Different, or simply more difficult? The role of reduced time in the construct of computerised C-Tests in English and German

Anna Timukova[1], Franziska Möller[1] and Anastasia Drackert[1,2]

[1] g.a.s.t. (*Society for Academic Study Preparation and Test Development*), Germany

[2] Ruhr University of Bochum, Germany

Speeded C-Tests are a time-constrained variation of the traditional C-Test, requiring test takers to complete partial word gaps in increasingly complex short texts within about two minutes per text, compared to the standard five. While the original C-Test is a well-established measure of global language proficiency, the speeded version – promising greater reliability and cost-effectiveness – remains underexplored. Reducing time typically increases test difficulty due to the speed-ability trade-off but may also shift the construct by altering the skills assessed. The study examined university students' performance on computerised C-Tests in L2 English (N = 237) and German (N = 191) under two timing conditions. Time constraints significantly lowered scores in both languages, although the effect was small. Less proficient learners were more affected, but the difference in point loss was not statistically significant. An interaction between time and text difficulty was observed in German, but not in English. The speeded C-Test showed slightly stronger correlations with oral proficiency and marginally better predictive power in both samples, particularly among more advanced learners. In sum, the speeded C-Test is more difficult and appears to emphasise automatised, oral-like language skills, with the construct being sensitive to proficiency and (in German) text difficulty.

Keywords: global language proficiency, placement test, speeded testing, computerised language testing

Email address for correspondence: anna.timukova@rub.de

© The Author(s) 2026. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits the user to copy, distribute, and transmit the work provided that the original authors and source are credited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

For over 40 years, C-Tests have been used in language assessment to measure global language proficiency by having test takers complete gaps in a series of short texts of increasing difficulty (Grotjahn & Drackert, 2022). C-Tests are economical in design, administration, and scoring, efficiently ranking examinees by proficiency in the tested language. Due to their high efficiency and practicability, C-Tests in around 30[2] languages are widely used for placement (e.g., Mozgalina & Ryshina-Pankova, 2015), for screening before expensive and time-consuming test batteries (e.g., Eckes, 2010), in educational monitoring (e.g., Harsch & Schröder, 2007) and as a proficiency measure in Second Language Acquisition (SLA) and (language) education research (e.g., Norris, 2018). Recently, the format has also seen increasing use in high-stakes proficiency-testing contexts (Naismith et al., 2025).

C-Tests were initially designed as power tests (Grotjahn, 2010) with a time limit intended to allow all participants to attempt all items. In the canonical (power) version of the test, five minutes are allocated for each short text, which generally provides learners with enough time to work without substantial time pressure, irrespective of their individual working speed and level of proficiency. However, in several studies other time limits were used: from 55 seconds up to four minutes per text (see "Time-reduced C-Tests" below). Time allocation in these studies was grounded in the test use rather than theoretically linked to the construct. Grotjahn et al. (2010) were the first to explicitly address the role of time in the C-Test construct, introducing the so-called speeded C-Test. While this time-constrained version shows promise for greater efficiency and precision, it has not yet been sufficiently researched. The present study addresses this gap by investigating how time constraints are reflected in C-Test performance and how this may relate to the underlying construct.

[2] This estimate is based on documented C-Tests in more than two dozen languages across major repositories and institutional test platforms (e.g., c-test.de, AELRC C-Test Repository, DLTPT project), in addition to several locally developed tests in university placement systems (e.g., Swedish, Polish, Czech, Ukrainian, Persian).

Literature review and research questions

C-Test construct

Previous research has shown that canonical C-Tests are "objective, highly reliable and very economical means for measuring global language proficiency" (Grotjahn, 2013, p. 181). These claims are supported by numerous studies on the C-Test construct, including correlational and factor-analytic analyses, logical task analysis, introspective research into cognitive processes, and analysis of learner responses (see Grotjahn, 2019 for an overview of C-Test validation methods).

C-Tests operationalise the principle of reduced linguistic redundancy, which relies on the ability of proficient language users to decipher damaged messages through anticipatory processing (Klein-Braley, 1997). Reconstructing the gaps involves processing at all levels from letters to whole passages (Sigott, 2004) and requires an integrated application of language knowledge and skills in written form (Hastings, 2002), while also incorporating strategies (Harsch & Schröder, 2007).

Learners must retrieve the partially deleted word from their mental lexicon, identify it based on the context, and reconstruct it in correct form (Drackert & Timukova, 2020). This process activates both declarative and procedural knowledge, as described in the Declarative/Procedural Model (Ullman, 2020). Consciously accessible word meanings, irregular forms, and learned rules are stored in declarative memory, whereas unconscious and automatised processing of rule-governed structures and predictable sequences, such as morphology and syntax, is supported by procedural memory.

Raatz (2002) found strong loadings on verbal intelligence in a factor analysis for German as a native language (L1), attributing C-Test performance to concept-driven processing, where greater language knowledge supports faster, more efficient reconstruction of missing segments. However, the specific cognitive processes involved in solving C-Tests and the balance of lower-level (microstructural) and higher-order (macrostructural) processing remain only partly understood and subject to ongoing debate (cf. Grotjahn & Schiller, 2014). Moreover, the C-Test construct appears malleable, placing different demands on learners depending on their proficiency and

the difficulty of individual texts. Sigott (2004) observed that some gaps require more high-level processing than others, and that more proficient test takers can use sentential context where less proficient ones may need the entire passage. This dependence of the amount of text-level processing on test takers' proficiency and specific text characteristics has led Sigott to describe the construct underlying C-Tests as *fluid*.

In summary, the C-Test construct is a composite of several dimensions of linguistic knowledge and subskills engaging cognitive processing at various levels that enables test takers to navigate reduced redundancy in short written passages – an ability reflecting their global language proficiency. However, the still limited understanding and consensus regarding the relative contribution and interaction of these components is further complicated by the fluid nature of the construct. Adding to this complexity is the fact that, depending on the test's intended purpose, target population, and target language, the canonical C-Test format is often modified (e.g., Mainzer-Murrenhoff & Drackert, 2022). Such modifications may involve construction principles, scoring methods, and the time limit allotted per text, the latter being the focus of the present study.

Time-reduced C-Tests

Initial research using shorter time limits for C-Tests originates from psychological studies on general intelligence, in which C-Tests were used to measure L1 proficiency (e.g., Raatz, 2002). Time reduction in these studies was introduced to make the C-Test more difficult for native language speakers and is reported to have effectively prevented ceiling effects (Wockenfuß, 2008, p. 148).

Other studies used time-reduced C-Tests with foreign/second language (L2) learners, though not explicitly for construct-related purposes. For example, Drackert and Felberg (2019) set a four-minute time limit per text in their validation of a paper-based Russian C-Test, based on earlier observations that most learners were able to demonstrate their true competence within that time, rather than using the full five minutes. Reichert et al. (2010) were the first in L2 research to adjust the time limit for individual texts based on their difficulty, allocating three minutes for easier texts and

four minutes for more difficult ones. Otherwise, no justification for the time reduction was provided.

The latest modification of the time variable has resulted in the so-called speeded C-Test (SC-Test). In their search for an appropriate instrument to measure global language proficiency in studies on age effects in SLA, Aguado et al. (2007) and Grotjahn et al. (2010) found that the canonical C-Test lacked discriminatory power among advanced L2 learners. They therefore introduced variable, text-specific time limits (1:05 – 1:55 minutes) – which increased test difficulty, discrimination, and reliability – but gathered no evidence regarding the construct of the speeded C-Test.

Construct of the speeded C-Test

The increase in difficulty resulting from reduced time limits is a function of the speed-ability trade-off typical of many time-restricted psychological tests: with fewer resources available for producing a response, the likelihood of a correct answer decreases (Goldhammer, 2015). Moreover, as participants must work faster, the measured ability may change not only quantitatively but also qualitatively, resulting, for example, in "different person orderings, reflecting a change in the exact attribute being measured" (Tijmstra & Bolsinova, 2018, p. 5). This raises the question of whether and how time reduction may influence the abilities required to solve C-Tests.

In educational and psychological testing, a key distinction is whether speededness is treated as a nuisance factor or as part of the construct. Alderson (2005, p. 260) argues that processing speed is a key component of language proficiency, especially in reading and listening, as it reflects automatic meaning processing, which supports comprehension and distinguishes highly proficient language users. He also suggests that speeded tests may better capture learners' implicit language knowledge, which echoes Ullman's (2020) predictions that unconscious and fast processing would indicate implicit procedural knowledge. Accordingly, we view unreached gaps in speeded C-Tests as construct-relevant and not requiring adjustment through scoring.

Our current understanding of the construct of time-constrained C-Tests in L2 research is based primarily on two publications: Grotjahn (2010) and Zimmermann (2019). Grotjahn suggested that a canonical C-Test with a five-minute time limit per text

primarily measures the amount of learners' declarative and procedural knowledge, while a speeded C-Test also taps into the automaticity of language skills and the efficiency of information processing. Since a partially different construct underlies the SC-Test, it "would correlate higher with measures of listening comprehension and speaking skills than a canonical C-Test because, similar to a SC-Test, listening and speaking occur under time pressure" (Grotjahn, 2010, p. 289).

Zimmermann (2019) tested this hypothesis by comparing correlations between a canonical and a speeded German C-Test and the listening and speaking sections of the Goethe Certificate exam (B2 level, CEFR; Council of Europe, 2020). The speeded C-Test showed slightly higher correlations with listening (r_s = 0.46) and speaking (r_s = 0.33) than the canonical version (r_s = 0.37 and r_s = 0.32 respectively), although the differences were not statistically significant. When the sample was divided into two proficiency groups, moderate correlations were found for the speeded condition in the higher proficiency group, while the correlations in the lower proficiency group were weak (listening) or nonexistent (speaking).

The correlations reported in Zimmermann (2019) draw exclusively on external criteria assessing participants' speaking and listening skills at a single proficiency level, thereby limiting the psychometric interpretability of the findings. Furthermore, both Zimmermann and Grotjahn (2010) focussed specifically on the construct of paper-based SC-Tests. However, there is a growing shift towards computer-based administration of C-Tests worldwide (e.g., Mozgalina & Ryshina-Pankova, 2015; Riggs & Maimone, 2018), which introduces new variables – most notably, typing skills – that may influence test performance. The impact of such factors remains largely unexplored but is crucial for interpreting the results from digital formats.

To date, only preliminary findings are available on cognitive processing in time-reduced C-Tests. Zimmermann (2019) asked 120 participants whether they completed the gaps in the given order. Fewer than half reported using linear processing, regardless of timing condition, although backtracking was more frequent in the canonical C-Test. Zimmermann concluded that, while time constraints affect test-taking behaviour, her method was insufficient to uncover the underlying cognitive processes.

To summarize, initial findings on the construct of time-reduced C-Tests in L2 research suggest that, beyond assessing language knowledge, these tests may also tap into its automatisation and processing efficiency. Since procedural knowledge is a prerequisite for automatisation (DeKeyser, 2015), time constraints may also shift the construct's focus, placing greater emphasis on procedural knowledge. Furthermore, because high language proficiency is characterised by automatic processing, time reduction may affect test takers' scores differently depending on their proficiency levels. Finally, given that oral skills also rely on real-time processing under time pressure, speeded C-Tests may engage similar underlying abilities, leading to potential overlaps with measures of oral proficiency.

The present study aims to gather further evidence on the construct of the speeded C-Test through performance-based analyses. In line with established validity frameworks (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), we use score comparisons and relationships with external measures to explore whether time constraints in C-Test administration affect not only test difficulty but also the nature of the construct being measured. We acknowledge, however, that our performance-based analyses and correlations with external measures provide only indirect evidence about the underlying construct and cannot, on their own, disentangle timing or mode effects from genuine construct differences. To reduce construct-irrelevant variance in the computerised administration, typing speed is taken into account when comparing performance under timed and untimed conditions.

We pose the following research questions for computerised C-Tests administered in English and German:

> (1) How does the time variable influence C-Test total scores for learners of different proficiency levels when their typing skills are taken into account?
>
> (2) How does the time variable influence scores on individual C-Test texts of varying difficulty?
>
> (3) How does the time variable influence the correlations between the C-Test and a global measure of oral skills?

(4) Which of the two versions of the C-Test (canonical or speeded) is more predictive of learners' scores in an oral skills test?

Methodology

Participants

Two groups of participants were recruited online to take part in the study. The 237 learners of L2 English (60.8% female, 38.8% male, 0.4% non-binary) were university students between the ages of 18 and 37 (M = 25.25, SD = 3.98). Of the 42 different L1s, German (N = 46), Russian (N = 26), Turkish (N = 25) and Arabic (N = 18) were the most frequent. The German L2 sample consisted of 191 university students (58.1% female, 41.9% male) between the ages of 18 and 40 (M = 25.46, SD = 3.92) with 47 different L1s, with Russian (N = 30), Turkish (N = 23), English and Spanish (N = 14 each) being the most frequent.

Participants were asked to self-assess their level of proficiency in the target language on the CEFR scale from A1 to C2. Most participants in the English sample rated their proficiency at B2 – C2, whereas most learners of German considered their language skills to be somewhere between A2 and C1 (Appendix A).

Instruments

C-Tests

Both versions of the C-Test (canonical and speeded) used in this study were constructed using calibrated texts from the item bank of onSET (g.a.s.t., n.d.) ordered according to Rasch-based difficulty measures. Ten texts were selected in each language to create two parallel tests per language of comparable difficulty and topic variety (Table 1). Additionally, one text for each language was chosen to serve as an ice-breaker at the beginning of the study.

Table 1. Characteristics of the C-Test texts (English and German)

Text	CEFR level	Text title	Logit value in the item bank	Text title	Logit value in the item bank
English		**C-Test**		**SC-Test**	
1	**A1**	Oscar Wilde	-1.23	Travel Insurance	-1.33
2	**A2**	Au Pair	-0.81	Orcas	-0.87
3	**B1**	Short-Sightedness	-0.19	Pidgin and Creole	-0.19
4	**B2**	Variety of Birds	0.21	The Vikings' Traces	0.21
5	**C1**	Taste Protects our Body	0.87	Car-Hacking	0.88
German		**C-Test**		**SC-Test**	
1	**A1**	Career Choices	-1.43	Stay Abroad Programmes	-1.46
2	**A2**	Music	-0.98	Sleep and the Brain	-0.97
3	**B1**	New Academic Degrees	-0.21	Triumph of Technology	-0.21
4	**B2**	Vegetarianism	0.15	Speaking and Seeing	0.15
5	**C1**	The Ocean as the Basis of Life	0.93	Amateurs in Research	0.96

In current C-Test research, texts (or passages) are routinely treated as polytomous Rasch items (super-items/testlets), and their difficulty is represented by the estimated Rasch item parameter on a common logit scale (Eckes, 2011). Accordingly, our texts are more or less difficult to the extent that they elicit lower or higher expected scores from learners located at the same point on the underlying proficiency continuum. Linguistic features such as sentence length, lexical diversity, or word frequency can be interpreted as potential sources of this empirically derived difficulty (cf. Beinborn et al., 2014; Kaufmann, 2016), but they do not constitute the difficulty metric used in our analyses.

Following the canonical principle, twenty deletions were made in each text: the second half of every second word, beginning with the second sentence, was removed. Each gap was worth one point, with a maximum score of 100 points per test (ice-breakers were not assessed). Test takers' responses were scored automatically (1 point for a correct response; 0 points otherwise) and subsequently reviewed by the researchers to ensure that valid alternative answers were accepted. Only responses that were both grammatically and orthographically correct received credit (see Drackert & Timukova, 2020, for discussion of related limitations and possible alternatives). For the purposes of item and reliability analyses, texts were treated as polytomous super-items (for alternative scoring approaches, see e.g. Effatpanah et al., 2024).

Five minutes were allocated for each text in the canonical C-Test, while the time limits in the speeded version ranged from 1:30 minutes for the easiest Text 1 to 2:30 minutes

for the most difficult Text 5. These time limits were determined through an online try-out with native speakers of English (N = 12) and German (N = 16)[3], who were asked to complete the C-Test texts in their respective native language as quickly and accurately as possible. The average processing times of the participants who filled at least 90% of the gaps correctly were calculated and then adjusted for L2 learners by adding 20%. The resulting times (ranging from 1:40 to 2:50 minutes) were used in the pilot study and further adjusted (reduced) based on the averages observed in the L2 pilot participants (N_{ENG} = 34*;* N_{GER} = 21) across proficiency levels.

The C-Tests were implemented on *Moodle* and demonstrated high reliability for both language samples (Appendix B).

Oral Elicited Imitation Test

To measure listening and speaking skills in a compact format, the oral elicited imitation task (OEIT) was used. In this test, participants listen to sentences in the target language presented as audio-recordings on a computer and are asked to repeat them after a pause. The responses are recorded and assessed by human raters. Although initially proposed as a measure of general language proficiency, several studies have demonstrated OEIT's applicability to assessing oracy or global oral proficiency in L2 (Drackert, 2016; McManus & Liu, 2022; Wu & Ortega, 2013).

We constructed the OEITs using stimulus sentences from the English (16 sentences) and German (17 sentences) tests created by Ortega et al. (2002). Since Kostromitina and Plonsky (2021) found that EIT dependability improves with longer and more numerous stimuli, we added sentences ranging from 19 (English) and 20 (German) to 24 syllables to extend beyond those used by Ortega et al. These new sentences were piloted with native speakers, who were asked to judge their naturalness and acceptability, and recorded by male native speakers of British English and German, respectively, to balance out the original recordings by females and in American English. The final version consisted of 20 sentences of gradually increasing length (number of syllables ranging from 7 to 24) in both languages.

[3] The sample consisted of international university students from various disciplines, recruited in Germany and abroad.

The OEIT was administered on an online platform for psychological experiments and tests (Testable, n.d.). Each stimulus was followed by a 2.5-second pause, after which test takers had 15 seconds to repeat the stimulus. Recorded responses were rated on a scale from 0 to 4 points based on how closely they resembled the original, using a scoring rubric originally developed by Ortega et al. (2002) and revised and adapted for this study[4]. The maximum score was 80 points. Response recordings (per stimulus) were distributed so that each learner was assessed by three different raters for greater objectivity. Raters participated in a calibration session beforehand, with the resulting inter-rater reliability (Fleiss' Kappa) for the selected 10 samples (200 ratings) averaging around .83 for English and .84 for German. The OEIT proved to be highly reliable for both language samples (Appendix B).

Test of typing speed

Typing speed was measured using a free commercial test (10FastFingers, n.d.). Participants typed individual words in the target language presented on the screen and then submitted their results (words per minute) on the study platform.

Grammatical Acceptability Judgment Test

To control for learners' proficiency level while correlating the C-Test results with the oral performance (RQ3), we used their scores on the Grammatical Acceptability Judgment Test (GAJT). We constructed the GAJT for English using the items from the tests presented in DeKeyser (2000) and Lu (2010) and then developed the German version following the same pattern[5]. The instrument has shown sufficient reliability for both language samples (Appendix B).

Establishing proficiency

Proficiency groups for RQ1 were based on Rasch person measures derived from OEIT scores, as elicited imitation tasks have long been used as a measure of language proficiency (Kostromitina & Plonsky, 2021). To ensure comparable group sizes as well as comparable and distinct ability ranges, only two groups were built from each

[4] The English and German OEIT items and the adapted scoring rubric used in the study are available in the IRIS database.
[5] The English and German versions of the GAJT used in the study are available in the IRIS database.

sample, including only a subset of the participants. In particular, participants at the edges and between groups were excluded. Reflecting differences in proficiency between samples, a medium (N = 51; 0 to +2.0 logits) and a higher (N = 58; +2.55 to +4.55 logits) proficiency group were built in English and a lower (N = 41; -4.0 to -0.9 logits) and a medium (N = 50; -0.5 to +2.5 logits) group in German.

For RQ3, we wanted to control for learners' proficiency while correlating the C-Test results with their oral performance as measured by the OEIT. Since both traditional measures of global language proficiency had already been assigned different roles in the study design, we resorted to the scores on the Grammatical Acceptability Judgment Test (GAJT). We acknowledge that grammatical judgment tests (GJTs) are not conventional measures of language proficiency. However, existing studies interpret GJT performance as indicative of learners' L2 linguistic ability, including both explicit and implicit grammatical knowledge (Gutiérrez, 2013). GJT-based knowledge measures have been found to correlate positively with external proficiency tests (Han & Ellis, 1998), and to be sensitive to proficiency level (Sándor, 2016). We therefore deemed GAJT scores sufficient to obtain a coarse differentiation between more and less proficient learners for our purposes and used them as a *proxy* for overall L2 proficiency, primarily reflecting grammatical knowledge.

Similar to groups for RQ1, the groups built using GAJT scores for RQ3 included only a subset of participants to ensure comparable group sizes and non-overlapping logit ranges. Also here the groups reflected proficiency differences between language samples in that, in English, the lower group had scores of 30–45, and the higher group, 46–60; in German, 20–40 and 41–60 respectively.

Data collection

Data collection took place as part of a larger study and was conducted online from August to October 2023. After registering, participants received study details for informed consent and text and video instructions. They were asked to complete the study on *Moodle* in one sitting of approximately 2.5 hours with at least one break in the middle.

The order of the instruments was fixed on the platform and the same for each participant: OEIT, canonical C-Test, GAJT, speeded C-Test and typing speed test. Due to limitations of the testing platform, the order of the instruments could not be randomised.

Data analysis

Data cleaning and preparation

The data was checked for outliers and cleaned (cf. Chapter 4 in Tabachnik & Fidell, 2014) prior to the main analyses. Descriptive statistics were calculated for the C-Tests in both language samples at the test and text level.

Since the dependent variable in the analysis of (co)variance should be normally distributed, the normality of C-Test score distributions was assessed. Scores on both test versions in each language sample showed roughly normal distributions. Deviations from normality (see Fig. C.1 – C.4 in Appendix C) remain within acceptable limits and are unlikely to impact the analyses, especially given the large sample sizes.

Item and reliability analyses were conducted for the C-Tests, the GAJT and the OEIT to ensure instrument quality. To check the suitability of the covariate, Spearman correlations were calculated between typing speed and C-Test and SC-Test scores. In English, typing speed showed satisfactory correlations with the variables of interest (r_s ranging from .40 to .50). In German, the correlations were deemed unsatisfactory ($r_s \leq .25$) and the covariate was not included in the analysis.

Finally, Rasch person measures were determined based on the OEIT scores using Winsteps (Linacre, 2023) to assign participants to proficiency groups for RQ1.

Main analyses

To address RQ1, a two-way mixed AN(C)OVA was conducted on C-Test scores across two time conditions (within-subjects factor) which included typing speed as a covariate (for English) and proficiency as a between-subjects factor. For RQ2, a two-way RM-ANOVA examined the interaction between time (first within-subjects factor) and text difficulty (second within-subjects factor), with *post hoc* one-way RM-ANOVAs for significant interactions. In this analysis, the dependent variable was text score

measured at five difficulty levels (Text 1, Text 2 etc.) under two timing conditions (canonical and speeded).

To answer RQ3, Spearman's ρ correlations[6] were calculated between OEIT and scores on the C-Test and SC-Test for the whole sample and separately for the two proficiency groups. Pearson and Filon's z was computed to statistically compare the correlations (Myers & Sirois, 2014). Simple linear regression was used to examine how C-Test and SC-Test scores predict OEIT scores (RQ4). All analyses were performed in IBM SPSS (Version 27), unless otherwise specified.

Results

RQ1: Time and test scores

Over 400 participants from both language samples (N_{ENG} = 222; N_{GER} = 183)[7] provided complete responses in both C-Test versions that could be analysed for descriptive parameters. As shown in Table 2, the speeded C-Tests were more difficult than their canonical counterparts for both samples, with average scores 3.7 points lower for English and 5.1 points lower for German under time constraints. A similar 4-point gap between the minimum and maximum scores in both tests was observed for German, but not for English, where the lowest C-Test score was 15 points higher than the lowest SC-Test score, with maximum scores nearly identical (96 vs. 95 points). Participants from the English sample outperformed the German learners on both C-Test versions and showed less score variability (*SD* = 15.2 – 17.7 in English vs. 21.4 – 21.9 in German).

[6] OEIT scores are 0 - 4 ratings that are best treated as ordinal. In our sample, the scores also deviated from normality. We therefore used Spearman's rank-order correlation (ρ) rather than Pearson's *r* when relating OEIT scores to C-Test and SC-Test performance, in line with SLA statistics texts recommending Spearman's ρ for ordinal or non-normally distributed variables (Hatch & Lazaraton, 1991; Larson-Hall, 2016).

[7] Discrepancies between these and the participant numbers in separate analyses result from missing and/or invalid data from other tests relevant to those analyses. For example, not all participants with valid C-Test and SC-Test results provided complete OEIT responses that could be included in the correlational and regression analyses.

Table 2. Descriptive parameters for C-Test and SC-Test scores (English and German)

		N	*M*	*SD*	*Min*	*Max*
English	C-Test	222	70.1	15.2	28	96
	SC-Test	222	66.4	17.7	13	95
German	C-Test	183	48.7	21.4	8	93
	SC-Test	183	43.6	21.9	4	89

Table 3 presents the mean C-Test and SC-Test scores for two proficiency groups in the English sample. Participants in the medium group lost an average of 6.2 points with the reduced time, whereas those in the higher group scored only 1.5 points lower on the speeded version.

Table 3. Descriptive parameters for C-Test scores for proficiency groups (English)

Proficiency group		**C-Test**		**SC-Test**	
	N	***M***	***SD***	***M***	***SD***
Medium	51	62.3	12.8	56.1	16.4
Higher	58	83.8	7.2	82.3	7.7

This tendency was also observable in the profile plots in Figure 1: the less proficient "medium" group (blue line; steeper slope) seemed to be more affected by the reduced time than the more proficient "higher" group (green line; less steep).

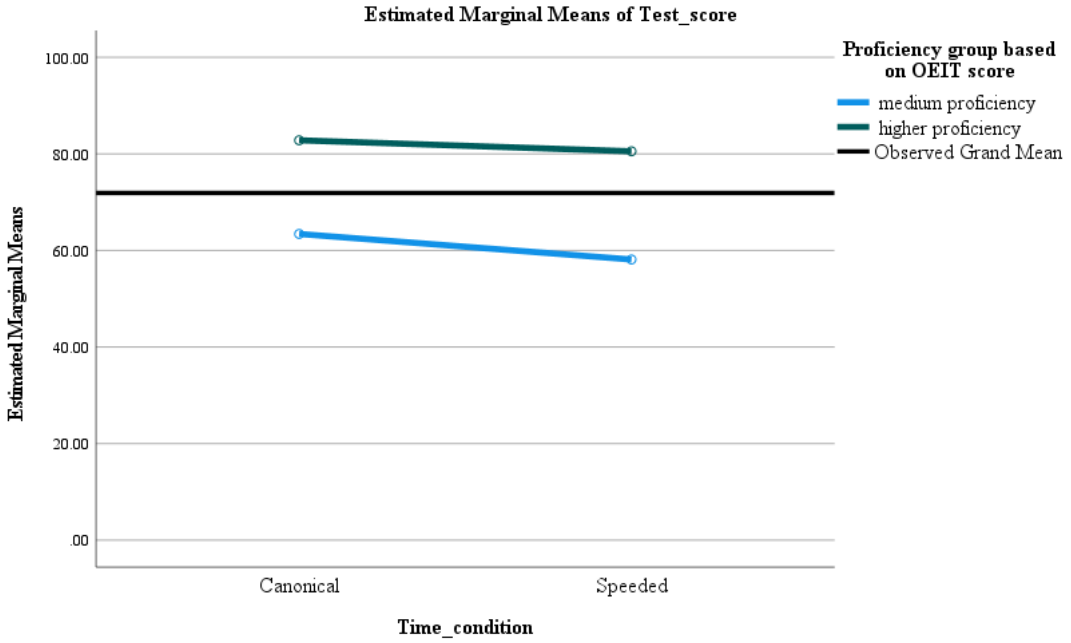

Figure 1. Profile plots for proficiency groups (English)

In the German sample, performance differences between proficiency groups were less pronounced: the lower group lost an average of 6.3 points under time constraints, while the medium group scored 4.6 points lower on the speeded C-Test than on the canonical version (Table 4).

Table 4. Descriptive parameters for C-Test scores for proficiency groups (German)

Proficiency group		**C-Test**		**SC-Test**	
	N	*M*	*SD*	*M*	*SD*
Lower	41	28.8	9.1	22.5	8.2
Medium	50	59.0	16.5	54.4	16.6

Similarly, the profile plots in Figure 2 show that the less proficient "lower" group (blue line; somewhat steeper slope) was only marginally more affected by the reduced time than the more proficient "medium" group (green line; somewhat less steep).

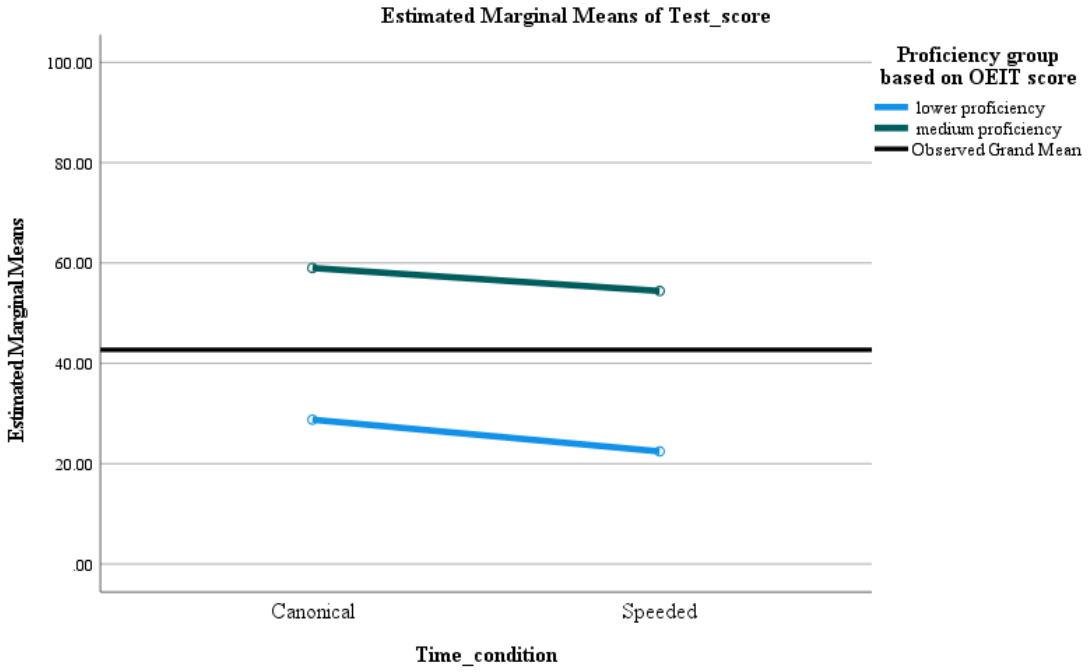

Figure 2. Profile plots for proficiency groups (German)

The two-way mixed ANCOVA[8] examining the effect of reduced time on learners' performance (RQ1) revealed no significant interaction between time condition and proficiency group in English when the typing speed of the participants was controlled

[8] Since the assumption of homogeneity was not upheld in both samples (Levene's test and Box's test were both statistically significant), we followed the recommendation by Tabachnick and Fidell (2014, p. 86) to set a more stringent α level of .01 for the analysis. The assumption of sphericity does not have to be met as the within-subjects factor only has two levels.

for, $F(1, 106) = 3.38$, $p = .069$, partial $\eta^2 = .031$. This indicates that the impact of time constraints was consistent across proficiency groups. However, both the main effect of time across proficiency groups ($F(1, 106) = 22.44$, $p < .001$, partial $\eta^2 = .175$) and the main effect of group membership regardless of time condition ($F(1, 106) = 115.19$, $p < .001$, partial $\eta^2 = .521$) on the C-Test scores were significant, with large[9] effect sizes.

Similarly, in German, the two-way mixed ANOVA showed no significant interaction between time condition and proficiency group, $F(1, 89) = 1.07$, $p = .303$, partial $\eta^2 = .012$, indicating that the effect of time constraints was the same for each proficiency group. Also here both main effects were statistically significant with large effect sizes: time condition, $F(1, 89) = 31.70$, $p < .001$, partial $\eta^2 = .322$; and group membership, $F(1, 89) = 129.06$, $p < .001$, partial $\eta^2 = .592$.

RQ2: Time and text difficulty

Research question 2 examines the role that the difficulty of individual C-Test texts might play alongside the time factor. In both languages, scores declined as text difficulty increased (i.e., from Text 1 to Text 5) under both timing conditions. However, boxplots of text scores in both samples (Figures 3 and 4) show greater score dispersion in the speeded C-Test version compared to the canonical one. This difference was especially evident in English and somewhat less pronounced in German, where overall score variability was higher. Descriptive data in Appendix D indicate that text score differences between timing conditions varied by text. In English, the largest difference appeared in text pair 3 (1.2 points), while the most difficult Text 5 showed almost no difference (0.4 points). In German, the differences were most pronounced for the easiest text pairs (Text 1 and especially Text 2, with 1.1 and 2.1 points respectively), with only 0.6 points difference for Text 4.

[9] Effect sizes are interpreted based on Cohen (1988), unless otherwise specified.

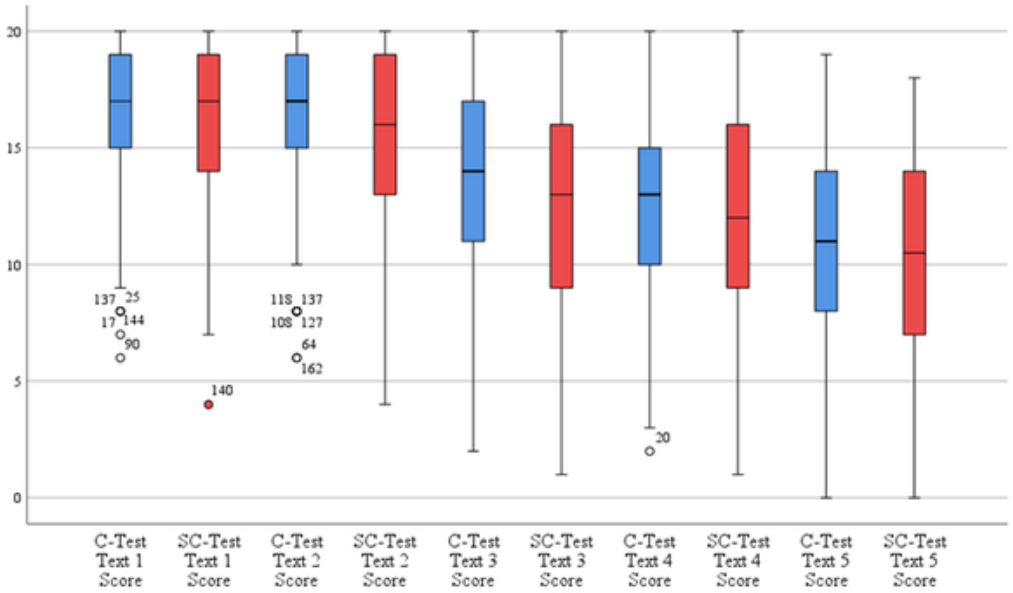

Figure 3. Boxplots of C-Test and SC-Test text scores in English

Note. Sample sizes differ because not all participants completed every text. Ns range from 225 to 229; exact numbers are provided in Appendix D.

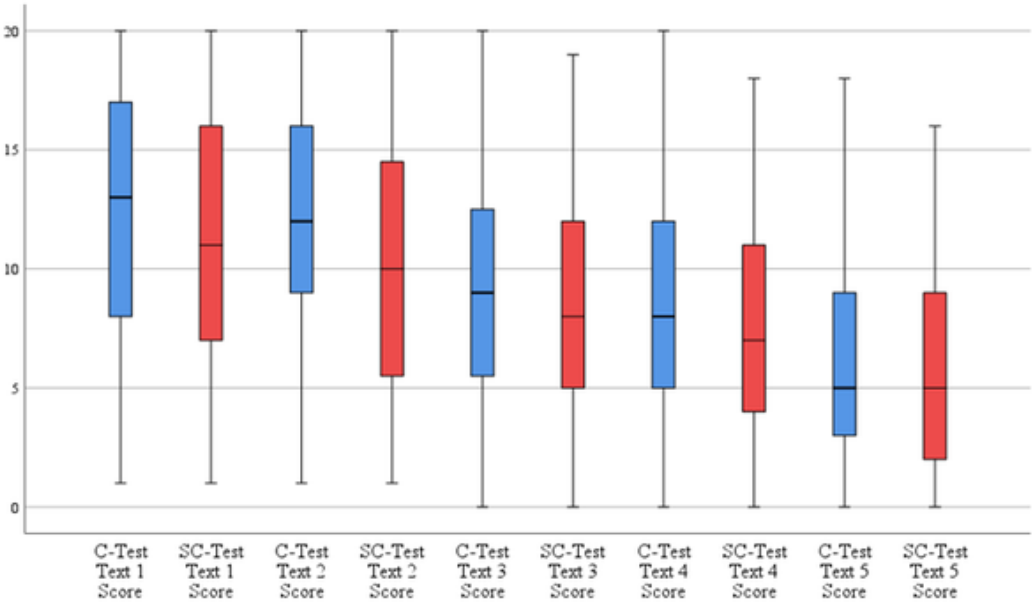

Figure 4. Boxplots of C-Test and SC-Test text scores in German

Note. Sample sizes differ because not all participants completed every text. Ns range from 184 to 188; exact numbers are provided in Appendix D.

To examine whether text difficulty moderates the effect of reduced time on test performance, a two-way RM-ANOVA was conducted. Mauchly's test confirmed that

the assumption of sphericity for the interaction term was met (p > .05) in both language samples.

In English, there was no statistically significant two-way interaction between time condition and text difficulty, $F(4, 876) = 2.29$, $p = .058$, partial $\eta^2 = .010$. Therefore, main effects were examined. The effect of time was significant, with lower mean text scores under time constraints and a medium effect size, $F(1, 219) = 31.73$, $p < .001$, partial $\eta^2 = .127$. Text difficulty also had a significant effect, with mean scores differing across texts and a large effect size, $F(4, 876) = 449.88$, $p < .001$, partial $\eta^2 = .673$.

In the German sample, the two-way interaction between time condition and text difficulty was statistically significant with a small effect size, $F(4, 704) = 8.56$, $p < .001$, partial $\eta^2 = .046$. Therefore, separate RM-ANOVAs were run for each text pair, applying Bonferroni adjustment for multiple comparisons, to examine the effect of text difficulty at each level of the time condition (i.e., simple main effects). Text scores differed significantly between the speeded and canonical conditions for all five text pairs (Appendix E). This indicates that, in German, the effect of time constraints on text scores varied by text difficulty.

RQ3: Time and correlations with oral skills

To examine the effect of time constraints on the relationship between C-Test performance and oral proficiency, Spearman's ρ correlations were calculated between both C-Test versions and the OEIT score in each sample. Results show strong, significant correlations for both C-Test formats in English and German.

Correlation coefficients were slightly higher for the speeded C-Test than for the canonical version, though their 95% confidence intervals partially overlapped (Table 5). The difference between the two C-Test correlations with the OEIT score was tested using Pearson and Filon's z and was not statistically significant in either sample (English: $z = -1.13$, $p = .259$; German: $z = -1.73$, $p = .084$).

Table 5. Correlations of C-Test and Speeded C-Test with OEIT (English and German)

		N	Spearman's ρ (95% CI)	*p*	r^2
English	C-Test	202	.70 (.61 - .76)	< .001	.48
	SC-Test	204	.73 (.65 - 79)	< .001	.53
German	C-Test	164	.86 (.82 - .90)	< .001	.75
	SC-Test	164	.89 (.85 - .92)	< .001	.79

To test whether the relationship between C-Test scores and a measure of oral proficiency is stronger among more proficient language learners, each sample was divided into higher and lower proficiency groups based on GAJT scores. Correlations between both C-Test versions and OEIT scores were then computed separately for each group.

In the English sample, all correlations were strong and significant across both proficiency groups. In the lower group, the C-Test and SC-Test showed similar coefficients with partially overlapping 95% confidence intervals (Table 6). In the higher group, the speeded C-Test correlated more strongly with OEIT than the canonical C-Test, though their confidence intervals also partially overlapped. Overall, both C-Test versions correlated more strongly with the OEIT score in the higher proficiency group, with the SC-Test yielding the highest correlation – exceeding that of the C-Test by 0.122.

Table 6. Correlations of C-Tests with OEIT for proficiency groups (English)

Proficiency group		*n*	Spearman's ρ (95% CI)	*p*	r^2
Low	C-Test	103	.42 (.24 - .57)	< .001	.17
	SC-Test	105	.44 (.27 - .59)	< .001	.20
High	C-Test	87	.50 (.32 - .65)	< .001	.25
	SC-Test	87	.62 (.47 - .74)	< .001	.38

In the German sample, both C-Test versions correlated significantly and strongly with the OEIT score across proficiency groups. In both groups, the SC-Test yielded slightly higher coefficients than the canonical version, though their 95% confidence intervals partially overlapped (Table 7).

Table 7. Correlations of C-Tests with OEIT for proficiency groups (German)

Proficiency group		*n*	Spearman's ρ (95% CI)	*p*	r^2
Low	C-Test	77	.63 (.47 - .75)	< .001	.40
	SC-Test	76	.66 (.50 - .77)	< .001	.43
High	C-Test	79	.71 (.58 - .81)	< .001	.51
	SC-Test	79	.73 (.60 - .82)	< .001	.53

Correlations were stronger in the higher proficiency group for both test versions. The SC-Test showed the highest correlation in this group, but the difference from the C-Test was minimal (0.019).

RQ4: Time and prediction of oral skills

To test which C-Test version better predicts oral proficiency, simple linear regressions were conducted. In the English sample, both the canonical C-Test ($F(1, 201) = 166.22$, $p < .001$, $R^2 = .45$) and the speeded C-Test ($F(1, 203) = 196.63$, $p < .001$, $R^2 = .49$) significantly predicted OEIT performance. A one-point increase in the C-Test score corresponded to a .499-point increase in OEIT, while a one-point increase in the SC-Test score corresponded to a .444-point increase. The C-Test explained about 45% of the variance in OEIT scores, and the SC-Test almost 50%, which according to Cohen (1992) indicates a large effect ($f = .91$ and $f = .98$, respectively).

Similarly, in the German sample, both C-Test versions significantly predicted OEIT scores (C-Test: $F(1, 162) = 462.62$, $p < .001$, $R^2 = .74$; SC-Test: $F(1, 162) = 570.12$, $p < .001$, $R^2 = .78$). Each additional point on the C-Test predicted a .751-point increase in OEIT, while each point on the SC-Test predicted a 1.018-point increase. The tests accounted for 74% and 78% of the variance in OEIT, respectively, again indicating large effects (Cohen's $f = 1.69$ for the C-Test and $f = 1.88$ for the SC-Test).

Discussion

Interpretation of the results

This study investigated how time constraints affect performance on the computerised C-Test in English and German, exploring whether reduced time primarily increases task difficulty or may also be associated with construct-relevant differences in

performance. By comparing learners' scores under canonical and speeded conditions, we examined how time shapes scores and their relationships with other measures and what this implies for interpreting C-Test results.

Our results confirm that time constraints affect learners' performance on the C-Test in both languages: test takers obtained lower scores under the speeded condition, with average differences of 3.7 points in English and 5.1 in German (out of 100). While this decrease was expected, its consistency across languages and statistical significance – even after controlling for typing speed in English – highlight the robustness of the time effect. At the same time, the modest size of these differences suggests that time constraints do not render the C-Test entirely different but rather modify its processing demands.

Differences between proficiency groups were inconclusive. While more proficient learners appeared less affected by time constraints – particularly the higher versus medium group in English – these differences were not statistically significant, suggesting that time pressure influences performance similarly across proficiency levels. The small numerical advantage for more advanced learners may indicate that the speeded format draws somewhat more on automatised or proceduralised knowledge, which tends to be more developed at higher proficiency levels. The absence of significant group interactions may thus reflect overlapping skill sets mobilised during test taking, rather than distinctly different demands.

The interaction between time constraints and individual text difficulty revealed language-specific differences. In English, no statistically significant interaction was found, indicating that time condition affected performance on all texts similarly. In German, however, text difficulty moderated the impact of time on scores across all five texts. This partially supports the assumption that the construct of the C-Test is not static but is co-constructed by test format, task features, and learner characteristics (cf. Sigott, 2004). The fact that this interaction emerged only in German likely reflects both the greater relative difficulty of the German test, due to the lower proficiency of that sample, and broader language-specific processing differences, which together may have amplified the effect of time on text difficulty.

Further analyses of text and gap characteristics – such as the proportion of gaps requiring text-level processing versus those dependent on micro-context – could clarify why texts behaved differently across languages. German's richer inflectional morphology and verb-final structures may require integrating information across larger stretches of text, whereas English often provides more local cues. In multilingual assessment, such language-specific features can influence how strongly text difficulty moderates timing effects, even for typologically related languages.

The assumption that the speeded C-Test would correlate more strongly with a measure of oral proficiency than the canonical C-Test was only partially supported. In both language samples, SC-Test scores showed slightly higher correlations with OEIT scores, but these differences were not statistically significant. Consistent with expectations and previous findings (Zimmermann, 2019), these correlations were somewhat higher for more proficient learners, especially in English. Speeded C-Test scores also accounted for slightly more of the OEIT score variability than the canonical C-Test, explaining approximately an additional 4% in both language samples.

These patterns suggest a modest but meaningful overlap between the speeded C-Test and oral proficiency, consistent with the interpretation that the SC-Test, like the OEIT, draws on proceduralised and automatised language knowledge and skills required for real-time comprehension and production. The stronger associations observed in more proficient learners support this interpretation, as higher proficiency is typically associated with more automatised language processing. Overall, the findings suggest that the speeded C-Test may place greater emphasis on proceduralised and automatized skills than the canonical version and that aspects of real-time language use, which are also important for oral skills, may play a somewhat stronger role in performance in the speeded format. These interpretations are tentative, as our performance-based and correlational analyses provide only indirect evidence about construct representation. These findings also further highlight the fluidity of the construct, as the test appears to elicit somewhat different aspects of knowledge and skills in more proficient than in less proficient learners.

Interestingly, the overlap between C-Tests and oral skills was considerably larger in the German sample compared to the English one, both in correlations and linear regressions. In the German group, both C-Test versions accounted for roughly 40%

more variance in OEIT scores than in the English group. This difference may stem from broader disparities between the two samples. Lower proficiency levels and greater proficiency heterogeneity in the German sample may have contributed to the higher correlations. Additionally, differences in learners' language proficiency profiles – particularly the balance between oral and written skills – may also play a role. Proficiency profiles are typically shaped by learning goals and methods, which depend on intended language use and likely differ between learners of English and German. It is plausible that the L2 German learners in our study had a more balanced distribution of oral and written skills than their English-learning counterparts[10].

Taken together, these performance-based findings suggest that time pressure does more than simply reduce scores: it may also subtly shift the processing demands of the test, particularly for more proficient learners and depending on text characteristics. The speeded C-Test appears to tap somewhat different aspects of linguistic abilities, and to do so to varying degrees depending on learner proficiency, text difficulty, and language. These results are consistent with a view of the C-Test as a fluid and context-sensitive measure whose interpretation should take account of test design and learner characteristics.

Limitations

The difference between the language samples discussed in the previous section represents one of the main limitations of this study. It prevents direct comparability of the results and limits both their interpretability and generalisability.

Another important limitation is the fixed order of the instruments: all participants completed the canonical C-Test before the speeded version due to logistical constraints. This design prevented us from controlling for potential training effects, leaving their impact on speeded C-Test scores unknown. Randomising the order of the instruments in future studies would help mitigate such effects and might not only lead to larger and more consistent score differences between time conditions but also

[10] These differences may be attributed to the recruitment channels. Most participants were recruited through German universities and language centres, with only a few from other European institutions. As a result, the German sample was largely composed of international students studying in Germany, who likely learned German for academic purposes. In contrast, the English sample included mostly German students as well as some international students abroad, many of whom likely learned English for a broader range of purposes and contexts.

influence the results of statistical tests – particularly those examining the interaction between time and text difficulty (in English) and proficiency (in both languages).

Although the difference in point loss due to reduced time between proficiency groups was not statistically significant, researchers should be cautious about accepting the null hypothesis solely because it was not rejected. It remains plausible that the difference observed in the present sample exists in the broader population, as more advanced learners, who possess more and more readily accessible linguistic abilities, should theoretically be better able to demonstrate these under time constraints than less proficient learners. In other words, having more linguistic resources should make advanced learners less susceptible to the speed-ability trade-off.

A further limitation concerns the availability of German keyboards outside Germany and the resulting quality of data from the typing speed test in German. Since data collection was conducted online, some participants may not have had access to a German keyboard and were therefore unable to type the characters Ä, Ü, Ö and ß. This was evident in their C-Test responses, where manual score adjustment could be made to compensate for missing umlauts. However, we had no access to participants' written responses on the typing speed test and had to rely on their self-reported results (words per minute). Consequently, we were unable to adjust these scores to account for keyboard limitations. As a result, the typing speed measures for German did not correlate strongly enough with C-Tests scores and could not be included in the ANCOVA for RQ1.

Conclusions

This study examined how time constraints affect C-Test performance, exploring whether reduced time mainly increases test difficulty or may also have construct-relevant effects. The computerised speeded C-Test in English and German has proven to be highly reliable and distinctly more difficult than its canonical counterpart. The pattern of results is consistent with the interpretation that the speeded format places somewhat greater emphasis on automatised language knowledge and skills, although this inference remains tentative as it rests on indirect evidence. The construct's inherent malleability appears to persist under time constraints, with the aspects of knowledge and skills elicited from the more proficient learners differing slightly from

those of less proficient learners. There is also some interplay between the time variable and text difficulty, at least in German.

In practical terms, a C-Test that requires less time and may better reflect automatised, oral-like language abilities could serve as a useful research or placement tool, especially for advanced learners and in adaptive testing contexts. For multilingual testing programmes that employ C-Tests in several languages, the findings likewise indicate that timing and text difficulty need to be calibrated and validated in a language-specific manner, as they may shape the construct differently across languages and thereby constrain straightforward score comparability.

Although modest, the findings contribute meaningfully to a broader, multi-method approach to construct validation and provide a foundation for further research. Importantly, the extent to which differences between differently timed C-Tests reflect a shift in the underlying construct – as opposed to a mode or timing effect – cannot be determined from score analyses and correlations alone. Future research should therefore investigate learners' response processes and the specific language knowledge components they draw upon under varying time constraints. This could involve combined analyses of response behaviour and response content, complemented by a systematic examination of the linguistic and structural characteristics of test texts and gaps. Future refinement of the experimental design – such as stricter time limits for the speeded C-Test and tighter control of completion time in the canonical version – would enhance comparability. Additionally, incorporating individual difference variables other than typing speed (e.g. personality traits like introversion/extraversion) as covariates may help account for other sources of construct-irrelevant variance and improve the interpretive validity of results.

Acknowledgements

This study was only possible with the support of student research assistants Darija Felberg, Ilka Plesse and Anna Poberezhnaia. We also thank Dr. Thomas Eckes, Mirka Mainzer-Murrenhoff, Dr. Anja Peters, Leska Schwarz and Dr. Sonja Zimmermann for reviewing the manuscript and providing valuable feedback prior to submission, as well as the anonymous reviewers for their constructive comments on earlier versions of this article.

Author disclosures

The authors reported no conflicts of interest in conducting the study.

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG) [grant number 462766474].

CRediT authorship contribution statement

Anna Timukova: Conceptualization, Methodology, Project administration, Investigation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing; **Franziska Möller:** Methodology, Project administration, Investigation, Data curation, Formal analysis, Writing – original draft; **Anastasia Drackert**: Conceptualization, Methodology, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

During the preparation of this work, *ChatGPT 4o* was used to identify opportunities for improving readability and conciseness. The authors carefully reviewed and edited the content after using the tool and take full responsibility for the final version of the text.

ORCID iDs

Anna Timukova https://orcid.org/0009-0000-1883-4343

Franziska Möller https://orcid.org/0000-0001-7435-6120

Anastasia Drackert https://orcid.org/0000-0002-9649-2972

References

Aguado, K., Grotjahn, R., & Schlak, T. (2007). Erwerbsalter und Sprachlernerfolg: Zeitlimitierte C-Tests als Instrument zur Messung prozeduralen sprachlichen Wissens [Age of language acquisition and learning success: Time-limited C-Tests for measuring procedural language knowledge]. In H. J. Vollmer (Ed.), *Synergieeffekte in der Fremdsprachenforschung. Empirische Zugänge, Probleme, Ergebnisse* (pp. 137–149). Peter Lang.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum. https://doi.org/10.5040/9781474212151

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, *2*, 517–530. https://doi.org/10.1162/tacl_a_00200

Cohen, J. (1988). *Statistical power analysis for the behaviorial sciences* (2nd ed.). Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Council of Europe (2020), *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*(4), 499–533. https://doi.org/10.1017/S0272263100004022

DeKeyser, R. M. (2015). Skill acquisition theory. In B. VanPatten, & J. Williams (Eds.), *Theories in second language acquisition – An introduction* (pp. 94–112). Routledge. https://doi.org/10.4324/9780429503986-5

Drackert, A. (2016). *Validating language proficiency assessments in second language acquisition research*. Peter Lang. https://doi.org/10.3726/978-3-653-06280-9

Drackert, A., & Felberg, D. (2019). Ausbildung der Testentwicklungskompetenz angehender LehrerInnen und LinguistInnen durch Forschendes Lernen [Enhancing test development competence in prospective teachers and linguists through research-based learning]. In A. Drackert, & K. B. Karl (Eds.), *Sammelband des 2. Arbeitskreises „Didaktik der Slawischen Sprachen“* (pp. 17–57). Innsbruck University Press. https://doi.org/10.15203/3187-80-1-03

Drackert, A., & Timukova, A. (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

Eckes, T. (2010). Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung [The Online Placement Test German as a Foreign Language (onDaF): Theoretical foundations, construction, and validation]. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 125–192). Peter Lang. https://www.onset.de/fileadmin/Redakteure/PDF/Publikationen/Eckes_onDaF_2010.pdf

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling*, *53*(4), 414–439.

Effatpanah, F., Baghaei, P., Tabatabaee-Yazdi, M., & Babaii, E. (2024). A new scoring method for item response theory analysis of C-Tests. *Language Testing*, *42*(2), 167-192. https://doi.org/10.1177/02655322241265350

g.a.s.t. (n.d.) *onSET: online-Spracheinstufungstest [online placement test]*. https://www.onset.de/en/language-placement-test-english-onset/

Goldhammer, F. (2015). Measuring ability, speed, or both? challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives, 13*(3–4), 133–164. https://doi.org/10.1080/15366367.2015.1100020

Grotjahn, R. (2010). Gesamtdarbietung, Einzeltextdarbietung, Zeitbegrenzung und Zeitdruck: Auswirkungen auf Item- und Testkennwerte und C-Test-Konstrukt [Total presentation, individual text presentation, time limitation, and time pressure: Impact on item and test parameters and C-Test construct]. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 265–296). Peter Lang.

Grotjahn, R. (2013). C-test. In M. Byram, & A. Hu (Eds.), *Routledge encyclopedia of language teaching and learning* (2nd ed., pp. 180–181). Routledge. https://doi.org/10.4324/9780203101513

Grotjahn, R. (2019). C-Tests. In S. Jeuk, & J. Settinieri (Eds.), *Sprachdiagnostik Deutsch als Zweitsprache: Ein Handbuch* (pp. 579–603). De Gruyter Mouton. https://doi.org/10.1515/9783110418712

Grotjahn, R., & Drackert, A. (2022). The electronic C-Test bibliography: version October 2022. Available at: https://www.gast.de/de/forschung-entwicklung/publikationen/veroeffentlichungen-von-gast

Grotjahn, R., & Schiller, C. S. (2014): Zur Rolle des Makrokontexts bei der Bearbeitung spanischer C-Test-Texte: Fehleranalysen ausgewählter Lückenwörter [The role of macrocontext in the processing of Spanish C-Test texts: Error analyses of selected gap words]. In R. Grotjahn (Ed.), *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* pp. 277-289). Peter Lang.

Grotjahn, R., Schlak, T., & Aguado, K. (2010). S-C-Tests: Messung automatisierter sprachlicher Kompetenzen anhand von C-Tests mit massiver textspezifischer Zeitlimitierung [S-C-Tests: Measuring automated linguistic competencies using C-Tests with substantial text-specific time limitation]. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 297–319). Peter Lang.

Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 35*(3), 423–449. https://doi.org/10.1017/S0272263113000041

Han, Y., & Ellis, R. (1998). Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research, 2*(1), 1–23. https://doi.org/10.1177/136216889800200102

Harsch, C., & Schröder, K. (2007). Textrekonstruktion: C-Test [Text reconstruction: C.Test]. In B. Beck, & E. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 212–225). Beltz. http://www.pedocs.de/volltexte/2010/3140/pdf/978_3_407_25398_9_1A_D_A.pdf

Hastings, A. J. (2002): Error analysis of an English C-Test: Evidence for integrated processing. In R. Grotjahn (Ed.), *Der C-Test. Theoretische Grundlagen und praktische Anwendungen* 4 (pp. 53-66). AKS-Verlag, 53–66.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Newbury House.

Kaufmann, N. (2016). Die Vorhersage der Schwierigkeit deutscher C-Test-Texte: Untersuchungen am Beispiel des onDaF [Predicting the difficulty of German C-test texts: Investigations using the example of the onDaF]. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *21*(2), 111-126.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing, 14(1),* 47-84. https://doi.org/10.1177/026553229701400104

Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: a meta-analysis. *Studies in Second Language Acquisition, 44*(3), 886–911. https://doi.org/10.1017/S0272263121000395

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). Routledge.

Linacre, J.M. (2023). Winsteps® (Version 5.6.0) [Computer Software] Winsteps.com. Available from https://www.winsteps.com

Lu, Y. (2010). *Cognitive factors contributing to Chinese EFL learners' L2 writing performance in timed essay writing* [Doctoral dissertation, Georgia State University]. ProQuest Dissertations & Theses Global.

Mainzer-Murrenhoff, M., & Drackert, A. (2022). Erfassung von Sprachkompetenzen im Deutschen von Schülerinnen und Schülern mittels C-Tests im Arbeitsfeld Sprachbildung: Fragen der Validität revisited [Assessing German language proficiency of school students using C-Tests in the field of language education: Revisiting questions of validity]. *Informationen Deutsch als Fremdsprache*, 49(5), 470–492. https://doi.org/10.1515/infodaf-2022-0067

McManus, K., & Liu, Y. (2022). Using elicited imitation to measure global oral proficiency in SLA research: A close replication study. *Language Teaching, 55(1)*, 116–135. https://doi.org/10.1017/S026144482000021X

Mozgalina, A., & Ryshina-Pankova, M. (2015). Meeting the challenges of curriculum construction and change: Revision and validity evaluation of a placement test. *The Modern Language Journal, 99*(2), 346–370. https://doi.org/10.1111/modl.12217

Myers, L., & Sirois, M. J. (2014). Spearman correlation coefficients, differences between. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels (Eds.), *Wiley StatsRef: Statistics reference online*. Wiley. https://doi.org/10.1002/9781118445112

Naismith, B., Cardwell, R., LaFlair, G. T., Nydick, S., & Kostromitina, M. (2025). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo.

Norris, J. M. (2018). Developing and investigating C-tests in eight languages: Measuring proficiency for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 7–33). Peter Lang. https://doi.org/10.3726/b13235

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research* [Conference presentation]. Second Language Research Forum, Toronto, Canada.

Raatz, U. (2002). C-Tests and intelligence. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-test* (pp. 169–185). AKS-Verlag. http://www.c-test.de/deutsch/index.php?lang=de&section=originalia

Reichert, M., Keller, U., & Martin, R. (2010). The C-test, the TCF and the CEFR: A validation study. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from current research* (pp. 205–231). Peter Lang.

Riggs, D., & Maimone, L. L. (2018). A computer-administered C-test in Spanish. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 265–294). Peter Lang. https://doi.org/10.3726/b13235

Sándor, K. (2016). Grammaticality judgement tests as measures of explicit knowledge. *Argumentum, 12*, 216–230.

Sigott, G. (2004). *Towards identifying the C-Test construct*. Peter Lang.

Tabachnick, B., & Fidell, L. (2014). *Using multivariate statistics* (6th ed.). Pearson Education.

10FastFingers. (n.d.). *Typing speed test*. https://10fastfingers.com/

Testable. (n.d.). *Testable* [online platform for creating, running, and managing experiments, surveys, and psychological tests]. https://www.testable.org/

Tijmstra, J., & Bolsinova, M. (2018). On the importance of the speed-ability trade-off when dealing with not reached items. *Frontiers in Psychology, 9*, Article 964. https://doi.org/10.3389/fpsyg.2018.00964

Ullman, M. T. (2020). The Declarative/Procedural Model: A neurobiologically-motivated theory of first and second language. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition* (3rd ed., pp. 128-161). Routledge. https://doi.org/10.4324/9780429503986-7

Wockenfuß, V. (2008). *Diagnostik von Sprache und Intelligenz bei Jugendlichen und jungen Erwachsenen: Eine empirische Untersuchung* [*Assessment of language and intelligence in adolescents and young adults: An empirical study*] [Dissertation, Universität Duisburg-Essen]. Universitätsbibliothek Duisburg-Essen.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 46*(4), 680–704. https://doi.org/10.1111/flan.12063

Zimmermann, K. (2019). *Keine Zeit für den C-Test? Eine empirische Untersuchung zum Einfluss einer Geschwindigkeitskomponente auf das Konstrukt des C-Tests* [*No time for the C-Test? An empirical study on the impact of a speed component on the C-Test construct*]. Universitätsverlag der TU Berlin. http://dx.doi.org/10.14279/depositonce-8288

Appendix A. Self-assessment of proficiency in English (*N* = 237) and German (*N* = 185)

	English		*German*	
	n	***%***	***n***	***%***
A1	2	0.8	26	14.1
A2	6	2.5	44	23.8
B1	23	9.7	43	23.2
B2	70	29.5	22	11.9
C1	102	43.0	41	22.2
C2	34	14.3	9	4.9

Note: Six participants in the German sample failed to answer the question about their proficiency.

Appendix B. Reliability of the instruments (English and German)

	C-Test	*SC-Test*	*OEIT*	*GAJT*	*TTS*
	Cronbach's α (95% CI)				
English	.89 (.87, .91)	.90 (.88, .92)	.92 (.91, .94)	.88 (.86, .90)	no data
German	.95 (.94, .96)	.95 (.94, .96)	.97 (.96, .97)	.91 (.90, .93)	no data

Note: OEIT - Oral Elicited Imitation Test; GAJT - Grammatical Acceptability Judgement Text; TTS - Test of Typing Speed.

Appendix C. Distributions of the C-Test and SC-Test total scores (English and German)

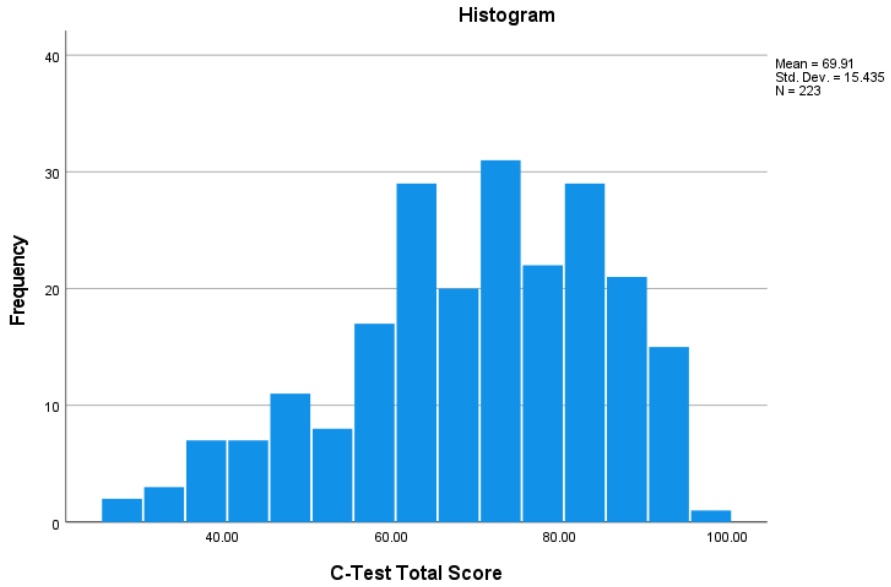

Fig. C.1 Distribution of the C-Test total scores in English

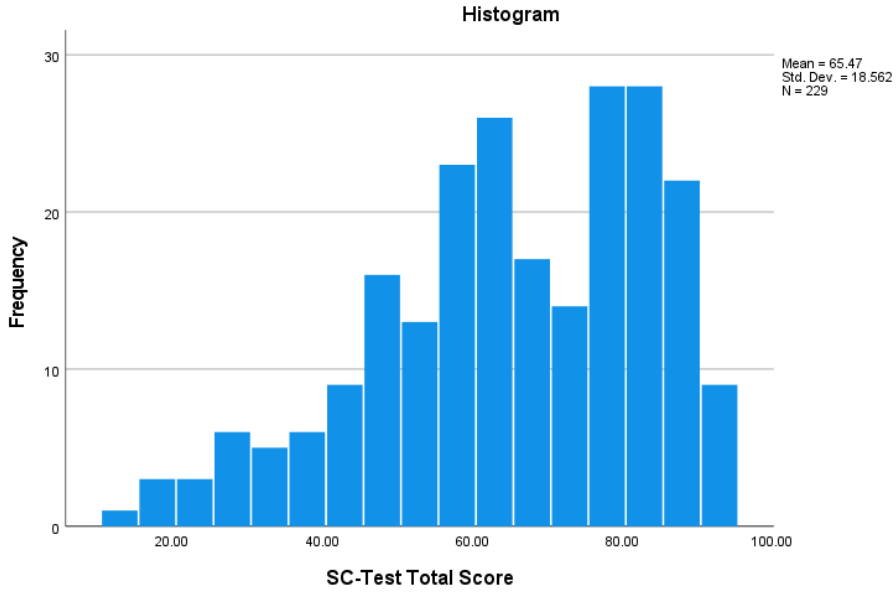

Fig. C.2 Distribution of the speeded C-Test total scores in English

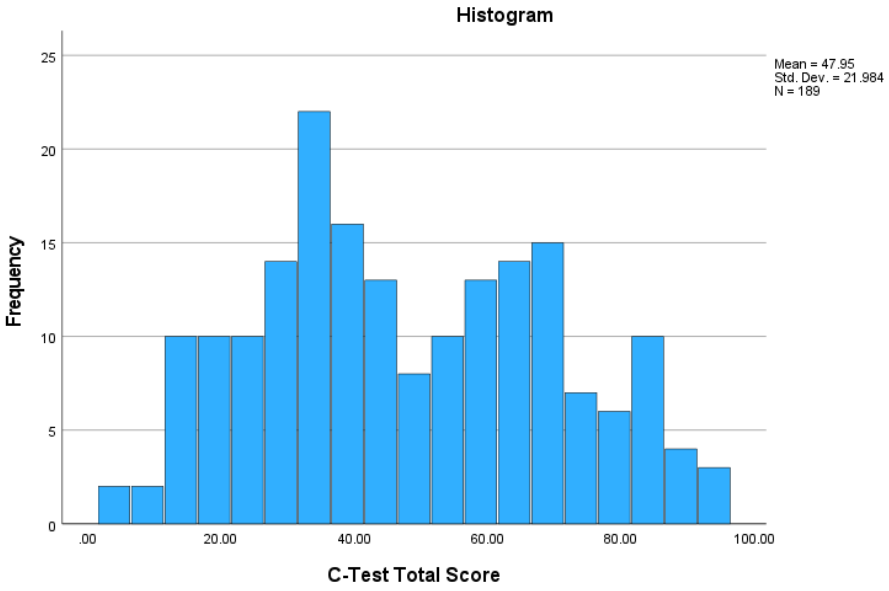

Fig. C.3 Distribution of the C-Test total scores in German

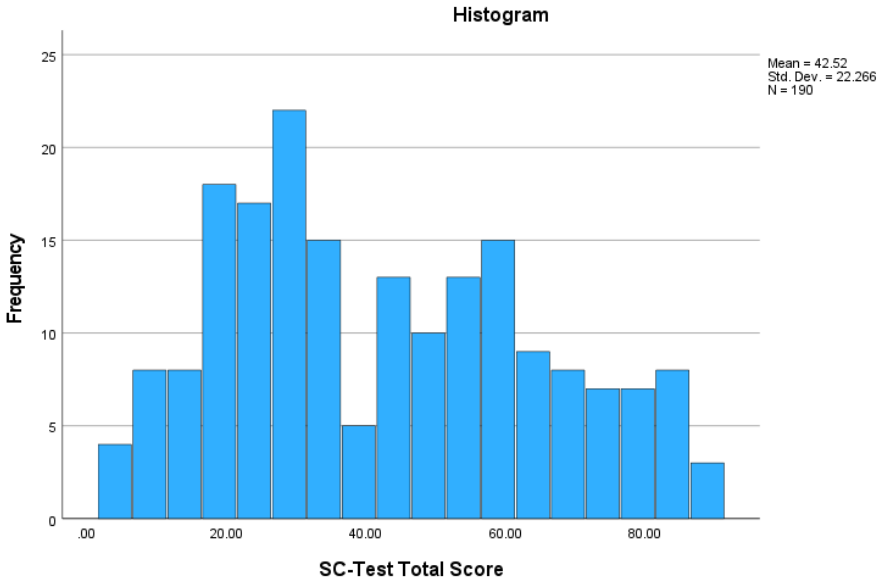

Fig. C.4 Distribution of the speeded C-Test total scores in German

Appendix D. Descriptive parameters for C-Test and SC-Test text scores (English and German)

Table D.1 Descriptive parameters for C-Test and SC-Test text scores in English

	Text 1		Text 2		Text 3		Text 4		Text 5	
	CT	**SCT**	**CT**	**SCT**	**C-T**	**SC-T**	**C-T**	**SC-T**	**C-T**	**SC-T**
N	229	229	228	228	227	228	225	228	225	228
M	16.6	15.8	16.3	15.6	13.5	12.3	12.3	11.8	10.7	10.3
SD	3.1	3.7	3.4	3.8	4.2	4.8	3.9	4.5	4.6	4.2

Table D.2 Descriptive parameters for C-Test and SC-Test text scores in German

	Text 1		Text 2		Text 3		Text 4		Text 5	
	CT	**SCT**	**CT**	**SCT**	**CT**	**SCT**	**CT**	**SCT**	**CT**	**SCT**
N	187	188	187	188	187	187	184	185	184	185
M	12.2	11.1	12.2	10.1	9.2	8.6	8.5	7.8	6.5	5.7
SD	5.2	5.1	4.7	5.3	4.6	4.7	4.7	4.5	4.5	4.1

Appendix E. Results of ANOVAs for mean differences in text scores (German)

Text pair	***N***	***F***	**Error (*MS*)**	***p***	**Part. η²**	**Mean difference**
Text 1	185	23.84	4.999	<.001	.115	1.1*
Text 2	185	101.98	4.135	<.001	.357	2.1*
Text 3	184	9.28	4.288	.003	.048	0.7*
Text 4	181	7.07	4.228	.009	.038	0.6*
Text 5	180	13.97	3.625	<.001	.072	0.8*

** The mean difference is significant at the .05 level; adjustment for multiple comparisons: Bonferroni.*