

So viele Lücken, so wenig Zeit: Die Rolle der Zeit im Konstrukt des deutschen C-Tests anhand der Analyse der Verarbeitungsprozesse

Anna Timukova, Oleksandra Yazdanfar & Anastasia Drackert

Abstract: Zeitreduzierte C-Tests sind eine relativ neue Variante des seit Jahrzehnten etablierten und erforschten Messverfahrens zur reliablen und ökonomischen Erfassung allgemeiner Sprachkompetenz. Die Rolle der Zeit im Konstrukt ist bislang wenig untersucht, insbesondere ihre Auswirkungen auf die Verarbeitungsprozesse. Die vorliegende Studie liefert Belege für die Prozessvalidität von zeitreduzierten C-Tests, indem sie das Bearbeitungsverhalten von 16 relativ fortgeschrittenen DaF-Lernenden beobachtet. Die Analyse von Bildschirmaufnahmen der Bearbeitung eines kanonischen und eines zeitreduzierten C-Tests zeigt eine vorwiegend lineare Bearbeitung unabhängig von der Zeit. Unter erhöhtem Zeitdruck lässt sich zudem weniger *Backtracking* (Zurückspringen im Text) beobachten, selbst bei kontextabhängigen Lücken. Der zeitreduzierte C-Test scheint die automatisierte Verarbeitung zu begünstigen und den Zugriff auf nicht-automatisiertes, deklaratives Sprachwissen zu erschweren. Dies legt nahe, dass er ein vom kanonischen C-Test teilweise abweichendes Konstrukt erfasst.

So Many Gaps, So Little Time: The Role of Time in the Construct of the German C-Test Based on the Analysis of Response Processes

Time-reduced C-Tests are a recent variation of the long-established and researched method for a reliable and efficient assessment of general language proficiency. The role of time in the construct remains underexplored, particularly its impact on response processes. This study provides evidence for the process validity of time-reduced C-Tests by observing the behaviour of 16 fairly advanced learners of German as a foreign language (DaF). Analysis of screen recordings from completing a canonical and a speeded C-Test reveals a predominantly linear approach to test completion, regardless of time. With increased time pressure, less backtracking is observed, even with context-dependent gaps. The time-reduced C-Test appears to favour automated processing, making access to non-automatised, declarative language knowledge more difficult. This suggests that it captures a construct that partially deviates from that of the canonical C-Test.

Schlagwörter: Sprachtestforschung, Testkonstrukt, Konstruktvalidität, kognitive Verarbeitung, Bildschirmaufzeichnung, deklaratives Wissen, Automatisierung; language assessment, test construct, construct validity, response processes, screen recording, declarative knowledge, automaticity

Timukova, Anna; Yazdanfar, Oleksandra & Drackert, Anastasia (2026):
So viele Lücken, so wenig Zeit: Die Rolle der Zeit im Konstrukt des deutschen
C-Tests anhand der Analyse der Verarbeitungsprozesse.
Zeitschrift für Interkulturellen Fremdsprachenunterricht 31: 1, 81–108.
<https://doi.org/10.48694/zif.4262>

1 Einleitung

Sprachenzentren und Sprachschulen weltweit stufen jährlich Tausende Lernende in ihre Sprachkurse ein. Dafür brauchen sie valide, reliable und praktikable Instrumente. Seit Jahrzehnten werden zu Einstufungszwecken C-Tests genutzt (Grotjahn/Drackert 2022). Sie können schnell erstellt und ausgewertet werden, wobei computerbasierte C-Tests auch eine (teil)automatisierte Erstellung und automatische Auswertung zulassen. Beim C-Test ergänzen Sprachlernende partiell getilgte Wörter in mehreren kurzen Texten und haben in der Regel fünf Minuten Bearbeitungszeit pro Text (Testbearbeitungszeit von bis 40 Minuten), was sich in der Praxis nicht immer als ökonomisch herausstellt. Was passiert, wenn man die Bearbeitungszeit reduziert? Werden Testteilnehmende dieselbe Leistung zeigen? Und vor allem – wird der Test dasselbe messen? Die Beantwortung dieser Fragen ist Teil der Validierung der neuen zeitreduzierten (*speeded*) Formatvariante des C-Tests.

Obwohl moderne Validitätsmodelle (z.B. Kane 2016) die Bedeutung der Testverwendung und somit der mit der Testleistung verbundenen Interpretationen und Entscheidungen betonen, bleibt es essenziell zu klären, was ein Testformat tatsächlich misst. Abhängig vom Testzweck werden dazu verschiedene Belege gesammelt, etwa zur Inhaltsvalidität, konvergenten Validität, Übereinstimmungsvalidität oder zur Validität der Verarbeitungsprozesse (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education 2014). Die letzte Art der Validität, auch substantielle Validität (*substantive validity*) genannt (Messick 1989) fokussiert auf Nachweise für die (kognitiven) Prozesse der Testbearbeitung. Sie umfasst theoretische Erklärungen für beobachtete Regelmäßigkeiten in den Testantworten, einschließlich Prozessmodellen der Aufgabenbearbeitung (Embretson 1983), sowie empirische Belege dafür, dass die postulierten kognitiven Prozesse bei der Aufgabenlösung tatsächlich ablaufen. Obwohl erste Erkenntnisse zum Konstrukt von zeitreduzierten C-Tests vorliegen (Timukova/Möller/Drackert im Druck; Zimmermann 2019), fehlen bislang Belege für ihre Prozessvalidität – eine Lücke, die diese Studie adressiert.

2 Verarbeitungsprozesse im Konstrukt des C-Tests

2.1 Konstrukt des C-Tests

C-Tests werden zur integrativen Messung allgemeiner Sprachkompetenz eingesetzt (Grotjahn/Drackert 2022). Sie operationalisieren das Prinzip der reduzierten sprachlichen Redundanz: Sprachkompetente Personen können durch antizipatorische Verarbeitung auch wenig redundante Mitteilungen entschlüsseln (Klein-Braley 1994). Für die Rekonstruktion der Lücken ist eine integrierte Anwendung sprachlicher Wissensbestände und Fertigkeiten in Schriftform erforderlich (Hastings 2002), wobei die Verarbeitung auf allen Ebenen vom

Buchstaben bis zum Text erfolgt (Sigott 2004) und Strategien einschließt (Harsch/Schröder 2007).

Lernende müssen das partiell getilgte Wort im mentalen Lexikon haben, es kontextbasiert identifizieren und grammatikalisch sowie orthographisch korrekt rekonstruieren (Drackert/Timukova 2020). Dabei wird sowohl deklaratives (z.B. Wortschatz, explizite Regeln) als auch prozedurales (z.B. syntaktische Verarbeitung, Rechtschreibung) Wissen aktiviert¹.

Raatz (2002) zeigt in einer Faktorenanalyse hohe Ladungen auf den Faktor *verbal intelligence* für Deutsch als Erstsprache und erklärt die C-Test-Lösung durch eine konzeptgesteuerte Verarbeitung. Dabei bildet die Analyse des vorhandenen sprachlichen Materials die Basis für Hypothesen zu fehlenden Wortteilen, während größeres Sprachwissen eine schnellere und effizientere Verarbeitung ermöglicht.

Zahlreiche empirische Befunde belegen die Übereinstimmungsvalidität (konkurrente Validität) von C-Tests, verstanden als der Grad der Übereinstimmung mit den Ergebnissen etablierter Sprachtestverfahren. McKay et al. (2021) analysierten Korrelationen zwischen C-Tests in verschiedenen Sprachen und anderen Sprachtests aus 239 Studien und kamen zum Schluss, dass das Konstrukt des C-Tests allgemeine Sprachkompetenz umfasst, die aus mehreren Komponenten besteht. Auch Faktorenanalysen stützen diese Annahme (z.B. Eckes/Grotjahn 2006).

Studien zu den kognitiven Prozessen beim Lösen von C-Tests sind seltener und ihre Ergebnisse uneinheitlich. Während die Komplexität der notwendigen Sprachverarbeitung eine Synthese verschiedener Verarbeitungsprozesse nahelegt, wird weiterhin diskutiert, ob sich Testpersonen primär auf mikro- oder makrostrukturelle Prozesse konzentrieren. Die Tiefe der Verarbeitung hängt zudem von Textschwierigkeit und Sprachkompetenz der Testpersonen ab – ein Merkmal, das Sigott (2004) als *fluid construct* bezeichnet (für mögliche Ausprägungen von diesem Merkmal im Kontext der herkunftssprachlichen Sprachkompetenz siehe z.B. Drackert/Timukova 2020). Da Verarbeitungsprozesse im Fokus der Untersuchung stehen, werden sie in Kap. 2.3 näher erläutert.

2.2 Konstrukt des C-Tests mit Zeitreduktion

Die zeitreduzierte Variante des C-Tests wurde ursprünglich in Intelligenzstudien zur Messung der Sprachkompetenz in der Erstsprache verwendet, um Deckeneffekte bei Muttersprachler*innen zu vermeiden (Raatz 2002; Wockenfuß 2008). Aguado, Grotjahn und Schlak (2005) setzten einen zeitlimitierten C-Test für Deutsch als Fremdsprache (L2) bei

¹ Diese Unterscheidung basiert auf dem *Declarative/Procedural Model* (Ullman 2020; Morgan-Short/Ullman 2022), das zwei miteinander interagierende Gedächtnissysteme beschreibt: Das deklarative Gedächtnis speichert bewusst zugängliches Wissen wie Wortbedeutungen, unregelmäßige Formen, *chunks* und gelernte Regeln. Das prozedurale Gedächtnis umfasst unbewusstes, automatisiertes Wissen und ist zentral für die Verarbeitung regelhafter sprachlicher Strukturen und vorhersagbarer Elemente und Sequenzen wie Morphologie und Syntax.

weit fortgeschrittenen Sprachlernenden ein, um ihr deklaratives Sprachwissen und dessen Prozeduralisierungsgrad integriert zu messen.

Erste Erkenntnisse zum Konstrukt zeitreduzierter C-Tests in der L2-Forschung zeigen, dass diese neben Sprachwissen auch Automatisierung sowie Verarbeitungseffizienz messen können. Zimmermann (2019) prüfte Grotjahns (2010) Hypothese, dass zeitreduzierte C-Tests stärker mit mündlichen Fertigkeiten zusammenhängen. Sie korrelierte die Punktwerte aus kanonischen und zeitreduzierten C-Tests mit den Ergebnissen aus Prüfungsteilen Hörverstehen und Sprechen des Goethe-Zertifikats bei DaF-Lernenden (B2-Niveau). Die Ergebnisse stützen Grotjahns Annahme teilweise: Die Korrelationen fielen etwas – jedoch nicht signifikant – höher für den zeitreduzierten C-Test aus.

Timukova et al. (im Druck) untersuchten Testergebnisse von über 400 Studierenden mit L2 Englisch oder Deutsch in computerbasierten C-Tests unter zwei Zeitbedingungen. Zeitlimitierte Tests erwiesen sich in beiden Sprachen als signifikant schwieriger und sensibel für das Sprachniveau der Lernenden sowie die Textschwierigkeit (im Deutschen). Sie korrelierten etwas stärker mit dem *Oral Elicited Imitation Test*² und sagten dessen Ergebnisse besser voraus. Die Autorinnen interpretieren diese Befunde als Hinweis auf einen höheren Anteil automatisierter Sprachfähigkeiten im zeitlimitierten Konstrukt, betonen jedoch die Notwendigkeit, das Antwortverhalten zu untersuchen.

Zu den Verarbeitungsprozessen bei zeitreduzierten C-Tests gibt es bislang nur erste, vorläufige Ergebnisse, die auf dem berichteten Verhalten der Testteilnehmenden basieren. In Zimmermanns Studie (2019) beantworteten 120 Proband*innen nach der Bearbeitung beider Testversionen (kanonisch und zeitreduziert) eine Einzelfrage zu ihrem Lösungsverhalten anhand vorgegebener Antwortoptionen. Weniger als die Hälfte gab an, die Lücken linear (d.h. in der Reihenfolge ihres Auftretens im Text) bearbeitet zu haben, unabhängig von der Zeit, während Rückwärtsbewegungen (*Backtracking*) im kanonischen C-Test häufiger als im zeitreduzierten vorkamen. Zimmermann schloss daraus, dass der Zeitfaktor das Lösungsverhalten zwar beeinflusst, die Methode jedoch keine eindeutigen Rückschlüsse auf zugrunde liegende Prozesse zulässt.

Diese Forschungslücke steht im Mittelpunkt der vorliegenden Studie. Da bereits Untersuchungen zu Verarbeitungsprozessen im kanonischen C-Test existieren, werden diese im folgenden Kapitel vorgestellt.

² Im *Oral Elicited Imitation Test* (OEIT) hören die Teilnehmenden gesprochene Sätze in der Zielsprache und wiederholen sie nach einer kurzen Pause; die Antworten werden aufgezeichnet und von Rater*innen bewertet. Ursprünglich zur Messung allgemeiner Sprachkompetenz entwickelt, wird der OEIT inzwischen auch zur Einschätzung globaler mündlicher Sprachkompetenz in der L2 eingesetzt (z.B. Drackert 2016).

2.3 Verarbeitungsprozesse in der C-Test-Forschung

2.3.1 Konzepte, Methoden, Klassifikationen

Die C-Test-Forschung stützt sich bei der Beschreibung von Verarbeitungsprozessen auf Modelle kognitiver Prozesse beim (Testen von) Lesen, welche Bottom-up- und Top-down-Verarbeitung, Schemaaktivierung sowie das Zusammenspiel zwischen den Eigenschaften der Lesenden und der Texte betonen (vgl. Jang 2017). Khalifa und Weir (2009) unterscheiden zwischen hierarchieniedrigen Prozessen wie Dekodieren, Worterkennung und Bedeutungsbildung auf Satzebene und hierarchiehohen Prozessen wie Textverknüpfung und inferentiellem Lesen. Hierarchieniedrige Prozesse können automatisiert werden und unbewusst ablaufen (Grabe/Yamashita 2022).

Trotz verwandter Begriffe lassen sich Erkenntnisse aus der Leseforschung nach unserer Einschätzung nur bedingt auf das Lösen von C-Tests übertragen: Die unvollständigen Texte müssen rekonstruiert werden, was die ablaufenden Leseprozesse notwendigerweise beeinflusst und zusätzlich eine produktive (z.B. morphologische) Verarbeitung erfordert. Ein übertragbarer Befund der Lesetestforschung betrifft jedoch die Interaktion zwischen Texteigenschaften und der Sprachkompetenz der Testteilnehmenden, welche die zum Einsatz kommenden Prozesse steuert.

Drei methodische Ansätze dominieren die Literatur zu Lösungsprozessen in C-Tests. Teilnehmende verbalisieren ihr Vorgehen direkt (Lautes Denken) oder retrospektiv (*stimulated recall*) (Babaii/Ansary 2001; Stemmer 1991), ihr Bearbeitungsverhalten wird z.B. mit Key-Stroke-Logging (Aufzeichnung von Tastenanschlägen) verfolgt und protokolliert (Germann/Grotjahn 1994), oder die Antworten werden im Nachhinein analysiert (Hastings 2002; Klein-Braley 1994; Sigott 2004). Diese Verfahren erlauben Zugang zu unterschiedlichen Untersuchungsebenen, weshalb Lösungsprozesse unterschiedlich klassifiziert werden.

Analysen der Lernendenantworten, wie etwa Fehleranalysen, richten den Fokus auf den Text und unterscheiden die Prozesse meist anhand des Umfangs des für die Lösung herangezogenen Kontexts. Dabei sprechen Forschende von Verarbeitung auf niedrigerer und höherer Ebene (z.B. Klein-Braley 1994; Sigott 2004), Mikro- vs. Makroebene (z.B. Beinborn/Zesch/Gurevych 2014), Textebene (*text-level processing*) vs. Satzebene und darunter (Sigott 2004) oder von Mikro- vs. Makrokontext (z.B. Grotjahn/Schiller 2014).

Beobachtungen des **Bearbeitungsverhaltens** liefern alternative Klassifikationen. Germann und Grotjahn (1994) untersuchen in einer Pilotstudie mit acht fortgeschrittenen Englischlernenden die Bearbeitungsprozesse mithilfe der Protokollierung von Tastatureingaben am Computer und identifizieren drei Vorgehensweisen: lineare Bearbeitung der Lücken nacheinander, wiederholtes Zurückspringen zu ungelösten Lücken vor Erreichen des

letzten Items und häufiges Überspringen ohne anschließende zeitnahe Bearbeitung (Germann/Grotjahn 1994: 288).

Introspektive Verfahren führen zwangsläufig zu eigenen, je nach Studie variierenden Klassifikationen. Stemmer (1991) identifizierte in den Aussagen von 30 Personen beim Lösen französischer C-Tests nur in 41% der Fälle eine der drei Verhaltensweisen: *recall strategies* (Abruf des gesuchten Elements durch Rückgriff auf das Element selbst oder den unmittelbaren Kontext), *evaluation strategies* (Einbezug weiterer Ko- und Kontexte zur Überprüfung der Passung der gefundenen Lösung) und „*superstrategy*“ (z.B. Vorlesen des (ganzen) Textes). Abhängig davon, ob sich diese Strategien innerhalb einer Proposition bewegen oder über Propositionsgrenzen hinausgehen, ordnete sie diese in *low-level* und *higher-level comprehension* ein. Babaii und Ansary (2001: 215–216) erstellten anhand retrospektiver verbaler Protokolle von 32 Teilnehmenden beim Lösen eines englischen C-Tests eine andere Klassifikation: 1) automatische Verarbeitung, 2) Nutzung des unmittelbaren Kontexts der Lücke (*lexical adjacency*), 3) Nutzung des grammatikalischen Kontexts und der Bedeutung benachbarter Sätze (*sentential cues*) und 4) Nutzung von Hintergrundwissen und des Gesamttexts (*top-down cues*). Die ersten beiden Lösungsprozesse ordnen sie der Mikro- und die letzten beiden der Makroebene zu, liefern jedoch keine Begründung für diese Zuordnung.

Diese Klassifikationen zeigen, dass Konzepte aus der Leseforschung von den Forschenden unterschiedlich interpretiert und verwendet werden, oft ohne explizite Begründung. Sigott (2004) folgt offenbar der Unterscheidung von Khalifa und Weir (2009) in *lower-* und *higher-level processing*, indem er die Grenze zwischen der Verarbeitung auf Text- und Satzebene zieht. Hingegen ordnen Babaii und Ansary (2001) ihre Kategorie „*sentential cues*“ der Makroebene zu. Beinborn et al. (2014) beschränken den Mikrokontext auf die Lücke mit je einem Wort links und rechts davon, alles andere zählt als Makrokontext.

Es überrascht kaum, dass unterschiedliche Methoden mit variierenden Klassifikationen zu uneinheitlichen Ergebnissen über die primär genutzten Verarbeitungsprozesse beim Lösen von C-Tests führen. Diese Ergebnisse werden im Folgenden kurz zusammengefasst.

2.3.2 Erkenntnisse

Introspektive Verfahren von Stemmer (1991) sowie Babaii und Ansary (2001) weisen auf eine bevorzugte Verarbeitung auf lokaler Ebene beim C-Test hin: Der unmittelbare Kontext wird am häufigsten (ca. 55 %) und die Makroebene seltener (*sentential cues* - 22 %, Top-down-Verarbeitung - ca. 6 %) einbezogen (Babaii/Ansary 2001). Beide Studien zeigen, dass die Makroebene vor allem dann aktiviert wird, wenn die automatische, serielle Bearbeitung scheitert. Stemmer (1991) beobachtete, dass der breitere Kontext meist zur Lösungsevaluation dient, während die Verarbeitung auf Textebene vor allem bei Schwierigkeiten in der lokalen Lösungssuche als problemlösende Strategie eingesetzt wird.

Klein-Braley (1994) und Hastings (2002) analysierten Lernendenantworten und betonten hingegen die Relevanz makrostruktureller Verarbeitungsprozesse. Lokale Verarbeitung dominierte bei Testpersonen mit schlechten Ergebnissen, während leistungsstärkere Lernende den Makrokontext berücksichtigten. Schwächere Teilnehmende erkannten oft nur einzelne Wörter und erfassen den Text nicht als Ganzes, was entweder zu einer vorschnellen Festlegung auf eine Lösung ohne Berücksichtigung der nachfolgenden Textbezüge (*early closure*) oder zu einer Bearbeitung mit ausschließlich lokalem Fokus auf die unmittelbare Lückenumgebung (*narrow focus*) führte (Klein-Braley 1994). Gleichzeitig wird argumentiert, dass die Kürze der C-Test-Texte eine eindeutige Bestätigung der Nutzung globaler, integrativer Verarbeitungsprozesse erschwert.

Mehrere Studien betonen den Zusammenhang zwischen den Verarbeitungsprozessen und Lernendenkompetenz. Grotjahn und Schiller (2014) zeigten, dass steigende Sprachkompetenz der Testpersonen mit besserer Nutzung größerer Rekonstruktionskontexte korrespondiert. Laut Babaii und Fatahi-Majd (2014) führen Kompetenzunterschiede zu verschiedenen Fehlerquellen: Während kompetente Lernende vor allem lexikalische Abrufprobleme haben und feine Grammatikfehler übersehen, gelingt es leistungsschwächeren Lernenden häufig nicht, den Gesamtzusammenhang des Textes aufzubauen, sodass sie sich auf unsystematische, lokale Rekonstruktionen stützen. Für Sigott (2004) stellt die Abhängigkeit der Verarbeitungsprozesse vom Kompetenzniveau sogar ein zentrales Merkmal des Konstrukts (*fluid construct*) dar. Er berichtete folgende Ergebnisse: Schwächere Lernende benötigen mehr Kontext, um fehlendes Sprachwissen zu kompensieren, während fortgeschrittene Lernende in der Lage sind, einige Lücken allein durch den unmittelbaren Kontext zu schließen.

Germann und Grotjahn (1994) sahen in der Top-down-Verarbeitung eine der erfolgreicheren Lösungsstrategien, betonten jedoch den Einfluss individueller und textueller Faktoren auf deren Wahl. Sie stellten nur exemplarische Bezüge zwischen beobachtetem Verhalten und zugrundeliegenden Prozessen her und bedauerten die fehlende Erhebung temporaler Daten.

Da die Eigenschaften von Texten und Lücken die Verarbeitungsprozesse beeinflussen, werden sie sowohl zur Kategorisierung der Prozesse herangezogen als auch direkt untersucht. Stemmer (1991) identifizierte z.B. die Anzahl möglicher Vervollständigungen der Lücke und kohäsive Verknüpfungen im Text als Faktoren, die die Verarbeitungslast erhöhen. Klein-Braley (1994) zeigte, dass Strukturwörter (*function words*) wie Präpositionen und Konjunktionen zwar oft leichter zu rekonstruieren sind als Inhaltswörter, die Lückenschwierigkeit jedoch von der Wort-Kontext-Interaktion abhängt. Sigott (2004) bestätigte, dass beide Wortarten je nach Kontext die Verarbeitung auf Textebene auslösen können. Nach Kaufmann (2016) stellen bestimmte Merkmale der deutschen Sprache, die für DaF-Lernende besonders herausfordernd sind, potenzielle Fehlerquellen in einem C-Test dar.

Für Beinborn et al. (2014) ergibt sich die Schwierigkeit einzelner Lücken aus Wortmerkmalen (z.B. Vertrautheit, Flexion) sowie Makrofaktoren wie Interlückendependenz und Textschwierigkeit.

3 Vorhaben dieser Studie

3.1 Untersuchungsgegenstand und Methode

Bisher liegen keine Erkenntnisse zu den Verarbeitungsprozessen in zeitlimitierten C-Tests vor (vgl. Zimmermann 2019). Daraus ergibt sich die Notwendigkeit einer theoretischen und empirischen Klärung hinsichtlich der Prozessvalidität – und damit auch die Begründung für unsere Studie. Da bestehende Studien zu Prozessen im kanonischen C-Test uneinheitliche Definitionen und teils vage Kategorisierungen verwenden (Kap. 2.3.1), präzisieren wir zunächst den Untersuchungsrahmen, definieren den Gegenstand unserer Analyse und erläutern die Beziehungen zwischen den Untersuchungsebenen.

Der Gegenstand unserer Untersuchung sind Bearbeitungsprozesse (*response processes*) nach Hubley und Zumbo (2017: 2): „the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation.“ Diese Definition umfasst neben kognitiven Prozessen auch das Verhalten sowie Motivation und Emotionen. Die Untersuchung der Bearbeitungsprozesse liefert über die Analyse der Testpunktwerte hinaus wertvolle Einblicke in das zugrunde liegende Konstrukt und trägt somit wesentlich zur Validierung bei.

Um diese Bearbeitungsprozesse empirisch zugänglich zu machen, verfolgen wir das Bearbeitungsverhalten, in dem sie sich manifestieren. Diese Methode ermöglicht einen direkteren Zugang zu den Prozessen im Vergleich zu einer Fehleranalyse. Gleichzeitig ist sie weniger intrusiv als das Verbalisieren mentaler Vorgänge und umgeht einige Einschränkungen introspektiver Methoden, wie etwa die Möglichkeit, nur bewusst ablaufende Prozesse zu verbalisieren, und den potenziellen Einfluss der Verbalisierung auf das Verhalten³ (Germann/Grotjahn 1994). Das beobachtbare Bearbeitungsverhalten wird durch Bildschirmaufnahmen der Testpersonen in Form von Videos dokumentiert. Da die Bearbeitung durch die Videoaufnahme in ihrem zeitlichen Ablauf erfasst wird, beseitigt diese Methode eine wichtige Einschränkung früherer Studien (vgl. Germann/Grotjahn 1994; Stemmer 1991).

Eine relativ homogene Stichprobe von Teilnehmenden mit Deutschkompetenz auf B2-C1-Niveau des Gemeinsamen europäischen Referenzrahmens für Sprachen (GeR; Europarat

³ Der Verzicht auf introspektive Verfahren schränkt unseren Untersuchungsgegenstand insofern ein, als komplexe Testlösungsverhaltensweisen wie Strategien und die emotionale Komponente der *response processes* nicht berücksichtigt werden.

2020) wurde gewählt, um den Einfluss der Sprachkompetenz auf die genannten Prozesse zu kontrollieren. Der Einfluss der Lückeneigenschaften auf die Verarbeitungsprozesse wird durch eine Klassifikation der Lücken nach dem minimal erforderlichen Lösungskontext berücksichtigt: Die Unterscheidung zwischen Mikro- und Makrolücken (s. Kap. 4.2) fließt in unsere Forschungsfragen ein.

3.2 Annahmen und Forschungsfragen

Der C-Test eliziert realitätsnahe Sprachverarbeitung: Umfangreicheres linguistisches Wissen führt zu besserem Umgang mit reduzierter Redundanz und schnellerer Verarbeitung (Kap. 2.1). Wir erwarten, dass hochkompetente Sprachnutzer*innen die meisten Lücken durch automatischen Abruf ihres Sprachwissens lösen. Manche Lösungen werden (schnell) evaluiert, während komplexe und zeitaufwendige Lösungssuchprozesse selten auftreten – etwa bei kontextabhängigen Lücken oder schwierigen Texten. Problemlösungsstrategien und strategisches Antwortverhalten (*test wiseness*) spielen nur in Ausnahmefällen eine Rolle. Mit abnehmender Sprachkompetenz verändert sich die Verteilung der Testlösungsprozesse: Geringeres und weniger prozeduralisiertes Sprachwissen erschwert und verlangsamt die Verarbeitung, die dann stärker auf deklarativem, nicht-automatisiertem Wissen basiert und häufiger Evaluations- und Suchprozesse erfordert.

Bei der C-Testbearbeitung werden niedrigere kognitive Prozesse der Worterkennung und der syntaktischen Analyse durch höhere Prozesse der Hypothesenbildung und Textverknüpfung ergänzt, um Redundanzverluste auszugleichen. Je nach Lückenart erfolgt die Wiederherstellung eines getilgten Wortteils automatisch über das mentale Lexikon oder erfordert eine syntaktische Verarbeitung der Umgebung (in Echtzeit oder verzögert). Lösungsevaluation und -suche können lokal durch syntaktische Analyse auf Teilsatzebene oder global durch Inferenzbildung auf Textebene ablaufen. Somit nähert sich die C-Testbearbeitung auf dem Kontinuum zwischen Leseverstehensaufgaben und *language-in-use tasks*⁴ tendenziell eher Letzteren an.

Aus diesem Grund und aufgrund der Kürze der Texte schlagen wir vor, zwischen der automatisierten Beseitigung reduzierter Redundanz und der aufmerksamen, zeit- sowie kognitiv anspruchsvolleren Verarbeitung zu unterscheiden. Diese Unterscheidung erscheint uns zielführender als eine Einteilung in hierarchieniedrige und -hohe Ebenen. Daher untersuchen wir in unserer Studie zwei Arten von Verarbeitungsprozessen beim Lösen eines C-Tests:

⁴ *Language-in-use tasks* sind Aufgabenformate, die sprachliche Kompetenz nicht isoliert, sondern im funktionalen Gebrauchskontext prüfen – etwa durch das Verknüpfen, Ergänzen oder Interpretieren sprachlicher Einheiten im Kontext. Sie erfordern typischerweise ein Zusammenspiel von grammatischem, lexikalischem, pragmatischem und strategischem Wissen (vgl. Ellis 2003; McCray/Brunfaut 2018).

- Automatisierte Verarbeitung: schnelle, serielle, unbewusste Verarbeitung überwiegend auf lokaler Ebene
- Kognitiv anspruchsvolle Verarbeitung: verzögerte, nicht-lineare, bewusste Verarbeitung, die i.d.R. einen breiteren Kontext analysiert, evaluiert und/oder integriert

Welche Prozesse dominieren, hängt von Faktoren wie der Lernendenkompetenz, den Lücken- und Texteigenschaften sowie der Bearbeitungszeit ab. Die Bearbeitung von Mikrolücken fördert automatisierte Prozesse, während Makrolücken eine bewusste und zeitintensive Verarbeitung begünstigen. Eine automatisierte Verarbeitung dominiert bei hoher Sprachkompetenz und nimmt mit geringerer Kompetenz ab, während Zeitreduktion automatisierte Prozesse verstärkt aktiviert.

Daraus ergeben sich folgende Forschungsfragen:

- 1) Wie unterscheidet sich die Bearbeitung vom kanonischen im Vergleich zum zeitreduzierten C-Test?

Hypothese: Der zeitreduzierte C-Test wird vorwiegend linear bearbeitet, mit wenig Überspringen und Zurückkehren. Beim kanonischen C-Test findet häufiger eine nicht-lineare Bearbeitung statt.

- 2) Wie unterscheidet sich das Antwortverhalten beim erfolgreichen Ausfüllen von Mikrolücken?

Hypothese: Die zügige, lineare erfolgreiche Bearbeitung der Mikrolücken wird im zeitreduzierten C-Test und im kanonischen C-Test vergleichbar häufig vorkommen.

- 3) Wie unterscheidet sich das Antwortverhalten beim erfolgreichen Ausfüllen von Makrolücken?

Hypothese: Im kanonischen C-Test wird die nicht-lineare sowie verzögerte erfolgreiche Bearbeitung der Makrolücken häufiger vorkommen als im zeitreduzierten C-Test.

4 Methodik

4.1 Proband*innen, Instrumente und Datenerhebung

An der Studie nahmen 16 ausländische Studierende (9 weiblich, 7 männlich; Alter zwischen 19 und 31 Jahren ($M = 22,75$ Jahre); unterschiedliche L1s) einer großen Universität in Deutschland teil, die ihre Deutschkenntnisse auf B2 (10 Personen) oder C1-Niveau (sechs Personen) des GeR einschätzten.

Um die Sprachkompetenz der Testpersonen zu kontrollieren, setzten wir die neue Variante der *Elicited Imitation Task* in schriftlicher Form ein. Der WEIT (*Written Elicited Imitation Task*)⁵ ist ein Messverfahren, das prozedurales Sprachwissen erfasst, welches sich im Umgang mit schriftsprachlichen Mustern entwickelt und bei der Bearbeitung der Testaufgaben in schriftlicher Form weitgehend automatisiert aktiviert wird. Die WEIT-Ergebnisse betragen zwischen 17 und 67 Punkten ($M = 45,8$; $SD = 13,7$ auf einer Skala von 80 Punkten).

Die eingesetzten C-Tests bestanden jeweils aus vier unterschiedlichen Texten mit je 20 Lücken. Insgesamt wurden also acht Texte aus der Datenbank des Online-Spracheinstufungstests (onSET)⁶ ausgewählt und nach aufsteigender Schwierigkeit angeordnet, sodass sich zwei parallele Testversionen mit vergleichbarem Schwierigkeitsgrad und thematischer Vielfalt ergaben. Mit „Schwierigkeit“ ist hier die Rasch-Itemschwierigkeit gemeint: Die Texte werden auf einer gemeinsamen Logit-Skala angeordnet, auf der schwierigere Texte niedrigere erwartete Punktwerte bei Personen mit derselben Ausprägung auf dem zugrunde liegenden Fähigkeitskontinuum hervorrufen (Eckes, 2011). Linguistische Merkmale (z.B. Satzlänge, lexikalische Diversität, Worthäufigkeit) sind mögliche Quellen dieser empirisch bestimmten Schwierigkeit (vgl. Beinborn et al. 2014; Kaufmann 2016).

Alle Testteilnehmenden haben alle acht Texte in den beiden Testversionen bearbeitet. Für die Analyse wurden Bildschirmaufnahmen der Texte 2 und 4 sowohl aus der kanonischen als auch aus der zeitreduzierten Testversion herangezogen. Text 1 diente als Icebreaker und zur Familiarisierung sowohl mit dem C-Testformat als auch mit dem Aufnahme-Tool. Der mittelschwere Text 3 wurde ausgeschlossen, um einen deutlicheren Kontrast in Bezug auf die Schwierigkeit zu erzielen.

Die Bearbeitungszeit war textspezifisch festgelegt: jeweils 5 Minuten für die Texte der kanonischen Testversion und zwischen 1:30 und 2:00 Minuten für die Texte der zeitreduzierten Version (1:40 Minuten für Text 2; 2:00 Minuten für Text 4)⁷. Eine vorzeitige Abgabe war möglich, sodass die durchschnittlichen Bearbeitungszeiten in der kanonischen Version bei 3:17 Minuten ($SA = 0,93$) für Text 2 und 3:51 Minuten ($SA = 0,94$) für Text 4 lagen. In der zeitreduzierten Version betragen sie 1:38 Minuten ($SA = 0,16$) für Text 2 und

⁵ Das Instrument für Englisch, Deutsch und Russisch wurde im Rahmen des DFG-Projekts (Drackert n.d.) entwickelt, pilotiert und eingesetzt (für Deutsch siehe Chifligarov/Laâguidi/Schellenberg/Dill/Timukova/Drackert/Laarmann-Quante 2025).

⁶ <https://www.onset.de/> (25.10.2022)

⁷ Die Zeitlimits wurden durch einen Pretest mit Muttersprachler*innen ($N = 16$) ermittelt, die gebeten wurden, die C-Test-Texte möglichst schnell und genau zu bearbeiten. Die durchschnittlichen Bearbeitungszeiten der Teilnehmenden, die mindestens 90 % der Aufgaben korrekt lösten, wurden berechnet und für L2-Lernende um 20 % erhöht. Diese Zeiten (1:40 bis 2:50 Minuten) wurden in der Pilotstudie verwendet und basierend auf den Durchschnittswerten der L2-Pilotteilnehmer*innen ($N = 21$) weiter angepasst.

1:58 Minuten ($SA = 0,07$) für Text 4. Die Reihenfolge der Texte war innerhalb einer Testversion fixiert.

Die Datenerhebung fand in einem Computerraum der Universität statt und dauerte eine Stunde. Nach einer Einführung in den Zweck der Studie und die Nutzung des Screen-Recorders bearbeiteten die Teilnehmenden die Tests auf der *Moodle*-Plattform. Die Bildschirme wurden bei der Bearbeitung mit dem vorinstallierten frei zugänglichen Screen-Recorder *DemoAir* (Wondershare; Wondershare n.d.) aufgenommen.

4.2 Klassifikation der Lücken

Alle Lücken in den vier analysierten C-Test-Texten wurden in die Kategorien „Mikro-“ und „Makrolücken“ unterteilt. Die Zuordnung erfolgte basierend auf dem minimalen Kontext, den hochkompetente Sprachnutzer*innen zur Lösung der Lücke benötigen. Mikrolücken können mithilfe des Mikrokontextes gelöst werden, während Makrolücken einen breiteren Kontext erfordern. Abweichend von Beinborn et al. (2014) definieren wir *Mikrokontext* als die Lücke selbst plus bis zu **zwei** Wörter davor und danach; alles darüber hinaus gilt als *Makrokontext*. Die Erweiterung des Mikrokontexts auf bis zu zwei Wörter vor und nach der Lücke ergibt sich aus der von den Autorinnen durchgeführten Lückenanalyse. Diese zeigte, dass viele Lücken zwar ein etwas weiter entferntes als das unmittelbar angrenzende Wort für die Rekonstruktion erfordern, dennoch lokal lösbar sind und nicht als Makrolücken (im Sinne textübergreifender, kontextuell komplexer Lücken; vgl. Beinborn et al. 2014) bezeichnet werden können.⁸

Der minimal erforderliche Kontext wurde durch Triangulation verschiedener Datenquellen bestimmt. Die Analyse der Lernendenantworten⁹ lieferte Schwierigkeitsindizes und häufige Antworten. Die Untersuchung jedes getilgten Wortes und seiner Position im Text zeigte, welche alternativen Fortsetzungen lokal möglich wären und welche durch den Kontext ausgeschlossen werden (vgl. Sigott 2004). Auf dieser Grundlage erfolgte die Zuordnung in Mikro- vs. Makrolücken sowie die Bestimmung der linguistischen Wissensbestände für das erfolgreiche Ausfüllen. Tabelle 1 veranschaulicht die analysierten Daten und die resultierende Zuordnung anhand von vier Lücken eines Textes.

⁸ So kann beispielsweise die Konjunktion „oder“, die zwei Substantive verbindet, mit sehr hoher Wahrscheinlichkeit lokal rekonstruiert werden. Sie kann jedoch auch in einem Satz wie „Ob im Konzertsaal od[er] im Kaufhaus, Musik begleitet uns täglich“ vorkommen, in dem eines der Substantive von einer Präposition begleitet wird und daher nicht unmittelbar an die Lücke grenzt, sondern das **zweite** Wort rechts davon ist.

⁹ Die Antworten der Lernenden konnten dem großen Datensatz ($N=188$) des Gesamtprojekts entnommen werden (vgl. Timukova et al. im Druck).

Tab. 1: Beispiel für die Lückenanalyse

Lücke	Schwierigkeit	häufigste falsche Antworten	Wortart	minimal erforderlicher Kontext	Interlückendependenz	Lückenart	erford. Wissen
i[m]	45,4	in(s) - 21,3 ich - 12,9 ist - 11,9	Präp	Ob i____ Konzertsaal od____ im Kaufh____, Musik begl ____	begl____	Makro (Kontext nachfolgend)	lexik.; syntakt.; morph.
od[er]	90,7	-	Konj	(im) Konzertsaal od____ im Kaufh____		Mikro	lexik.; syntakt.
begl[eitet]	33,0	keine Antwort - 13,9 begleiten - 11,3	Verb	Einleitungssatz und Satzteil vor der Lücke, sonst "beglückt" möglich	Ob i____ Konzertsaal od____ im Kaufh____	Makro (Kontext vorangehend)	lexik.; syntakt.; morph.
Emot[ionen]	72,2	-	Subst	Emot____		Mikro	lexik.

Die Ergebnisse der Lückeklassifikation in Tabelle 2 zeigen, dass Makrolücken etwa ein Drittel aller Lücken (14 von 40) in den analysierten Texten in beiden Testversionen ausmachen, wobei es in den schwierigeren Texten (Text 4) mehr Makrolücken gibt (jeweils acht) als in den einfacheren Texten (Text 2, jeweils sechs). Makrolücken waren im Durchschnitt schwieriger als Mikrolücken, wie die niedrigeren Schwierigkeitsindices¹⁰ verdeutlichen ($M_{makro} = 39,1$ vs. $M_{mikro} = 51,2$; $\approx 24\%$ niedriger).

Tab. 2: Verteilung der Mikro- und Makrolücken in den analysierten Texten

	kanonischer C-Test			zeitreduzierter C-Test		
	Text 2	Text 4	Gesamt	Text 2	Text 4	Gesamt
Mikro	14	12	26	14	12	26
Makro	6	8	14	6	8	14
Gesamt	20	20	40	20	20	40

¹⁰ Schwierigkeitsindex gibt die Lösungsrate an, hier in einen Prozentwert umgerechnet.

4.3 Kodieren der Bildschirmaufnahmen

Das Bearbeitungsverhalten der Teilnehmenden wurde in den Bildschirmaufnahmen anhand folgender Daten dokumentiert¹¹: Cursorbewegung im Text; Entwicklung und Änderungen einer Antwort (z.B. Korrekturen, Löschen) und temporale Daten (Verweildauer in einer Lücke, Bearbeitungszeiten).

Das Kodierschema zur Erfassung des Bearbeitungsverhaltens wurde iterativ auf Grundlage bestehender Theorien und vorliegender Daten entwickelt. Die Hauptkategorien „Lineare Bearbeitung“, „Überspringen“ u.Ä. wurden *a priori* festgelegt (Germann/Grotjahn 1994; Zimmermann 2019). Nach der ersten Anwendung auf die Daten (fünf Aufnahmen) durch drei Autorinnen und eine weitere Kodierperson wurde das vorläufige Schema in einem Treffen diskutiert und verfeinert. So entstand z.B. die Unterkategorie „Direkte Revision“ in der Kategorie „Lineare Bearbeitung“. Verhaltensweisen wie „Verschlimmbesserungen“ (*disimprovement*) und längeres Verweilen in einer Lücke vor der Eingabe einer Antwort wurden ebenso als Beobachtungskategorien definiert.

Anschließend kodierten zwei Personen die restlichen der 64 Aufnahmen. Die beiden Kodierenden machten sich im Vorfeld mit dem Kodierschema vertraut und trafen sich während des Verfahrens mehrfach, um eine einheitliche Kodierung zu gewährleisten. Tabelle 3 veranschaulicht das Kodierschema.

¹¹ Eine Beispielaufnahme ist auf der Projektseite verfügbar (Drackert n.d.). Die Grenze zwischen „schneller“ und „langsamer“ Bearbeitung wurde empirisch anhand der Verteilung der Bearbeitungszeiten bestimmt: Werte oberhalb des 75. Perzentils wurden als „langsam“ klassifiziert – ein Vorgehen, das in der empirischen Forschung bei fehlenden theoretisch fundierten Schwellenwerten häufig Anwendung findet (vgl. MacCallum et al. 2002: 19).

Tab. 3: Kodierschema für Bildschirmaufnahmen

Code	Bearbeitungsvorgang
<i>Correct in 1st linear attempt</i>	Richtige Antwort im ersten linearen Bearbeitungsdurchgang
<i>Wrong in 1st linear attempt</i>	Falsche Antwort im ersten linearen Bearbeitungsdurchgang
<i>Direct revision</i>	Überarbeitung einer Antwort unmittelbar nach der Eingabe
<i>Skipped</i>	Überspringen (keine Bearbeitung) einer Lücke im ersten Bearbeitungsdurchgang
<i>Backtracking: delayed revision</i>	Zurückkehren zu und nachträgliche Revision einer bereits ausgefüllten Lücke
<i>Backtracking to skip-ped gaps</i>	Zurückkehren zu und nachträgliches Ausfüllen einer noch nicht ausgefüllten Lücke
<i>Not attempted</i>	Lücke, die aufgrund abgelaufener Testzeit nicht ausgefüllt wurde
<i>Disimprovement</i>	Korrektur einer Antwort (unmittelbar oder nachträglich) zu einer falschen Antwort

Zusätzlich zu den in Tabelle 3 dargestellten Bearbeitungsvorgängen wurden temporale Daten erfasst. Ermittelt wurde die Verweilzeit in einer Lücke gemessen als Zeitspanne zwischen dem Aufleuchten des Lückenkästchens und der Eingabe der Antwort, sofern diese mehr als 3 Sekunden betrug. Diese Daten wurden in die Analysen einbezogen, um zwischen zügigen und verzögerten Verarbeitungsprozessen zu unterscheiden. Verzögerte Verarbeitung wurde weiter unterteilt in „schnelle“ (bis 6 Sekunden) und „langsame“ (ab 7 Sekunden) Bearbeitungsvorgänge¹².

4.4 Analysen

Die Häufigkeiten der verschiedenen Kategorien des Bearbeitungsverhaltens wurden berechnet und in *Excel* erfasst. Die statistischen Auswertungen (Durchschnittswerte, Standardabweichungen, t-Tests) erfolgten in *Excel* oder *R*.

¹² Die Grenze zwischen „schneller“ und „langsamer“ Bearbeitung wurde empirisch anhand der Verteilung der Bearbeitungszeiten bestimmt: Werte oberhalb des 75. Perzentils wurden als „langsam“ klassifiziert – ein Vorgehen, das in der empirischen Forschung bei fehlenden theoretisch fundierten Schwellenwerten häufig Anwendung findet (vgl. MacCallum et al. 2002: 19).

5 Ergebnisse

5.1 Unterschiede in der Bearbeitung allgemein

Tabelle 4 gibt einen Überblick über die unterschiedlichen Bearbeitungsvorgänge nach Testversion, die als Referenzwerte dienen.

Tab. 4: Bearbeitungsvorgänge nach Testversion

		kanonischer C-Test	zeitreduzierter C-Test
Lineare Bearbeitung		583	584
Nicht-lineare Bearbeitung	Überspringen	113	93
	Nachträgliche Revision	25	14
	Nachträgliches Ausfüllen	119	48
Bearbeitungsvorgänge gesamt		840	739

Die lineare Bearbeitung dominiert in beiden Testformen mit knapp 70 – 80 % aller Bearbeitungsvorgänge (Abb. 1). Im kanonischen C-Test tritt nicht-lineare Bearbeitung häufiger als im zeitreduzierten C-Test auf: Der Unterschied beträgt rund 9 % und ist laut t-Test für abhängige Stichproben signifikant mit einem großen Effekt: $t(15) = 2,67, p = ,017, d = 0,67$.

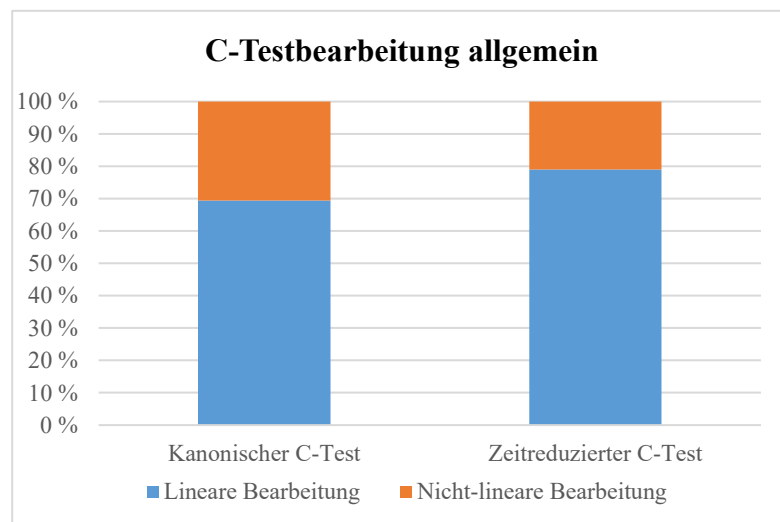


Abb. 1: C-Testbearbeitung allgemein

Beim Vergleich der Kategorien nicht-linearer Bearbeitung zwischen den Testformen (Abb. 2) zeigt sich, dass Teilnehmende in der kanonischen Form häufiger Lücken überspringen und zu ihnen zurückkehren. Der Unterschied im „Nachträgliches Ausfüllen“ ist statistisch

signifikant mit einer großen Effektstärke: $t(15) = 3,51, p < ,005, d = 0,88$. Hingegen unterscheiden sich die Testformen wenig in der Häufigkeit nachträglicher Revision bereits ausgefüllter Lücken, die insgesamt selten auftritt. Im kanonischen C-Test zeigt sich das nachträgliche Ausfüllen von Lücken häufiger als das Überspringen, da übersprungene Lücken beim *Backtracking* mehr als einmal wieder bearbeitet werden.

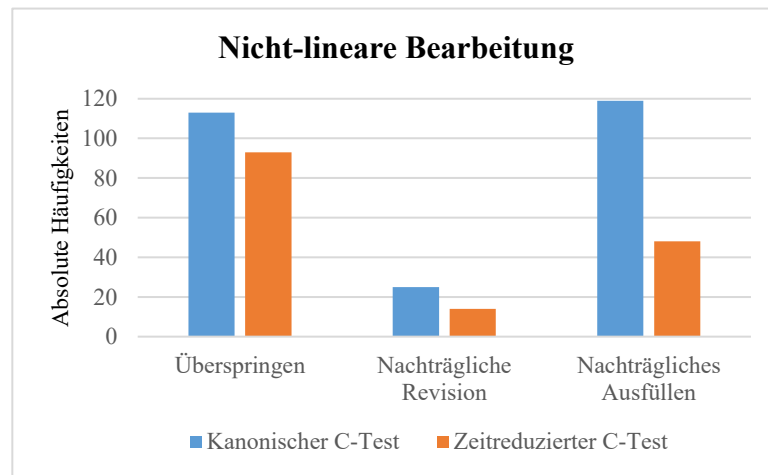


Abb. 2: Nicht-lineare Bearbeitung nach Kategorien

5.2 Antwortverhalten beim Bearbeiten von Mikrolücken

Für diese Analyse wurden nur Bearbeitungsvorgänge gezählt, die zu einer korrekten Antwort führten. In beiden Testformen werden Mikrolücken überwiegend linear ausgefüllt (über 95 %, Abb. 3). Eine zügige lineare Bearbeitung mit korrektem Ergebnis tritt im zeitreduzierten C-Test etwas häufiger auf (89 % vs. 83 % der Fälle), wobei der Unterschied nicht signifikant ist mit einer mittleren Effektgröße: $t(15) = 1,91, p = ,080, d = 0,48$.

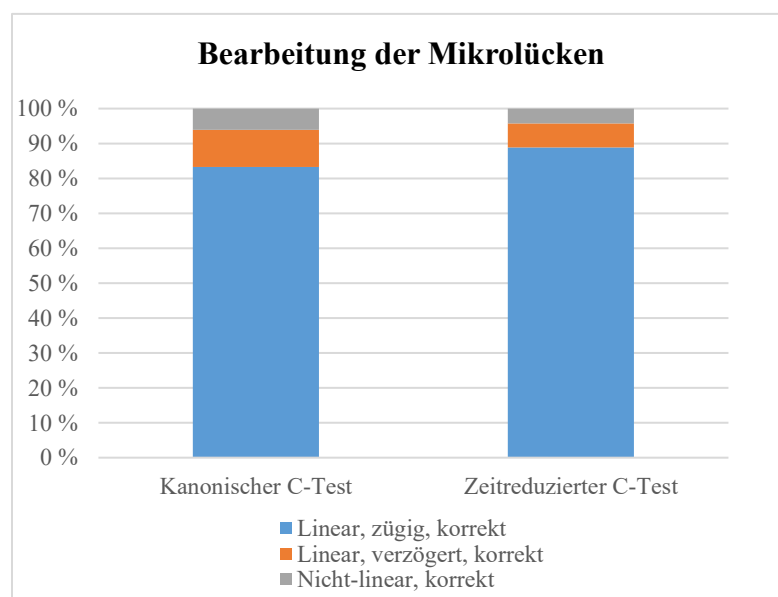


Abb. 3: Bearbeitung der Mikrolücken

5.3 Antwortverhalten beim Bearbeiten von Makrolücken

Beim Bearbeiten von Lücken, die einen breiteren Kontext für die Lösung erfordern, zeigten sich keine signifikanten Unterschiede im linearen Vorgehen zwischen den C-Testformen, obwohl ein kleiner bis mittlerer Effekt erkennbar war ($t(15) = -1,26, p = ,228, d = 0,32$). In der Häufigkeit der nicht-linearen Bearbeitung (vgl. Abb. 4) hingegen zeigte sich nicht nur ein signifikanter Unterschied zugunsten des kanonischen C-Tests (26 % vs. 10 %), sondern auch eine mittlere bis starke Effektgröße ($t(15) = 2,42, p = ,029, d = 0,61$). In die Analyse gingen nur korrekt ausgefüllte Lücken ein.

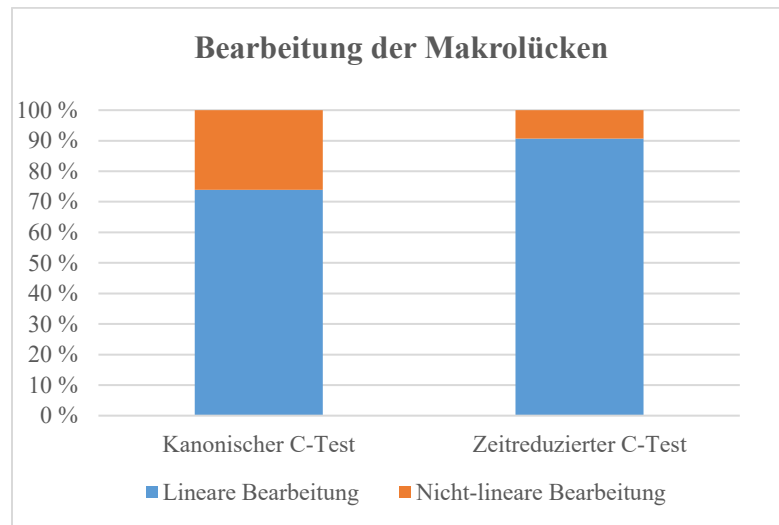


Abb. 4: Bearbeitung der Makrolücken

Rücksprünge zu Makrolücken treten im kanonischen C-Test etwa 2,5-mal so häufig auf wie im zeitreduzierten Format (74 vs. 29 Bearbeitungsvorgänge; vgl. Abb. 5). Dieser Unterschied ist statistisch signifikant mit großem Effekt: $t(15) = 4,58, p < ,001, d = 1,10$. Solche Rücksprünge erfolgen entweder zur Revision einer bereits bearbeiteten Lücke oder zum nachträglichen Ausfüllen einer bislang unbeantworteten Lücke und können als erfolgreich (richtige Lösung) oder erfolglos (falsche oder keine Lösung) klassifiziert werden. Unabhängig vom Testformat liegt die Erfolgsquote dieser nachträglichen Bearbeitungsvorgänge deutlich unterhalb der 50 %-Marke – bei 43,2 % im kanonischen und 41,4 % im zeitreduzierten C-Test.

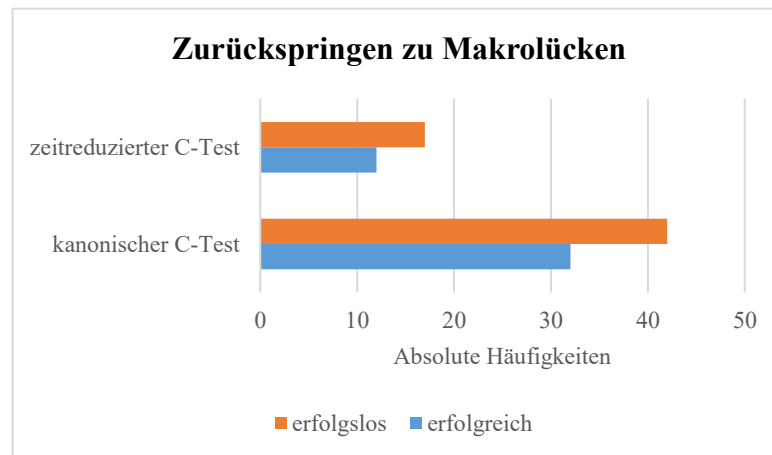


Abb. 5: Zurückspringen zu Makrolücken

Wird die rückläufige Bearbeitung der Makrolücken nach Verweildauer in „schnell“ (bis zu 6 Sekunden) und „langsam“ (ab 7 Sekunden) unterteilt, zeigt sich eine tendenzielle Präferenz für die langsame Bearbeitung im kanonischen C-Test (45 vs. 29 Vorgänge), $t(15) = 2,03, p = ,060, d = 0,51$. Dabei sinkt die Erfolgsrate leicht von 48 % bei schneller auf 40 % bei langsamer Bearbeitung (Abb. 6), ohne signifikanten Unterschied ($t(15) = 0,94, p = ,360, d = 0,23$).

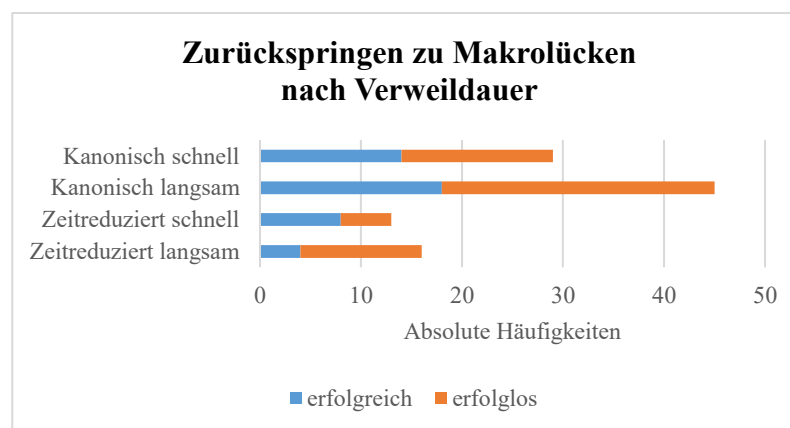


Abb. 6: Zurückspringen zu Makrolücken nach Verweildauer

Bei Zeitreduktion treten „schnelle“ und „langsame“ Bearbeitungen der Makrolücken in etwa gleich häufig auf (13 vs. 16 Vorgänge), $t(15) = 0,53, p = ,610, d = 0,13$. Die Erfolgsrate sinkt mit zunehmender Verweildauer deutlich – von 62 % bei kürzerer auf nur 25 % bei längerer Verweildauer in der Lücke (Abb. 6). Dieser Unterschied ist statistisch jedoch nicht signifikant, $t(15) = -1,07, p = ,300, d = -0,27$.

6 Diskussion

6.1 Interpretation der Ergebnisse

Deutsche C-Tests mit und ohne Zeitbegrenzung werden von relativ fortgeschrittenen DaF-Lernenden überwiegend linear rekonstruiert: Rund 70 – 80 % aller Bearbeitungsvorgänge erfolgen linear, bei Mikrolücken sogar über 95%. Lineares Vorgehen ist somit die dominante Bearbeitungsform, die unabhängig von den Zeitbedingungen stabil bleibt. Dieses Ergebnis widerspricht den Befunden von Zimmermann (2019) und spricht für die Konstruktvalidität des C-Testformats: Offenbar wird das Format von den Testteilnehmenden mit fortgeschrittenen Sprachkenntnissen primär als kohärenter Text und nicht als Sprachrätsel wahrgenommen und bearbeitet. Die Dominanz linearer Bearbeitung dürfte auch mit den Texteigenschaften zusammenhängen, da etwa zwei Drittel der Lücken in den analysierten Texten Mikrolücken darstellen, die nur minimalen Kontext erfordern und lokal rekonstruierbar sind.

Gleichzeitig zeigen die Daten, dass Zeitbedingungen bestimmte Bearbeitungsweisen im C-Test beeinflussen, was auf Unterschiede in den zugrunde liegenden kognitiven Verarbeitungsprozessen hindeuten kann. Im zeitreduzierten C-Test werden Lücken seltener übersprungen oder nachträglich bearbeitet, was wiederum die Hypothese von Zimmermann (2019) stützt. Bereits bearbeitete Lücken werden dagegen nur selten revidiert, selbst wenn mehr Zeit zur Verfügung steht, was möglicherweise darauf zurückzuführen ist, dass Erstlösungen häufig auf automatisierten, prozeduralen Prozessen beruhen und der zusätzliche kognitive Aufwand die Wahrscheinlichkeit einer Revision reduziert. Insgesamt scheint die längere verfügbare Bearbeitungszeit im kanonischen Format flexiblere Bearbeitungsmuster zu ermöglichen, die exploratives Ergänzen offener Lücken begünstigen und dabei einen vermeintlichen Zugriff auf deklaratives Wissen nahelegen.

Bei Mikrolücken, die in beiden Formaten fast ausschließlich linear bearbeitet werden, scheint die Verarbeitung unabhängig von Zeitdruck stark automatisiert abzulaufen. Der geringe, nicht signifikante Unterschied zwischen kanonischer und zeitreduzierter Form legt nahe, dass diese lokal gebundenen Lücken vor allem routinisierte sprachliche Prozesse aktivieren, die gegenüber Zeitrestriktionen vergleichsweise robust sind.

Ein anderes Bild ergibt sich bei Makrolücken, die einen breiteren Kontext für die Lösung erfordern. Zwar dominiert auch hier die lineare Bearbeitung (75 % im kanonischen vs. knapp 90 % im zeitreduzierten C-Test)¹³, doch zeigen sich im kanonischen Format signifikant häufiger nicht lineare Bearbeitungen, bei denen Teilnehmende Lücken überspringen

¹³ Dabei ist zu berücksichtigen, dass nicht alle Makrolücken rückläufige Bearbeitung erfordern. Nach unserer Klassifikation benötigt etwa die Hälfte dieser Lücken den nachfolgenden Kontext zur Lösung, während die andere Hälfte aus dem vorangehenden Textteil erschließbar ist und vermutlich auch linear – gegebenenfalls mit Verzögerung – gelöst werden kann.

und später nacharbeiten. Dies könnte darauf hindeuten, dass das kanonische Format eher bewusst-strategisches Vorgehen bei diesen anspruchsvolleren Lücken begünstigt, das möglicherweise stärker auf deklarativem Wissen basiert.

Die vorwiegend lineare Bearbeitungsweise der Makrolücken im zeitreduzierten C-Test-Format könnte mehrere Ursachen haben. Denkbar ist, dass Zeitdruck voreilige Eingaben begünstigt, bevor der notwendige Kontext berücksichtigt wird (*early closure*), oder Raten als konstruktirrelevante Bearbeitungsstrategie provoziert. Dagegen spricht jedoch die vernachlässigbar geringe Differenz in der Gesamtzahl korrekt gelöster Makrolücken: 121 im kanonischen vs. 119 im zeitreduzierten C-Test. Naheliegender erscheint daher, dass Zeitrestriktion zu einem schnelleren, lineareren Bearbeitungsstil zwingt, der stärker automatisierte Prozesse und prozeduralisiertes Wissen aktiviert und dadurch eine schnelle antizipatorische Verarbeitung ermöglicht.

Obwohl Teilnehmende im kanonischen C-Test häufiger zu leeren oder bereits ausgefüllten Makrolücken zurückkehren, bleibt die Erfolgsrate beim *Backtracking* in beiden Testformen vergleichbar niedrig (ca. 43 % im kanonischen vs. ca. 41 % im zeitreduzierten C-Test). Zudem verweilen die Proband*innen im kanonischen C-Tests tendenziell länger in den Makrolücken, wenngleich der Unterschied nicht signifikant ist. Eine längere Verweildauer führt jedoch nicht zu mehr richtigen Antworten, und zwar in den beiden Testformen. Im Gegenteil: Es zeigt sich ein negativer Zusammenhang (längere Verweildauer – geringere Erfolgsrate), besonders unter erhöhtem Zeitdruck. Im zeitreduzierten C-Test führte eine Verweildauer von weniger als 7 Sekunden in 62 % der Fälle zu einer richtigen Antwort, während beim längeren Verweilen die betroffene Makrolücke in 75 % der Fälle leer oder falsch ausgefüllt blieb. Zwar erreichte auch dieser Unterschied keine statistische Signifikanz, dennoch könnte der Befund darauf hindeuten, dass längeres Verweilen in einer Lücke weniger mit gründlicher Bearbeitung als mit Lösungsschwierigkeiten zusammenhängt. Dies könnte wiederum darauf hinweisen, dass das kanonische Format eher mühevoll, auf deklarativem Wissen basierende Prozesse fördert, während Zeitdruck stärker automatisierte Prozesse begünstigt, ohne dabei den Abruf notwendigen Wissens zu behindern – vorausgesetzt, dieses Wissen ist vorhanden und automatisiert.

Eine exemplarische Analyse der Lernendenantworten illustriert dies. Für die korrekte Lösung der Makrolücke „a[uf]“ im zeitreduzierten C-Test-Text 4 muss die Phrase „d ___ Gehirn a ___ die akust ___ Signale vorbe ___“ bis zum Verb „vorbereiten“ rekonstruiert werden, welches die Präposition „auf“ regiert. Auffällig ist jedoch, dass die Hälfte der Teilnehmenden „an“ einsetzte, obwohl fünf von diesen acht Personen „vorbereiten“ korrekt rekonstruierten. In diesem Fall verhinderte eher fehlendes grammatikalisches Wissen die richtige Lösung als die für das Erfassen des notwendigen Kontexts mangelnde Zeit. Zudem können Interdependenzen der Lücken und fehlende oder falsche Antworten in den benachbarten Lücken die Lösung erschweren (vgl. Beinborn et al. 2014).

Unsere Ergebnisse stützen auch die Beobachtung von Stemmer (1991), dass Verarbeitung des breiteren Kontexts vor allem als problemlösende Strategie eingesetzt wird. *Backtracking* (insbesondere mit längerem Verweilen) tritt auch bei schwierigen Lücken auf, die eigentlich nur die Berücksichtigung des Mikrokontexts erfordern – etwa wenn das gesuchte Wort vermutlich nicht direkt aus dem mentalen Lexikon abrufbar (z.B. Vielf[*alt*], tiri[*scher*]) oder das notwendige grammatikalische Wissen nicht prozeduralisiert bzw. nicht vorhanden war (wie im Beispiel oben mit „auf“).

Schließlich verdeutlicht die Kategorie der „Verschlimmbesserungen“ (korrekt gelöste Lücken, die nachträglich falsch korrigiert wurden), dass zusätzliche Bearbeitungszeit nicht automatisch zu besseren Ergebnissen führt. Im kanonischen Test traten sieben solcher Fälle auf, im zeitreduzierten nur einer – auch wenn insgesamt im Durchschnitt mehr richtige Antworten ohne Zeitdruck produziert wurden.

6.2 Einschränkungen und Ausblick

Unsere Studie analysierte das Bearbeitungsverhalten von 16 Proband*innen anhand von vier deutschsprachigen C-Test-Texten. Replikationen mit anderen Texten, mehr Teilnehmenden unterschiedlicher Kompetenz und in weiteren Sprachen wären sinnvoll. Zudem könnten weitere Datenanalysen erfolgen oder ergänzende Methoden wie retrospektive Interviews eingesetzt werden.

Weitere Beobachtungen und Daten würden eine präzisere Differenzierung in der Lückenklassifikation ermöglichen. Besonders relevant wäre eine separate Analyse von ausgefüllten Makrolücken, welche die Berücksichtigung des vorangehenden und nachfolgenden Kontexts erfordern, sowie die Einführung von Unterkategorien für Mikro- und insbesondere für Makrolücken (z. B. auf Satzteil-, Satz- oder Textebene). Darüber hinaus wäre es sinnvoll, zwischen Lücken zu unterscheiden, die gezielte Aufmerksamkeit auf sprachliche Form und deren Manipulation erfordern, und solchen, die es nicht tun.

Auch die Sprachkompetenz sollte in zukünftigen Studien stärker berücksichtigt werden. Unsere Annahmen beruhen auf der Vorstellung einer Testperson mit einer vollständig ausgebildeten Kompetenz in Deutsch als Fremdsprache, wie sie beispielsweise bei gebildeten Muttersprachler*innen vorhanden ist. Allerdings entsprach unsere Stichprobe – trotz der Selbsteinschätzung auf B2-C1-Niveau – nur teilweise diesem Maßstab (vgl. WEIT-Ergebnisse in Kap. 4.1). Eine Analyse der C-Testbearbeitung durch gebildete Muttersprachler*innen könnte als wertvolle Referenz dienen. Zudem wäre ein Vergleich des Bearbeitungsverhaltens von sehr fortgeschrittenen Lernenden (C1-C2) und wenig kompetenten Lernenden (A2-B1) sinnvoll.

Weitere Analysen könnten das Korrekturverhalten untersuchen, etwa wie oft und wann (direkte vs. nachträgliche Revision) sowie was (Inhalt vs. Form) bevorzugt korrigiert wird, und welche Lückentypen dabei betroffen sind. Zudem ist zu klären, ob sich die Bearbeitung

von bereits ausgefüllten Lücken und noch nicht ausgefüllten Lücken chronologisch unterscheidet. So könnte es beispielsweise sein, dass nachträgliche Korrekturen häufiger im ersten Bearbeitungsdurchgang erfolgen, während das Ausfüllen leerer Lücken eher nach Abschluss des ersten Durchgangs stattfindet, wenn mehr Kontext zur Verfügung steht.

Diese Analysen des Korrekturverhaltens sollten mit systematischen Antwortanalysen kombiniert werden, die zusätzlichen Aufschluss über die eingesetzten und fehlenden Wissensbestände geben können. Dadurch ließe sich genauer untersuchen, welche Aspekte sprachlicher Fähigkeiten und Kenntnisse der C-Test im Allgemeinen und unter Zeitdruck erfasst.

Zukünftige Studien sollten ein theoretisches Modell anstreben, das das Bearbeitungsverhalten systematisch mit den zugrunde liegenden Verarbeitungsprozessen und den dabei aktivierten linguistischen Wissensbeständen verknüpft und eine feinere Beschreibung der Verarbeitungsprozesse ermöglicht, anstelle der bisherigen dichotomen Unterscheidung in automatisierte vs. nicht-automatisierte Bearbeitung. Dazu sind neben der Beobachtung des Bearbeitungsverhaltens auch introspektive Verfahren wie *stimulated recall* erforderlich. Der Einsatz solcher Verfahren würde zudem ermöglichen, den Einfluss personenspezifischer Merkmale wie Bearbeitungspräferenzen auf die Verarbeitungsprozesse zu untersuchen.

6.3 Zusammenfassung

Die Studie belegt die prozessbezogene Validität des C-Testformats, indem sie eine vorwiegend lineare Rekonstruktion der deutschen Texte durch eine kleine Stichprobe relativ fortgeschrittener Lernender als kohärente sprachliche Einheiten und nicht als Sprachrätsel nachweist. Abweichungen von der linearen Bearbeitung treten vor allem dann auf, wenn einzelne Lücken nur unter Einbezug eines breiteren Kontexts geschlossen werden können oder die lokale, serielle Wiederherstellung scheitert.

Je nach Zeitbedingung rücken bestimmte Verarbeitungsprozesse und Wissensbestände unterschiedlich stark in den Vordergrund. Im kanonischen Format treten signifikant häufiger nicht-lineare Bearbeitungen auf, was auf inferentielles, wissensgeleitetes Vorgehen verweist – besonders, aber nicht ausschließlich bei Makrolücken. Während hier auch weniger automatisierte, deklarative Wissensbestände stärker genutzt werden, sind diese unter Zeitdruck nur eingeschränkt zugänglich, sodass automatisierte Prozesse und prozeduralisiertes Wissen eine größere Rolle übernehmen. Der zeitreduzierte und der kanonische C-Test erfassen damit ein stark überlappendes Kernkonstrukt, aktivieren jedoch automatisierte bzw. deklarative Verarbeitungsprozesse in unterschiedlichem Ausmaß. Die höheren Punktwerte im kanonischen C-Test erklären sich daher weniger durch den bloßen Zeitvorteil, sondern durch den stärkeren Einbezug deklarativen Sprachwissens, das unter Zeitrestriktion nur

begrenzt zugänglich ist. Dies stützt den Befund von Timukova et al. (im Druck), dass zeitlimitierte C-Tests einen etwas größeren Anteil automatisierter Sprachfähigkeiten im Konstrukt abbilden als das kanonische Format.

Diese Erkenntnisse haben auch praktische Implikationen. Der Einsatz zeitreduzierter C-Tests zur Einstufung und in der Forschung kann nicht nur Zeit sparen, sondern auch Ergebnisse liefern, die das automatisierte und prozeduralisierte Sprachwissen der Teilnehmenden besser abbilden und damit genauere Aussagen über deren abrufbereite Sprachkompetenz ermöglichen.

Literatur

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014): *Standards for educational and psychological testing*. American Educational Research Association.
- Aguado, Karin; Grotjahn, Rüdiger & Schlak, Torsten (2005): Erwerbssalter und Sprachlernerfolg. Theoretische und methodologische Grundlagen eines empirischen Forschungsprojekts. *Zeitschrift für Fremdsprachenforschung* 16: 2, 275–293.
- Babaii, Esmat & Ansary, Hasan (2001): The C-test: A valid operationalization of reduced redundancy principle? *System* 29: 2, 209–219.
- Babaii, Esmat & Fatahi-Majd, Mosayeb (2014): Failed restorations in the C-test: Types, sources, and implications for C-test processing. In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*. Frankfurt am Main: Lang, 263–276.
- Beinborn, Lisa; Zesch, Torsten & Gurevych, Iryna (2014): Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics* 2, 517–529. <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/414/88>
- Chifligarov, Mihail; Laâguidi, Jammila; Schellenberg, Max; Dill, Alexander; Timukova, Anna; Drackert, Anastasia & Laarmann-Quante, Ronja (2025): Automated Scoring of a German Written Elicited Imitation Test. In: *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, 237–247, Vienna, Austria. Association for Computational Linguistics. <https://aclanthology.org/2025.bea-1.18/>
- Drackert, Anastasia (2016): *Validating language proficiency assessments in second language acquisition research*. Peter Lang Verlag.
- Drackert, Anastasia & Timukova, Anna (2020): What does the analysis of C-test gaps tell us about the construct of a C-test? A comparison of foreign and heritage language learners' performance. *Language Testing* 37: 1, 107–132. <https://doi.org/10.1177/0265532219861042>

- Drackert, Anastasia (n.d.): *Time to fill the gaps: A mixed-methods study into the role of the time variable in the construct of computerized C-Tests in three languages*. [DFG project 462766474, 2022 - 2025]. Gesellschaft für Akademische Studienvorbereitung und Testentwicklung (g.a.s.t.). <https://www.gast.de/de/forschung-entwicklung/forschung/forschungsprojekte/c-test-zeitfaktor/english-version>
- Eckes, Thomas & Grotjahn, Rüdiger (2006): A closer look at the construct validity of C-tests. *Language Testing* 23: 3, 290–325.
- Ellis, Rod (2003): *Task-Based Language Learning and Teaching*. Oxford University Press.
- Embretson, Susan. E. (Whitely) (1983): Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin* 93: 1, 179–197. <https://doi.org/10.1037/0033-2909.93.1.179>
- Europarat (2020): *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Begleitband mit neuen Deskriptoren*. Council of Europe Publishing.
- Germann, Ulrich & Grotjahn, Rüdiger (1994): Das Lösen von C-Tests auf dem Computer: Eine Pilotuntersuchung zu den Bearbeitungsprozessen. In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* Bd. 2. Bochum: Brockmeyer, 279–304.
- Grabe, William & Yamashita, Junko (2022): *Reading in a Second Language: Moving from Theory to Practice*. [2. Aufl.]. Cambridge: Cambridge University Press. (Cambridge Applied Linguistics)
- 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/The C-Test: Contributions from current research*. Frankfurt am Main: Lang, 265–296.
- Grotjahn, Rüdiger & Schiller, Cordula S. (2014): Zur Rolle des Makrokontexts bei der Bearbeitung spanischer C-Test-Texte: Fehleranalysen ausgewählter Lückenwörter. In: Grotjahn, Rüdiger (Hrsg.): *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*. Frankfurt am Main: Lang, 277–289.
- Grotjahn, Rüdiger & Drackert, Anastasia (2022): *The electronic C-Test bibliography: version October 2022*. <http://www.c-test.de> & <https://www.gast.de/de/forschung-entwicklung/publikationen/veroeffentlichungen-von-gast> (19.10.2024)
- Harsch, Claudia & Schröder, Konrad (2007): Textrekonstruktion: C-Test. In: Beck, Bärbel & Klieme, Eckhard (Hrsg.): *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim u.a.: Beltz, 212–225. <https://doi.org/10.25656/01:3140>
- 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* 4. Bochum: AKS-Verlag, 53–66.

- Hubley, Anita M. & Zumbo, Bruno D. (2017): Response Processes in the Context of Validity: Setting the Stage. In: Zumbo, Bruno D. & Hubley, Anita M. (Hrsg.): *Understanding and Investigating Response Processes in Validation Research*. 1–12. https://link.springer.com/chapter/10.1007/978-3-319-56129-5_1
- Jang, Eunice (2017): Cognitive Aspects of Language Assessment. In: Shohamy, Elana; Or, Iair & May, Stephen (Hrsg.): *Language Testing and Assessment*. Cham: Springer, 163–179. https://doi.org/10.1007/978-3-319-02326-7_11-1
- Kane, Michael T. (2016): Explicating validity. *Assessment in Education: Principles, Policy, & Practice* 23: 2, 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kaufmann, Nicole (2016): Die Vorhersage der Schwierigkeit deutscher C-Test-Texte: Untersuchungen am Beispiel des onDaF. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 21: 2, 111–126. <http://tujournals.ulb.tu-darmstadt.de/index.php/zif/> (25.03.2025).
- Khalifa, Hanan & Weir, Cyril J. (2009): *Examining Reading: Research and Practice in Assessing Second Language Reading* (Studies in Language Testing 29). Cambridge: Cambridge University Press.
- Klein-Braley, Christine (1994): *Language testing with the C-Test. A linguistic and statistical investigation into the strategies used by C-Test takers, and the prediction of C-Test difficulty*. Habilitationsschrift, Universität-Gesamthochschule Duisburg.
- MacCallum, Robert C.; Zhang, Shaobo; Preacher, Kristopher J. & Rucker, Derek D. (2002): On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1), 19–40. <https://doi.org/10.1037/1082-989X.7.1.19>
- McCray, Gareth & Brunfaut, Tineke (2018): Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing* 35: 1, 51–73. <https://doi.org/10.1177/0265532216677105>
- McKay, Todd; Teimouri, Yasser; Sağdıç, Ayşenur; Salen, Bradford; Reagan, Derek & Malone, Margaret E. (2021): The cagey C-test construct: Some evidence from a meta-analysis of correlation coefficients. *System* 99. <https://doi.org/10.1016/j.system.2021.102526>
- Messick, Samuel (1989): Validity. In: Linn, Robert L. (Hrsg.): *Educational measurement*. [3. Aufl.]. New York: American Council on Education and Macmillan, 13–103.
- Morgan-Short, Kara & Ullman, Michael T. (2022): Declarative and Procedural Memory in Second Language Learning. Psycholinguistic considerations. In: Godfroid, Aline & Hopp, Holger (Hrsg.): *The Routledge Handbook of Second Language Acquisition and Psycholinguistics*. New York: Routledge, 322–334. <https://doi.org/10.4324/9781003018872-30>
- 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. <http://www.c-test.de/deutsch/index.php?lang=de§ion=originalia>
- Sigott, Günther (2004): *Towards identifying the C-Test construct*. Habilitationsschrift, Frankfurt am Main: Lang.

- Stemmer, Brigitte (1991): *What's on a C-test taker's mind: Mental processes in C-test taking*. Bochum: Brockmeyer.
- Timukova, Anna; Möller, Franziska & Drackert, Anastasia (im Druck): Different, or simply more difficult? The role of reduced time in the construct of computerised C-Tests in English and German. *Studies in Language Assessment*.
- Ullman, Michael T. (2020): The Declarative/Procedural Model: A Neurobiologically-Motivated Theory of First and Second Language. In: VanPatten, Bill; Keating, Gregory D. & Wulff, Stefanie (Hrsg.): *Theories in Second Language Acquisition*. [3. Aufl.]. New York: Routledge, 128–161.
- Wondershare. (n.d.): *DemoAir* [Computer software]. Wondershare Technology Group Co., Ltd. <https://demoair.wondershare.com/>
- Wockenfuß, Verena (2008): *Diagnostik von Sprache und Intelligenz bei Jugendlichen und jungen Erwachsenen: Eine empirische Untersuchung*. Dissertation, Universität Duisburg-Essen.
- Zimmermann, Kerstin (2019): *Keine Zeit für den C-Test? Eine empirische Untersuchung zum Einfluss einer Geschwindigkeitskomponente auf das Konstrukt des C-Tests*. Berlin: Universitätsverlag der TU Berlin. <http://dx.doi.org/10.14279/depositonce-8288>.

Kurzbio:

Anna Timukova ist wissenschaftliche Mitarbeiterin am Sprachenzentrum der Ruhr-Universität Bochum und verantwortet dort die Qualitätssicherung verschiedener Testverfahren – von Einstufungstests bis zu Kursabschlussprüfungen. Ihre Forschungsinteressen umfassen die Konstruktvalidierung und kognitive Validität von Testverfahren, mit einem Schwerpunkt auf C-Tests. Darüber hinaus befasst sie sich mit dem Testen bildungs- und akademischer Sprachkompetenzen, mit Ansätzen des *Elicited-Imitation*-Testens sowie mit der automatisierten Auswertung von Testaufgaben.

Oleksandra Yazdanfar ist MA-Studentin der Sprachverarbeitung an der Ruhr-Universität Bochum. Im Rahmen ihres Studiums untersuchte sie die Organisation des lexikalischen Wissens bei bilingualen Personen sowie den Zusammenhang zwischen dem *Written Elicited Imitation Test* und Gedächtnisprozessen. In ihrer Bachelorarbeit beschäftigte sie sich mit NLP-Ansätzen zur Klassifikation des Informationsstatus im Deutschen.

Anastasia Drackert ist wissenschaftliche Direktorin bei der Gesellschaft für Akademische Studienvorbereitung und Testentwicklung (g.a.s.t. e.V.) und verantwortet dort die Forschung sowie die Weiterentwicklung der Testverfahren und Lernangebote. Gleichzeitig ist sie Professorin für Sprachtesten und Digitales Lernen an der Ruhr-Universität Bochum. In zahlreichen empirischen Studien untersuchte sie unter anderem den Einsatz von KI beim Testen und Lernen von Fremdsprachen, die Leistungsbeurteilungskompetenz (*Assessment Literacy*) von Fremdsprachenlehrkräften, den Einsatz von Selbstevaluation im Fremdsprachenunterricht sowie das Konstrukt des C-Tests.

Anschrift:

Anna Timukova
Sprachenzentrum der Ruhr-Universität Bochum
Universitätsstraße 150
D-44801 Bochum
anna.timukova@rub.de

Oleksandra Yazdanfar
Ruhr-Universität Bochum
Universitätsstraße 150
D-44801 Bochum
oleksandra.yazdanfar@rub.de

Anastasia Drackert
Gesellschaft für Akademische Studienvorbereitung und Entwicklung (g.a.s.t.)
Universitätsstraße 134
D- 44799 Bochum
drackert@gast.de