl GROTJAHN 2011: 132). Vielmehr existiert eine Vielzahl von Varianten, die im Folgenden vorgestellt und diskutiert werden. Hierbei lassen sich zunächst C-Test-Versionen für verschiedene (Fach)sprachen unterscheiden (vgl. Kapitel 2.3.1 und 2.3.2). Des Weiteren gibt es Ansätze, das kanonische Tilgungsprinzip des C-Tests zu modifizieren (vgl. Kapitel 2.3.3). Schließlich werden auch die Darbietungsbedingungen des C-Tests variiert (vgl. Kapitel 2.3.4). Hier fügt sich auch der bereits in Kapitel 2.2.5 besprochene *speeded*-C-Test ein.

2.3.1 Sprachvarianten

Wenngleich das C-Test-Format mit Hinblick auf die Sprachen Deutsch und Englisch entwickelt wurde, existieren inzwischen C-Test-Versionen für viele verschiedene Sprachen, darunter nicht nur flektierende, sondern auch agglutinierende und isolierende. Aufgrund des unterschiedlichen Aufbaus der Sprachen sind oftmals idiosynkratische Änderungen am kanonischen Tilgungsprinzip des C-Tests nötig.

Für den Bereich der romanischen Sprachen wurde vergleichsweise viel publiziert. Bereits in den 1980er Jahren wurden C-Tests für das **Französische** entwickelt (vgl. u. a. GROTJAHN & STEMMER 1984; WARD 1987) so-

wie für das **Spanische** (vgl. LÜTTICKEN 1985). Auch für das **Italienische** wurden C-Tests erstellt (vgl. GROTJAHN et al. 1994). Da die romanischen Sprachen von ihrer Typologie her ebenso wie das Deutsche und (mit Einschränkungen) das Englische flektierende Sprachen darstellen, ist die Übertragbarkeit des kanonischen C-Test-Prinzips hier weitestgehend unproblematisch. Sprachspezifische Regelungen müssen jedoch auch hier getroffen werden, unter anderem für die im Italienischen vorkommenden enklitischen Pronomen, beispielsweise die Präposition *di* (Dt. von) + *la* (Dt. die) = *della* oder der Infinitiv *vedere* (Dt. sehen) + *la* (Dt. sie/die) = *vederla*.

Auch für die slawischen Sprachen[14] gibt es einige Untersuchungen zum C-Test. STEURER (1986) berichtet vom ersten Einsatz eines **russischen C-Tests** im Rahmen des Bundeswettbewerbs für Fremdsprachen. Bei der Entwicklung des C-Tests wurden die Vorgaben von Raatz und Klein-Braley strikt befolgt und ein Set von vier Texten à 20 Lücken entwickelt. Das Testset wurde im Anfängerunterricht (eine genauere Angabe des Niveaus liegt nicht vor) im schulischen Kontext bei Schülern der Sekundarstufe II eingesetzt. Am kanonischen Tilgungsprinzip mussten offenbar keine Modifikationen vorgenommen werden. Statistische Ergebnisse werden nicht dokumentiert, jedoch schlussfolgert die Autorin, dass der C-Test zum Messen des Sprachstands im Russischunterricht geeignet sei (vgl. STEURER 1986: 88).

Für das **Persische** unternimmt BAGHAEI (2014) den Versuch einen C-Test zu entwickeln und zu validieren. Da das Persische über ein alphabetbasiertes Schriftsystem verfügt und zudem zu den flektierenden Sprachen zählt, konstatiert Baghaei, dass „Persian lends itself very well to the C-Test principle" (BAGHAEI 2014: 302). Er erstellt einen fünfteiligen C-Test mit je 25 Lücken und erprobt diesen bei iranischen *High School*

14 Zum C-Test für die polnische Sprache siehe die unveröffentlichte Masterarbeit von KUJAWKA (2011).

Schülern (n = 158), d. h. Muttersprachlern des Persischen. Für Cronbachs Alpha konnte ein beachtlicher Wert von $\alpha = 0{,}95$ ermittelt werden. Die C-Test-Texte weisen untereinander Korrelationen zwischen 0,85 und 0,91 auf, woraus BAGHAEI (2014: 304) auf die Einheitlichkeit des gemessenen Konstrukts schließt. Eine weitergehende Untersuchung an Lernern des Persischen steht derzeit noch aus.

Für Sprachen, die einem agglutinierenden Sprachbau folgen, zeigen sich hingegen größere Schwierigkeiten bei der Anwendung des C-Test-Prinzips. Da hier die grammatische Funktion eines Worts durch „Ankleben" (lat. *agglutinare* = kleben) eines oder mehrerer Suffixe bestimmt wird, würden bei strenger Anwendung des C-Test-Prinzips ggf. zu viele Informationen gelöscht, wie das folgende Beispiel (Tab. 3) aus dem Finnischen illustriert:

toimistoissani (in meinen Büros), toimist[oissani]

toimisto	-i	-ssa	-ni
Grundwort	Plural	Inessiv	Possessivsuffix

Tab. 3: Beispiel agglutinierender Sprachbau

Bei einer regelkonformen Tilgung des Wortes *toimistoissani* würde folglich die Information über den Numerus, den Kasus und das Possessivsuffix gelöscht sowie ein Teil des Grundworts selbst. Dennoch wurde der Versuch gemacht, auch für agglutinierende Sprachen C-Tests zu entwickeln. DALLER et al. (2002) wenden auch für das **Türkische** das kanonische *rule of two* an. Jedoch zeigt sich bei ihrem Test bei Muttersprachlern eine Lösungsquote von nur 75 %, was deutlich unter den zu erwartenden 90 % bis 95 % (vgl. GROTJAHN 2002: 216; URL 1: C-Test. Der Sprachtest) liegt. Als ein Hauptproblem bei der Anwendung des klassischen Tilgungsprinzips auf das Türkische nennen die Autoren die in der türkischen Sprache weitaus weniger stark vertretende Redundanz. Als Beispiel wird der Unter-

schied zwischen „du gehst“ und „gidiyorsun“ (DALLER et al. 2002: 191) angeführt, anhand dessen deutlich wird, dass die Information über die vorliegende 2. Person Singular im Deutschen auch durch das vorangehende Pronomen kodiert ist, was jedoch im Türkischen (sowie in allen *Pro-Drop*-Sprachen) nicht gegeben ist. Dennoch halten die Autoren das klassische Tilgungsprinzip für auf das Türkische anwendbar. Daller et al. variieren das klassische Tilgungsprinzip jedoch auch und schlagen das sogenannte *morpheme principle* vor, bei dem sich die Tilgung an Morphemen und nicht an Buchstaben orientiert. Jedes dritte Morphem soll hierbei gelöscht werden (vgl. DALLER et al 2002: 193 f.). Eine weitere Variante ist das *syllable principle*, bei dem, wie der Name bereits erahnen lässt, die Silben der Wörter die Grundlage der Tilgung bilden. Bei dieser Form wird jede dritte Silbe gelöscht. Schon BAUR und MEDER (1994) hatten diese Variation des C-Test-Prinzips im Rahmen des Projekts „Zweisprachigkeit und Schulerfolg ausländischer Kinder“ vorgeschlagen. Eine letzte Variante ist das *middle principle*, bei dem der mittlere Teil des zweiten Wortes gelöscht wird.

CAPREZ-KROMPÀK & GÖNÇ (2006) lehnen das klassische C-Test-Prinzip für agglutinierende Sprachen gänzlich ab. Sie schlagen das sogenannte *first suffix*-Prinzip vor, bei dem dem Aufbau eines Wortes mit mehreren Suffixen folgend jeweils das erste Suffix gelöscht wird (vgl. Tab. 4).

arkadaşlarımıza (unseren Freunden)			
arkadaş	-lar	-ımız	-a
Freund	Plural	Possessivsuffix	Dativ

Tab. 4: Beispiel first-suffix-Tilgung (vgl. CAPREZ-KROMPÀK & GÖNÇ 2006: 254)

Die Autoren erreichten mit diesem Tilgungsverfahren eine Reliabilität von $\alpha = 0{,}74$ (bei $n = 18$), was unter dem erwartbaren Wert von $\alpha = 0{,}8$ liegt (vgl. GROTJAHN 2002: 214). Zu beachten ist hier außerdem die äußerst

kleine Probandenzahl, weshalb der relativ niedrige Wert für Cronbachs Alpha durchaus erwartbar ist.

Ebenso wie das Türkische folgt auch das **Japanische** dem agglutinierenden Sprachbau. Im Fall des Japanischen kommt jedoch in Bezug auf die C-Test-Erstellung noch erschwerend hinzu, dass es sich bei den japanischen Schriftsystemen Hiragana und Katakana um Silbenschriften handelt. Daneben werden auch die Logogramme des chinesischen Kanji verwendet (vgl. ROOS 1994: 64). Bei dem Versuch, eine geeignete Variante des C-Tests für die japanische Sprache zu finden, erprobte ROOS (1994) insgesamt sieben verschiedene Tilgungsmuster. Sie kam zu dem Schluss, dass eine Version, bei der bei jedem zweiten Silbenzeichen die untere (!) Hälfte gelöscht wird, am besten geeignet sei (vgl. ROOS 1994: 83).

ARRAS und GROTJAHN (1994) überprüfen die Übertragbarkeit des C-Test-Prinzips auf die **chinesische Sprache**. Das Chinesische zählt zu den isolierenden Sprachen, und sein Zeichensystem Kanji bildet mit jedem Zeichen ein Wort ab (vgl. ARRAS & GROTJAHN 1994: 2 f.). Eine weitere Besonderheit stellt das Fehlen von Kategorien wie Kasus, Numerus, Genus oder Tempus dar (vgl. ibid.). Das C-Test-Prinzip besteht daraus, dass die Hälfte der Buchstaben eines Worts gelöscht wird. Im Chinesischen aber müssen für eine Übertragung des C-Test-Prinzips die Einheiten „Wort" und „Buchstabe" zunächst definiert werden (vgl. ARRAS & GROTJAHN 1994: 10). Eine Möglichkeit besteht darin, Striche der einzelnen Schriftzeichen zu entfernen. In einer Pilotstudie mit vier chinesischen Muttersprachlern sowie 24 Chinesischlernern der Universitäten Bochum und Heidelberg setzen ARRAS und GROTJAHN (1994) zwei derartige C-Tests ein, die eine Reliabilität von etwa $\alpha = 0{,}9$ erreichen sowie Korrelationen von bis zu $r = 0{,}77$ mit der Kursabschlussklausur aufweisen. Die Autoren kommen zu dem Schluss, dass das C-Test-Prinzip durch diese Modifikation für das Chinesische anwendbar gemacht werden kann, jedoch „[a]ls Methode zur

Messung globaler Sprachbeherrschung dürfte sich der C-Test im Chinesischen – zumindest in der vorliegenden Form – allerdings nur sehr bedingt eignen" (ARRAS & GROTJAHN 1994: 43). Vielmehr erfasse er die Schreibkompetenz der Testteilnehmer (vgl. ARRAS & GROTJAHN 1994: 18). Zudem ist die Testökonomie aufgrund des deutlich höheren Aufwands sowohl bei der Testerstellung als auch bei der Auswertung deutlich geringer als bei flektierenden Sprachen (vgl. ARRAS & GROTJAHN 1994: 28 & 43), was dem C-Test im Chinesischen einen seiner Vorteile gegenüber anderen Testformaten nimmt.

Auch für das **Hebräische** wurde versucht, das C-Test-Prinzip zu adaptieren. Wenngleich Hebräisch zu den flektierenden Sprachen gezählt wird, können hier Artikel oder Präpositionen in Form von Affixen an Wörter angehängt werden (vgl. COHEN et al. 1984: 221). Eine weitere Besonderheit stellt das Schriftsystem der hebräischen Sprache dar, welches von rechts nach links gelesen wird. Es handelt sich hierbei um eine Konsonantenschrift, d. h. nur die Konsonanten werden ausgeschrieben, nicht jedoch die Vokale (vgl. MCTAGGART & KLAR o. J.: 5). Dies führt dazu, dass ein Leser auch ohne Beschädigung des Textes die Informationen mental vervollständigen muss. SEGAL (1983[15]) übersetzte von RAATZ und KLEIN-BRALEY (1982) eingesetzte C-Tests ins Hebräische und stellte fest, dass einige Texte für die 42 Probanden (angehende Lehrer) zu schwer waren. Daher fügt er bei der Hälfte der Lücken die normalerweise nicht ausgeschriebenen Vokale ein und erreicht dadurch zufriedenstellende Ergebnisse: Eine Reliabilität von KR-20 > 0,8 und Korrelationen mit einem Grammatiktest von 0,87 sowie mit einem Leseverstehenstest von 0,69 (vgl. COHEN et al. 1984: 223). (Die verwendeten Korrelationskoeffizienten werden nicht genannt.)

15 Diese Quelle ist in hebräischer Sprache verfasst und somit für die Autorin nicht rezipierbar. Daher wird hier auf COHEN et al. (1984) verwiesen, die eine englischsprachige Zusammenfassung von SEGAL (1983) liefern.

Wie die vorangehenden Ausführungen gezeigt haben, sind für nahezu alle Sprachen Modifikationen des kanonischen C-Test-Prinzips oder doch zumindest ergänzende Regelungen (beispielsweise für den Umgang mit Enklitika) unabdingbar. Je stärker sich eine Sprache strukturell von Englisch und Deutsch, wofür der C-Test entwickelt wurde, unterscheidet, desto umfangreicher fallen diese Modifikationen aus. Die Frage, ob sich durch die veränderten Tilgungsregeln auch ein sprachspezifisches Testkonstrukt ergibt, muss für jede Sprache individuell untersucht und diskutiert werden.

2.3.2 Fachsprachliche C-Tests

Neben den bereits diskutierten C-Test-Versionen in diversen Sprachen, existieren weitere Varianten des C-Tests, die sich hinsichtlich ihres textlichen Inhalts voneinander unterscheiden. Hierzu zählen die Entwicklung und der Einsatz fachsprachlicher C-Tests. Dies ist jedoch ein noch weitgehend unerforschtes Gebiet, zu dem nur vereinzelt wissenschaftliche Arbeiten vorliegen.

ANCKAERT und BEECKMANS (1992) erstellten für die Zielsprache Niederländisch einen C-Test, dessen Texte wirtschaftswissenschaftliche Themen zum Inhalt hatten. Jedoch war die fachsprachliche Ausrichtung der Texte nicht Ziel dieser Untersuchung, ging es den Autoren doch vielmehr darum, zielgruppenadäquate Texte für die wirtschaftswissenschaftliche Probandengruppe zu konzipieren. BOLTEN (1992) nutzt erfolgreich einen C-Test zur Einstufung in einen Wirtschaftsdeutschkurs. Wie allgemein- oder fachsprachlich orientiert die Texte des Testsets hierbei waren, gibt der Autor jedoch nicht an. Zwischen dem C-Test und der Düsseldorfer Zertifikatsprüfung Wirtschaftsdeutsch fand er Korrelationen von $r = 0{,}69$ (Hörverstehen) und $r = 0{,}92$ (Textproduktion). Ungeachtet der Inhalte des C-Tests erscheint der Wert von 0,92 beachtlich.

In der Untersuchung von DALLER (1996) sollten allgemeinsprachliche mit fachsprachlichen C-Tests verglichen werden. Hierzu wurden zwei C-Tests erstellt: Dem einen dienten Texte aus einer türkischen Zeitschrift für Deutschlerner (*Bizim Almanca*-Unser Deutsch) als Grundlage, welche Alltagssprache vertreten sollte. Der andere C-Test wurde aus Texten germanistischer Fachliteratur erstellt. Daller geht es hierbei offensichtlich um eine Trennung von *Cognitive Academic Language Proficiency* (CALP) und *Basic Interpersonal Communication Skills* (vgl. CUMMINS 1979):

> Es ist davon auszugehen, daß bei der Lösung von C-Tests, die auf diesen beiden Textsorten basieren, andere Sprachfähigkeiten eine Rolle spielen. Werden akademische Texte als Basis genommen, sind akademische Sprachfähigkeiten nötig, um den […] Test zu lösen. Umgekehrt sind alltagssprachliche Fähigkeiten nötig, um C-Tests zu lösen, die auf der Basis von alltagssprachlichen Texten erstellt wurden. (DALLER 1996: 348)

Wenngleich diese Feststellung zunächst lapidar klingt, stellt diese Studie von Daller doch einen ersten Schritt in Richtung fachsprachliche C-Tests dar. Für jedes der beiden Testsets ermittelt DALLER (1996: 349) eine Reliabilität von $\alpha = 0{,}86$.

Auch BAUR et al. (2010) befassen sich mit fachsprachlichen C-Tests. Sie erstellten einen vierteiligen C-Test, der zwei allgemeinsprachliche aus Tageszeitungen sowie zwei fachsprachliche Texte aus einem Lehrbuch für die Berufsschule enthielt. Dieses Testset wurde 67 Schülern eines Berufskollegs vorgelegt. Die zugrundeliegende Hypothese der Autoren war, dass für ein erfolgreiches Erschließen komplexer fachsprachlicher Texte zunächst die Strukturen allgemeinsprachlicher Texte beherrscht werden müssen. Diese Hypothese sahen Baur et al. bestätigt, da die Ergebnisse der fachsprachlichen Texte (Gruppe 1: 49,8 (n = 49); Gruppe 2: 62,8 (n = 18)) im Mittel nicht besser ausfielen als die der allgemeinsprachlichen Texte (Gruppe 1: 64,4 (n = 49); Gruppe 2: 67,3 (n = 18)) (vgl. BAUR et al. 2010: 34). Dieses Ergebnis bleibt auch bestehen, wenn der Worterkennungswert ermittelt

wird. Zwischen den beiden Textsorten ermittelten die Autoren eine sehr hohe Korrelation von r = 0,98. Dieser hohe Wert kann als ein Hinweis darauf verstanden werden, dass zumindest die hier verwendeten fach- und allgemeinsprachlichen Texte ein sehr ähnliches Konstrukt haben.

Ein gezielter Versuch, die Eignung des C-Tests zur Messung fachsprachlicher Kompetenz heranzuziehen, stellt die Untersuchung von SCHLAK et al. (2010) dar. Um herauszufinden, ob C-Tests mit fachsprachlichen Inhalten dazu geeignet sind, auch die Fachsprachenkompetenz einer Person zu ermitteln, wurde in Anlehnung an die Themen der Prüfung Wirtschaftsdeutsch (PWD) ein C-Test entwickelt und zunächst an muttersprachlichen Studenten des Fachbereichs Wirtschaft erprobt, ehe er Lernern in Wirtschaftsdeutschkursen zweimal, zu Kursbeginn und am Kursende, vorgelegt wurde.[16]

Auch SCHÖLER (2016) plädiert für den Einsatz des C-Tests zur Erhebung fachsprachlicher Kompetenzen im Unterricht und erstellt einen exemplarischen C-Test aus einem Sachtext, bleibt jedoch die empirische Überprüfung schuldig.

HÖTTECKE et al. (2017) erstellten im schulischen Kontext neben allgemeinsprachlichen C-Tests auch solche für die Fächer Sport und Physik. Jedoch muss angemerkt werden, dass aufgrund der linksseitigen Tilgung in den diskursspezifischen Testsets hier nicht von einem klassischen C-Test gesprochen werden kann, weshalb die Studie an dieser Stelle keine weitere Beachtung erfährt (vgl. Kapitel 2.3.3).

16 Aufgrund des Todes des Projektleiters Prof. Torsten Schlak konnte das Projekt leider nicht zu Ende geführt werden.

2.3.3 Weitere Varianten des C-Tests

Es existiert eine ganze Reihe weiterer Testvarianten, die unter dem Namen C-Test auftauchen, und mehr oder weniger eng mit dem originalen Testformat verwandt sind und deren Einfluss und Bedeutung sehr unterschiedlich zu bewerten ist. Sie sollen im Folgenden zumindest kurz Erwähnung finden.

BAUR & SPETTMANN (2008b) berichten von sogenannten **Teilfertigkeitstests**. Hierbei wird das Tilgungsprinzip des C-Tests derartig variiert, dass in jedem Text ein fokussiertes Element abgefragt wird. Bei einem auf Fachlexik abzielenden Teilfertigkeitstest wird nunmehr die erste Hälfte jedes zweiten bis dritten Wortes gelöscht, wie Abbildung 4 zeigt.

Das Gebiss des Wolfes

Das Gebiss des Wolfes ist besonders gut zum Fleischfressen geeignet. Die dolchartigen ___kzähne dienen zum ___sthalten und Töten der ___te. Die meisten ___kenzähne sind ___itz und haben ___rfe Kanten. Zum ___leinern großer ___schstücke oder zum ___cken von ___chen sind die ___ißzähne im ___erkiefer und ___terkiefer bestens ___bildet. Mit den ___nteren Backenzähnen kann der ___lf ___nzenkost zerquetschen mit den ___eidezähnen gelingt es den Tieren auch das letzte Stückchen Fleisch von einem Knochen zu ___aben. Tiere, bei denen die Zähne so beschaffen sind, bezeichnet man auch als ___ubtiere. Das Fleischfressergebiss wird auch Raubtiergebiss genannt.

Abb. 4: Fachwortschatztest (BAUR & SPETTMANN 2008b: 439)

Das veränderte Testformat dient dazu, Fachwortschatz zu üben. Bemerkenswert sind in diesem Zusammenhang zwei Aspekte: Zum einen, dass durch das veränderte Tilgungsschema Flexionsendungen gar keine Rolle mehr spielen und der Fokus eindeutig auf Semantik bzw. Wortschatz liegt.

Zum anderen wird der C-Test hier nicht als Test, sondern als ein Übungs- und Förderungsinstrument verwendet. Es bleibt jedoch zu diskutieren, ob ein so gestalteter Text den Titel C-Test überhaupt noch tragen sollte, ist doch das Prinzip der reduzierten Redundanz in allen bisher angeführten Varianten des C-Tests beibehalten worden, während es hier jedoch nicht mehr zu erkennen ist. SIGOTT & KÖBERL (1993) nennen diese Testvariante X-Test.

Der Versuch eines **Multiple-Choice-C-Tests** findet sich bei JAKSCHIK und KLEMMERT (2006). Hier wurde ein aus sechs Texten bestehendes C-Test-Set entwickelt, bei dem jeder Text zwischen 18 und 22 Lücken hatte. Diese Lücken waren anders als beim herkömmlichen C-Test nicht frei ausfüllbar, sondern es standen den Probanden neben Bestantwort vier Distraktoren, also insgesamt fünf Antwortoptionen zur Verfügung. Ihren Erwartungen entsprechend fanden Jakschik und Klemmert heraus, dass der Multiple-Choice-C-Test leichter war als das gleiche Textset in Form eines gewöhnlichen C-Tests (vgl. JAKSCHIK & KLEMMERT 2006: 198).

Zwischen beiden Testformaten fanden die Autoren eine Korrelation von r = 0,9. Wie das folgende Textbeispiel (Abb. 5) deutlich macht, ist ein Multiple-Choice-C-Test in der Entwicklung sehr viel aufwendiger, da es durchaus eine Herausforderung darstellt, akzeptable Distraktoren zu finden.

	männer		ikel		che	
	benzin		ute		iert	
Der Super	leicht	hat he	imisch	ganz fris	ches	Obst.
	nova		izen		tlos	
	markt		ktisch		chen	

Abb. 5: Beispiel für einen MC-C-Test (JAKSCHIK & KLEMMERT 2006: 197)

Einige der hier angebotenen Distraktoren wirken nicht sehr plausibel. Die Testökonomie betreffend ist der Multiple-Choice-C-Test dem normalen C-Test aufgrund des Aufwands bei der Erstellung der Distraktoren unterlegen. Dies kann ein Grund dafür sein, dass der Multiple-Choice-C-Test bislang keine Verbreitung gefunden hat.

Eigens zu Forschungszwecken erstellte GROTJAHN (1996) sogenannte ***Scrambled* C-Tests**, bei denen die Sätze innerhalb eines C-Tests unterschiedlich stark permutiert wurden. Dieses Vorgehen diente zur Überprüfung der Annahme STEMMERs (1991; 1992), dass zur Lösung von C-Test-Texten hauptsächlich der Mikrokontext eine Rolle spiele. Sollte sich durch die Permutation der Sätze in einem C-Test-Text die Schwierigkeit erhöhen, so könne daraus der Schluss gezogen werden, dass bei der Bearbeitung der Makrokontext sehr wohl eine Rolle spielt (vgl GROTJAHN 1996: 98). Grotjahn kommt zu dem Ergebnis, dass C-Tests zwar hauptsächlich auf der Mikroebene testen, dass aber bei der Bearbeitung längerer Texten auch makrostrukturelle Prozesse relevant sein können (vgl. ibid. 114). In der Praxis haben *scrambled* C-Tests keine Relevanz.

Der C-Test wird nicht nur bei erwachsenen Sprechern, sondern ebenfalls bei Kindern eingesetzt. Im schulischen Bereich entwickelten KNIFFKA & LINNEMANN (2014) einen C-Test mit dem Ziel einer zuverlässigeren Einschätzung der Deutschkompetenz von Kindern mit Migrationshintergrund, nachdem sie festgestellt hatten, dass die Lehrerurteile trotz Kenntnis des Gemeinsamen Europäischen Referenzrahmens (vgl. EUROPARAT 2001) wenig aussagekräftig waren (vgl. KNIFFKA & LINNEMANN 2014: 240). Das finale Testset besteht aus fünf Texten à 20 Lücken und erreicht eine hohe Reliabilität von $\alpha = 0{,}96$ ($n = 120$). Im Gegensatz zu anderen Autoren nutzen Kniffka und Linnemann das klassische Tilgungsprinzip und nehmen keine Modifikationen hieran vor.

Auch BAUR und SPETTMANN (2008b) verwendeten einen **C-Test im Schulkontext**, d. h. in der Unterstufe. Sie modifizieren das kanonische C-Test-Prinzip dahingehend, dass nicht jedes zweite Wort, sondern nur jedes dritte Wort getilgt wird (vgl. BAUR & SPETTMANN 2008b: 432). Den Testteilnehmern steht somit zum Lösen der Lücken mehr Kontext zur Verfügung.

2.3.4 Varianten der Darbietungsbedingungen

Weitere Varianten des C-Tests ergeben sich aus Modifikationen an den Durchführungs- bzw. Darbietungsbedingungen. Die Rahmenbedingungen und Durchführungsbestimmungen für den Einsatz von C-Tests waren lange Zeit nicht näher bestimmt und sind auch heute noch uneinheitlich. So war es zunächst üblich, die Texte eines C-Test-Sets den Testteilnehmern in Form der sogenannten **Gesamtdarbietung** zu präsentieren (vgl. GROTJAHN 2010: 271). Das bedeutet, dass die Texte theoretisch in beliebiger Reihenfolge bearbeitet werden können und ein Vor- und Zurückblättern möglich ist. Die Verantwortung, die zur Verfügung stehende Bearbeitungszeit auf die einzelnen Texte in sinnvoller Weise zu verteilen, liegt somit in den Händen der Prüflinge. Dies kann zur Folge haben, dass ein Teilnehmer aufgrund ungünstigen Zeitmanagements nicht die Gelegenheit hat, alle Texte zu bearbeiten. Dies mindert nicht nur die Reliabilität der Testergebnisse, sondern wirkt sich ebenfalls negativ auf die Berechnung von Aufgabenschwierigkeiten aus (vgl. GROTJAHN 2010: 271), da rechnerisch nicht erkennbar ist, ob ein Text wegen mangelnder Kompetenz des Testteilnehmers oder aufgrund von Zeitmangel nur oberflächlich oder gar nicht bearbeitet werden konnte.

Aus diesem Grund hat sich – insbesondere im Rahmen computeradministrierter C-Tests – die sogenannte **Einzeldarbietung** etabliert. Hierbei wird die den Teilnehmern zur Verfügung stehende Bearbeitungszeit nicht

für das gesamte Test-Set anberaumt, sondern für jeden einzelnen Text. Für gewöhnlich ist die Bearbeitungszeit pro Text konstant über das Test-Set verteilt. GROTJAHN (2010: 276) spricht von einer Bearbeitungszeit von üblicherweise vier bis sechs Minuten pro Text. In der praktischen Durchführung von Datenerhebungen zeigt sich, dass bereits bei einer Bearbeitungszeit von vier oder fünf Minuten pro Text kaum ein Proband die volle ihm zur Verfügung stehende Zeit ausnutzt. Im Rahmen des onDaF besteht für die Testteilnehmer daher auch die Möglichkeit, die maximale Bearbeitungszeit von fünf Minuten pro Text à 20 Lücken eigenmächtig vorzeitig zu beenden (vgl. URL 26: onSET: Teilnehmende). Wichtig ist indes, dass die Bearbeitungszeit so großzügig bemessen ist, dass alle Kandidaten ausreichend Zeit haben, um alle Texte vollständig zu bearbeiten. Die Einzeldarbietung ist somit nicht nur der Durchführungsobjektivität, sondern auch der Reliabilität förderlich (vgl. GROTJAHN 2010: 276; vgl. Kapitel 2.2.4).

Eine weitere Unterscheidung bei der Darbietung von C-Tests ist die zwischen einem herkömmlichen **Papier-und-Bleistift-Test und einer computeradministrierten** Variante. Da in der vorliegenden Studie beide Formate parallel eingesetzt werden, wird die Vergleichbarkeit beider Testformate im Folgenden ausführlich betrachtet.

Schon GERMANN und GROTJAHN (1994) untersuchen die Prozesse bei der Bearbeitung von C-Tests am Computer über eine sehr simple und frühe Variante des *Key Logging*, d. h. über Tastenprotokollierung. Sie ließen acht Oberstufenschüler einen englischsprachigen C-Test am Computer mit angeschlossener Tastatur lösen und analysierten die durch die Enter-Taste abgeschickten Lösungen einzelner Lücken sowie den Gebrauch der Befehle PASSON, um eine Lücke zu überspringen, oder PASSBACK, um zu einer vorangehenden Lücke zurückzukehren (vgl. GERMANN & GROTJAHN 1994: 280). Weitere sieben Oberstufenschüler legten als Kontrollgruppe den gleichen Test als Papier-und-Bleistift-Version ab. Die

arithmetischen Mittel der Teilnehmer der Experimental- ($\bar{x}$ = 93,25) und der Kontrollgruppe ($\bar{x}$ = 90,14) lagen sehr nah beieinander. Die Standardabweichung der Kontrollgruppe liegt mit SD = 16,66 etwas höher als die der Experimentalgruppe (SD = 11,22) (vgl. GERMANN & GROTJAHN 1994: 287). Dies kann jedoch auch auf die sehr geringe Stichprobengröße zurückgeführt werden. Wie die Autoren selbst betonen, handelt es sich lediglich um eine Pilotuntersuchung, und die Frage nach der Vergleichbarkeit von Computer- und Papierversionen des C-Tests bleibt hier zunächst offen.

BISPING und RAATZ (2002) widmen sich in ihrem Artikel „Sind computerisierte und Papier&Bleistift-Versionen des C-Tests äquivalent?" jedoch genau dieser Frage. Ein zuvor pilotierter englischsprachiger C-Test wurde 41 Oberstufenschülern als Papier-und-Bleistift-Version präsentiert, während weitere 47 Probanden das gleiche Test-Set am Computer zu bearbeiten hatten (vgl. BISPING & RAATZ 2002: 137). Die arithmetischen Mittel sind mit einem Wert von $\bar{x}$ = 75 für die Papier-und-Bleistift-Gruppe und $\bar{x}$ = 75,2 für die Computergruppe nahezu identisch (vgl. ibid: 136). Gleiches gilt für die Reliabilität, für die Werte für Cronbachs Alpha von α = 0,84 für die Papier- und α = 0,85 für die Computergruppe errechnet wurden.

Über einen Fragebogen erhoben die Autoren auch Daten über das Erleben der beiden Testversionen durch die Probanden. Es zeigte sich, dass in beiden Gruppen negative Emotionen wie Nervosität nur in geringem Maße vorhanden waren. Im Gegensatz dazu wurde die Computervariante des C-Tests von den Probanden als deutlich interessanter, moderner und professioneller empfunden (vgl. ibid: 139–140).

Besondere Beachtung erfährt bei BISPING und RAATZ (2002) der mögliche Einfluss des Faktors Computerangst und -erfahrung. Dieser wird operationalisiert als Computerängstlichkeit, Computerbesitz, Tipphäufigkeit und am Computer verbrachte Zeit (vgl. ibid: 141). Es wurde jedoch in kei-

nem Fall, d. h. weder bei der Computer- noch bei der Papiergruppe eine signifikante Korrelation zwischen einem dieser Aspekte und dem Testergebnis gefunden (vgl. ibid: 143). Die Autoren schlussfolgern darum, dass ein Einfluss auf die Validität durch den Faktor Computerangst und -erfahrung nicht gegeben ist.

BISPING (2006) widmet sich der Frage nach der Validität von Computer-C-Tests. Hierzu vergleicht er zwei C-Test-Sets mit unterschiedlichen Texten. Von diesen lag ein Test-Set als Papier-und-Bleistift-Version vor und hatte eine Bearbeitungszeit von fünf Minuten pro Text. Das andere Test-Set wurde am Computer abgelegt und die Bearbeitungszeit pro Text betrug lediglich drei Minuten. Die Reliabilitäten der beiden Test-Sets waren mit Werten von $\alpha = 0{,}9$ (Computer-C-Test) und $\alpha = 0{,}92$ (Papier-C-Test) sehr gut. Von den maximal 125 erreichbaren Punkten erreichten die Probanden im ersten Test-Set (Papierversion) durchschnittlich $\bar{x} = 98{,}6$ Punkte, im zweiten bearbeitungszeitreduzierten *speeded* Test-Set (Computer) waren es nur $\bar{x} = 83{,}4$. Der Autor interpretiert diese Ergebnisse dahingehend, dass die Messmethode zu diesen unterschiedlichen Werten führt (vgl. ibid: 157). Da jedoch die Zeitbemessung für den computerbasierten C-Test knapper war, könnte dies ebenso ein Grund für das kleinere arithmetische Mittel sein. Eine weitere Erklärungsmöglichkeit stellen unterschiedliche Schwierigkeitsgrade der einzelnen Texte dar (vgl. ibid: 161).

Die vorangegangenen Ausführungen machen deutlich, dass es nicht *den* C-Test gibt (vgl. GROTJAHN 2011: 132), sondern dass neben dem ursprünglichen C-Test-Format inzwischen zahlreiche Varianten existieren, die das klassische Tilgungsprinzip je nach Erhebungskontext oder Zielsprache angepasst haben. Darüber hinaus führen auch die genannten Unterschiede in der Darbietung von C-Tests streng genommen zu weiteren C-Test-Varianten. Im Rahmen der vorliegenden Arbeit soll nun überprüft werden, ob

sich der *speeded* C-Test hier als eine gangbare Variante in *low-stakes*-Prüfungssituationen einreihen kann.

2.4 Einsatzmöglichkeiten des C-Tests

Der C-Test stellt ein sehr vielseitiges Messinstrument dar und konnte sich daher in zahlreichen Einsatzbereichen etablieren. Aufgrund seiner Zeitökonomie und Auswertungsobjektivität wird er sehr häufig als Instrument zur **Einstufung** von Sprachkursinteressenten verwendet, insbesondere im Hochschulbereich (vgl. beispielsweise URL 11: ZEMS. TU Berlin; URL 12: ZFA. Ruhr-Universität Bochum; URL 13: Sprachenzentrum der LMU München; URL 14: Zentraleinrichtung Sprachenzentrum der Humboldt-Universität zu Berlin; URL 15: Sprachenzentrum der Westfälischen Wilhelms-Universität Münster; URL 16: Sprachenzentrum der Philipps-Universität Marburg; URL 17: Sprachenzentrum der Universität Passau). Wie aus den Verweisen hervorgeht, werden hierbei von den universitären Sprachenzentren zahlreiche der unter Kapitel 2.3 abgehandelten Varianten abgedeckt. Bei der Einstufung geht es darum, ein globales Bild der Sprachkompetenz der Testteilnehmer zu erhalten, um sie der richtigen Kursstufe zuordnen zu können, weshalb der C-Test in diesem Kontext besonders geeignet ist.

Jedoch eignet sich der C-Test nicht nur zur Einstufung im Hochschulkontext: JAKSCHIK (1994) nutzt den C-Test als Einstufungsinstrument für erwachsene Zweitsprachler, die auf dem zweiten Bildungsweg den Hauptschulabschluss nachholen möchten. Hier dient der C-Test dazu, anhand der

Deutschkompetenz der künftigen Schüler Leistungsgruppen definieren und bilden zu können.[17]

Der onDaF (vgl. URL 3: onDaF) wurde als **Screening**-Test für den TestDaF (vgl. URL 4: TestDaF) entwickelt. Die Leistung im onDaF soll potentiellen TestDaF-Teilnehmern also einen ersten Hinweis darauf geben, ob ihre Sprachkenntnisse bereits so weit fortgeschritten sind, dass sie den (deutlich teureren) TestDaF erfolgreich ablegen können. In diesem Kontext wird also davon profitiert, dass der C-Test mittlere bis hohe Korrelationen mit den den TestDaF umfassenden Fertigkeiten Hörverstehen, Leseverstehen, Sprechen und Schreiben aufweist. Nur so ist eine Einschätzung eines erfolgreichen Ablegens des TestDaF überhaupt möglich. Auch der von SCHLAK et al. (2010) erstellte fachsprachlich ausgerichtete „C-Test Wirtschaftsdeutsch" war als Screening-Test für die Prüfung Wirtschaftsdeutsch International (PWD) (vgl. URL 5: DIHK – Gesellschaft für berufliche Weiterbildung: Prüfung Wirtschaftsdeutsch International) konzipiert.

Ebenfalls Verwendung findet der C-Test als **Diagnoseinstrument**. An der Ruhr-Universität Bochum wurde ein C-Test für die Zielsprache Französisch entwickelt (vgl. GROTJAHN 1986), der eine bis dahin eingesetzte mehrteilige diagnostische Testbatterie ergänzen und Auskunft über die Fremdsprachenkompetenz von Studienanfängern des Fachs Französisch liefern sollte. Seine Eignung als diagnostischer Test verdankt der C-Test den unterschiedlichen Möglichkeiten seiner Auswertung. So wird im Bereich Deutsch als Zweitsprache beispielsweise zwischen einem sogenannten Richtig-Falsch-Wert und einem Worterkennungswert unterschieden. Ers-

17 Bemerkenswerterweise konnte sich der C-Test an Volkshochschulen seiner Ökonomie zum Trotz bislang nicht als Einstufungsinstrument etablieren. Hier wird auf andere Verfahren wie beispielsweise Selbsteinschätzung oder Einstufungstests von Lehrwerksverlagen zurückgegriffen (vgl. z. B. URL 20: Hamburger Volkshochschule; URL 21: VHS Einstufungstest).

terer bezieht sich auf die Anzahl der korrekt ausgefüllten Lücken, während letzterer auch diejenigen Lücken mitzählt, die zwar nicht korrekt ausgefüllt wurden, aus deren Lösungsversuch jedoch erkennbar ist, dass der Testteilnehmer das gesuchte Wort semantisch erkannt hat (vgl. BAUR & SPETTMANN 2008a: 97 ff.). Aus der Differenz zwischen Richtig-Falsch-Wert und Worterkennungswert werden Rückschlüsse auf das Verhältnis von rezeptiven und produktiven Fertigkeiten eines Lerners gezogen. Auf diese Weise kann für jeden Schüler ein individueller Förderbedarf ermittelt werden. Eine größere Differenz zwischen Worterkennungs- und Richtig-falsch-Wert lässt beispielsweise auf ein hinreichendes Textverständnis mit Schwierigkeiten im Bereich Rechtschreibung und/oder Morphosyntax schließen (vgl. BAUR et al. 2013: 9). Beachtenswert ist in diesem Kontext, dass der C-Test nicht nur zur Sprachstandsdiagnostik bei Kindern mit Deutsch als Zweitsprache eingesetzt wird, sondern dass ebenfalls bei muttersprachlichen Schülern eine Sprachstandsdiagnose mittels C-Tests durchgeführt wird, um einen etwaigen Förderbedarf zu ermitteln.

So setzt das Bundesland Hamburg C-Tests flächendeckend ab der 2. Klasse als Lesetest bei Grundschulkindern ein (vgl. MAY & BENNÖHR 2013). Im Rahmen der sogenannten Testbatterie KEKS (**K**ompetenz**e**rfassung in **K**indergarten und **S**chule) (vgl. URL 18: Cornelsen: KEKS), die vom Landesinstitut für Lehrerbildung und Schulentwicklung Hamburg entwickelt wurde, wird eine für die Zielgruppe modifizierte C-Test-Version eingesetzt. Hierbei wird mit Rücksichtnahme auf die sehr jungen Testteilnehmer das kanonische Tilgungsprinzip gelockert, so dass zur Regulierung der Testschwierigkeit die Tilgung nicht immer exakt die Hälfte eines Wortes trifft, sondern je nach Schwierigkeit etwas mehr oder weniger (vgl. MAY et al. 2014: 264). Auch wird nicht jedes zweite Wort getilgt, wie das Beispiel in Abbildung 6 aus einem KEKS-Musterheft zeigt.

Da der C-Test im Rahmen von KEKS als Lesetest genutzt wird, gibt es auch bei der Auswertung eine Abweichung von der klassischen Vorgehensweise: Um Aufschluss darüber zu erhalten, wie weit ein Testteilnehmer den Text verstanden hat, werden die ausgefüllten Lücken nicht nach den Kategorien ‚richtig' oder ‚falsch' bewertet, sondern orthographische wie morphologische Fehler werden hingenommen, sofern aus der Lösung Textverständnis zu erkennen ist (vgl. MAY et al. 2014: 265).

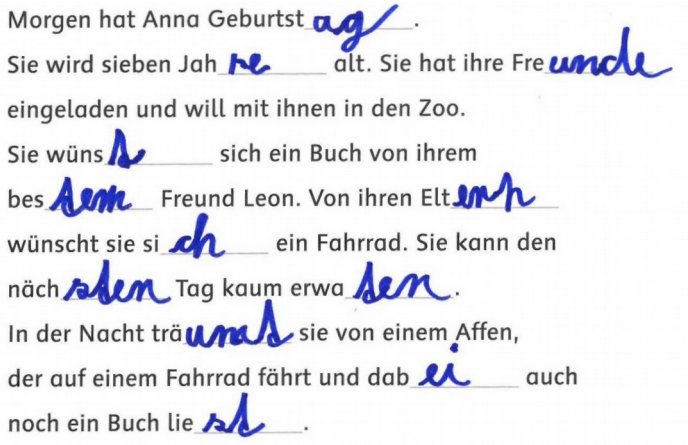

Morgen hat Anna Geburtst ag .
Sie wird sieben Jah re alt. Sie hat ihre Fre unde
eingeladen und will mit ihnen in den Zoo.
Sie wüns ht sich ein Buch von ihrem
bes ten Freund Leon. Von ihren Elt ern
wünscht sie si ch ein Fahrrad. Sie kann den
näch sten Tag kaum erwa rten.
In der Nacht trä umt sie von einem Affen,
der auf einem Fahrrad fährt und dab ei auch
noch ein Buch lie st .

Abb. 6: Beispiel für einen C-Test bei KEKS (KEKS Beispielheft, URL 19)

Die Autoren geben an, mit auf diese Weise erzeugten C-Tests die Lesekompetenz der Schüler zuverlässiger als mit anderen Leseverstehenstests messen zu können (vgl. MAY et al. 2014: 265). An dieser Stelle wird bereits deutlich, dass die unterschiedlichen Versionen des C-Test-Formats in einem ursächlichen Zusammenhang mit dem Kontext, in dem sie eingesetzt werden, stehen.

Eine weitere Möglichkeit der Sprachstandsdiagnostik mittels C-Test besteht darin, eine Klassifikation der gemachten Fehler vorzunehmen, wie

beispielsweise bei einer Typisierung der nicht korrekt ergänzten Lücken nach Rechtschreibfehler, Kasusfehler, Numerusfehler, Genusfehler oder ähnlichen Kategorien. Dieses Vorgehen wurde in AGUADO et al. (2007) und GROTJAHN et al. (2010) verwendet. Es erlaubt, ein Profil der für den betreffenden Testteilnehmer noch schwierigen Sprachbereiche zu erstellen oder beispielsweise herauszufiltern, ob es einen speziellen sprachlichen Bereich gibt, der das Gesamtergebnis des C-Tests negativ beeinflusst. Diese Art der C-Test-Auswertung ist im Vergleich zu herkömmlichen Methoden weniger ökonomisch, zumal im Einzelfall die Entscheidung zwischen zwei Fehlerkategorien uneindeutig sein kann. Eine maschinelle Auswertung ist daher kaum möglich, was folglich die Auswertungsobjektivität herabsetzt. Dennoch ist in diesem Verfahren ein großes diagnostisches Potential zu erkennen.

Nicht nur im Bereich der Sprachstandsdiagnostik findet der C-Test Verwendung, sondern er kann auch gezielt zur **Sprachförderung** eingesetzt werden, denn „[d]as Lösen von C-Tests sensibilisiert die SuS [Schülerinnen und Schüler; Anmerkung K. Z.] einerseits für den korrekten Sprachgebrauch und fördert die Aktivierung von Lösungs- und Lesestrategien“ (BAUR et al. 2013: 13). Je nach Komplexität des Textes erfordert das Lösen eines C-Tests Erschließungsstrategien auf der Wort-, Satz- und Textebene (vgl. ibid.). BAUR & SPETTMANN (2008b: 439) argumentieren, dass der C-Test zur Leseförderung im Bereich Deutsch als Zweitsprache besonders gut geeignet sei, weil er im Gegensatz zu anderen Formaten das Gedächtnis nicht belaste. Neben der Anwendung von Hintergrundwissen, selektivem Leseverstehen und dem Erlernen einer Selbstkontrolle nennen die Autoren lexikalische, morphologische, syntaktische und textbezogene Strategien, die durch das Lösen von C-Tests geschult werden können (vgl. ibid.). Die Verwendung von C-Tests zu Zwecken der Sprachförderung im

schulischen Bereich scheint für erst- wie zweitsprachliche Kinder sein Potential noch nicht ausgeschöpft zu haben.

Der C-Test hat sich darüber hinaus in zahlreichen Studien als **Forschungsinstrument** etabliert. Das bedeutet, neben Studien, die den C-Test selbst als zentralen Untersuchungsgegenstand haben, wird er auch genutzt, um beispielsweise den Lernfortschritt zu dokumentieren (vgl. SCHLAK et al. 2010: 14) oder den Sprachstand von Probanden zu ermitteln (vgl. z. B. LAUBER 2013). Aufgrund der allgemein akzeptierten Annahme, der C-Test messe allgemeine Sprachkompetenz, wird er auch als Operationalisierung globaler Sprachkompetenz in Studien eingesetzt, um den Zusammenhang mit anderen Variablen wie beispielsweise Intelligenz (vgl. z. B. RAATZ 2002) zu untersuchen.

Eine neuere Arbeit unter Einbezug eines C-Tests als Forschungsinstrument ist die Studie von DRACKERT (2015). Sie nutzt einen russischsprachigen C-Test neben anderen Testformaten zur Validierung eines anderen Testformats, dem sogenannten *Elicited Imitation Test* in der russischen Sprache.

Auch in qualitativen Forschungsarbeiten findet der C-Test Verwendung: LAUBER (2013) untersucht das kommunikative Verhalten bilingualer Paare mit der Sprachkombination Deutsch und Spanisch. Sie ermittelt mithilfe eines C-Tests den Sprachstand der drei teilnehmenden Paare, um die Ergebnisse der Interviews mit der Sprachkompetenz in der jeweiligen L2 der Probanden in Beziehung setzen zu können.

Wenngleich der C-Test im Gegensatz zum Cloze-Test von Anfang an für den Einsatz bei Fremdsprachenlernern intendiert war, so zeigt sich, dass er dennoch ebenso in erstsprachlichen Kontexten Verwendung findet. Wie bereits diskutiert wird der C-Test im schulischen Bereich nicht nur bei Zweitsprachlern, sondern bei allen Schülern zum Zweck der Sprachstandsdiagnostik eingesetzt. WOCKENFUSS & RAATZ (2006) untersuchen den

Zusammenhang von C-Test-Leistung und Klassenstufe. Ihrer groß angelegten Studie ($n = 1051$) liegt die Annahme zugrunde, dass C-Tests zwischen der 2. Klasse der Grundschule als frühestem Einsatzzeitpunkt und dem Erwachsenenalter immer einfacher zu lösen sind, bis die üblichen Lösungsraten von Muttersprachlern beobachtbar sind (vgl. WOCKENFUSS & RAATZ 2006: 211). Diese Hypothese fanden die Autoren durch ansteigende Mittelwerte der erreichten C-Test-Punkte für die Klassenstufen fünf bis zehn bestätigt. Dieses Ergebnis konnte für drei Schulformen gleichermaßen beobachtet werden: Gymnasium sowie Real- und Hauptschule, wobei die Mittelwerte der Gymnasiasten relativ zu den anderen Schulformen am höchsten, die der Hauptschüler am niedrigsten ausfallen (vgl. WOCKENFUSS & RAATZ 2006: 223). Zwischen der C-Test-Leistung und der besuchten Schulform der Probanden besteht folglich ein deutlicher Zusammenhang.

Die Dissertation von MASHKOVSKAYA (2014) trägt den Titel „Der C-Test als Lesetest bei Muttersprachlern". Mashkovskaya stellt die Frage, ob C-Tests genutzt werden können, um bei gebildeten Erwachsenen zwischen stärkeren und schwächeren Lesern zu differenzieren. Hierzu ordnet sie die Lücken eines C-Tests jeweils der Wort-, Satz- oder Textebene zu. Die Annahme ist hier, dass stärkere Leser C-Tests auf der höheren Satz- und Textebene lösen, während es schwächeren Lesern meist nur gelingt, die Lücken auf der Wortebene zu füllen. Die Daten von 887 Lehramtsstudenten bestätigen Mashkovskayas Hypothese. Jedoch hat MASHKOVSKAYA (2014) ebenso wie BAUR & SPETTMANN (2008b) in ihrem Teilfertigkeitstest die erste Hälfte der Wörter getilgt.

Einige weitere Studien, die den (S-)C-Test als Forschungsinstrument bei Muttersprachlern einsetzen, finden sich in Kapitel 3.2.

Zusammenfassend lässt sich festhalten, dass der C-Test sowohl bei Erwachsenen als auch bei Kindern im erst- sowie zweit- und fremdsprachli-

chen Bereich Verwendung findet. Nach umfassender Diskussion der Gütekriterien, der Frage nach dem Konstrukt des C-Tests und seinen zahlreichen Varianten und Einsatzmöglichkeiten, fasst folgendes Zitat von GROTJAHN (1996: 96) die Situation treffend zusammen:

> Den C-Test mit einer ein für allemal feststehenden Konstruktvalidität gibt es nicht. Es gibt vielmehr das C-Test-Prinzip (mit einer Reihe von Variationen) sowie eine Vielzahl verschiedener C-Tests in unterschiedlichen Sprachen und mit unterschiedlichen Texten, wobei streng genommen die Konstruktvalidität jedes einzelnen Tests für jede Population und jede spezifische Zielsetzung gesondert nachzuweisen ist.

In Bezug auf diese Studie bedeutet das konsequenterweise, dass überprüft werden soll, ob sich der S-C-Test im Kontext von Einstufungsverfahren als ein gültiges und zuverlässiges Messinstrument eignet.

3 Stand der Forschung

Ein Einsatz oder gar die Validierung des S-C-Tests bei erwachsenen Fremdsprachenlernern mittleren Sprachniveaus hat bislang nicht stattgefunden. Bisherige Studien, die im Rahmen dieses Forschungsprojekts relevant sind, fokussieren zwei unterschiedliche Aspekte:

- Korrelationen des C-Tests (als reinem Niveautest) mit verschiedenen sprachlichen Teilkompetenzen
- Einsatz mehr oder weniger stark zeitbegrenzter C-Tests an Muttersprachlern und sehr weit fortgeschrittenen Fremd- und Zweitsprachenlernern

Beide Gruppen von Forschungsarbeiten sollen näher beleuchtet werden (vgl. Kapitel 3.2 und 3.3). Zuvor gilt es jedoch herauszustellen, weshalb die hier fokussierten Fertigkeiten Hörverstehen und Sprechen im Hinblick auf den S-C-Test von besonderem Interesse sind.

3.1 Zur Relevanz von Hörverstehen und Sprechen beim S-C-Test

Die Rolle der vier Fertigkeiten Hörverstehen, Leseverstehen, Sprechen und Schreiben hat sich mit dem Wandel der Methodenkultur des Fremdsprachenunterrichts verändert. Standen zur Zeit der Grammatik-Übersetzungs-Methode neben der Kenntnis von Sprachregeln und Wortschatz die Fertigkeiten Lesen und Schreiben im Zentrum des Unterrichts (vgl. LARSEN-FREEMAN 2000: 18), so zeigt sich seit dem Ende des 19. Jahrhunderts und dem Aufkommen der Direkten Methode ein zunehmendes Interesse an

sprachlichen Alltagskompetenzen, was eine Verschiebung der Aufmerksamkeit weg vom Sprachwissen und hin zum Sprachgebrauch mit sich brachte (vgl. NEUNER & HUNFELD 1993: 34), was wiederum eine Fokussierung der mündlichen Fertigkeiten Hörverstehen und Sprechen bedingte. Der Anspruch an die zu erwerbenden Kompetenzen und die praktische Umsetzung im Fremdsprachenunterricht waren jedoch noch weit voneinander entfernt: Wie selbst Lehrbuchtexte aus den 1960er Jahren zeigen, war die dort vermittelte mündliche Sprache noch ausgesprochen steif und routinenhaft (vgl. z. B. SCHLIMBACH 1964; GRIESBACH & SCHULZ 1967). Der Erwerb der mündlichen Sprachproduktion sollte über Nachahmen erreicht werden (vgl. FAISTAUER 2001: 866). Auch in der Audiolingualen Methode findet keine Ablösung von den starren Kommunikationsmustern statt, denn die Lerner sollen hier zielsprachliche, mehr oder minder authentische Dialoge der Alltagssprache auswendig lernen, um auf diese Weise ein Modell für echte Kommunikation an der Hand zu haben (vgl. FAISTAUER 2001: 866). Sogenannte Satzschalttafeln und *pattern drills* waren an der Tagesordnung (vgl. LARSEN-FREEMAN 2000: 48 f.). Erst mit der kommunikativen Wende in den 1970er Jahren (vgl. SOLMECKE 2001: 893) und dem Aufblühen der kommunikativ orientierten Didaktik verliert sich diese Formelhaftigkeit, und Fremdsprachenlerner werden zu aktiven und kreativen Benutzern der Zielsprache.

Die Anforderungen an Teilnehmer kommunikativ orientierter Sprachkurse sind folglich kaum mehr mit denen zu Zeiten der Grammatik-Übersetzungs-Methode zu vergleichen. Die erfolgreiche Einstufung von Sprachkursinteressenten muss die an die Teilnehmer gestellten Anforderungen abbilden, um gelingen zu können. Im kommunikativen Fremdsprachenunterricht bedeutet dies, dass die Fertigkeiten Hörverstehen und Sprechen repräsentiert werden müssen.

Hörverstehen und Sprechen als die beiden mündlichen Fertigkeiten sind in der Praxis untrennbar miteinander verbunden. Als wechselseitige Bestandteile kommunikativer Handlungen erfordern sie von einem Sprecher sowohl das Verstehen der sprachlichen Äußerung seines Gegenübers als auch das spontane und adäquate Reagieren darauf. Jedoch erst mit der kommunikativen Wende in der Fremdsprachendidaktik kam die Einsicht, dass es sich beim fremdsprachlichen Hörverstehen mitnichten um eine passive Tätigkeit handelt, sondern dass Hörverstehen „eine eigenständige, sehr aktive und außerordentlich komplexe Vorgänge umfassende Fertigkeit ist" (SOLMECKE 2001: 893).

Der Hörverstehensvorgang setzt sich aus zweierlei Prozessen zusammen: Als aufsteigende (*bottom-up*) Prozesse bezeichnet man die Einwirkung des Hörtextes auf den Hörer. Ein Beispiel für einen solchen *bottom-up*-Prozess ist das Analysieren und Verarbeiten des auf den Hörer einwirkenden Lautstroms (vgl. HUNEKE & STEINIG: ^{5}2010: 139). Hierzu gehört beispielsweise das Identifizieren von Wortgrenzen. Zum anderen bezeichnen absteigende (*top-down*) Prozesse die Wechselwirkung zwischen dem Hörtext und dem Wissen, das ein Hörer einsetzt, um das Gehörte zu verstehen (vgl. SOLMECKE 2001: 895). Im *top-down*-Verfahren werden beispielsweise Wörter (wieder)erkannt (vgl. HUNEKE & STEINIG ^{5}2010: 139). Die beim Hörer eingehenden Schallwellen sorgen dafür, dass das Gedächtnis aktiviert wird und das Gehörte mit bereits Bekanntem bzw. Gelerntem in Zusammenhang gebracht wird (vgl. SOLMECKE 2001: 894).

Dass die Lautstruktur einer Sprache beim Lösen von C-Tests ebenfalls eine Rolle spielt, wird erst auf den zweiten Blick offensichtlich. SOLMECKE (2001: 895) stellt die Bedeutung der sprachlichen Redundanz für den Hörverstehensprozess heraus, insbesondere was die Verstehensgeschwindigkeit betrifft:

> Die Tatsache, dass bestimmte Laute und Lautkombinationen, Wörter und Satzteile mit größerer Wahrscheinlichkeit auftreten als andere, erleichtert dem Hörer ihre Identifikation auf schmaler Informationsbasis.

So zeigt sich in der Praxis, dass beim Versuch, die Lücken zu füllen, Schwierigkeiten auftreten können, wenn die Erwartungshaltung des Testteilnehmers hinsichtlich der grapho-phonemischen Struktur des gesuchten Worts nicht erfüllt wird (vgl. auch SCHÖLER 2016: 240). Im Deutschen kann dies zum Beispiel der Fall sein, wenn bei Buchstabenkombinationen „ck", „ch" oder „sch" ein Teil getilgt ist, so dass der Testteilnehmer durch inneres Verbalisieren auf eine falsche Fährte gelockt wird. Neben diesen Fähigkeiten auf lautlicher Ebene erfordert es selbstredend weiterer Kenntnisse auf morphologischer und syntaktischer Ebene, um gesprochene Sprache zu verstehen (vgl. SOLMECKE 2001: 896). Hier zeigt sich nun deutlich, dass das Konstruktionsprinzip des C-Tests dazu geeignet ist, Fähigkeiten zu erfassen, die für das erfolgreiche Verstehen fremdsprachlicher Texte unabdingbar sind. Das Prinzip der reduzierten Redundanz, das dem C-Test zugrunde liegt, findet in der alltäglichen Realität Entsprechungen, gerade wenn es um das Hörverstehen geht. An dieser Stelle sei noch einmal an die Beispiele aus Kapitel 2.1 verwiesen, den einfahrenden Zug, der die Bahnhofsdurchsage stört, oder eine schlechte Verbindung am Telefon.

HUNEKE und STEINIG ([5]2010: 139) stellen als wichtiges Merkmal des Hörverstehens den Zeitdruck heraus:

> Dem Hörer dagegen wird das Tempo seiner verstehenden Sprachverarbeitung vom Gesprächspartner, vom Vortragenden, vom Radiosprecher usw. diktiert. Das kann zu Stress führen: eine eher ungünstige Ausgangslage, wenn es darum geht, zur Lösung einer schwierigen Sprachverwendungsaufgabe die Fähigkeiten des Gedächtnisses optimal zu nutzen.

Das obige Zitat macht deutlich, welcher Bezug sich zwischen dem Hörverstehen und dem S-C-Test ergibt, welcher für den herkömmlichen C-Test nicht gegeben ist: Durch das Hinzufügen einer Geschwindigkeitskompo-

nente zum C-Test wird nicht nur das Arbeitstempo, sondern auch der Stress erhöht, dem sich die Testteilnehmer ausgesetzt sehen:

> Ein Zeitdruck kann Stress und Testangst produzieren, so dass die „wahren" Fähigkeiten der Testperson überdeckt werden. (NEUGEBAUER et al. 2012: 200)

Das Testergebnis wird also zumindest in Teilen davon abhängen, wie gut ein Teilnehmer unter den gegebenen Testbedingungen auf sein Gedächtnis zugreifen kann. Die äußeren Bedingungen, denen ein Fremdsprachenlerner beim Hörverstehen meist ausgesetzt ist, werden folglich durch den S-C-Test nachgebildet.

Wenngleich das Hörverstehen eine rezeptive Fertigkeit ist und das Sprechen eine produktive, so ist beiden Fertigkeiten jedoch eine Sache gemein: Sowohl das Hörverstehen als auch das Sprechen laufen in Echtzeit ab. Das bedeutet, anders als beim Lese- und Schreibprozess steht dem Sprecher keine Planungszeit zur Verfügung bzw. ist es dem Hörer in den allermeisten Situationen nicht möglich, den Text ein weiteres Mal zu hören. Wenngleich ein Sprecher das Verarbeitungstempo theoretisch selbst bestimmen kann (vgl. AGUADO 2003: 13), wird zum Ziel einer gelungenen Kommunikation ein angemessenes Sprechtempo angestrebt. Während monologischem Sprechen durchaus ein Planungsprozess vorgeschaltet sein kann, sind die Äußerungen in dialogischen Kommunikationssituationen stets spontan.

Dass der C-Test auch mündliche Sprachkompetenzen erfasst, wurde lange Zeit kritisch betrachtet (vgl. GROTJAHN 1995: 53). Wie jedoch in Kapitel 3.2 gezeigt wird, weist der C-Test mittelstarke Korrelationen mit diversen Operationalisierungen der mündlichen Sprachkompetenz auf. Die kommunikative Sprachkompetenz beinhaltet zahlreiche Ebenen, darunter eine pragmatische, soziolinguistische, phonetische und morphosyntaktische.

AGUADO (2003) nennt drei kognitive Prozesse, die an mündlichen Produktionen in der Fremdsprache beteiligt seien: Aufmerksamkeit, *Monitoring* und Automatisierung. Durch die Steuerung der Aufmerksamkeit können sprachliche Informationen in das Langzeitgedächtnis überführt werden (vgl. AGUADO 2003: 14). Die Rolle der Aufmerksamkeit für den L2-Erwerb unterstreicht auch die *Noticing*-Hypothese von SCHMIDT (1990). Diese besagt, dass „subliminal language learning is impossible, and that intake is what learners notice" (SCHMIDT 1990: 149). AGUADO (2003: 12) stellt zudem heraus, dass es bei der Produktion von fremdsprachlichem Output anders als beim Hörverstehen nicht genügt, wenn sich ein Lerner auf die lexiko-semantische Ebene fokussiert und die Morphosyntax ausblendet. Das sogenannte *Monitoring* bezeichnet einen Prozess, bei dem der fremdsprachliche Output durch den Sprecher selbst hinsichtlich seiner Korrektheit auf inhaltlicher wie morphosyntaktischer Ebene kontrolliert und ggf. Selbstkorrekturen eingeleitet werden (vgl. AGUADO 2003: 17). Die Beobachtung solcher Monitorprozesse gibt einen Einblick in den Spracherwerbsprozess eines Lerners:

> Je weniger automatisiert die Sprachbeherrschung ist, desto mehr bewußte Kontrolle ist erforderlich. Die Menge der für das Monitoring zur Verfügung stehenden Aufmerksamkeit beeinflusst den Verlauf und die Effizienz dieses Prozesses maßgeblich. Wenn zwei (oder mehr) Prozesse Aufmerksamkeit benötigen und um diese konkurrieren, so hat dies Auswirkungen auf die Performanz. […] Durch einen höheren Automatisierungsgrad auf den einzelnen sprachlichen Ebenen wird mehr Aufmerksamkeitskapazität für andere Bereiche der Sprachproduktion freigesetzt. (AGUADO 2003: 18 f.)

Die dritte von AGUADO (2003) genannte kognitive Komponente des Fremdsprachenerwerbs ist die Automatisierung. Diese stellt für die sprachliche Produktion in Echtzeit, d. h. für die mündliche Sprachproduktion einen wichtigen Faktor dar, denn „automaticity refers to the absence of attentional control in the execution of a cognitive activity" (SEGALOWITZ &

HULSTIJN 2005: 371). Durch die Automatisierung sprachlicher Elemente werden Kapazitäten im Arbeitsgedächtnis frei, die wiederum für andere Prozesse genutzt werden können, die ebenfalls Aufmerksamkeit benötigen (vgl. AGUADO 2003: 19), beispielsweise für den Einsatz von Lösestrategien beim Ablegen eines C-Tests.

Das sogenannte *power-law of practice* stammt aus der Psychologie und besagt, dass eine „Aufgabe (*task*) mit zunehmender Übung immer schneller und mit immer weniger Fehlern" ausgeführt werden kann (BÄRENFÄNGER 2002: 129). Hierzu passend zeigt BÄRENFÄNGER (2002: 130) eine Graphik (vgl. Abb. 7) der Automatisierung sprachlichen Wissens, die gemäß dem *Adaptive Control of Thought*-Modell (vgl. ANDERSON 1983; ANDERSON & LEBIERE 1998) drei Phasen unterscheidet.

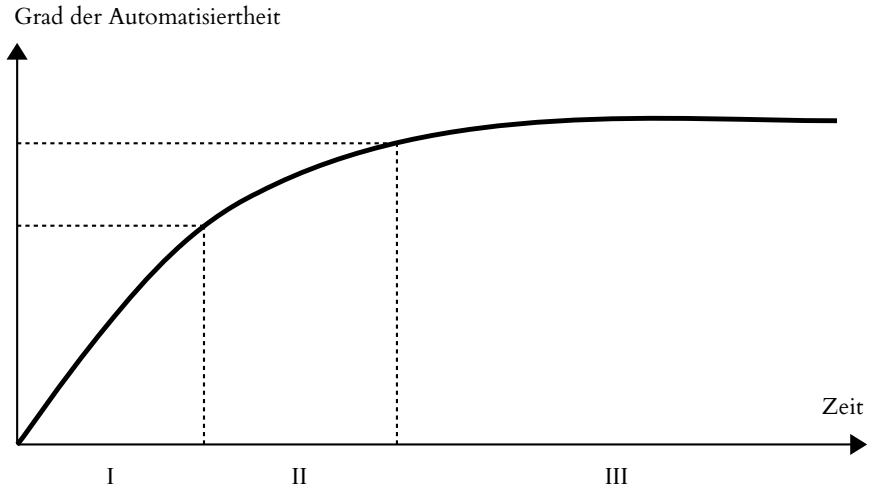


Abb. 7: Verlauf einer Lernkurve (BÄRENFÄNGER 2002: 130)

Während in der ersten Phase, der sogenannten „kognitiven Phase", sprachliche Aspekte regelhaft gelernt werden, findet in der zweiten „assoziativen Phase" die Anwendung derselben statt. Erst in der dritten „autonomen Phase" ist ein Fremdsprachenlerner dazu in der Lage, die gelernten Regeln implizit anzuwenden, d. h. ohne die Anwendung der Regeln im Geiste explizieren zu müssen (vgl. SEGALOWITZ 2003: 395). Der Lerner befindet sich somit auf dem Übergang vom *knowing that* (deklaratives Wissen) zum *knowing how* (vgl. ibid.).

Zwar weist BÄRENFÄNGER (2002: 129) darauf hin, dass die empirischen Daten, die dieser Lernkurve zugrunde liegen, mit sehr einfachen Aufgaben und teilweise gar tierischen Probanden gewonnen wurden, jedoch verweist er ebenfalls auf die Ergebnisse einer Untersuchung von DEKEYSER (1997), aus der hervorgeht, dass zumindest der Erwerb fremdsprachlicher Syntax der Lernkurve aus Abb. 7 zu entsprechen scheint. DEKEYSER (1997) untersucht den Prozess der Automatisierung bei explizit gelernten grammatischen Regeln. Hierzu lernten 61 Probanden vier morphosyntaktische Regeln sowie 32 lexikalische Einheiten einer eigens für die Studie entwickelten agglutinierenden Kunstsprache (*Autopractan*) (vgl. DEKEYSER 1997: 200). Sowohl die explizite Regelvermittlung als auch die sich anschließenden Übungseinheiten und die Überprüfung des Lernstands fanden am Computer und mithilfe von Bildern statt. DeKeyser fand, dass sowohl bei Verständnis- als auch bei Produktionsaufgaben mit zunehmender Übung zunächst ein steiler und dann asymptotischer Abfall der Reaktionszeit sowie der Fehlerhäufigkeit zu verzeichnen war. Auch AGUADO (2003: 20) weist darauf hin, dass es sich bei dem Phänomen der Automatisierung um ein Kontinuum handelt und es folglich unterschiedliche Ausprägungsgrade gibt.

In Bezug auf den C-Test kann folglich die Hypothese aufgestellt werden, dass ein Lerner morphosyntaktische Regeln der Phase II des Modells

von BÄRENFÄNGER (2002) bei einem herkömmlichen C-Test mit großzügig bemessener Arbeitszeit erfolgreich abrufen und anwenden kann. Beim Lösen eines C-Tests unter Zeitdruck könnte dies je nach Schwierigkeitsgrad des zu bearbeitenden Textes nicht mehr gelingen. Hier könnte die Fähigkeit aus Phase III, sprachliches Wissen implizit anzuwenden, für das Abschneiden im Test entscheidend sein. Um diese Hypothese zu überprüfen wird in Kapitel 4.4.3.5 ein Vergleich der stärkeren und schwächeren Hälfte der Probanden vorgenommen. Dem zugrunde liegt die Annahme, dass mit fortschreitender Fremdsprachenkompetenz auch der Grad der Automatisierung zunimmt. An dieser Stelle sei noch einmal an das bereits in Kapitel 2.2.5 erwähnte Phänomen des *Speed-Accuracy-Trade-Off* verwiesen, welches besagt, dass beim Arbeiten unter Zeitdruck Fehler gemacht werden, die ohne diesen nicht auftreten würden (vgl. ZIEGLER & BÜHNER 2012: 70).

PARADIS (2009: X) stellt jedoch heraus, dass es ein Trugschluss sei, von der Arbeitsgeschwindigkeit auf eine vorliegende Automatisierung zu schließen, da auch kontrollierte Prozesse durchaus mit hoher Geschwindigkeit ablaufen können:

> Conscious production **can** be more or less speeded-up, that is, more or less efficiently controlled. Control admits degrees of velocity in the performance of a task. But we cannot have more or less control over computational procedures that we are unaware of. Hence, automaticity does not admit to degrees. (PARADIS 2009: X; vgl. auch Seite 64 f.)

Des Weiteren erklärt er, dass metasprachliches Wissen (*metalinguistic knowledge*) und implizite Sprachkompetenz (*implicit language competece*) niemals ineinander überführt werden könnten, da es sich um getrennte Systeme handele, die nicht miteinander interagieren. Dies sei dadurch zu erklären, dass sie auf unterschiedlichen Teilen des Gedächtnisses basieren (vgl. ibid. XI). Vielmehr sei es so, dass die Systeme nebeneinander existieren und die

implizite Sprachkompetenz erst ab einem bestimmten Punkt für den Lerner verfügbar wird (vgl. Abb. 8).

Abb. 8: Metasprachliches Wissen (MLK) und implizites Wissen (ILC) (Paradis 2009: 68)

Auch TSCHIRNER (2001) teilt die Auffassung, dass eine Automatisierung bewussten Regelwissens durch Üben nicht möglich sei, da es sich bei dem gelernten Wissen und den mentalen Regeln um unterschiedliche Systeme handele.

Für Paradis ergibt die Schlussfolgerung:

> Inciental acquisition through practice is the only way to internalize implicit linguistic competence. But it is not the only way to become a proficient, fluent speaker of L2: Explicit learning may lead to speeded-up controlled use of a second language. (PARADIS 2019: 8)

LIST (2002: 127) hingegen verweist auf MARKOWITSCH (1996), um herauszustellen, dass es eine neurophysiologische Grundlage für die Unterscheidung zwischen deklarativen und prozeduralen Wissensbeständen gebe. Markowitsch führt das Sprechen als ein Beispiel dafür an, dass verschiedene Formen des Gedächtnisses im Zusammenspiel aktiv sind, wobei das prozedurale Gedächtnis für die Koordination der Sprechwerkzeuge benötigt wird, während sich der Inhalt des Gesprochenen auf deklarative Wissensbestände stützt. Jedoch scheint das prozedurale Gedächtnis noch auf einer anderen Ebene in die Sprachproduktion eingebunden zu sein, denn MARKOWITSCH (1996: o. S.) formuliert „daß man Wörter und grammatische Regeln einer fremden Sprache oder Geschichtsdaten und ähnliche Fakten automatisch beherrschen, also über das prozedurale Gedächtnis re-

gelrecht trainieren muß, wenn sie ohne Mühe parat sein sollen". LIST (2002: 129) wiederum konstatiert, dass „implizite Lernvorgänge von Grund auf, also ein Aufbau prozeduralen Wissens im eigentlichen Sinne, […] wegen der bereits hergestellten kortikosubkortikalen Kreisläufe nicht mehr in der gleichen Weise möglich" seien, wie dies beim kindlichen Erstsprachenerwerb der Fall sei. Daraus lässt sich schlussfolgern, dass der Erwerb prozeduralisierten (!) fremdsprachlichen Wissens, d. h. der Aufbau von automatisierten Wissensbeständen in der Fremdsprache über den Umweg des deklarativen Wissens über umfassendes Üben des Lerngegenstands erreicht wird. Auch dies steht mit der Lernkurve aus BÄRENFÄNGER (2002: 130) im Einklang.

AGUADO (2003: 14) stellt folgenden Zusammenhang zwischen der Automatisierung sprachlicher Elemente und der Produktionsgeschwindigkeit her:

> Je häufiger bestimmte sprachliche Äußerungen verwendet werden, desto ineffizienter wird es, sie bei jedem Gebrauch erneut regelhaft zu bilden. Daher werden insbesondere hochfrequente sprachliche Ausdrücke ganzheitlich memorisiert und ebenso abgerufen – was sich nicht zuletzt in einer höheren Produktionsgeschwindigkeit niederschlägt.

Hieraus lässt sich ableiten, dass ein höherer Grad an Automatisierung in der Fremdsprache eine höhere Arbeitsgeschwindigkeit beim Lösen von C-Tests zulässt, weil bestimmte Phrasen, Kollokationen oder idiomatische Ausdrücke als automatisierte Einheiten vorliegen und nicht über das Anwenden einer morphosyntaktischen Regel verarbeitet oder produziert werden müssen.

Es zeigt sich also, dass die Fertigkeiten Hörverstehen und Sprechen durch die ihnen innewohnende Verarbeitungsgeschwindigkeit in Bezug auf den S-C-Test von besonderem Interesse sind. Wenngleich auch der Zusammenhang des durch eine Geschwindigkeitskomponente modifizierten Testformats mit den schriftlichen Fertigkeiten sowie anderen fremdsprachlichen Kompetenzen wie Wortschatz und Grammatik ebenfalls er-

gründet werden sollte, kann in der vorliegenden Studie aus Gründen der Kapazität nur ein Blick auf einen Ausschnitt des fremdsprachlichen Kompetenzspektrums geworfen werden.

Im Folgenden werden die beiden Gruppen von Studien – d. h. herkömmliche Korrelationsstudien, die den C-Test mit den Fertigkeiten Hörverstehen und Sprechen in Beziehung setzen sowie Studien, die in ganz unterschiedlicher Weise einen S-C-Test beinhalten – betrachtet. Hierbei wird ein besonderes Augenmerk auf die verwendeten Instrumente gelegt, da diese die gefundenen Korrelationen maßgeblich beeinflussen können. Im Falle der Studien zum S-C-Test wird darüber hinaus das Ziel dessen Einsatzes herausgestellt.

3.2 Korrelationsstudien zum C-Test mit den Fertigkeiten Hörverstehen und Sprechen

Der Frage, inwieweit der C-Test mit unterschiedlichen sprachlichen Teilkompetenzen korreliert, wurde in einigen Untersuchungen nachgegangen, hauptsächlich im Rahmen von Validierungsstudien, die einen C-Test gegen unterschiedliche Außenkriterien korrelieren. Während sich die Fertigkeit Hörverstehen als Komponente einiger Studien findet, ist die Datenlage beim mündlichen Sprachgebrauch – vermutlich aufgrund des deutlich höheren Aufwands bei der Datenerhebung und -aufbereitung – relativ dünn (vgl. Tab. 3.3). Im Folgenden werden Studien, die Korrelationen zwischen einem C-Test und den Fertigkeiten Hörverstehen und Sprechen ermitteln und somit relevant für dieses Projekt sind, vorgestellt. Die meisten dieser Studien beinhalten auch Daten zum Verhältnis zwischen C-Test und anderen sprachlichen Kompetenzen, beispielsweise Leseverstehen oder Wort-

schatz. Diese werden im Folgenden jedoch nicht thematisiert. Auch weitere Korrelationsstudien, die jedoch die hier fokussierten Fertigkeiten außer Acht lassen, werden nachfolgend nicht behandelt.[18]

Chapelle und Abraham (1990) vergleichen die Reliabilität eines C-Tests und seine Korrelationen mit verschiedenen sprachlichen Teilkompetenzen, u. a. Hörverstehen mit der Reliabilität verschiedener Varianten des damals noch recht populären Cloze-Tests (vgl. Kapitel 2.2). Die Autoren erstellten einen fünfteiligen C-Test, dessen Teile – anders als bei C-Tests üblich – aus einer einzigen Textgrundlage, einem Artikel aus *Scientific American*, stammen. Jeder Textabschnitt hatte 15 Lücken. Für den C-Test ermitteln Chapelle und Abraham (1990) mittels der Kuder-Richardson-Formel eine Reliabilität von KR-20 = 0,81. Als Außenkriterium fungierte der *Iowa State English Placement Test*, welcher aus drei Teilen besteht: Leseverstehen, Hörverstehen und Wortschatz. Der Testteil Hörverstehen besteht aus 35 Multiple-Choice-Items. Die Autoren beschreiben die Items weiterhin wie folgt: „Some items require students to discern the grammatical details of what they heard while others require overall comprehension of discourse." (Chapelle & Abraham 1990: 129). Die insgesamt 201 Probanden waren fortgeschrittene Englischlerner mit unterschiedlicher L1. Sie legten den *Iowa State English Placement Test* ab und füllten einen von vier möglichen Lückentexten aus. Der C-Test wurde folglich von etwa 50 Probanden bearbeitet. Hierzu erhielten sie eine Bearbeitungszeit von 25 Minuten. Chapelle und Abraham ermitteln zwischen Hörverstehen und C-Test eine Korrelation von 0,472, geben jedoch nicht an, welchen Korrelationskoeffizienten sie für die Rechnung genutzt haben.

Dörnyei und Katona (1992) streben mit ihrer Studie einen Beitrag zur Validierung des C-Test-Formats an. Sie legten 102 ungarischen Anglis-

18 Weitere Korrelationsstudien zum C-Test, die die Fertigkeiten Sprechen und Hören nicht berücksichtigen, finden sich in Grotjahn (2014).

tikstudenten zu Beginn ihres Studiums an der Eötvös Universität in Budapest eine mehrteilige Testbatterie vor. Der eingesetzte C-Test wurde eigens für diese Studie entwickelt und bestand aus vier Texten mit variierender Anzahl an Lücken zwischen 17 und 24. Der C-Test wies in der beschriebenen Probandengruppe eine Reliabilität von Cronbachs Alpha $\alpha = 0{,}75$ auf. Das ist ein relativ niedriger Wert, denn da das C-Test-Format generell sehr messgenau ist, erwartet man von einem C-Test für gewöhnlich eine Reliabilität von mindestens 0,8 (vgl. GROTJAHN 2004: 538). Eine Vorerprobung des C-Tests hat nicht stattgefunden (vgl. DÖRNYEI & KATONA 1992: 202). Neben dem institutseigenen Einstufungstest bestehend aus je einem Aufgabenteil zu Wortschatz, Grammatik und Hörverstehen legten die Probanden den Hör- und Leseverstehenstest des TOEIC (vgl. URL 2: TOEIC) sowie ein mündliches Interview ab. Letzteres dauerte etwa 10 bis 15 Minuten und wurde von zwei Prüfern abgenommen, von denen einer Englisch als Muttersprache sprach. Die Bewertung des Interviews erfolgte hinsichtlich der Kriterien Korrektheit, Wortschatz und Flüssigkeit auf einer fünf-Punkt-Skala. Der Hörverstehensteil des TOEIC setzt sich aus vier Teilen (*Photographs, Question-Response, Conversations, Short Talks*) zusammen, die alle aus Multiple-Choice-Aufgaben mit drei bzw. vier Antwortmöglichkeiten bestehen. Der gesamte Hörverstehenstest dauert 45 Minuten (vgl. EDUCATIONAL TESTING SERVICE 2015: 9). Eine Besonderheit bei diesem Testformat ist, dass die Testteilnehmer weder den Stamm der Items lesen können noch die Antwortoptionen. Beides wird lediglich einmalig gehört. DÖRNYEI und KATONA (1992) fanden statistisch signifikante Korrelationen zwischen dem C-Test und den Leistungen im Hörverstehen zwischen 0,33 (institutseigener Test) und 0,51 (TOEIC) ($p < 0{,}001$). Die Korrelation des C-Tests mit dem Interview lag bei 0,43 ($p < 0{,}001$) (vgl. DÖRNYEI & KATONA 1992: 192). Mit welchem Korrelationskoeffizienten diese Werte ermittelt wurden, geben die Autoren nicht

an. Bemerkenswert ist darüber hinaus, dass die Autoren berichten, Schwierigkeiten bei der Suche nach ausreichend schwierigen Texten für sehr kompetente Fremdsprachenlerner gehabt zu haben (vgl. DÖRNYEI & KATONA 1992: 190). Über die Zeitbemessung der schlussendlich eingesetzten C-Test-Texte machen die Autoren ebenfalls keine Angabe.

Die Studie von **CHIHARA, CLINE und SAKUMI (1996)** setzt sich zum Ziel, die Überlegenheit des C-Tests gegenüber dem Test über eine Validierung des C-Tests mit dem TOEFL zu überprüfen. Hierzu erstellen die Autoren zunächst zwölf C-Test-Texte à 25 Lücken und erproben diese an 180 Studenten eines japanischen *Junior College*. Resultierend aus dieser Pilotierung stellen CHIHARA et al. (1996) zwei parallele Testsets aus jeweils vier Texten zusammen (Basic Form A und Basic Form B). Hinzu kommen zwei sogenannte experimentelle C-Test-Formen. Diese beruhen auf der gleichen Textgrundlage wie die erstgenannten C-Tests, jedoch beginnt die Tilgung hier nicht dem ursprünglichen Prinzip folgend beim zweiten Wort des zweiten Satzes, sondern bereits beim ersten Wort des zweiten Satzes (Experimental Form A und Experimental Form B). Jeweils 220 Probanden legten jedes so zusammengestellte C-Test-Set ab. Eine Angabe über die den Probanden zur Verfügung gestellte Bearbeitungszeit liefern die Autoren nicht. Eine Analyse der Reliabilität mittels der Kuder-Richardson-Formel ergab Werte zwischen 0,748 und 0,810. Der Teil Hörverstehen des TOEFL beinhaltet 31 bis 54 Items, wofür den Testteilnehmern 60 bis 90 Minuten Bearbeitungszeit zur Verfügung stehen. Inhaltlich besteht der Test aus Vorlesungen und Seminardiskussionen zu denen Fragen im Multiple-Choice-Format mit vier Antwortmöglichkeiten beantwortet werden müssen (vgl. URL 27: TOEFL ibt: Test Questions). Bei einer Korrelation aller vier C-Test-Sets gegen den TOEFL ergaben sich folgende Werte für Pearsons Produkt-Moment-Korrelationskoeffizienten (vgl. Tab. 5):

	Basic Form A (n = 82)	Basic Form B (n = 93)	Experimen. Form A (n = 93)	Experimen. Form B (n = 82)
TOEFL (Hörverstehen)	r = 0,417	r = 0,569	r = 0,611	r = 0,357

Tab. 5: Pearson-Korrelationen zwischen C-Test-Sets und Subtest Hörverstehen des TOEFL[19]

Da sich für die anderen Subtests des TOEFL (Sprachstrukturen, Wortschatz und Leseverstehen) ebenso mittlere positive Korrelationen finden, schlussfolgern die Autoren, dass der C-Test ein Messinstrument der allgemeinen Sprachkompetenz darstellt.

Ähnlich wie DÖRNYEI und KATONA (1992) ist auch das Ziel **HUHTAs (1996)** die Validierung eines C-Tests. Jedoch geht es hier um einen konkreten C-Test, der die bisherige Testbatterie zum Messen des konkreten Kursziels innerhalb des Anglistikstudiums an der Universität Jyväskylä (Finnland) ablösen sollte. Untersucht wurden 129 Englischstudenten der Universität Jyväskylä. Der eingesetzte C-Test wurde eigens zu diesem Zweck erstellt und bestand aus fünf Texten à 20 Lücken. Dies entspricht einem sehr üblichen C-Test-Format. Ihm voran wurde noch ein Eisbrechertext gestellt, um die Probanden mit dem Testformat vertraut zu machen. Dieser wird nicht bewertet. Unüblich ist hingegen das Vorhandensein von direkter Rede bzw. dialogischem Textmaterial innerhalb des C-Tests (vgl. HUHTA 1996: 201). Eine Vorerprobung an sieben Personen diente u. a. dazu, die Texte in aufsteigender Schwierigkeit anzuordnen. Das Außenkriterium bestand aus mehreren Teilen: einem aus etwa 250 Wörtern bestehendem Diktat, einem Cloze-Test mit 35 Lücken sowie aus 100 Items bestehender Multiple-Choice- bzw. richtig-falsch-Tests zur Grammatik und zum Wortschatz. Die Bearbeitungszeit für den Cloze- und C-Test gemeinsam wird mit einer Stunde angegeben. Für einen Teil der

19 Es fehlen die Angaben über das Signifikanzniveau.

Probanden lagen darüber hinaus die Ergebnisse des Studieneingangstests vor, der aus je einem Teil zum Leseverstehen, Grammatik, Wortschatz und Hörverstehen bestand. Über den Hörverstehenstest berichtet der Autor lediglich, dass es sich um einen Test im Multiple-Choice-Format handelt. HUHTA (1996) findet abhängig von der Auswertungsmethode des C-Tests Korrelationen mit dem Hörverstehen zwischen 0,16 (*lenient scoring*) und 0,19 (*exact scoring*). Wenngleich HUHTA (1996) zu den verwendeten Korrelationskoeffizienten keine Angabe macht, sind die von ihm gefundenen Korrelationen zwischen C-Test und Hörverstehen auffallend gering (vgl. Tab. 18 auf S. 143).

Da es nur wenige Studien bzgl. der Korrelation des C-Tests mit mündlichen Fertigkeiten gibt, soll an dieser Stelle auch die Untersuchung von **JAKSCHIK (1996)** Erwähnung finden. Er korrelierte im Rahmen beruflicher Bildungsmaßnahme die C-Test-Ergebnisse erwachsener DaZ-Lerner mit den Leistungseinschätzungen von Lehrern. Auch diese Studie hat zum Ziel, den C-Test zu validieren, hier speziell für die Zielgruppe der erwachsenen DaZ-Sprecher. Der hier eingesetzte C-Test besteht aus sechs Texten à 20 Lücken und wurde zuvor in einigen Arbeitsämtern erprobt (vgl. JAKSCHIK 1992). Die Bearbeitungszeit des C-Tests war unlimitiert. Die Einschätzung der Lehrer bzgl. der „Sprachlichen Fertigkeiten im mündlichen Bereich" wurde auf einer fünf-Punkt-Skala gegeben, die sich von „deutlich überdurchschnittlich" bis hin zu „deutlich unterdurchschnittlich" bewegte. Bezugspunkt hier war das Kursniveau. Die Probanden waren nach angestrebten Berufsgruppen unterteilt. Es zeigen sich folgende Ergebnisse (vgl. Tab. 6).

	Kaufmännischer Bereich (n = 412)	Metallbereich (n = 142)	Elektrobereich (n = 124)
C-Test	ρ = -0,38*	ρ = -0,48*	ρ = -0,54*

* $p < 0{,}001$ (einseitig)

Tab. 6: Spearman-Rho-Korrelationen zwischen Lehrereinschätzung der „Sprachlichen Fertigkeiten im mündlichen Bereich" und C-Test (JAKSCHIK 1996: 258 ff.)

Die negativen Werte der hier aufgeführten Rangkorrelationen ergeben sich aus der Polung der Einschätzungsskala für die Lehrerurteile. Entscheidend ist der Betrag der angeführten Werte. Für alle Berufsgruppen zeigen sich schwache signifikante Korrelationen.

ARRAS, ECKES und GROTJAHN (2002) untersuchen die Eignung des C-Tests als Ankertest im Rahmen der Erprobungen neuer TestDaF-Aufgaben. Hierzu wurden zuvor Grammatik- und Lexiklückentexte eingesetzt, die die *University of Cambridge Local Examinations Syndicate* (UCLES) entwickelt hatte und deren Schwierigkeit bekannt war (vgl. ARRAS et al. 2002: 176). Entsprechend der Zielgruppe des TestDaF besteht die Probandengruppe dieser Studie aus Deutschlernern mit unterschiedlichen Erstsprachen, die entweder bereits ein Studium in einem deutschsprachigen Land aufgenommen haben oder dies beabsichtigen zu tun. Nach Pilotierung mehrerer C-Test-Texte besteht der finale C-Test bei ARRAS et al. (2002) aus vier Texten à 20 Lücken. Die Themen der C-Test-Texte stammen der Zielgruppe des TestDaF entsprechend aus dem akademischen Kontext und sind größtenteils populärwissenschaftlichen Zeitschriften und ähnlichen Textquellen entnommen. Zur Bearbeitung des C-Tests hatten die 187 Testteilnehmer fünf Minuten pro Text. Die Reliabilitätsanalyse ergibt einen Wert von Cronbachs Alpha von α = 0,842 für den C-Test. Der Subtest Hörverstehen des TestDaF besteht aus drei Teilen mit insgesamt 25 Items, die aus richtig-falsch-Aufgaben sowie dem Bearbeiten von Fragen per Kurzantwort bestehen (vgl. URL 4: TestDaF). Das Format des mündlichen

Ausdrucks beim TestDaF ist an dieser Stelle hervorzuheben: Es handelt sich um ein sogenanntes *Simulated Oral Proficiency Interview* (SOPI) (vgl. STANSFIELD 1989). Hierbei legen die Testteilnehmer den Prüfungsteil an einem Computer ab. Es erfolgt zunächst ein verbaler oder graphischer Stimulus, woraufhin die Testteilnehmer ihre Ausführungen über ein Mikrophon in den Computer sprechen. Es gibt insgesamt sieben Stimuli, auf die die Testteilnehmer sprachlich reagieren sollen. Auf den Stimulus folgt zunächst eine Vorbereitungszeit, die ebenso lang ist wie die sich anschließende Redezeit. Diese beträgt jeweils zwischen 0:30 und 1:30 Minuten (vgl. URL 4: TestDaF). Der sprachliche Output ist hier folglich rein monologisch, eine wechselseitige Kommunikation findet nicht statt. Für den Testteil mündlicher Ausdruck sind insgesamt 25 Minuten vorgesehen. Die Ergebnisse des C-Tests werden gegen alle Subtests des TestDaF korreliert. Für die Fertigkeiten Hörverstehen und mündlicher Ausdruck ergeben sich mittlere signifikant positive Korrelationen (vgl. Tab. 7):

	TestDaF Hörverstehen (n = 187)	TestDaF mündlicher Ausdruck (n = 145)
C-Test (Spearman Rho)	ρ = 0,636	ρ = 0,640
C-Test (Kendalls Tau b)	τ = 0,483	τ = 0,521

Alle Korrelationen sind auf dem Niveau 0,01 zweiseitig signifikant.

Tab. 7: Korrelationen zwischen C-Test und den Subtests Hörverstehen und mündlicher Ausdruck des TestDaF

Wie zu erwarten fallen die Werte für Kendalls Tau b etwas geringer aus als die Werte von Spearman Rho (vgl. HILL & LEWICKI 2006: 37). ARRAS et al. (2002) heben insbesondere die recht hohe Korrelation des C-Tests mit dem Subtest mündlicher Ausdruck (MA) hervor und schlussfolgern aus den Werten von Spearman Rho und Kendalls Tau b, dass „C-Test und MA in nicht unbeträchtlichem Maße die gleichen (zugrundeliegenden) Fähigkeiten zu erfassen“ scheinen (ARRAS et al. 2002: 201).

Jafarpur (2002) behandelt in seiner Studie die Frage, ob der C-Test ein besseres Instrument zur Messung der Sprachkompetenz ist als der Test. Dazu wird 146 iranischen Studenten ein Cloze-Test, ein C-Test sowie eine als Außenkriterium fungierende Testbatterie vorgelegt und deren Ergebnisse gegeneinander korreliert. Der hier eingesetzte C-Test besteht aus vier Texten, die aus didaktischem Material für Zweit- und Fremdsprachenlerner des Englischen stammen. Dies ist ungewöhnlich, da für C-Tests laut Grotjahn (2002: 222) ausschließlich authentische Texte Verwendung finden sollten. Weiterhin wiesen die vier Texte eine unterschiedliche Anzahl von Lücken auf (Text 1: 29 Lücken; Text 2: 20 Lücken; Text 3: 29 Lücken; Text 4: 22 Lücken). Ein weiteres ungewöhnliches Vorgehen ist, dass die getilgten Wortenden mit einem Unterstrich pro fehlendem Buchstaben ersetzt wurden. Dieses Vorgehen gibt den Testteilnehmern einerseits einen Hinweis auf das zu rekonstruierende Wort. Andererseits ist die Möglichkeit für korrekte Alternativlösungen hierdurch sehr stark eingeschränkt. Der C-Test erreicht eine Reliabilität von KR-21 = 0,92. Als Außenkriterium dient der sogenannte *English Placement Test, Form B* (Corrigan et al. 1978). Er besteht aus vier Teilen: Hörverstehen, Leseverstehen, Grammatik und Wortschatz. Der Subtest zum Hörverstehen besteht aus 20 Multiple-Choice-Items, bei denen die Testteilnehmer entweder die Antwort auf eine im Input gestellte Frage ankreuzen oder einen Satz mit gleicher Bedeutung identifizieren müssen. Der Hörverstehenstest erreicht lediglich eine Reliabilität von KR-21 = 0,6, was nicht zuletzt in der relativ geringen Anzahl der Testitems begründet liegen kann. Die mittels Pearsons-Produkt-Moment-Korrelationskoeffizienten errechnete Korrelation zwischen dem eingesetzten C-Test und Hörverstehen liegt bei $r = 0{,}65$ ($p < 0{,}001$).

Krekelers (2002) Untersuchung zielt darauf ab, zwei Deutschprüfungen für den Hochschulzugang und deren Kompetenzstufen miteinander zu vergleichen: den standardisierten TestDaF und die nicht standardisierte

DSH (Deutsche Sprachprüfung für den Hochschulzugang). Das Hörverstehen der DSH besteht aus einem Vortrag, an den sich unterschiedliche Aufgaben anschließen können, zum Beispiel Fragen zum Text oder das Anfertigen eines Resümees (vgl. HRK & KMK 2011: 12 f.). (Zum Format des Hörverstehens und Sprechens beim TestDaF siehe URL 4: TestDaF). Ausländische Studienbewerber mit unterschiedlicher Herkunft und L1 (n = 67) legten hierzu zunächst eine Erprobungsfassung des TestDaF ab, die standardmäßig auch einen C-Test als Anker enthält. Über den C-Test macht der Autor nur wenige Angaben. Es handelt sich jedoch um einen vom TestDaF-Institut entwickelten Test mit 80 Lücken (der somit vermutlich aus vier Texten à 20 Lücken besteht). Mit drei Monaten Abstand legten die Probanden eine an der Fachhochschule Konstanz erstellte DSH-Prüfung ab. Die Korrelation der Testergebnisse wurde mittels Spearman-Rho ermittelt und liegt im mittleren Bereich (vgl. Tab. 8; vgl. KREKELER 2002: 35).

	Hörverstehen – DSH	Hörverstehen – TestDaF	Sprechen – TestDaF
C-Test	ρ = 0,437*	ρ = 0,439**	ρ = 0,415**

* signifikant auf dem Niveau 0,01 signifikant (2-seitig)
** signifikant auf dem Niveau 0,05 signifikant (2-seitig)

Tab. 8: Korrelationen zwischen C-Test und den Subtests Hörverstehen der DSH und TestDaF und mündlichem Ausdruck des TestDaF

Bemerkenswert an diesen Ergebnissen ist, dass die ermittelten Korrelationen für das Hörverstehen von der DSH und dem TestDaF sehr nah beieinanderliegen, obwohl sich die Prüfungsformate voneinander unterscheiden.

DALLER und PHELAN (2006) untersuchen, ob C-Tests sich dazu eignen, den Lernfortschritt in Intensivsprachkursen zu messen. Dazu ließen sie 32 französische Studenten, die an einem intensiven Englischkurs in England teilnahmen, zu Beginn und zum Ende des elfwöchigen Kurses folgende Tests bearbeiten: Einen aus sechs Texten à 20 Lücken bestehenden

C-Test, der sich inhaltlich mit landeskundlichen und gesellschaftlichen Themen Englands befasst. Die Reliabilität des C-Tests lag bei $\alpha = 0{,}836$. Darüber hinaus legten die Probanden *The Test of English for International Communication* (TOEIC) (vgl. URL 2: TOEIC) ab. Dieser Test hat den Vorteil, ein etabliertes Testverfahren zu sein. Jedoch testet der TOEIC (in der Version von 2002/2003) ausschließlich die rezeptiven Fertigkeiten Hörverstehen und Leseverstehen – ein auch von den Autoren angesprochener Nachteil dieses Tests. Hörverstehen und Leseverstehen sind innerhalb des TOEIC gleich gewichtet. Der Hörverstehensteil besteht aus vier Subtests mit insgesamt 100 Fragen und hat eine Durchführungsdauer von 45 Minuten. Im ersten Aufgabenteil hören die Testteilnehmer je vier Sätze und müssen herausfinden, welcher gehörte Satz am besten zu einem vorgegebenen Foto passt (40 Items). Der zweite Aufgabenteil besteht aus 20 Multiple-Choice-Items, bei denen die Probanden jeweils die richtige Antwort zu einer Frage finden müssen. Im dritten Aufgabenteil sollen die Probanden im Multiple-Choice-Format Fragen zu einem gehörten Gespräch beantworten (30 Items). Im vierten und letzten Hörverstehensteil hören die Teilnehmer einen kurzen Monolog und beantworten im Anschluss wiederum im Multiple-Choice-Format zwei bis drei Fragen dazu (30 Items). Das Leseverstehen umfasst drei Aufgabenteile mit insgesamt 90 Fragen und dauert 75 Minuten. Im ersten Aufgabenteil sollen die Testteilnehmer unvollständige Sätze über eine Mehrfachauswahl vervollständigen (40 Items). Der zweite Leseverstehensteil besteht aus 20 Sätzen, in denen je vier Wörter unterstrichen sind. Es ist die Aufgabe der Probanden herauszufinden, welche Wörter korrekt bzw. falsch sind. Im dritten Aufgabenteil lesen die Teilnehmer kurze Textabschnitte und beantworten im Anschluss je zwei bis fünf Multiple-Choice-Fragen dazu (30 Items). Daller und Phelan ermittelten eine signifikante Korrelation von $r = 0{,}775$ zwischen dem Eingangs- und Ausgangs-C-Test. Lerner, die also im Eingangstest besser abschnitten,

taten dies auch im Ausgangstest, was ein zu erwartendes Ergebnis darstellt. Ähnlich verhält es sich mit dem Eingangs- und Ausgangs-TOEIC, welche eine ebenfalls signifikante Korrelation von r = 0,473 aufweisen. Zwischen TOEIC und C-Test fanden die Autoren folgende Korrelationen (vgl. Tab. 9; vgl. DALLER & PHELAN 2006: 113):

	C-Test (Kursbeginn)	C-Test (Kursende)
Hörverstehen (Kursbeginn)	**r = 0,455 (*p* = 0,011, 2-seitig)**	r = 0,368 (*p* = 0,045)
Hörverstehen (Kursende)	r = 0,555 (*p* = 0,001)	**r = 0,415 (*p* = 0,023, 2-seitig)**

Tab. 9: Pearson-Korrelationen zwischen C-Tests und dem Subtest Hörverstehen des TOEIC

Da zwischen Kursbeginn und Kursende ein Lernzuwachs zu verzeichnen sein sollte, sind hier vorrangig die fettgedruckten Tabellenfelder von Belang.

Die Autoren schlussfolgern, dass es möglich ist, einen C-Test als Lernfortschrittstest in Intensivkursen einzusetzen, da sowohl der TOEIC als auch der C-Test beim Posttest einen höheren Punktwert der Lerner gegenüber der Leistung im Prätest zeigt (vgl. Tab. 10; vgl. DALLER & PHELAN 2006: 111).

	$\bar{x}$	SD	Differenz in Prozent
TOEIC Hörverstehen Kursbeginn	335,67	64,84	18,42 %
TOEIC Hörverstehen Kursende	397,5	48,35	
C-Test Kurseingang	56,97	13,30	24,92 %
C-Test Kursende	71,17	8,71	

Tab. 10: Arithmetisches Mittel und Standardabweichung bei C-Test und dem Subtest Hörverstehen des TOEIC

Unklar bleibt jedoch, inwieweit Gedächtniseffekte hier eine Rolle spielen, denn bei Prä- und Posttest wurden identische Tests eingesetzt. Unterschiedliche Gedächtnisleistungen unter den Probanden könnten auch eine Erklärung für beim Posttest niedriger ausfallende Korrelationen zwischen C-Test und TOEIC-Subtests bieten. Kritisch anzumerken bleibt außerdem, dass die Teilnehmerzahl mit $n = 32$ recht klein ist. Über die Darbietungsform des C-Tests und eine etwaige Zeitlimitierung werden keine Angaben gemacht. Zudem sollte bedacht werden, dass das hier genutzte Außenkriterium, der TOEIC, einen Berufsbezug hat und somit eher fach- als allgemeinsprachlich orientiert ist.

Die Studie von **Lei (2008)** stellt den Versuch dar, einen C-Test bei chinesischen Lernern des Englischen zu validieren. Die Hypothese des Autors ist, dass der C-Test allgemeine Sprachkompetenz misst, die sich aus verschiedenen Komponenten zusammensetzt (vgl. Lei 2008: 123). Zur Überprüfung dieser Hypothese bekamen 265 Erstsemesterstudenten der Huazhong Universität folgende Tests vorgelegt: einen eigens für die Studie erstellten C-Test mit fünf Texten à 20 Lücken. Die verwendeten Texte stammten aus *21st Century College English* (1999) und behandeln diverse Themen (vgl. Lei 2008: 139 f.). Die Reliabilität des C-Tests lag bei $\alpha = 0{,}747$. Dieser Wert liegt unter den bei C-Test zu erwartenden 0,8 (vgl. Grotjahn 2004: 538), ist jedoch noch zufriedenstellend (vgl. Moosbrugger & Kelava 2007: 11). Neben dem C-Test legten die Teilnehmer den Test des sogenannten *College English Book I* ab. Dieser besteht aus sechs Teilen: Leseverstehen, Hörverstehen, Wortschatz, Cloze-Test, Übersetzung und einer Schreibaufgabe. Der Testteil Hörverstehen besteht aus Kurzdialogen und kurzen Textpassagen. Lei (2008: 130) ermittelt eine signifikante Korrelation zwischen C-Test und Hörverstehen von 0,318, gibt jedoch nicht an, für welchen Korrelationskoeffizienten.

Reichert, Keller und Martin (2010) haben in ihrer Studie angestrebt, einen C-Test mit dem Gemeinsamen Europäischen Referenzrahmen (Europarat 2001) zu verbinden. Hierzu wählten sie ein indirektes Vorgehen, indem sie einen C-Test mit einer bereits am Gemeinsamen Europäischen Referenzrahmen ausgerichteten, standardisierten Prüfung korrelierten. Die Probanden waren 288 luxemburgische Schüler von vier verschiedenen Schulen, die kurz vor ihrem Abschluss standen. Das verwendete Außenkriterium war der *Test de Connaissance du Français* (TCF), einem vom französischen Bildungsministerium herausgegebenen standardisiertem Sprachtest. Es handelt sich dabei um eine skalierte Sprachprüfung, die den Kompetenzbereich von A1 bis C2 abdeckt. Sie besteht aus drei obligatorischen (Hörverstehen, Sprachstrukturen, Leseverstehen) und zwei fakultativen (mündlicher Ausdruck, schriftlicher Ausdruck) Prüfungsteilen. Der Test zum Hörverstehen besteht aus vier Teilen, die alle aus Multiple-Choice-Items mit vier Antwortoptionen bestehen. Die Höraufgaben reichen von bildgestütztem Verständnis über Alltagstexte wie Bahnhofsdurchsagen über kurze Konversationen bis hin zu Vorträgen wie sie beispielsweise im Radio zu hören sind (vgl. URL 28: CIEP: TCF). Der Subtest zur mündlichen Ausdrucksfähigkeit besteht neben der Vorstellung des Kandidaten aus einer kommunikativen (dialogischen) Aufgabe und dem Ausdrücken des eigenen Standpunkts zu einem Thema. Die in der Studie genutzten C-Tests stammten größtenteils aus der TESTATT-Studie (Raatz et al. 2006) und waren somit bereits zuvor erfolgreich eingesetzt worden. Raatz et al. (2006) dokumentieren Reliabilitäten zwischen $\alpha = 0{,}87$ und $\alpha = 0{,}93$ für verschiedene Testversionen und Teilnehmergruppen. Reichert et al. (2010) erstellten zusätzlich weitere C-Test-Texte. Alle in der Studie genutzten Texte wurden vor Durchführung der Untersuchung zweifach vorerprobt und ihre Schwierigkeit ermittelt. Das finale Test-Set bestand aus neun Texten. Jeder Text hatte 20 Lücken, d. h. es waren maxi-

mal 180 Punkte erreichbar. Die C-Test-Texte wurden den Teilnehmern einzeln dargeboten. Ein Vor- oder Zurückblättern war nicht gestattet. Die Probanden bekamen für die ersten fünf Texte drei Minuten Bearbeitungszeit, für die letzten vier Texte vier Minuten pro Text. Zuvor bekamen die Teilnehmer einen Übungs-C-Test, um mit dem Format vertraut zu werden. Leider machen die Autoren keine Angaben dazu, wie es zu der Zeitbemessung von drei bzw. vier Minuten pro Text kam. Die Autoren fanden folgende Korrelationen zwischen dem eingesetzten C-Test und dem *Test de Connaissance du Français* (vgl. Tab. 11):

	TCF – Hörverstehen	TCF – mündlicher Ausdruck
C-Test ($p < 0{,}01$ (zweiseitig))	r = 0,665	r = 0,529

Tab. 11: Pearson-Korrelationen zwischen C-Test und den Subtests Hörverstehen und mündlicher Ausdruck des Test de Connaissance du Français (vgl. REICHERT et al. 2010: 220)

Diese Werte sind vergleichsweise hoch (vgl. Tab. 3.3). Unter den genannten Studien, die Zusammenhänge von C-Tests und Leistungen in Testteilen zum Hörverstehen und/oder zum mündlichen Ausdruck behandeln, ist die von REICHERT et al. (2010) durchgeführte Untersuchung besonders fundiert.

Der verwendete C-Test wurde zumindest in Teilen bereits in einem anderen Forschungsprojekt verwendet, die anderen Texte wurden zuvor erprobt. Es sind Angaben zur Darbietungsform und zur Zeitbemessung vorhanden. Der als Außenkriterium genutzte *Test de Connaissance du Français* ist ein standardisierter und skalierter Test. Die Probandenzahl ist erfreulich groß, und es wurden Korrelationen mit allen vier Fertigkeiten ermittelt. Jedoch muss auch einschränkend erwähnt werden, dass die Probanden den *Test de Connaissance du Français* bereits sieben Monate vor dem C-Test abgelegt haben, so dass Veränderungen des Sprachstands in beide

Richtungen möglich sind, weshalb die Ergebnisse mit Vorsicht interpretiert werden sollten.

Von besonderem Interesse für die vorliegende Studie ist auch die Untersuchung von **ECKES (2014)**. Er überprüft den Zusammenhang zwischen dem als Screening-Test entwickelten onDaF mit dem TestDaF. Im Rahmen der TestDaF-Erprobungen wird stets ein C-Test als Anker zur Ermittlung der Aufgabenschwierigkeit in das Testset integriert. Dieser C-Test besteht aus acht immer gleichen Texten aus der onDaF-Datenbank, die bereits erprobt und kalibriert wurden. Die Texte haben jeweils 20 Lücken und sind in aufsteigender Schwierigkeit angeordnet. Der TestDaF besteht aus vier Teilen: Leseverstehen, Hörverstehen, schriftlicher Ausdruck und mündlicher Ausdruck. Insgesamt 1467 Teilnehmer nahmen bei sechs verschiedenen Erprobungen an der Studie teil. Die Reliabilität des onDaF wurde für alle sechs Erprobungen separat ermittelt und rangiert zwischen $\alpha = 0{,}92$ und $\alpha = 0{,}95$. Die Reliabilität der Subtests Hörverstehen bewegt sich zwischen $\alpha = 0{,}79$ und $\alpha = 0{,}87$, und die Reliabilität des mündlichen Ausdrucks liegt zwischen 0,94 und 0,96 und ist damit bemerkenswert hoch. Tabelle 12 zeigt die ermittelten Korrelationen zwischen den Ergebnissen des TestDaF-Teils Hörverstehen mit dem C-Test bei den Erprobungsläufen E033 bis E038.

	E033	E034	E035	E036	E037	E038	gesamt
Hörverstehen	$n = 258$ r = 0,68	$n = 206$ r = 0,68	$n = 222$ r = 0,64	$n = 258$ r = 0,63	$n = 243$ r = 0,82	$n = 253$ r = 0,73	$n = 1440$ r = 0,69

$p < 0{,}01$

Tab. 12: Pearson-Korrelationen zwischen dem Hörverstehen des TestDaF und dem onDaF (vgl. ECKES 2014: 152)

Während alle errechneten Korrelationen recht hoch sind und nah beieinander liegen, ist insbesondere der Wert von r = 0,82 bei der Erprobung

E037 hervorzuheben. Ein höherer Wert für eine Korrelation zwischen dem C-Test und einer sprachlichen Teilfertigkeit findet sich in der Literatur nicht. Für den mündlichen Ausdruck hat ECKES (2014) die Korrelationen nicht mit dem von den Testteilnehmern erreichten Punktwert errechnet, sondern über die Stufenzuordnung zum Gemeinsamen Europäischen Referenzrahmen (EUROPARAT 2001) und die TestDaF-Niveaustufen TDN3 bis TDN5, welche sich am Referenzrahmen orientieren. Wenngleich Eckes auch hier Pearson-Korrelationen berechnet, sollen hier aufgrund der vorliegenden ordinalskalierten Daten lediglich die Korrelationen mit Kendalls Tau b betrachtet werden. Diese Daten liegen ebenfalls für das Hörverstehen vor (vgl. Tab. 13).

	E033	E034	E035	E036	E037	E038	gesamt
Hör-verstehen	n = 258 τ = 0,55	n = 206 τ = 0,54	n = 222 τ = 0,49	n = 258 τ = 0,47	n = 243 τ = 0,61	n = 253 τ = 0,58	n = 1440 τ = 0,56
mündlicher Ausdruck	n = 222 τ = 0,64	n = 179 τ = 0,47	n = 118 τ = 0,53	n = 207 τ = 0,50	n = 128 τ = 0,57	n = 140 τ = 0,57	n = 994 τ = 0,55

$p < 0{,}01$

Tab. 13: Kendalls Tau-Korrelationen zwischen der Stufenzuordnung des Hörverstehens und mündlichem Ausdruck des TestDaF und dem onDaF (vgl. ECKES 2014: 152 & 154)

Wie aus Tabelle 13 hervorgeht, schwanken die Korrelationen für den mündlichen Ausdruck stärker als diejenigen für das Hörverstehen. Dies mag unter anderem darin begründet liegen, dass auch die Teilnehmerzahlen bei den einzelnen Erprobungen beim mündlichen Ausdruck stärker voneinander abweichen. Nichtsdestotrotz sind die hier errechneten Korrelationen zwischen τ = 0,47 und τ = 0,64 im mittleren Bereich anzusiedeln.

Wie gezeigt wurde, existiert eine ganze Reihe von Studien, die im Rahmen unterschiedlicher Zielsetzungen Korrelationen zwischen einem C-Test und einem Mehrkomponentenkriterium ermitteln. Für die zum

Teil sehr unterschiedlichen Ergebnisse lassen sich mehrere mögliche Ursachen anführen:

- Der Stichprobenumfang schwankt innerhalb der Studien enorm, nämlich zwischen $n = 32$ (vgl. DALLER & PHELAN 2006) und $n = 412$ (vgl. JAKSCHIK 1996) bzw. $n = 1440$ (vgl. ECKES 2014). Kleine Probandengruppen sind jedoch nicht unproblematisch, da sie zu Stichprobenfehlern führen können (vgl. BÜHNER 2011: 169).
- Die eingesetzten C-Tests unterscheiden sich hinsichtlich der Textquellen, der Anzahl der Texte und der Anzahl der Lücken pro Text. Teilweise waren die C-Tests kalibriert (vgl. ECKES 2014) oder zumindest vorerprobt (vgl. JAKSCHIK 1996). Andere Autoren machen keine Angaben darüber oder schreiben, dass keine Pilotierung stattgefunden habe (vgl. DÖRNYEI & KATONA 1992).
- Die gewählten Außenkriterien unterscheiden sich ebenfalls stark. Es finden sich sowohl Studien, die als Kriterium einen etablierten und standardisierten Sprachtest heranziehen (vgl. REICHERT et al. 2010; ECKES 2010; ECKES 2014) als auch solche, über deren Außenkriterien nur wenig bekannt ist, da es sich beispielsweise um universitätsinterne Prüfungen handelt (vgl. CHAPELLE & ABRAHAM 1990).
- Die Testformate beim Sprechen bzw. mündlichen Ausdruck sind äußerst verschieden. Das beschriebene Spektrum reicht von wenig standardisierten Interviews (vgl. DÖRNYEI & KATONA 1992) über Zertifikatsprüfungen mit standardisierten Aufgaben im Prüfer-Prüfling-Gespräch (vgl. REICHERT et al. 2010) bis hin zum sogenannten computeradministrierten SOPI (vgl. ECKES 2010; ECKES 2014). Auch der Faktor des monologischen Sprechens gegenüber dialogischem Sprechen spielt hier vermutlich eine Rolle.

- Auch die Reliabilitäten der verwendeten Testinstrumente variiert von Studie zu Studie. Einige Autoren ermitteln die Reliabilität ihres C-Tests mittels Kuder-Richardson (vgl. CHAPELLE & ABRAHAM 1990; CHIHARA 1996; JAFARPUR 2002), was jedoch aufgrund der Abhängigkeit der Lücken voneinander eigentlich nicht zulässig ist (vgl. KLEIN-BRALEY & RAATZ 1984: 135). Zudem liefern mehrere Studien gar keine Informationen über die Messgenauigkeit der eingesetzten Tests, insbesondere fehlt diese Angabe häufig bei den gewählten Außenkriterien. Da die Reliabilität eine notwendige, aber nicht hinreichende Voraussetzung für die Validität ist (vgl. BÜHNER 2006: 42 f.), ist eine akzeptable Reliabilität eine Grundvoraussetzung zur adäquaten Interpretation der Ergebnisse.
- Die Verschiedenheit der Messinstrumente bringt Unterschiede in der Auswertung bzw. Bewertung der Aufgaben mit sich, was in einer unterschiedlichen Skalierung der Daten resultiert. Dies wiederum macht die Verwendung entsprechender Korrelationskoeffizienten notwendig: Pearson (r) für mindestens metrisch skalierte Daten, Spearman Rho (ρ) oder Kendalls Tau (τ) für ordinalskalierte Daten. Der Pearson-Korrelationskoeffizient fällt generell höher aus als Spearman Rho und letzterer wiederum höher als Kendalls Tau (vgl. HILL & LEWICKI 2006: 37). Eine Angabe über den herangezogenen Korrelationskoeffizienten ist folglich von großer Bedeutung, jedoch nicht in allen beschriebenen Studien gegeben.

Für die hier durchgeführte Studie ist darüber hinaus von Bedeutung, dass in den meisten aufgeführten Vorgängerstudien keinerlei Angaben zur Zeitbemessung des C-Tests gemacht wurde, meist auch nicht zur Darbietungsform. In einigen Studien weisen die eingesetzten C-Tests Abweichungen von der Norm auf, beispielsweise die Erstellung mehrerer

C-Test-Texte auf der gleichen Textgrundlage (vgl. CHAPELLE & ABRAHAM 1990), oder die Verwendung nicht authentischer Texte (vgl. JAFARPUR 2002). Aus den genannten Gründen ist eine Vergleichbarkeit zwischen der hier durchgeführten Untersuchung mit der aufgearbeiteten Literatur nur sehr bedingt möglich.

In der folgenden Übersicht werden alle hier diskutierten Korrelationsstudien zum C-Test tabellarisch zusammengefasst (vgl. Tab. 14). Diese tabellarische Aufstellung der Studien verdeutlicht noch einmal die ungleiche Datenlage was die Anzahl der Studien anbelangt bzgl. Hörverstehen bzw. Sprechen.

Tab. 14: Übersicht über Korrelationsstudien zum C-Test

Studie	*n*	C-Test	Kriterium	Korrelationen	Bemerkungen
CHAPELLE & ABRAHAM (1990)	*n* = 201 (in 4 Gruppen)	5 Texte à 15 Lücken. Alle Texte aus gleicher Textgrundlage erstellt. KR-20 = 0,81 25 min Bearbeitungszeit	*Iowa State University English Placement Test*	Hörverstehen 0,472	C-Test bestand nur aus einem langen Text
DÖRNYEI & KATONA (1992)	*n* = 102	4 Texte à 17-24 Lücken α = 0,75	Institutseigener Test; TOEIC; *oral interview*	Hörverstehen (Institutstest) 0,33 Hörverstehen (TOEIC) 0,51 Sprechen 0,43	*oral interview* nicht standardisiert; niedrige Reliabilität des C-Tests, keine Pilotierung
CHIHARA (1996)	*n* = 220 (Form A, *n* = 82) (Form B, *n* = 93)	4 Texte à 25 Lücken (Form A & Form B) KR-20= 0,748 bis 0,810	TOEFL	Hörverstehen r = 0,417 (Form A) Hörverstehen r = 0,569 (Form B)	
HUHTA (1996)	*n* = 129 (insgesamt) *n* = 47 (in dieser Berechnung)	5 Texte à 20 Lücken plus Eisbrecher	Eingangstest vom *English Department*	Hörverstehen (*exact*) 0,19 Hörverstehen (*lenient*) 0,16 Hörverstehen (*more lenient*) 0,18	Reliabilität des C-Tests unbekannt; relativ kleines *n*

Studie	*n*	C-Test	Kriterium	Korrelationen	Bemerkungen
JAKSCHIK (1996)	*n* = 678 (Kaufmann, *n* = 412) (Metall, *n* = 142) (Elektro, *n* = 124)	6 Texte à 20 Lücken Bearbeitungszeit unlimitiert	Lehrereinschätzungen auf 5-Punkt-Skala	Sprechen (Kaufm.) ρ = −0,38 Sprechen (Metall) ρ = −0,48 Sprechen (Elektro) ρ = −0,54	Reliabilität des C-Tests unbekannt Kriterium subjektiv
ARRAS, ECKES & GROTJAHN (2002)	*n* = 187 (Hörverstehen, *n* = 187) (Sprechen, *n* = 145)	4 Texte à 20 Lücken α = 0,842 5 min Bearbeitungszeit pro Text	TestDaF	Hörverstehen ρ = 0,636; τ = 0,483 Sprechen ρ = 0,640; τ = 0,521	
JAFARPUR (2002)	*n* = 146	4 Texte à 20–29 Lücken KR-21 = 0,92	*English Placement Test* (CORRIGAN ET AL. 1978)	Hörverstehen r = 0,65	geringe Reliabilität des Hörverstehenstests (KR-21 = 0,6)
KREKELER (2002)	*n* = 67	80 Lücken (vermutlich 4 Texte à 20 Lücken)	TestDaF; DSH	Hörverstehen (DSH) ρ = 0,437 Hörverstehen (TestDaF) ρ = 0,439 Sprechen (TestDaF) ρ = 0,415	

Studie	*n*	C-Test	Kriterium	Korrelationen	Bemerkungen
Daller & Phelan (2006)	*n* = 32	6 Texte à 20 Lücken α = 0,836	TOEIC	Hörverstehen (*enter*) r = 0,455 Hörverstehen (*exit*) r = 0,415	kleines *n*; keine Pilotierung des C-Tests
Lei (2008)	*n* = 265	5 Texte à 20 Lücken α = 0,747	Test aus *College Englisch Book I*	Hörverstehen 0,318*	Niedrige Reliabilität des C-Tests
Reichert, Keller & Martin (2010)	*n* = 228	9 Texte à 20 Lücken (z. T. aus TESTATT-Studie (Raatz et al. 2006) α = 0,87–0,93 3–4 min Bearbeitungszeit pro Text	*Test de Connaissance du Français*	Hörverstehen, r = 0,665 Sprechen, r = 0,529	6 Monate zwischen zwei Testsitzungen
Eckes (2014)	*n* = 1467 (6 Erhebungen) E033, *n* = 258 E034, *n* = 206 E035, *n* = 222 E036, *n* = 258 E037, *n* = 243 E038, *n* = 253)	onDaF (8 Texte à 20 Lücken) α = 0,92–0,95	TestDaF	Hörverstehen, r = 0,69 (gesamt) E033, r = 0,68 E034, r = 0,68 E035, r = 0,64 E036, r = 0,63 E037, r = 0,82 E038, r = 0,73	

Sofern der Korrelationskoeffizient hier nicht angegeben ist, wird er in den Quelltexten nicht berichtet.

3.3 Studien zum C-Test mit Geschwindigkeitskomponente

Neben den bereits genannten im Rahmen dieser Studie relevanten Korrelationsstudien existieren ebenfalls einige Untersuchungen, die mit einem (stark) zeitlimitierten C-Test arbeiten. Während die oben aufgeführten Korrelationsstudien ähnliche Ziele und eine grundlegend ähnliche Methodik aufweisen, unterscheiden sich die nachfolgend angeführten Studien mit Verwendung eines C-Tests unter Zeitdruck erheblich voneinander. Dies zeigt sich insbesondere in den unterschiedlichen Zwecken, zu denen der C-Test mit einer *speed*-Komponente versehen wird.

Zu den Studien, die einen C-Test unter *speed*-Bedingungen einsetzen, zählt jene von **Raatz (2002)**. Er untersucht den Zusammenhang von allgemeiner Sprachkompetenz und Intelligenz. Die Probanden waren zwei Gruppen von muttersprachlichen Studenten (n = 113, 46 Männer, 67 Frauen, Durchschnittsalter 24,3 Jahre)[20] und (n = 61, 22 Männer, 39 Frauen, Durchschnittsalter 23,3 Jahre) sowie eine Gruppe von ebenfalls muttersprachlichen Schülern (n = 112, 50 Jungen, 62 Mädchen, Durchschnittsalter 11,3). Die Schülergruppe wurde in die Studie integriert, um etwaige alters- bzw. entwicklungsbedingte Unterschiede im Zusammenhang zwischen den oben genannten Faktoren zu ermitteln. Die hier eingesetzten parallelen C-Tests (Form A für die erste Studentengruppe, Form B für die zweite Studentengruppe sowie die Schülergruppe) stellen die Operationalisierung der allgemeinen Sprachkompetenz dar. Sie weichen insofern von sonst üblichen C-Tests ab, als dass es sich jeweils um einen einzigen Text mit 100 Lücken handelt. Der C-Test wird sowohl als S-C-Test als auch als reiner

20 Diese Probandengruppe lag aus einer unpublizierten Studie von van den Bruck vor (vgl. Raatz 2002: 170).

Niveautest eingesetzt. Die Probanden bearbeiten den Text zehn Minuten lang (Niveau-Bedingung), wechseln jedoch nach fünf Minuten den Stift (*speed*-Bedingung), um zwischen der Testleistung unter diesen verschiedenen Bedingungen differenzieren zu können. Warum genau fünf Minuten bzw. zehn Minuten Bearbeitungszeit gewählt wurden, geht aus der Studie nicht hervor. Der C-Test dient hier gemeinsam mit der letzten Deutschnote als Operationalisierung der Deutschkompetenz (LI). Die mentale Verarbeitungsgeschwindigkeit wurde mit zwei Tests erhoben, die beide auf Zahlen basieren. Im „Test d2" (vgl. BRICKENKAMP 1981) sollen die Testteilnehmer den Buchstaben „d" so oft er vorkommt markieren. Dies teste auch die Konzentrationsfähigkeit. Im ZVT-Test (vgl. OSWALD & ROTH 1987) müssen die Testteilnehmer die Zahlen 1 bis 90 in aufsteigender Reihenfolge miteinander verbinden. Intelligenz wurde über mehrere Subtests der Testbatterie von HORN (1983) erhoben. Hierbei lag der Fokus auf verbaler Intelligenz.[21] Um herauszufinden, ob die Schreibgeschwindigkeit (*mechanical writing speed*) einen Einfluss auf die Performanz bei den Geschwindigkeitstests hat, wurde auch diese mit einem eigens erstellten Test erhoben. Hierbei müssen die Probanden in einer Minute zwei vorgegebene Buchstaben so oft wie möglich schreiben (vgl. RAATZ 2002: 185). RAATZ (2002: 172 & 176) berichtet, dass kein Einfluss der *speed*-Komponente auf die Testperformanz festgestellt werden konnte, auch nicht bei der *speed*-Version des C-Tests ($r = 0{,}05$; $r = 0{,}22$; $r = 0{,}04$). Die Ergebnisse zeigen weiterhin, dass der C-Test bei keiner Probandengruppe Ladungen auf den Faktor mentale Verarbeitungsgeschwindigkeit aufweist (vgl. RAATZ 2002: 175 f.). Jedoch besteht ein Zusammenhang mit dem Konstrukt der verbalen Intelligenz, denn diese war der beste Prädiktor für die Leistung im C-Test. Während die Daten der erwachsenen Probanden sowohl beim C-Test als auch beim

21 Eine Übersicht der relevanten Konstrukte und derer Operationalisierungen findet sich in RAATZ (2002: 185).

S-C-Test nur einen Zusammenhang zwischen Testleistung und verbaler Intelligenz aufwiesen, zeigte sich bei den Kindern ein weiterer relevanter Faktor beim C-Test, nämlich *Reasoning*. Beim S-C-Test bildet das *Reasoning* bei den Kindern sogar den einzigen Faktor, die verbale Intelligenz spielt dann augenscheinlich keine Rolle mehr. Raatz führt diese Unterschiede zwischen den Kindern und den erwachsenen Probanden darauf zurück, dass die metasprachlichen Prozesse bei Kindern noch nicht ausgereift seien und sie somit beim Lösen des C-Tests auf andere Fähigkeiten und Strategien zurückgriffen (vgl. RAATZ 2002: 182). Er schlussfolgert, „it could be problematical to set a time limit when using C-Tests for children in L1 since this may change the validities" (RAATZ 2002: 177).

Der Artikel von **AGUADO, GROTJAHN & SCHLAK (2007)** leistet Vorarbeiten für eine spätere Untersuchung über den Zusammenhang von Sprachlernerfolg und Erwerbsalter. Die Autoren stellen die Hypothese auf, dass deklaratives Sprachwissen allein nicht genügt, um Sprache unter Echtzeitbedingungen erfolgreich zu verwenden, sondern dass hierzu auch prozeduralisiertes Wissen erforderlich sei. Aus diesem Grund entwickeln sie einen C-Test mit einer textspezifischen, starken *speed*-Komponente, der neben deklarativem auch prozedurales Wissen erfassen und somit im obersten Kompetenzbereich zwischen sehr erfolgreichen Lernern des Deutschen und deutschen Muttersprachlern differenzieren soll. Die *speed*-Komponente dient hierbei dazu, einen Deckeneffekt zu vermeiden, da bei C-Tests normalerweise von einer Lösungsrate von mindestens 90 % bei Muttersprachlern ausgegangen wird (vgl. AGUADO et al. 2007: 143 f.). AGUADO et al. (2007) legten 11 teilweise vorerprobte C-Tests einer Gruppe von 20 Studenten vor, die sich aus muttersprachlichen (n = 10) und fortgeschrittenen Lernern des Deutschen (n = 8) sowie bilingualen Sprechern (n = 2) zusammensetzte. Die Probanden bekamen pro C-Test-Text eine Bearbeitungszeit von drei Minuten. Die minimalen Bearbeitungszeiten lagen je nach

Text zwischen 1:25 min und 2:35 min (vgl. AGUADO et al. 2007: 146). Welche Lösungsrate die in dieser Zeit bearbeiteten C-Test-Texte aufwiesen, geben die Autoren nicht an. Jedoch wurden neben der Originallösung auch korrekte Alternativlösungen akzeptiert. Die Autoren stellten fest, dass sich die Zeitlimitierung positiv auf die Differenzierungsfähigkeit der C-Tests auswirkt. Das Test-Set erreichte mit $\alpha = 0{,}97$ eine ausgesprochen hohe Reliabilität.

Eine Fortsetzung der Studie von AGUADO et al. (2007) findet sich in **GROTJAHN, SCHLAK & AGUADO (2010)**. Die Autoren zielen darauf ab, über das Hinzufügen einer *speed*-Komponente die Differenzierungsfähigkeit des C-Tests bei hoch kompetenten Deutschlernern zu erhöhen. Auf Basis der ersten Untersuchungsergebnisse wurden hier 33 Deutschlernern und 9 Muttersprachlern des Deutschen, die alle DaF-Studenten waren, sieben Texte in aufsteigender Schwierigkeit präsentiert, denen noch ein Eisbrechertext voran ging. Den Teilnehmern wurde vor der Testdurchführung angekündigt, dass sie pro C-Test-Text zwischen einer und zwei Minuten Zeit haben würden. Tatsächlich forderten die Testleiter die Probanden jedoch dann zum Weiterblättern zum nächsten Text auf, nachdem der dritte Proband angezeigt hatte, dass er mit der Bearbeitung fertig sei. So ergaben sich für jeden Text spezifische Bestzeiten, die zwischen 57 Sekunden und 110 Sekunden lagen. Über die Bearbeitungszeit der Person, die als drittes einen C-Test-Text abschließt, wird hier für die eingesetzten C-Test-Texte eine Bearbeitungszeit zwischen 1:05 und 1:55 Minuten empirisch ermittelt. Während sich die arithmetischen Mittelwerte der C-Test-Texte bei maximal 25 erreichbaren Punkten pro Text bei den Muttersprachlern zwischen $\bar{x} = 20{,}22$ und $\bar{x} = 23{,}89$ bewegen, liegen diese bei den fortgeschrittenen Deutschlernen zwischen $\bar{x} = 13{,}64$ und $\bar{x} = 15{,}27$ (vgl. GROTJAHN et al. 2010: 308). Diese Ergebnisse deuten darauf hin, dass das Ziel einer besseren Differenzierbarkeit im oberen Kompetenzbereich durch

den S-C-Test erreicht wurde. Eine Reliabilitätsanalyse unter Einbezug aller muttersprachlichen sowie nicht-muttersprachlichen Probanden ergab einen Wert für Cronbachs Alpha von $\alpha = 0{,}97$. Dieser Wert ist somit beachtlich hoch, insbesondere vor dem Hintergrund der geringen Probandenzahl.

Basierend auf den Arbeiten von AGUADO et al. (2007) und GROTJAHN et al. (2010) nutzt auch **HEINE (2017)** einen S-C-Test. Sie beschäftigt sich mit der Frage, durch welche Eigenschaften und Faktoren sich sehr erfolgreiche Späterwerber des Deutschen auszeichnen. Hierbei ist der S-C-Test nur eins von vielen Erhebungsinstrumenten. Sein Einsatz dient dazu herauszufinden, ob es Späterwerber gibt, die innerhalb des Spektrums der Ergebnisse bilingualer Muttersprachler abschneiden. Die angesetzte textspezifische Bearbeitungszeit mit Zeiten zwischen 1:05 min und 1:55 min wurde von GROTJAHN et al. (2010) übernommen. Heine fand heraus, dass 14,6 % der Späterwerber beim S-C-Test im Bereich bilingualer Muttersprachler abschneiden (vgl. HEINE 2017: 164). Die monolingualen Muttersprachler zeigen beim C-Test verhältnismäßig homogene Ergebnisse, während ihre Testergebnisse beim S-C-Test deutlich weiter streuen (vgl. Tab. 15).

	min	max (von 150)	$\tilde{x}$	$\bar{x}$	SD
C-Test	130	150	144	143,5316	5,07845
S-C-Test	79	149	129	126,2875	13,27336

Tab. 15: Deskriptive Statistiken zu C-Test und S-C-Test bei Muttersprachlern (vgl. HEINE 2017: 130)

Auch diese Ergebnisse zeigen, dass die Geschwindigkeitskomponente des S-C-Tests für eine höhere Differenzierfähigkeit im oberen Kompetenzbereich sorgt. Derselbe Effekt tritt in der Gruppe der bilingualen Muttersprachler auf. Bei den Späterwerbern unterscheiden sich zwar die Maße der Lage bei C-Test und S-C-Test deutlich voneinander, die Streuung der Testergebnisse ist jedoch deutlich geringer (vgl. Tab. 16).

	min	max	$\tilde{x}$	$\bar{x}$	SD
C-Test	27	140	83	84,1124	24,23860
S-C-Test	15	112	50	50,7865	22,00024

Tab. 16: Deskriptive Statistiken zu C-Test und S-C-Test bei Späterwerbern (vgl. HEINE 2017: 130)

In der Dissertationsschrift von **WOCKENFUSS (2009)** wird ein zeitbeschränkter C-Test bei muttersprachlichen Probanden im Schulalter eingesetzt, um den Verlauf ihrer sprachlichen Entwicklung zu erfassen. Die Autorin prüft die Hypothese, dass ein Zusammenhang zwischen dem Alter der Testteilnehmer und ihrer Leistung im C-Test besteht. Der hier eingesetzte C-Test stammt aus einer Vorarbeit von GUMMICH (1997) und besteht aus zwei parallelen C-Tests mit jeweils fünf Texten à 20 Lücken. Die Textgrundlage bildeten Schulbücher unterschiedlicher Klassenstufen. In seiner ursprünglichen Fassung stand den Testteilnehmern für das gesamte Testset 30 Minuten Bearbeitungszeit zur Verfügung. Um einen Deckeneffekt zu vermeiden, wird in WOCKENFUSS (2009) die Bearbeitungszeit drastisch reduziert. Um eine geeignete Zeitbemessung zu finden, wurde das Testset 137 studentischen Probanden mit einer Minute Bearbeitungszeit pro Text vorgelegt. Die Probanden erreichten Mittelwerte von $\tilde{x} = 64{,}3$ (Studenten der Psychologie) bzw. $\tilde{x} = 64{,}9$ (Studenten der Soziologie), weshalb die Autorin davon ausgeht, dass die erprobte Bearbeitungszeit von einer Minute für die jüngere Probandengruppe (Durchschnittsalter 13,6 bzw. 13,7 Jahre) zu kurz sei. Daher wurde die Zeitbemessung auf eine Minute und 15 Sekunden pro C-Test-Text erhöht, ohne ein weiteres Mal erprobt zu werden. Die Reliabilitätsanalyse liefert für die beiden parallelen C-Test-Versionen für Cronbachs Alpha Werte von $\alpha = 0{,}93$ und $\alpha = 0{,}91$. Die Autorin konstatiert unter anderem, dass der C-Test aufgrund seiner Korrelationen mit verbalen Intelligenztests ein Messinstrument globaler Sprachfähigkeit sei.

Berger & Zimmermann (in Vorbereitung) untersuchen, ob der S-C-Test ein für Links- und Rechtshänder gleichermaßen faires Testformat ist. Durch die mechanische Abdeckung des unmittelbaren Kontexts beim Schreiben mit der rechten Hand könnten Rechtshänder gegenüber Linkshändern benachteiligt sein. Um dieser Frage nachzugehen, vergleichen Berger und Zimmermann muttersprachliche Links- und Rechtshänder miteinander. Den Probanden wird eine Kurzversion des C-Tests aus Aguado et al. (2007) vorgelegt. Dieser wurde zunächst mit den von Aguado et al. (2007) empirisch festgelegten textspezifischen Zeitbegrenzungen administriert und direkt im Anschluss noch einmal ohne jede Zeitbegrenzung. Zusätzlich füllen die Probanden einen Fragebogen aus, der für die Fragestellung relevante Informationen über den Gebrauch der beiden Hände, die Lese- und Schreibgewohnheiten der Teilnehmer Auskunft gibt.

Bemerkenswert an den angeführten Studien ist, dass sich alle mit der Zielsprache Deutsch befassen. Die Probanden sind ausnahmslos Muttersprachler oder sehr weit fortgeschrittene Lerner des Deutschen.

Die Geschwindigkeitskomponente wird dem C-Test meist hinzugefügt, um mit geringem Aufwand seinen Schwierigkeitsgrad zu erhöhen. Die behandelten Fragestellungen sind unterschiedlich. Mal ist die Verwendung der *speed*-Komponente ein reines Mittel zum Zweck (vgl. Wockenfuss 2009), mal wird sie bewusst genutzt, um das C-Test-Format für einen spezifischen Kontext zu optimieren (vgl. Grotjahn et al. 2010).

Einzig die Untersuchung von **Bisping (2006)** stellt hier eine bemerkenswerte Ausnahme dar. Er setzt zwei deutschsprachige C-Tests mit fünf Texten à 25 Lücken ein. Für eines der beiden Test-Sets bekamen die Probanden, 56 slowakische Studenten, fünf Minuten Bearbeitungszeit pro Text. Dieses lag als Papier-und-Bleistift-Version vor. Beim zweiten Test-Set, welches am Computer zu bearbeiten war, betrug die Bearbeitungszeit pro Text hingegen nur drei Minuten. Das Ziel der Studie war es, zu ermit-

teln, ob sich die Reliabilität und Validität der am Computer gelösten C-Tests ändert, wenn man die Zeitbemessung reduziert. Der Autor verweist auf STERNBERG (1999) und AGUADO et al. (2005), die durch eine Geschwindigkeitskomponente sowohl die Reliabilität als auch die Schwierigkeit des C-Tests erhöht sehen. Ein Vergleich des arithmetischen Mittels zeigt, dass der zeitreduzierte Computer-C-Test ($\bar{x}$ = 83,4) schwieriger war als der gewöhnliche Papier-und-Bleistift-C-Test ($\bar{x}$ = 98,6), was die Hypothese von STERNBERG (1999) und AGUADO et al. (2005) stützt. Der zeitreduzierte computerbasierte C-Test erreicht zudem eine sehr zufriedenstellende Reliabilität von α = 0,9, die Reliabilität des herkömmlichen C-Tests liegt im Gegensatz dazu bei α = 0,92. Vor dem Hintergrund der Teilnehmerzahl (n = 56) ist dieser Unterschied als äußerst gering zu bewerten.

BISPING (2006) setzt zusätzlich einen Konzentrationstest ein, der den Probanden sowohl als Papier-und-Bleistift-Version als auch als Computer-Test vorgelegt wurde. Hier ging es darum, in kurzer Zeit möglichst viele Rechenaufgaben im Kopf zu lösen. Aufgrund des geringen Aufgabenniveaus handelt es sich bei diesem Konzentrationstest um einen reinen Geschwindigkeitstest (vgl. Kapitel 2.2.5.1). Beim Korrelieren der vier Testteile gegeneinander fand BISPING (2006: 159) deutlich höhere Pearson-Korrelationen zwischen den beiden C-Test-Varianten (r = 0,79) als zwischen dem Computer-C-Test und dem Computer-Konzentrationstest (r = 0,54). Die konvergente Validität des S-C-Tests scheint hier folglich gegeben zu sein. BISPING (2006: 162) schließt mit dem Urteil, dass der C-Test sehr robust sei und sowohl computerbasiert als auch zeitreduziert eingesetzt werden könne. Es bleibt jedoch diskutierbar, ob eine durchschnittliche Bearbeitungszeit von 7,2 Sekunden pro Lücke wenig genug ist, um hier von einem S-C-Test zu sprechen (vgl. Kapitel 3.2).

Die jüngste und zugleich in diesem Kontext beachtenswerteste Untersuchung ist die von **Fadaeipour und Zohoorian (2017)**.[22] Die Autorinnen verweisen darauf, dass C-Tests oftmals einen Deckeneffekt aufweisen, weil sie für Muttersprachler und weit fortgeschrittene Lerner zu einfach seien und es ihnen so an Trennschärfe in diesen Probandengruppen fehle (vgl. Fadaeipour & Zohoorian 2017: 42). Der S-C-Test wird als ein möglicher Lösungsansatz für diese Schwierigkeit präsentiert und hinsichtlich seiner Eignung untersucht. Hierzu wurde 100 iranischen Studenten unterschiedlichen Fremdsprachenniveaus ein C-Test, ein S-C-Test und ein Test zum Leseverstehen in englischer Sprache vorgelegt. Der C-Test bestand aus zwei Texten mit jeweils 25 Lücken. Die Bearbeitungszeit betrug insgesamt zehn Minuten. Der S-C-Test bestand ebenfalls aus zwei Texten mit jeweils 25 Lücken. Die Bearbeitungszeit lag hier jedoch bei lediglich fünf Minuten für beide Texte. Die Basis für die C-Tests bildeten Texte eines Lehrwerks für Englisch als Fremdsprache. Die Reliabilität des gesamten C-Test-Sets lag bei $\alpha = 0{,}7$. Die Werte für den herkömmlichen C-Test ($\alpha = 0{,}59$) und den S-C-Test ($\alpha = 0{,}41$) für sich genommen liegen jedoch darunter und sind als nicht zufriedenstellend einzuordnen (vgl. Bühner 2006: 140). Allerdings muss bedacht werden, dass aufgrund der geringeren Testlänge eine Abnahme der Reliabilität gegenüber dem gesamten Testset durchaus zu erwarten ist. Zum Testen des Leseverstehens wurde der entsprechende Subtest des *Pearson Test of General English* eingesetzt: vier Kurztexte mit insgesamt 20 Multiple-Choice-Items und offenen Antworten.

Die deskriptive Analyse zeigte, dass die Streuungsmaße für den C-Test (min = 3; max = 43; r = 40; SD = 7,81) und den S-C-Test (min = 2; max = 41; r = 39; SD = 7,69) bemerkenswert nah beieinander liegen. Das arithmeti-

22 Aufgrund des jüngeren Erscheinungsdatums war diese Studie der Autorin zum Zeitpunkt der Planungen zu diesem Dissertationsprojekt nicht bekannt.

sche Mittel liegt für den S-C-Test mit $\bar{x}$ = 15,43 erwartungsgemäß unter dem Wert des C-Tests mit $\bar{x}$ = 20,14. Bei der Berechnung der Korrelationen zwischen den beiden C-Test-Varianten und dem Leseverstehenstest zeigte sich, dass der C-Test mit großzügiger Arbeitszeitbemessung einen höheren Wert für den Pearson-Produkt-Moment-Koeffizienten erreicht (vgl. Tab. 17).

	C-Test	S-C-Test
Leseverstehen	r= 0,65**	r= 0,55**

** signifikant auf dem Niveau 0,01.

Tab. 17: Korrelation von C-Test und S-C-Test mit Hörverstehen

Die Autorinnen schlussfolgern, dass der herkömmliche C-Test besser dazu geeignet sei, Leseverstehen zu testen. Weitergehende Regressions- und Faktoranalysen bekräftigen diesen Eindruck. Das hier vorliegende Ergebnis ist nicht überraschend insofern, als sowohl der C-Test unter Niveaubedingungen als auch der Leseverstehenstest den Testkandidaten genügend Zeit lassen, um die Aufgaben planvoll zu lösen. Umso interessanter ist die Frage, ob die hier gefundene Überlegenheit des C-Tests gegenüber dem S-C-Test auch Bestand hat, wenn Korrelationen mit den in Echtzeit ablaufenden Fertigkeiten Hörverstehen und Sprechen korreliert werden.

Das hier beschriebene Dissertationsprojekt betritt insofern Neuland, als erstmalig ein *speeded*-C-Test bei (Deutsch-)Lernern einer mittleren Kompetenzstufe eingesetzt wird und dessen Zusammenhang mit zwei sprachlichen Teilkompetenzen, Hörverstehen und Sprechen, untersucht wird, um die von GROTJAHN et al. (2010: 315) aufgestellte Hypothese zu überprüfen, dass sich über eine *speed*-Komponente die Korrelationen mit den in Echtzeit ablaufenden Teilkompetenzen Hörverstehen und Sprechen erhöhen lassen könnten. Dabei wird der S-C-Test nicht als ein Instrument einge-

setzt, das dazu dient, eine nicht direkt mit ihm verbundene Forschungsfrage zu beantworten, sondern er steht selbst im Fokus der Fragestellung.

Die sich anschließende Tabelle 18 liefert einen Überblick über die wenigen Studien, die sich mit dem S-C-Test befassen oder ihn zumindest einsetzen.

Tab. 18: Übersicht über Studien mit einem S-C-Test

Studie	*n*	C-Test	Zeitbemessung	Zielsetzung	Bemerkungen
RAATZ (2002)	*n* = 174 (Studenten) *n* = 112 (Kinder) L1 Deutsch	Form A und Form B, je ein Text à 100 Lücken	*non-speeded condition* 10 min Bearbeitungszeit, *speeded condition* 5 min Bearbeitungszeit	C-Test als Operationalisierung der L1-Kompetenz	C-Test bestand nur aus einem langen Text; Zeitbemessung scheinbar willkürlich
BISPING (2006)	*n* = 56 (Studenten) L1 Slowakisch	2 Sets mit 5 Texten à 25 Lücken	5 min (C-Test) 3 min (S-C-Test)	Ändern sich Reliabilität und Validität bei weniger Bearbeitungszeit am Computer?	C-Test und S-C-Test computeradministriert; beide Test-Sets nicht normiert
AGUADO, GROTJAHN & SCHLAK (2007)	*n* = 10 L1 Deutsch *n* = 8 L2 Deutsch *n* = 2 Bilinguale	11 teilweise vorerprobte Texte à 25 Lücken; $\alpha = 0{,}97$	3 min, Bestzeiten zwischen 1:25 min und 2:35 min	Vorarbeiten für GROTJAHN et al. (2010)	Lösungsquoten unklar
GROTJAHN, SCHLAK & AGUADO (2010)	*n* = 9 L1 Deutsch *n* = 33 L2 Deutsch	7 Texte à 25 Lücken plus Eisbrecher; $\alpha = 0{,}97$	Textspezifisch zwischen 1:05 und 1:55, empirisch ermittelt über drittbeste Lösungszeit	Prozedurales Wissen erfassen, um zwischen hochkompetenten Sprechern zu differenzieren	

Studie	*n*	C-Test	Zeitbemessung	Zielsetzung	Bemerkungen
HEINE (2017)	*n* = 219	6 Texte à 25 Lücken	Textspezifisch zwischen 1:05 und 1:55, empirisch ermittelt über drittbeste Lösungszeit	Erreichen Späterwerber das muttersprachliche Spektrum?	Texte aus AGUADO et al. (2007)
WOCKENFUSS (2009)	*n* = 531 (Set A) *n* = 524 (Set B) L1 Deutsch	2 parallele Sets mit 5 Texten à 20 Lücken	1:15 min pro Text	*speed*-Faktor um Deckeneffekt zu vermeiden	Texte aus GUMMICH (1997)
BERGER & ZIMMERMANN (in Vorbereitung)	angestrebt: *n* = 50 (Rechtshänder) *n* = 50 (Linkshänder) L1 Deutsch	4 Texte à 25 Lücken	Textspezifisch zwischen 1:05 und 1:25 min	Überprüfung ob S-C-Test faires Format	Texte und spezifische Zeitbemessung aus AGUADO et al. (2007)
FADAEIPOUR & ZOHOORIAN (2017)	*n* = 100 L2 Englisch (iranische Studenten)	C-Test & S-C-Test: je 2 Texte à 25 Lücken; Textbasis Englischlehrbuch	C-Test 10:00 min S-C-Test 5:00 min	Vergleich von C-Test und S-C-Test in Bezug auf Leseverstehen	Zeitbemessung bei S-C-Test relativ großzügig

4 Die Studie

4.1 Forschungslücke, Forschungsfragen und Relevanz der Studie

Wie aus den vorangegangenen Kapiteln hervorgeht, existiert eine Vielzahl von Studien, die den C-Test gegen verschiedene, zum Teil etablierte Außenkriterien und teilkompetenzspezifische Subtests (Wortschatz, Grammatik) korrelieren und somit die Konstruktvalidität des C-Tests belegen (vgl. Kapitel 2.2.4 und 3.2). Ebenso existiert eine Reihe von Untersuchungen, die einen mit einem Geschwindigkeitsfaktor versehenen C-Test einsetzen. Hierbei geht es meist darum, die Schwierigkeit eines bestehenden C-Test-Sets zu erhöhen (vgl. Kapitel 2.3). Die vorliegende Studie betritt daher in mehreren Punkten Neuland:

Zunächst steht der S-C-Test selbst im Mittelpunkt des Erkenntnisinteresses und ist nicht nur Mittel zum Zweck bei der Beantwortung anderer Forschungsfragen. So soll zunächst geklärt werden, ob der C-Test auch unter *speed*-Bedingungen über eine akzeptable Reliabilität verfügt. Da durch den Geschwindigkeitsfaktor weniger Gelegenheit bleibt, beim Sitznachbarn abzuschreiben oder *Test-Wiseness* anzuwenden, könnte sich dies positiv auf die Reliabilität auswirken. Wie ausgeführt wurde, rechnen auch GROTJAHN et al. (2010: 303) mit einem Anstieg der Reliabilität durch den Geschwindigkeitsfaktor (vgl. Kapitel 2.2.5.2). Daher lautet die **Forschungsfrage 1**: Verfügt der S-C-Test über eine akzeptable Reliabilität?

Die vorliegende Untersuchung stellt eine Korrelationsstudie unter Einsatz eines S-C-Tests dar. Sie hat das Ziel, einen Beitrag zu der zentralen Frage nach dem Einfluss des *speed*-Faktors auf das Konstrukt des C-Tests

zu leisten, und die in GROTJAHN et al. (2010: 315) aufgestellte Hypothese, dass sich die Korrelationen mit den in Echtzeit ablaufenden Teilfertigkeiten Hörverstehen und Sprechen durch das Hinzufügen einer Geschwindigkeitskomponente erhöhen könnten, zu überprüfen. Der Hypothese zugrunde liegt die Ansicht, dass sowohl beim S-C-Test als auch beim fremdsprachlichen Hören und Sprechen Sprachverwendung in Echtzeit stattfindet:

> Über den Aufbau von Zeitdruck sollte der zeitlichen Dimension mündlich-sprachlicher Rezeptions- und Produktionsprozesse Rechnung getragen werden und die Korrelation zu Hörverstehen und Sprechen erhöht werden. (GROTJAHN et al. 2010: 301–302)

Somit lautet die **Forschungsfrage 2a**: Besteht ein Zusammenhang zwischen den S-C-Test-Ergebnissen und den Fertigkeiten Hörverstehen und Sprechen?

Darüber hinaus bietet das Design der Untersuchung die Möglichkeit, die Ergebnisse der gleichen Probandengruppe beim S-C-Test mit denen eines äquivalenten C-Tests zu vergleichen. Es lässt sich auf diese Weise direkt vergleichen, welche Testvariante höher mit den oben genannten Fertigkeiten korreliert. Dementsprechend lautet die **Forschungsfrage 2b**: Besteht ein stärkerer Zusammenhang zwischen den Ergebnissen des S-C-Tests und den Fertigkeiten Hörverstehen und Sprechen als zwischen den Ergebnissen von C-Test und den Fertigkeiten Hörverstehen und Sprechen?

Durch die Wahl des Außenkriteriums, einer mündlichen Prüfung des Goethe-Zertifikats B2 (vgl. Kapitel 4.3.1.2), das zwischen monologischem und dialogischem Sprechen differenziert, kann zudem der Frage nachgegangen werden, mit welcher Art des mündlichen Sprachgebrauchs der S-C-Test höher korreliert. Da beim dialogischen Sprechen mehr Spontaneität gefordert ist, lässt dies die Vermutung zu, dass die Korrelation hier ver-

gleichsweise höher ausfällt. Daraus ergibt sich die **Forschungsfrage 3**: Besteht ein stärkerer Zusammenhang zwischen den Ergebnissen des S-C-Tests mit dialogischem Sprechen oder mit monologischem Sprechen?

Weil der S-C-Test in bisherigen Untersuchungen ausschließlich bei Muttersprachlern oder weit fortgeschrittenen L2-Lernern zum Einsatz kam, wird ebenfalls untersucht, ob sich der S-C-Test auch für niedrigere Niveaustufen eignet bzw. ob der S-C-Test für Lerner unterschiedlichen Sprachniveaus gleichermaßen geeignet ist. Dies führt zur **Forschungsfrage 4**: Unterscheidet sich die Stärke der Korrelationen des S-C-Tests mit den Fertigkeiten Hörverstehen und Sprechen bei unterschiedlich weit fortgeschrittenen Lernern?

Mündliche Sprachkompetenz besteht aus verschiedenen Teilaspekten. Einer davon ist Flüssigkeit. MICHEL (2017: 50) schreibt, „Fluency refers to the smooth, easy, and eloquent production of speech". SKEHAN (1996: 46) bezieht Flüssigkeit auf „the learners capacity to mobilize an interlanguage system to communicate meanings in real time". Dieser Aussage folgend kann angenommen werden, dass Flüssigkeit ein Bestandteil des Konstrukts des S-C-Tests ist, da dieser Sprachverwendung in Echtzeit simuliert. Im Rahmen dieser Studie soll daher anhand weniger Probanden explorativ untersucht werden, ob es lohnenswert erscheint, weitere Studien in diese Richtung zu konzipieren. So lautet die **Forschungsfrage 5:** Gibt es einen Zusammenhang zwischen dem Ergebnis beim S-C-Test und der Flüssigkeit mündlicher Sprachproduktion?

Wie in Kapitel 2.2.3.3 dargelegt, ist die relativ geringe Augenscheinvalidität des C-Tests eine Schwachstelle dieses Prüfungsformats. Es lässt sich vermuten, dass die Erweiterung des C-Tests um eine Geschwindigkeitskomponente das Vertrauen in das Testformat nicht unbedingt erhöht. Aus diesem Grund lautet die **Forschungsfrage 6**: Verändert die *speed*-Komponente die Augenscheinvalidität des C-Tests?

Nicht zuletzt kann auf der Grundlage eines Begleitfragebogens ermittelt werden, ob die *speed*-Komponente beim C-Test eine systematische Verzerrung hervorruft, beispielsweise insofern, als Testteilnehmer, die in einem anderen Schriftsystem alphabetisiert wurden, schlechter abschneiden als Teilnehmer, die von Kindesbeinen an mit der lateinischen Schrift vertraut sind. Zeigt sich eine systematische Verzerrung beim S-C-Test, sodass bestimmte Probandengruppen generell benachteiligt sind, so ließe sich schlussfolgern, dass das S-C-Test-Format keine Chancengleichheit gewährleistet. Die **Forschungsfrage 7** lautet: Hat das zuerst erlernte Schriftsystem einen Einfluss auf den Erfolg bei einem deutschsprachigen S-C-Test?

C-Tests werden für gewöhnlich mit einer sehr großzügig bemessenen Arbeitszeit versehen. Durch den Aufbau von Zeitdruck ist es möglich, dass die Testteilnehmer die Lücken des C-Test-Textes nicht mehr in derselben Art und Weise bearbeiten wie ohne Zeitdruck, beispielsweise weil die Zeit für ein Vor- und Zurückspringen im Text und damit verbundene Re-Analysen nicht mehr ausreicht. Daher lautet die letzte **Forschungsfrage 8**: Unterscheidet sich die Lösungsreihenfolge beim Lösen von C-Tests und S-C-Tests?

Sollte diese Studie zu dem Ergebnis kommen, dass der S-C-Test eine zufriedenstellende Reliabilität aufweist und dem C-Test in Bezug auf die Korrelationen mit den Fertigkeiten Hörverstehen und Sprechen überlegen ist, so bietet sich das S-C-Test-Format an, um damit Sprachkursinteressenten in einen adäquaten Sprachkurs einzustufen. Dies erscheint nicht nur deshalb sinnvoll, weil in einem kommunikativ orientierten Fremdsprachenunterricht die Kompetenzen Hörverstehen und Sprechen besonders wichtig sind, um dem Unterricht erfolgreich folgen zu können. Universitäre Sprachenzentren müssen zum Teil eine große Anzahl von Kursinteres-

senten einstufen.[23] Aufgrund der bislang üblichen Zeitbemessung von rund 5 Minuten pro C-Test-Text ist dieses Einstufungsverfahren jedoch recht zeitintensiv. Von einer Reduktion der Testzeit in derartigen *low stakes*-Situationen würden folglich nicht nur die Testzentren durch eine Schonung der Ressourcen, sondern auch die Teilnehmer selbst enorm profitieren.

4.2 Forschungsethische Aspekte

Unabdingbare Voraussetzung einer jeden Forschungsarbeit ist das Einhalten ethischer Grundsätze. Werden Daten unter ethisch nicht einwandfrei vertretbaren Bedingungen erhoben, so sollten sie nicht genutzt und publiziert werden. BACH und VIEBROCK (2012: 25) stellen jedoch fest: „In der Fremdsprachenforschung ist Forschungsethik ein noch unbestelltes Feld." So verfüge auch die Deutsche Gesellschaft für Fremdsprachenforschung (DGFF) über keinen Kriterienkatalog ethischen Forschens (vgl. VIEBROCK 2015: 23). BACH und VIEBROCK (2012: 18) weisen zudem darauf hin, dass im Gegensatz zum englischsprachigen Raum, wo Forschungsvorhaben meist durch eine universitätsinterne Ethikkommission freigegeben werden müssen, die Verantwortung für ethisches Handeln in der Forschung in Deutschland in den Händen der Forscher selbst liegt. Jedoch existiert inzwischen eine Ethikkommission der Deutschen Gesellschaft für Sprachwissenschaft (vgl. URL 22: DGS: Ethikkommission der Deutschen Gesellschaft für Sprachwissenschaft).

Die Herausforderung besteht folglich zunächst darin, die für die eigene Forschung relevanten Richtlinien zu identifizieren, da es keine einheitli-

23 Vgl. dazu beispielsweise SCHWINDELER (2013: 31 f.), die in ihrer unpublizierten Masterarbeit erläutert, dass im Rahmen ihrer Untersuchung 245 internationale Studierende an den Intensivkursen der ZEMS an der TU Berlin teilnahmen.

chen Kriterienkataloge gibt. So stellt die *International Language Testing Association* (URL 9: ILTA Code of Ethics) neun ethische Prinzipien zusammen, die es beim Testen von Sprache zu beachten gelte, darunter beispielsweise der stets respekt- und würdevolle Umgang mit den Testteilnehmern. Jedoch können die Prinzipien der *International Language Testing Association* für die vorliegende Arbeit nicht in dieser Form übernommen werden, da sie einen anderen intentionalen Fokus aufweisen und somit zu kurz greifen. STROHM KITCHENER und KITCHENER (2009: 13 ff.) nennen fünf Prinzipien ethischen Forschens für die empirische Sozialforschung: *non-maleficence, beneficence, respects for persons, fidelity, justice*. Hier zeigen sich einige Parallelen mit dem gemeinsamen Ethik-Kodex der Deutschen Gesellschaft für Soziologen (DGS) und des Berufsverbands Deutscher Soziologen (BDS) (URL 10: Ethik-Kodex der DGS und BDS). LEGUTKE & SCHRAMM (2016: 109 ff.) nennen eine Reihe von ethischen Richtlinien für die Fremdsprachenforschung. Hier wird beispielsweise auch die Gestaltung von Forschungsbeziehungen angesprochen, was in qualitativen Studien eine größere Rolle spielt als in der vorliegenden quantitativ ausgerichteten Untersuchung. Im Folgenden werden einige der wichtigsten forschungsethischen Aspekte herausgegriffen und es wird beschrieben, inwiefern ihnen in dieser Studie Rechnung getragen wird.

Die Forderung nach **Objektivität und Integrität** verpflichtet die Forscher dazu, gewonnene Daten nicht zu verfälschen, beispielsweise durch Vernichtung unerwünschter Teilergebnisse (vgl. VON UNGER 2014: 19).

Ein unbedingt zu beachtender Punkt ist die **Freiwilligkeit der Teilnahme** der Probanden an der Studie (vgl. LEGUTKE & SCHRAMM 2016: 111). Jedem Probanden muss es freigestellt werden, die Teilnahme zu jedem beliebigen Zeitpunkt (und ohne Angabe von Gründen und ohne daraus resultierende persönliche Nachteile) abbrechen zu können. Diesem Grundsatz wurde in der vorliegenden Studie gefolgt. Alle potentiellen Pro-

banden wurden vor der ersten Datenerhebung darauf hingewiesen, dass die Studienteilnahme auf freiwilliger Basis erfolgt und es jedem Kursteilnehmer frei stehe, zu gehen. Diese Möglichkeit wurde zu Studienbeginn vereinzelt von Deutschlernern genutzt. Im späteren Verlauf der drei Testtage umfassenden Studie brach jedoch eine größere Anzahl von Probanden die Teilnahme ab und erschien nicht zur Sprechaufgabe am dritten Testtag (vgl. Abb. 10, S. 186). Des Weiteren besteht eine Pflicht, die Probanden vor der Teilnahme ausreichend darüber zu informieren, was sie im Verlauf der Studie erwartet, und zu welchem Zweck die bei ihnen erhobenen Daten verwendet werden. Das sogenannte **Prinzip der informierten Einwilligung** (vgl. URL 10: Ethik-Kodex der DGS und BDS) wurde durch ein den potentiellen Probanden zu Beginn der Studie ausgehändigtes Informationsblatt befolgt (vgl. Anhang E). Die Zustimmung wurde mündlich von den Probanden eingeholt. Dieses Vorgehen wurde gewählt, da den Studienteilnehmern absolute **Anonymität** zugesichert wurde (vgl. VON UNGER 2014: 10), die durch eine zu unterschreibende Einverständniserklärung gefährdet gewesen wäre. Die Namen der Probanden waren und sind nicht bekannt, stattdessen wurde mit einem durch die Probanden eigenhändig erstellten Teilnehmer-Code gearbeitet, um die verschiedenen Testteile mit ein und derselben Person identifizieren zu können (vgl. Kapitel 4.3.1.5).

Am dritten Testtag wurden die Testteilnehmer einer mündlichen Prüfung unterzogen. Auch hier wurde die Einverständniserklärung der Probanden mündlich eingeholt, die Prüfungsgespräche zum Zweck der Forschung auf Tonband aufzunehmen. Kein Proband zeigte hier Bedenken, ausnahmslos alle stimmten der Tonaufnahme zu.[24]

24 Auf den Aufnahmen der mündlichen Prüfungen sind häufig Vornamen zu hören. Diese werden beim *Warming Up* von den Probanden freiwillig genannt. Eine Aufforderung hierzu bestand nicht. Alle Namen wurden zur Wahrung der Anonymität mit dem Audio-Editor Audacity mit einem Piepton überschrieben.

Auch das **Prinzip der Nicht-Schädigung** der Studienteilnehmer ist beachtet worden (vgl. STROHM KITCHENER & KITCHENER 2009: 12 f.). So wurde den Probanden zugesichert, dass ihre Testergebnisse keinesfalls an ihre Deutschkursdozenten weitergetragen würden und sich somit auch nicht negativ auf den Scheinerwerb bzw. die Abschlussnote des Kurses auswirken konnten. Nach Beendigung der mündlichen Prüfung erhielten die Testpersonen auf Wunsch ein adäquates, kompetenzorientiertes und konstruktives Feedback durch die Beurteiler.

4.3 Methode

Um zu ermitteln, ob und wie sich das Konstrukt des C-Tests unter *speed*-Bedingungen verändert, wird ein Studiendesign gewählt, bei dem die Experimental- und Kontrollgruppe identisch ist. Das Ziel ist es, einen Vergleich zwischen dem bereits etablierten C-Test-Format mit 5 Minuten Bearbeitungszeit pro Text und dem S-C-Test unter hohem Zeitdruck ziehen zu können. Hierzu wird der gleichen Probandengruppe in fixer Reihenfolge zunächst ein C-Test mit normaler Bearbeitungszeit und dann ein S-C-Test vorgelegt. Dies hat den Vorteil, dass die Teilnehmer zum Zeitpunkt der Durchführung des S-C-Tests bereits mit dem allgemeinen Format des C-Tests vertraut sind und auf einen Eisbrechertext verzichtet werden kann. Als Außenkriterium dienen eine Hörverstehensaufgabe und eine Aufgabe zum mündlichen Ausdruck, mit denen die Ergebnisse aus den beiden C-Test-Versionen korreliert werden.

Im Folgenden werden die verwendeten Forschungsinstrumente vorgestellt, sowie das Setting und der Aufbau der Studie konkret erläutert. Es folgt eine Beschreibung der Probandengruppe auf Grundlage des Begleitfragebogens sowie eine statistische Auswertung der vorliegenden Testergebnisse.

4.3.1 Instrumente

4.3.1.1 onDaF & S-C-Test

Der C-Test stellt das zentrale Messinstrument dieser Untersuchung dar. Um die Leistung der Probanden in einem C-Test und einem S-C-Test miteinander vergleichen zu können, gibt es verschiedene Möglichkeiten. So kann zum Beispiel ein C-Test-Set zweimal bei der gleichen Probandengruppe eingesetzt werden, einmal mit und einmal ohne eine (starke) Zeitbegrenzung. Dieses Vorgehen hat den Vorteil, dass der Schwierigkeitsgrad der Texte sich nicht schon ohne unterschiedliche Darbietungsbedingungen voneinander unterscheidet. Ein offensichtlicher Nachteil besteht jedoch in der Tatsache, dass Gedächtniseffekte nicht ausgeschlossen werden können und die Probanden somit im zweiten Testdurchlauf bereits einen Vorteil haben. Versucht man hier gegenzusteuern, indem man den zeitlichen Abstand zwischen den beiden Testläufen erhöht, besteht wiederum die Gefahr, dass sich der Sprachstand der Probanden in der Zwischenzeit verbessert oder auch verschlechtert hat, sodass etwaige Unterschiede in der Testleistung nicht eindeutig den verschiedenen Testbedingungen zugeschrieben werden können.

Eine elegantere Lösung ist es also, zwei möglichst parallele Tests einzusetzen. Eine solche Möglichkeit bietet der Einsatz des vom TestDaF-Institut entwickelten onDaF (vgl. URL 3: onDaF; URL 25: onSET: über onSET). Der onDaF ist ein computeradministrierter C-Test. Die Testteilnehmer bekommen „on the fly“ per Zufallsgenerator aus einem sehr großen Pool erprobter und kalibrierter C-Tests nacheinander acht Texte präsentiert, wobei die ersten beiden Texte für einen Lerner auf dem Niveau A2 lösbar sein sollen, die nächsten beiden Texte für einen Lerner auf dem Niveau B1. Dementsprechend folgen zwei Texte auf dem Niveau B2 sowie

zwei weitere auf dem Niveau C1 lösbare C-Test-Texte. Alle Texte im onDaF-Pool sind nicht nur einer Niveaustufe gemäß GER (EUROPARAT 2001), sondern ebenfalls bestimmten thematischen Kategorien zugeordnet. Somit kommt dem Zufallsgenerator auch die Funktion zu, eine Ballung thematisch ähnlicher Texte zu vermeiden. Probanden, die den onDaF eventuell schon zuvor in einem anderen Kontext abgelegt haben, entsteht somit kein Vorteil. Im Rahmen der Intensivkurse an der ZEMS an der TU Berlin wird dieser Test ohnehin zur Einstufung der neu eingeschriebenen Sprachkursteilnehmer verwendet. Der onDaF erfüllt im Rahmen der vorliegenden Studie zwei Funktionen: einerseits dient er als Vorauswahl von Probanden der anvisierten Niveaustufe, andererseits stellt er bereits einen Teil der hier verwendeten Testbatterie dar.

Den Teilnehmern des onDaF stehen für jeden Text fünf Minuten Bearbeitungszeit zur Verfügung. Es ist jedoch auch möglich, per Maus-Klick bereits vor Ablauf dieser fünf Minuten mit der Bearbeitung des folgenden Textes zu beginnen. Jeder der acht Texte hat 20 Lücken, die von den Testteilnehmern ausgefüllt werden sollen. Insgesamt werden den Testteilnehmern also 160 Lücken präsentiert. Es handelt sich beim onDaF um einen skalierten Test, der den Testteilnehmern neben dem erreichten Punktwert auch eine Information über die erreichte Niveaustufe gemäß GER (EUROPARAT 2001) ausgibt. Die Angabe der Niveaustufe hat den Vorteil, dass über den onDaF somit mit hinreichend hoher Wahrscheinlichkeit davon ausgegangen werden kann, dass die Probanden wirklich über das in dieser Untersuchung angestrebte B2-Niveau verfügen – tun sie das nicht, so können sie von der weiteren Studie ausgeschlossen werden.

Jeder Proband löst also beim onDaF acht unterschiedliche Texte. Um nun dennoch eine möglichst gute Vergleichbarkeit zwischen dem onDaF und einem S-C-Test erreichen zu können, fiel die Wahl auf den Einsatz

von acht onDaF-Texten als Papier-und-Bleistift-Test unter *speed*-Bedingungen.[25]

Der Einsatz von einem computeradministrierten C-Test (onDaF) und einem Papier-und-Bleistift-Test (S-C-Test) nebeneinander stellt in dieser konkreten Kombination kein Problem dar. So vergleichen BISPING & RAATZ (2002) einen Papier-und-Bleistift-C-Test mit einem computeradministrierten C-Test und finden dabei keine signifikanten Unterschiede.

Auch REICHERT et al. (2010) nutzen in ihrer Validierungsstudie ebenfalls sowohl computeradministrierte C-Tests als auch eine Papier-und-Bleistift-Version. Mittels einer Varianzanalyse überprüfen sie, ob die beiden Versionen äquivalent sind. Zwar konnte kein statistisch signifikanter Effekt der Darbietungsform festgestellt werden (F [1,232] = 0,75, p = 0,389), jedoch geben die Autoren an, dass sie die Äquivalenz der beiden Testformate nicht abschließend beurteilen können, da sie eine *interaction* von (F [1,232] = 7,26, p < 0,701), zwischen den beiden Faktoren gefunden hatten.

Die Ergebnisse einer Studie von BISPING (2006) weisen ebenfalls darauf hin, dass ein paralleler Einsatz von computeradministrierten und analogen C-Test-Versionen keinen Methodeneffekt mit sich bringt. Er folgert, dass „computerisierte C-Tests nicht weniger valide sind als traditionelle" (BISPING 2006: 162).

Die Durchführung am Computer scheint also zumindest bei zeitlich unlimitierten C-Tests oder solchen mit einer großzügig bemessenen Zeitbegrenzung keinen Einfluss auf die Testperformanz zu haben. Ob und wie sich hingegen die Testdurchführung am Computer auf die Ergebnisse bei einem S-C-Test auswirkt, wurde noch nicht untersucht. Damit wird auch begründet, dass in der vorliegenden Untersuchung der S-C-Test als Pa-

25 Mein herzlicher Dank gilt dem TestDaF-Institut in Bochum und Herrn Dr. Thomas Eckes, die mir freundlicherweise acht kalibrierte onDaF-Texte inkl. Lösungsschablonen zur Verwendung in dieser Studie anvertraut haben.

pier-und-Bleistift-Test eingesetzt wird. Wenngleich BISPING (2006) die Bearbeitungszeit seines Computer-C-Tests auf drei Minuten pro Text reduziert hat, besteht dennoch ein deutlicher Unterschied zu den hier angesetzten Zeiten zwischen 0:52 Minuten und 1:14 Minuten (vgl. Kapitel 4.3.2). Beim Einsatz derartig beschleunigter C-Tests am Computer können Faktoren wie zum Beispiel die Fähigkeit, blind mit zehn Fingern tippen zu können, die generelle Tippgeschwindigkeit (Zeichen/Minute) oder der Grad der Vertrautheit mit der deutschen Tastaturbelegung einen Einfluss auf das Testergebnis haben und so die Ergebnisse verzerren. Da in der vorliegenden Studie jedoch der onDaF mit fünf Minuten Bearbeitungszeit pro Text am Computer abgelegt wird und der S-C-Test derjenige ist, der als Papier-und-Bleistift-Test präsentiert wird, ist diese Kombination aus Darbietungsformaten vertretbar.

Die Tatsache, dass jeder Proband beim onDaF ein anderes, zufällig generiertes Testset bearbeitet, während beim S-C-Test alle Studienteilnehmer die gleichen Texte vorgelegt bekommen, wird in Kauf genommen. Obgleich die Parallelität der Tests nicht nachgewiesen werden kann, ist aufgrund der ausgiebigen Erprobung und Kalibrierung der onDaF-Texte davon auszugehen, dass beide Testformen hinreichend vergleichbar sind.

4.3.1.2 Subtest mündlicher Ausdruck

Der Zusammenhang zwischen der Leistung im C-Test mit der Leistung in Testformaten zum mündlichen Ausdruck bzw. Sprechen wurde – wenngleich im Vergleich mit anderen Fertigkeiten und Teilkompetenzen recht unterrepräsentiert – von einigen Autoren untersucht (vgl. Kapitel 3.2). Dabei ist das gewählte Außenkriterium jedoch häufig nicht standardisiert (vgl. z. B. DÖRNYEI & KATONA 1992) oder es ist sogar nur sehr wenig über den verwendeten Test bekannt (vgl. z. B. HUHTA 1996). In anderen Studien ist

der gewählte Test zum mündlichen Ausdruck nicht kommunikativ bzw. interaktiv orientiert (vgl. ARRAS et al. 2002).

Im Rahmen dieser Studie werden C-Test und S-C-Test gegen ein Deutsch-Zertifikat des Goethe-Instituts korreliert. Das Goethe-Institut hat sich bereits ab 1962 als Prüfungsinstitution etabliert und verfügt heutzutage über ein umfassendes Prüfungsportfolio (vgl. GOETHE-INSTITUT 2007: 3). Als Gründungsmitglied der *Association of Language Testers in Europe* (ALTE) hat sich das Goethe-Institut verpflichtet, die Anforderungen der ALTE zur Qualitätssicherung bei Sprachprüfungen zu erfüllen. Diese umfassen unter anderem Vorgaben zur Entwicklung, Erstellung und Beurteilung von Zertifikaten (vgl. GOETHE-INSTITUT 2007: 4). So werden Testsätze an ca. 200 Probanden erprobt und die Ergebnisse statistisch analysiert, um das Erfüllen der Testgütekriterien zu gewährleisten (vgl. GOETHE-INSTITUT 2007: 5). Ein wichtiger Aspekt zur Erfüllung der Gütekriterien ist die Standardisierung von Tests. Diese wird bei den Goethe-Zertifikaten durch spezifische Angaben zu den Durchführungsbestimmungen, dem zeitlichen Rahmen, dem geforderten quantitativen Umfang, inhaltlichen Leitpunkten sowie einem ausdifferenzierten Raster mit Bewertungskriterien und Lösungsschablonen gesichert.

Die Struktur des Subtests des Goethe-Zertifikats B2 zum mündlichen Ausdruck ist neben diesen Aspekten der Testqualität entscheidend: Im Gemeinsamen Europäischen Referenzrahmen für Sprachen (EUROPARAT 2001: 36) wird zwischen „zusammenhängendem Sprechen" und „an Gesprächen teilnehmen" unterschieden. Für das avisierte Niveau B2 bedeutet dies:

> Zusammenhängendes Sprechen:
>
> Ich kann zu vielen Themen aus meinen Interessengebieten eine klare und detaillierte Darstellung geben. Ich kann einen Standpunkt zu einer aktuellen Frage erläutern und Vor- und Nachteile verschiedener Möglichkeiten angeben.

> An Gesprächen teilnehmen:
>
> Ich kann mich so spontan und fließend verständigen, dass ein normales Gespräch mit einem Muttersprachler recht gut möglich ist. Ich kann mich in vertrauten Situationen aktiv an einer Diskussion beteiligen und meine Ansichten begründen und verteidigen. (ebd.)

Der Subtest des Goethe-Zertifikats B2 zum mündlichen Ausdruck verfügt über zwei Teilaufgaben, von denen eine monologisches Sprechen (Aufgabe 1) und eine dialogisches Sprechen, d. h. sprachliches Agieren und Reagieren erfordert (Aufgabe 2) (vgl. FREY 2012: 43 ff.). Dies ermöglicht es, einerseits beide Aufgaben als Gesamtheit zu betrachten und andererseits Berechnungen getrennt nach monologischem und dialogischem Sprechen vorzunehmen. Dies ist nicht nur aufgrund der Unterteilung in „zusammenhängendes Sprechen" und „an Gesprächen teilnehmen" im Referenzrahmen sinnvoll. Da beim Teilnehmen an Gesprächen spontaner agiert und auf den Redebeitrag des Gegenübers eingegangen werden muss, hat der Sprecher hierbei weniger Zeit, seine Äußerung zu planen als beim zusammenhängenden Sprechen. Dies entspricht daher noch mehr einer Sprachverwendung in Echtzeit.

Beim Goethe-Zertifikat B2 handelt es sich hierbei um eine standardisierte Prüfung mit festen Durchführungsbestimmungen und Bewertungsrichtlinien, die sich am Gemeinsamen Europäischen Referenzrahmen für Sprachen (EUROPARAT 2001) orientieren. So liefert der Referenzrahmen etwa zur Beurteilung mündlicher Sprachkompetenz ein Raster, das die Kategorien Spektrum, Korrektheit, Flüssigkeit, Interaktion und Kohärenz umfasst. Für das Niveau B2 beinhaltet dies (vgl. Tab. 19):

Spektrum	Verfügt über ein ausreichend breites Spektrum von Redemitteln, um in klaren Beschreibungen oder Berichten über die meisten Themen allgemeiner Art zu sprechen und eigene Standpunkte auszudrücken; sucht nicht auffällig nach Worten und verwendet einige komplexe Satzstrukturen.
Korrektheit	Zeigt eine recht gute Beherrschung der Grammatik. Macht keine Fehler, die zu Missverständnissen führen, und kann die meisten eigenen Fehler selbst korrigieren.
Flüssigkeit	Kann in recht gleichmäßigem Tempo sprechen. Auch wenn er/sie eventuell zögert, um nach Strukturen oder Wörtern zu suchen, entstehen kaum auffällig lange Pausen.
Interaktion	Kann Gespräche beginnen, die Sprecherrolle übernehmen, wenn es angemessen ist, und das Gespräch beenden, wenn er/sie möchte, auch wenn das möglicherweise nicht immer elegant gelingt. Kann auf vertrautem Gebiet zum Fortgang des Gesprächs beitragen, indem er/sie das Verstehen bestätigt, andere zum Sprechen auffordert usw.
Kohärenz	Kann eine begrenzte Anzahl von Verknüpfungsmitteln verwenden, um seine/ihre Äußerungen zu einem klaren, zusammenhängenden Beitrag zu verbinden; längere Beiträge sind möglicherweise etwas sprunghaft.

Tab. 19: Beurteilungsraster des Gemeinsamen Europäischen Referenzrahmens (EUROPARAT 2001: 37)

Das Bewertungsschema des Goethe-Zertifikats B2 differenziert die Aspekte Erfüllung der Aufgabenstellung (Inhalt und Angemessenheit), Kohärenz und Flüssigkeit (Verknüpfungen, Sprechtempo), Ausdruck (Wortwahl, Umschreibungen, Wortsuche), Korrektheit (Morphologie, Syntax) sowie Aussprache und Intonation (Laute, Wortakzent, Satzmelodie) (vgl. FREY 2012: 41 f.). In jedem dieser Bereiche kann ein Testteilnehmer auf einer fünfstufigen Skala zwischen 2,5 und 0 Punkten erreichen. Das Goethe-Zertifikat bietet somit die Möglichkeit einer analytischen Bewertung der mündlichen Sprachkompetenz. Der im Bewertungsraster des Goethe-Zer-

tifikats B2 auftretende Aspekt Erfüllung der Aufgabenstellung erscheint im Gemeinsamen Europäischen Referenzrahmen naturgemäß nicht. Des Weiteren sind die Punkte Kohärenz und Flüssigkeit im Goethe-Zertifikat zusammengefasst, während sie im Referenzrahmen getrennt aufgeführt werden (vgl. EUROPARAT 2001: 36).

Es zeigt sich, dass sich das Goethe-Zertifikat B2 – ebenso wie die Goethe-Zertifikate der anderen Niveaustufen – sehr eng an den Gemeinsamen Europäischen Referenzrahmen anlehnt und somit als valide Prüfung zu den GER-Stufen betrachtet werden kann.[26]

Einer von vier zur Verfügung stehenden Modellsätzen wurde gemeinsam mit zwei erfahrenen Dozenten für Deutsch als Fremdsprache der ZEMS ausgewählt, die über jahrelange Unterrichtserfahrung verfügen und mit den Anforderungen an das B2-Niveau bestens vertraut sind. Das Goethe-Zertifikat B2 gilt als bestanden, wenn ein Teilnehmer insgesamt 60 % der Punkte erreicht.[27] Zudem müssen in den schriftlichen Prüfungsteilen (Leseverstehen, Hörverstehen und schriftlicher Ausdruck) zusammengenommen mindestens 45 von maximal 75 Punkten erreicht werden und im mündlichen Ausdruck mindestens 15 von maximal 25 Punkten (vgl. FREY 2012: 7). Während folglich Schwächen in einem schriftlichen Prüfungsteil durch die beiden anderen schriftlichen Prüfungsteile ausgeglichen werden können, hat die Fertigkeit Sprechen hier eine Sonderstellung.

26 Zu Hintergründen und Entstehung des GER vgl. z. B. LITTLE (2007).

27 Punktwerte zur Benotung des Goethe-Zertifikats B2: 100–90 Punkte = sehr gut; 89,5–80 Punkte = gut; 79,5–70 Punkte = befriedigend; 69,5–60 Punkte = ausreichend; < 60 Punkte = nicht bestanden (vgl. FREY 2012: 7).

4.3.1.3 Flüssigkeit

Ergänzend zur Bewertung des mündlichen Ausdrucks mithilfe des Bewertungsrasters vom Goethe-Zertifikat B2, wird ein Teil der Tonaufnahmen hinsichtlich der Flüssigkeit des mündlichen Ausdrucks der Teilnehmer untersucht. Flüssigkeit ist sowohl eine Bewertungskategorie im Raster des Goethe-Zertifikats (vgl. Tab. 19) als auch eine der drei bereits seit den 1980er bzw. 1990er Jahren vorgeschlagenen „three fundamental dimensions characterizing L2 usage" (MICHEL 2017: 50), nämlich *complexity, accuracy* und *fluency* (vgl. z. B. HOUSEN et al. 2012).

Flüssigkeit kann von unterschiedlichen Standpunkten aus betrachtet werden. SEGALOWITZ (2010: 46 f.) unterscheidet zum einen die kognitive Flüssigkeit, welche es einem Sprecher ermögliche, Gedanken fließend in Sprache zu überführen. Des Weiteren nennt er die wahrgenommene Flüssigkeit, welches ein subjektives Maß auf der Seite des Hörers darstellt. In dieser Arbeit wird nur die dritte Art von Flüssigkeit, die sogenannte *utterance fluency* betrachtet, da diese objektiv messbare Parameter enthält.

Generell können drei unterschiedlichen Typen von Flüssigkeitsparametern unterschieden werden:

> In terms of language processing, speed is associated with control of and access to proceduralized knowledge; breakdown is thought to reflect the planning and conceptualization stages of language production; while repair fluency is seen as an indicator of monitoring processes. (MICHEL 2017: 56)

Alle drei Kategorien werden in der vorliegenden Arbeit abgedeckt; durch das Zählen von Wörtern und Silben pro Minute (*speed fluency*), das Zählen von Wortwiederholungen (*repair fluency*) sowie das Zählen von Verzögerungsindikatoren (*breakdown fluency*) (vgl. Kapitel 4.4.3.6).

4.3.1.4 Subtest Hörverstehen

Der Testteil zum Hörverstehen stammt aus dem gleichen Modellsatz des Goethe-Zertifikats B2 wie der Testteil zum mündlichen Ausdruck (vgl. FREY 2012). Da der Fokus der Untersuchung auf dem mündlichen Ausdruck lag, war es vorrangiges Ziel, ein für das Sprechen geeignetes Testformat zu finden. Um die verwendeten Außenkriterien möglichst vergleichbar zu machen, wurde für das Hörverstehen auf die gleiche Prüfung zurückgegriffen.

Der Subtest zum Hörverstehen des Goethe-Zertifikats B2 (vgl. FREY 2012) besteht aus zwei Aufgaben. Bei der ersten Aufgabe liegt den Testteilnehmern eine Art Stunden- oder Zeitplan vor, den es zu ergänzen bzw. zu korrigieren gilt. Nachdem die Teilnehmer Zeit bekommen, um sich diesen Plan anzusehen, hören sie einmal einen scheinbar auf den Anrufbeantworter gesprochenen Text, aus dem die fehlenden bzw. zu korrigierenden Informationen zu entnehmen sind. Die Tondatei dauert zwei Minuten. Den Testteilnehmern wird eine Kurzantwort abverlangt, d. h. sie müssen eigenständig Wörter oder Zahlen in die entsprechenden Felder eintragen. Vorgegebene Antwortmöglichkeiten gibt es bei dieser Aufgabe nicht. Insgesamt fünf Informationseinheiten gilt es auf dem Zeitplan zu bearbeiten.

Die zweite Aufgabe des Hörverstehenstests besteht aus zehn Multiple-Choice-Items, die jeweils über drei Antwortmöglichkeiten (Bestantwort plus zwei Distraktoren) verfügen. Auch hier bekommen die Testteilnehmer vor Start der Tondatei Zeit, sich die Items durchzulesen. Der gehörte Text besteht aus einem Gespräch zwischen drei Personen. Ein Interviewer befragt einen Mann und eine Frau, die unterschiedliche Perspektiven auf ein Thema repräsentieren. Diese Tondatei dauert etwa 7:30 Minuten und wird zweimal gehört. Beim zweiten Hören ist die Datei auf dem Tonträger in drei Abschnitte unterteilt, die jedoch lediglich von der Ansage des nächsten Abschnitts unterbrochen werden.

4.3.1.5 Fragebogen

Der Begleitfragebogen (vgl. Anhang F) dient dazu, einerseits Hintergrundinformationen über die Probandengruppe zu erheben, andererseits die Einschätzung der Testteilnehmer zu den verwendeten C-Test-Formaten einzuholen.

Der Fragebogen ist in drei Teilbereiche gegliedert: Der erste Teil des Fragebogens ist überschrieben mit „Persönliche Daten" und erfragt das Alter, das biologische Geschlecht, die L1, das Studienfach bzw. die Studienfächer und das Herkunftsland. Darüber hinaus wird über ein Item erfasst, mit welchem Schriftsystem die Probanden alphabetisiert wurden. Dies könnte relevant sein, falls sich die Schreibgeschwindigkeit in Kombination mit den weniger gewohnten Buchstaben nachteilig auf die Testperformanz im S-C-Test auswirkt. Ein letztes Item in diesem ersten Fragebogenteil erfasst die durchschnittliche wöchentliche Stundenzahl, die ein Proband seiner Selbsteinschätzung folgend mit Tippen am PC verbringt. Zwar ist der eingesetzte computeradministrierte C-Test mit so großzügiger Arbeitszeit bemessen, dass die Vertrautheit mit der PC-Tastatur eine untergeordnete Rolle spielen sollte, jedoch können etwaige Ausreißer- oder Extremwerte eventuell retrospektiv hierdurch erklärt werden.

Der zweite Teil des Fragebogens befasst sich mit den Lernerfahrungen der Probanden. Hier wird zunächst der Lernkontext des Deutschen als Fremdsprache bzw. Zweitsprache (mit Möglichkeit der Mehrfachnennung) abgefragt sowie das Alter zum Zeitpunkt des Lernbeginns. Es folgen zwei Items, die sich auf die Fertigkeit Sprechen beziehen. Zum einen werden die Teilnehmer danach gefragt, wie häufig sie die Gelegenheit haben, Deutsch zu sprechen. Zum anderen können sie offen darauf antworten, ob es Situationen gibt, in denen sie sich unwohl fühlen, Deutsch zu sprechen. Diese Angabe ist für den Testteil des mündlichen Ausdrucks von Interesse, denn sie könnte Aufschluss darüber geben, ob ein Testteilnehmer sich tendenzi-

ell in fremdsprachlichen Kommunikationssituationen unwohl fühlt. Im Folgenden werden noch die Nutzung deutschsprachiger Medien sowie Kenntnisse in weiteren Fremdsprachen abgefragt. Ein letztes Item ermittelt, ob die Probanden vor der Studienteilnahme schon einmal einen C-Test abgelegt haben und somit mit diesem Testformat vertraut sind.

Der dritte und letzte Teil des Fragebogens widmet sich den Erfahrungen der Probanden mit dem onDaF bzw. S-C-Test. Für beide Testvarianten wird abgefragt, ob die Teilnehmer der Meinung sind, dass der Test ihre Deutschkenntnisse korrekt wiedergibt. Außerdem wird (mit Möglichkeit der Mehrfachnennung) erhoben, wie die Probanden die C-Tests bearbeitet haben (z. B. lineare Arbeitsweise). Für den onDaF werden zudem die erreichten Punkte erfasst, während beim S-C-Test noch ein Item erfragt, ob die Teilnehmer aufgrund ihrer mechanischen Schreibgeschwindigkeit Schwierigkeiten beim Lösen des S-C-Tests hatten.

Nach der Anmerkung einer Probandin, dass es ebenfalls aufschlussreich sein könnte, ob jemand mit der linken oder rechten Hand schreibt, da ein Rechtshänder beim Schreiben den unmittelbar folgenden Kontext mit der Schreibhand verdeckt, während ein Linkshänder den nachfolgenden Text schon während des Schreibens lesen kann, wurde der Fragebogen um zwei Items erweitert, die zum einen nach der Schreibhand fragen, und zum anderen danach, ob man sich durch die Händigkeit beim Ablegen des S-C-Tests benachteiligt fühlte. Zwar ist es bedauerlich, dass diese Items nicht von Anbeginn in den Fragebogen integriert waren. Jedoch sind Linkshänder in der Gesamtpopulation im Vergleich zu Rechtshändern prozentual in so

geringer Zahl vertreten, dass mit 125 Probanden ohnehin nur mit größten Schwierigkeiten Berechnungen zu diesem Aspekt möglich gewesen wären.[28]

Tabelle 20 bietet einen Überblick über die Operationalisierungen der Konstrukte sowie die jeweils angesetzte Bearbeitungszeit pro Testteil.

Um den Fragebogen und die unterschiedlichen Testteile sicher ein und derselben Person zuordnen zu können, wurde jeder Testbogen von den Probanden mit einem selbst erstellten Teilnehmer-Code versehen. Dieser war stets achtstellig und bestand aus den ersten beiden Buchstaben des Vornamens der Mutter, den ersten beiden Buchstaben des Vornamens des Vaters, dem Geburtstag und dem Geburtsmonat des Probanden. Auf diese Weise war nicht nur die Anonymität der Probanden gesichert, sondern es war ebenfalls gewährleistet, dass die Testteilnehmer sich auch am zweiten und dritten Tag der Datenerhebung noch an ihren persönlichen Code erinnern bzw. ihn erneut erzeugen konnten. Diese Art der Teilnehmercodeerstellung wurde von AGUADO et al. (2007) bzw. GROTJAHN et al. (2010) übernommen.

28 Die Anmerkung der oben genannten Probandin führte jedoch dazu, das Thema der Schreibhand beim S-C-Test in einem anderen Projekt zu beleuchten. Hierbei werden gezielt muttersprachliche Links- und Rechtsschreiber akquiriert und ihre Leistungen im S-C-Test miteinander verglichen (vgl. BERGER & ZIMMERMANN (in Vorbereitung)).

Variable	Quelle	Bearbeitungszeit	Konstrukt	Aufgabe
C-Test	onDaF	5 min pro Text, 40 min insgesamt	Allgemeine Sprachkompetenz	pro Text 20 Lücken füllen
S-C-Test	Papier-und-Bleistift-Version des onDaF	Text 1 & 2: 52 sec Text 3 & 4: 58 sec Text 5 & 6: 67 sec Text 7 & 8: 74 sec insgesamt: 8 min 37 sec (vgl. Kapitel 4.3.2)	Untersuchungsgegenstand	pro Text 20 Lücken füllen
Hörverstehen	Modellsatz des Goethe-Zertifikat B2 (vgl. FREY 2012)	Aufgabe 1: 8 min Aufgabe 2: 22 min (Tondateien: Aufgabe 1: 2 min Aufgabe 2: 15 min)	Hörverstehen	Aufgabe 1: Zeitplan per Kurzantwort ergänzen und korrigieren Aufgabe 2: 10 Multiple-Choice-Items zu einem Radio-Interview
Sprechen	Modellsatz des Goethe-Zertifikat B2 (vgl. FREY 2012)	15 min Vorbereitung, 3–4 min (Monolog) bzw. 6–7 min (Dialog)	Monologisches und dialogisches Sprechen	Aufgabe 1: monologisches Sprechen über ein vorgegebenes Thema Aufgabe 2: dialogisches Sprechen, sich über etwas einig werden und argumentieren

Tab. 20: Übersicht über die Erhebungsinstrumente[29]

29 Diese Tabelle ist angelehnt an die Übersicht in RAATZ (2002: 185).

4.3.2 Zeitpilotierung

C-Tests werden normalerweise unter Niveaubedingungen durchgeführt. Um aus dem C-Test einen *speeded*-C-Test zu machen, ist es nötig, die den Testteilnehmern zur Verfügung stehende Arbeitszeit pro Text so zu reduzieren, dass ein deutlicher Zeitdruck erzeugt wird. Jedoch muss die Zeit zugleich so bemessen sein, dass eine vollständige Bearbeitung der C-Test-Texte zumindest theoretisch im Rahmen der vorgegebenen Arbeitszeit möglich ist.

Zur Bestimmung der den Deutsch lernenden Probanden zur Verfügung stehenden Bearbeitungszeit pro C-Test-Text wurde das C-Test-Set an Muttersprachlern des Deutschen erprobt (vgl. Kapitel 4.3.2). Für diese Pilotierung der Bearbeitungszeit wurden 20 Personen getestet, die teilweise im beruflichen, teilweise im privaten Kontext oder über Dritte vermittelt gefunden wurden oder sich auf einen Aushang gemeldet hatten (vgl. Anhang A). Die muttersprachlichen Probanden waren Akademiker im Alter von 22 bis 54 Jahren ($\tilde{x}$ = 32 Jahre; $\bar{x}$ = 33,3 Jahre), darunter 10 Männer und 10 Frauen (jeweils 50 %). Vier Probanden (20 %) hatten bereits zuvor einen C-Test abgelegt.

Das zu erprobende Test-Set bestand aus den acht vom TestDaF-Institut zur Verfügung gestellten, kalibrierten onDaF-Texten, die nach aufsteigendem Schwierigkeitsgrad angeordnet waren. Diesen acht Texten wurde ein weiterer, selbst erstellter und bereits in anderen Studien (vgl. SCHLAK et al. 2010) zum gleichen Zweck verwendeter C-Test als Eisbrechertext vorangestellt (vgl. Anhang D). Hierdurch sollte verhindert werden, dass ein Proband beim ersten Text relativ zu den anderen Texten schlechter abschneidet, weil er sich erst an das Testformat gewöhnen muss. Zwischen den einzelnen Texten wurde je ein Zwischenblatt mit dem Text „Bitte warten!“ eingefügt, um zu verhindern, dass im Falle unerwünschten vorzeitigen Weiterblätterns bereits ein erster Blick auf den nachfolgenden

Text geworfen werden konnte. Auf dem Deckblatt des Test-Sets fanden sich detaillierte Anweisungen zur Bearbeitung des Tests (vgl. Anhang D). Durch die schriftlichen Testanweisungen sollte die Durchführungsobjektivität sichergestellt werden.

Im Anschluss an den C-Test füllten die Muttersprachler einen kurzen Fragebogen aus, der Informationen über ihr Alter und Geschlecht, über den höchsten Bildungsabschluss und das Studienfach, die aktuelle berufliche Tätigkeit und Fremdsprachenkenntnisse erfasste. Außerdem wurden die Teilnehmer gebeten eine Angabe darüber zu machen, ob sie schon vor der Teilnahme an dieser Studie einen C-Test abgelegt hatten.

Die Zeitpilotierung dieser Studie ist an das in GROTJAHN et al. (2010: 301) beschriebene Verfahren angelehnt. Die Autoren ermittelten eine textspezifische Zeitbemessung für ihren C-Test, indem sie zwanzig sowohl muttersprachliche als auch nicht-muttersprachliche Probanden eine Reihe von C-Test-Texten mit einer maximalen Bearbeitungszeit von drei Minuten pro Text bearbeiten ließen. Die Probanden wurden zudem dazu aufgefordert, aufzuzeigen, sobald sie mit der Bearbeitung eines Textes fertig waren. Die benötigte Zeit der jeweils als drittes fertigen Person wurde als Zeitbemessung für den entsprechenden Text festgelegt. Die Autoren liefern hingegen keine Angaben dazu, ob die Lösungsrate der Probanden hierbei eine Rolle spielte. Wie bereits erwähnt, wird bei muttersprachlichen Probanden von einer Lösungsquote von 90 % ausgegangen. Bei den C-Test-Texten der Studie von GROTJAHN et al. (2010) wurden Texte mit jeweils 25 Lücken verwendet, was einer korrekten Lösung von mindestens 22 (rechnerisch: 22,5) Lücken entspricht. Sollte die Lösungsquote nicht beachtet worden sein, kann nicht rekonstruiert werden, ob die Probanden in der ermittelten Zeit den C-Test nur vollständig bearbeitet oder auch (weitestgehend) korrekt gelöst haben.

Ein erster Testdurchlauf wurde zunächst an einer gebildeten Muttersprachlerin vorgenommen. Auf diese Weise sollte zum einen sichergestellt werden, dass die Anweisungen auf dem Deckblatt des Testsets eindeutig sind und verstanden werden. Dies beinhaltet auch, dass ein muttersprachlicher Proband nach Lesen der Testanweisung sorgfältig genug arbeitet, um die angestrebte Lösungsrate von 90 % bei Muttersprachlern (hier entsprechend 18 von 20 korrekten Lösungen pro C-Test-Text) (vgl. GROTJAHN 2002: 216) zu erreichen, d. h. dass die Lösungsgüte nicht zugunsten der Arbeitsgeschwindigkeit vernachlässigt wird. Das Ergebnis dieses ersten Durchlaufs war sehr positiv: Die Probandin erreichte in allen Texten bei Zeiten zwischen 0:48 min und 1:33 min eine Lösungsquote von 100 % mit Ausnahme nur eines Textes, bei dem immerhin eine Lösungsquote von 95 % erreicht wurde. Modifikationen am Aufbau des Testsets waren somit nicht erforderlich. Es kam jedoch der Hinweis, dass das Anschauungsbeispiel des C-Test-Formats auf dem Deckblatt recht komplex sei und durch ein sprachlich einfacheres ersetzt werden könnte. Diesem Vorschlag wurde nachgegangen.

Im Anschluss wurde das sonst unveränderte Test-Set an gebildeten Muttersprachlern ohne sprachwissenschaftlichen Bildungshintergrund erprobt. Da bei Muttersprachlern alle Prozesse automatisiert sind, ist davon auszugehen, dass sie dazu in der Lage sind, mit reduzierter Redundanz umzugehen und somit C-Test-Texte zügig zu rekonstruieren. Ziel war es, von mindestens zehn Probanden Daten zu erhalten, und aus diesen die Zeitbemessungen für Lerner des Deutschen zu ermitteln. Die muttersprachlichen Probanden wurden einzeln getestet. Auf diese Weise konnten die Zeiten für jede Person und jeden einzelnen Text getrennt erhoben werden. Es zeigte sich schnell, dass hierfür deutlich mehr als zehn Probanden nötig waren, da trotz der deutlichen Anweisung, dass das korrekte Lösen der Lücken wichtiger sei als ein besonders schnelles Arbeitstempo, ei-

nige Probanden nicht in allen C-Test-Texten die angestrebte Lösungsquote von 90 % (entsprechend 18 Lücken) erreichten. Dies war insbesondere beim letzten und somit schwersten Text der Fall. Da es sich jedoch um bereits erprobte und kalibrierte Texte des TestDaF-Instituts handelt, ist davon auszugehen, dass die Ursache für das Verfehlen der üblichen muttersprachlichen Lösungsquote damit zusammenhängt, dass die entsprechenden Probanden möglichst schnell sein wollten und sich somit Fehler eingeschlichen haben (vgl. *Speed-Accuracy-Trade-Off*, Kapitel 2.2.5.1). Die Tatsache, dass den Probanden während der Datenerhebung die Testleiterin mit einer Stoppuhr in der Hand gegenüber saß, könnte ein nervositäts- und somit fehlererzeugender Faktor gewesen sein. Insgesamt wurden also Daten von 20 Muttersprachlern erhoben. Tabelle 21 gibt eine Übersicht über die von den muttersprachlichen Probanden erreichten Punkte pro C-Test-Text. Die dunkelgrau unterlegten Kästchen zeigen Texte an, die nicht mit der bei Muttersprachlern erwartbaren Lösungsquote von 90 % gelöst wurden (vgl. Tab. 21). Diese Probanden wurden nicht in die weitere Berechnung einbezogen. Von den verbleibenden elf Personen wurden für jedes Textniveau, also A2, B1, B2 und C1, die zehn Topzeiten ausgewählt und aus diesen das arithmetische Mittel gebildet (vgl. Tab. 22).

Die auf diese Weise berechneten Werte wurden aufgerundet und jeweils als Zeitbemessung für die Texte einer Niveaustufe gemäß Gemeinsamem Europäischen Referenzrahmen festgesetzt. Die Entscheidung, nur die in Tabelle 22 dunkelgrau unterlegten *Top Ten* der Bearbeitungszeiten zur Berechnung der für die Lernergruppe der Hauptstudie anzusetzende Bearbeitungszeit heranzuziehen, liegt in Erfahrungen bei der Datenerhebung mit den muttersprachlichen Probanden begründet: Trotz der Anweisung so schnell (aber so sorgfältig) wie möglich zu arbeiten, hielten sich einzelne Probanden bis zu 20 Sekunden bei einer letzten noch nicht ausgefüllten

und für sie offensichtlich schwierigen Lücke auf, ehe sie den Stift niederlegten und so das Zeichen zum Stoppen der Arbeitszeit gaben.

Proband	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8
M1	20	19	**17**	20	19	20	20	**15**
M2	19	19	19	19	19	19	18	**9**
M3	20	20	19	20	18	18	18	18
M4	20	20	20	20	20	20	20	19
M5	20	20	20	20	19	20	20	20
M6	20	20	18	18	18	**15**	**17**	**15**
M7	20	20	20	18	20	20	20	18
M8	20	19	**15**	18	**17**	18	19	18
M9	20	20	20	20	20	20	20	19
M10	20	20	19	20	20	20	20	19
M11	20	20	20	19	20	**17**	**17**	**16**
M12	20	20	19	**17**	18	19	18	18
M13	19	20	20	19	18	20	**16**	18
M14	20	20	19	20	19	19	18	20
M15	20	19	20	20	20	20	20	**17**
M16	20	20	19	19	19	18	18	18
M17	20	20	20	20	20	20	20	18
M18	20	20	20	20	20	20	20	20
M19	20	20	20	20	20	19	19	19
M20	20	20	**16**	20	18	19	20	**17**

Tab. 21: Korrekt gelöste Lücken nach Text und Proband

Proband	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8
Niveau	A2		B1		B2		C1	
M3	67	75	84	91	104	202	101	132
M4	**48**	**50**	**47**	**56**	**55**	**58**	**67**	**60**
M5	61	52	**64**	**60**	**68**	**76**	**69**	125
M7	**38**	**44**	**49**	**57**	**56**	**59**	**55**	**75**
M9	75	71	70	72	85	99	105	182
M10	**47**	**55**	**63**	72	**70**	91	**74**	**96**
M14	67	80	91	96	87	219	139	166
M16	66	73	83	98	**77**	125	109	**100**
M17	77	**62**	**55**	**69**	85	88	**75**	186
M18	**58**	69	**64**	73	**77**	98	**66**	123
M19	**54**	**63**	69	100	**76**	81	133	189
$\bar{x}$	51,9		58,4		67,2		73,7	

Tab. 22: Lösungszeit nach Text und Proband in Sekunden

GROTJAHN et al. (2010) ermittelten in ihrer Studie eine massive textspezifische Zeitlimitierung für die darin verwendeten C-Tests. Textspezifische Zeitbemessungen lassen jedoch nur wenige allgemeine Aussagen über den Einfluss einer konkreten Bearbeitungszeit auf das C-Test-Konstrukt zu. Da es sich bei den genutzten onDaF-Texten um Texte sehr verschiedenen Schwierigkeitsgrads handelt, wäre jedoch auch eine einzige, für alle C-Test-Texte gültige Zeitbegrenzung problematisch. Denn so liefe man Gefahr, dass eine gewählte Zeitbemessung entweder die Bearbeitung der einfacheren Texte zu einfach macht, oder sie sich für die schweren Texte als zu anspruchsvoll erweist. In diesem Spektrum stellen nach Kompetenzniveaus gestufte Zeitbemessungen einen Mittelweg dar, der einerseits der variierenden Textschwierigkeit Rechnung trägt, es andererseits aber mög-

lich macht, allgemeinere Aussagen über die Bearbeitungszeit für beispielsweise Texte eines B1-Niveaus treffen zu können.

4.3.3 Probanden

Die Probanden für diese Studie sind ausnahmslos Lerner des Deutschen als Fremdsprache, die an der Zentraleinrichtung Moderne Sprachen, dem Sprachenzentrum der Technischen Universität Berlin, Deutschkurse auf B2- und (in wenigen Ausnahmen) C1-Niveau belegt haben.[30] Sie besuchten sowohl Intensivkurse (3 Wochen à 4 Unterrichtseinheiten pro Tag) als auch semesterbegleitende Sprachkurse (15 Wochen à 4 Unterrichtseinheiten pro Woche). Insgesamt wurden Daten in elf Sprachkursen erhoben. Das B2- bzw. C1-Niveau wurde aus mehreren Gründen gewählt: Der wichtigste ist, dass der S-C-Test bisher nur bei Muttersprachlern und sehr fortgeschrittenen Fremdsprachenlernern eingesetzt wurde, meist, um die Schwierigkeit eines C-Tests zu erhöhen. Was für Ergebnisse der S-C-Test bei Lernern der Mittelstufe (*Vantage* oder *Upper Intermediate,* vgl. EUROPARAT 2001: 44) liefert, ist noch völlig unbekannt. Auf die darunter liegenden Kompetenzstufen trifft dies folglich in noch stärkerem Maße zu. Die Auswirkung des S-C-Tests auf die Testergebnisse bei Lernern des A1-, A2- und B1-Niveaus kann und soll im Rahmen dieser Studie nicht untersucht werden. Da C1-Lerner des Deutschen nur sehr vereinzelt in Sprachkursen der ZEMS zu finden sind und dies die Suche nach Probanden und infolgedessen auch die mehrtägige Datenerhebung erheblich verkompliziert hätte, wurde auf das höchste verfügbare Niveau, nämlich B2, zurückgegriffen.

30 Ich bedanke mich herzlich bei Dr. Almut Schön, Geschäftsführerin der ZEMS an der TU Berlin, für die Möglichkeit, meine umfangreiche Datenerhebung dort durchführen zu dürfen. Mein Dank gilt ebenfalls den zahlreichen Dozenten, in deren Deutschkursen die Datenerhebungen stattfanden.

Unter den 125 Teilnehmern der Studie befinden sich lediglich 58 Personen, die sowohl das anvisierte Niveau B2 gemäß onDaF erreichten als auch alle vier Testteile (onDaF, S-C-Test, Hörverstehen und mündlicher Ausdruck) ablegten. Aus dieser Tatsache ergeben sich unterschiedliche Subgruppen für die Berechnung der Reliabilität und den Korrelationen mit den Fertigkeiten Hörverstehen und mündlicher Ausdruck (siehe unten).

Der folgende Boxplot (Abb. 9) zeigt die Berechnung des Alters zu Beginn des Deutschlernens unter Einbezug aller 125 Teilnehmer.

*Abb. 9: Lernbeginn in Jahren (*n = *125)*

Während die beiden Extremwerte nach oben (Lernbeginn mit jeweils 31 Jahren) als unproblematisch einzustufen sind, nicht nur weil sie sehr nah an der oberen Antenne liegen, sondern auch, weil das Alter des Lernbeginns für die behandelte Fragestellung weniger relevant ist, müssen die Ausreißer nach unten anders bewertet werden. Fünf Testpersonen haben einen Lern-

beginn der deutschen Sprache mit 0 Jahren (n = 3), 1 Jahr (n = 1) und 2 Jahren (n = 1) angegeben. In der Diskussion um die Abgrenzung von Erst-, Zweit- und Fremdsprache wird ein Alter von drei bis maximal vier Jahren als Grenze zwischen Erst- und Zweitsprachenerwerb angesetzt. Das bedeutet, findet der bedeutsame Erstkontakt zu einer Sprache bis zum dritten oder vierten Lebensjahr statt, so spricht man gemeinhin vom Erstsprachenerwerb, danach hingegen von Zweitsprachenerwerb (vgl. AHRENHOLZ 2010: 5 f; RÖSCH 2011: 11). Die oben genannten Probanden mit einem Erstkontakt zur deutschen Sprache vor dem dritten Lebensjahr werden aus sämtlichen weiteren Berechnungen ausgeschlossen, da sie nicht als Fremd- oder Zweitsprachler eingestuft werden können.

Unter den verbleibenden 120 Probanden befinden sich 49 Männer (40,8 %) und 69 Frauen (57,5 %). Zwei Probanden (1,7 %) haben keine Angabe zu ihrem Geschlecht gemacht. Die Probanden hatten zum Zeitpunkt der Datenerhebung ein durchschnittliches Alter von 23,73 Jahren ($\tilde{x}$ = 23, Min = 19, Max = 35). Die Gruppe setzt sich aus Sprechern verschiedener Erstsprachen zusammen, wobei Chinesisch die am häufigsten vertretene Erstsprache ist (vgl. Tab. 23).[31]

31 Sieben Probanden haben zwei Erstsprachen angegeben. Die genannten Kombinationen sind Russisch/Ukrainisch, Russisch/Lettisch, Galicisch/Spanisch, Kantonesisch/Englisch, Telugu/Hindi (jeweils einmal), Katalan/Spanisch (zweimal).

Sprache	Häufigkeit	Prozent	Sprache	Häufigkeit	Prozent
Chinesisch	22	18,3	Schwedisch	3	2,5
Polnisch	11	9,2	Arabisch	2	1,7
Portugiesisch	9	7,5	Dänisch	2	1,7
Spanisch	9	7,5	Indonesisch	2	1,7
Englisch	8	6,7	Ukrainisch	2	1,7
Französisch	8	6,7	Slowenisch	2	1,7
Finnisch	4	3,3	Bulgarisch	1	0,8
Russisch	4	3,3	Estnisch	1	0,8
Türkisch	4	3,3	Galicisch	1	0,8
Ungarisch	4	3,3	Kantonesisch	1	0,8
Griechisch	3	2,5	Koreanisch	1	0,8
Italienisch	3	2,5	Kurdisch	1	0,8
Niederländisch	3	2,5	Rumänisch	1	0,8
Norwegisch	3	2,5	Telugu	1	0,8
Persisch	3	2,5	Vietnamesisch	1	0,8

Tab. 23: Erstsprachen der Probanden wie angegeben auf Fragebogen (n = 120)

Entsprechend den Erstsprachen ergibt sich die in Tab. 24 gezeigte Verteilung der Herkunftsländer der Probanden.

Da die Probanden in DaF-Kursen angeworben wurden, die allen eingeschriebenen Studenten der Technischen Universität Berlin offenstehen, zeigt sich ein recht großes Spektrum an Studienfächern. Die relativ stark vertretenden Studienrichtungen Architektur und Ingenieurwissenschaften erklären sich durch zwei fachsprachlich orientierte Deutschkurse mit Ausrichtung auf eben diese Fachbereiche. Die im einzelnen vertretenen Studienfächer zeigt Tabelle 25.

Land	Häufigkeit	Prozent
China	23	19,2
Polen	11	9,2
Brasilien	8	6,7
Spanien	7	5,8
Frankreich	7	5,8
Finnland	4	3,3
Türkei	4	3,3
Ungarn	4	3,3
Italien	3	2,5
Iran	3	2,5
Niederlande	3	2,5
Norwegen	3	2,5
Schweden	3	2,5
Russland	3	2,5
USA	3	2,5
Dänemark	2	1,7
Griechenland	2	1,7
Indonesien	2	1,7
Slowenien	2	1,7
Ukraine	2	1,7
Algerien	1	0,8

Land	Häufigkeit	Prozent
Australien	1	0,8
Belgien	1	0,8
Bulgarien	1	0,8
Chile	1	0,8
Deutschland	1	0,8
Estland	1	0,8
Großbritannien	1	0,8
Irak	1	0,8
Kanada	1	0,8
Kolumbien	1	0,8
Lettland	1	0,8
Libanon	1	0,8
Mexiko	1	0,8
Neuseeland	1	0,8
Portugal	1	0,8
Rumänien	1	0,8
Südkorea	1	0,8
Syrien	1	0,8
Vietnam	1	0,8
Zypern	1	0,8

*Tab. 24: Herkunftsland der Probanden wie angegeben auf Fragebogen (*n = *120)*

Studienfach	Häufigkeit	Prozent
Architektur	18	15
Informatik	9	7,5
Elektrotechnik/Elektroingenieurwesen	7	5,8
Prozess-, Energie- und Umweltsystemtechnik	7	5,8
Germanistik	6	5
Wirtschaftsingenieurwesen	5	4,2
Bauingenieurwesen	5	4,2
Maschinenbau	5	4,2
Chemieingenieurwesen	5	4,2
Wirtschaft	4	3,3
Produktionstechnik	4	3,3
Industrial Engineering	3	2,5
Kulturwissenschaften	2	1,7
Technischer Umweltschutz	2	1,7
Fremdsprachen	2	1,7
Landschaftsarchitektur	2	1,7
Umweltplanung	2	1,7
Mathematik	2	1,7
Soziologie	2	1,7
Städtebau	2	1,7
Deutsch als Fremdsprache	2	1,7
Informationstechnik	2	1,7
Wirtschaftsmathematik	2	1,7
Ingenieurswissenschaften	2	1,7
Volkswirtschaftslehre	1	0,8
Finanzierung und Rechnungswesen	1	0,8
Luft- und Raumfahrttechnik	1	0,8

Studienfach	Häufigkeit	Prozent
Brauerei- und Getränketechnologie	1	0,8
Jura	1	0,8
Pädagogik	1	0,8
Übersetzen/Dolmetschen	1	0,8
Automotive Systems Engineering	1	0,8
Neurowissenschaften	1	0,8
Medizintechnik	1	0,8
Physik	1	0,8
Telekom (?)	1	0,8
Management	1	0,8
Physikalische Ingenieurswissenschaften	1	0,8
Verkehrstechnik	1	0,8
Wirtschaftsinformatik	1	0,8
Strömungsmechanik	1	0,8
Lebensmitteltechnologie	1	0,8

Tab. 25: Studienfach der Probanden wie angegeben auf Fragebogen (n = *120)*

Neun Probanden haben zwei Studienfächer angegeben. Die genannten Kombinationen sind Elektrotechnik bzw. Elektroingenieurwesen/Informatik, Germanistik/Übersetzen bzw. Dolmetschen, Germanistik/Pädagogik, Fremdsprachen/fremde Kulturen, Bauingenieurwesen/Akustik, Mathematik/Informatik, Maschinenbau/Management (je einmal), Maschinenbau/Informationstechnik (zweimal).

66 Personen (55 %) gaben an, dass sie bereits zuvor einen C-Test abgelegt hatten, dem gegenüber stehen 51 Probanden (42,5 %), die noch keine Erfahrungen mit dem C-Test-Format sammeln konnten (3 Personen machten hierzu keine Angabe). 53 Probanden (44,2 %) waren der Meinung, dass der onDaF ihre Deutschkenntnisse korrekt widerspiegelt, 57 Personen (47,5 %)

waren gegensätzlicher Meinung (10 Teilnehmer machten keine Angabe). Während das Verhältnis von Misstrauen und Vertrauen in das Testformat recht ausgeglichen ist, zeigten sich die Probanden bezüglich des S-C-Tests skeptischer. Nur 35 Teilnehmer (29,2 %) kreuzten an, dass sie glauben, dass der S-C-Test ihre Deutschkenntnisse gut wiedergibt, während 80 Personen (66,7 %) dies nicht glaubten (5 Probanden machten keine Angabe).

4.3.4 Datenerhebung

Die Daten wurden im Wintersemester 2013/2014 und im Sommersemester 2014 an der Technischen Universität Berlin erhoben (vgl Kapitel 4.3.3). Die Studienteilnahme wurde allen Kursteilnehmern freigestellt. Wer kein Interesse daran hatte, an der Untersuchung teilzunehmen, konnte den Kurs für die Dauer der Datenerhebung verlassen. Von dieser Möglichkeit wurde vereinzelt Gebrauch gemacht. Die oben beschriebenen Instrumente wurden aufgrund der Rahmenbedingungen an der ZEMS stets in der gleichen Reihenfolge abgelegt. Eine Permutation der Testteile war aus organisatorischen Gründen nicht möglich. Tabelle 26 fasst den zeitlichen Ablauf der Datenerhebung zusammen.

Intensivkurse		Semesterbegleitende Kurse	
Tag 1 onDaF (40 min)	Woche vor Kursbeginn	**Tag 1** onDaF (40 min)	Möglichst früh nach Kursbeginn
Tag 2 S-C-Test (9 min) Fragebogen (10 min) Hörverstehen (30 min)	erste Kurswoche	**Tag 2** S-C-Test (9 min) Fragebogen (10 min) Hörverstehen (30 min)	eine Woche später
Tag 3 Sprechen (15 min Vorbereitung + 15 min Prüfung)	1–3 Tage später	**Tag 3** Sprechen (15 min Vorbereitung + 15 min Prüfung)	eine weitere Woche später

Tab. 26: Zeitlicher Ablauf der Datenerhebung

Allen Probanden wurde vor Beginn der Datenerhebung erklärt, dass es sich um eine wissenschaftliche Studie handelt, bei der ihre Leistung in den einzelnen Testteilen und Antworten im Fragebogen nicht bewertet wird. Darüber hinaus wurde ihnen Anonymität zugesichert, die über die Verwendung eines durch die Probanden selbst erstellten Teilnehmercodes sichergestellt werden konnte. Dieses Vorgehen bietet darüber hinaus die Möglichkeit, die einzelnen Testteile ein und derselben Person zuordnen zu können. Etwaige Rückfragen der Testteilnehmer wurden beantwortet.

Testtag 1: Während in den Intensivkursen die Ergebnisse des onDaF bereits vorlagen, da dieser von der ZEMS zur Einstufung in das richtige Kursniveau genutzt wird, musste der onDaF in den semesterbegleitenden Kursen noch erhoben werden. Da der ZEMS die onDaF-Ergebnisse der Intensivkursteilnehmer vorlagen, konnten die Probanden ihr Ergebnis erfragen, sofern sich ein Proband selbst nicht mehr daran erinnern konnte. Auf diese Weise wurde sichergestellt, dass der Testleiterin die Namen der Teilnehmer unbekannt blieben. In den semesterbegleitenden Kursen wurden die Teilnehmer zu Beginn des ersten Testtags darauf hingewiesen, dass sie sich die im onDaF erreichte Punktzahl bis zur kommenden Woche merken und notieren sollten, da diese dann in einen Fragebogen einzutragen sei. Auch hier konnte im Falle des Vergessens ein Ergebnis über die ZEMS durch die Probanden selbst erfragt werden.

Testtag 2: Der zweite Testtag begann mit dem S-C-Test. Alle Probanden waren zu diesem Zeitpunkt bereits durch das Ablegen des onDaF mit dem C-Test-Format vertraut. Die Probanden wurden darauf hingewiesen, dass ihnen bei den folgenden C-Tests nur sehr wenig Zeit zur Verfügung steht und dass es kaum möglich sei, die Texte in dieser Zeit vollständig zu bearbeiten. Hierdurch sollte verhindert werden, dass durch den hohen Zeitdruck Frust bei den Probanden aufkommt und sie sich nicht mehr (bestmöglich) anstrengen. Tatsächlich schien viele Probanden die sehr

kurze Bearbeitungszeit eher zu erheitern, worauf ein Kichern und Lachen nach dem STOP-Signal des ersten Textes hindeutete. Eine Probandin meldete am Ende der S-C-Test-Erhebung sogar zurück, dass der enorme Zeitdruck ihren Ehrgeiz geweckt habe. Jedoch darf man hier nicht unerwähnt lassen, dass es sich für die Probanden um eine absolute *low stakes*-Situation handelte und sie keine Konsequenzen etwaigen schlechten Abschneidens zu befürchten hatten.

Damit sich die Teilnehmer nach dem S-C-Test etwas erholen konnten, wurde als Nächstes der Fragebogen ausgeteilt. Gemeinsam mit dem Fragebogen bekamen alle Probanden eine kleine Tüte Gummibärchen, welchen eine motivierende Funktion zukam. Zwar waren für den Fragebogen zehn Minuten angesetzt, jedoch wurde in der Praxis so lange gewartet, bis wirklich alle Teilnehmer mit dem Ausfüllen fertig waren. Die ungefähre Bearbeitungszeit von zehn Minuten war jedoch wichtig, damit die Kursleiter in etwa abschätzen konnten, wann sie die Datenerhebung am besten in ihren Unterricht integrieren. Der letzte Aufgabenteil am zweiten Testtag war das Hörverstehen. Die Audio-Dateien wurden je nach Unterrichtsraum von einem mit fest installierten Boxen verbundenen PC oder mit einem gewöhnlichen Abspielgerät von der CD wiedergegeben. Zu Beginn des Prüfungssatzes ertönt eine Melodie sowie die Ansage, dass es sich um einen Modellsatz des Goethe-Zertifikats B2 handelt. Diese einführenden und nicht zur eigentlichen Prüfung gehörenden Sätze wurden dazu genutzt, die Lautstärke so einzustellen, dass alle Teilnehmer rückmeldeten, gut hören zu können. Bei der weiteren Durchführung des Hörverstehens wurde sich an die vom Goethe-Institut vorgeschriebenen Durchführungsbedingungen gehalten. Die Teilnehmer bearbeiteten die beiden Aufgaben in der vorgegebenen Reihenfolge und erhielten jeweils Zeit, um sich die Aufgaben durchzulesen.

Testtag 3: Die mündlichen Prüfungen wurden von zwei mit den Prüfungen des Goethe-Instituts vertrauten und in der Abnahme derartiger Prüfungen erfahrenen Prüferinnen abgenommen. Die vom Goethe-Institut vorgegebenen Durchführungsbestimmungen wurden streng befolgt. Die Probanden erhielten zunächst das Aufgabenblatt mit beiden Aufgabenteilen und konnten sich in einem ruhigen Raum unter Aufsicht 15 Minuten lang auf die Prüfung vorbereiten. Im Anschluss wurden sie direkt ins Prüfungszimmer gebracht, wo die Prüfung abgenommen wurde. Die Prüfungsgespräche wurden auf Band aufgenommen.[32] Die Probanden wurden über die geplante Aufnahme vor Beginn der Prüfung informiert und gaben ausnahmslos ihre mündliche Zustimmung.

Für die mündliche Prüfung wurden mehrere Termine angeboten, für die die Probanden sich anmelden konnten. Pro *Time Slot* konnten sich zwei Personen (unter Verwendung ihres Teilnehmercodes) eintragen. Idealerweise wird der Testteil zum mündlichen Ausdrucks des Goethe-Zertifikats B2 als Paarprüfung abgenommen. Um eine möglichst hohe Vergleichbarkeit zu gewährleisten wurde bei den Einzelprüfungen der Part des zweiten Gesprächspartners bei Aufgabe 2 stets von der gleichen Prüferin übernommen, die über mehr Unterrichtserfahrung auf diesem Sprachniveau verfügte.

Die **Auswertung** des onDaF erfolgt automatisch. Das System gibt neben der erreichten Punktzahl von maximal 160 erreichbaren Punkten auch eine Information über das erreichte Niveau gemäß GER (EUROPARAT 2001) aus. Die *Cut Scores* des onDaF zum Zeitpunkt der Datenerhebung[33] zeigt Tabelle 27.

32 Zwei Tondateien brechen nach kurzer Zeit ab, da das Aufnahmegerät bedauerlicherweise zweimal den Dienst versagt hat.

Punkte	Niveaustufe
< 36	unter A2
36 bis 59	A2
60 bis 95	B1
96 bis 129	B2
> 129	C1 oder höher

Tab. 27: Cut Scores *des onDaF zum Zeitpunkt der Datenerhebung*

Bei der Auswertung des S-C-Tests wurde die vom TestDaF-Institut zur Verfügung gestellte Lösungsschablone genutzt. Wie auch beim onDaF bekommt ein Teilnehmer einen Punkt pro korrekter Original- oder Alternativlösung. Rechtschreibfehler werden geahndet, während alte und neue Rechtschreibung parallel gelten. Des Weiteren wurde bei der Überführung der Testergebnisse in SPSS eine unterschiedliche Kodierung für falsch bzw. gar nicht ausgefüllte Lücken vorgenommen. Auf diese Weise lässt sich später erahnen, ob ein Text fehlerhaft bearbeitet wurde, ob er ggf. abgebrochen wurde, oder ob die Lücken einfach nicht korrekt ausgefüllt werden konnten.

Die Auswertung der Hörverstehensaufgaben erfolgte gemäß der dazugehörigen Lösungsschablone des Goethe-Zertifikats B2. Bei Aufgabe 1 kam es darauf an, dass die gegebene Kurzantwort inhaltlich stimmte. Etwaige morphosyntaktische oder orthographische Fehler wurden hier nicht sanktioniert. Bei Aufgabe 2 des Hörverstehens lag eine eindeutige Antwortschablone vor, die jeweils die Bestantwort angab.

33 Da der Textpool des onDaF bzw. jetzt onSET stetig erweitert wird und immer neue Texte kalibriert werden, kann es zu leichten Verschiebungen der *Cut Scores* kommen. Die hier angegebenen Punktwerte stammen daher nicht aus offizieller Quelle, sondern waren aufgrund der Punkteverteilung der Testkandidaten ablesbar.

Für die Bewertung der beiden Sprechaufgaben lag ein Kriterienraster des Goethe-Instituts vor. Beide Aufgaben werden getrennt voneinander bewertet. Das Kriterienraster war beiden Prüferinnen lange vor der ersten Datenerhebung bekannt. Zudem existiert auf den Seiten des Goethe-Instituts die Videoaufnahme einer Musterprüfung für das Goethe-Zertifikat B2 (vgl. URL 6: Goethe-Institut), das vorbereitend auf die mündlichen Testaufgaben geschaut wurde. Beide Prüferinnen hatten zudem Erfahrung im Abnehmen von Goethe-Zertifikaten und hatten in der Vergangenheit an Prüferschulungen des Goethe-Instituts teilgenommen.

Das Bewertungsraster erfasst für jede der beiden Aufgaben fünf Aspekte. Bei Aufgabe 1 (monologisches Sprechen) bestehen diese aus der qualitativen und quantitativen Erfüllung der Aufgabenstellung, Kohärenz und Flüssigkeit der Äußerung, dem Ausdrucksvermögen, der morphosyntaktischen Korrektheit sowie Aussprache und Intonation. Für jeden dieser fünf Aspekte können maximal 2,5 Punkte vergeben werden. Das Bewertungsraster der Aufgabe 2 (dialogische Aufgabe) weicht nur geringfügig ab. Anders als bei Aufgabe 1 kommt es hier beim Aspekt der Aufgabenbewältigung auch auf die Diskussionsfähigkeit an.

4.4 Ergebnisse

Die Daten wurden mit IBM SPSS Statistics, Versionen 22 und 24 ausgewertet. Der Entscheidung über geeignete Analysemethoden gehen eine deskriptive Analyse der Daten sowie Tests auf Normalverteilung voraus. Die zur Berechnung der einzelnen Testteile herangezogenen Subgruppen von Probanden ergeben sich aus der folgenden Darstellung (Abb. 10):

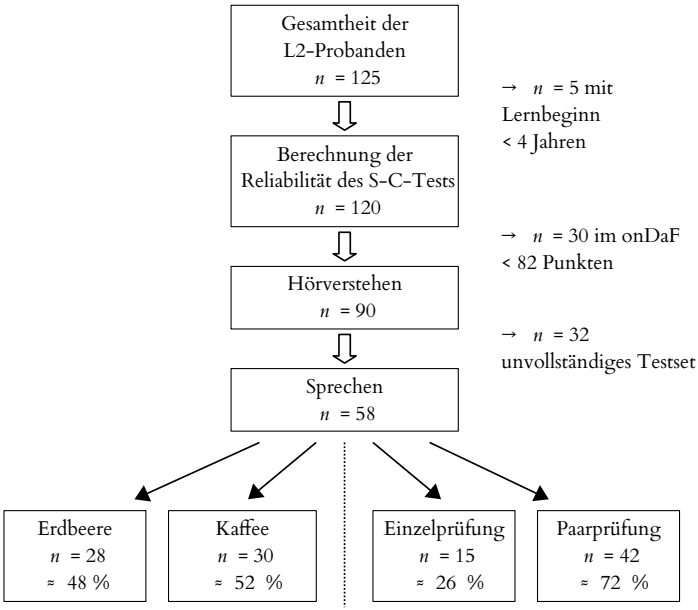


Abb. 10: Flowdiagramm aller fremdsprachlichen Probanden

4.4.1 Deskriptive Analyse

Im Folgenden werden zunächst die Ergebnisse der Probanden bei allen vier Testteilen deskriptiv analysiert.

4.4.1.1 C-Test und S-C-Test

Da es sich bei dem hier eingesetzten S-C-Test um eine Papier-und-Bleistift-Version des onDaF handelt, ist das mögliche Punktespektrum bei beiden Tests identisch und reicht von jeweils 0 bis maximal 160 Punkten. Die folgenden Boxplots (Abb. 11) zeigen die von den Probanden erreichten Punkte in den Testteilen onDaF und S-C-Test.

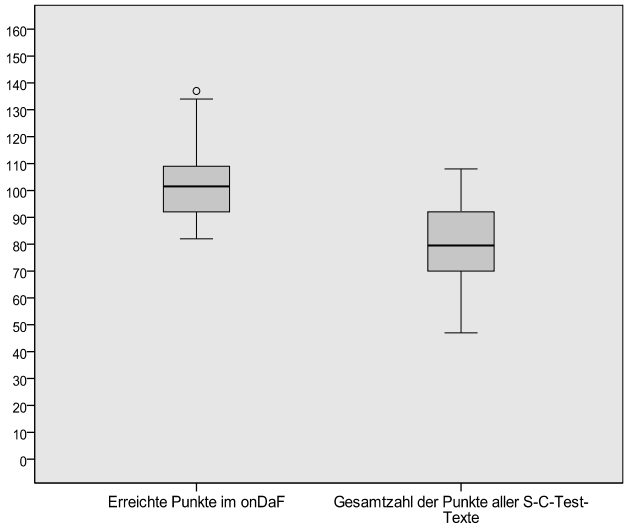


Abb. 11: Erreichte Punkte im onDaF und im S-C-Test (n = *90)*

Bei der graphischen Gegenüberstellung der Ergebnisse aus beiden C-Test-Varianten wird deutlich, dass die erreichten Punkte beim S-C-Test erwartungsgemäß deutlich niedriger ausfallen als beim onDaF. Beim onDaF erreichten die Probanden zwischen 82 (min) und 137 (max) Punkte (R = 55). Hierzu ist anzumerken, dass die Untergrenze von 82 durch Berechnung einer Vertrauensgrenze beim *Cut Score* zwischen B1- und B2-Niveau festgelegt worden war (vgl. Kapitel 4.4.3.2). Der Median des onDaF liegt bei $\tilde{x}$ = 101,5 und das arithmetische Mittel beträgt $\bar{x}$ = 102,67 (SD = 12,72). Beim S-C-Test erreichten die Testteilnehmer zwischen 47 (min) und 103 (max) Punkten (R = 61). Der Median liegt mit $\tilde{x}$ = 79,5 sehr nah am arithmetischen Mittel von $\bar{x}$ = 80,88 (SD = 14,17). Der Standardfehler für die beiden Testversionen ist mit SE = 1,18 für den onDaF und SE = 1,58 für den S-C-Test ähnlich. Tabelle 28 fasst diese Ergebnisse noch einmal zusammen.

	min	max	R	x̃	x̄	SD	SE
onDaF	82	137	55	101,5	102,67	12,72	1,78
S-C-Test	47	103	61	79,5	80,88	14,17	1,58

Tab. 28: Deskriptive Analysen onDaF und S-C-Test

Für die Berechnung der Itemschwierigkeit existieren unterschiedliche Formeln, je nachdem, ob es sich um einen Niveau- oder einen Geschwindigkeitstest handelt. Während bei der Formel für Niveautests lediglich der Mittelwert der korrekten Antworten herangezogen wird, berücksichtigt die Berechnung der Itemschwierigkeit für Geschwindigkeitstests die Tatsache, dass aufgrund des Zeitdrucks Aufgaben unbearbeitet bleiben, so dass hier zwischen falsch gelösten und gar nicht gelösten Items unterschieden wird (vgl. MOOSBRUGGER & KELAVA 2012: 78).

Dies scheint zunächst für den S-C-Test eine geeignete Vorgehensweise zu sein. Da beim C-Test jedoch jeder Text als ein Super-Item betrachtet wird, erfolgt die Berechnung der Item-Schwierigkeit mit der Formel für mehrstufige Items (vgl. BORTZ & DÖRING 2006: 219). Diese setzt die Anzahl der von allen Testteilnehmern erreichten Punkte bei einem Item ins Verhältnis zu den durch alle Probanden maximal erreichbaren Punkten:

$$p_i = \frac{\sum_{m=1}^{n} x_{\mathrm{im}}}{k_i \cdot n}$$

Die ermittelten Werte liegen stets zwischen 0 und 1. Je größer der gefundene Wert p ist, desto einfacher ist das Item. Laut BORTZ & DÖRING (2006: 219) sind generell Items mit einer mittleren Schwierigkeit mit Werten von p zwischen 0,2 und 0,8 erstrebenswert. Ist ein Item zu leicht und kann von (nahezu) allen Teilnehmern gelöst werden, ist es zum Differenzieren der Probanden ebenso wenig geeignet wie ein Item, dass aufgrund

einer hohen Schwierigkeit von keiner Testperson korrekt gelöst werden kann.

Unter Anwendung dieser Formel ergeben sich folgende Werte für den S-C-Test (n = 120) (vgl. Tab. 29):

Text 1	0,7487
Text 2	0,5725
Text 3	0,4533
Text 4	0,4762
Text 5	0,4266
Text 6	0,3533
Text 7	0,4075
Text 8	0,3154
gesamt	0,4692

Tab. 29: P_i die Texte des S-C-Tests

Für den onDaF ist aufgrund der computerisierten Gesamtauswertung keine Berechnung der Schwierigkeit für die einzelnen Texte möglich, sondern nur für das gesamte Testset (n = 120) (vgl. Tab. 30):

onDaF	0,5878

Tab. 30: P_i für den onDaF

Es zeigt sich, dass die Schwierigkeit des Testsets mit Zeitdruck erwartungsgemäß höher ist als beim onDaF mit großzügig bemessener Bearbeitungszeit. Zugleich zeigen die Schwierigkeitsindizes jedoch auch, dass alle Texte des S-C-Tests eine mittlere Schwierigkeit gemäß BORTZ und DÖRING (2006) aufweisen, so dass die Texte durch die Geschwindigkeitskomponente nicht zu schwierig für die Testteilnehmer wurden.

Die Trennschärfe gibt Auskunft darüber, wie gut ein Item, d. h. im Fall des C-Tests ein Text (Super-Item) zwischen den Probanden differenziert. Die Trennschärfe bewegt sich zwischen −1 und 1, wobei ein Wert von 0 bedeutet, dass ein Item gar nicht differenziert (vgl. MOOSBRUGGER & KELAVA 2012: 86). Tabelle 31 gibt einen Überblick über die für den S-C-Test ermittelten Trennschärfen. Diese werden als korrigierte Item-Skala-Korrelation berechnet, d. h. als Korrelation des jeweiligen C-Test-Texts mit dem gesamten Test-Set abzüglich des betreffenden C-Tests selbst.

Text 1	0,559
Text 2	0,749
Text 3	0,660
Text 4	0,663
Text 5	0,634
Text 6	0,676
Text 7	0,553
Text 8	0,676

Tab. 31: Trennschärfen für den S-C-Test

Im Allgemeinen werden Trennschärfen zwischen 0,4 und 0,7 als gut eingestuft (vgl. MOOSBRUGGER & KELAVA 2012: 86), was von allen C-Test-Texten erreicht und von Text 2, der einen noch höheren Wert für die Trennschärfe aufweist, übertroffen wird. Es lässt sich festhalten, dass das hier eingesetzte C-Test-Set auch unter Geschwindigkeitsbedingungen hinreichend zwischen den Probanden differenziert.

4.4.1.2 Hörverstehen und Sprechen

Sowohl beim Hörverstehenstest als auch bei den Aufgaben zum mündlichen Ausdruck konnten maximal 25 Punkte erreicht werden. Abbildung 12 liefert einen optischen Eindruck der Verteilung der Daten aus diesen beiden Testteilen.

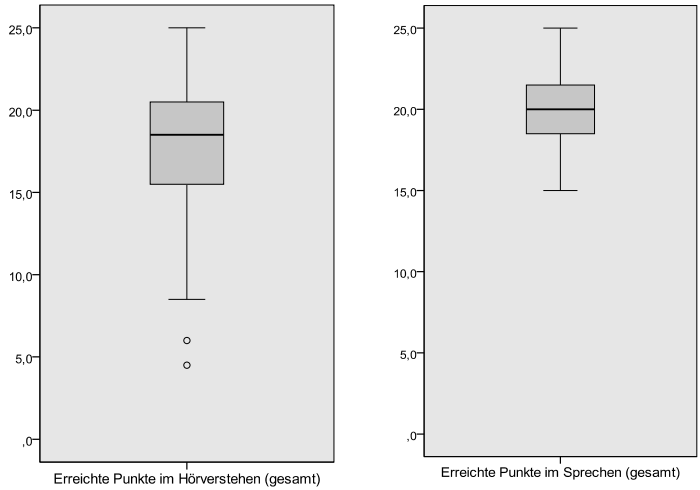


Abb. 12: Erreichte Punkte Hörverstehen (n = 90) und im Sprechen (n = 58)

Beim Betrachten der Boxplots fällt auf, dass die Ergebnisse beider Testteile im oberen Bereich des möglichen Punktespektrums liegen. Beim Hörverstehen erreichten die Probanden zwischen 4,5 (min) und 25 (max) Punkte (R = 20,5). Der Median liegt bei $\tilde{x}$ = 18,5 und das arithmetische Mittel beträgt $\bar{x}$ = 17,87 (SD = 4,14). Beim Sprechen ist dieser Effekt noch deutlicher: Es wurden zwischen 15 (min) und 25 (max) Punkte erreicht (R = 10). Der Median liegt bei $\tilde{x}$ = 20 und das arithmetische Mittel bei $\bar{x}$ = 19,97 (SD = 2,11). Die Streuung der Daten um das arithmetische Mittel ist beim Sprechen deutlich geringer als beim Hörverstehen. Der Minimalwert liegt

hier höher und die Spannweite ist deutlich kleiner, so dass man beim mündlichen Ausdruck durchaus von einem Deckeneffekt sprechen kann. Der Standardfehler liegt beim Hörverstehen (SE = 0,44) höher als beim Sprechen (SE = 0,28). Eine Zusammenfassung dieser Ergebnisse liefert Tabelle 32.

	min	max	R	$\tilde{x}$	$\bar{x}$	SD	SE
Hörverstehen	4,5	25	20,5	18,5	17,87	4,14	0,44
Sprechen	15	25	10	20	19,97	2,11	0,28

Tab. 32: Deskriptive Statistik Hörverstehen und Sprechen

Die Lagemaße für das Hörverstehen und den mündlichen Ausdruck sind auffällig, insbesondere vor dem Hintergrund, dass auch Lerner des oberen B1-Niveaus in die Berechnung mit einbezogen wurden. Beim Hörverstehen liegen der hohe Median und das hohe arithmetisches Mittel vermutlich zum Teil in der sich durch das Multiple-Choice-Format der zweiten Aufgabe ergebenden Ratewahrscheinlichkeit begründet. Beim mündlichen Ausdruck wird das Punktespektrum nur zu einem geringen Maß ausgefüllt, da zu Beginn der Datenerhebungsphase eine gewisse Selbstauslese der Probanden erkennbar war. Während zahlreiche Probanden die ersten beiden Testtage mitmachten, erschienen nur sehr wenige Testpersonen am dritten Testtag zur Sprechaufgabe. Diejenigen, die kamen, schnitten relativ gut ab. Erst nach Einführung einer Aufwandsentschädigung von 6,- Euro, die die Probanden nach erfolgter Teilnahme an der Sprechaufgabe bar erhielten, schienen sich auch weniger geübte Sprecher zum dritten Testtag einzufinden. Empirisch nachweisbar ist dieser sich den beiden Prüferinnen bietende Eindruck jedoch nicht. Die guten Ergebnisse in der Hörverstehens- und Sprechaufgabe könnten jedoch auch darauf zurückzuführen sein, dass sich die Probanden zum Zeitpunkt der Datenerhebung in Deutschland befanden und an einem Erasmus-Semester oder ähnlichem Programm teilnahmen.

Exkurs: Sprechangst und Teilnahme an der Sprechaufgabe

Die Hypothese, dass Probanden, die das Fragebogenitem „Gibt es Situationen, in denen Sie sich unwohl fühlen, Deutsch zu sprechen" mit „ja" beantwortet haben, weniger häufig an der Sprechaufgabe teilgenommen hätten, konnte rechnerisch nicht bestätigt werden.

Von den 69 Probanden, die das Item mit „ja" beantwortet haben, nahmen 46 Personen an der Sprechaufgabe teil (67 %). Unter den 20 Probanden, die das Item mit „nein" beantwortet haben, nahmen 12 Personen teil (60 %). Ein Vergleich der Mittelwerte beim onDaF und S-C-Test zeigt Folgendes: Die beiden Teilgruppen von Probanden, die auch an der Sprechaufgabe teilnahmen und diejenigen, die nicht daran teilnahmen unterscheiden sich beim onDaF kaum (mit Sprechaufgabe (n = 58): $\tilde{x}$ = 104; $\bar{x}$ = 103,72; SD = 12,74; ohne Sprechaufgabe (n = 32): $\tilde{x}$ = 99; $\bar{x}$ = 100,75; SD = 12,67), was auf ein ähnliches Sprachniveau hinweist. Im Gegensatz dazu zeigen sich jedoch beim S-C-Test deutlichere Unterschiede in den Mittelwerten (mit Sprechaufgabe (n = 58): $\tilde{x}$ = 82,5; $\bar{x}$ = 83,81; SD = 13; ohne Sprechaufgabe (n = 32): $\tilde{x}$ = 71,5; $\bar{x}$ = 75,56; SD = 14,85).

Es hat demzufolge den Anschein, dass Probanden, die sich mit dem S-C-Test-Format schwer taten, nicht zum nächsten Testtag erschienen sind.

Die folgenden Boxplots (Abb. 13) zeigen das Abschneiden der Probanden (n = 58) bei der monologischen Aufgabe 1 und der dialogischen Aufgabe 2 im Vergleich zueinander.

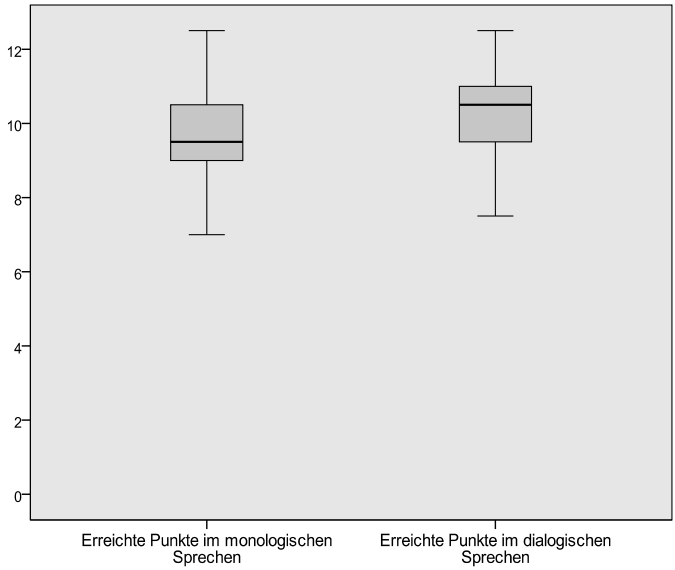


Abb. 13: Erreichte Punkte bei Aufgabe 1 und Aufgabe 2 des mündlichen Ausdrucks (n = *58)*

Wie zu sehen ist, sind diese beiden Verteilungen recht ähnlich, lediglich der Median liegt bei Aufgabe 2 etwas höher ($\tilde{x}$ = 10,5) als bei Aufgabe 1 ($\tilde{x}$ = 9,5). Das arithmetische Mittel beträgt für Aufgabe 1 $\bar{x}$ = 9,78 und für Aufgabe 2 $\bar{x}$ = 10,2. Die Standardabweichung beider Testteile liegt mit SD = 1,15 für Aufgabe 1 und SD = 1,09 für Aufgabe 2 nah beieinander.

Die monologische Aufgabe 1 zum mündlichen Ausdruck bestand wie in Kapitel 4.3.1.2 beschrieben aus zwei unterschiedlichen Stimulationstexten. Abbildung 14 zeigt das Abschneiden der Probanden bei dieser Aufgabe in Abhängigkeit des ihnen zugewiesenen Textes. 28 Personen bearbeiteten den Impulstext „Erdbeere“ und 30 Personen das Thema „Kaffee“.

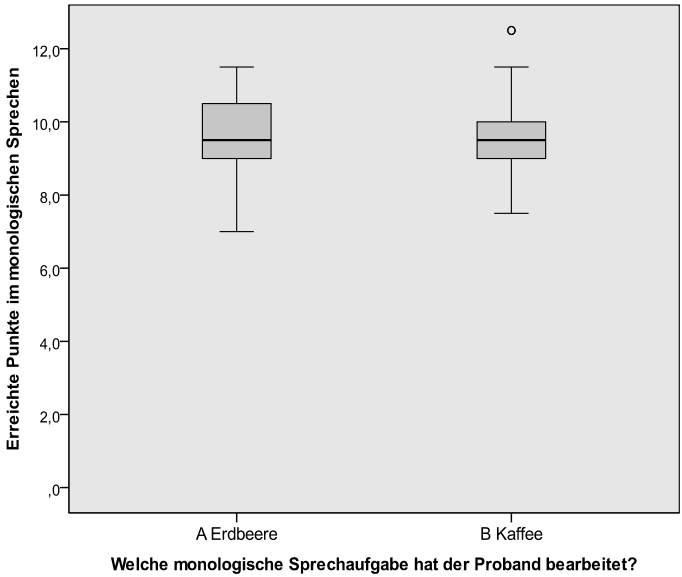


*Abb. 14: Erreichte Punkte bei Aufgabe 1 nach Stimulustext (*n *„Erdbeere“ = 28;* n *„Kaffee“ = 30)*

Aus den Boxplots wird ersichtlich, dass die Daten der Erdbeer-Gruppe (min = 7; max = 11,5; R = 4,5; SD = 1,07) und der Kaffee-Gruppe (min = 7,5; max = 12,5; R = 5; SD = 1,24) ähnlich verteilt sind. Das arithmetische Mittel beider Gruppen unterscheidet sich mit $\bar{x} = 9{,}73$ (Erdbeere) und $\bar{x} = 9{,}82$ (Kaffee) jedoch kaum. Der Median ist mit $\tilde{x} = 9{,}5$ für beide Gruppen identisch.

Die zweite Aufgabe des Testteils zum mündlichen Ausdruck aus dem Goethe-Zertifikat B2 ist als dialogische Paarprüfung angelegt. Zwei Probanden sollten sich über ein geeignetes Foto für einen Werbeprospekt einig werden. Die Termine zur Sprechaufgabe wurden mit größtmöglicher Flexibilität vereinbart, um den Probanden so weit wie möglich entgegen-

zukommen. Dennoch war es nicht möglich, alle Probanden in Zweierteams zu prüfen. In dem Fall, dass zu einem Testtermin nur ein Proband erschien, wurde der zweite Diskussionspartner gemäß den Durchführungsbestimmungen des Goethe-Instituts von einer Prüferin übernommen. Diese war bei allen Einzelprüfungen dieselbe, und zwar diejenige, die über mehr Erfahrung im praktischen DaF-Unterricht verfügt. Dieses Vorgehen sollte die Vergleichbarkeit erhöhen. Die folgenden Boxplots (Abb. 15) zeigen das Abschneiden der Probanden bei der dialogischen Aufgabe des Subtests zum Sprechen in Abhängigkeit von der Prüfung als Einzelprüfung (n = 15) oder Paarprüfung (n = 42).[34]

Ein Vergleich der Ergebnisse aus den Einzel- und Paarprüfungen zeigen kaum Unterschiede zwischen diesen beiden Gruppen bei Median ($\tilde{x}$ = 10,5 (Einzelprüfung); $\tilde{x}$ = 10,25 (Paarprüfung)) und arithmetischem Mittelwert ($\bar{x}$ = 10,43 (Einzelprüfung); $\bar{x}$ = 10,11 (Paarprüfung)). Die Streuung beider Gruppen ist mit Werten von R = 3,5 (Einzelprüfung) und R = 4,5 (Paarprüfung) für die Spannweite und SD = 1,18 (Einzelprüfung) und SD = 1,07 (Paarprüfung) für die Standardabweichung ebenfalls ähnlich. Lediglich der niedrigste gemessene Wert bei der Paarprüfung liegt mit min = 7,5 ein wenig unterhalb des Minimums von min = 9 bei der Einzelprüfung. Die Höchstpunktzahl von max = 12,5 Punkten wurde nur in der Einzelprüfung erreicht, der höchste Wert der Paarprüfung liegt bei max = 12.

Tabelle 33 (S. 198) fasst alle Ergebnisse der deskriptiven Analysen der verwendeten Testinstrumente und ggf. ihrer Subtests noch einmal übersichtlich zusammen.

34 Bei einem Probanden fehlt aufgrund einer technisch bedingten Aufnahmepanne die Angabe zur Paar- oder Einzelprüfung. Daher werden hier nur 57 Probanden aufgeführt.

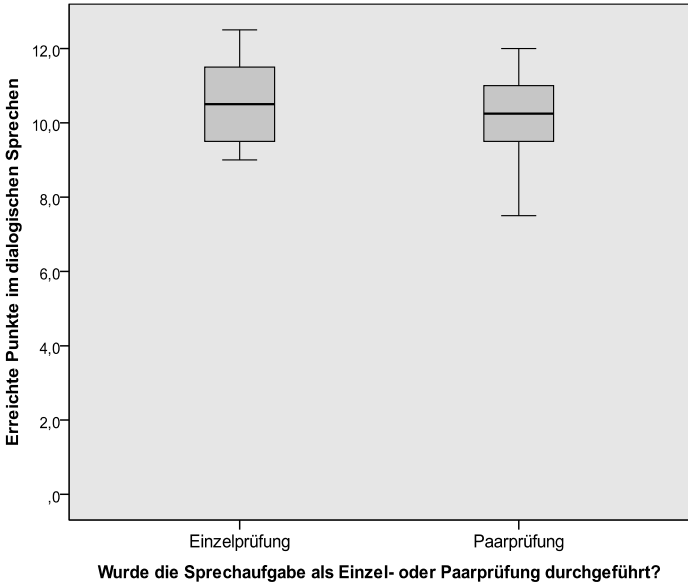


*Abb. 15: Erreichte Punkte bei Aufgabe 2 nach Einzel- und Paarprüfung (*n = *58)*

Nach dieser deskriptiven Auswertung der Testergebnisse wird im Folgenden schließende Statistik genutzt, um die in Kapitel 4.1 ausgeführten Forschungsfragen zu beantworten. Vorab soll die Berechnung der Reliabilität Auskunft über die Zuverlässigkeit des S-C-Tests geben. Da die Reliabilität eine notwendige Voraussetzung für ein valides Testinstrument ist, würde eine niedrige Reliabilität weitere Berechnungen bzgl. etwaiger Korrelationen des S-C-Tests mit anderen Kriterien erübrigen.

	n	min	max	R	$\tilde{x}$	$\bar{x}$	SD	SE
onDaF	90	82	137	55	101,5	120,67	12,72	1,78
S-C-Test	90	47	103	61	79,5	80,88	14,17	1,58
Hörverstehen	90	4,5	25	20,5	18,5	17,87	4,14	0,44
Sprechen (gesamt)	58	15	25	10	20	19,97	2,11	0,28
Monologisches Sprechen (gesamt)	58	7	12,5	5,5	9,5	9,78	1,15	0,15
Monologisches Sprechen (Erdbeere)	28	7	11,5	4,5	9,5	9,73	1,07	0,2
Monologisches Sprechen (Kaffee)	30	7,5	12,5	5	9,5	9,82	1,24	0,23
Dialogisches Sprechen (gesamt)	58	7,5	12,5	5	10,5	10,2	1,09	0,14
Dialogisches Sprechen (Einzelprüfung)	15	9	12,5	3,5	10,25	10,43	1,18	0,29
Dialogisches Sprechen (Paarprüfung)	42	7,5	12	4,5	10,25	10,11	1,07	0,18

Tab. 33: Übersicht über deskriptive Analysen aller Subtests

4.4.2 Verteilung der Daten

Für die weiteren Berechnungen mit Verfahren der Inferenzstatistik zur Beantwortung der Forschungsfragen, muss zunächst überprüft werden, ob die vorliegenden Datensätze normalverteilt sind, um entscheiden zu können, welche Rechenverfahren geeignet und zulässig sind. Um eine möglichst zuverlässige Aussage darüber treffen zu können, ist es sinnvoll, sich nicht allein auf graphische Darstellungen oder statistische Tests zu verlassen, sondern beide Verfahren zu kombinieren (vgl. LARSEN-HALL 2010: 74 und 84 f.). Unter den möglichen statistischen Verfahren zur Überprüfung der Verteilung fällt die Wahl auf den Kolmogorov-Smirnov-Test. Dieser wird im Folgenden auf alle vier Testteile angewendet und mit dem opti-

schen Eindruck der Histogramme und Q-Q-Diagramme kombiniert. Die Verwendung zweier graphischer Darstellungen ist sinnvoll, da der optische Eindruck des Histogramms bekanntlich je nach Festlegung der Inkremente bei der Klasseneinteilung sehr unterschiedlich ausfallen kann.

Der Kolmogorov-Smirnov-Test nimmt das Vorliegen einer Normalverteilung an und überprüft, ob ein Datensatz mit hoher Wahrscheinlichkeit davon abweicht. Auf dem Signifikanzniveau von 5 % wird ab einem Wert von $p < 0{,}05$ die Nullhypothese, d. h. die Annahme, dass eine Normalverteilung der Daten vorliegt, verworfen. Der Kolmogorov-Smirnov-Test eignet sich auch für kleinere Datenmengen (vgl. JANSSEN & LAATZ 2013: 495). Ein Q-Q-Diagramm vergleicht viele verschiedene Quantile des vorliegenden Datensatzes mit denen der Normalverteilung. Abweichungen von einer geraden Linie deuten somit eine Abweichung von der Normalverteilung an (vgl. LARSEN-HALL 2010: 82). Die folgenden Abbildungen zeigen ein Histogramm (Abb. 16) sowie ein Q-Q-Diagramm (Abb. 17) für die Daten des onDaF.

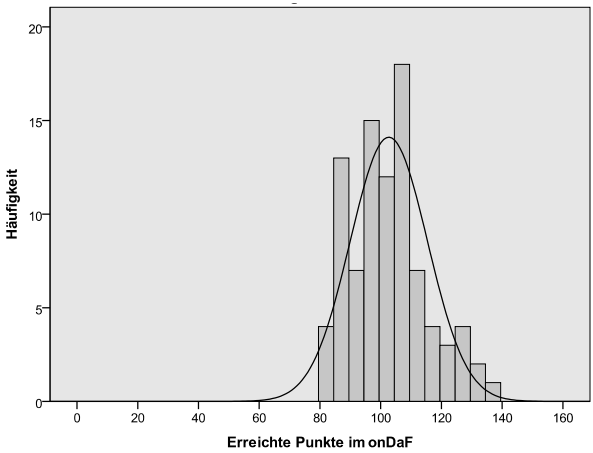


Abb. 16: Histogramm onDaF (n= 90)

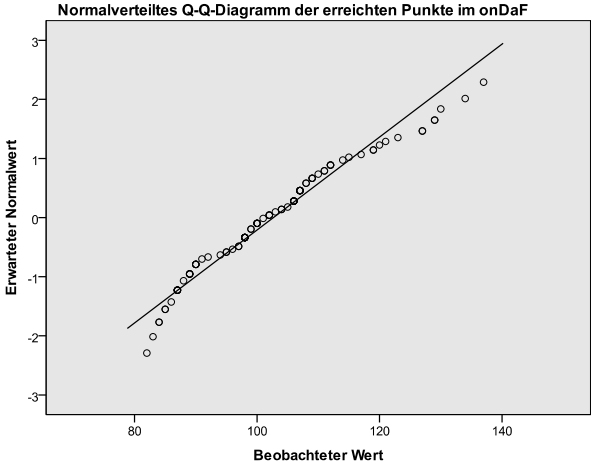


Abb. 17: Q-Q-Diagramm onDaF (n = 90)

Das Histogramm der Ergebnisse beim onDaF legt bereits die Vermutung nahe, dass die Daten von der Normalverteilung abweichen. Dies deckt sich mit den Ergebnissen des Kolmogorov-Smirnov-Tests, der einen Wert von 0,078 liefert (Sig. = 0,2). Dieses Ergebnis hängt vermutlich damit zusammen, dass die festgelegte Punktgrenze von 82 onDaF-Punkten zur Teilnahme an der Studie zu einem abrupten Abfall am linken Ende des Datenspektrums führt.

Wenngleich das Vorliegen eines nicht normalverteilten Datensatzes in der Konsequenz bereits dazu führt, dass auf nicht-parametrische Rechenverfahren zurückgegriffen werden muss, werden auch die Daten zum S-C-Test, dem Hörverstehen und der Sprechaufgabe auf ihre Verteilung hin untersucht. Auch für den S-C-Test sollen ein Histogramm (Abb. 18) und ein Q-Q-Diagramm (Abb. 19) einen graphischen Eindruck von der Verteilung der Daten liefern.

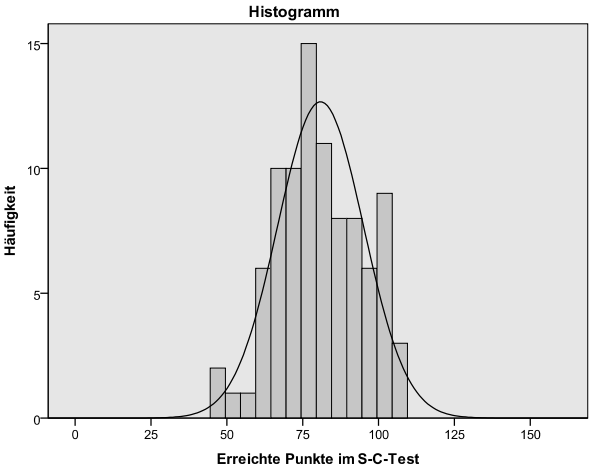


Abb. 18: Histogramm S-C-Test (n = 90)

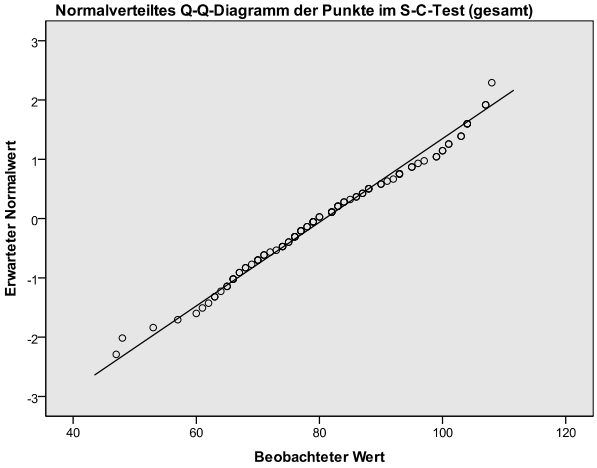


Abb. 19: Q-Q-Diagramm S-C-Test (n = 90)

Das Histogramm des S-C-Tests scheint einer Normalverteilung recht nahe zu kommen. Die Abweichungen von der Geraden beim Q-Q-Diagramm fallen deutlich geringer aus als beim onDaF.

Der Wert von 0,55 (Sig. 0,2) des Kolmogorov-Smirnov-Tests zeigt jedoch eine (wenngleich geringe) Abweichung von der Normalverteilung an.

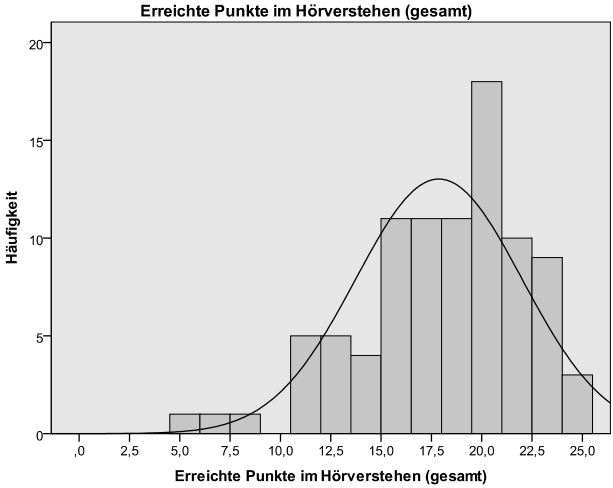


Abb. 20: Histogramm Hörverstehen (n = 90)

Im Folgenden wird die Verteilung der Daten beim Hörverstehenstest untersucht. Das Histogramm der Hörverstehensergebnisse (Abb. 20) scheint der Normalverteilungskurve nicht zu folgen. Zudem zeigen sich im Q-Q-Diagramm (Abb. 21) deutliche Abweichungen von der Geraden, insbesondere im unteren Wertebereich.

Eine rechnerische Überprüfung über den Kolmogorov-Smirnov-Test liefert einen Wert von 0,098 (Sig. 0,034). Es muss folglich von einer signifikanten Abweichung von der Normalverteilung ausgegangen werden.

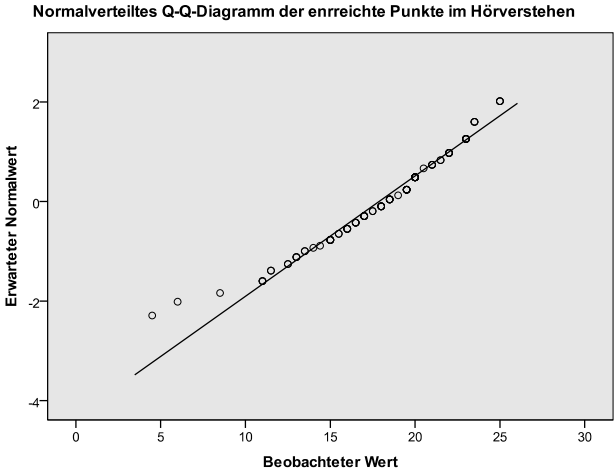


Abb. 21: Q-Q-Diagramm Hörverstehen (n = 90)

Da in der weiteren Analyse die monologische und dialogische Aufgabe des Testteils zum mündlichen Ausdruck getrennt gemeinsam sowie getrennt voneinander untersucht werden sollen, wird der Test auf Normalverteilung für alle drei Varianten durchgeführt. Für die Gesamtpunktzahl aus beiden Aufgabenteilen zeigen das Histogramm (Abb. 22) und das Q-Q-Diagramm (Abb. 23) bereits wenig Ähnlichkeit mit einer Normalverteilung.

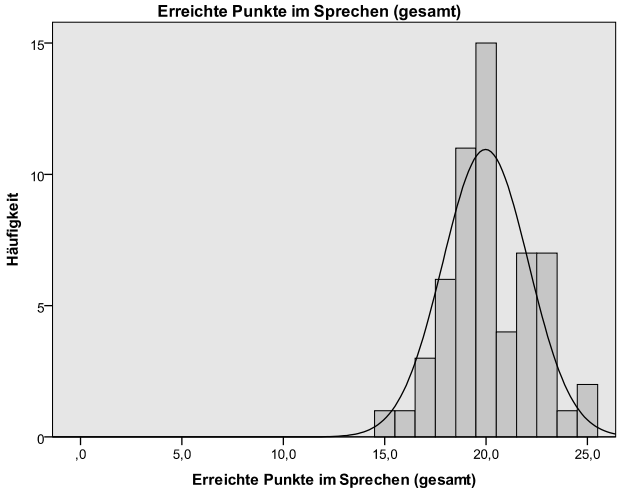


Abb. 22: Histogramm Sprechen (n = 58)

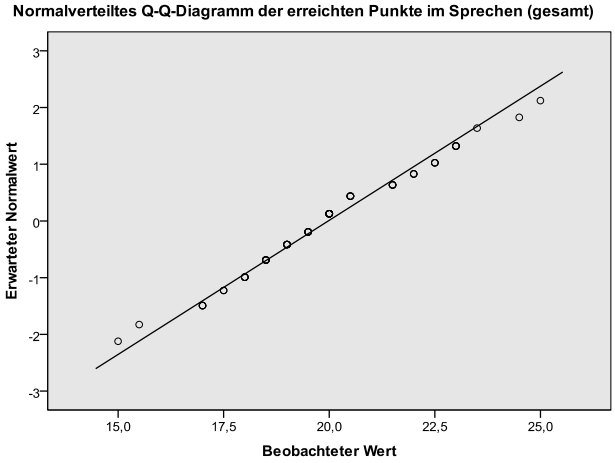


Abb. 23: Q-Q-Diagramm Sprechen (n = 58)

Rechnerisch liefert der Kolmogorov-Smirnov-Test einen Wert von 0,133 (Sig. 0,012). Die Annahme, dass eine Normalverteilung vorliegt, ist folglich abzulehnen.

Für die getrennte Betrachtung der Aufgaben 1 (monologisches Sprechen) und 2 (dialogisches Sprechen) ergibt sich ein ähnliches Bild. Sowohl die graphischen Darstellungen der Verteilung von Aufgabe 1 (Abb. 24 und 25) als auch von Aufgabe 2 (Abb. 26 und 27) weisen auf eine Abweichung von der Normalverteilung hin.

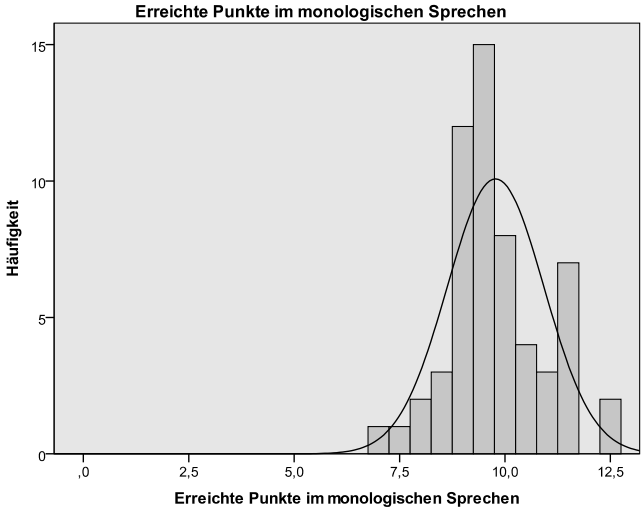


Abb. 24: Histogramm monologisches Sprechen (n = *58)*

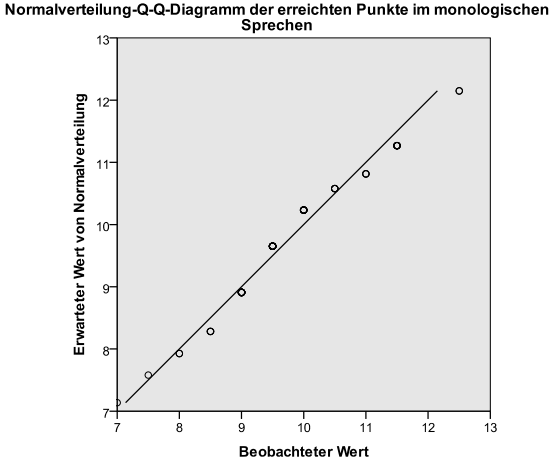


Abb. 25: Q-Q-Diagramm monologisches Sprechen (n = *58)*

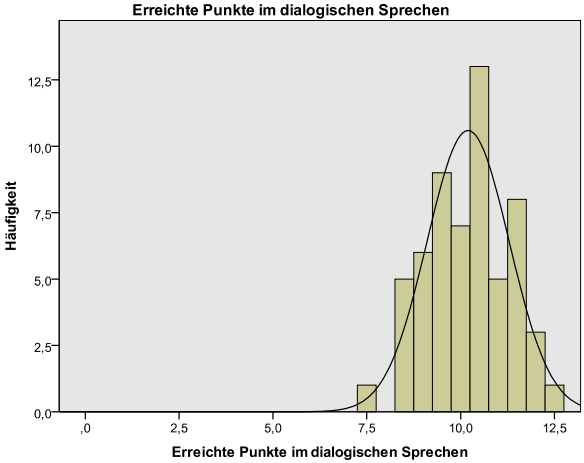


Abb. 26: Histogramm dialogisches Sprechen (n = *58)*

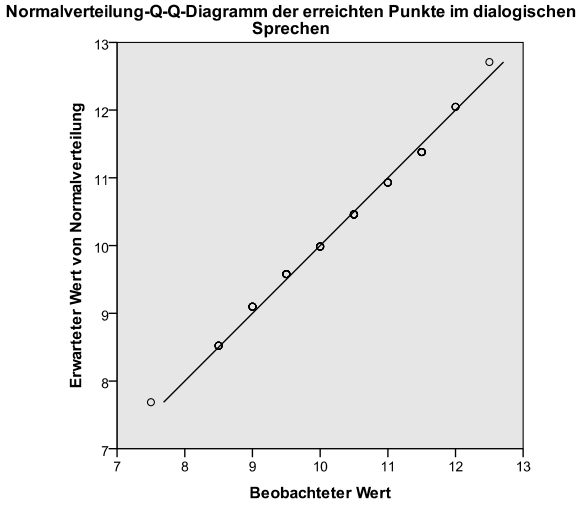


*Abb. 27: Q-Q-Diagramm dialogisches Sprechen (*n = *58)*

Auch der Kolmogorov-Smirnov-Test mit einem Wert von 0,181 (Sig. 0,000) für Aufgabe 1 bzw. 0,126 (Sig. 0,022) für Aufgabe 2 zeigt an, dass die Nullhypothese zu verwerfen ist und eine Abweichung von der Normalverteilung angenommen werden muss.

4.4.3 Beantwortung der Forschungsfragen

Tabelle 34 bietet einen Überblick über die zur Beantwortung der in Kapitel 4.1 aufgestellten Forschungsfragen durchgeführten statistischen Analysen. Die Wahl der Methoden wird im Anschluss erläutert.

Forschungsfrage	Analyse
(1) Verfügt der S-C-Test über eine akzeptable Reliabilität?	Reliabilitätsanalyse mittels Cronbachs Alpha
(2a) Besteht ein Zusammenhang zwischen den S-C-Test-Ergebnissen mit den Fertigkeiten Hörverstehen und Sprechen?	Korrelationsanalyse mittels Spearman Rho und Kendalls Tau
(2b) Besteht ein stärkerer Zusammenhang zwischen den Ergebnissen von S-C-Test und den Fertigkeiten Hörverstehen und Sprechen als zwischen den Ergebnissen von onDaF und den Fertigkeiten Hörverstehen und Sprechen?	Korrelationsanalyse mittels Spearman Rho und Kendalls Tau, Konfidenzintervalle
(3) Besteht ein stärkerer Zusammenhang zwischen den Ergebnissen von S-C-Test und dialogischem oder monologischem Sprechen?	Korrelationsanalyse mittels Spearman Rho und Kendalls Tau, Konfidenzintervalle
(4) Unterscheidet sich die Stärke der Korrelationen des S-C-Tests mit den Fertigkeiten Hörverstehen und Sprechen bei unterschiedlich weit fortgeschrittenen Lernern?	Korrelationsanalyse mittels Spearman Rho und Kendalls Tau, Konfidenzintervalle
(5) Gibt es einen Zusammenhang zwischen dem Ergebnis beim S-C-Test und der Flüssigkeit mündlicher Sprachproduktion?	Silben pro Minute Wörter pro Minute Verzögerungsindikatoren
(6) Verändert die *speed*-Komponente die Augenscheinvalidität des C-Tests?	Kreuztabellen Kontingenzkoeffizient C Mann-Whitney-U-Test
(7) Hat das zuerst erlernte Schriftsystem einen Einfluss auf den Erfolg bei einem deutschsprachigen S-C-Test?	Mann-Whitney-U-Test
(8) Unterscheidet sich die Lösungsreihenfolge beim Lösen von C-Tests und S-C-Tests?	Häufigkeitsverteilungen

Tab. 34: Übersicht über Forschungsfragen und Analysemethoden

4.4.3.1 Verfügt der S-C-Test über eine akzeptable Reliabilität?

Wie in Kapitel 2.2.3 diskutiert, ist eine akzeptable Reliabilität eine unabdingbare Voraussetzung für ein erfolgreiches Testformat. Eine Beurteilung des S-C-Test-Formats ohne Ermittlung der Reliabilität ist folglich nicht möglich.

Um die Reliabilität eines Tests zu ermitteln, gibt es verschiedene Möglichkeiten. Eine Variante ist die **Retestreliabilität**. Diese wird durch das zweifache Einsetzen eines (identischen) Tests bei der gleichen Probandengruppe ermittelt. Im Anschluss wird die Korrelation zwischen den beiden Testergebnissen errechnet (vgl. MOOSBRUGGER & KELAVA 2012: 122). Dieses Vorgehen hat jedoch den Nachteil, dass man Gedächtniseffekte nicht ausschließen kann, wenn die beiden Testungen zeitlich relativ nah beieinander liegen. Liegen die Testungen hingegen zeitlich weiter auseinander ist ein Lernzuwachs der Probanden wahrscheinlich, was das Testergebnis verzerren kann (vgl. LIENERT & RAATZ 1998: 180 f.). MOOSBRUGGER & KELAVA (2012: 123) sprechen von einer „Veränderung der wahren Werte".

Eine weitere Möglichkeit ist die Berechnung der **Paralleltestreliabilität**. Die Messgenauigkeit zweier paralleler Testsets wird hierbei in Beziehung gesetzt (vgl. LIENERT & RAATZ 1998: 182). Zwar können Gedächtniseffekte ausgeschlossen werden, jedoch benötigt man zwei parallele Versionen eines Tests, was im vorliegenden Fall nicht gegeben ist.

Bei der sogenannten ***split half*-Reliabilität** wird ein einziger Test in zwei Hälften aufgeteilt und die Reliabilität über die Korrelation der beiden Testteile ermittelt. Es werden quasi Paralleltests simuliert. Da die Reliabilität jedoch mit zunehmender Testlänge steigt, ist der sich hier ergebende Wert eine Unterschätzung des wahren Werts (vgl. LIENERT & RAATZ 1998: 182 ff.). Die Bestimmung der *split half*-Reliabilität eignet sich für län-

gere Tests (vgl. MOOSBRUGGER & KELAVA 2012: 128). Im vorliegenden Fall besteht der S-C-Test jedoch aus lediglich acht Super-Items.

Die in SPSS integrierte Möglichkeit zur Berechnung der Reliabilität ist der Test auf **interne Konsistenz**. Diese Variante wird im Folgenden genutzt, da, wie oben angeführt, im vorliegenden Fall weder eine Retest- noch eine Paralleltestreliabilität ermittelt werden kann und das Testset relativ kurz ist (vgl. LIENERT & RAATZ 1998: 191 ff.). Voraussetzung zur Ermittlung der internen Konsistenz ist, dass die Items eines Testsets dasselbe Merkmal erfassen (vgl. MOOSBRUGGER & KELAVA 2012: 130). Da die Lücken eines C-Tests stochastisch voneinander abhängig sind, wird bei der Berechnung ein Text als ein (Super-)Item behandelt. Bei einer Betrachtung der Lücken als Items würde die Reliabilität aufgrund dieser Abhängigkeit überschätzt.

In der Testforschung gilt im Allgemeinen, dass der Wert für Cronbachs Alpha nicht unter 0,7 liegen sollte, damit ein Test als zufriedenstellend reliabel bezeichnet werden kann (vgl. MOOSBRUGGER & KELAVA 2007: 11). Bei C-Tests liegt die Reliabilität laut GROTJAHN (2004: 538) in der Regel zwischen $\alpha = 0{,}8$ und $\alpha = 0{,}9$. Für den *speeded*-C-Test wird daher auch ein Wert von mindestens 0,8 angestrebt.

Da der hier verwendete S-C-Test eine Papier-und-Bleistift-Version des etablierten onDaF ist, interessiert zunächst ein Blick auf die Reliabilität des onDaF. Diese wird in der Literatur mit einem äußerst hohen Wert für Cronbachs Alpha zwischen 0,94 und 0,98 angegeben (vgl. ECKES 2010: 125). Da der onDaF über einen Pool zahlreicher Texte verfügt und die in diesem Pool befindlichen Texte in verschiedenen Testsets erprobt werden, ergeben sich je nach Textauswahl für den onDaF als Gesamtteil leicht unterschiedliche Werte. Die in Tabelle 35 angegebenen Werte des onDaF stammen aus den für diese Arbeit vom TestDaF-Institut erstellten Berechnungen vom Oktober 2013.

Die Berechnung der **Reliabilität als interne Konsistenz** des S-C-Tests mit SPSS (Versionen 22 & 24) unter Einbezug von 120 Probanden ergibt einen Wert von Cronbachs Alpha von 0,874. Tabelle 35 stellt die Reliabilitäten des onDaF und des S-C-Tests gegenüber.

	onDaF	S-C-Test
n	11745	120
Cronbachs Alpha	0,97	0,874

Tab. 35: Reliabilität für den onDaF und den S-C-Test

In Anbetracht der Tatsache, dass in die Berechnung der Reliabilität des onDaF Daten von ungleich mehr Probanden mit sehr unterschiedlicher fremdsprachlicher Kompetenz eingeflossen sind als bei der Berechnung der Reliabilität des S-C-Tests, ist ein Cronbachs Alpha von 0,874 folglich ein sehr zufriedenstellender Wert. Aufgrund des Verfahrens zur Probandenauswahl (nur Teilnehmer, die mindestens 82 von 160 Punkten im onDaF erzielt haben, vgl. Kapitel 4.4.3.2) ist die Varianz der Testergebnisse beim S-C-Test (SD = 14,17) geringer als sie bei einer Erprobung mit Testteilnehmern aller Kompetenzstufen ist, wie es beim onDaF der Fall ist. Es ist daher anzunehmen, dass unter Einbezug weiterer Probanden niedrigerer Kompetenzniveaus die Reliabilität weiter steigt.

Die Frage, ob der S-C-Test über eine ausreichend hohe Reliabilität verfügt, kann also eindeutig positiv beantwortet werden. Eine wichtige Voraussetzung für die Eignung des S-C-Tests als Messinstrument ist somit gegeben.

Exkurs: Reliabilität des Hörverstehenstests

Da die Reliabilität eine notwendige Voraussetzung für die Validität eines Tests darstellt, ist es wichtig, diese für die einzelnen Messinstrumente zu ermitteln, um so die hiermit erzielten Ergebnisse richtig interpretieren zu können.

Weil es sich bei dem hier verwendeten Hörverstehenstest (vgl. Kapitel 4.3.1.4) um Aufgaben handelt, deren Items eine dichotome richtig-falsch-Kodierung in der Lösungsschablone erhalten, wird zur Berechnung der Reliabilität Cronbachs Alpha herangezogen (vgl. Tab. 36). Es ergeben sich folgende Werte:

n = 120	Cronbachs Alpha	Anzahl der Items
Aufgabe 1	0,447	5
Aufgabe 2	0,440	10
gesamt	0,571	15

Tab. 36: Reliabilitäten Hörverstehen

Ein Wert für die innere Konsistenz eines Tests unter 0,8 ist als niedrig zu bewerten. In der Testforschung werden Werte ab 0,7 als akzeptabel angesehen (vgl. MOOSBRUGGER & KELAVA 2007: 11). Vor diesem Hintergrund ist der für den Hörverstehenstest ermittelte Wert von $\alpha = 0{,}571$ als problematisch zu bewerten.

Grundvoraussetzung für die Messgenauigkeit eines Tests ist Objektivität auf den Ebenen der Durchführung und Bewertung. Dies ist beim vorliegenden Hörverstehenstest aufgrund einheitlicher Durchführungsbedingungen gemäß den Vorgaben des Goethe-Instituts sowie vorgegebenen Lösungsschablonen gegeben und kann somit nicht ursächlich für die geringe Reliabilität des Hörverstehenstests sein.

Ein weiterer Einflussfaktor ist die Testlänge. Hierzu ist anzumerken, dass der eingesetzte Hörverstehenstest lediglich 15 Items beinhaltet, deren Reliabilität für sich betrachtet noch deutlich unter der für beide Aufgaben gemeinsam ermittelten Reliabilität liegt. Die Testlänge kann daher als eine mögliche Ursache für die niedrigen Werte von Cronbachs Alpha angesehen werden.

Es gilt zu beachten, dass die geringe Reliabilität des Hörverstehenstests dazu führt, dass die im Folgenden errechneten Korrelationen den wahren Wert unterschätzen (vgl. ARRAS et al. 2002: 201).

4.4.3.2 Besteht ein Zusammenhang zwischen den S-C-Test-Ergebnissen mit den Fertigkeiten Hörverstehen und Sprechen?

Zur Beantwortung der Frage, ob der S-C-Test mit den in Echtzeit ablaufenden Fertigkeiten Hörverstehen und Sprechen korreliert, wird diese Forschungsfrage in mehrere Hypothesen untergliedert, die es zu testen gilt.

In die Berechnung der Korrelation werden nur diejenigen Probanden aus der Kerngruppe der 120 Testpersonen einbezogen, die im ersten Testteil onDaF mindestens 82 Punkte erreicht haben. Diese Punktgrenze ergibt sich aus dem *Cut Score* zwischen Niveaustufe B1 und B2, der zum Zeitpunkt der Datenerhebung zwischen 95 und 96 Punkten lag. Da das Kriterium, gegen das die Leistung in beiden C-Test-Varianten korreliert werden sollte, Subtests aus einer Sprachprüfung auf B2-Niveau waren, wurden neben den als B2-Lerner klassifizierten Probanden auch Lerner in die Untersuchung aufgenommen, die im Deutschen ein hohes B1-Niveau erreicht hatten. Hierzu wurde für den Punktwert 96 die untere Vertrauensgrenze bestimmt, aus der ersichtlich wird, wie weit unterhalb eines gemessenen Testwerts von 96 ein Testteilnehmer noch als möglicher B2-Proband eingestuft werden kann (vgl. LIENERT & RAATZ 1998: 365 f.). Es ergibt sich für eine Irrtumswahrscheinlichkeit von 5 % folgende Rechnung:

$Clx = X_i +/- 1{,}96 \times S_e$
$Clx = 96 - 1{,}96 \times 6{,}89$
$Clx = 96 - 13{,}5$
$Clx = 82{,}5$

Die Mindestpunktzahl für die Teilnahme an der Studie wurde folglich bei 82 Punkten festgesetzt.[35]

Da nicht alle Probanden die dreitägige Datenerhebung vollständig mitgemacht haben, fehlt bei einer Reihe von Testpersonen die Aufgabe zum mündlichen Ausdruck. Es ergeben sich daher unterschiedliche Subgruppen für die Berechnung der Korrelationen mit Hörverstehen und Sprechen. Da für das Hörverstehen mehr Datensätze in die Rechnung einbezogen werden konnten, sind die hier ermittelten Werte als zuverlässiger einzustufen als die Korrelationen mit der Fertigkeit Sprechen.

Das Punktdiagramm in Abbildung 28 gibt bereits einen ersten Eindruck davon, ob ein linearer Zusammenhang zwischen dem S-C-Test und der Fertigkeit Hörverstehen besteht. Bei einem perfekten Zusammenhang zwischen zwei Merkmalen befinden sich alle Punkte auf einer Linie. Je weiter die Punkte sich von der Anpassungslinie entfernt befinden bzw. je weiter die Punkte um diese Linie streuen, desto schwächer ist der Zusammenhang zwischen den beiden Merkmalen.

Die Punktwolke deutet auf einen positiven linearen Zusammenhang hin, d. h. darauf, dass beide Merkmale, erreichte Punkte im S-C-Test und im Hörverstehenstest, gemeinsam ansteigen. Aus diesem Grund kann in der Korrelationsanalyse ein einseitiger Signifikanztest durchgeführt werden. Die aufgetragenen Punkte streuen jedoch sichtbar um die Anpassungslinie, was auf eine nur mäßig starke Korrelation hinweist (vgl. BAMBERG et al. 2012: 34).

35 Der Standardmessfehler S_e berechnet sich aus der Reliabilität (0,97) und der Standardabweichung (39,8) eines Tests. Beide Werte wurden mir freundlicherweise von Dr. Thomas Eckes vom TestDaF-Institut zur Verfügung gestellt (Stand 10/2013).

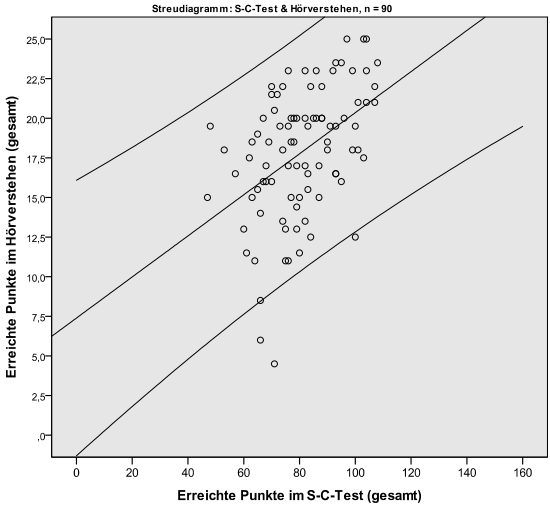


Abb. 28: Streudiagramm S-C-Test und Hörverstehen (n = 90)

Aufgrund der Abweichungen von der Normalverteilung muss zur Berechnung der Korrelationen zu nonparametrischen Verfahren gegriffen werden, so dass zur Beantwortung der Forschungsfrage der Rangkorrelationskoeffizient Spearman Rho zum Einsatz kommt. Dieser eignet sich zur Berechnung von Korrelationen, wenn das gemessene Merkmal lediglich ordinalskaliert ist, oder, wie im vorliegenden Fall, der Datensatz nicht der Normalverteilung folgt.

Da jedoch der Spearman-Rho-Rangkorrelationskoeffizient und der Pearson-Bravais-Prudukt-Moment-Koeffizient ineinander überführt werden können, weil es sich dabei um eine „Produkt-Moment-Korrelation der Rangwerte“ (EID et al. 2015: 551) handelt, werden ebenfalls die Werte für Kendalls Tau b angegeben. Dieser Koeffizient hat den Vorteil, dass er keine Annahmen über die Verteilung der zugrundeliegenden Daten macht und die Korrelation rein kombinatorisch über einen „vollständigen Paarver-

gleich aller Merkmalsträger im Hinblick auf ihre jeweiligen Rangplätze" (EID et al. 2015: 543) ermittelt. Kendalls Tau fällt generell etwas niedriger aus als Spearman Rho (vgl. HILL & LEWICKI 2006: 37).

Die Berechnung der beiden Korrelationskoeffizienten bestätigt den optischen Eindruck des Streudiagramms (vgl. Tab. 37; vgl. BÜHL 2006: 342):

	$n = 90$
Spearman Rho	$\rho = 0{,}459^*$
	Sig. (1-seitig) 0,000
Kendalls Tau	$\tau = 0{,}322^*$
	Sig. (1-seitig) 0,000

* Die Korrelation ist auf dem Niveau 0,01 signifikant (1-seitig).

Tab. 37: Spearman Rho und Kendalls Tau Korrelation zwischen S-C-Test und Hörverstehen

Beträge der Werte des Korrelationskoeffizienten zwischen 0,2 und 0,5 werden laut BÜHL (2006: 263) als geringe Korrelationen bezeichnet. Es besteht folglich ein geringer positiv linearer Zusammenhang zwischen der Leistung im S-C-Test und im Hörverstehen.

Auch für die Fertigkeit Sprechen soll zunächst ein Streudiagramm (vgl. Abb. 29) einen ersten optischen Eindruck geben. Hier scheint ebenso ein mäßiger linearer Zusammenhang zwischen den S-C-Test-Ergebnissen und der Leistung in der Sprechaufgabe zu bestehen. Die Berechnung von Spearmans Rangkorrelationskoeffizienten und Kendalls Tau vervollständigt dieses Bild (vgl. Tab. 38). Es zeigt sich ein geringer positiv linearer Zusammenhang der Ergebnisse aus beiden Testteilen. Die Frage, ob der S-C-Test mit den Fertigkeiten Hörverstehen und Sprechen korreliert, kann somit positiv beantwortet werden.

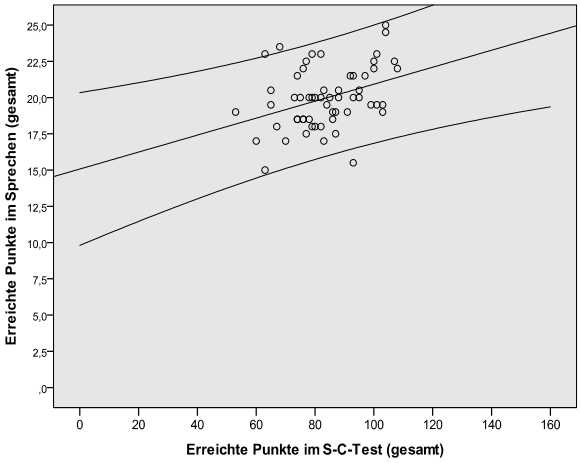


Abb. 29: Streudiagramm S-C-Test und Sprechen (n = 58)

	n = 58
Spearman Rho	ρ = 0,325*
	Sig. (1-seitig) 0,006
Kendall Tau	τ = 0,243*
	Sig. (1-seitig) 0,005

* Die Korrelation ist auf dem Niveau 0,01 signifikant (1-seitig).

Tab. 38: Spearman Rho und Kendalls Tau Korrelationen zwischen S-C-Test und Sprechen

4.4.3.3 Besteht ein stärkerer Zusammenhang zwischen den Ergebnissen des S-C-Tests und den Fertigkeiten Hörverstehen und Sprechen als zwischen den Ergebnissen von onDaF und den Fertigkeiten Hörverstehen und Sprechen?

Um diese Frage zu beantworten, müssen die Korrelationen zwischen dem onDaF und Hörverstehen sowie Sprechen berechnet und anschließend mit den Korrelationswerten des S-C-Tests mit diesen beiden Fertigkeiten verglichen werden. Dazu werden zunächst wieder Streudiagramme herangezogen. Das Punktdiagramm in Abbildung 30 zeigt den ermittelten Zusammenhang der Ergebnisse des onDaF mit der Fertigkeit Hörverstehen.

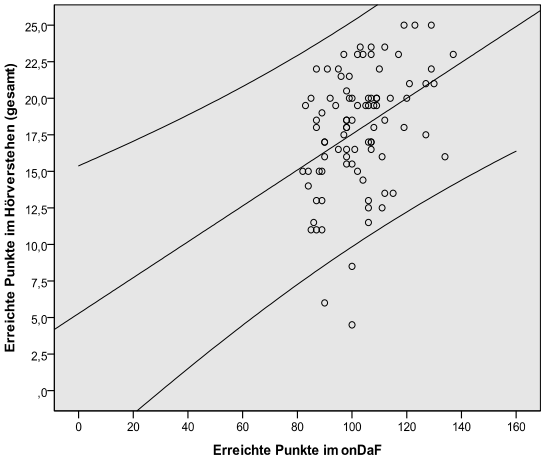


Abb. 30: Streudiagramm C-Test und Hörverstehen (n = *90)*

Bei der Betrachtung dieser Graphik muss beachtet werden, dass der abrupte Abfall der Punkte zur linken Seite der X-Achse daher rührt, dass die Punktgrenze von 82 Punkten beim onDaF als Bedingung zur (weiteren) Teil-

nahme an der Studie festgesetzt wurde. Rechnerisch ergeben sich folgende Werte (vgl. Tab. 39):

	n = 90
Spearman Rho	ρ = 0,371*
	Sig. (1-seitig) 0,000
Kendalls Tau	τ = 0,266*
	Sig. (1-seitig) 0,000

* Die Korrelation ist auf dem Niveau 0,01 signifikant (1-seitig).

Tab. 39: Spearman Rho und Kendalls Tau Korrelationen zwischen C-Test und Hörverstehen

Der errechnete Wert von ρ = 0,371 liegt relativ mittig zwischen den in der Literatur angegebenen Korrelationen von 0,16 und 0,69. Kendalls Tau liegt mit τ = 0,266 erwartungsgemäß niedriger. Beide Werte sind als schwache positive Korrelation zu interpretieren.

Für die Fertigkeit Sprechen ergibt sich in Kombination mit den Ergebnissen aus dem onDaF folgendes Punktdiagramm (vgl. Abb. 31).

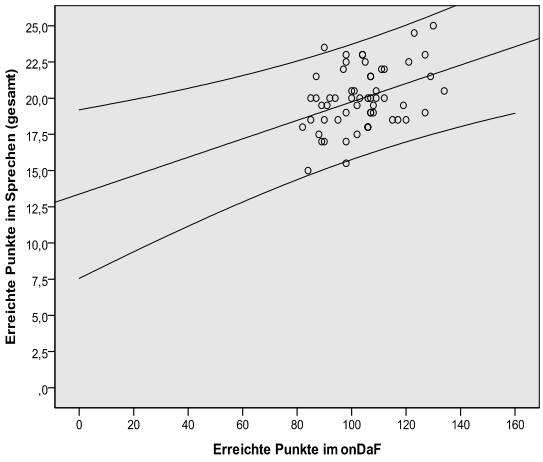


Abb. 31: Streudiagramm C-Test und Sprechen (n = 58)

Die Berechnung der Korrelationskoeffizienten bestätigt den sich aus dem Streudiagramm ergebenden optischen Eindruck eines schwachen positiv linearen Zusammenhangs (vgl. Tab. 40).

	n = 58
Spearman Rho	ρ = 0,323**
	Sig. (1-seitig) 0,007
Kendalls Tau	τ = 0,233**
	Sig. (1-seitig) 0,006

**Die Korrelation ist auf dem Niveau 0,01 signifikant (1-seitig).

Tab. 40: Spearman Rho und Kendalls Tau Korrelationen zwischen C-Test und Sprechen

Mit Werten von ρ = 0,323 und τ = 0,233 liegen die in dieser Studie gefundenen Korrelationen zwischen dem onDaF und der Fertigkeit Sprechen unterhalb der in der Literatur zu findenden Werte zwischen −0,38 und 0,640.

Um eine Aussage darüber treffen zu können, ob der S-C-Test höher als der C-Test mit den in Echtzeit ablaufenden Fertigkeiten Hörverstehen und Sprechen korreliert, werden nun die ermittelten Korrelation gemeinsam betrachtet. Tabelle 41 gibt einen Überblick.

Sowohl für das Hörverstehen als auch für den mündlichen Ausdruck liegen die Werte für Spearman Rho und Kendalls Tau nah beieinander. Im Falle des mündlichen Ausdrucks liegen die errechneten Korrelationen mit dem S-C-Test und dem onDaF sehr nah beieinander, insbesondere für Spearman Rho. Beim Hörverstehen hingegen fallen die Korrelationen mit dem S-C-Test etwas aus höher als die mit dem onDaF.

	S-C-Test	C-Test (onDaF)
Hörverstehen (n = 90)	ρ = 0,459** Sig. (1-seitig) 0,000	ρ = 0,371** Sig. (1-seitig) 0,000
	τ = 0,322** Sig. (1-seitig) 0,000	τ = 0,266** Sig. (1-seitig) 0,000
Sprechen (n = 58)	ρ = 0,325** Sig. (1-seitig) 0,006	ρ = 0,323** Sig. (1-seitig) 0,007
	τ = 0,243* Sig. (1-seitig) 0,005	τ = 0,233** Sig. (1-seitig) 0,006

** Die Korrelation ist auf dem Niveau 0,01 signifikant (1-seitig).

Tab. 41: Übersicht Spearman Rho Korrelationen zwischen C-Test/S-C-Test und Hörverstehen/Sprechen

Da die Fisher-Z-Transformation zur Ermittlung signifikanter Unterschiede zwischen der Stärke zweier Korrelationen die Normalverteilung der Daten voraussetzt (vgl. BORTZ & DÖRING 2002: 606), wird stattdessen darauf zurückgegriffen, Konfidenzintervalle um die Werte von Kendalls Tau zu legen, um zumindest näherungsweise eine Aussage darüber zu machen, ob sich zwei Korrelationen signifikant voneinander unterscheiden.[36] Tabelle 42 zeigt die über die Werte von Kendalls Tau berechneten Vertrauensintervalle für die Korrelationen der beiden C-Test-Varianten mit dem Testteil Hörverstehen (vgl. LENHARD & LENHARD 2014).

S-C-Test		C-Test (onDaF)	
τ = 0,322		τ = 0,266	
0,062	0,448	0,123	0,496

Tab. 42: Konfidenzintervalle für die Korrelationen mit Hörverstehen (95 %)

36 Zu den Vorteilen der Verwendung von Konfidenzintervallen gegenüber Signifikanztests siehe BRANDSTÄTTER (1999).

Es zeigt sich, dass die Intervalle sich in weiten Bereichen überschneiden und sich die Korrelationswerte somit nicht signifikant voneinander unterscheiden. Für den Testteil Sprechen ergibt sich ein ähnliches Bild, wie Tabelle 43 zeigt.

S-C-Test		C-Test (onDaF)	
τ = 0,243		τ = 0,233	
−0,016	0,472	−0,027	0,463

Tab. 43: Konfidenzintervalle für die Korrelationen mit Sprechen (95 %)

Die Konfidenzintervalle für die Korrelationen des S-C-Tests und herkömmlichen C-Tests mit der Fertigkeit Sprechen überlappen sich fast vollständig. Auch hier ist auf der Grundlage dieses näherungsweisen Tests kein signifikanter Unterschied in der Stärke des Zusammenhangs erkennbar. Eine eindeutige Antwort auf die Forschungsfrage kann anhand der vorliegenden Daten nicht gegeben werden. Zwar liegen die berechneten Korrelationen des S-C-Tests sowohl mit Hörverstehen als auch mit Sprechen höher als die des C-Tests mit den beiden Fertigkeiten. Jedoch konnte nicht nachgewiesen werden, dass es sich dabei um einen statistisch bedeutsamen Unterschied handelt.

4.4.3.4 Besteht ein stärkerer Zusammenhang zwischen den Ergebnissen des S-C-Tests mit dialogischem oder monologischem Sprechen?

Das gewählte Testformat zur Fertigkeit Sprechen ermöglicht eine differenzierte Betrachtung des Zusammenhangs zwischen dem Abschneiden beim S-C-Test und monologischem bzw. dialogischem Sprechen. Dies ist so beabsichtigt, da in bisherigen Studien nicht zwischen unterschiedlichen Arten der mündlichen Sprachproduktion unterschieden wurde. Das dialogische Sprechen erfordert von den Testpersonen ein Verstehen dessen, was

der Gesprächspartner äußert sowie eine spontane Reaktion darauf. Man kann daher zu der Hypothese gelangen, dass der S-C-Test durch seine Zeitkomponente einen höheren Zusammenhang mit dialogischem als mit monologischem Sprechen aufweist.

Zur Beantwortung dieser Forschungsfrage werden im Folgenden Korrelationen des S-C-Tests mit den beiden Testteilen der Sprechaufgabe, d. h. Teil 1, monologisches Sprechen, und Teil 2, dialogisches Sprechen, gerechnet und miteinander verglichen (vgl. Tab. 44):

	Spearman Rho	Kendalls Tau
Monologisches Sprechen *n* = 58	ρ = 0,247* Sig. (1-seitig) 0,031	τ = 0,187* Sig. (1-seitig) 0,026
Dialogisches Sprechen *n* = 58	ρ = 0,365** Sig. (1-seitig) 0,002	τ = 0,279** Sig. (1-seitig) 0,002

*Die Korrelation ist auf dem Niveau 0,05 signifikant (1-seitig).
**Die Korrelation ist auf dem Niveau 0,01 signifikant (1-seitig).

Tab. 44: Spearman Rho Korrelationen zwischen S-C-Test und monologischem/dialogischem Sprechen

Der Hypothese entsprechend ergibt sich ein stärkerer Zusammenhang zwischen dem S-C-Test und den Ergebnissen aus dem dialogischen Aufgabenteil als mit den Ergebnissen aus dem monologischen Aufgabenteil. Man muss hier allerdings beachten, dass durch die rechnerische Splittung der Sprechaufgabe in zwei Teile die Varianz der Testergebnisse geringer ausfällt als für die gesamte Sprechaufgabe. Die geringere Varianz geht mit einer niedrigeren Reliabilität einher (vgl. BORTZ & DÖRING 2006: 196), was eine Erklärung für die insgesamt sehr gering ausfallenden Korrelationswerte sein kann.

Auch hier soll mittels der Konfidenzintervalle um die Werte für Kendalls Tau überprüft werden, ob sich diese beiden Korrelationen in ihrer Stärke signifikant voneinander unterscheiden (vgl. Tab. 45).

monologisches Sprechen		dialogisches Sprechen	
τ = 0,187		τ = 0,279	
−0,057	0,425	−0,022	0,501

Tab. 45: Konfidenzintervalle für den S-C-Test (95 %)

Wie aus Tabelle 45 ersichtlich ist, überschneiden sich die Konfidenzintervalle sehr deutlich. Dies ist als ein Hinweis darauf zu werten, dass sich die Stärke der beiden Korrelationen nicht signifikant voneinander unterscheiden.

4.4.3.5 Unterscheidet sich die Stärke der Korrelationen des S-C-Tests mit den Fertigkeiten Hörverstehen und Sprechen bei unterschiedlich weit fortgeschrittenen Lernern?

Bisher wurden S-C-Tests nur bei sehr fortgeschrittenen Lernern oder bei Muttersprachlern eingesetzt. Daher steht die Frage im Raum, ob der S-C-Test auch für andere Niveaustufen geeignet ist. Zur Beantwortung dieser Frage wurde der vorhandene Datensatz in eine stärkere und eine schwächere Gruppe aufgeteilt und die Korrelationen mit den Fertigkeiten Hörverstehen und Sprechen getrennt voneinander berechnet. Um die Einteilung der Probanden in die beiden Gruppen vorzunehmen wurde der Median des onDaF herangezogen, da es sich hierbei um das zuverlässigste und am besten erprobte Instrument in dieser Untersuchung handelt. Da sich die Teilnehmerzahlen beim Hörverstehen und beim Sprechen unterscheiden, wird der Median jeweils separat ermittelt.

Für die Probanden, die am Hörverstehenstest teilgenommen haben, liegt der Median des onDaF bei $\tilde{x} = 101{,}5$. Die Teilnehmer in der stärkeren Gruppe haben in diesem Testteil folglich mindestens 102 Punkte erreicht, die Teilnehmer der schwächeren Gruppe höchstens 101 Punkte.

Die Spearman Rho und Kendalls Tau Korrelationen zwischen dem S-C-Test und dem Testteil Hörverstehen werden in Tabelle 46 getrennt für die stärkere und schwächere Probandengruppe berichtet.

	Schwächere Gruppe (n = 45)	Stärkere Gruppe (n = 45)
S-C-Test	ρ = 0,273*	ρ = 0,506**
	Sig. (1-seitig) = 0,037	Sig. (1-seitig) = 0,000
	τ = 0,173	τ = 0,351**
	Sig. (1-seitig) = 0,053	Sig. (1-seitig) = 0,001

* auf dem Niveau 0,05 signifikant (einseitig)
** auf dem Niveau 0,01 signifikant (einseitig)

Tab. 46: Spearman Rho und Kendalls Tau Korrelationen zwischen S-C-Test und Hörverstehen aufgeteilt nach stärkeren und schwächeren Probanden

Zunächst fällt auf, dass in der schwächeren Gruppe über Kendalls Tau keine signifikante Korrelation zwischen den Ergebnissen des S-C-Tests und dem Hörverstehen gefunden werden konnte (τ = 0,173; Sig. 0,053). Für den Rangkorrelationskoeffizienten von Spearman findet sich mit ρ = 0,273 eine schwache, signifikante Korrelation.

In der stärkeren Gruppe werden sowohl mit Spearman Rho als auch mit Kendalls Tau signifikante Korrelationen zwischen den Ergebnissen des S-C-Tests mit der Fertigkeit Hörverstehen erreicht: Kendalls Tau liegt bei τ = 0,351 und Spearman Rho liegt mit einem Wert von ρ = 0,506 noch höher und lässt sich bereits als mittelstarker Zusammenhang bewerten. Eine näherungsweise Überprüfung des Unterschieds zwischen den beiden Werten mittels Konfidenzintervallen um die Werte von Kendalls Tau entfällt. Es lässt sich festhalten, dass für den S-C-Test in der Gruppe der stärkeren Lerner ein deutlich stärkerer Zusammenhang mit der Fertigkeit Hörverstehen zu bestehen scheint als bei der schwächeren Gruppe.

Bei den Teilnehmern der Sprechaufgabe liegt der Median des onDaF bei $\tilde{x}$ = 104. Die schwächere Gruppe wurde mit weniger als 104 Punkten im onDaF definiert, die Teilnehmer der stärkeren Gruppe hatten demzufolge 104 oder mehr Punkte im onDaF erreicht. Für das Sprechen konnten auf diese Weise folgende Werte ermittelt werden (vgl. Tab. 47):

	Schwächere Gruppe (n = 28)	Stärkere Gruppe (n = 30)
S-C-Test	ρ = −0,066 Sig. (einseitig) = 0,369	ρ = 0,561** Sig. (einseitig) = 0,001
	τ = −0,069 Sig. (einseitig) = 0,309	τ = 0,447** Sig. (einseitig) = 0,000

** auf dem Niveau 0,01 signifikant (einseitig)

Tab. 47: Spearman Rho und Kendalls Tau Korrelationen zwischen S-C-Test und Sprechen aufgeteilt nach stärkeren und schwächeren Probanden

Die Ergebnisse für die Sprechaufgabe unterscheiden sich von denen des Hörverstehens. Beim S-C-Test findet sich in der schwächeren Gruppe kein signifikanter Wert (ρ = −0,066; Sig. 0,369; τ = −0,069; Sig. 0,309). Zwischen dem S-C-Test und der Sprechaufgabe in der stärkeren Teilnehmergruppe zeigen sich jedoch signifikante Korrelationen von τ = 0,447 und ρ = 0,561. Letzterer Wert ist als Zeichen eines mittelstarken Zusammenhangs zu interpretieren. Auch hier entfällt aufgrund der nicht signifikanten Werte für Kendalls Tau die Ermittlung der Konfidenzintervalle.

Es zeigt sich, dass der S-C-Test im höheren Kompetenzbereich offenkundig besser dazu geeignet ist, die Sprachkompetenz der Teilnehmer in den Bereichen Hörverstehen und Sprechen abzubilden. Die in der stärkeren Probandengruppe gefundenen Korrelationen fallen zudem in allen Fällen höher aus als die der gesamten Probandengruppe, wie Tabelle 48 noch einmal veranschaulicht.

	schwache Gruppe	alle Probanden	starke Gruppe
Hörverstehen	ρ = 0,273 τ = 0,173 (*n* = 45)	ρ = 0,459 τ = 0,322 (*n* = 90)	ρ = 0,506 τ = 0,351 (*n* = 45)
Sprechen	ρ = −0,066 τ = −0,069 (*n* = 28)	ρ = 0,325 τ = 0,243 (*n* = 58)	ρ = 0,561 τ = 0,447 (*n* = 30)

Tab. 48: Vergleich der Korrelationen des S-C-Tests mit Hörverstehen und Sprechen nach Probandengruppe

Diese Ergebnisse weisen in die Richtung, dass ein Einsatz des S-C-Tests auf niedrigeren Niveaustufen weniger geeignet ist. Allerdings kann auch das aus der Aufteilung der Probanden resultierende kleine *n* hierbei eine Rolle spielen.

4.4.3.6 Gibt es einen Zusammenhang zwischen dem Ergebnis beim S-C-Test und der Flüssigkeit mündlicher Sprachproduktion?

Zur Beantwortung dieser Forschungsfrage wurden zunächst jeweils zwei Probanden ausgewählt, die beim S-C-Test innerhalb der Probandengruppe im oberen, mittleren und unteren Punktebereich abgeschnitten haben. Für eine bessere Vergleichbarkeit der Daten wurde bei der Auswahl darauf geachtet, dass alle sechs Probanden bei der monologischen Sprechaufgabe den gleichen Stimulustext („Kaffee") bekommen hatten. Von diesen ausgewählten Testteilnehmern wurde die erste Minute der monologischen Aufgabe des Goethe-Zertifikats B2 als Minimaltranskript nach GAT 2 (vgl. Selting et al. 2009; Hagemann & Henle 2014) transkribiert (siehe Anhang I).[37]

37 Ich danke Dr. Dörte Grunzig für die Unterstützung beim Anwenden von GAT 2.

In der Literatur finden sich zahlreiche Verfahren zum Ermitteln der Flüssigkeit mündlicher Sprachproduktion. Als Maße der *breakdown fluency* werden die Anzahl und Länge der Sprechpausen mittels dazu geeigneter Software wie beispielsweise Praat ermittelt. Die sogenannte *repair fluency* bezieht sich auf die Anzahl von Wiederholungen und Selbstkorrekturen während einer mündlichen Äußerung. Zuletzt kann als *speed fluency* die Artikulationsrate erhoben werden, beispielsweise als Anzahl gesprochener Silben innerhalb eines festgesetzten Zeitrahmens (vgl. TAVAKOLI & SKEHAN 2005: 254 f.).

In der vorliegenden Arbeit wird auf ein sehr einfaches Verfahren zurückgegriffen: Im Vordergrund steht die *speed fluency,* welche als Artikulationsrate über die Anzahl der (bedeutungsvollen) Silben pro Minute operationalisiert wird. Da die Wörter „Fettleberrisiko" und „Alkoholkonsum" verhältnismäßig häufig vorkommen und sechs bzw. fünf Silben umfassen, werden zusätzlich auch die Wörter pro Minute gezählt.

Die Anzahl direkt aufeinanderfolgender Wortwiederholungen wie im nachfolgenden Beispiel illustriert, stehen für die *repair fluency* :

> 01 S1 okay also in meinem text äh geht es um äh **die die** möglichkeit dass ähm kaffeekonsum ah das risiko einer alkoholbändigten fettleber reduzieren kann
>
> (Proband mit laufender Teilnehmernummer 12, vgl. Anhang I)

Des Weiteren werden auch die Verzögerungssignale „äh" und „ähm" als ein Maß der *breakdown fluency* gezählt. Tabelle 49 gibt eine Übersicht über die von den sechs ausgewählten Probanden erreichten Punkte im S-C-Test sowie die in ihrer Sprechaufgabe ermittelten Flüssigkeitsmaße.

Lfd. Teilnehmer-Nr.[38]	89	7	28	8	19	12
	Schwach im S-C-Test		Mittel im S-C-Test		Gut im S-C-Test	
Punkte im S-C-Test	53	53	82	83	104	108
Silben pro Minute	133	151	204	177	145	167
Wörter pro Minute	87	78	121	123	92	90
Wortwiederholungen	3	4	0	1	1	4
„äh", „ähm"	11	22	8	11	18	14

Tab. 49: Übersicht über Punkte im S-C-Test und Flüssigkeitsmaße

Wie die Tabelle zeigt, steigt die Anzahl der gesprochenen Silben bei den untersuchten Probanden nicht mit zunehmender Zahl von Punkten im S-C-Test. Ebenso verhält es sich mit den gesprochenen Wörtern pro Minute. Abbildung 32 veranschaulicht das Verhältnis. Zunächst fällt auf, dass jeweils die beiden Probanden einer Gruppe in Bezug auf die gesprochenen Silben und Wörter relativ nah beieinander liegen. Entgegen der Annahme, dass ein besseres Abschneiden im S-C-Test mit einer flüssigeren Sprechweise zusammenhinge, zeigt die Abbildung 32, dass die Probanden des mittleren S-C-Test-Punktbereichs die meisten Silben pro Minute gesprochen haben.

Die Anzahl der Wortwiederholungen ist bei den beiden Probanden der mittleren Kategorie am geringsten, während je ein Proband der schwachen und der starken Kategorie das beobachtete Maximum von vier Wiederholungen aufweist. Auch die Zählung der Verzögerungssignale zeigt keinen linearen Anstieg in Bezug auf die im S-C-Test erreichten Punkte.

38 Um sicherzustellen, dass die Transkripte keiner Person zugeordnet werden können, wird hier lediglich die laufende Teilnehmernummer der Probanden angegeben, nicht jedoch der Teilnehmercode.

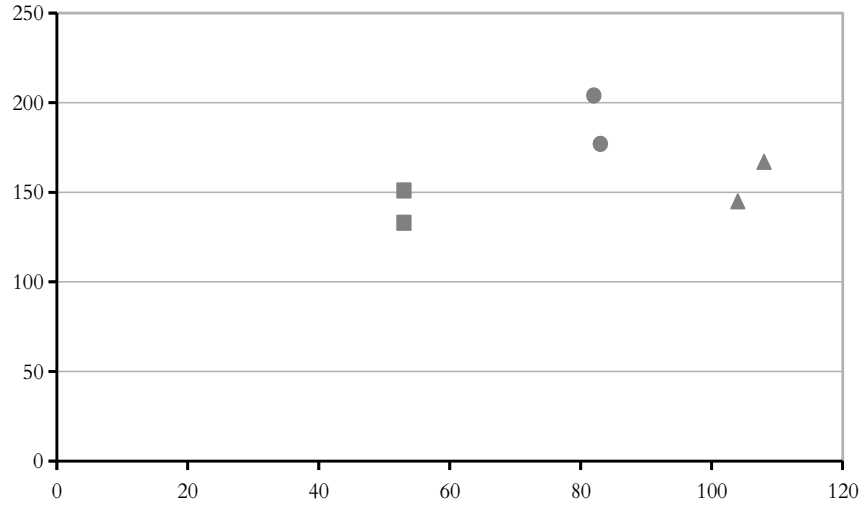


Abb. 32: Punktdiagramm über Punkte im S-C-Test (x-Achse) und Silben pro Minute (y-Achse)

Die hier analysierte Sprechaufgabe beinhaltet, wie in Kapitel 4.3.1.2 beschrieben, einen Stimulustext, der zunächst wiedergegeben und um Beispiele sowie die eigene Meinung ergänzt werden sollte. Da die Testteilnehmer sich auch an den Inhalt des Textes erinnern mussten, ist es nicht möglich, Sprechverzögerungen auf eine schlechtere sprachliche Verarbeitung zurückzuführen, da es ebenso denkbar ist, dass diese inhaltlicher Natur sind.

Das erwartete Muster zeigt sich bei den wenigen hier analysierten Datensätzen nicht. Es könnte einerseits vermutet werden, dass für die Flüssigkeit beim Sprechen und das Lösen von C-Tests unter massivem Zeitdruck, jeweils unterschiedliche Fähigkeiten zentral sind. Andererseits kann aufgrund der wenigen untersuchten Fälle jedoch auch nicht ausgeschlossen werden, dass die Ergebnisse rein zufällig sind. Eine eigens zur Beantwortung dieser

Frage angelegte Studie mit deutlich mehr Datensätzen könnte hier belastbarere Ergebnisse liefern.

4.4.3.7 Verändert die *speed*-Komponente die Augenscheinvalidität des C-Tests?

Eine beim C-Test viel diskutierte Frage ist die nach seiner Augenscheinvalidität (*face validity*). Wie bereits diskutiert wurde (vgl. Kapitel 2.2.3) ist die Augenscheinvalidität des C-Tests ein kritischer Punkt. Eine geringe Augenscheinvalidität kann aus dem Grund problematisch sein, dass Testteilnehmer, die der Meinung sind, dass ein Test ihre Sprachkompetenz ohnehin nicht korrekt wiedergibt, sich beim Lösen der Aufgaben möglicherweise weniger anstrengen als Teilnehmer, die dem Testformat grundsätzlich vertrauen. In diesem Zusammenhang ist es zunächst von Interesse zu untersuchen, ob es einen Zusammenhang zwischen dem Vertrauen in den C-Test und den S-C-Test gibt. Darüber gibt nach Auswertung des Fragebogens die folgende Kreuztabelle (vgl. Tab. 50) Aufschluss:[39]

		„Denken Sie, dass der *speeded* C-Test Ihre Deutschkompetenz korrekt wiedergibt?“		
		nein	ja	gesamt
„Denken Sie, dass der onDaF Ihre Deutschkompetenz korrekt wiedergibt?“	nein	47 43,52 %	9 8,34 %	56 51,82 %
	ja	29 26,86 %	23 21,30 %	52 48,15 %
	gesamt	76 70,37 %	32 29,63 %	108 100 %

Tab. 50: Vertrauen in das Testformat C-Test/S-C-Test (n = *120)*

39 12 Probanden haben keine Angaben gemacht.

Die mit Abstand am häufigsten vorzufindende Antwortkombination ist die, dass weder dem C-Test noch dem S-C-Test zugetraut wird, die eigene Deutschkompetenz korrekt wiederzugeben. Von den knapp 50 % der Testteilnehmer, die dem onDaF vertrauen, ist die Einschätzung des S-C-Tests mit 29 (nein) zu 23 (ja) Stimmen einigermaßen ausgeglichen. Der Fall, dass Probanden zwar dem S-C-Test vertrauen, nicht aber dem C-Test unter Niveaubedingungen ist mit nur 9 (8,34 %) Nennungen am seltensten. Das Vertrauen in das Testformat nimmt im Allgemeinen ab, wenn der C-Test unter großem Zeitdruck gelöst werden muss.

BISPING und RAATZ (2002) legten ihren Probanden einen computeradministrierten C-Test sowie einen C-Test als Papier-und-Bleistift-Version vor und fanden heraus, dass der computerbasierte C-Test in ihrer Probandengruppe eine bessere Augenscheinvalidität hatte als die herkömmliche Darbietungsform. Auch hierin könnte ein Grund dafür liegen, dass der onDaF eine höhere Akzeptanz bei den Probanden dieser Studie hat als der S-C-Test – unabhängig vom Zeitfaktor.

Der Kontingenzkoeffzient C nach Pearson eignet sich zur Berechnung von Zusammenhängen auf nominalem Messniveau, wie es hier der Fall ist (vgl. JANSSEN & LAATZ ^{9}2017: 268). Seine Berechnung bestätigt den sich aus der Kreuztabelle ergebenden Eindruck: Mit einem Wert von C = 0,294 ($p = 0{,}001$; $C_{max} = 0{,}71$) liegt zwischen den beiden Fragebogenitems ein schwacher Zusammenhang vor. In Anbetracht der Tatsache, dass es sich um das gleiche Testformat und die gleiche Testlänge handelt und nur die Darbietungsform (computeradministriert bzw. Papier-&-Bleistift-Test) und die Bearbeitungszeit variiert wurden, fällt dieser Wert geringer aus als man hätte erwarten können.

Des Weiteren lässt sich aus der Kreuztabelle ableiten, dass die Augenscheinvalidität für den S-C-Test ein größeres Problem darstellt als für den herkömmlichen C-Test. Maßgeblich ist jedoch, ob sich zwischen diesen

beiden Gruppen, also Probanden, die dem S-C-Test vertrauen, und solchen, die das nicht tun, Leistungsunterschiede zeigen. Dies ist nach Durchführung des Mann-Whitney-U-Tests, der einen Wert von 595,000 (sig. 0,05) aufweist, nicht der Fall, so dass zumindest im vorliegenden Datensatz die geringe Augenscheinvalidität kein Problem für den S-C-Test darstellt.

4.4.3.8 Hat das zuerst erlernte Schriftsystem einen Einfluss auf den Erfolg bei einem deutschsprachigen S-C-Test?

Ein möglicherweise relevanter Faktor ist das Schriftsystem, das die Testteilnehmer zuerst erlernt haben. Es wäre möglich, dass Probanden, die bereits als Kinder mit dem lateinischen Alphabet vertraut gemacht wurden, bzgl. ihrer Schreib- als auch ihrer Lesegeschwindigkeit einen Vorteil gegenüber Teilnehmern haben, die mit anderen Schriftsystemen alphabetisiert wurden. Es zeigt sich unter den 120 Probanden, die den S-C-Test abgelegt haben, folgende Verteilung (vgl. Tab. 51):

„In welchem Schriftsystem sind Sie alphabetisiert worden?“

	absolut	relativ
lateinische Schrift	84	70,0 %
arabische Schrift	5	4,2 %
kyrillische Schrift	6	5,0 %
Hanzi	19	15,8 %
griechische Schrift	2	1,7 %
kurdische Schrift	1	0,8 %
koreanische Schrift	1	0,8 %
keine Angabe	2	1,7 %

Tab. 51: Alphabetisierung der Probanden (n = *120)*

Wie aus der Tabelle hervorgeht, ist der Großteil der Probanden mit dem lateinischen Alphabet vertraut gewesen. Andere Alphabete und Schriftsysteme sind mit Ausnahme von Hanzi nur vereinzelt vertreten, so dass hier keine sinnvollen Gruppenvergleiche möglich sind. Um hierfür dennoch hinreichend große Gruppen zu erhalten, werden die Teilnehmer in zwei neue Kategorien eingeteilt, wie Tabelle 52 zeigt.

	ja	nein
Sind Sie mit lateinischer Schrift alphabetisiert worden?	84	34

Tab. 52: Alphabetisierung mit lateinischer Schrift (n = *118)*

Ein Gruppenvergleich für unabhängige Stichproben mittels Mann-Whitney-U-Test zeigt jedoch keinen signifikanten Unterschied zwischen der durchschnittlichen Testleistung in beiden Gruppe (690,0 (sig. 0,05)), so dass auch dieser Faktor beim S-C-Test keine Rolle zu spielen scheint.

4.4.3.9 Unterscheidet sich die Lösungsreihenfolge der Probanden bei C-Tests und S-C-Tests?

Ein weiterer interessanter Aspekt, der jedoch im Rahmen dieser Arbeit nicht erschöpfend behandelt werden kann, ist die Frage, ob sich durch das Hinzufügen der Geschwindigkeitskomponente beim C-Test das Lösungsverhalten der Testteilnehmer gegenüber dem C-Test mit großzügiger Bemessung der Arbeitszeit verändert. Die Hypothese hierbei war, dass die Testteilnehmer durch den Zeitdruck dazu gedrängt werden, die 20 Lücken eines C-Test-Textes eher linear zu bearbeiten. Um dieser Frage nachzugehen wurden die Probanden mit der Frage „Wie haben Sie den onDaF gelöst?“ bzw. „Wie haben Sie den *speeded* C-Test gelöst?“ konfrontiert. Es handelte sich hierbei um zwei Mehrfachauswahlitems im Fragebogen, deren Ergebnisse in Tabelle 53 zusammengefasst werden.

	Wie haben Sie den onDaF gelöst?	Wie haben Sie den S-C-Test gelöst?
Ich habe alle Lücken in der vorgegebenen Reihenfolge bearbeitet.	48	53
Ich bin innerhalb eines SATZES vor- und zurückgesprungen.	44	31
Ich bin innerhalb eines TEXTES vor- und zurückgesprungen.	51	27
Ich habe den Text zunächst überflogen und die leichteren Lücken ausgefüllt.	53	57
Ich habe mich längere Zeit mit schwierigen Lücken beschäftigt.	51	14
Ich habe schwierige Lücken leer gelassen.	45	66

Tab. 53: Häufigkeiten der Antworten aufgeschlüsselt nach C-Test und S-C-Test (Mehrfachnennung möglich, Versalien wie auf dem Fragebogen), n = *120*

Es zeigt sich, dass sowohl für den C-Test als auch für den S-C-Test weniger als die Hälfte der Befragten angab, die Lücken in der vorgegebenen Reihenfolge zu beantworten. Die Hypothese, dass die Geschwindigkeitskomponente hier einen deutlichen Einfluss haben könnte, muss folglich verworfen werden. Auch das Überfliegen der Texte und der Beginn mit leichteren Lücken ist mit 53 (C-Test) und 57 (S-C-Test) Nennungen fast gleich häufig.

Bemerkenswert ist hingegen, dass sich sehr wohl ein Effekt in den anderen Kategorien verzeichnen lässt. Sowohl das Vor- und Zurückspringen auf Satzebene als auch auf Textebene wird beim S-C-Test gegenüber dem C-Test deutlich weniger. Insbesondere auf Textebene ist dieser Unterschied mit 51 Nennungen für den C-Test gegenüber nur 27 Nennungen für den S-C-Test besonders deutlich.

Auch das Verweilen bei einzelnen, als schwierig empfundenen Lücken lässt durch das Hinzufügen der Geschwindigkeitskomponente deutlich nach. Während beim C-Test 51 Teilnehmer angaben, sich längere Zeit mit

schwierigen Lücken beschäftigt zu haben, antworteten in Bezug auf den S-C-Test lediglich 14 Probanden entsprechend. Dass in der Folge schwierige Lücken beim S-C-Test (66 Nennungen) häufiger unausgefüllt bleiben als beim C-Test (45 Nennungen) passt zu diesen Angaben.

Es scheint also Unterschiede im Lösungsverhalten der Testteilnehmer zu geben, abhängig davon, ob der C-Test mit großzügiger oder knapp bemessener Bearbeitungszeit präsentiert wird. Es gilt zu Bedenken, dass die hier gewählte Formulierung der Antwortmöglichkeiten „Ich bin innerhalb eines SATZES vor- und zurückgesprungen." bzw. „Ich bin innerhalb eines TEXTES vor- und zurückgesprungen." nicht präzise genug sind, um eine Aussage über die zugrunde liegenden Arbeitsprozesse zu treffen. Unklar bleibt, ob lediglich der Versuch, eine Lücke zu lösen sprunghaft erfolgt, oder ob hiermit auch das Lesen des Kontexts gemeint war. Zielführend wäre es, zum Beleuchten der Lösestrategien Probanden C-Tests am Computer unter Nutzung eines *Eye Trackers* lösen zu lassen. Neben dem Aufzeichnen der Augenbewegung könnte auch die Methode des *Key Loggings* Aufschluss darüber geben, wie die einzelnen Testteilnehmer bei der Bearbeitung der Texte vorgehen.

4.4.4 Diskussion

Ziel der vorliegenden Studie ist es, den Einfluss einer starken Begrenzung der Bearbeitungszeit auf das Konstrukt des C-Tests zu untersuchen. Zentral hierbei sind die Fragen nach der Reliabilität des C-Tests unter *speed*-Bedingungen sowie die von GROTJAHN et al. (2010) aufgestellte Hypothese, dass eine drastische Verkürzung der Arbeitszeit die Korrelation des C-Tests durch die Simulation von Sprachverwendung in Echtzeit mit den Fertigkeiten Hörverstehen und Sprechen erhöhen könnte und sich somit das Konstrukt des C-Tests durch den Geschwindigkeitsfaktor verändert.

Die Reliabilität des in dieser Studie eingesetzten S-C-Test-Sets war mit $\alpha = 0{,}874$ äußerst zufriedenstellend, sowohl was die allgemeinen Anforderungen an einen Test als auch was die in der Literatur zu findenden Erwartungen an den C-Test betrifft. Somit ist eine erste Grundlage zur Nutzung und sinnvollen weiteren Erforschung des S-C-Tests gegeben.

Die Frage nach der Validität und dem Konstrukt des S-C-Tests ist hingegen weitaus schwieriger zu beantworten: Die Testergebnisse der Teilnehmer fielen beim S-C-Test erwartungsgemäß schlechter aus als beim onDaF, d. h. der Median und das arithmethische Mittel lagen für den S-C-Test niedriger ($\tilde{x} = 79{,}5$ und $\bar{x} = 80{,}88$) als für den onDaF ($\tilde{x} = 101{,}5$ und $\bar{x} = 102{,}67$). Die Maße der Streuung zeigen, dass die Ergebnisse des S-C-Tests ($r = 61$; $SD = 14{,}17$) eine größere Varianz aufweisen als die des C Tests ($r = 55$; $SD = 12{,}72$). Ein Decken- oder Bodeneffekt ist bei der teilnehmenden Kohorte von B2-Lernern nicht zu erkennen.

Die Korrelationsanalyse lieferte einen signifikanten Zusammenhang beider C-Test-Formate mit den Ergebnissen der Testteile zum Hörverstehen und zum Sprechen. Hierbei zeigte sich, dass der S-C-Test eine stärkere Korrelation mit den Ergebnissen des Hörverstehenstests aufweist ($\rho = 0{,}459$; $\tau = 0{,}322$) als der herkömmliche C-Test unter Niveaubedingungen ($\rho = 0{,}371$; $\tau = 0{,}266$). In der untersuchten Teilnehmergruppe bildet der S-C-Test die auch für das Hörverstehen benötigten Fähigkeiten somit besser ab als der C-Test ohne Zeitdruck. Die Hypothese von GROTJAHN et al. (2010) wird hierdurch bekräftigt: Der S-C-Test scheint Sprachverwendung in Echtzeit simulieren zu können.

In Bezug auf die Fertigkeit Sprechen lieferten der C-Test und der S-C-Test einen nahezu identischen Wert für Spearman-Rho von $\rho = 0{,}323$ bzw. $\rho = 0{,}325$ ($\tau = 0{,}243$ bzw. $\tau = 0{,}233$). Bei einer Differenzierung nach monologischem und dialogischem Sprechen zeigte sich jedoch deutlich, dass der S-C-Test einen größeren Zusammenhang mit dialogischem

(ρ = 0,365; τ = 279) als mit monologischem Sprechen (ρ = 0,247; τ = 0,187) aufweist, wenngleich beide Korrelationen als schwach zu bewerten sind. In der Tendenz entspricht jedoch auch dieses Ergebnis der Hypothese Grotjahns, denn beim Aufgabenteil zum monologischen Sprechen hatten die Testteilnehmer mehr Planungszeit und -sicherheit und mussten, anders als im Aufgabenteil zum dialogischen Sprechen, nicht spontan auf Input des Gesprächspartners reagieren.

Eine Aufteilung der Probanden entlang des Medians zeigte deutliche Unterschiede zwischen der Gruppe der stärkeren und der Gruppe der schwächeren Lerner. Während in der schwächeren Gruppe nur ein schwacher Zusammenhang von ρ = 0,273 bzw. τ = 0,173 zwischen den Ergebnissen des S-C-Tests und des Hörverstehenstests gefunden werden konnte, liegen die ermittelten Werte für die stärkere Gruppe mit ρ = 0,506 bzw. τ = 0,351 deutlich höher. Für die Fertigkeit Sprechen konnte in der schwächeren Gruppe gar kein statistisch signifikanter Zusammenhang mit den Ergebnissen des S-C-Tests festgestellt werden. In der stärkeren Gruppe zeigten sich jedoch mittelstarke Korrelationen von ρ = 0,561 bzw. τ = 0,447. Daraus lässt sich schlussfolgern, dass das Niveau der Testteilnehmer eine zentrale Rolle spielt, und dass der S-C-Test bei fortgeschrittenen Lernern besser dazu geeignet ist, die Fertigkeiten Hörverstehen und Sprechen abzubilden.

Alle ermittelten Korrelationen für den S-C-Test reihen sich in das Spektrum der in der Literatur für den herkömmlichen C-Test zu findenden Werte ein. Tabelle 54 liefert einen Überblick über den jeweils höchsten und niedrigsten gefundenen Wert im Relation zu den Werten für den S-C-Test in der vorliegenden Studie.

	Literatur C-Test (niedrigster Wert)	S-C-Test vorliegende Studie	Literatur C-Test (höchster Wert)
Hörverstehen	0,16 (HUHTA 1996)	ρ = 0,459 τ = 0,322	0,69 (ECKES 2014)
Sprechen	−0,38 (JAKSCHIK 1996)	ρ = 0,325 τ = 0,243	ρ= 0,640 (ARRAS et al. 2002)

Tab. 54: Höchste und niedrigste in der Literatur gefundene Korrelation von C-Test mit Hörverstehen und Sprechen

An dieser Stelle sei noch einmal auf die geringe Reliabilität des Hörverstehenstests hingewiesen. Diese liegt bei α = 0,571 (vgl. Kapitel 4.4.3.1) und kann aufgrund der Tatsache, dass Messgenauigkeit eine notwendige Bedingung für Validität ist, dazu beigetragen haben, dass in der vorliegenden Studie nur mäßige Zusammenhänge zwischen den verschiedenen Testteilen gefunden wurden.

Wie in Kapitel 3.2 diskutiert erklärt sich das breite Spektrum an Werten durch Variationen in der Länge der eingesetzten C-Tests, der Reliabilität und Validität der verwendeten Außenkriterien sowie den unterschiedlichen Probandengruppen und -größen. Im Vergleich mit den aus der Literatur entnommenen Daten lässt sich festhalten, dass der S-C-Test dem C-Test pauschal weder über- noch unterlegen ist. Vielmehr gilt es herauszustellen, in welchem Kontext und in Bezug auf welche Aspekte (fremd)sprachlicher Kompetenz er gegenüber dem *power*-C-Test einen Mehrwert hat. Betrachtet man die vorliegenden Ergebnisse gemeinsam mit denen aus der Untersuchung von FADAEIPOUR und ZOHOORIAN (2017), die für ihren S-C-Test einen schwächeren Zusammenhang mit Leseverstehen fanden als für einen C-Test unter Niveaubedingungen, so zeichnet sich ab, dass sich durch die Geschwindigkeitskomponente eine Verschiebung des S-C-Konstrukts in Richtung der in Echtzeit ablaufenden Fertigkeiten ergibt.

Die schon beim C-Test unter Niveaubedingungen oft bemängelte niedrige Augenscheinvalidität bildet den analysierten Daten zufolge auch beim S-C-Test einen Schwachpunkt. Entscheidend ist jedoch, dass zwischen den Testergebnissen der Teilnehmer, die angaben, dem Format des S-C-Tests zu vertrauen und jenen, die ihm misstrauen, mittels des Mann-Whitney-U-Tests kein statistisch signifikanter Unterschied gefunden werden konnte. Somit lässt sich festhalten, dass die Augenscheinvalidität für den S-C-Test zwar nicht gut ist, dass sie aber keinen Einfluss auf das Testergebnis zu haben scheint.

Die Geschwindigkeitskomponente hat ebenfalls Auswirkungen auf das Lösungsverhalten der Probanden. Während die Testteilnehmer sowohl beim C-Test als auch beim S-C-Test am häufigsten angaben, die Lücken in linearer Folge bearbeitet zu haben, zeigte die Auswertung des Fragebogens, dass beim S-C-Test sowohl auf Satz- als auch auf Textebene weniger vor- und zurückgesprungen wurde. Dieses Ergebnis überrascht nicht, da durch den Zeitdruck weniger Arbeitszeit bleibt, um ggf. eine Re-Analyse von einzelnen Lücken oder ganzen Sätzen durchzuführen. Die durch die *speed*-Komponente geradezu erzwungene lineare Textbearbeitung bildet die Sprachverarbeitung beim Hörverstehen ab, wo aufgrund des Faktors Echtzeit ebenfalls kein Vor- und Zurückspringen möglich ist. Dies deckt sich mit den stärkeren Korrelationen zwischen S-C-Test und Hörverstehen gegenüber den Werten beim herkömmlichen C-Test. Auch das Verweilen bei schwierigen Lücken fand beim S-C-Test, wie zu erwarten war, seltener statt.

5 Grenzen und Desiderata

Die vorliegende Untersuchung hat das Ziel, einen Beitrag zur Erforschung des S-C-Tests zu leisten, erreicht. Während der C-Test seit seiner Einführung vor mehr als 35 Jahren Gegenstand zahlloser Studien war und sich längst als ökonomisches und zuverlässiges Messinstrument für das Testen allgemeiner Sprachkompetenz etabliert hat, ist der S-C-Test noch weitestgehend unerforscht. Die wenigen Studien, die einen S-C-Test einsetzen, nutzen diesen meist als Mittel zum Zweck, etwa um die Schwierigkeit eines bestehenden C-Test-Sets zu erhöhen.

Mit diesem Dissertationsprojekt wurde ein neuer Ansatz verfolgt, um dem Konstrukt des C-Tests unter *speed*-Bedingungen näherzukommen. Auf Basis der vorliegenden Daten konnte gezeigt werden, dass der S-C-Test eine gute Reliabilität aufweist und als integrativer Sprachtest mit dem C-Test unter Niveaubedingungen mithalten kann. Naturgemäß zeichnen die gewonnen Erkenntnisse jedoch kein vollständiges oder gar abschließendes Bild des S-C-Tests.

Im Folgenden wird aufgezeigt, welche Aspekte im Rahmen der vorliegenden Untersuchung nicht oder zu wenig beachtet wurden. Zugleich bieten diese Einschränkungen Anknüpfungspunkte für künftige Forschungsprojekte.

Zunächst ist der gewählte Fokus der Forschungsarbeit zu nennen: Die Fertigkeiten Hörverstehen und Sprechen wurden ausgewählt, da sie in Echtzeit ablaufen – eine Sprachverwendungssituation, die der S-C-Test durch den Zeitdruck bei der Bearbeitung der Lückentexte nachbildet. Sprachkompetenz besteht jedoch aus sehr viel mehr Aspekten. Neben der Arbeit von FADAEIPOUR und ZOHOORIAN (2017), die das Leseverstehen

fokussieren, sind weitere Untersuchungen, die den Zusammenhang des S-C-Tests mit der Fertigkeit Schreiben sowie der Wortschatz- oder Grammatikkompetenz thematisieren, wünschenswert und unbedingt notwendig, um das Konstrukt des S-C-Tests besser beleuchten zu können. Liegen hierzu Daten vor, wäre zudem eine Faktorenanalyse, wie sie für den herkömmlichen C-Test von ECKES und GROTJAHN (2006) durchgeführt wurde auch für den S-C-Test erstrebenswert. Auf diese Weise könnte das Konstrukt des S-C-Tests weiter ergründet und überprüft werden, ob der S-C-Test beispielsweise für die mündlichen Kompetenzen eine signifikant bessere Anpassungsgüte aufweist als für die schriftlichen Fertigkeiten.

Wie in den Kapiteln 4.4.3.2 bis 4.4.3.5 deutlich wurde, sind die ermittelten Korrelationen sowohl zwischen dem C-Test als auch zwischen dem S-C-Test und den Ergebnissen aus den Tests zum Sprechen und zum Hörverstehen lediglich schwach bis mittelstark. Beim onDaF, der ein erprobtes und gut validiertes Messinstrument ist, fallen die Korrelationen schwächer aus als in den Untersuchungen des TestDaF-Instituts (vgl. ARRAS et al. 2002). Eine mögliche Ursache hierfür ist, dass das hier gewählte Außenkriterium, ein Modellsatz des Goethe-Instituts, womöglich weniger gut validiert wurde als real eingesetzte Testbatterien. Ebenso ist der Zusammenhang von C-Tests mit dem GER noch nicht hinreichend erforscht. Es besteht zudem ein geringes Restrisiko, dass der eine oder andere Proband zufällig schon einmal mit den hier eingesetzten Aufgaben konfrontiert wurde, da der verwendete Modellsatz im Rahmen eines Übungsheftes zur Vorbereitung auf das Goethe-Zertifikat B2 publiziert wurde.

Ein zweiter Faktor in diesem Kontext ist, dass nur Daten von 58 Probanden (Sprechen) bzw. 90 Probanden (Hörverstehen) in die Korrelationsberechnungen einfließen konnten. Das ursprüngliche Ziel, Daten von mindestens 100 Testpersonen zur Berechnung der Korrelationen heranzuziehen, wurde aufgrund von Komplikationen bei der Datenerhebung nicht

erreicht (vgl. Kapitel 4.3.4). Eine Replikation der Studie mit einer größeren Probandengruppe ist daher erstrebenswert.

Der Fokus lag in dieser Studie auf einem spezifischen Sprachniveau, nämlich B2 gemäß GER. In bisherigen Studien, die einen S-C-Test verwendeten, war das Sprachniveau der fremd- und zweitsprachlichen Probanden entweder sehr weit fortgeschritten oder es handelte sich gar um Muttersprachler. Der Fokus auf ein mittleres Sprachniveau ist somit neu. Die Korrelationsanalyse nach rechnerischer Teilung der Probanden in eine stärkere und eine schwächere Gruppe zeigt, dass der Zusammenhang des S-C-Tests mit den getesteten Fertigkeiten Hörverstehen und Sprechen in der Gruppe der stärkeren Deutschlerner deutlich höher ist als in der Gruppe der schwächeren Lerner. Hieraus den Schluss zu ziehen, dass der S-C-Test generell für die niedrigeren Niveaustufen nicht geeignet sei, greift jedoch zu kurz. Um eine fundierte Aussage hierüber machen zu können, sollten S-C-Tests an Lernern der Niveaustufen A1 bis B1 gemäß GER erprobt werden, die ein für diese Zielgruppe adäquates Niveau haben.

Der Zusammenhang der S-C-Test-Leistung mit verschiedenen Maßen von Flüssigkeit mündlicher Sprachproduktion wurde explorativ anhand weniger Probanden untersucht und ließ kein Muster erkennen. Eine systematische Untersuchung der drei Aspekte Flüssigkeit, Korrektheit und Akkuratheit in Zusammenhang mit dem S-C-Test fehlt bislang und ist somit für künftige Forschung von großem Interesse. Dies trifft insbesondere auf die Unterscheidung zwischen monologischer und dialogischer Sprachproduktion zu. Denn wie TAVAKOLI (2016: 141 f.) herausstellt, unterscheidet sich interaktionales Sprechen durch *Turn Taking*, gleichzeitiges Sprechen und gegenseitiges Unterbrechen so stark von monologischem Sprechen, dass bestehende Flüssigkeitsparameter überdacht werden müssen und es fraglich sei, ob diese Flüssigkeitsmaße in beiden Sprecharten überhaupt das gleiche Konstrukt erfassen (vgl. ebd. 146).

Für Probanden sind Datenerhebungen zu Forschungszwecken ausgesprochene *low stakes*-Situationen. So auch hier. Die Teilnahme an der Studie war freiwillig und insbesondere beim Ablegen des S-C-Tests zeigten sich einige Teilnehmer äußerst motiviert. Dennoch bleibt ungewiss, wie viel Mühe sich das Gros der Probanden gegeben hat bzw. ob sie unter größerer Anstrengung zu einer besseren Leistung fähig gewesen wären. Auf der anderen Seite muss auch bedacht werden, dass es unabhängig von der Fremdsprachenkompetenz und dem Grad der Automatisierung sprachlichen Wissens einzelne Personen geben kann, die mit der Testsituation unter enormem Zeitdruck überfordert sind. So verweist HEINE (2017: 130) auf einen Ausreißer in ihrem Datensatz, der unter Niveaubedingungen ein sehr gutes C-Test-Ergebnis erzielt hat (146 von möglichen 150 Punkten), jedoch in der *speed*-Bedingung zum Teil 50 % leere Lücken in der jeweils zweiten Texthälfte aufweist. Sie zieht daraus den Schluss, dass es sich bei diesem Probanden um jemanden handelt, der mit der Testbedingung des Zeitdrucks – unabhängig von der offenbar gut ausgebildeten Sprachkompetenz – große Schwierigkeiten hatte.

Der C-Test wird an zahlreichen Universitäten zu Einstufungszwecken eingesetzt und dabei häufig computeradministriert abgelegt (z. B. am TestDaF-Institut, am Sprachenzentrum der Humboldt-Universität zu Berlin sowie an der ZEMS der TU Berlin), um im Sinne der Testökonomie die sich anschließende Auswertung automatisieren zu können. Die Eingabe der fehlenden Wortteile über eine Tastatur stellt jedoch einen Aspekt dar, dessen Tauglichkeit für den S-C-Test gesondert untersucht werden muss. Zwar haben die Studien von BISPING und RAATZ (2002) und BISPING (2006) nahegelegt, dass Papier-und-Bleistift-C-Tests und computerbasierte C-Tests äquivalent sind, jedoch besteht unter *speed*-Bedingungen die Gefahr, dass die Vertrautheit mit der ggf. anders als in der L1 aufgebauten Tastatur und das Beherrschen des Tippens mit zehn Fingern einen signifi-

kanten Einfluss auf das Ergebnis des C-Tests haben könnten. Dies wäre als testirrelevantes Konstrukt höchst problematisch und würde eine Verwendung des digitalen S-C-Tests ausschließen.

Der Faktor Motorik sollte bei einem Test mit Geschwindigkeitskomponente nicht ganz außer Acht gelassen werden. Die motorische Fertigkeit (ggf. Buchstaben in einer anderen Schrift zu schreiben) wird durch den Zeitfaktor relevant. Eine Variante, um den C-Test noch weiter zu beschleunigen ohne auf die Motorik der Schreibhand oder das Geschick an der Computertastatur Rücksicht nehmen zu müssen, wäre, Probanden den C-Test mündlich lösen zu lassen. Zwar können bei einem solchen Setting Orthographiefehler nicht erkannt werden, dafür könnte aber Aufschluss über die Bearbeitungsprozesse auf der Mikro- und Makroebene beim Lösen eines C-Tests mit und ohne Zeitdruck gegeben werden. Hierzu sollten Daten aus dem Protokoll der mündlichen Lösung mit durch *Eye-Tracking* gewonnenen Daten trianguliert werden. Die Methode des *Eye-Trackings* bietet die Möglichkeit, exakt zu untersuchen, auf welche Satzteile ein Proband seine Aufmerksamkeit richtet. Aus Gründen der Praktikabilität eignet sich dieser Ansatz jedoch ausschließlich zu Forschungszwecken.

Ein weiterer Punkt für künftige Forschung ist die Frage nach einer optimalen Zeitbemessung. In der vorliegenden Studie wurde die Bearbeitungszeit der L2-Probanden durch eine Pilotierung mit L1-Probanden festgelegt. Hierbei wurde für je zwei Texte einer Niveaustufe eine gemeinsame Bearbeitungszeit zwischen 0:52 Minuten und 1:14 Minuten ermittelt (vgl. Kapitel 4.3.2). Mit dieser Zeitbemessung wurde für das eingesetzte S-C-Test-Set eine Reliabilität von $\alpha = 874$ sowie schwache bis mittlere signifikante Korrelationen mit den Fertigkeiten Hörverstehen und Sprechen erzielt. Da der Schwierigkeitsgrad eines jeden C-Test-Texts stark variieren kann, wäre idealerweise eine textspezifische Zeitbemessung anzustreben wie sie von AGUADO et al. (2007) und GROTJAHN et al. (2010)

eingeführt wurde. In der Praxis würde dies jedoch zu umfangreichen Vorerprobungen vor Einsatz eines C-Tests führen. Zudem wäre die Bearbeitungszeit für die Teilnehmer intransparent. Dennoch ist es erstrebenswert, die Zeitbemessung in Abhängigkeit der Textschwierigkeit zu optimieren. In der vorliegenden Studie lag die Bearbeitungszeit für den ersten auf Niveaustufe A2 lösbaren Text bei 0:52 Sekunden. Trotz dieses immensen Zeitdrucks liegt das arithmetische Mittel für diesen Text bei $\bar{x} = 14{,}98$. Für die untersuchte Probandengruppe von B2-Lernern hätte die Bearbeitungszeit daher noch knapper bemessen sein können. Künftige Studien sollten daher daran ansetzen und verschiedene Arbeitszeitbemessungen für ein und denselben C-Test-Text miteinander vergleichen.

Unklar bleibt die hier nicht behandelte Frage, ob der S-C-Test tatsächlich dazu geeignet ist, prozeduralisiertes von deklarativem Sprachwissen zu trennen. Wie in Kapitel 3.1 gezeigt wurde, kann eine hohe Verarbeitungsgeschwindigkeit auch ohne automatisierte Sprache erreicht werden. So bleibt in diesem Kontext zunächst zu klären, welche Rolle die Prozeduralisierung von sprachlichen Entitäten auf den verschiedenen Niveaustufen spielt, und welchen Einfluss sie auf den Erfolg im (kommunikativen) Fremdsprachenunterricht haben, d. h. wie wichtig die Prozeduralisierung für die Fertigkeiten Hörverstehen und Sprechen tatsächlich ist.

6 Fazit

Der C-Test ist seit mehr als 35 Jahren als ein objektives, reliables, valides und ökonomisches Messinstrument zum Testen allgemeiner Sprachkompetenz bekannt. Wenngleich er mit Hinblick auf die Sprachen Englisch und Deutsch entwickelt wurde, gibt es inzwischen Versionen in vielen verschiedenen Sprachen. Teilweise funktioniert das kanonische Prinzip des C-Tests gut, teilweise wird nach Variationen des Tilgungsschemas gesucht, um dem andersartigen Sprachbau gerecht zu werden.

Die Einsatzmöglichkeiten des C-Tests sind zahlreich: In Form des onDaF als ein *Screening*-Instrument für den TestDaF. Er wird an universitären Sprachenzentren eingesetzt, um Sprachkursinteressenten dem für sie passenden Kursniveau zuzuweisen. Im schulischen Bereich, insbesondere im Bereich der Sprachbildung/Deutsch als Zweitsprache, wird der C-Test als Diagnoseinstrument sowie zum Zweck der Sprachförderung eingesetzt. In zahlreichen Studien wird der C-Test als Forschungsinstrument genutzt, beispielsweise um den Sprachstand der Probanden zu ermitteln.

Je nach Zielsetzung und Einsatzgebiet existieren verschiedene Varianten des C-Tests. So wird bei Schülern teilweise nur jedes dritte Wort getilgt, um mehr Kontext zu erhalten und den Text leichter zu machen. Andererseits nutzen Forscher C-Tests mit *speed*-Komponente, um bei muttersprachlichen Probanden den Schwierigkeitsgrad zu erhöhen und einen Deckeneffekt zu verhindern.

Die vorliegende Studie hat den *speeded*-C-Test in den Fokus genommen und untersucht, ob er gegenüber dem herkömmlichen *power*-C-Test ein besserer Indikator dafür ist, wie ein Testteilnehmer die Fertigkeiten Hörverstehen und Sprechen beherrscht, da diese Kompetenzen in einem modernen, kommunikativ orientierten Fremdsprachenunterricht unab-

dingbar sind, um erfolgreich am Unterrichtsgeschehen teilnehmen zu können. Mit der Beantwortung dieser Frage wird auch ein Beitrag zur Erforschung des Konstrukts des C-Tests geleistet.

Es wurde gezeigt, dass der S-C-Test eine sehr zufriedenstellende Reliabilität aufweist, selbst in der auf B2-Niveau relativ homogenen Probandengruppe. Darüber hinaus wurden Korrelationen zwischen dem S-C-Test und den Fertigkeiten Hörverstehen und Sprechen gefunden. In Bezug auf das Hörverstehen weisen die analysierten Daten darauf hin, dass der S-C-Test dem C-Test überlegen ist. Für das Sprechen ist das gezeichnete Bild weniger deutlich: Zwar zeigt sich auch hier beim S-C-Test ein stärkerer Zusammenhang als beim gewöhnlichen C-Test; die gefundenen Zusammenhänge sind jedoch für beide C-Test-Versionen nur schwach. Bei einer Differenzierung nach monologischem und dialogischem Sprechen zeigt sich ein stärkerer Zusammenhang des S-C-Tests mit dialogischem als mit monologischem Sprechen.

Eine Aufteilung aller Probanden in eine stärkere und eine schwächere Hälfte führte zu dem Ergebnis, dass die Korrelationen sowohl für das Hörverstehen als auch für das Sprechen in der Gruppe der weiter fortgeschrittenen Lerner deutlich stärker sind.

Hinweise auf einen Zusammenhang zwischen der Leistung im S-C-Test und einem höheren Maß an Flüssigkeit in der mündlichen Sprachproduktion konnten anhand einer sehr kleinen Auswahl von Probanden nicht gefunden werden.

Zusammenfassend lässt sich festhalten, dass der S-C-Test sowohl Aspekte des Hörverstehens als auch des Sprechens erfasst. In welchem Umfang sich die Konstrukte der verwendeten Tests überlappen, kann aufgrund der verhältnismäßig kleinen Probandenzahl nicht abschließend gesagt werden.

Wenngleich die Augenscheinvalidität des S-C-Tests nicht gut ist, konnte kein Zusammenhang zwischen der Einstellung zum Testformat und dem Ergebnis der Probanden im S-C-Test gefunden werden. Auch das Schriftsystem, mit dem die Testteilnehmer alphabetisiert worden sind, zeigte in dieser Studie keinen Zusammenhang mit dem Testergebnis der Teilnehmer.

Durch Auswertung des Fragebogens konnte zudem ermittelt werden, dass die Probanden beim Lösen des S-C-Tests tendenziell weniger vor- und zurückspringen und die Lücken eher linear bearbeiten.

Während eine Vergleichbarkeit der hier durchgeführten Untersuchung mit vorangegangenen Studien (vgl. Kapitel 3.2) aufgrund der unterschiedlichen Populationen, Niveaustufen, Messinstrumente und Außenkriterien nur sehr eingeschränkt möglich ist, deuten die zusammengefassten Ergebnisse darauf hin, dass der S-C-Test eine berechtigte und valide Variante des C-Tests darstellt, die dem herkömmlichen C-Test in gewissem Maße sogar überlegen ist, insbesondere bezüglich der Testökonomie.

Im kommunikativen Fremdsprachenunterricht haben die Fertigkeiten Hörverstehen und Sprechen gegenüber den klassischen Vermittlungsmethoden an Relevanz gewonnen. Die Fremdsprache wird in Echtzeit verwendet. Ein C-Test mit Geschwindigkeitskomponente kann im Sinne der *predictive validity* die Einstufung in Sprachkurse zuverlässiger machen, da er das, was im Unterricht von den Lernern verlangt wird, besser abbilden kann als ein *power*-C-Test.

C-Tests werden für gewöhnlich nicht in *high stakes*-Testsituationen eingesetzt. Die Geschwindigkeitskomponente beim C-Test hat somit keine weitreichenden Folgen, sollte ein Testteilnehmer hierdurch Schwierigkeiten beim Ablegen des Tests erfahren. Auf der anderen Seite bedeutet die drastische Reduktion der Testzeit für Sprachenzentren und andere Testabnahmestellen eine enorme Einsparung von Ressourcen. Ein Ablösen des

C-Tests durch den S-C-Test ist weder angestrebt noch sinnvoll: Insbesondere wenn es um den diagnostischen Einsatz von C-Tests geht, sollte dieser als reiner Niveautest, also ohne zusätzlichen Zeitdruck, eingesetzt werden.

Der *speeded*-C-Test hat sich somit als eine gangbare Variante des C-Tests erwiesen, die in bestimmten Kontexten dem herkömmlichen C-Test gleichgestellt, wenn nicht gar überlegen ist. Dennoch bleiben zahlreiche Fragen zum S-C-Test, seinem Konstrukt und seinem Potential offen, was einen guten Ausgangspunkt für weitere Forschungsprojekte liefert.

Literaturverzeichnis

AGUADO, Karin (2003): „Kognitive Konstituenten mündlicher Produktion in der Fremdsprache: Aufmerksamkeit, Monitoring, Automatisierung". In: *Fremdsprachen Lehren und Lernen* (FluL), 32. 11–26.

AGUADO, Karin/GROTJAHN, Rüdiger/SCHLAK, Torsten (2005): „Erwerbsalter und Sprachlernerfolg: Theoretische und methodologische Grundlagen eines empirischen Forschungsprojekts." In: *Zeitschrift für Fremdsprachenforschung*, 16 (2). Baltmannsweiler: Schneider Verlag Hohengehren. 275–293.

AGUADO, Karin/GROTJAHN, Rüdiger/SCHLAK, Torsten (2007): „Erwerbsalter und Sprachlernerfolg: Zeitlimitierte C-Tests als Instrument zur Messung prozeduralen sprachlichen Wissens". In: Vollmer, Helmut J. (Hrsg.): *Synergieeffekte in der Fremdsprachenforschung. Empirische Zugänge, Probleme, Ergebnisse*, 27. Kolloquium Fremdsprachenunterricht. Frankfurt am Main: Peter Lang. 137–149.

AHRENHOLZ, Bernt (2010): „Erstsprache – Zweitsprache – Fremdsprache". In: Ahrenholz, Bernt & Oomen-Welke, Ingelore (Hrsg.): *Deutsch als Zweitsprache.* Baltmannsweiler: Schneider Verlag Hohengehren. 3–16.

ALDERSON, J. Charles (1979): „The Effect on the Cloze Test of Changes in Deletion Frequency". In: *Journal of Research in Reading*, 2 (2). 108–119.

ALDERSON, J. Charles (1983): „The Cloze Procedure and Proficiency in English as a Foreign Language". In: Oller, John W. (Hrsg.): *Issues in Language Testing Research.* Rowley/Massachusetts: Newbury House Pubishers. 205–217.

ALTE (Association of Language Testers in Europe) (2005): *Materials for the Guidance of Test Item Writers.* Online: http://www.alte.org/attachments/files/item_writer_guidelines.pdf (Zugriff 17.09.2014)

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (Hrsg.) (2002): *Standards for educational and psychological testing.* Washington D. C.

ANCKAERT, Philippe & BEECKMANS, Renaud (1992): „Le C-Test. Difficulté intrinsèque, pouvoir discriminant et validité de contenu". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Bochum: Brockmeyer. 145–172.

ANDERSON, D. F. (1953): „Tests of Achievement in the English Language". In: *English Language Teaching*. Vol. 7, No. 2. 37–69.

ANDERSON, John R. (1983): „A spreading activation theory of memory". In: *Journal of Verbal Learning and Verbal Behavior*, 22. 261–295.

ANDERSON, John R. & LEBIERE, Christian (1998): „Introduction". In: Anderson, John R. & Lebiere, Christian (Hrsg.): *The atomic components of thought*. Mahwah, NJ : Erlbaum. 1–17.

ARISTOTELES (1995): *Philosophische Schriften in sechs Bänden.* Band 6. P*hysik: Vorlesung über die Natur*. Übersetzt von Hans Günter Zekl. Hamburg: Felix Meiner Verlag.

ARRAS, Ulrike/ECKES, Thomas & GROTJAHN, Rüdiger (2002): „C-Tests im Rahmen des „Test Deutsch als Fremdsprache" (TestDaF): Erste Forschungsergebnisse". In: Grotjahn, Rüdiger (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Vol. 4. Bochum: AKS-Verlag. 175–209.

ARRAS, Ulrike & GROTJAHN, Rüdiger (1994): „Der C-Test im Chinesischen". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer: 1–60.

BABAII, Esmat & SHAHRI, Somaliyeh (2010): „Psychometric rivalry: The C-test and the cloze test interacting with test takers' characteristics". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Beiträge aus der aktuellen Forschung*. Frankfurt am Main: Peter Lang. 41–56.

BACH, Gerhard & VIEBROCK, Britta (2012): „Was ist erlaubt? Ethik in der Fremdsprachenforschung." In: Doff, Sabine (Hrsg.): *Fremdsprachenunterricht empirisch erforschen. Grundlagen – Methoden – Anwendung*. Tübingen: Narr Verlag. 17–33.

BACHMAN, Lyle F. (1990): *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

BAGHAEI, Purya (2014): „Construction and Validation of a C-Test in Persian". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen*. Frankfurt am Main: Peter Lang. 299–312.

BAGHAEI, Purya & GROTJAHN, Rüdiger (2014): „The Validity of C-Tests as Measures of Academic and Everyday Language Proficiency: A Multidimensional Item Response Modeling Study". In: Grotjahn, Rüdiger (Hrsg.) *Der C-Test. Aktuelle Tendenzen*. Frankfurt am Main: Peter Lang. 163–171.

BAMBERG, Günter/BAUR, Franz & KRAPP, Michael (2012): *Statistik*. 17. überarbeitete Auflage. München: Oldenbourg Verlag.

Bärenfänger, Olaf (2002): „Automatisierung der mündlichen L2-Produktion: Methodische Überlegungen". In: Börner, Wolfgang & Vogel, Klaus (Hrsg.): *Grammatik und Fremdsprachenerwerb. Kognitive, psycholinguistische und erwerbstheoretische Perspektiven*. Tübingen: Gunter Narr. 119–140.

Baur, Rupprecht S./Goggin, Melanie & Wrede-Jackes, Jennifer (2013): *Der C-Test: Einsatzmöglichkeiten im Bereich DaZ*. Universität Duisburg-Essen: pro-DaZ. http://www.uni-due.de/imperia/md/content/prodaz/c_test_einsatzmoeglichkeiten_daz.pdf (Zugriff 10.08.2016)

Baur, Rupprecht S./Mashkovskaya, Anna/Spettmann, Melanie (2010): „Der C-Test als Instrument zur Ermittlung allgemeinsprachlicher und fachsprachlicher Fähigkeiten im Berufskolleg." In: Berndt, Annette & Kleppin, Karin (Hrsg.): *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn*. Frankfurt am Main: Peter Lang. 23–38.

Baur, Rupprecht S. & Meder, Gregor (1994): „C-Tests zur Ermittlung der globalen Sprachfertigkeit im Deutschen und in der Muttersprache ausländischer Schüler in der Bundesrepublik Deutschland". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 151–178.

Baur, Rupprecht & Spettmann, Melanie (2008a): „Screening – Diagnose – Förderung: Der C-Test im Bereich DaZ". In: Ahrenholz, Bernt Ahrenholz, Bernt (Hrsg.): *Deutsch als Zweitsprache: Voraussetzungen und Konzepte für die Förderung von Kindern und Jugendlichen mit Migrationshintergrund*. Freiburg im Breisgau: Fillibach. 95–110.

Baur, Rupprecht S. & Spettmann, Melanie (2008b): „Sprachstandsmessung und Sprachförderung mit dem C-Test". In: Ahrenholz, Bernt & Oomen-Welke, Ingelore (Hrsg.): *Deutsch als Zweitsprache*. Baltmannsweiler: Schneider Verlag Hohengehren. 430–441.

Berger, Carina & Zimmermann, Kerstin (in Vorbereitung): „Sind Linkshänder die besseren C-Test-Löser?".

Bickley, A. C./Ellington, Billie J. & Bickley, Rachel T. (1970): „The Cloze Procedure: A Conspectus". In: *Journal of Literacy Research,* Vol 2, Issue 3. 232 –249.

Bisping, Meikel (2006): „Zur Validität von Computer-C-Tests." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen*. Frankfurt am Main: Peter Lang. 149–166.

BISPING, Meikel & RAATZ, Ulrich (2002): „Sind Computerisierte und Papier&Bleistift-Versionen des C-Tests äquivalent?“ In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: AKS-Verlag. 131–155.

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

BORTZ, Jürgen & DÖRING, Nicola (2002): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 3. Auflage. Berlin et al.: Springer.

BORTZ, Jürgen & DÖRING, Nicola (2006): *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 4. überarbeitete Auflage. Berlin et al.: Springer.

BRANDSTÄTTER, Eduard (1999): „Konfidenzintervalle als Alternative zu Signifikanztests“. In: *Methods of Psychological Research Online*, Vol. 4, No. 2. 1–17.

BRICKENKAMP, Rolf (1981): *Test d2: Aufmerksamkeits-Belastungs-Test. Handanweisung. Durchführung, Auswertung, Interpretation*. Göttingen et al.: Verlag für Psychologie, Hogrefe.

BROWN, James D. (1994): „A closer look at cloze: Validity and reliability“. In: Oller, John W. & Jonz, Jon (Hrsg.): *Cloze and coherence*. Lewisburg, PA: Associated University Presses. 189–196. [Reprinted by permission from the original: Brown, James D. (1983).]

BROWN, James D. (2002): „Do Cloze Tests Work? Or, is it just an Illusion?“. In: *Second Language Studies*, 21 (1). 79–125.

BÜHL, Achim (2006): *SPSS 14. Einführung in die moderne Datenanalyse.* Hallbergmoos: Pearson. 10. überarbeitete und erweiterte Auflage.

BÜHNER, Markus (2011): *Einführung in die Test- und Fragebogenkonstruktion*. 3. aktualisierte und erweiterte Auflage. München et al.: Pearson Studium.

BÜHNER, Markus (2006): *Einführung in die Test- und Fragebogenkonstruktion*. 1. Auflage. München et al.: Pearson Studium.

CAI, Hongwen (2012): „Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis“. In: *Language Testing*, 30 (2). 177–199

CAPREZ-KOMPÀK, Edina & GÖNÇ, Mesut (2006): „Der C-Test im Albanischen und Türkischen: Theoretische Überlegungen und empirische Befunde“. In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen*. Frankfurt/Main: Peter Lang. 243–260.

CHEN, Jing (2011): „Language assessment: Its development and future — An Interview with Lyle F. Bachman“. In: *Language Assessment Quarterly*, 8 (3). 277–290.

CHIHARA, Tetsuro/CLINE, William D. & SAKURAI, Toshiko (1996): „If the cloze test is a question , is the C-test the answer?". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Band 3. Bochum: Brockmeyer. 183–195.

CRANNEY, A. Garr (1972): „The Construction of Two Types of Cloze Reading Tests for College Students". In: *Journal of Literacy Research* (5). 60–64.

CAULFIELD, Joan & SMITH, William C. (1981): „The Reduced Redundancy Test and the Cloze Procedure as Measures of Global Language Proficiency". In: *The Modern Language Journal*, 65. 54–58.

CHAPELLE, Carol A. & ABRAHAM, Roberta G. (1990): „Cloze-Method: What difference does it make?" In: *Language Testing.* No. 7, (2). 121–146.

COHEN, Andrew/SEGAL, Michael & BAR-SIMAN-TO, Ronit (1984): „The C-test in Hewbrew". In: *Language Testing*, 1. 221–225.

COLEMAN, James A. (1994): „Profiling the advanced language learner: The C-Test in British further and higher education." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen. Band 2.* Bochum: Brockmeyer. 217–237.

CORRIGAN, A./ DOBSON, Barbara/KELLMAN, E./SPAAN, Mary W. & TYMA, S. (1978): „English Placement Test (Form B)". Ann Arbot: Testing and Certification Division, University of Michigan.

CUMMINS, Jim (1979): „Linguistic interdependence and the educational development of bilingual children". In: *Review of Educational Research* 49/79. 222–251.

DALLER, Helmut (1996): „Der C-Test als Meßinstrument alltagssprachlicher und akademischer Sprachfähigkeiten türkischer Remigranten". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 343–366.

DALLER, Helmut & GROTJAHN, Rüdiger (1999): „The Language Proficiency of Turkish returnees from Germany: An Empirical Investigation of Academic and Everyday Language Proficiency.". In: *Language, Culture and Curriculum*, 12 (2). 156–172.

DALLER, Helmut/TREFFERS-DALLER, Jeanine/ÜNALDI-CEYLAND, Aylin & YILDIZ, Cemal (2002): „The Development of a Turkish C-Test". In: Coleman, James A./Grotjahn, Rüdiger & Raatz, Ulrich (Hrsg.): *University Language Testing and the C-Test*. Bochum: AKS-Verlag. 187–199.

DALLER, Helmut & PHELAN, David (2006): „The C-test and TOEIC as measures of students' progress in intensive short courses in EFL." In Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen*. Frankfurt am Main: Peter Lang. 101–119.

DEKEYSER, Robert M. (1997): „Beyond Explicit Rule Learning: Automatizing Second Language Morphosyntax". In: *Studies in Second Language Acquisition*, 19. 195–221.

DÖRNYEI, Zoltán & KATONA, Lucy (1992): „Validation of the C-test amongst Hungarian EFL learners". In: *Language Testing* 9: 187. 187–206.

DRACKERT, Anastasia (2015): *Validating Language Proficiency Assessment in Second Language Acquisition Research. Applying an Argument-Based Approach.* Frankfurt am Main: Peter Lang.

ECKES, Thomas. (2007): „Konstruktion und Analyse von C-Tests mit Ratingskalen-Rasch-Modellen". In: *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: Peter Lang. 125–192.

ECKES, Thomas (2014): „Die onDaF-TestDaF-Vergleichsstudie: Wie gut sagen Ergebnisse im onDaF Erfolg oder Misserfolg beim TestDaF vorher?". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen*. Frankfurt am Main: Peter Lang. 137–162.

ECKES, Thomas & GROTJAHN, Rüdiger (2006): „A closer look at the construct validity of C-tests". In: *Language Testing* 23: 3. 290–325.

EDUCATIONAL TESTING SERVICE (Hrsg.) (2015): *TOEIC Examinee Handbook: Listening & Reading*. Princeton, NJ. (Abrufbar unter https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf) (Zugriff 14.04.2016)

EID, Michael/GOLLWITZER, Mario & SCHMITT, Manfred (2015): *Statistik und Forschungsmethoden*. 4. überarbeitete und erweiterte Auflage. Weinheim & Basel: Beltz.

EISENBERG, Peter (2017): „Das missbrauchte Geschlecht". In: *Süddeutsche Zeitung*, Ausgabe vom 2. März 2017. (http://www.sueddeutsche.de/kultur/essay-das-missbrauchte-geschlecht-1.3402438) (Zugriff 11.04.2017)

EUROPARAT. Rat für kulturelle Zusammenarbeit. (Hrsg.) (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen*. Berlin et al.: Langenscheidt.

FADAEIPOUR, Anita & ZOHOORIAN, Zahra (2017): „Comparing the Psychometric Characteristics of Speeded and Standard C-Tests". In: I*nternational Journal of Language Testing*. Vol. 7, No. 1. 40–50.

FAISTAUER, Renate (2001): „Zur Rolle der Fertigkeiten". In: Helbig, Gerhard/Götze, Lutz/Henrici, Gerd & Krumm, Hans-Jürgen (Hrsg.): *Deutsch als Fremdsprache. Ein internationales Handbuch*. 2. Halbband. Berlin & New York: DeGruyter. 864–871.

FREESE, Hans-Ludwig (1994): „Was mißt und was leistet ‚Leistungsmessung mittels C-Tests'?". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 305–311.

FREY, Evelyn (2012): *Fit fürs Goethe-Zertifikat B2. Prüfungstraining*. Ismaning: Hueber.

GERMANN, Ulrich (1996): „C-Tests automatisch erstellen – mit Word für Windows 6.0". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 419–434.

GERMANN, Ulrich & GROTJAHN, Rüdiger. (1994): „Das Lösen von C-Tests Auf Dem Computer: Eine Pilotuntersuchung zu Den Bearbeitungsprozessen." In: Grotjahn, Rüdiger: *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 279–304.

GOETHE-INSTITUT (Hrgs.) (2007): *Goethe-Zertifikat B2. Prüfungsziele. Testbeschreibung*. München. Abrufbar unter: https://www.goethe.de/resources/files/pdf62/Pruefungsziele_Testbeschreibung_B2.pdf

GRADMAN, Harry L. & SPOLSKY, Bernard (1975): „Reduced Redundancy as a Tool. A Porgress Report.". In: Jones, Randall L. & Spolsky, Bernard (Hrsg.): *Testing Language Proficiency*. Arlington, Va.. 59–70.

GRIESBACH, Heinz & SCHULZ, Dora (1967): *Deutsche Sprachlehre für Ausländer: Grundstufe in einem Band*. 2. Neuauflage. München: Hueber.

GROTJAHN, Rüdiger (1986): „Der Bochumer Einstufungstest ‚Französisch'". In: Bausch, Karl-Richard/Königs, Frank G./Kogelheide, Rainer (Hrsg.): *Probleme und Perspektiven der Sprachlehrforschung. Bochumer Beiträge zum Fremdsprachenunterricht in Forschung und Lehre.* Frankfurt am Main: Scriptor. 313–324.

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

GROTJAHN, Rüdiger (1995): „Der C-Test: State of the Art". In: *Zeitschrift für Fremdsprachenforschung,* 6 (2). 37–60.

GROTJAHN, Rüdiger (1996): „'Scrambled' C-Tests. Untersuchungen zum Zusammenhang zwischen Lösungsgüte und sequentieller Textstruktur". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 95–125.

GROTJAHN, Rüdiger (2000): *Studieneinheit Leistungsmessung und Leistungsbeurteilung*. Patras: Hellenic Open University. Online abrufbar unter: http://herder.philol.uni-leipzig.de/temp/lehrende/schirnetr/testen/grundlag.pdf (Zugriff 24.09.2016)

GROTJAHN, Rüdiger (2002): „Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Vol. 4. Bochum: AKS-Verlag. 211–225.

GROTJAHN, Rüdiger (2004): „Der C-Test: Aktuelle Entwicklungen". In: Wolff, Armin/ Ostermann, Torsten & Chlosta, Christoph (Hrsg.), *Integration durch Sprache. Beiträge der 31. Jahrestagung DaF 2003*. Regensburg: Fachverband Deutsch als Fremdsprache. 535–550.

GROTJAHN, Rüdiger (2010): „Gesamtdarbietung, Einzeltextdarbietung, Zeitbegrenzung und Zeitdruck: Auswirkungen auf Item- und Testkennwerte und C-Test-Konstrukt". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Beiträge aus der aktuellen Forschung*. Frankfurt am Main: Peter Lang. 265–296.

GROTJAHN, Rüdiger (2011): „C-Tests – Aspekte der Validität". In *Deutsch als Fremdsprache,* 48 (3). 131–137.

GROTJAHN, Rüdiger (2012): C-Test. In: Byram, Michael & Hu, Adelheid (Hrsg.): *Routledge Encyclopedia of Language Teaching and Learning*. London & New York: Routledge. 2. überarbeitete Auflage. 180–181.

GROTJAHN, Rüdiger (2014): „The C-Test bibliography: version January 2014". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen.* Frankfurt am Main: Peter Lang. 325–363

GROTJAHN, Rüdiger/TÖNSHOFF, Wolfgang & HOHENBLEICHER, Heike (1994): „Der C-Test im Italienischen. Theoretische Überlegungen und empirische Analysen." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 115–149.

GROTJAHN, Rüdiger/KLEIN-BRALEY, Christine & RAATZ, Ulrich (2002): „C-Tests: an overview". In: Coleman, James/ Grotjahn, Rüdiger/ Raatz, Ulrich (Hrsg.): *University Language Testing and the C-Test.* Bochum: AKS-Verlag. 93–114.

GROTJAHN, Rüdiger/SCHLAK, Torsten & AGUADO, Karin (2010): „S-C-Test: Messung automatisierter sprachlicher Kompetenzen anhand von C-Tests mit massiver textspezifischer Zeitlimitierung". In: Grotjahn, Rüdiger (Hrsg.): D*er C-Test: Beiträge aus der aktuellen Forschung.* Frankfurt am Main: Peter Lang. 297–319.

GROTJAHN, Rüdiger/TÖNSHOFF, Wolfgang & HOHENBLEICHER, Heike (1994): „Der C-Test im Italienischen. Theoretische Überlegungen und empirische Analysen". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*, Vol. 2. Bochum: Brockmeyer. 115–149.

GROTJAHN, Rüdiger & STEMMER, Brigitte (1984): „Entwicklung und Einsatz eines C-Tests ‚Französisch'". In: Kühlwein, Wolfgang (Hrsg.): *Sprache, Kultur, Gesellschaft. Kongreßberichte der 14. Jahrestagung der Gesellschaft für Angewandte Linguistik, GAL e. V.* Tübingen: Narr: 101–102.

GUMMICH, Verena (1997): *C-Test-Leistung, Schultyp, Schulstufe*. Unveröffentlichte Diplomarbeit im 2. Nebenfach Psychologie, Gerhard Mercator-Universität-GH-Duisburg: Integrierter Studiengang Sozialwissenschaften.

HABER, Fred (1974): *An introduction to information and communication theory.* Reading/Massachusetts: Addison-Wesley Publishing Company.

HAGEMANN, Jörg & HENLE, Julia (2014): „Transkribieren nach GAT 2 (Minimal- und Basistranskript) - Schritt für Schritt". [https://www.ph-freiburg.de/fileadmin/dateien/mitarbeiter/hagemannfr/Transkribieren_nach_GAT_2.pdf] (Zugriff 05.06.2018)

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* (Vol. 4). Bochum: AKS-Verlag. 53–66.

HEINE, Simone (2017): *Fremd- und Zweitsprachenlernerfolg und seine Erklärung durch Erwerbsalter, kognitive, affektiv-motivationale und sozio-kulturelle Variablen. Eine empirische Studie*. Kassel: kassel university press.

HILL, Thomas & LEWICKI, Pawel (2006): *Statistics: Methods and Applications : a Comprehensive Reference for Science, Industry, and Data Mining*. Tulsa: StatSoft.

HINOFOTIS, Frances B. (1980): „Cloze as an Alternative Method of ESL Placement and Proficiency Testing". In: Oller, John W. & Perkins, Kyle (Hrsg.): *Research in Language Testing*. Rowley/Massachusetts: Newbury House Publishers. 121–128.

HINOFOTIS, Frances B. & SNOW, Becky G. (1980): „An Alternative Cloze Testing Procedure: Multiple-Choice Format". In: Oller, John W. & Perkins, Kyle (Hrsg.): *Research in Language Testing*. Rowley/Massachusetts: Newbury House Publishers. 129–133.

HUHTA, Ari (1996): „Validating an EFL C-test for Students of English Philology." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und Praktische Anwendungen.* Vol. 3. Bochum: Brockmeyer. 97–234.

HOCHSCHULREKTORENKONFERENZ & KULTUSMINISTERKONFERENZ (2011): *Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen (RO-DT).* (Abrufbar unter http://www.kmk.org/fileadmin/Dateien/pdf/ZAB/Hochschulzugang_Beschluesse_der_KMK/Rahmenordnung_dtSprachp.pdf) (Zugriff 03.03.2014)

HORN, Wolfgang (1983): *L-P-S Leistungsprüfsystem.* 2. Auflage. Göttingen: Hogrefe.

HÖTTECKE, Dietmar/EHMKE, Timo/KRIEGER, Claus & KULIK, Marta Anna (2017): „Vergleichende Messung fachsprachlicher Fähigkeiten in den Domänen Physik und Sport". In: *Zeitschrift für Didaktik der Naturwissenschaften.*

HOUSEN, Alex/KUIKEN, Folkert & VEDDER, Ineke (2012): „Complexity, accurancy and fluency. Definitions, measurement and research". In: Housen, Alex/Kuiken, Folkert & Vedder, Ineke (Hrsg.): *Dimensions of L2 performance. Complexity, accuracy and fluency in SLA.* Amsterdam/Philadelphia: John Benjamins. 1–20.

HUNEKE, Hans-Werner & STEINIG, Wolfgang (2010): *Deutsch als Fremdsprache. Eine Einführung.* Berlin: Erich Schmidt Verlag. 5. Auflage.

IRVINE, Patricia/ATAI, Parvin & OLLER, John W. (1974): „Cloze, Dictation, and the Test of English as a Foreign Language". In: *Language Learning*, Vol. 24, No. 2. 245–252.

JAFARPUR, Abdoljavad (1995): „Is C-testing superior to cloze?". In: *Language Testing*, 12. 194–216.

JAFARPUR, Abdoljavad (2002): „A Comparative Study of a C-Test and a Cloze Test." In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und Praktische Anwendungen. Vol. 4.* Bochum: AKS-Verlag. 31–51.

JAKSCHIK, Gerhard (1992): „Zum Einsatz des C-Tests in den Psychologischen Diensten der Arbeitsämter. Ein C-Test für Deutsch als Zweitsprache". In: Grotjahn, Rüdiger (Hrsg.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Vol. 1. Bochum: Brockmeyer. 297–311.

JAKSCHIK (1994): „Der C-Test für erwachsene Zweitsprachler als Einstufungsinstrument bei der Schulausbildung". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Bochum: Brockmeyer. 259–278.

JAKSCHIK, Gerhard (1996): „Validierung des C-Tests für erwachsene Zweitsprachler. Eine Längsschnittuntersuchung bei Trägern von beruflichen Bildungsmaßnahmen". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen.* Vol. 3. Bochum: Brockmeyer. 235–277.

JAKSCHIK, Gerhard & KLEMMERT, Hella (2006): „Erste Erprobung eines Multiple Choice C-Tests". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen*. Frankfurt am Main: Peter Lang. 195–210.

JOHANSSON, Stig (1973): „An Evaluation of the Noise Test: A Method for Testing Overall Second Language Proficiency by Perception under Masking Noise". In: *IRAL*, 11 (2). 107–133.

JANSSEN, Jürgen & LAATZ, Wilfried (2013): *Statistische Datenanalyse mit SPSS für Windows*. Hamburg: Springer.

JANSSEN, Jürgen & LAATZ, Wilfried (2017): *Statistische Datenanalyse mit SPSS für Windows*. Hamburg: Springer. 9. Auflage.

KIENCKE, Uwe & EGER, Ralf (2008): *Systemtheorie für Elektrotechniker.* 7. Auflage. Berlin & Heidelberg: Springer.

KLEIN-BRALEY, Christine (1983): „A Cloze is a Cloze is a Question". In: Oller, John W. (Hrsg.): *Issues in Language Testing.* Rowley: Newbury House Publishers. 218–228.

KLEIN-BRALEY, Christine (1985a): „A Cloze-Up on the C-Test: A Study in the Construct Validation of Authentic Tests." In: *Language Testing*, 76 (2).76–104.

KLEIN-BRALEY, Christine (1985b): „C-Tests and Construct Validity". In: *Fremdsprachen und Hochschule* 13/14. Thematischer Teil: C-Tests in der Praxis. Bochum: AKS-Verlag. 55–65.

KLEIN-BRALEY, Christine (1996): „Hunting unicorns: C-Tests, cloze and the RIP (Removal of Information Procedure)". In: Grotjahn, Rüdiger (Hrsg): D*er C-Test. Theoretische Grundlagen und praktische Anwendungen*. Vol. 3. Bochum: Brockmeyer. 127–137.

KLEIN-BRALEY, Christine (1997): „C-Tests in the Context of Reduced Redundancy Testing. An Appraisal." In: *Language Testing*, 14 (1). 47–84.

KLEIN-BRALEY, Christine & RAATZ, Ulrich (1984): „A Survey of Research on the C-Test". In: *Language Testing* (1). 134–146.

KNIFFKA, Gabriele & LINNEMANN, Markus (2014): „A German C-Test for Migrant Children". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen*. Frankfurt am Main: Peter Lang. 239–259.

KOKKOTA, Valmar (1988): „Letter-Deletion Procedure: A Flexible Way of Reducing Text Redundancy". In: *Language Testing*, 5 (1). 115–119.

KOLLER, Gerhard & ZAHN, Rosemary (1996): „Computer Based Construction and Evaluation of C-Tests". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen*. Bochum: Brockmeyer. 401–418.

KONTRA, Edit H. & KORMOS, Judit (2006): „Strategy Use and the Construct of C-Tests". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen*. Frankfurt am Main: Peter Lang. 122–138.

KOW YIP CHENG, Karen/BIGLAR BAYGI, Amir & SOLAYMANI, Mesod (2009): „The Effect of Test Authenticity on the Performance of Iranian EFL Students in a C-Test". In: *Research in Language*, Vol. 7. 61–74.

KRAUTH, Joachim (1995): *Testkonstruktion und Testtheorie*. Weinheim: Beltz.

KREKELER, Christian (2002): „TestDaF und DSH – ungleiche Sprachtests im Vergleich." In: *EliSe*, 2(2). 19–50.

KUJAWKA, Justyna (2011): *Die Einschätzung von fremdsprachlichen Kompetenzen anhand eines C-Tests für die polnische Sprache.* Schriftliche Hausarbeit für die Masterprüfung, Ruhr-Universität Bochum, Fakultät für Philologie, Seminar für Sprachlehrforschung, Bochum.

LADO, Robert (1961): *Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book.* London: Longmans.

LARSEN-FREEMAN, Diane (2000): *Techniques and Principles in Language Teaching*. Oxford: Oxford University Press.

LARSEN-HALL, Jenifer (2010): *A Guide to Doing Statistics in Second Language Acquisition Research using SPSS.* New York et al.: Routledge.

LAUBER, Julia (2013): *Das Sprachverhalten bilingualer Paare – eine qualitative Studie zur Kommunikation in deutsch-spanischen Beziehungen.* Unveröffentlichte Masterarbeit am Fachgebiet Deutsch als Fremdsprache. Berlin: Technische Universität.

LEGENHAUSEN, Lienhardt (1989): „Zur face validity des C-Tests". In: Finkenstaedt, Thomas & Schröder, Konrad: *Zwischen Empirie und Machbarkeit. Erstes Symposium zum Bundeswettbewerb Fremdsprachen*. Augsburg: Universitätsverlag Augsburg. 70–81.

LEGUTKE, Michael K. & SCHRAMM, Karen (2016): „Forschungsethik". In: Caspari, Daniela/Klippel, Friederike/Legutke, Michael & Schramm, Karen (Hrsg.): *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch.* Tübingen: Narr. 108–117.

LEI, Lei (2008): „Validation of the C-Test among Chinese ESL learners". In: *The Journal of Asia TEFL*, Vol. 5, No. 2. 117–140.

LENHARD, Wolfgang & LENHARD, Alexandra (2014): *Signifikanztests bei Korrelationen*. Verfügbar unter https://www.psychometrica.de/korrelation.html. Bibergau: Psychometrica. DOI: 10.13140/RG.2.1.2954.1367 (Zugriff 06.05.2018)

LENNON, Paul (1990): „Investigating fluency in EFL: A quantitative approach." In: *Language Learning*, 40 (3). 387–417.

LIENERT, Gustav A. & RAATZ, Ulrich (1998): *Testaufbau und Testanalyse.* (6. Auflage). Weinheim: Beltz, Psychologie Verlags Union.

LIST, Gudula (2002): „‚Wissen' und ‚Können' beim Spracherwerb – dem ersten und den weiteren". In: Barkowski, Hans & Faistauer, Renate (Hrsg.): *… in Sachen Deutsch als Fremdsprache. Sprachenpolitik und Mehrsprachigkeit, Unterricht, interkulturelle Begegnung. Festschrift für Hans-Jürgen Krumm zum 60. Geburtstag*. Baltmannsweiler: Schneider Verlag Hohengehren. 121–131.

LITTLE, David (2007): „The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy". In: *Modern Language Journal,* 91 (4). 645–655.

LÜTTICKEN, Martha (1985): „Der C-Test in Spanischkursen an der Volkshochschule". In: *Fremdsprachen und Hochschule* 13/14. Bochum: AKS-Verlag. 130–131.

MARCHWACKA, Maria A. (2012): „Gesundheitsförderung – eine pädagogische Herausforderung ?". In: Marchwacka, Maria A. (Hrsg.): *Gesundheitsförderung im Setting Schule*. Wiesbaden: Springer. 11–28.

MARKOWITSCH, Hans (1996): „Neuropsychologie des menschlichen Gedächtnisses". In: Spektrum der Wissenschaft, 9. 52–61. Abrufbar unter: http://www.spektrum.de/magazin/neuropsychologie-des-menschlichen-gedaechtnisses/823249 (Zugriff 07.10.2017)

MAY, Peter & BENNÖHR, Jasmine (Hrsg.) (2013): *Kompetenzerfassung in Kindergarten und Schule. Handbuch Konzept, theoretische Grundlagen und Normierung*. Berlin: Cornelsen.

MAY, Peter/BENNÖHR, Jasmine & BERGER, Carina (2014): „Lernentwicklungsmonitoring mit KEKS". In: Hasselhorn, Marcus/Schneider, Wolfgang & Trautwein, Ulrich (Hrsg.): *Lernverlaufsdiagnostik. Tests und Trends*, Neue Folge Band 12. Göttingen: Hogref. 257–280.

MCTAGGART, Jonathan & KLAR, Nicole (ohne Jahr): *Sprachbeschreibung Hebräisch.* Universität Duisburg-Essen: pro-DaZ. (https://www.uni-due.de/prodaz/einzelsprachen.php) (Zugriff 28.09.2017)

MICHEL, Marije (2017): „Complexity, Accuracy and Fluency in L2 production". In: Loewen, Shawn & Sato, Masatoshi (Hrsg.) *Routledge Handbook of Instructed Second Language Acquisition*. London: Routledge. 50–68.

MOOSBRUGGER, Helfried & KELAVA, Augustin (2007): *Testtheorie und Fragebogenkonstruktion.* Berlin: Springer Verlag.

MOOSBRUGGER, Helfried & KELAVA, Augustin (2011): *Testtheorie und Fragebogenkonstruktion.* 2. Auflage. Berlin: Springer Verlag.

MÜNZEL, Helmut (1989): „Wie erleben Schüler den Klausurtag im Einzelwettbewerb?" In: Finkenstaedt, Thomas & Schröder, Konrad (Hrsg.): *Zwischen Empirie und Machbarkeit. Erstes Symposium zum Bundeswettbewerb Fremdsprachen*. Augsburg: Universitätsverlag. 103–111.

NEUGEBAUER, Bettina/GROTJAHN, Rüdiger/TESCH, Bernd (2012): „Zeitbegrenzung in Lesetests. Auswirkungen auf das Testkonstrukt, testmethodische Konsequenzen und didaktisches Potential am Beispiel der VERA-8-Leseaufgaben im Fach Französisch". In: *Zeitschrift für Fremdsprachenforschung*, 23 (2). 195–241.

NEUNER, Gerhard & HUNFELD, Hans (1993): *Methoden des fremdsprachlichen Deutschunterrichts: Eine Einführung*. Berlin et al.: Langenscheidt bei Klett.

OLLER, John W. (1971): „Dictation as a Device for Testing Foreign-Language Proficiency". In: *English Language Teaching*, 25 (3). 254–259.

OLLER, John W. (1973): „Cloze Tests of Foreign Language Proficiency and what they Measure". In: *Language Learning,* 23 (1). 105–118.

OLLER, John W. (1979): *Language Tests at School.* Longman.

OLLER, John W. (1983): „Evidence for a General Language Proficiency Factor: an Expectancy Grammar". In: Oller, John W. (Hrsg.): *Issues in Language Testing Research*. Rowley/Massachusetts: Newbury House Publishers. 3–10.

OLLER, John W. & CONRAD, Christine A. (1971): „The Cloze Technique and ESL Proficiency". In: *Language Learning,* 21 (2). 183–194.

OLLER, John W. & JONZ, Jon (1994): „Why Cloze Procedure?". In: Oller, John W. & Jonz, Jon (Hrsg.): *Cloze and Coherence*. Lewisburg, PA: Associated University Presses. 1–20.

OLLER, John W. & STREIFF, Virginia (1975): „Dictation: A Test of Grammar Based Expectancies". In: Jones, Randall L. & Spolky, Bernard (Hrsg.): *Testing Language Proficiency*. Arlington: Center for Applied Linguistics.

OSWALD, Wolf D. & ROTH, Erwin (1987): *Der Zahlen-Verbindungs-Test (ZVT). Ein sprachfreier Intelligenz-Test zur Messung der „kognitiven Leistungsgeschwindigkeit". Handanweisung*. (2. überarbeitete und erweiterte Auflage). Göttingen: Hogrefe.

PANG, Lee Yick (1984): „Is There a Global Factor of Language Proficiency? A Critique of Oller and Hinofotis (1980)". In: *IRAL*, 22 (3). 203–228.

PARADIS, Michael (2009): *Declarative and procedural determinants of second languages.* Amsterdam: Benjamins.

RAATZ, Ulrich (1985): „The Factorial validity of C-Tests". In: *Fremdsprachen und Hochschule 13/14. Thematischer Teil: C-Tests in der Praxis.* Bochum: AKS-Verlag. 42–54.

RAATZ, Ulrich (2002): „C-Tests and Intelligence." In: Coleman, James A./ Grotjahn, Rüdiger/ Raatz, Ulrich (Hrsg.): *University Language Testing and the C-Test*. Bochum: AKS-Verlag. 169–185.

RAATZ, Ulrich & KLEIN-BRALEY, Christine (1994):„Analyse der Ergebnisse im Einzelwettbewerb des Bundeswettbewerbs Fremdsprachen: Wettbewerb Sekundarstufe I Frühjahr 1991". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*, Bd. 2. Bochum: Brockmeyer. 239–258.

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. 75–91.

RAATZ, Ulrich & WOCKENFUSS, Verena (2006): „Das TESTATT-Projekt: Entwicklung von C-Tests zur Evaluation des Fremdsprachenerfolgs". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen*. Frankfurt am Main: Peter Lang. 85–99.

REICHERT, Monique/KELLER, Ulrich & MARTIN, Romain (2010): „The C-test, the TCF and the CEFR: a Validation Study." In: Grotjahn, Rüdiger: *Der C-Test: Beiträge aus der Aktuellen Forschung.* Frankfurt am Main: Peter Lang. 205–231.

REICHERT, Monique/BRUNNER, Martin & MARTIN, Romain (2014): „Do test takers with different language backgrounds take the same C-test? The effect of native language on the validity of C-tests". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen.* Frankfurt am Main: Peter Lang. 109–135.

ROOS, Undine (1995): *Ein C-Test für Lerner der japanischen Sprache: Entwicklung, Erprobung und Validierung.* Bochum: AKS-Verlag.

RÖSCH, Heidi (2011): *Deutsch als Zweit- und Fremdsprache.* Berlin: Akademie Verlag.

SCHLAK, Torsten/ZIMMERMANN, Kerstin/MOLNÁR, Heike (2010): „Zur Entwicklung eines ‚C-Tests Wirtschaftsdeutsch': Hinweise und Erfahrungen aus der Praxis". In: Berndt, Annette & Kleppin, Karin (Hrsg.): *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn.* Frankfurt am Main: Peter Lang. 13–22.

SCHLIMBACH, Alice (1964): *Kinder lernen Deutsch: Die Familie Schiller*. Band 1. München: Hueber.

SCHMIDT, Richard (1990): „The Role of Consciousness in Second Language Learning". In: *Applied Linguistics*, Vol. 11, No. 2. 129–158.

SCHÖLER, Marianne (2016): „Der C-Test als Instrument zur Erfassung der fachsprachlichen Kompetenz". In: Drumbl, Hans/Kletschko, Dmitri/ Sorrentino, Daniela & Zanin, Renata (Hrsg.): *Tagungsband IDT* 2013, Band 7: Lerngruppenspezifik in DaF, DaZ, DaM. Bozen: Bozen-Bolzano University Press. 233–246.

SCHÖN, Almut/ZIMMERMANN, Kerstin/JOHNSON, Natalia (2012): „Intrauniversitäre Kooperation – zur gemeinsamen Entwicklung eines C-Tests durch Sprachenzentrum und Sprachlehrforschung. In: *Fremdsprachen und Hochschule,* 86. 61–79.

SCHWINDELER, Nicole (2013): *Sprachgebrauch und Sprachbedarf: Sprachbezogene Voraussetzungen und Erwartungen ausländischer Studierender an der Technischen Universität Berlin*. Unveröffentlichte Masterarbeit. TU Berlin, Fachgebiet Deutsch als Fremdsprache.

SEGAL, Michael (1983): *The C-Test*. School of Education, Hewbrew University, seminar paper.

SEGALOWITZ, Norman (2003): „Automaticity and Second Languages". In: Doughty, Catherine J. & Long, Michael H. (Hrsg.): *The Handbook of Second Language Acquisition.* Malden, MA: Blackwell. 383–408.

SEGALOWITZ, Norman (2010): *Cognitive bases of second language fluency*. New York: Routledge.

SEGALOWITZ, Norman & HULSTIJN, Jan H. (2005): „Automaticity in bilingualism and second language learning". In: Kroll, Judith F. & DeGroot, Annette M. B. (Hrsg.): *Handbook of bilingualism: Psycholinguistic approaches*. Oxford: Oxford University Press.

SELTING, Margret/AUER, Peter/BARTH-WEINGARTEN, Dagmar/BERGMANN, Jörg/BERGMANN, Pia/BIRKNER, Karin/COUPER-KUHLEN, Elizabeth/DEPPERMANN, Arnulf/GILLES, Peter/GÜNTHNER, Susanne/HARTUNG, Martin/KERN, Friederike/MERTZLUFFT, Christine/MEYER, Christian/MOREK, Miriam/OBERZAUCHER, Frank/PETERS, Jörg/QUASTHOFF, Uta/SCHÜTTE, WILFRIED/STUKENBROCK, Anja & UHMANN, Susanne (2009): „Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)". In: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion.* Ausgabe 10 (2009). 353–402. (www.gespraechsforschung-ozs.de)

SHANNON, Claude E. & WEAVER, Warren (1949): *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.

SIGOTT, Günther (2004): *Towards identifying the C-Test construct*. Frankfurt am Main: Peter Land.

SIGOTT, Günther (2006): „How fluid is the C-Test Construct?". In: In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen.* Frankfurt am Main: Peter Lang. 139–146.

SKEHAN, Peter (1996): „A Framework for the Implementation of Task-based Instruction". In: *Applied Linguistics*, 17 (1). 38–62.

SOLMECKE, Gert (2001): „Hörverstehen". In: Helbig, Gerhard/Götze, Lutz/Henrici, Gert & Krumm, Hans-Jürgen (Hrsg.): *Deutsch als Fremdsprache. Ein internationales Handbuch*. Berlin: DeGruyter. 893–900.

SOMARATNE, W. R. P. (1965): *Aids and tests in the teaching of English as a second language.* London: Oxford University Press.

SPOLSKY, Bernard (1986): „Preliminary studies in the development of techniques for testing overall second language proficiency". In: *Language Learning*, 18 (23). 79–101.

SPOLSKY, Bernard/SIGURD, Bengt/SATO, Masahito/WALKER, Edward/ARTERBURN, Catherine (1968): Preliminary Studies in the Development of Techniques for Testing Overall Second Language Proficiency. In: *Language Learning*, 3. 79–101.

STANSFIELD, Charles W. (1985): „A History of Dictation in Foreign Language Teaching and Testing". In: *The Modern Language Journal*, 69 (2). 121–128.

STANSFIELD, Charles W. (1989): *Simulated Oral Proficiency Interviews.* ERIC digest.

STEMMER, Brigitte (1991): *What's on a C-Test Taker's Mind? Mental Processes in C-Test Taking*. Bochum: Brockmeyer.

STEMMER, Brigitte (1992): „An Alternative Approach to C-Test Validation". In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen. Band 1.* Bochum: Brockmeyer. 97–131.

STERNBERG, Gerald (1999): *Zusammenhänge zwischen der Position von Texten und ihrem Schwierigkeitsgrad bei muttersprachlichen C-Tests*. Unveröffentlichte Staatsexamensarbeit, Gerhard-Mercator-Universität Duisburg, Fachbereich Psychologie.

STEURER, Veronika (1986): „Ein neues Verfahren beim ‚Bundeswettbewerb Fremdsprachen': der C-Test. - Auch im schulischen Russischunterricht einsetzbar?". In: *Zielsprache Russisch: Zeitschrift für den Russischunterricht*. Ismaning & München: Hueber. 83–90.

STROHM KITCHENER, Karen & KITCHENER, Richard F. (2009): „Social Science Research Ethics. Historical and Philosophical Issues". In: Mertens, Donna M. & Ginsberg, Pauline E. (Hrsg.): *The Handbook of Social Research Ethics*. Thousand Oaks: Sage. 5–22.

STUBBS, J. B., & TUCKER, G. Richard (1974): „The cloze test as a measure of ESL proficiency for Arab students". In: *Modern Language Journa*l, 58. 239–241.

TAVAKOLI, Parvaneh (2016): „Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency". In: *International Review of Applied Linguistics in Language Teaching*, 54 (2). 133–150.

TAVAKOLI, Parvaneh & SKEHAN, Peter (2005): „Strategic planning, task structure, and performance testing". In: Ellis, Rod (Hrsg.): *Planning and task performance*. Amsterdam & Philadelphia: John Benjamins. 125–144.

TAYLOR, Wilson L. (1953): „'Cloze-procedure': A new tool for measuring readability." In: *Journalism Quarterly*, Vol. 30. 415–433.

TSCHIRNER, Erwin (2001): „Kompetenz, Wissen, mentale Prozesse: Zur Rolle der Grammatik im Fremdsprachenunterricht". In: Funk, Hermann & König, Michael (Hrsg.): *Kommunikative Didaktik in Deutsch als Fremdsprache – Bestandsaufnahme und Ausblick. Festschrift für Gerhard Neuner.* München: Iudicium. 106–125.

VON UNGER, Hella (2014): „Forschungsethik in der qualitativen Forschung: Grundsätze, Debatten und offene Fragen". In: von Unger, Hella/Narimani, Petra & M'Bayo, Rosaline (Hrsg.): *Forschungsethik in der qualitativen Forschung. Reflexivität, Perspektiven, Positionen.* Wiesbaden: Springer. 15–39.

VALLETTE, Rebecca M. (1977): *Modern Language Testing*. 2. Ausgabe. San Diego et al.: Harcort Brace Jovanovich.

VASEGHI, Saeed V. (2000): *Advanced Digital Signal Processing and Noise Reduction*. 2. Ausgabe. Chichester et al.: John Wiley & Sons.

VIEBROCK, Britta (2015): *Forschungsethik in der Fremdsprachenforschung. Eine systemische Betrachtung*. Frankfurt am Main: Peter Lang.

VOLLMER, Helmut J. (1981): „Why are we interested in 'general language proficiency'?". In: Klein-Braley, Christine & Stevenson, Douglas, K. (Hrsg.): *Practice and problems in language testing I.* Frankfurt/Main: Lang. 96–123.

VOLLMER, Helmut J. (1982): *Spracherwerb und Sprachbeherrschung. Untersuchungen zur Struktur von Fremdsprachenfähigkeit.* Tübingen: Narr.

WARD, Magdalen (1987): „A study of the C-test in French". In: *Language Testing Update*, 4, 23.

WERTHEIMER, Max (1924): „Gestalt Theory". Zuerst publiziert in: Gestalt Theory, Vol 21, No. 3. (Nov. 1999). 181–182. Abrufbar unter: http://www.gestalttheory.net/cms/uploads/pdf/archive/1910_1933/gestalt_theory_wertheimer.pdf (Zugriff 18.10.2014)

WHITESON, Valerie & SELIGER, Herbert W. (1962): An Integrative Approach to the ‚Noise' Test. In: *Audio-Visual Language Journal. Journal of Applied Linguistics and Language Teaching Technology.* Birmingham: Organ of the Audio-Visual Language Association. 17–18.

WILHELM, Oliver & SCHULZE, Ralf (2002): „The relation of speeded and unspeeded reasoning with mental speed". In: *Intelligence*, 30. 537–554.

WOCKENFUSS, Verena (2009): *Diagnostik von Sprache und Intelligenz bei Jugendlichen und jungen Erwachsenen.* Aachen: Shaker Verlag.

WOCKENFUSS, Verena & RAATZ, Ulrich (2006): „Über den Zusammenhang zwischen Testleistung und Klassenstufe bei muttersprachlichen C-Tests". In: Grotjahn Rüdiger (Hrsg.): *Der C-Test: Theorie, Empirie, Anwendungen.* Frankfurt am Main: Peter Lang. 211–242.

ZIEGLER, Matthias & BÜHNER, Markus (2012): *Grundlagen Der Psychologischen Diagnostik.* Wiesbaden: Springer Verlag für Sozialwissenschaften.

Internetquellen

URL 1: C-Test: Der Sprachtest (http://www.c-test.de/deutsch/index.php?lang=de&content=konstruktion&section=ctest) Abrufdatum 10.03.2018

URL 2: TOEIC (http://www.ets.org/toeic) Abrufdatum 10.03.2018

URL 3: onDaF (http://www.ondaf.de/) Abrufdatum 12.05.2013

URL 4: TestDaF (http://www.testdaf.de/) Abrufdatum 12.05.2013

URL 5: DIHK – Gesellschaft für berufliche Weiterbildung: Prüfung Wirtschaftsdeutsch International (https://www.dihk-bildungs-gmbh.de/weiterbildung/pruefungen-von-a-z/weitere-pruefungskategorien/wirtschaftsdeutsch/) Abrufdatum 10.03.2018

URL 6: Goethe-Institut: Übungsmaterial für das Goethe-Zertifikat B2 (http://www.goethe.de/lrn/prj/pba/bes/gb2/mat/deindex.htm) Abrufdatum 14.01.2018

URL 7: Kultusministerkonferenz: Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen (https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_06_25_RO_DT.pdf) Abrufdatum 13.08.2017

URL 8: Wikimedia Commons: Gesetz der Geschlossenheit und Gesetz der Kontinuität (https://commons.wikimedia.org/w/index.php?title=File:Gestalt_ley_de_cierre.png&oldid=141668925) Abrufdatum 02.02.2018

URL 9: International Language Testing Association: ILTA Code of Ethics (http://www.iltaonline.com/page/CodeofEthics) Abrufdatum 05.03.2018

URL 10: DGS – Deutsche Gesellschaft für Soziologie: Ethik-Kodex der Deutschen Gesellschaft für Soziologie (DGS) und des Berufsverbandes Deutscher Soziologinnen und Soziologen (BDS) (http://www.soziologie.de/de/die-dgs/ethik/ethik-kodex.html) Abrufdatum 06.06.2017

URL 11: ZEMS – Zentraleinrichtung Moderne Sprachen der TU Berlin: Einstufung (https://www.zems.tu-berlin.de/einstufung/) Abrufdatum 17.03.2018

URL 12: Zentrum für Fremdsprachen Ausbildung der Ruhr-Universität Bochum: Elektronische Einstufungstests (http://www.zfa.ruhr-uni-bochum.de/lehre/einstufung/testformat.html.de) Abrufdatum 08.06.2017

URL 13: Sprachenzentrum der LMU München: Einstufungstets (http://www.sprachenzentrum.uni-muenchen.de/teiln_bed_anmeldung/einstufungstests/index.html) Abrufdatum 08.06.2017

URL 14: Zentraleinrichtung Sprachenzentrum der Humboldt-Universität zu Berlin: Einstufungstests (https://www.sprachenzentrum.hu-berlin.de/de/studium-und-lehre/einstufungstests-online) Abrufdatum 08.06.2018

URL 15: Sprachenzentrum der Westfälischen Wilhelms-Universität Münster: C-Test (http://spz.uni-muenster.de/ctest) Abrufdatum 25.06.2017

URL 16: Sprachenzentrum der Philipps-Universität Marburg: Beispiele des Einstufungstests für die Englisch-Kurse (https://www.uni-marburg.de/sprachenzentrum/lehrangebot/englisch/bsp_test) Abrufdatum 09.06.2017)

URL 17: Sprachenzentrum der Universität Passau: Einstufungstests für die Sprachkurse (http://www.zim.uni-passau.de/login/ilias/c-test/) Abrufdatum 10.03.2018

URL 18: Cornelsen: KEKS - Kompetenzerfassung in Kindergarten und Schule (http://www.cornelsen.de/keks/) Abrufdatum 05.04.2017

URL 19: Cornelsen: KEKS - Kompetenzerfassung in Kindergarten und Schule: Beispieltest (http://www.cornelsen.de/keks/1.c.3403254.de) Abrufdatum 05.04.2017

URL 20: Hamburger Volkshochschule: Online-Einstufungstests (https://www.vhs-hamburg.de/infocenter/sprachenberatung/online-einstufungstests-443) Abrufdatum 04.02.2018

URL 21: VHS Einstufungstest (http://www.vhseinstufungstest.de/) Abrufdatum 10.03.2018

URL 22: Deutsche Gesellschaft für Sprachwissenschaft – DGS: Ethikkommission der Deutschen Gesellschaft für Sprachwissenschaft (https://dgfs.de/de/inhalt/ueber/ethikkommission.html) Abrufdatum 29.03.2017

URL 23: C-Test: Der Sprachtest: Wie sieht ein C-Test aus? (http://www.c-test.de/deutsch/index.php?lang=de&content=beschreibung_aussehen&section=ctest) Abrufdatum 05.11.2014

URL 24: onSET: Beispieltest (https://www.onset.de/home/beispieltest-onset-online-spracheinstufungstest) Abrufdatum 21.03.2018

URL 25: onSET: über onSET (https://www.onset.de/home/ueber-onset/) Abrufdatum 21.03.2018

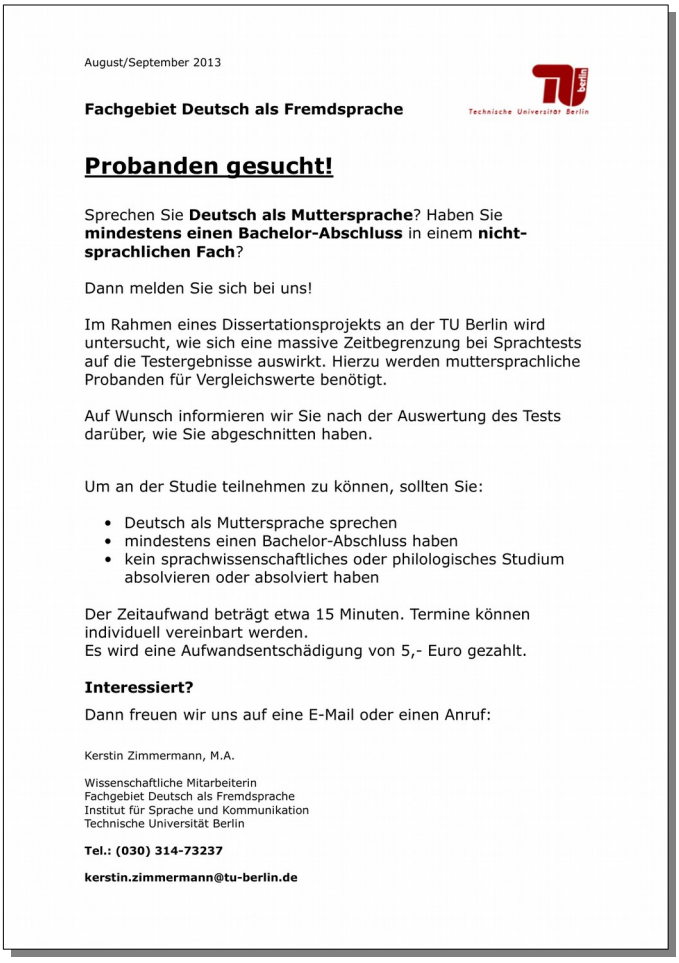
URL 26: onSET: Teilnehmende (https://www.onset.de/home/teilnehmende/) Abrufdatum 21.03.2018

URL 27: TOEFL ibt: Test Questions (https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf) Abrufdatum 18.03.2018

URL 28: CIEP – Centre International d`Études Pédagogique: TCF (http://www.ciep.fr/de/tcf-tout-public) Abrufdatum 18.03.2018

Anhang

Anhang A: Probandensuche Zeitpilotierung



August/September 2013

Fachgebiet Deutsch als Fremdsprache

Technische Universität Berlin

Probanden gesucht!

Sprechen Sie **Deutsch als Muttersprache**? Haben Sie **mindestens einen Bachelor-Abschluss** in einem **nicht-sprachlichen Fach**?

Dann melden Sie sich bei uns!

Im Rahmen eines Dissertationsprojekts an der TU Berlin wird untersucht, wie sich eine massive Zeitbegrenzung bei Sprachtests auf die Testergebnisse auswirkt. Hierzu werden muttersprachliche Probanden für Vergleichswerte benötigt.

Auf Wunsch informieren wir Sie nach der Auswertung des Tests darüber, wie Sie abgeschnitten haben.

Um an der Studie teilnehmen zu können, sollten Sie:

- Deutsch als Muttersprache sprechen
- mindestens einen Bachelor-Abschluss haben
- kein sprachwissenschaftliches oder philologisches Studium absolvieren oder absolviert haben

Der Zeitaufwand beträgt etwa 15 Minuten. Termine können individuell vereinbart werden.
Es wird eine Aufwandsentschädigung von 5,- Euro gezahlt.

Interessiert?

Dann freuen wir uns auf eine E-Mail oder einen Anruf:

Kerstin Zimmermann, M.A.

Wissenschaftliche Mitarbeiterin
Fachgebiet Deutsch als Fremdsprache
Institut für Sprache und Kommunikation
Technische Universität Berlin

Tel.: (030) 314-73237

kerstin.zimmermann@tu-berlin.de

Anhang B: Durchführungsplan Zeitpilotierung

Teilnehmer-Code

„Zeitbemessung beim C-Test“
Fragebogen für Muttersprachler

C-Test für Muttersprachler: Testablauf

Der Test wird in einer ruhigen, störungsfreien Umgebung abgelegt.
Dem Probanden wird erklärt, dass untersucht wird, was passiert, wenn man Deutschlernern zur Bearbeitung eines C-Test nur sehr wenig Zeit lässt. Dazu muss man zunächst herausfinden, wie viel Zeit Muttersprachler für den Test benötigen.

Das Testheft wird ausgeteilt.

Der Proband soll das Deckblatt des Testheftes lesen und ggf. Fragen stellen.

Vor Beginn noch einmal darauf hinweisen, dass zwar die Zeit gestoppt wird, es aber von größter Wichtigkeit ist, sorgfältig zu arbeiten.

Der Proband blättert um.
➔ Im Moment des Umblätterns die Stoppuhr starten

Der Proband sagt STOP.

➔ Sofort die Stoppuhr anhalten

➔ Gemessene Zeit auf dem Blatt „Datenerhebung Muttersprachler“ vermerken

Der Proband darf erst nach Aufforderung umblättern. Bevor er dies tut, bitte die Stoppuhr wieder zurücksetzen. Dann das Kommando zum umblättern geben, etc. s.o.

Im Anschluss den Begleitfragebogen austeilen. Darauf achten, daß alle Felder ausgefüllt werden.

Anmerkungen (falls Nachfragen kommen):

- Bei den Fremdsprachen muss kein hohes Niveau vorhanden sein, um sie auflisten zu dürfen.
- Beim Beruf geht es darum, womit man sich derzeit täglich fachlich befasst, d.h. Elternzeit oder kein Job wegen Studienendphase ist auch eine wichtige Info!
- Der C-Test kann auch in einer anderen Sprache abgelegt worden sein. Es geht hier nur darum, ob jemand das Testformat schonmal mitgemacht (und nicht nur gesehen!) hat.

Anhang C: Fragebogen (Zeitpilotierung/Muttersprachler)

„Zeitbemessung beim C-Test"
Fragebogen für Muttersprachler

Teilnehmer-Code

1 Alter: _____ Jahre

2 Geschlecht: o männlich
o weiblich

3 Welche Sprachen sprechen Sie?
Muttersprache(n) _____
Fremdsprachen _____

4 Was haben Sie studiert? _____

5 Was ist Ihr höchster Bildungsabschluss?
o Bachelor
o Master
o Magister
o Diplom
o Promotion
o Habilitation

6 In welchem Beruf arbeiten Sie? _____

7 Haben Sie vorher schon einmal einen C-Test abgelegt? o ja
o nein

Anhang D: Deckblatt/Eisbrechertext (Zeitpilotierung)

„Zeitbemessung beim C-Test“
Testset für Muttersprachler

C-Test

Teilnehmer-Code:

Hinweise für die Teilnehmer

In den folgenden 9 Texten fehlen bei einer Reihe von Wörtern am Wortende jeweils einige Buchstaben. Bitte ergänzen Sie möglichst sämtliche Lücken in sinnvoller Weise. Bitte geben Sie sich dabei größte Mühe.

Es handelt sich um eine Untersuchung mit einer ausschließlich wissenschaftlichen Zielsetzung. Ihre Leistung wird nicht benotet!

Bitte arbeiten Sie so schnell wie möglich, aber auch so sorgfältig wie möglich!
Bitte konzentrieren Sie sich immer nur auf die Bearbeitung eines einzigen Textes und blättern Sie nicht vor oder zurück. Die Testleiter teilen Ihnen mit, wann Sie jeweils umblättern und mit der Bearbeitung des nächsten Textes beginnen sollen.
Wenn Sie mit der Bearbeitung eines Textes fertig sind, sagen Sie bitte laut und deutlich STOP.

Schreiben Sie bitte leserlich!

Beispiel:

Spiele sind untrennbar mit der menschlichen Kultur verbunden. Heute erl***ebt*** das gemei***nsame*** Spielen ei***ne*** Renaissance. De***nn*** Spiele si***nd*** ein Ausd***ruck*** von Gesell***igkeit*** und Gemein***schaft***.

Vielen Dank für Ihre Mitarbeit!

0. Nach der Arbeit zum Sport

Viele Menschen arbeiten im Büro, wo sie eine überwiegend sitzende Tätigkeit ausüben. Umso wich_____ ist e_____, zum Ausg_____ Sport z_____ treiben. M_____ Sport erh_____ man se_____ körperliche Leistungsf_____ nämlich erhe_____. Aber au_____ die geis_____ Fertigkeiten wer_____ durch regelm_____ Sport verbe_____. Das hä_____ damit zusa_____, dass d_____ Sauerstoffversorgung d_____ Blutes du_____ die Bewe_____ wesentlich effek_____ funktioniert. Auch die Durchblutung sämtlicher Körperregionen ist bei Sportlern erheblich stärker gewährleistet, und weil damit das Gehirn besser mit Sauerstoff versorgt wird, ist auch seine Leistung höher.

Bitte erst nach Aufforderung umblättern!

Bitte warten!

Anhang E: Informationsblatt und Teilnehmer-Code (Lerner)

„Zeitbemessung beim C-Test“
Infoblatt & Teilnehmer-Code

Liebe Teilnehmer!

Mit dieser Studie wollen wir herausfinden, welchen Einfluss eine extreme Zeitbegrenzung auf die Zuverlässigkeit von Sprachtests mit Lückentexten (C-Tests) hat. Dazu vergleichen wir die Ergebnisse aus den Lückentexten mit den Ergebnissen von Aufgaben zum Hören und Sprechen.

Wir möchten Sie bitten, sich bei allen Aufgaben bestmöglich anzustrengen.

Die Studie dient rein wissenschaftlichen Zwecken. Alle erhobenen Informationen bleiben selbstverständlich anonym! Anstelle Ihres Namens werden wir einen Teilnehmer-Code verwenden, den Sie selbst erstellen und der für die gesamte Studie gilt.
Der Teilnehmer-Code besteht aus:

- den ersten zwei Buchstaben des Vornamens Ihrer Mutter
- den ersten zwei Buchstaben des Vornamens Ihres Vaters
- Ihrem Geburtstag
- Ihrem Geburtsmonat

Hier sehen Sie ein Beispiel:

Vorname der Mutter	Vorname des Vaters	Geburtstag	Geburtsmonat
MARIA	**PA**UL	**12**.04.1990	12.**04**.1990
MA	PA	12	04

Bitte tragen Sie nun Ihren Teilnehmer-Code ein:

Vorname der Mutter	Vorname des Vaters	Ihr Geburtstag	Ihr Geburtsmonat

Wir danken Ihnen herzlich für Ihre Mitarbeit und Unterstützung!

Anhang F: Fragebogen (Lerner)

„Zeitbemessung beim C-Test"
Fragebogen für Deutschlerner

Liebe Teilnehmer!
Im Folgenden finden Sie einige Fragen zu Ihrer Person, Ihren Lernerfahrungen und Ihren Erfahrungen mit dem C-Test (Lückentext). Bitte lesen Sie die Fragen genau und antworten Sie ehrlich. Ihre Angaben dienen ausschließlich wissenschaftlichen Zwecken und werden nicht bewertet.
Vielen Dank!

Persönliche Angaben

Teilnehmer-Code

1 Alter: _____ Jahre

2 Geschlecht: ☐ männlich ☐ weiblich

3 Muttersprache(n) _____

4 In welchem Land sind Sie hauptsächlich aufgewachsen? _____

5 Was studieren Sie? _____

6 In welchem Schriftsystem sind Sie alphabetisiert worden?
(z. B. lateinische Schrift, arabische Schrift, kyrillische Schrift, Hanzi, Hiragana, etc.)

_____

7 Wie viele Stunden schreiben („tippen") Sie täglich am Computer? _____ Stunden

Lernerfahrungen

8 Wie haben Sie bisher hauptsächlich Deutsch gelernt? (Mehrfachnennung möglich)

☐ Unterricht an der Schule
☐ Unterricht an der Universität im Heimatland
☐ Unterricht an der Universität in Deutschland, Österreich oder der Schweiz
☐ Unterricht an einer Sprachschule im Heimatland
☐ Unterricht an einer Sprachschule in Deutschland, Österreich oder der Schweiz
☐ 1-zu-1-Unterricht/Privatunterricht
☐ durch deutschsprachige Freunde
☐ durch deutschsprachige Familienmitglieder
☐ mit einem Tandem-Partner (über das Internet)
☐ mit einem Tandem-Partner (*face-to-face*)
☐ autodidaktisch/im Selbststudium

„Zeitbemessung beim C-Test"
Fragebogen für Deutschlerner

9 In welchem Alter haben Sie begonnen, Deutsch zu lernen? _____Jahre

10 Wie oft haben Sie die Möglichkeit, Deutsch zu sprechen?

täglich	wöchentlich	monatlich	seltener
☐	☐	☐	☐

11 Gibt es Situationen, in denen sie sich unwohl fühlen, Deutsch zu sprechen? ☐ja ☐nein

Wenn ja, welche? (z. B. am Telefon, im Unterricht, etc.) _____

_____

12 Wie oft nutzen Sie deutschsprachige Medien?

	täglich	wöchentlich	monatlich	seltener
TV	☐	☐	☐	☐
Radio	☐	☐	☐	☐
deutschsprachige Musik	☐	☐	☐	☐
Zeitungen	☐	☐	☐	☐
Zeitschriften	☐	☐	☐	☐
Bücher	☐	☐	☐	☐

13 Welche Fremdsprachen sprechen Sie noch? Schätzen Sie bitte Ihre Fremdsprachenkenntnisse auf einer Skala von 1 (sehr niedrig, Anfänger) bis 6 (fast muttersprachlich) ein.

Sprache	1 sehr niedrig	2 niedrig	3 mittel	4 hoch	5 sehr hoch	6 fast mutter-sprachlich
_____	☐	☐	☐	☐	☐	☐
_____	☐	☐	☐	☐	☐	☐
_____	☐	☐	☐	☐	☐	☐

14 Haben Sie zuvor schon einmal einen C-Test (Lückentext) gemacht? ☐ja ☐nein

„Zeitbemessung beim C-Test"
Fragebogen für Deutschlerner

Die folgenden Fragen beziehen sich auf den onDaF (C-Test am Computer):

15 erreichte Punkte im onDaF

16 Denken Sie, dass der onDaF Ihre Deutschkenntnisse korrekt wiedergibt? ☐ ja ☐ nein

17 Wie haben Sie den onDaF gelöst? (Mehrfachnennung möglich)

☐ Ich habe alle Lücken in der vorgegebenen Reihenfolge bearbeitet.
☐ Ich bin innerhalb eines Satzes vor- und zurückgesprungen.
☐ Ich bin innerhalb eines Textes vor- und zurückgesprungen.
☐ Ich habe den Text zunächst überflogen und die leichteren Lücken ausgefüllt.
☐ Ich habe mich längere Zeit mit schwierigeren Lücken beschäftigt.
☐ Ich habe schwierige Lücken leer gelassen.

Die folgenden Fragen beziehen sich auf den *speeded*-C-Test (C-Test auf dem Papier):

18 Denken Sie, dass der *speeded*-C-Test Ihre Deutschkenntnisse gut wiedergibt? ☐ ja ☐ nein

19 Hatten Sie aufgrund ihrer Schreibgeschwindigkeit Schwierigkeiten beim Lösen des *speeded*-C-Test? ☐ ja ☐ nein

20 Wie haben Sie den *speeded*-C-Test gelöst? (Mehrfachnennung möglich)

☐ Ich habe alle Lücken in der vorgegebenen Reihenfolge bearbeitet.
☐ Ich bin innerhalb eines Satzes vor- und zurückgesprungen.
☐ Ich bin innerhalb eines Textes vor- und zurückgesprungen.
☐ Ich habe den Text zunächst überflogen und die leichteren Lücken ausgefüllt.
☐ Ich habe mich längere Zeit mit schwierigeren Lücken beschäftigt.
☐ Ich habe schwierige Lücken leer gelassen.

21 Sind Sie Linkshänder oder Rechtshänder?

☐ Linkshänder ☐ Rechtshänder

22 Hatten Sie aufgrund ihrer Händigkeit Schwierigkeiten beim Lösen des *speeded*-C-Test?

☐ ja ☐ nein

Anhang G: Bewertung Hörverstehen

„Zeitbemessung beim C-Test"
Bewertung Hörverstehen

Teilnehmer-Code

Aufgabe 1

1	_____	große Halle	☐
2	_____	50 und 100 Meter	☐
3	_____	14 bis 17 Uhr	☐
4	_____	Kinder	☐
5	_____	0170/5255286	☐

Punkte _____ x 2 = _____

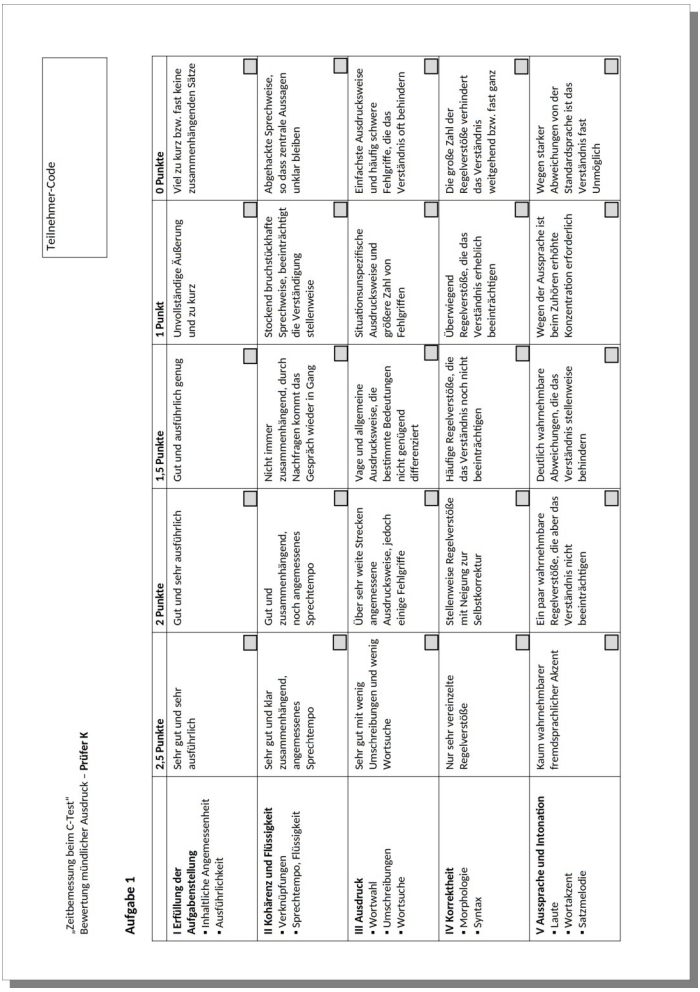
Aufgabe 2

6	A	B	C	☐
7	A	B	C	☐
8	A	B	C	☐
9	A	B	C	☐
10	A	B	C	☐
11	A	B	C	☐
12	A	B	C	☐
13	A	B	C	☐
14	A	B	C	☐
15	A	B	C	☐

Punkte _____ x 1,5 = _____

Punkte gesamt _____

Anhang H: Bewertungsvorgaben mündlicher Ausdruck



„Zeitbemessung beim C-Test"
Bewertung mündlicher Ausdruck - **Prüfer K**

Teilnehmer-Code

Aufgabe 1

	2,5 Punkte	2 Punkte	1,5 Punkte	1 Punkt	0 Punkte
I Erfüllung der Aufgabenstellung • Inhaltliche Angemessenheit • Ausführlichkeit	Sehr gut und sehr ausführlich	Gut und sehr ausführlich	Gut und ausführlich genug	Unvollständige Äußerung und zu kurz	Viel zu kurz bzw. fast keine zusammenhängenden Sätze
II Kohärenz und Flüssigkeit • Verknüpfungen • Sprechtempo, Flüssigkeit	Sehr gut und klar zusammenhängend, angemessenes Sprechtempo	Gut und zusammenhängend, noch angemessenes Sprechtempo	Nicht immer zusammenhängend, durch Nachfragen kommt das Gespräch wieder in Gang	Stockend bruchstückhafte Sprechweise, beeinträchtigt die Verständigung stellenweise	Abgehackte Sprechweise, so dass zentrale Aussagen unklar bleiben
III Ausdruck • Wortwahl • Umschreibungen • Wortsuche	Sehr gut mit wenig Umschreibungen und wenig Wortsuche	Über sehr weite Strecken angemessene Ausdrucksweise, jedoch einige Fehlgriffe	Vage und allgemeine Ausdrucksweise, die bestimmte Bedeutungen nicht genügend differenziert	Situationsunspezifische Ausdrucksweise und größere Zahl von Fehlgriffen	Einfachste Ausdrucksweise und häufig schwere Fehlgriffe, die das Verständnis oft behindern
IV Korrektheit • Morphologie • Syntax	Nur sehr vereinzelte Regelverstöße	Stellenweise Regelverstöße mit Neigung zur Selbstkorrektur	Häufige Regelverstöße, die das Verständnis noch nicht beeinträchtigen	Überwiegend Regelverstöße, die das Verständnis erheblich beeinträchtigen	Die große Zahl der Regelverstöße verhindert das Verständnis weitgehend bzw. fast ganz
V Aussprache und Intonation • Laute • Wortakzent • Satzmelodie	Kaum wahrnehmbarer fremdsprachlicher Akzent	Ein paar wahrnehmbare Regelverstöße, die aber das Verständnis nicht beeinträchtigen	Deutlich wahrnehmbare Abweichungen, die das Verständnis stellenweise behindern	Wegen der Aussprache ist beim Zuhören erhöhte Konzentration erforderlich	Wegen starker Abweichungen von der Standardsprache ist das Verständnis fast Unmöglich

„Zeitbemessung beim C-Test"
Bewertung mündlicher Ausdruck – **Prüfer S**

Teilnehmer-Code

Aufgabe 1

	2,5 Punkte	**2 Punkte**	**1,5 Punkte**	**1 Punkt**	**0 Punkte**
I Erfüllung der Aufgabenstellung ▪ Inhaltliche Angemessenheit ▪ Ausführlichkeit	Sehr gut und sehr ausführlich ☐	Gut und sehr ausführlich ☐	Gut und ausführlich genug ☐	Unvollständige Äußerung und zu kurz ☐	Viel zu kurz bzw. fast keine zusammenhängenden Sätze ☐
II Kohärenz und Flüssigkeit ▪ Verknüpfungen ▪ Sprechtempo, Flüssigkeit	Sehr gut und klar zusammenhängend, angemessenes Sprechtempo ☐	Gut und zusammenhängend, noch angemessenes Sprechtempo ☐	Nicht immer zusammenhängend, durch Nachfragen kommt das Gespräch wieder in Gang ☐	Stockend bruchstückhafte Sprechweise, beeinträchtigt die Verständigung stellenweise ☐	Abgehackte Sprechweise, so dass zentrale Aussagen unklar bleiben ☐
III Ausdruck ▪ Wortwahl ▪ Umschreibungen ▪ Wortsuche	Sehr gut mit wenig Umschreibungen und wenig Wortsuche ☐	Über sehr weite Strecken angemessene Ausdrucksweise, jedoch einige Fehlgriffe ☐	Vage und allgemeine Ausdrucksweise, die bestimmte Bedeutungen nicht genügend differenziert ☐	Situationsunspezifische Ausdrucksweise und größere Zahl von Fehlgriffen ☐	Einfachste Ausdrucksweise und häufig schwere Fehlgriffe, die das Verständnis oft behindern ☐
IV Korrektheit ▪ Morphologie ▪ Syntax	Nur sehr vereinzelte Regelverstöße ☐	Stellenweise Regelverstöße mit Neigung zur Selbstkorrektur ☐	Häufige Regelverstöße, die das Verständnis noch nicht beeinträchtigen ☐	Überwiegend Regelverstöße, die das Verständnis erheblich beeinträchtigen ☐	Die große Zahl der Regelverstöße verhindert das Verständnis weitgehend bzw. fast ganz ☐
V Aussprache und Intonation ▪ Laute ▪ Wortakzent ▪ Satzmelodie	Kaum wahrnehmbarer fremdsprachlicher Akzent ☐	Ein paar wahrnehmbare Regelverstöße, die aber das Verständnis nicht beeinträchtigen ☐	Deutlich wahrnehmbare Abweichungen, die das Verständnis stellenweise behindern ☐	Wegen der Aussprache ist beim Zuhören erhöhte Konzentration erforderlich ☐	Wegen starker Abweichungen von der Standardsprache ist das Verständnis fast Unmöglich ☐

Aufgabe 2

	2,5 Punkte	2 Punkte	1,5 Punkte	1 Punkt	0 Punkte
I Erfüllung der Aufgabenstellung ▪ Diskussionsfähigkeit	Sehr gut und sehr interaktiv ☐	Gut und interaktiv ☐	Gesprächsfähigkeit vorhanden, aber nicht sehr aktiv ☐	Beteiligung nur auf Anfrage ☐	Große Schwierigkeiten, sich überhaupt am Gespräch zu beteiligen ☐
II Kohärenz und Flüssigkeit ▪ Verknüpfungen ▪ Sprechtempo, Flüssigkeit	Sehr gut und klar zusammenhängend, angemessenes Sprechtempo ☐	Gut und zusammenhängend, noch angemessenes Sprechtempo ☐	Nicht immer zusammenhängend, durch Nachfragen kommt das Gespräch wieder in Gang ☐	Stockend bruchstückhafte Sprechweise, beeinträchtigt die Verständigung stellenweise ☐	Abgehackte Sprechweise, so dass zentrale Aussagen unklar bleiben ☐
III Ausdruck ▪ Wortwahl ▪ Umschreibungen ▪ Wortsuche	Sehr gut mit wenig Umschreibungen und wenig Wortsuche ☐	Über sehr weite Strecken angemessene Ausdrucksweise, jedoch einige Fehlgriffe ☐	Vage und allgemeine Ausdrucksweise, die bestimmte Bedeutungen nicht genügend differenziert ☐	Situationsunspezifische Ausdrucksweise und größere Zahl von Fehlgriffen ☐	Einfachste Ausdrucksweise und häufig schwere Fehlgriffe, die das Verständnis oft behindern ☐
IV Korrektheit ▪ Morphologie ▪ Syntax	Nur sehr vereinzelte Regelverstöße ☐	Stellenweise Regelverstöße mit Neigung zur Selbstkorrektur ☐	Häufige Regelverstöße, die das Verständnis noch nicht beeinträchtigen ☐	Überwiegend Regelverstöße, die das Verständnis erheblich beeinträchtigen ☐	Die große Zahl der Regelverstöße verhindert das Verständnis weitgehend bzw. fast ganz ☐
V Aussprache und Intonation ▪ Laute ▪ Wortakzent ▪ Satzmelodie	Kaum wahrnehmbarer fremdsprachlicher Akzent ☐	Ein paar wahrnehmbare Regelverstöße, die aber das Verständnis nicht beeinträchtigen ☐	Deutlich wahrnehmbare Abweichungen, die das Verständnis stellenweise behindern ☐	Wegen der Aussprache ist beim Zuhören erhöhte Konzentration erforderlich ☐	Wegen starker Abweichungen von der Standardsprache ist das Verständnis fast unmöglich ☐

Punkte aus Aufgabe 1	Punkte aus Aufgabe 2	Punkte gesamt

Aufgabe 2

	2,5 Punkte	2 Punkte	1,5 Punkte	1 Punkt	0 Punkte
I Erfüllung der Aufgabenstellung ▪ Diskussionsfähigkeit	Sehr gut und sehr interaktiv ☐	Gut und interaktiv ☐	Gesprächsfähigkeit vorhanden, aber nicht sehr aktiv ☐	Beteiligung nur auf Anfrage ☐	Große Schwierigkeiten, sich überhaupt am Gespräch zu beteiligen ☐
II Kohärenz und Flüssigkeit ▪ Verknüpfungen ▪ Sprechtempo, Flüssigkeit	Sehr gut und klar zusammenhängend, angemessenes Sprechtempo ☐	Gut und zusammenhängend, noch angemessenes Sprechtempo ☐	Nicht immer zusammenhängend, durch Nachfragen kommt das Gespräch wieder in Gang ☐	Stockend bruchstückhafte Sprechweise, beeinträchtigt die Verständigung stellenweise ☐	Abgehackte Sprechweise, so dass zentrale Aussagen unklar bleiben ☐
III Ausdruck ▪ Wortwahl ▪ Umschreibungen ▪ Wortsuche	Sehr gut mit wenig Umschreibungen und wenig Wortsuche ☐	Über sehr weite Strecken angemessene Ausdrucksweise, jedoch einige Fehlgriffe ☐	Vage und allgemeine Ausdrucksweise, die bestimmte Bedeutungen nicht genügend differenziert ☐	Situationsunspezifische Ausdrucksweise und größere Zahl von Fehlgriffen ☐	Einfachste Ausdrucksweise und häufig schwere Fehlgriffe, die das Verständnis oft behindern ☐
IV Korrektheit ▪ Morphologie ▪ Syntax	Nur sehr vereinzelte Regelverstöße ☐	Stellenweise Regelverstöße mit Neigung zur Selbstkorrektur ☐	Häufige Regelverstöße, die das Verständnis noch nicht beeinträchtigen ☐	Überwiegend Regelverstöße, die das Verständnis erheblich beeinträchtigen ☐	Die große Zahl der Regelverstöße verhindert das Verständnis weitgehend bzw. fast ganz ☐
V Aussprache und Intonation ▪ Laute ▪ Wortakzent ▪ Satzmelodie	Kaum wahrnehmbarer fremdsprachlicher Akzent ☐	Ein paar wahrnehmbare Regelverstöße, die aber das Verständnis nicht beeinträchtigen ☐	Deutlich wahrnehmbare Abweichungen, die das Verständnis stellenweise behindern ☐	Wegen der Aussprache ist beim Zuhören erhöhte Konzentration erforderlich ☐	Wegen starker Abweichungen von der Standardsprache ist das Verständnis fast unmöglich ☐

Punkte aus Aufgabe 1	Punkte aus Aufgabe 2	Punkte gesamt

Anhang I: Transkripte des mündlichen Ausdrucks

S1: weiblich (Proband)
S2: männlich (Proband)
Laufende Teilnehmer-Nummer: 12
Thema: Kaffee
Bemerkung: Paarprüfung
Ausschnitt gesamt: 9:46 min
Ausschnitt Sequenz: 2:43–3:43 min

01	S1	okay also in meinem text äh geht es um äh die die möglichkeit dass ähm kaffeekonsum ah das risiko einer alkoholbändigten fettleber reduzieren kann
02	S2	((lacht))
03	S1	und ähm also es gab m einige forschungen darüber so einige wissenschaftler haben geforscht also über (.) ((schluckt)) über dieses äh ähm über dieses thema geforscht
04		und sie haben entdeckt dass wer nach ä ähm also zu viele alkoholkonsum kaffee trinkte ähm also ähm dann ähm de das risiko an eine an einer einer fettleber zu erkranken ähm reduziert wird
05		also ist reduziert
06		und ähm tja also ahm es kann sein ich äh ich hatte nie diese (.) also nicht vorher etwas so gelesen oder so

---

S1: weiblich (Proband)
S2: weiblich (Prüfer)
Laufende Teilnehmer-Nummer: 19
Thema: Kaffee
Bemerkung: Einzelprüfung
Ausschnitt gesamt: 6:56 min
Ausschnitt Sequenz: 0:59–1:59 min

01	S1	ähm auf dem text steht alkolbedingte fettleber äh also wenn man ähm (--) wenn man nach dem alkoholkonsum ä kaffee trinkt
02		dann kann man diese ä risiko von alkoholbedingte fettleber äh verringern

03		ähm und äm aber die ärzte warnen auch ä vor ein ä herz und kreislauf ä problem
04	S2	hm_hm
05	S1	wenn man so viel äm kaffee ä konsumiert
06		und ähm also mir fällt dann diese beispiel auf
07		zum beispiel hier wird tee als etwas ä richtig gesundes erhalten
08	S2	hm_hm
09	S1	und wenn man probleme hat soll man tee trinken und muss man kaffee weg lassen und ofter tee trinken
10		aber bei uns hört man ef ä ofter dass wenn man äh schwarzen tee trinkt (.) ähm

---

S1: weiblich (Proband)
Laufende Teilnehmer-Nummer: 8
Thema: Kaffee
Bemerkung: Paarprüfung
Ausschnitt gesamt: 8:18 min
Ausschnitt Sequenz: 4:08 min bis 5:08 min

01	S1	äh ja
02		äh mein text handelt von äh kaffeekonsum
03		oder das ist das thema
04		kaffee zu trinken
05		und es steht auch doch hier dass äh ä kaffeekon ä konsum kann das risiko für äh (.) für fettleber reduziern also dass es gesund ist ein bisschen kaffee zu trinken
06		und besonders wenn man zu viel alkohol getrunken hat dann ist es klug ein bisschen kaffee zu trinken weil es gut für äh die leber ist
07		und äh noch was stehts hier dass äh die ärzte warnen davör dass man nicht zu viel kaffee trinken kann weil das nicht gut für das herz ist
08		und äh ich hab eigentlich nicht so viele beispiele gefunden aber wenn man zum beispiel einen kater hat dann ist es sehr sehr gut oder notwendig vielleicht auch einen kaffee zu trinken
09		und

S1: weiblich (Proband)
Laufende Teilnehmer-Nummer: 28
Thema: Kaffee
Bemerkung: Einzelprüfung
Ausschnitt gesamt: 6:21 min
Ausschnitt Sequenz: 0:28–1:28 min

Ausschnitt: 0:28 min bis 1:28 min

01	S1	ähm okay
02		also es wurde immer diskutiert ob kaffee hat positive oder negative wirkungen
03		und es gibt unterschiedliche meinungen aber es wisch die wissenschaftler haben auch bewiesen das kann auch positive äh wirkung auf körper haben
04		und vor allem es vehindert das risiko einer alkoholdingte fettleber ((schnalzt)) ähm genau das dann
05		aber es gibt auch ä_also manche sache es ist wie alles kann zu viel es ist nicht so gut und es kann auch negativ wirken
06		und zum beispiel dass es (hier) äh ä das risko für herz und kreislauf äh kann bestehen
07		ähm was war noch ähm
08		zum beispiel ja es ist kann man auch über andere sache nachdenken wie ungesunde zähne oder
09		ja wie mans spricht zum beispiel über manche andere sache das ist zum beispiel wie wein es ist

---

S1: weiblich (Proband)
Laufende Teilnehmer-Nummer: 7
Thema: Kaffee
Bemerkung: Einzelprüfung
Ausschnitt gesamt: 8:18 min
Ausschnitt Sequenz: 1:41 min 2:41 min

01	S1	okay ähm

02 nach mehrere forschung ähm könne kaffeekonsum das risiko einer fettleber äh reduzieren
03 aber ähm gleichzeitig ähm das kaffeekonsum kann auch zu schäde am herz und ähm kreislauf ähm nä äh kreislauf äh führen
04 und ähm (.) meiste chinesisch trinken nicht so viel kaffee
05 wir trinken ä tea ja
06 ähm und ich glaube äh auch äh auch das teil der chinesischen die fettleber haben ist relativ kleiner als die äh europäischen
07 äh aber das aber äh ich glaube das ist äh ähm das hat auch vielleicht etwas zu tun mit ä unser diät
08 wir äh wir essen vielleicht mehr äh mehr

S1: männlich (Proband)
Laufende Teilnehmer-Nummer: 89
Thema: Kaffee
Bemerkung: Paarprüfung
Ausschnitt gesamt: 13:44 min
Ausschnitt Sequenz: 4:46 min 5:46 min

01 S1 es handelt sich um äh kaffeekomsum
02 also äh es gibt ein also vor kurz äh oh_nä schuldigung
03 also äh man weiß wenn man also ganz viel alkohol trinkt
04 dann gibts ein risiko dass man unter dem also fettleber erkranken könnte
05 und ah äh vor kurzem gibts eine forschung und nach dieser forschung sagt sagten also also die forscher dass wenn man ganz viel also genügend kaffee nach ganz viel alkoholkonsum trinken
06 dann wird diese ri risiko dieser äh fettleberrisiko also mm erheblich reduziert
07 m äh das war dieser forschung
08 und äh äh viele ärzte sagen auch dagegen dass äh zwar das risiko

Keine Zeit für den C-Test?

Der C-Test gilt als ein objektives, reliables und valides Instrument zum Erheben allgemeiner Sprachkompetenz. Zahlreiche Studien belegen den Zusammenhang des C-Tests mit verschiedenen sprachlichen Teilfertigkeiten. Im Gegensatz dazu ist der S-C-Test – eine stark beschleunigte Variante des Testformats – noch kaum erforscht. Diesem Desiderat kommt der vorliegende Band nach. Er beschäftigt sich mit der Frage, wie sich eine drastische Verkürzung der Bearbeitungszeit auf den C-Test auswirkt. Zugrunde liegt die Hypothese, dass eine Geschwindigkeitskomponente dazu beitrage, Sprachverwendung in Echtzeit zu simulieren und sich so der Zusammenhang zwischen der Leistung in Hörverstehens- und mündlichen Tests mit den C-Test-Ergebnissen erhöhen könnte.

ISBN 978-3-7983-3076-4 (print)
ISBN 978-3-7983-3077-1 (online)

9 783798 330764

http://verlag.tu-berlin.de