There are C-Tests and C-Tests: Digitalised Formats and Reduced Times - Changed Constructs?

Anastasia Drackert
Anna Timukova
Franziska Möller

03.07.2024

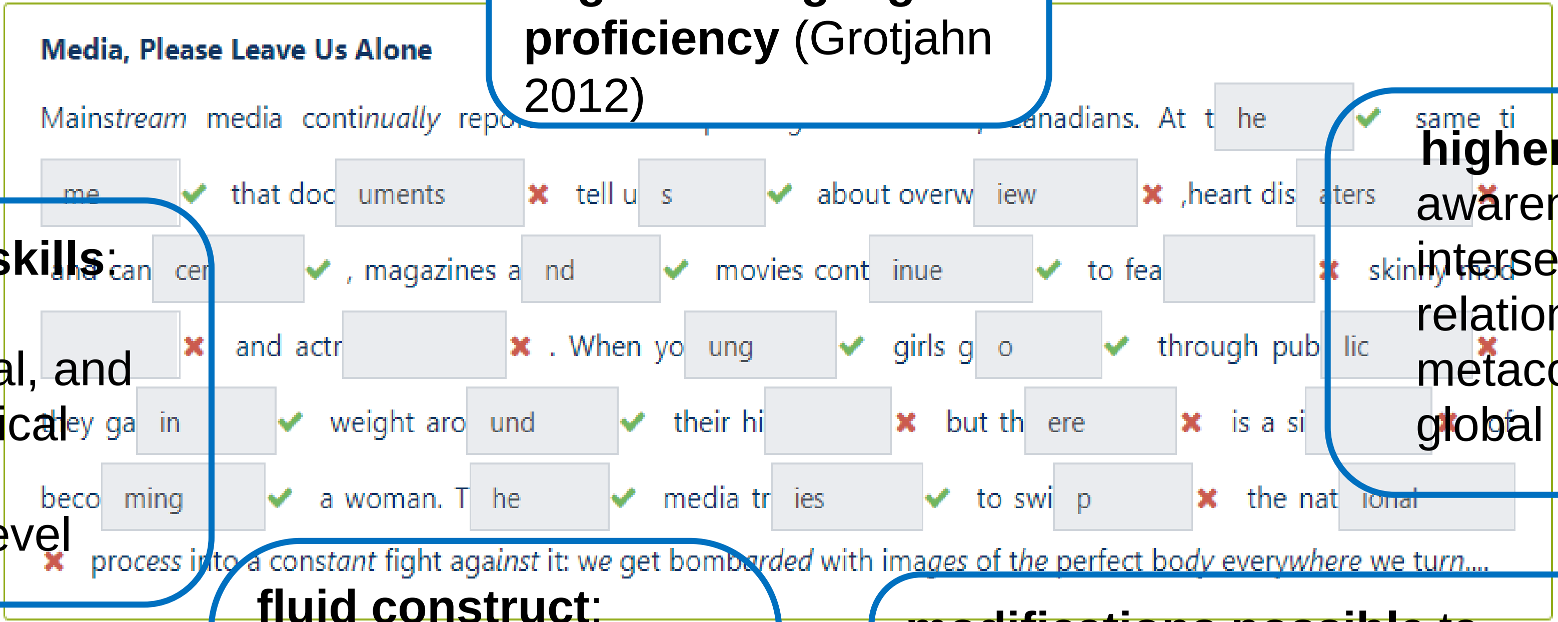
Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

project number 462766474

g.a.s.t.

C-Test & its construct

objective, reliable, economical measure of **global language proficiency** (Grotjahn 2012)



Media, Please Leave Us Alone

Mainstream media continually repo... Canadians. At t he ✓ same ti

me ✓ that doc uments ✗ tell u s ✓ about overw iew ✗ ,heart dis aters ✗

and can cer ✓ , magazines a nd ✓ movies cont inue ✓ to fea ✗ skinny mod

✗ and actr ✗ . When yo ung ✓ girls g o ✓ through pub lic ✗

they ga in ✓ weight aro und ✓ their hi ✗ but th ere ✗ is a si ✗ of

beco ming ✓ a woman. T he ✓ media tr ies ✓ to swi p ✗ the nat ional

✗ process into a constant fight against it: we get bombarded with images of the perfect body everywhere we turn....

low-level skills: lexical, grammatical, and orthographical skills at the sentence level

higher order skills: awareness of intersentential relationships, metacognitive strategies, global reading skills etc.

fluid construct: amount of text-level processing depends on test takers' proficiency and characteristics of the individual text (Sigott 2002; 2006)

modifications possible to construction principles, scoring and **time** to adjust to the target group, language and purpose

Construct of the Speeded C-Test

5 mins
per text

Grotjahn (2010):

- **canonical** C-Test measures the amount of learners’ **declarative and procedural knowledge**
- **speeded** C-Test additionally measures the **degree of automaticity** of their skills and the **efficiency of information processing** (cf. p. 285).

1:30 - 2:30
mins per text

Hypotheses:

- SC-Test would correlate higher with measures of listening comprehension and speaking skills (both under time pressure);
- SC-Test would correlate weaker with learners’ writing and reading skills if measured under generous time conditions than a canonical C-Test (p. 289)

Objective of the study

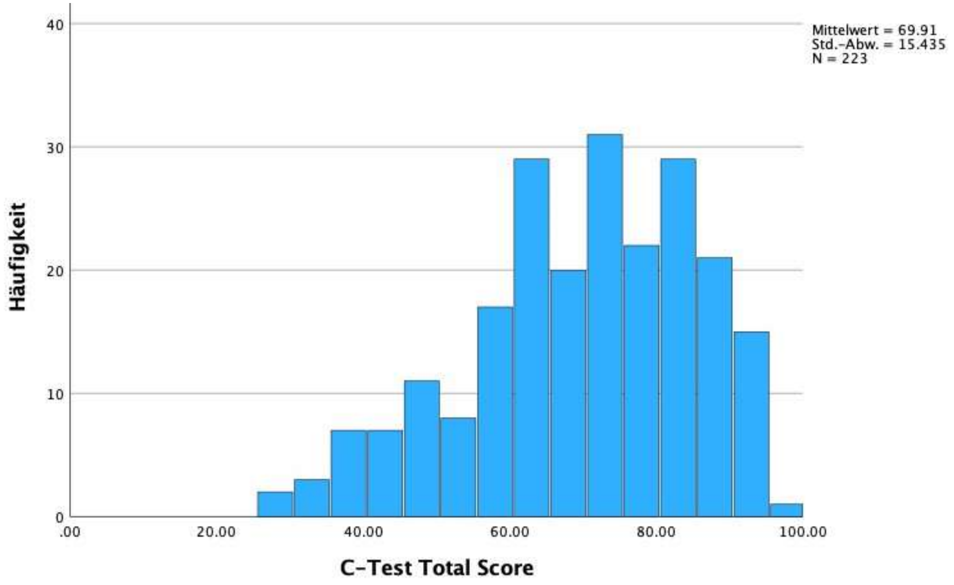
Using **different methods** gather **various types of evidence** to answer a range of questions to investigate **the role of the time variable** in the C-Test construct in a **comprehensive** way to allow for a higher degree of **generalizability** of the results for learners of different levels of proficiency; multiple languages (English, German, Russian); computer-administered C-Tests.

RQ	Method(s)
1. How does the time variable influence the **reliability** of computerised C-Tests?	IRT analysis; Cronbach’s alpha
2. How does the time variable influence **learners’ scores** depending on their **proficiency level** and **text difficulty**?	ANCOVA
3. Which components of L2 proficiency (**declarative, procedural knowledge** and **automaticity**) are better predictors of differently timed C-Tests?	Linear regression analysis; SEM
4. ... **correlations** between a C-Test and an integrated measure of **oral proficiency...**	Correlation; regression
5. How does the time variable influence the **strategies** deployed by learners?	Video-based analysis

Main study

- Data collection **online** (*Moodle*; *testable*) August – October 2023
- **Participants**: English (*N* = 229); German (*N* = 191); Russian (*N* = ca. 60)
- Instruments: **10 tests per language** (2 C-Tests; Oral Elicited Imitation Test (OEIT); test of typing speed; 6 tests of declarative and procedural knowledge)
- Fixed order of tests

N	**Age *M***	**L1**
229	25.25	42 different L1s: German (*n* = 46) Russian (*n* = 26) Turkish (*n* = 25) Arabic (*n* = 18)



RESULTS RQ1, RQ2 & RQ3

RQ1: HOW DOES THE TIME VARIABLE INFLUENCE THE RELIABILITY OF COMPUTERISED C-TESTS?

Method: IRT analysis; Cronbach’s alpha

Hypothesis: The reliability of the C-Test will be influenced by the time factor and learners’ L2 proficiency.

	IRT reliability estimates		***Cronbach‘s alpha***	***N of items***
	Person reliability	***Real separation***		
C-Test	.9 (*N* = 229)	3.05	.903 (N = 223)	5
Speeded C-Test	.91 (*N* = 230)	3.16	.911 (N = 226)	5

RQ1: HOW DOES THE TIME VARIABLE INFLUENCE THE RELIABILITY OF COMPUTERISED C-TESTS?

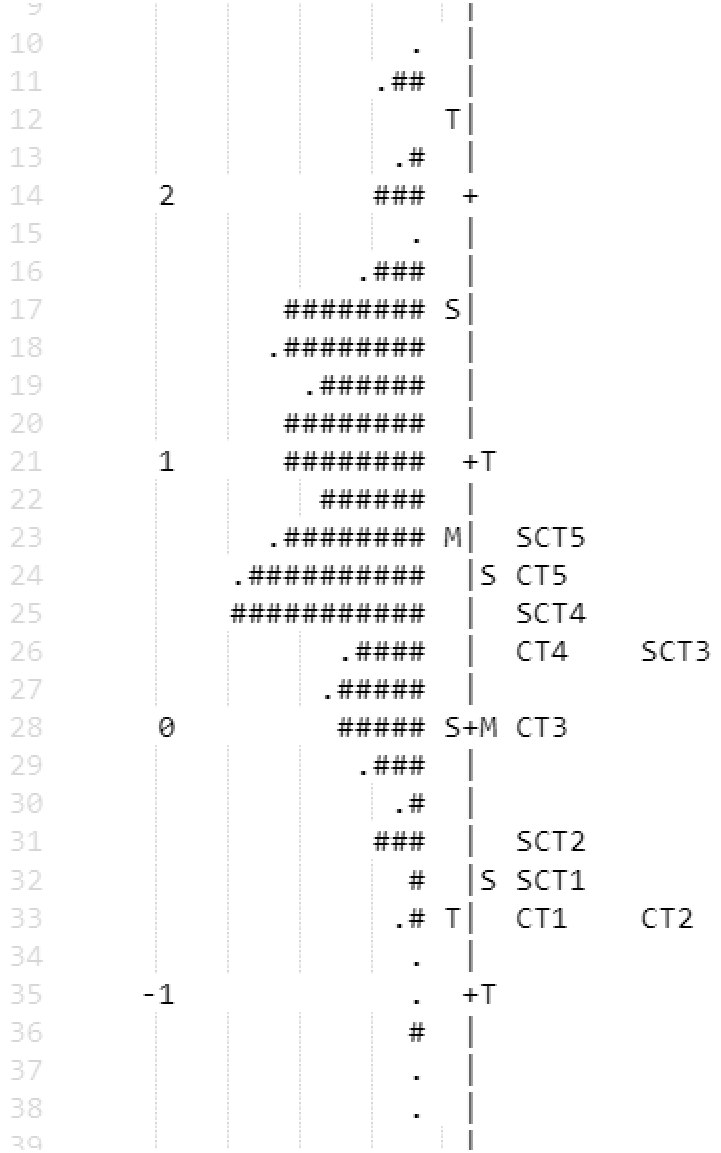
Learners’ proficiency

Group allocation based on IRT person measures derived from OEIT scores as produced by Winsteps 5.2.3.0. Reliability for OEIT (20 items): .91; REAL SEP.: 3.09

	Logit range
higher	*+2.55 to +4.55 logits*
medium	*0 to +2.0 logits*
lower	-

	N	**Cronbach‘s alpha C-Test**	**Cronbach‘s alpha Speeded C-Test**
Higher Prof.	60	.782	.684
Medium Prof.	55	.837	.876

WRIGHT MAP (C-Test & SC-Test texts)



```
                     |
                  .  |
                .##  |
                    T|
                 .#  |
 2              ###  +
                  .  |
               .###  |
           ######## S|
          .########  |
            .######  |
           ########  |
 1         ########  +T
             ######  |
          .######## M|  SCT5
        .##########  |S CT5
        ###########  |  SCT4
              .####  |  CT4     SCT3
             .#####  |
 0           ##### S+M CT3
               .###  |
                 .#  |
                ###  |  SCT2
                  #  |S SCT1
                 .# T|  CT1     CT2
                  .  |
-1                .  +T
                  #  |
                  .  |
                  .  |
                     |
```

RQ2: HOW DOES THE TIME VARIABLE INFLUENCE LEARNERS' SCORES?

Hypothesis 1: All learners' scores will **increase** with **additional time** irrespective of their typing skills and proficiency.

Hypothesis 2: All learners' scores will increase with additional time. The **amount of gain** in the scores will depend on learners' **level of proficiency**.

Hypothesis 3: Additional time will play a different role depending on the **difficulty of the C-Test texts**.

RQ2: HOW DOES THE TIME VARIABLE INFLUENCE LEARNERS' SCORES? (H1)

Descriptives

	N	*M*	*SD*	*Min.*	*Max.*
C-Test	222	70.10	15.21	28	96
Speeded C-Test	222	66.37	17.67	13	95

	N	*F*	*Part. Eta Squared*	*p*	
RM Within-Subjects ANCOVA with typing skills & proficiency (OEIT scores) as covariates	201	29.327	.129	< .001	Interaction with CVs significant (for TS p=.002; for proficiency p=.015)

RQ2: HOW DOES THE TIME VARIABLE INFLUENCE LEARNERS' SCORES DEPENDING ON THEIR PROFICIENCY LEVEL? (H1)

Descriptives

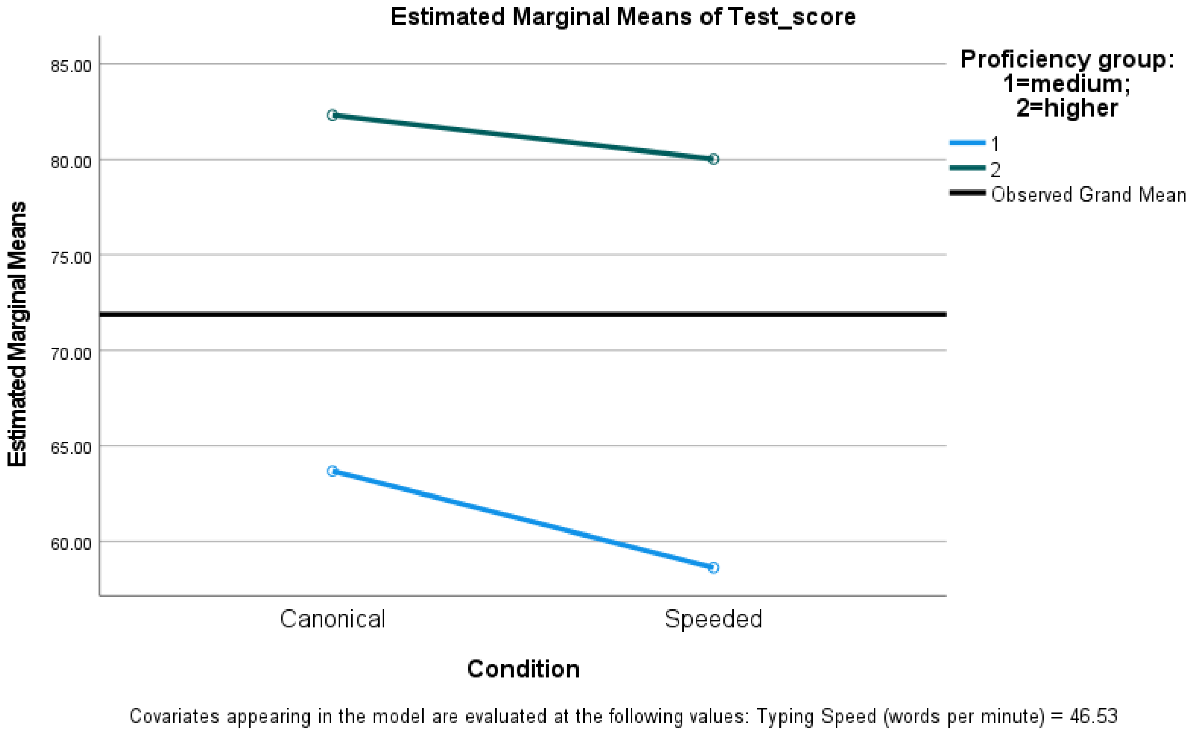
	Medium Proficiency* (N = 51)	**Higher Proficiency* (N = 59)**
C-Test *M*	62.6 (*SD* 13.3)	83.3 (*SD* 8.2)
Speeded C-Test *M*	56.6 (*SD* 17.1)	81.8 (*SD* 8.7)

	N	***F***	***Part. Eta Squared***	***p***	
RM Mixed Between-Within-Subjects ANCOVA (prof. group as between-subject factor; typing skills as a CV)	110	22.326	.173	< .001	interaction with TS significant (p=.001); interaction with prof group not significant (p=.092)

g.a.s.t.

RQ2: HOW DOES THE TIME VARIABLE INFLUENCE LEARNERS' SCORES DEPENDING ON THEIR PROFICIENCY LEVEL? (H2)

Profile plots for group* comparison

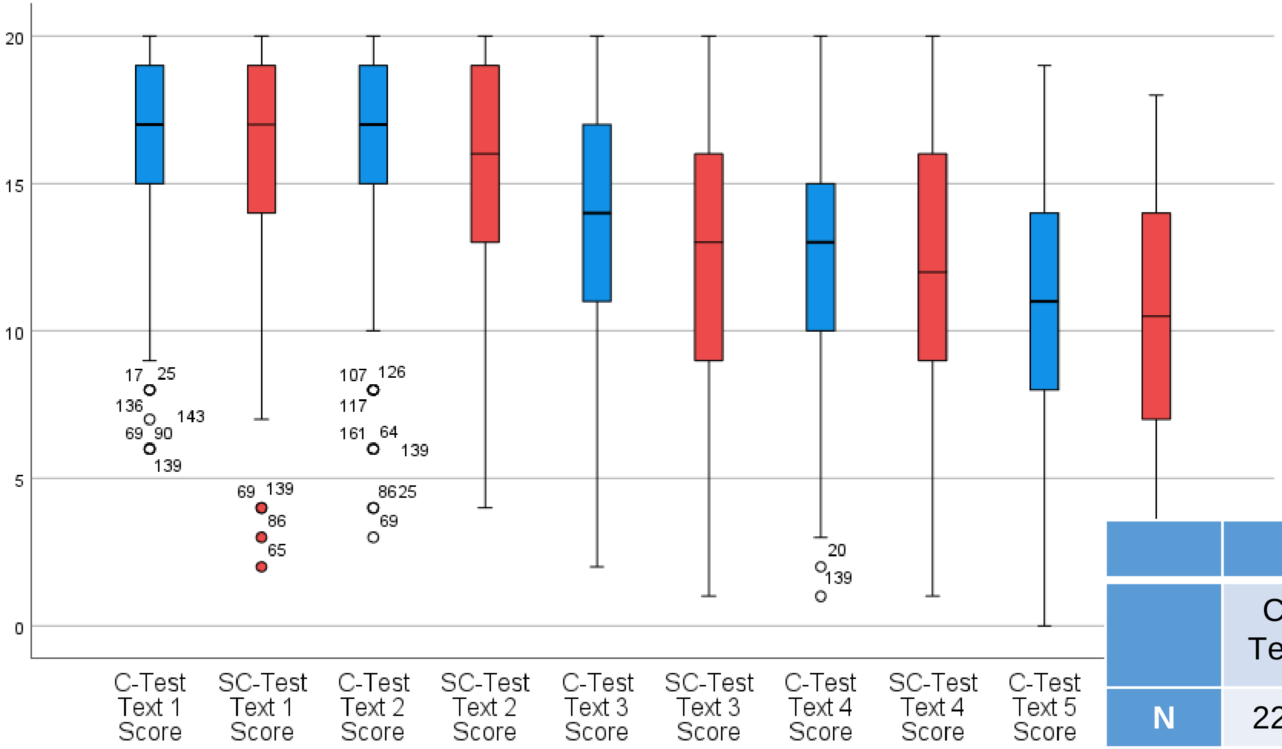


1.5 pts av. difference

6.0 pts av. difference

g.a.s.t.

RQ2: HOW DOES THE TIME VARIABLE INFLUENCE LEARNERS' SCORES RELATED TO THE TEXT DIFFICULTY? (H3)



	Text 1		Text 2		Text 3		Text 4		Text 5	
	C-Test	SC-Test	C-Test	SC-Test	C-Test	SC-Test	C-Test	SC-Test	C-Test	SC-Test
N	229	229	228	228	227	228	225	228	225	228
Mean	16.58	15.78	16.31	15.56	13.52	12.32	12.26	11.74	10.66	10.31
SD	3.06	3.66	3.39	3.83	4.17	4.84	3.85	4.52	4.55	4.19

RQ2: HOW DOES THE TIME VARIABLE INFLUENCE LEARNERS' SCORES RELATED TO THE TEXT DIFFICULTY? (H3)

RM Within-Subjects ANCOVAs (typing skills & proficiency as CVs)

Text pair	N	F	p	Part. Eta Squared	comment
1	199	**11.1190**	**<.001**	.063	Interaction with **both** CVs **not** significant (p=.063 for Prof; p=.076 for TS)
2	197	**15.195**	**<.001**	.073	Interaction with with Prof **significant** (p=.014); with TS **not** significant (p=.219)
3	197	**21.562**	**<.001**	.100	Interaction with with Prof **not** significant (p=.081); with TS **significant** (p<.001)
4	196	**5.115**	**.025**	.026	Interaction with **both** CVs **not** significant (p=.170 for Prof; p=.405 for TS)
5	196	**0.015**	**.902**	.000	Interaction with **both** CVs **not** significant (p=.378 for Prof; p=.275 for TS)

Interpretation & discussion RQ 1 & 2

RQ1:

- both C-Tests highly reliable; reliability values almost the same;
- lower reliability values for all prof. groups (homogeneity); lowest reliability of SC-Test for higher prof group (ability not captured; large degrees of error; but why C-Test lower?)

RQ2:

- scores increase with additional time; difference significant with TS & proficiency adjusted for
- increase consistent & statistically significant across two proficiency groups
- increase statistically significant for Texts 1-4 but not Text 5
- medium proficiency learners gain considerably more points with additional time than higher proficiency learners
- Possible mode effect (speed-ability trade-off)

RQ 3: WHICH COMPONENTS OF L2 PROFICIENCY (DECLARATIVE, PROCEDURAL KNOWLEDGE AND AUTOMATICITY) ARE BETTER PREDICTORS OF DIFFERENTLY TIMED C-TESTS?

Method: Linear regression analysis, SEM

Hypothesis 1: **Performance** on a canonical C-Test can be better **predicted** by measures of declarative and procedural knowledge, whereas performance on a speeded C-Test can be better **predicted** by measures of (procedural knowledge and) automaticity.

Hypothesis 2: A larger share of Declarative and Procedural Knowledge can be found in Slow Proficiency (**construct** measured by CT), whereas a larger share of Automaticity can be found in Fast Proficiency (**construct** measured by SCT).

Measures of declarative and procedural knowledge (RQ3)

	Test	Format	Construct	Source/Author
DK	**Vocabulary Size Test** (VST)	Match words to definitions (*untimed*)	Declarative (receptive) knowledge of vocabulary (breadth of vocabulary)	Institut für Testforschung und Testentwicklung e.V. Leipzig (Nation, 1990)
	Grammatical Acceptability Judgment Test (GAJT)	Decide whether sentences are grammatically acceptable or not (*untimed*)	Declarative (receptive) knowledge of grammar	ENG: DeKeyser (2000) & Lu (2010) - > GER/RUS: Drackert et al. (project)
	Grammar Correction Task (GCT)	Correct highlighted parts of sentences (*untimed*)	*Declarative(?)* (productive) knowledge of grammar	ungrammatical sentences from GAJT
PK	**Orthographic Choice Task** (OCT)	Decide whether words are spelled correctly or not (*timed*)	*Procedural(?)* (word-specific) knowledge of orthography	Drackert et al. (Olson et al., 1994)
	Self-Paced Reading Test (SPRT)	Read sentences part by part; answer questions about their content (distractors) and grammaticality (items) (*timed*)	Procedural (receptive) knowledge of grammar	versions of sentences used in GAJT (targeting same phenomena) (Marsden et al., 2017)
	Written Elicited Imitation Test (WEIT)	Reconstruct written stimuli in writing (*timed*)	Procedural integrated linguistic knowledge & skills	Drackert et al. (project); concept by AT

Measure of automaticity (RQ3)

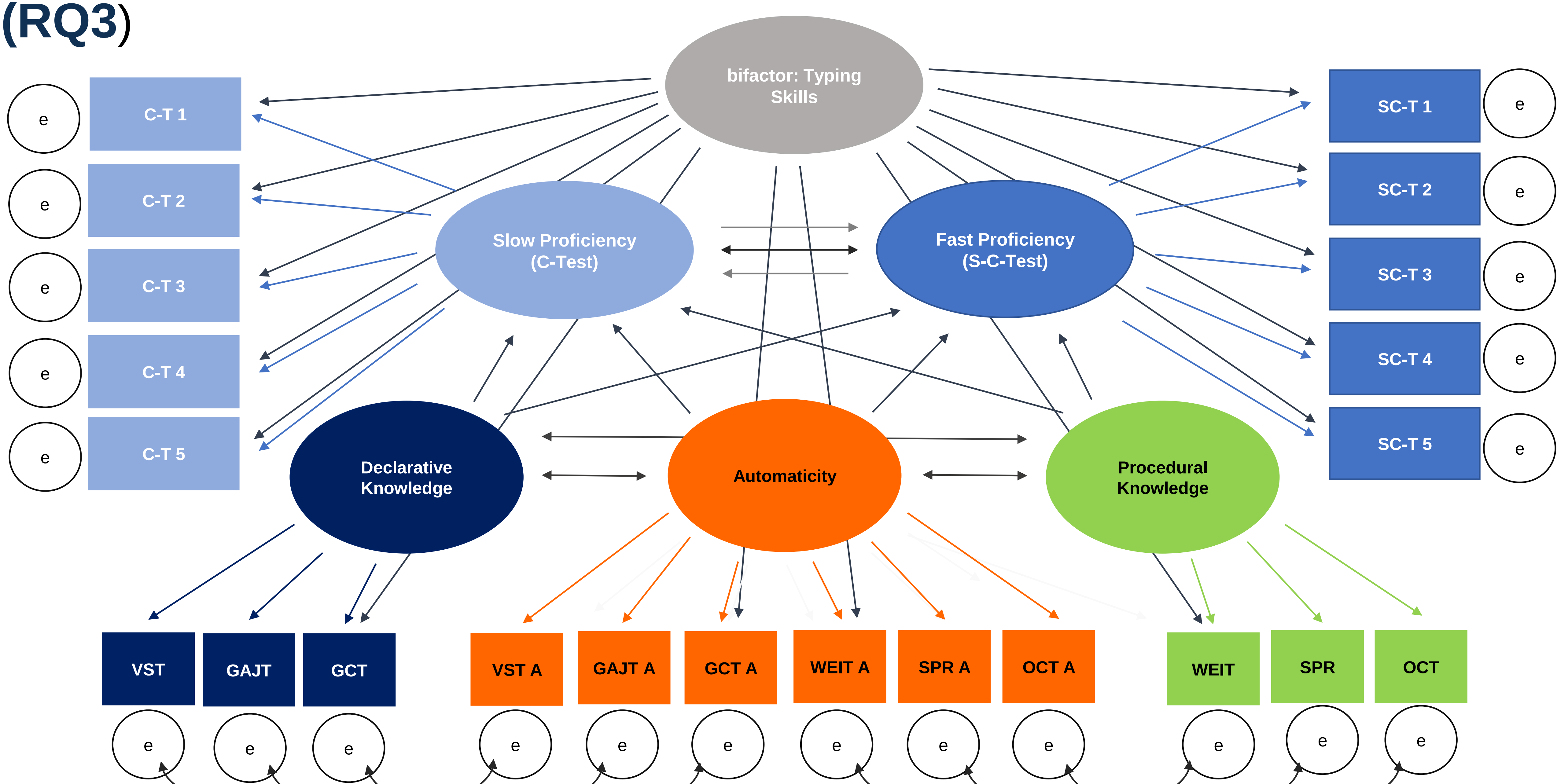
- processing speed -> reaction times for correctly solved items
- accuracy -> scores

total score on a test / mean reaction time for correctly solved items

Example:

ID	GAJT_score	GAJT_RT	GAJT_Automaticity
pe0103_03	62	2693	.023
pe0103_01	62	4648	.013
pe0103_01	52	13767	.004
pe2402_11	33	7310	.005

(RQ3)


bifactor: Typing Skills
C-T 1
C-T 2
C-T 3
C-T 4
C-T 5
SC-T 1
SC-T 2
SC-T 3
SC-T 4
SC-T 5
Slow Proficiency (C-Test)
Fast Proficiency (S-C-Test)
Declarative Knowledge
Automaticity
Procedural Knowledge
VST
GAJT
GCT
VST A
GAJT A
GCT A
WEIT A
SPR A
OCT A
WEIT
SPR
OCT
e

g.a.s.t.

Correlations between instruments (RQ3)

	VST	GAJT	GCT	OCT	SPRT	WEIT
VST	1.000	.773*	.688*	.628*	.692*	.712*
GAJT	.773*	1.000	.836*	.601*	.783*	.771*
GCT	.688*	.836*	1.000	.594*	.745*	.736*
OCT	.628*	.601*	.594*	1.000	.568*	.578*
SPRT	.692*	.783*	.745*	.568*	1.000	.716*
WEIT	.712*	.771*	.736*	.578*	.716*	1.000

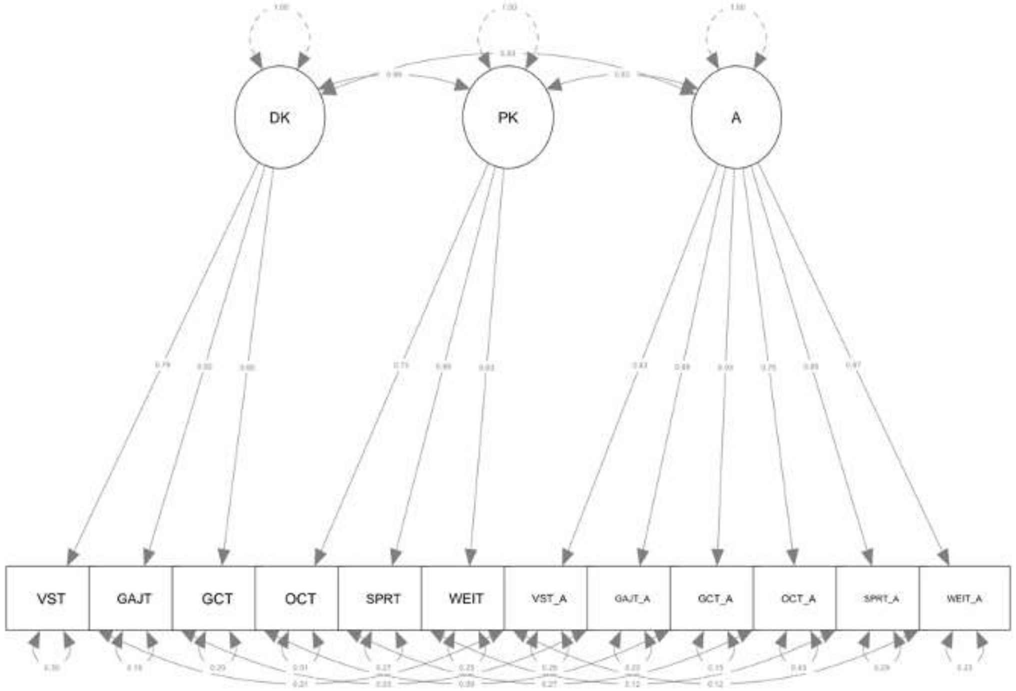
*significant (*p* <.001)

Initial Model-1 for CFA (3 factors: DK, PK & A)

Model estimation:

Estimator MLM (Satorra-Bentler due to non-normally distributed data)

- Chi square test: $\chi^2(45) = 153.260$, ***p* = .000** -> model does not perfectly mirror reality
- Robust **CFI: .959**; **TLI: .940** -> acceptable (Hu & Bentler, 1999)
- Robust **RMSEA: .110** -> not sufficient (Hu & Bentler, 1999; MacCallum et al., 1996)
- **SRMR: .047** -> acceptable (Hu & Bentler, 1999)

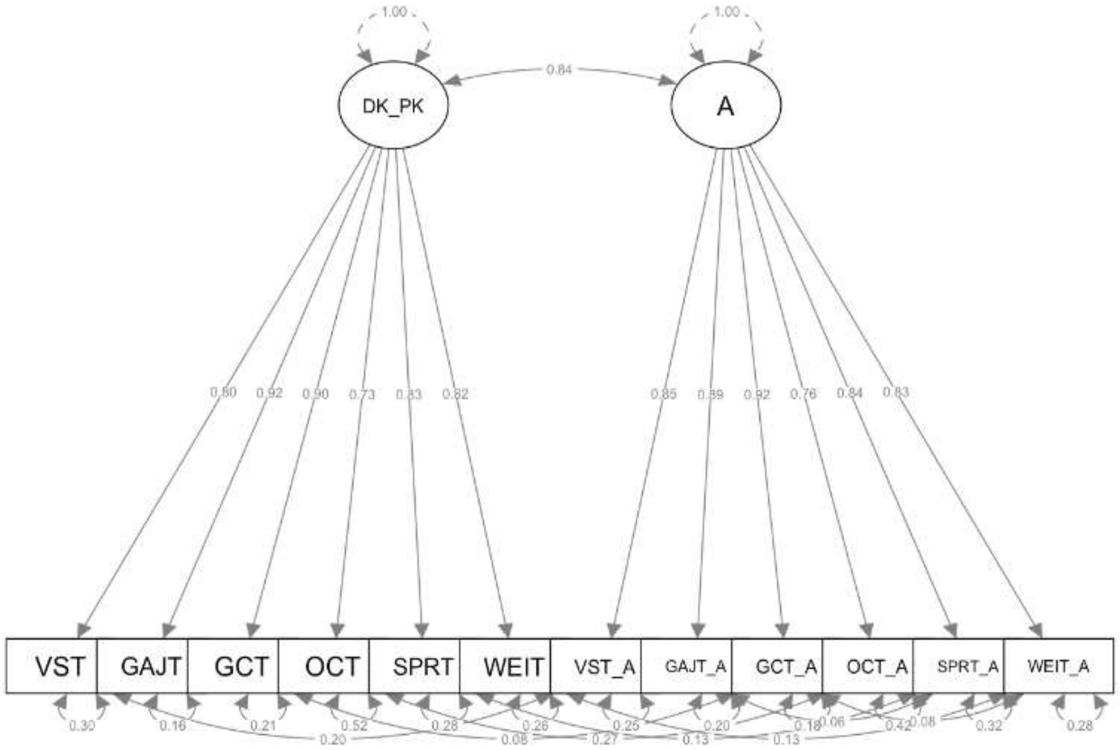


Respecified Model-1 for CFA (2 factors: DK/PK & A)

Model estimation:

Estimator MLM (Satorra-Bentler due to non-normally distributed data)

- Chi square test: $\chi^2(47) = 150.97$, ***p* < .001** -> model does not perfectly mirror reality
- Robust **CFI: .985; TLI: .941** -> acceptable (Hu & Bentler, 1999)
- Robust **RMSEA: .107** -> not sufficient (Hu & Bentler, 1999; MacCallum et al., 1996)
- **SRMR: .048** -> acceptable (Hu & Bentler, 1999)



Comparing the two models for DK, PK & A

```
Scaled Chi-Squared Difference Test (method = “satorra.bentler.2001”)

                   Df    AIC     BIC      Chisq    Chisq diff Df diff Pr(>Chisq)

fit_cfa_A_robust   45  4268.1  4376.9  153.26
fit_cfa_A1_robust  47  4264.4  4366.7  153.60       0.36088       2   0.8349
```

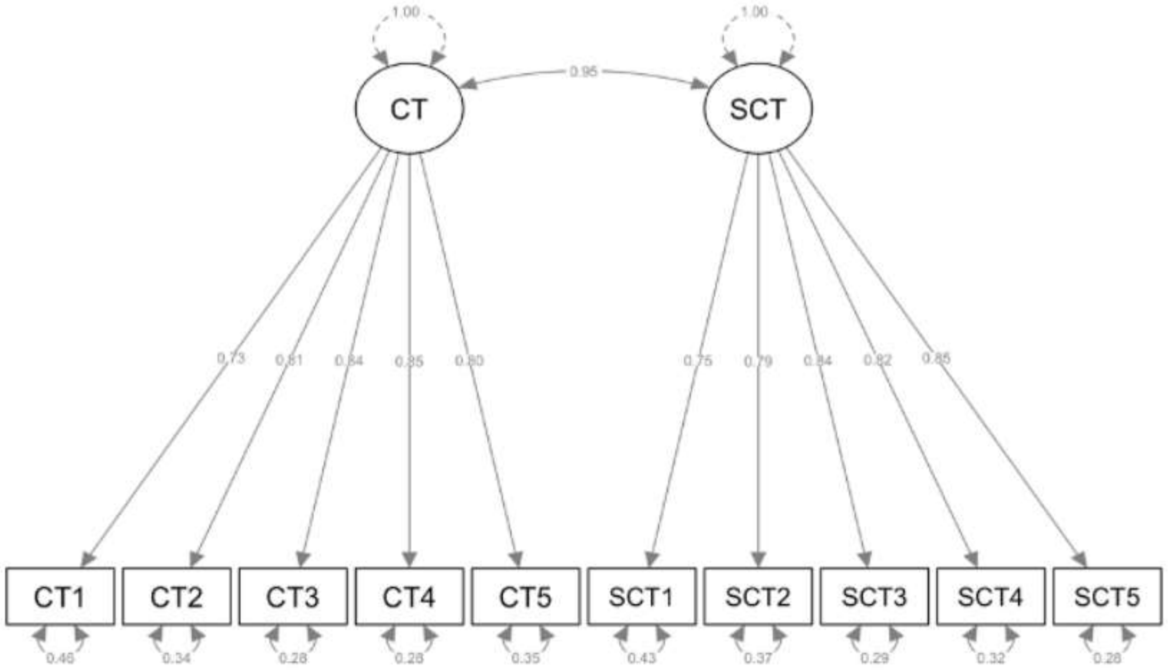
-> **Neither of the models (3-factor & 2-factor) fits better to the data than the other**

Initial Model-2 for CFA (2 factors: „Slow“ & „Fast“ Proficiency)

Model estimation:

Estimator MLM (Satorra-Bentler due to non-normally distributed data)

- Chi square test: $\chi^2(34) = 67.86$, ***p* < .001** -> model does not perfectly mirror reality
- Robust **CFI: .976**; **TLI: .968** -> good (Hu & Bentler, 1999)
- Robust **RMSEA: .073** -> acceptable following MacCallum et al. (1996) (interval lower = .047)
- **SRMR: .030** -> good (Hu & Bentler, 1999)



Loadings:
from 0.729 (CT1) to
0.846 (SCT5)

Summary and discussion of the results RQ 3 - SEM

DK & PK cannot be separated in our data collected with our instruments. Possible if:

- instruments separate insufficiently -> other instruments? (realistic?)
- alternative measure of automaticity (less correlated)?

  -> **coefficient of variation** (Segalowitz & Segalowitz, 1993)

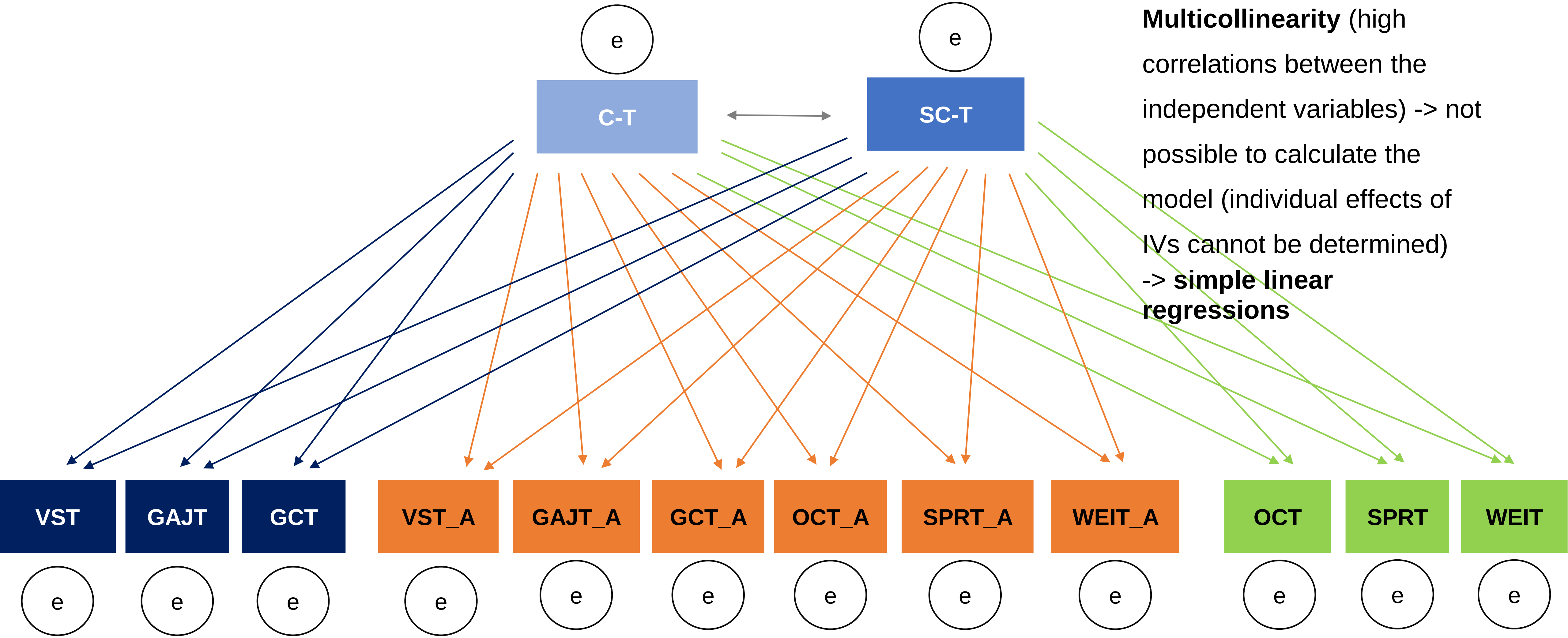
C-Tests load on **two factors** separating canonical and speeded **texts:**

- 1 factor (Global Language Proficiency) - worse fit than 2 factors (Slow & Fast Proficiency) as confirmed by Scaled Chi-Squared Difference Test (method = “satorra.bentler.2001”)
- Possible (to be checked): 2 factors Medium and High Proficiency

SEM to be continued (also with GER data)

CV = SD of all RTs of an individual divided by their mean (SD/Mean RT). Reveals **processing** variablity (**stability**). Can be used as a measure of automaticity when **analysed together with RT** data (if a **positive CV-RT correlation** found)*

Regression Model (observed level)



Multicollinearity (high correlations between the independent variables) -> not possible to calculate the model (individual effects of IVs cannot be determined)
-> **simple linear regressions**

g.a.s.t.

Overview regression C-Tests ~ DK & PK measures

	C-Test			***SC-Test***		
	R^2	Std. Err.	*p*	R^2	Std. Err.	*p*
VST	.463	6.9548e-02	.000*	.455	5.519e-02	.000*
GAJT	.584	4.819e-02	.000*	.616	4.630e-02	.000*
GCT	.527	5.141e-02	.000*	.567	4.916e-02	.000*
OCT	.275	6.364e-02	.000*	.355	6.004e-02	.000*
SPRT	.433	5.628e-02	.000*	.494	5.316e-02	.000*
WEIT	.579	4.850e-02	.000*	.638	4.497e-02	.000*

Overview regression C-Tests ~ Automaticity measures

	C-Test			***SC-Test***		
	R^2	Std. Err.	*p*	R^2	Std. Err.	*p*
VST_A	.336	6.092e-02	.000*	.402	5.781e-02	.000*
GAJT_A	.256	6.445e-02	.000*	.377	5.899e-02	.000*
GCT_A	.384	5.869e-02	.000*	**.499**	5.289e-02	.000*
OCT_A	.118	7.022e-02	.000*	.222	6.595e-02	.000*
SPRT_A	.339	5.824e-02	.000*	.459	5.498e-02	.000*
WEIT_A	.385	5.863e-02	.000*	**.569**	4.906e-02	.000*

Summary and discussion of the results RQ 3 - Regression

- All of the measures (scores on instruments and automaticity measures) predict the performance on both C-Test versions significantly
- Only instrument with higher R^2 for canonical C-Test: VST
- All automaticity measures with higher R^2 for SC-Test

Thank you!
Vielen Dank!
Спасибо!

drackert@gast.de

ENG

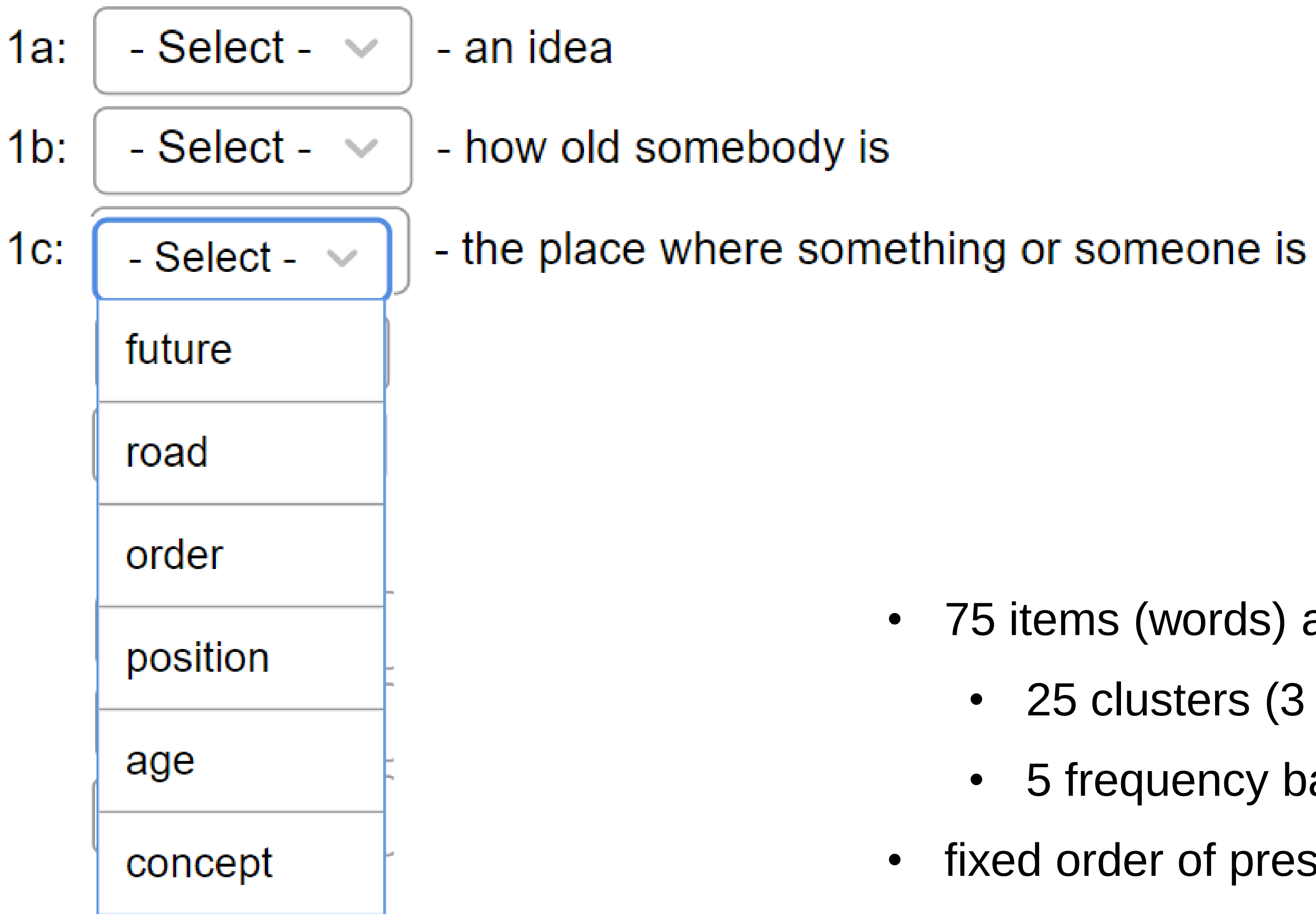
Predictors

- **GAJT:**
  - 62 grammatically correct or incorrect sentences to be judged by button response
  - acceptable – not acceptable – I don’t know
  - 20 sec time limit

- **WEIT:**
  - 20 sentences presented one by one on the computer screen for 2 to 6 seconds (depending on the length of the sentence)
  - after 2.5 sec pause, participants have to repeat the sentence by typing on the keyboard
  - max. response time: 30 sec

- **GCT:**
  - 32 ungrammatical sentences (from GAJT) to be corrected by participants (text box)
  - parts of the sentence are highlighted (mistake included)
  - 40 sec response window

VST (Vocabulary Size Test) ENG



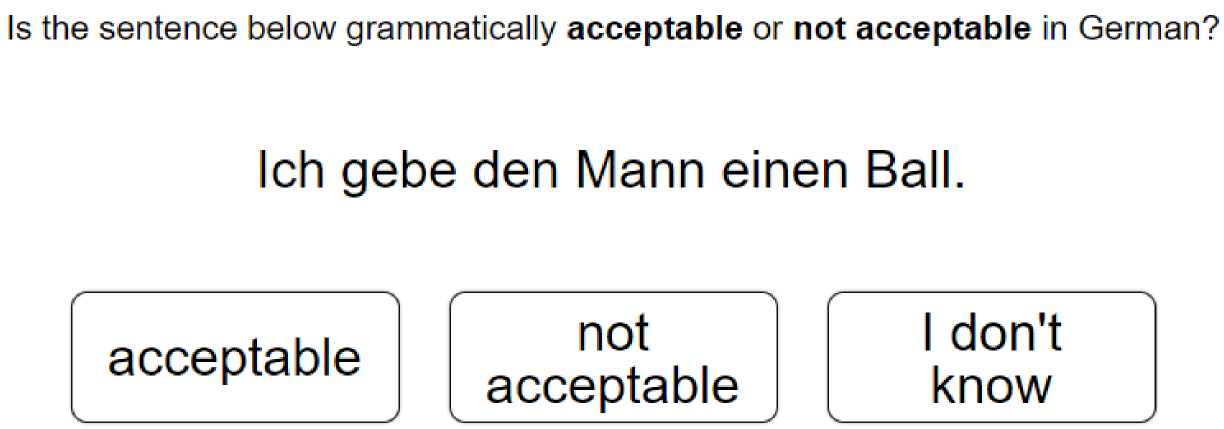
1a: - Select - - an idea

1b: - Select - - how old somebody is

1c: - Select - - the place where something or someone is

- 75 items (words) arranged in
  - 25 clusters (3 targets with 3 definitions + 3 distractors);
  - 5 frequency bands;
- fixed order of presentation

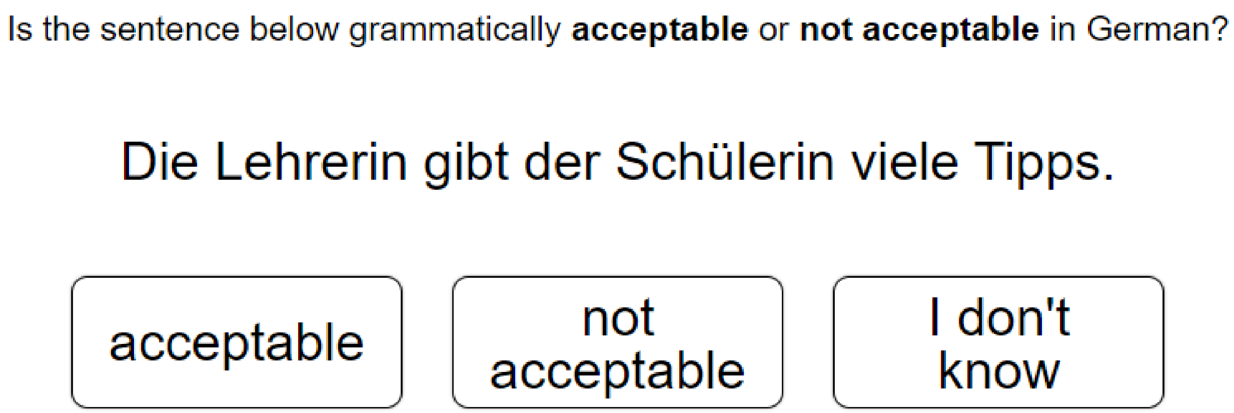
GAJT (Grammatical Acceptability Judgment Test) GER



Is the sentence below grammatically **acceptable** or **not acceptable** in German?

Ich gebe den Mann einen Ball.

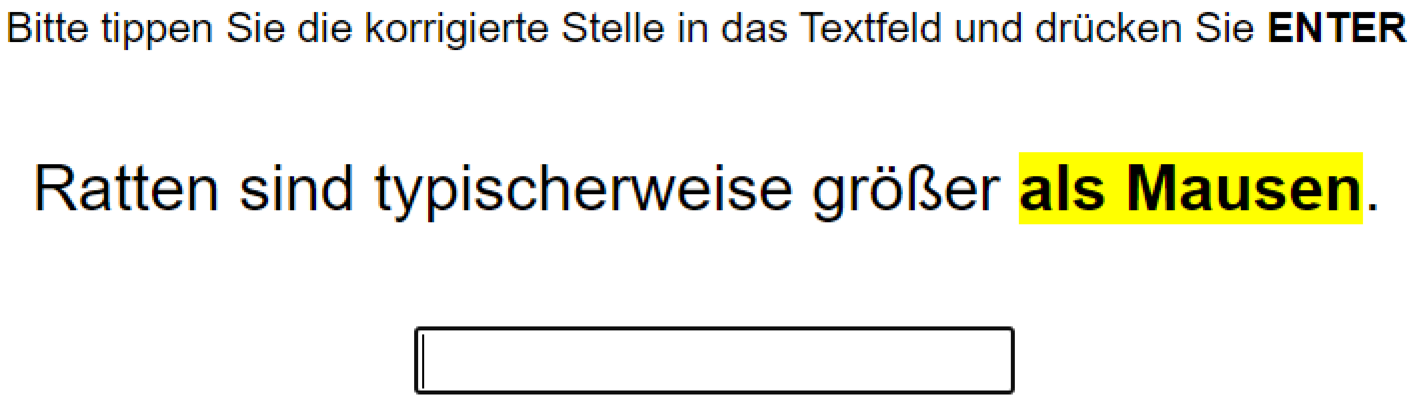
- 72 - 86 items;
- pairs of grammatical / ungrammatical sentences
- randomized order of presentation



Is the sentence below grammatically **acceptable** or **not acceptable** in German?

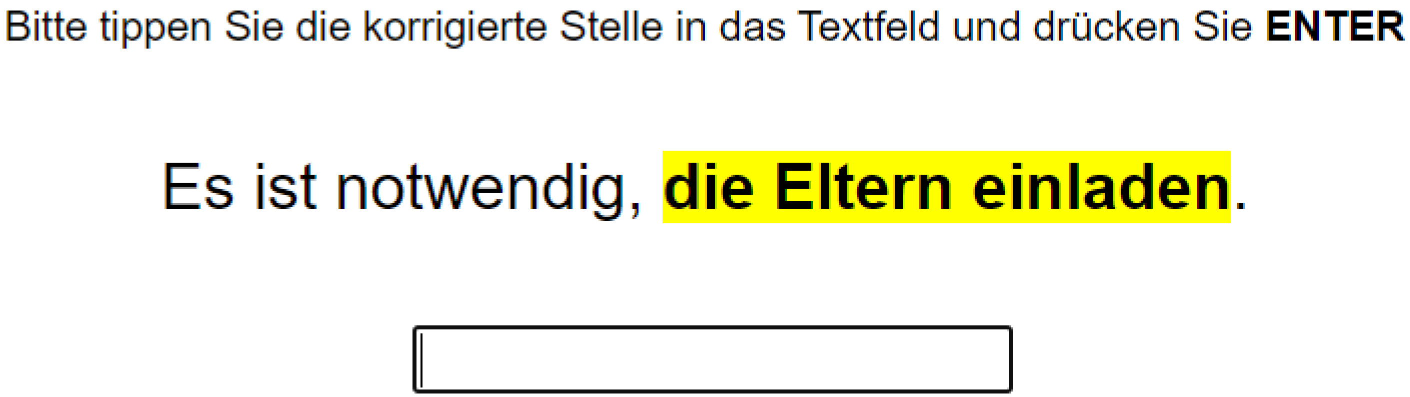
Die Lehrerin gibt der Schülerin viele Tipps.

GCT (Grammar Correction Task) GER


Bitte tippen Sie die korrigierte Stelle in das Textfeld und drücken Sie **ENTER**

Ratten sind typischerweise größer **als Mausen**.

- 35 - 36 items (ungrammatical sentences from GAJT)
- randomized order of presentation

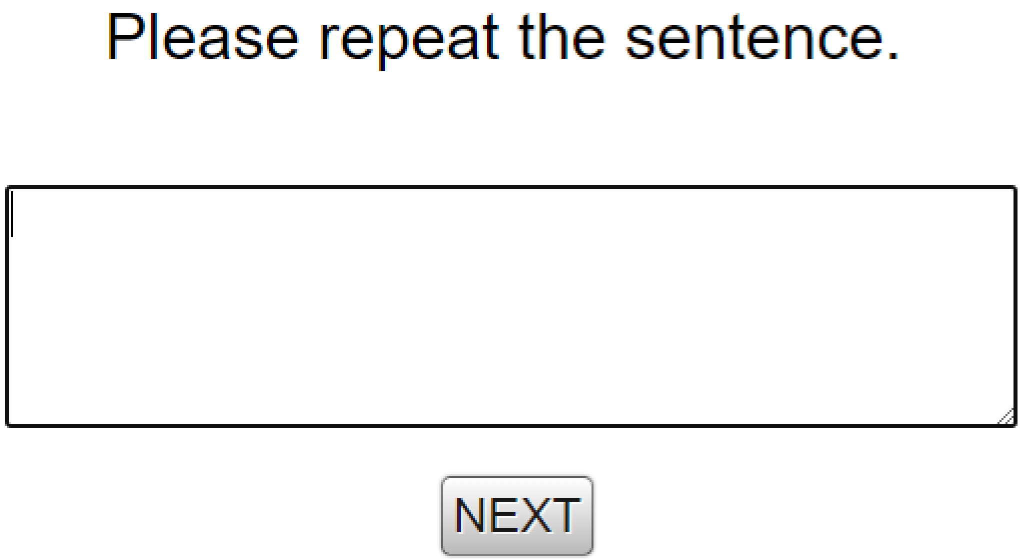

Bitte tippen Sie die korrigierte Stelle in das Textfeld und drücken Sie **ENTER**

Es ist notwendig, **die Eltern einladen**.

WEIT (Written Elicited Imitation Task) ENG

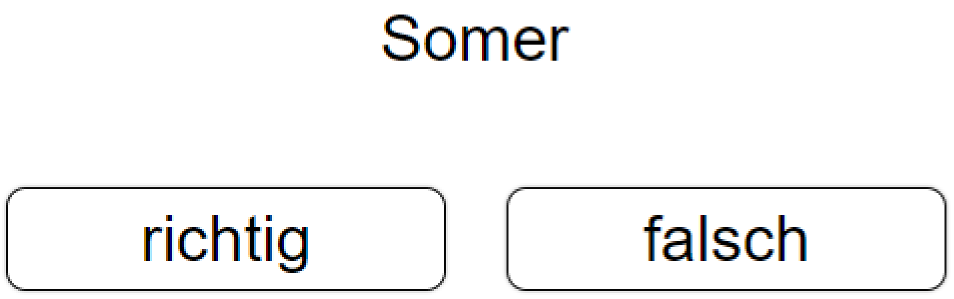
- newly developed EIT format
- 20 items (sentences)
- fixed order from shorter to longer sentences

The streets in this city are wide.



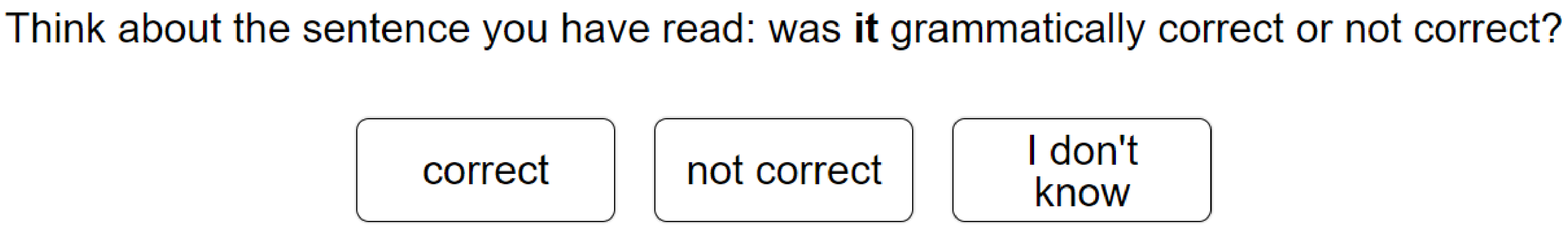
Please repeat the sentence.

OCT GER



SPRT ENG

I will buy new furnitures for my new apartment.



Think about the sentence you have read: was **it** grammatically correct or not correct?