

Katharina Karges

Modelling Item Difficulty in Several Languages

**On the Equivalence of Reading
Comprehension Items in a
Translated Foreign Language
Assessment**



DOI : <https://doi.org/10.51363/unifr.lth.2026.064>

Katharina Karges  <https://orcid.org/0009-0001-7205-0096>

© Katharina Karges, 2026



This work is published under a Creative Commons Attribution 4.0 International License (CC BY-NC-SA 4.0): <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Universität Freiburg (Schweiz)

Modelling Item Difficulty in Several Languages

On the Equivalence of Reading Comprehension Items
in a Translated Foreign Language Assessment

Katharina Karges

Heimatort: Wilsdruff, Deutschland

Dissertation zur Erlangung der Doktorwürde
an der Philosophischen Fakultät
der Universität Freiburg (Schweiz)

Genehmigt von der Philosophischen Fakultät
auf Antrag der Professorinnen und Professoren
Thomas Studer, Katrin Wisniewski und Tineke Brunfaut

Freiburg, den 20. Februar 2026
Prof. Dominik Schöbi, Dekan

Für Anna, Charlie, Emma, Gabriel, Ginny, Jakob, Mila und Käthe.
Möge die Komplexität unserer Welt für euch immer ein endloser Quell von Faszination sein.

Danksagung

Das Erstellen einer Dissertation kann eine ausgesprochen einsame Angelegenheit sein. Viel zu viel Zeit verbringt man am eigenen Schreibtisch und vor allem im eigenen Kopf. Umso mehr bin ich all den Menschen dankbar, die mich im Laufe der vergangenen Jahre mit so viel Zuneigung, Geduld und sogar Enthusiasmus dabei unterstützt (und dazu gebracht) haben, diese Arbeit tatsächlich zu beenden.

Am meisten davon, und noch viel mehr, hatten mein Mann, Patrick Karges, meine Eltern, Katrin und Torsten Berger, mein Bruder, Christoph Berger und meine beste Freundin, Malgorzata Barras. Ich danke euch für all die Wochenenden, an denen ich trotz Diss mit euch Zeit verbringen durfte (und die, an denen es deshalb nicht ging), all die (virtuellen) Kaffees, Mittagessen und Kuchenstücke, die ihren Weg auf meinen Schreibtisch gefunden haben, all die Nerven, die ihr ertragen habt, und die Tränen, die ihr getrocknet habt. Ohne euch wäre das alles sinnlos.

Absolut bedeutend für diese Arbeit war auch Katrin Wisniewski, die nicht locker gelassen und mir alle Möglichkeiten eingeräumt hat, um dieses Projekt tatsächlich zu realisieren. Dein Vertrauen, deine Beratung und HALL haben das alles überhaupt erst möglich gemacht. Hier ist nun das Ergebnis. Ich hoffe, ich konnte all dem wenigstens ein bisschen gerecht werden.

Thomas Studer hat diese Diss in all ihren Phasen unterstützt und an wichtigen Stellen die richtigen Fragen gestellt. Dabei hatte er immer alle Geduld der Welt, bis ich die Arbeit nun endlich einreichen konnte. Ich freue mich auf all die Punkte, über die wir in der Zukunft noch diskutieren können.

Einige Menschen haben ihre unmittelbaren Spuren in dieser Arbeit hinterlassen. Allen voran haben sich Anna Becker und Nina Hicks immer wieder Zeit genommen, um Kognate, Verneinungen und NIs zu identifizieren, über Variablen zu spekulieren und Operationalisierungen festzuzurren. Auch Abstracts, Präsentationen und Poster haben sie unermüdlich kommentiert. Vor allem hatten sie aber immer die richtigen Worte für mich übrig, um mir Mut zu geben. Eure Zuversicht hat mich weit gebracht und die Ergebnisse all unserer Diskussionen sind hoffentlich in dieser Arbeit zu sehen. Nicola Störr und Katia Rey haben ebenfalls alle Items begutachtet, Kognate und NIs gesucht und es mir so ermöglicht, die Variablen auf eine solidere Basis zu stellen. Sarah Graul hat sich durch eine frühe Fassung der Arbeit gekämpft und ist mir auch sonst an vielen kleinen und grossen Momenten zur Seite gestanden.

In einer frühen Phase der Operationalisierung hat sich Detmar Meurers viele Stunden Zeit genommen, um schwierige Fragen zu diskutieren und nützliche Literatur zu empfehlen.

Christine Czinglar hat mir dann während der Analysephase die Möglichkeit gegeben, mich an ihren Forschungsdaten auszuprobieren und so wichtige Werkzeuge und Methoden auszuwählen, was für meine Arbeit wegweisend war.

Die Kolleginnen und Kollegen aus HALL haben mich in zahlreichen Diskussionen an ihrer Expertise und ihrer Neugierde teilhaben lassen und so wichtige Grundlagen für diese Arbeit mitgeschaffen. Auch in vielen anderen Kolloquien, Workshops, SIGs und an Konferenzen wurde ich von freundlichen Zuhörenden immer wieder durch interessierte Nachfragen und enthusiastisches Feedback in meinem Vorhaben bestärkt.

Ich danke auch meinen ehemaligen Kolleginnen und Kollegen der Firma GITG, die mich an meinen Diss-Tagen in Ruhe gelassen haben und an den anderen Tagen immer wussten, wann sie nach meinen Fortschritten fragen können und wann lieber nicht.

Jan Vanhove hat nur dann nach der Diss gefragt, wenn ich ihn darum gebeten habe. Stattdessen hat er mich auf die wunderbarsten Wanderungen mitgenommen und im Scrabble zuverlässig besiegt. Ich weiss es sehr zu schätzen.

Die Spuren, die Peter Lenz in meiner Doktorarbeit hinterlassen hat, sind schliesslich kaum zu überblicken. Er hat die Forschungsprojekte und die Aufgabenentwicklung geleitet, aus denen die hier analysierten Aufgaben entstanden sind. Die Daten konnte ich nur dank seiner Unterstützung erhalten. Während unserer Zusammenarbeit am Institut für Mehrsprachigkeit hat er mich bei meinen ersten statistischen Analysen an der Hand gehalten, alle meine Fragen beantwortet und meine Arbeit jahrelang mit kritischen Anmerkungen besser gemacht. Ich habe mich nie eingeengt gefühlt, und ich habe so viel gelernt!

13.08.2025

Zusatz für die publizierte Version

Die Verteidigung dieser Arbeit am 20.02.2026 war ein grosser Moment, an dem viele Leute Anteil hatten. Ich danke allen, die dabei waren, lange Wege auf sich genommen haben, Daumen gedrückt haben und sich mit mir gefreut haben.

Allen voran haben mich Malgorzata, Frédéric und Gabriel Barras beherbergt, beschäftigt und vor allem beruhigt. Ich durfte die Woche vor meiner Verteidigung ganz in Ruhe und entspannt durch's Leben gehen – Worte können nicht beschreiben, wie viel wert das ist!

Auch bei Proben der Verteidigung haben sich viele Menschen Zeit genommen, um mich zu unterstützen. Allen voran hat mir Malgorzata Barras den entscheidenden Tipp gegeben, und Zorana Sokolovska, Audrey Bonvin und Nina Hicks hatten wichtige Ideen, die mir ungemein dabei geholfen haben, einen vernünftigen und vor allem verständlichen Vortrag zu halten. Jan Vanhove hat sich noch extra Zeit genommen und mich zu besseren Grafiken und genaueren Formulierungen überredet. Ruth Kessler, Anna Becker und Christine Renker haben schon eher wichtige Hinweise gegeben, die ich gern umgesetzt habe.

Nicht zuletzt sei meinen Gutachtenden gedankt: Thomas Studer, Katrin Wisniewski und Tineke Brunfaut sowie Raphael Berthele stellten viele nützliche Fragen und gaben wichtige Hinweise zu meiner Arbeit, die ich in diese publizierte Version einfliessen lassen konnte. Sie ist dadurch noch besser geworden. Alle übriggebliebenen Fehler und Schwächen sind meine eigenen.

21.03.2026

Contents

1	Introduction	18
1.1	Context of the study	19
1.1.1	The ÜGK foreign language assessment	19
1.1.2	Item difficulty and test translation research	21
1.2	Research questions	22
1.3	Overview	22
I	Literature review	24
2	Important concepts and terminology	25
2.1	Validity and validation	25
2.2	Languages	26
2.3	Competences	27
2.4	Assessment	29
2.4.1	Educational measurement	29
2.4.2	Tasks and items	30
2.4.3	Question, stem and answer options	30
2.5	Translation	31
3	Reading comprehension	32
3.1	Models of reading	34
3.1.1	The Construction-Integration Model	34
3.1.2	The Simple View of Reading	35
3.1.3	The Reading Systems Framework	36
3.1.4	The Cognitive Model of Reading for assessment	38
3.1.5	Conclusion	40
3.2	Components of reading	40
3.2.1	Word recognition	41
3.2.2	Word identification	42
3.2.2.1	Semantic knowledge	42
3.2.2.2	Lexical knowledge	43
3.2.2.3	Morphological knowledge	43
3.2.3	Sentence-level and discourse comprehension	44
3.2.3.1	Building the text model	44
3.2.3.2	Building the situation model	45
3.2.4	Metalinguistic awareness	46
3.2.5	Conclusion	46
3.3	Reading in an additional language	47
3.3.1	Cross-linguistic transfer in L2 reading	48
3.3.1.1	Linguistic knowledge	48

3.3.1.2	Higher-order skills	49
3.3.2	Reading at a lower proficiency level	49
3.3.3	Conclusion	50
3.4	Summary	51
4	Measuring reading	53
4.1	The input text	54
4.1.1	Text organisation	54
4.1.2	Topic and text genre	55
4.1.3	Linguistic complexity	56
4.1.3.1	Defining linguistic complexity	56
4.1.3.2	Measuring linguistic complexity	57
4.1.4	Conclusion	59
4.2	Task formats	60
4.3	The test taker	62
4.4	Conclusion	63
5	Measuring the lexical characteristics of reading texts	65
5.1	The role of lexical knowledge in reading comprehension: empirical findings	65
5.2	Lexical characteristics of reading texts	67
5.2.1	Lexical diversity	68
5.2.2	Lexical density	69
5.2.3	Lexical sophistication	70
5.2.3.1	Lexical frequency	71
5.2.3.2	Lexical difficulty	72
5.3	Measuring cross-linguistic characteristics of reading texts	74
5.3.1	Cross-linguistic transfer	74
5.3.2	Cognates	75
5.3.3	Operationalising transfer in a reading assessment	76
5.4	Conclusion	77
6	(Item) Difficulty: definition and measurement	78
6.1	A critical look at item difficulty	78
6.2	Item analyses in psychometrics	79
6.2.1	Item difficulty and item discrimination	79
6.2.1.1	Classical Test Theory	80
6.2.1.2	Item Response Theory	82
6.2.1.3	Other methods of estimating item difficulty	84
6.2.2	Item fit	84
6.2.2.1	Item fit statistics in IRT	85
6.2.2.2	Differential item functioning	85
6.3	Explanatory item response modelling	88
6.4	Conclusion	89
7	Research on multilingual assessments	90
7.1	Translated foreign language assessments	90
7.2	Cross-cultural foreign language assessment	91
7.3	Translated assessments and surveys	93
7.3.1	Methods of test translation	94
7.3.2	Measuring the effect of test translation	95
7.3.3	Empirical research on translated assessments and surveys outside education	97
7.4	Empirical research on translated educational assessments	99

7.5	Conclusion	102
8	Research on item difficulty	103
8.1	Research designs and statistical methods	104
8.1.1	Correlation, regression modelling and related methods	104
8.1.2	Factor analysis	107
8.1.3	Cognitive Diagnostic Assessment	107
8.1.4	Regression trees and artificial neural networks	108
8.1.5	Summary	110
8.2	Variables in item difficulty research	110
8.2.1	Sources of variables	110
8.2.2	Text, item and text-by-item variables	111
8.2.3	The Necessary Information	112
8.2.4	Linguistic variables	113
8.2.5	Variables related to the assessment of reading comprehension	116
8.2.5.1	Task format	116
8.2.5.2	Position effects	117
8.2.5.3	Overlap between input text and item	117
8.2.5.4	MCQ: plausibility and falsifiability of the answer options	118
8.2.5.5	Comprehension processes	119
8.2.6	Test-taker variables	120
8.2.7	Summary	120
8.3	Results of item difficulty research	121
8.3.1	Primarily linguistic modelling	121
8.3.2	Modelling response processes	123
8.3.3	Methodological studies	125
8.3.4	Test taker characteristics	125
8.3.5	Languages other than English	126
8.3.6	Summary	127
8.4	The role of other languages in foreign language assessment	129
8.5	Conclusion	131
II	Presentation of the study	132
9	Data	133
9.1	Context	133
9.1.1	Foreign language learning in Switzerland	133
9.1.2	ÜGK	134
9.1.3	The test takers	136
9.2	The foreign language assessment	137
9.2.1	The construct	137
9.2.2	The tasks	138
9.2.2.1	The items	141
9.2.2.2	The input texts	141
9.2.2.3	The language versions	142
9.2.2.4	The language of schooling in foreign language assessment	143
9.2.3	Other components of the ÜGK 2017 assessment	143
9.2.4	Item development	144
9.3	The variables	145
9.3.1	Linguistic variables	151
9.3.1.1	Descriptive variables	151

9.3.1.2	Lexical diversity and lexical density	152
9.3.1.3	Lexical sophistication and difficulty	153
9.3.1.4	Cross-linguistic variables	156
9.3.2	Text variables	157
9.3.3	Item and text-by-item variables	158
9.3.3.1	Item variables	158
9.3.3.2	Text-by-item variables	159
9.3.3.3	Identification of the NI	160
9.3.3.4	Overlap variables: operationalising multilingual test items	161
9.3.4	Test taker variables	162
9.3.5	The dependent variable	163
10	Analysis	164
10.1	Data preparation and variable extraction	164
10.1.1	Item data	165
10.1.2	Item response data	166
10.1.3	Test taker data	167
10.2	Statistical analyses	167
10.2.1	Descriptive statistics	168
10.2.2	Inferential statistics	169
10.2.2.1	Interactions between variables	169
10.2.2.2	Item difficulty prediction: correlational analyses	170
10.2.2.3	Item difficulty prediction: regression analyses	170
10.2.2.4	Item difficulty prediction: further exploratory statistics	171
III	Results	172
11	Descriptive statistics	173
11.1	Text, item and text-by-item variables	173
11.1.1	Categorical variables	174
11.1.2	Linguistic variables: descriptives	174
11.1.2.1	Input texts and NIs	174
11.1.2.2	Questions and answer options	176
11.1.3	Lexical variables	177
11.1.3.1	Input texts	177
11.1.3.2	NIs, questions and answer options	180
11.1.4	Cross-linguistic variables	180
11.1.4.1	Input texts and NIs	180
11.1.4.2	Items	181
11.1.5	Overlap variables	181
11.1.6	Changes to the variables	184
11.2	Item parameters	184
11.2.1	Response patterns	184
11.2.2	CTT item parameters	187
11.2.3	IRT item parameters	188
11.2.4	DIF across language versions	189
11.2.5	Item difficulty across language versions	191
11.2.6	Conclusion	191
11.3	Person attributes	191
11.3.1	Performance at the assessment	192
11.3.2	Test taker variables	192

11.3.3	Test taker variables in the IRT model	193
11.4	Summary	193
12	Inferential statistics	195
12.1	Interactions between variables	195
12.1.1	Categorical variables	196
12.1.2	Linguistic characteristics of the input text	198
12.1.2.1	Lexical diversity	199
12.1.2.2	Lexical sophistication	199
12.1.2.3	Lexical difficulty	199
12.1.2.4	Cross-linguistic variables	200
12.1.3	Linguistic characteristics of the NIs	201
12.1.4	Linguistic characteristics of the items	201
12.1.5	Overlap variables	202
12.2	Data reduction	203
12.2.1	Factor analyses of linguistic characteristics of input texts	203
12.2.2	Cluster analyses of POS variables of NIs and item parts	204
12.3	Prediction of item difficulty	205
12.3.1	Correlations with item difficulty	205
12.3.2	Correlations with mean task difficulty	208
12.3.3	Regression modelling	208
12.3.3.1	Item difficulty prediction in the English test	208
12.3.3.2	Item difficulty prediction in the German test	209
12.3.3.3	Item difficulty prediction in the French test	209
12.3.3.4	Conclusion	211
12.3.4	Multilingual prediction of item difficulty	211
12.3.4.1	Crosslinguistic validation	211
12.3.4.2	Mixed-effects modelling	211
12.3.5	Triangulation: exploratory random forest modelling	213
12.3.6	Beyond assessment characteristics: Bayesian regression	215
12.4	Summary	218
IV	Discussion and conclusion	220
13	Discussion	221
13.1	Summary of the main findings	221
13.1.1	RQ 1: Prediction of item difficulty in foreign language reading assessment	222
13.1.2	RQ 2: Prediction of item difficulty in foreign language reading <i>across</i> languages	223
13.1.3	RQ 3: Test taker variables in test response prediction	225
13.2	Additional findings	226
13.2.1	Transferability of findings from research on English to other languages	226
13.2.2	Measuring lexical diversity	227
13.2.3	Using reference corpora: measuring lexical sophistication and difficulty	228
13.2.4	L2 reading as <i>reading with several languages</i>	230
13.2.5	Multilingual assessment	230
13.2.6	Is L2 reading a language problem or a reading problem?	230
13.2.7	Additional open questions	231
13.3	Limitations of the study	232
13.3.1	Limited item sample	232
13.3.2	Variables	233

13.3.3	Young language learners	233
13.3.4	Sociocultural aspects	234
13.3.5	Methodological shortcomings	234
13.3.6	Artificial intelligence	235
13.4	Implications	235
14	Conclusion	238
	References	240
	Glossary	268
A	Use of Artificial Intelligence	273
A.1	Erklärung zur Nutzung von künstlicher Intelligenz	273
A.2	Declaration on the Use of Artificial Intelligence	273
B	Additional results	275
B.1	Descriptive analyses	275
B.1.1	Correlations of incidences of POS across language versions	275
B.1.2	Descriptive statistics of lexical variables in input texts	276
B.1.3	Correlations of lexical variables and token counts across languages	279
B.1.4	Lexical difficulty and other linguistic characteristics of input texts	280
B.1.5	Lexical diversity: MTLT across input texts and language versions	281
B.1.6	MFR: model fit	282
B.1.7	Wright Maps: Item difficulties per language version	283
B.1.8	Factor analyses of the linguistic characteristics of the input texts	286
B.1.8.1	FRit language version	286
B.1.8.2	FRde language version	286
B.1.8.3	DEfr language version	287
B.2	Inferential statistics	287
B.2.1	Regression models of French test data	287
B.2.1.1	Final model based on FRit data	287
B.2.1.2	Final model based on FRde data	287
B.2.2	Decision tree based on random forests analysis	288
B.2.3	Comparison of item difficulties from Bayesian and parametric IRT models	290
C	Electronic appendix: R project	291
C.1	Access to electronic appendices	291
C.2	File index	291
C.3	Overview of tables	300
C.3.1	Original item data	300
C.3.2	Original response data	301
C.3.3	Original test taker data	302

List of Tables

1.1	Overview of the test versions in the foreign language component of ÜGK 2017 . . .	20
4.1	Linguistic domains which are assessed in Coh-Metrix and CTAP	59
5.1	Operationalisation of subconstructs of lexical diversity according to Jarvis (2013) and Vanhove et al. (2019)	69
5.2	Levenshtein distance of German <i>Geister</i> and English <i>ghosts</i>	76
8.1	The most relevant predictors of item difficulty identified in the literature	128
9.1	ÜGK 2017 administration: schedule	135
9.2	Number of test takers per language version	136
9.3	Overview of reading tasks in ÜGK L2 reading assessment	139
9.4	Overview of predictor variables, hypothesised impact of item difficulty and relation to construct	150
9.5	Operationalisation of lexical diversity measures	154
9.6	Operationalisation of lexical rarity measures	155
9.7	Operationalisation of text variables	157
9.8	Operationalisation of text-by-item variables	159
10.1	Filename conventions for data analysis	164
11.1	Descriptive statistics of categorical item variables	175
11.2	Text length in tokens of input texts and NIs	176
11.3	Significance of difference between means of lexical variables in input text across foreign languages	178
11.4	Descriptive statistics of proportion correct by language version	187
11.5	Descriptive statistics of MFR item difficulties by language version	188
11.6	Descriptive statistics of MFR person parameters by language version	192
12.1	Correlations of POS_NAM_Q with linguistic characteristics of input texts . . .	202
12.2	Factor loadings from EFA on linguistic characteristics of ENde input texts . . .	204
12.3	Correlations of item and text variables with item difficulty	206
12.4	Linear regression model predicting ENde item difficulty	209
12.5	Linear regression model predicting DEfr item difficulty	210
12.6	RMSEs from crosslinguistic model tests	211
12.7	Summary of mixed-effects linear regression model on multilingual data	212
12.8	Summary of Bayesian multilevel logistic regression with covariates	216
12.9	Summary of Bayesian multilevel logistic regression with interactions	217
B.1	Descriptive statistics of lexical variables in input texts per language version . . .	278
B.2	Factor loadings from EFA on linguistic characteristics of FRit input texts . . .	286
B.3	Factor loadings from EFA on linguistic characteristics of FRde input texts . . .	286

B.4	Factor loadings from EFA on linguistic characteristics of DEfr input texts	287
B.5	Linear regression model predicting FRit item difficulty	287
B.6	Linear regression model predicting FRde item difficulty	288
C.1	File list of electronic appendix	299
C.2	Description of the data file 1_a_ItemData.xlsx	301
C.3	Description of the data file 1_b_ResponseData.csv	302

List of Figures

3.1	Profiles of readers proposed by the SVR model	36
3.2	Perfetti & Stafura’s Reading Systems Framework	38
3.3	Khalifa & Weir’s cognitive model of reading	39
4.1	Lexical complexity at three levels of construct specification	57
6.1	A typical item characteristic curve under the Rasch model	82
6.2	A hypothetical 4-PL item characteristic curve, annotated	83
6.3	A hypothetical item that displays uniform DIF, annotated	86
6.4	A hypothetical item that displays substantial non-uniform DIF, annotated	87
6.5	Overview of four explanatory item response models	88
7.1	Percent of items with at least moderate DIF in the PISA 2000 Reading data	100
8.1	Embretson & Wetzel’s Information-Processing Model for Evaluating the Response Alternatives	118
9.1	Map of Switzerland indicating the first foreign language taught at public schools	135
9.2	Reading item LV03E_dE_1, annotated	140
9.3	Reading item LV03E_dE_2	140
9.4	Reading item LV08G_dE_1, annotated	141
9.5	Screenshot of additional questions about reading item LV03E_dE_1	160
9.6	Partial screenshot of NI identification for reading item LV03E_dF_1	161
11.1	Lexical difficulty variables across languages	179
11.2	Cognate variables across language versions: human ratings and mean Levenshtein-Damerau distance	182
11.3	Overlap of correct answer option with NI, compared to overlap of incorrect answer option with NI and input text (FRde)	183
11.4	Typical response patterns of a multiple-choice item	185
11.5	Graphical depiction of DIF: facet item difficulties per language version	190
11.6	Item difficulty across language versions	191
12.1	Heat map of correlations of all variables in ENde language version	196
12.2	Scatterplot of variables ContinuousText, ItemType and TokenCount of input texts	197
12.3	Scatterplot of variables ContinuousText, POS_N and TokenCount of input texts	198
12.4	Scatterplots of cross-linguistic variables and the proportion of verbs in the input texts	201
12.5	Scatterplots of item difficulty and lexical sophistication of input texts	207
12.6	Comparison of fitted and actual effects of linear regression on ENde data set	210
12.7	Plot of random effects of mixed-effects linear regression model based on multilingual data set	213
12.8	Variable importance plot of random forest regression model	214

B.1	Correlations of incidences of parts of speech in input texts	276
B.2	Correlations of token count and lexical variables in NIs	279
B.3	Scatterplots of CEFRlex lexical difficulty variables, POS_N and ContinuousText across languages	280
B.4	Scatterplots of Decaleon lexical difficulty variables, POS_N and ContinuousText across languages	280
B.5	Boxplots of MTLDS of input texts	281
B.6	MTLD by input text and language version	282
B.7	Wright Map of item difficulties in language version DEfr	284
B.8	Wright Map of item difficulties in language version ENde	284
B.9	Wright Map of item difficulties in language version FRde	285
B.10	Wright Map of item difficulties in language version FRit	285
B.11	Single tree based on random forest regression modelling	289
B.12	Comparison of item difficulties from Bayesian and parametric IRT modelling . . .	290
C.1	Access to public appendix	291

Abstract

This thesis investigates the extent to which the difficulty of reading comprehension items in a foreign language reading assessment can be predicted by characteristics of the assessment, how these interact with certain attributes of the test takers, and, most importantly, whether such predictions can be made across languages. The dataset stems from a large-scale foreign language assessment administered to 6th-graders in Switzerland (ÜGK) and was developed in four parallel language versions: test takers in German-speaking Switzerland were assessed in English or French, test takers in French-speaking Switzerland solved an assessment in German, and the students in the Italian-speaking canton Ticino were given a French reading test. With minor exceptions, the four language versions were identical except for the language in which they were presented.

After an introduction, the thesis begins with definitions of key concepts and an in-depth discussion of reading comprehension, including a discussion of theoretical models of reading comprehension, possible components of the construct, and specific considerations for reading in a foreign language and at a low proficiency level. It then turns to the measurement of reading comprehension, examining the roles of input texts, tasks, and test takers. It also takes a closer look at the measurement of lexical characteristics of texts, especially across languages or in a multilingual context. This is complemented by an overview of psychometric approaches to estimating item difficulty, including both CTT and IRT. The literature review concludes with an examination of empirical studies on multilingual assessment and item difficulty, summarising their methodological approaches, variables, and key findings. Overall, it can be shown that foreign language reading is hardly ever considered in terms of its inherent multilingual nature and that not much is known about the linguistic effects of translating educational assessments.

The empirical study attempts to shed some light on these issues. The input texts and items of the assessment were operationalised in terms of their linguistic characteristics, focusing on lexical aspects given the low proficiency level of the test takers (e.g. text length, several lexical diversity measures, lexical sophistication and difficulty). In addition, certain assessment characteristics and a few selected attributes of the test takers were examined. All variables were extracted for each of the four language versions. Item difficulties for all items were estimated on the same scale using a many-facet Rasch model. This confirms the existence of DIF due to the language version in a majority of the items. It also shows that the items which include related languages (English and German, French and Italian) are on average 0.7 logits easier than their counterparts in the assessment versions including French and German. Qualitative examinations of selected items point to an influence of cognates and certain keywords to explain differences between language versions.

The results section presents descriptive statistics of all variables included in the study. Of particular interest is the behaviour of some linguistic variables across languages, which suggests problems with their operationalisation. Investigations of multicollinearity reveal numerous interdependences between variables describing the input texts and items. In light of these issues, the relationship between item difficulty, the assessment variables and, to a lesser extent, test taker attributes were then investigated using multiple statistical approaches. Correlation and regression analyses showed that text length and text type, along with lexical difficulty and a measure of cognateness were good predictors in all language versions. Characteristics of both the input text and the items normally played a role. In the details, however, the language versions often varied

notably. Also, variables which attempted to operationalise the cognitive processes involved in solving the items were not usually retained in the models. A random forests analysis and an exploratory analysis of a sample of the dataset using Bayesian statistics largely confirmed these findings. In general, the effects of all predictions are weak or, at best, moderate. Information about the test takers usually performs much better at predicting L2 reading ability. Reasons for these and some more unexpected results are discussed at length in the final chapters of this thesis.

Both the theoretical and the empirical part of this thesis underline the multivariate and partly unpredictable nature of reading comprehension and reading assessment, particularly when several languages are involved. No single variable can adequately predict item difficulty, and those which are among the best predictors can be shown to be influenced by a multitude of factors depending on context and language. Nevertheless, the results of several analyses in this thesis strongly suggest that other languages play an important role in foreign language reading comprehension, particularly at the lower proficiency levels. This has often been overlooked in prior research.

Chapter 1

Introduction

This thesis investigates the following question: *if a reading assessment is administered in different foreign languages but is otherwise the same, why are the items in the language versions not of equal difficulty?*

It is based on an empirical observation which arose from Switzerland's very first national educational monitoring, the *Überprüfung der Erreichung der Grundkompetenzen (ÜGK)*: the comprehension items in the four foreign language assessment versions were identical except for the language they were presented in. Their empirical item difficulty, however, differed substantially. The assessment was part of a larger national harmonisation effort, and meant to verify whether students in compulsory school reached the basic competences prescribed by the state. Following established practice in international educational assessments (e.g. the Programme for International Student Assessment (PISA)), the entire assessment, including its foreign language component, was developed once and then administered in parallel in the three major linguistic regions: the content stayed the same, only the language was changed. For the foreign language assessment, reading and listening tasks were developed alongside each other in the three foreign languages taught in the participating schools: English, French and German. They were presented to students with three different languages of schooling: French, German and Italian.

This assessment was administered in the spring of 2017 to $N = 20,177$ students at the end of primary school (grade 8¹, age 12), alongside an assessment of the language of schooling and a background questionnaire. Since psychometric analyses revealed significant differences between the item properties in the three foreign language versions of the assessment, all test results were analysed and reported separately for each language version. Nevertheless, the data is now available for scientific research, and lends itself to an investigation of item difficulty across language versions.

Very generally speaking, item difficulty is a measure of the probability with which a test taker will solve an item correctly. Given the nature of educational assessment, item difficulty is always a function of a variety of factors, including components of the test construct, characteristics of the test and the test items, as well as aspects unrelated to the test construct. For instance, an L2 reading item's difficulty may be influenced by the topic and lexical sophistication of the input text, the need to use knowledge of text cohesion, but also, possibly less desirable in terms of the test construct, by its item format or difficult wordings of test questions. In addition to that, attributes of the test takers may play a role: some test items may be easier to solve for a subgroup of test takers, e.g. because of their language knowledge, cultural background, age or even gender.

The functioning and interplay of the factors related to the test items and input texts is at the heart of research on item difficulty. Born out of an interest in construct validation, this research field now spans 40 years and a variety of applications: both reading and listening assessments have been analysed and a lot of variables have been investigated, though with varying success.

¹According to the HarmoS agreement (see below), school years in Switzerland are counted from 1 to 11, starting at age 4 in kindergarten. Grade 8 equals grade 6 in most German-speaking school systems.

There have also been various statistical approaches to the problem, including the use of artificial neural networks in the last few years and even some rare experimental studies. The language of the assessment, though, has rarely been an issue: almost all studies on item difficulties have been conducted on assessments of English, either as a (often advanced) foreign language or as the language of schooling. There are a few studies on other languages, but none of those address the question whether results from research on assessments of English are transferable to assessments of other languages.

This is particularly interesting as there is evidence that the languages involved in an assessment do, in fact, influence the way the underlying construct is assessed. First, studies on the impact of the L1 in L2 comprehension tests point to the fact that some test takers are advantaged over others, especially when lexical competence is assessed or a major part of the construct under investigation². Second, studies on international large-scale assessments, surveys and questionnaires which were translated to be used with various L1 groups regularly reveal differential item functioning related to the language of the assessment for a considerable number of items³.

The present study is an in-depth analysis of the ÜGK foreign language reading assessment data to address this important gap in current research. Specifically, it examines input text and item characteristics across four language versions and investigates their relationship with item difficulty. In the following, the context of the study is briefly described in more detail, and the research questions are laid out. The introduction concludes with an overview of this thesis.

1.1 Context of the study

The objective of this study is a deeper understanding of foreign language⁴ reading item difficulty *across languages*. As such, this thesis draws its empirical foundation from two fields of research: the study of item difficulty and the study of multilingual test adaptation. Item difficulty research has been conducted in both L1 and L2 assessment contexts, though researchers rarely took into account more than one language at a time (though see J. L. D. Clark 1965, whose study is the most similar to this one; see also Beinborn, Zesch, and Gurevych 2015; Bergqvist, Theens, and Österholm 2018). Empirical studies of test items in several languages are mostly found in sociological and psychological research, where researchers examined the comparability of tests or questionnaires which had been translated to be used by speakers of a different L1 group.

The study is also embedded in the physical context of the assessment under investigation: the Swiss school system, and, more precisely, its compulsory foreign language policy, which brought about the multilingual ÜGK⁵ foreign language assessment.

1.1.1 The ÜGK foreign language assessment

Switzerland is a multilingual country. Its four national languages (German, French, Italian and Romansh) are protected by the constitution and used by a majority in separate parts of the country. Many more languages have entered the country through migration and are now part

²For instance, Allen (2019b) could show that in a test of lexical competence, items which assessed cognates with the test takers' L1 were much easier to solve. If such a test contains relatively more cognates than the actual language in use, scores on this test will be inflated for competent speakers of the L1.

³This has been shown repeatedly for assessments such as PISA: in a study of the 2000/2001 PISA reading assessment, according to Grisay and Monseur (2007), up to 40% of all items differed substantially in their item difficulty across translated language versions (e.g. 29% of all items in the French PISA reading assessment were at least 0.5 logits harder or easier than the same English items, *ibid.*, p. 72).

⁴*Foreign language* is understood here as a language which is being studied as a non-intensive subject in compulsory school. See section 2.2 for a short discussion of the terminology surrounding languages.

⁵In French: Vérification de l'atteinte des compétences fondamentales (COFO), in Italian: Verifica del raggiungimento delle competenze fondamentali (VeCoF). In the following, the German name and abbreviation will be adopted for clarity's sake. Henceforth, the same rule applies to any other institutional terms for which there is no official English denomination.

of the linguistic landscape⁶. Education, including language instruction, has traditionally been managed by the Swiss federal states, the cantons, but over the past two decades, efforts toward national coordination have led to the Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS) (Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) 2010). This framework outlines common educational goals across language regions, including minimum requirements for foreign language instruction and the development of shared curricula. Most relevant for this thesis, it also established a national educational monitoring; what would eventually be ÜGK.

ÜGK took place for the first time in 2016: students at the end of secondary school had their mathematics skills tested (Konsortium ÜGK 2019a). A year later, in 2017, both the language of schooling and the first foreign language were assessed at the end of primary school (Konsortium ÜGK 2019b). In both years, the main interest of the study was to determine the extent to which students in Switzerland reached a selection of the basic competences (German: *Grundkompetenzen*) established through HarmoS (ibid., p. 5).

Both the mathematics and the language assessment were developed in parallel for the three major language regions of the country⁷ (Angelone and Keller 2019b, p. 5, 2019a, pp. 5, 7), which is a well-established practice in international large-scale assessments such as PISA (e.g. R. Adams and M. Wu 2002, pp. 57 sqq.; OECD 2019c, chapter 5). As a result, the foreign language assessment was administered in four language versions: the three *foreign* languages taught in school (English, French and German) combined with instructions and items written in the three languages of *schooling* (German, French and Italian, see table 1.1). With the exception of these differing languages, all texts and items were intended to be identical.

		Language of schooling	German	French	Italian
Target language	English	ENde			
	French		FRde		FRit
	German			DEfr	
Curricular region		ZH, BKZ, East	Passepartout	Romandie	Ticino

Table 1.1: Overview of the test versions in the foreign language component of ÜGK 2017. Only the *first* foreign language was assessed. The abbreviations are used by ÜGK to distinguish between the language versions. In those, the lowercase characters represent the language of schooling and the capital letters designate the foreign language. The letters are taken from the German words for the languages: DEutsch, FRanzösisch, ITalienisch, ENglisch.

Despite the fact that the tasks and items were direct translations of each other, analyses of the test results⁸ clearly show that psychometrically speaking, the language versions were not part of the same assessment: a large number of items shows regional Differential Item Functioning (DIF), which implies that test takers from the different curricular regions perceive the items in different ways (Angelone and Keller 2019b, p. 14). As a result, the three foreign language assessments

⁶The linguistic situation in Switzerland is too complex to be described here adequately. A lot has been written about it, though, both in scientific publications and outside of it. A good overview with numerous references can be found on the French version of Wikipedia (Wikipédia 2025).

⁷The German-, French- and Italian-speaking parts of Switzerland. The multilingual Grisons canton only participated in the assessment of mathematics and German as the language of schooling. The Romansh- and Italian-speaking schools in the canton did not participate in the project.

⁸In fact, this observation had already been made during a representative piloting session of the assessment in 2016 (F. Keller, personal communication, 2016-09-29).

were analysed separately in the main study, though items which showed strong regional DIF were still eliminated (ibid., pp. 13 sq.).

The unique make-up of this assessment—input texts and test items in parallel language versions—lends itself to an analysis of item difficulty across languages, which is the aim of this thesis.

1.1.2 Item difficulty and test translation research

Foreign language assessments are not usually translated (or, if they are, they are not researched or publications about them are hard to find). Hence, this study of item difficulty across languages is unusual and cannot draw on similar prior research (with two dated exceptions: J. L. D. Clark 1965; Manuel 1953). Instead, two lines of empirical research are considered: item difficulty research and the study of translated (educational) assessments.

Item difficulty research investigates the relationship between empirical item difficulty and characteristics of an assessment. Based on theoretical models of language use, related empirical research, and considerations of the assessment construct, researchers have been establishing possible explaining variables and study whether they can be used to predict the differences in item difficulty of a certain assessment (e.g. Buck, Kumi Tatsuoka, and Kostin 1997; Carr 2006; Drum, Calfee, and Cook 1981; Embretson and Wetzel 1987; Freedle and Kostin 1993, 1999; Neumann 2012). The assessments under investigation, the chosen variables and the methods differ widely, though most of these studies are quantitative in nature, study reading or listening comprehension, and focus on English (both as an L1 or an L2). The results of these studies show that solving a language assessment is the result of a complex interaction between the test taker and the various parts of an assessment, i.e. its instructions, input texts and items (see section 8.3). Problematically, none of these studies take into account languages other than the main language of the assessment, even if a foreign language assessment (e.g. the Test of English as a Foreign Language (TOEFL)) is studied.

Separate from this field of research is the *study of translated assessments*: international educational assessments (e.g. PISA), as well as psychological, sociological or medical tests and questionnaires have been translated for use in other linguistic contexts for decades (ElAtia 2013, p. 194). It has long been known that such a translation (or adaptation, cf. Sireci, Y. Yang, et al. 2006, note 2; Van de Vijver and Tanzer 2004, p. 122) does not automatically create an assessment which is equally valid as the original version (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, pp. 60 sq.). Hence, researchers have studied the comparability of different language versions of assessments, both in education and in other fields (e.g. Elosua and López-Jáuregui 2008; Ercikan 1998; Grisay, Jong, et al. 2007; Sireci and Allalouf 2003). Given the sometimes large differences between test versions (e.g. in PISA, cf. Grisay, Jong, et al. 2007; Kankaraš and Moors 2014), more elaborate methods of translation were developed, and statistical approaches discussed. Some authors also tried to investigate the reasons for these translation effects, with varying success (e.g. Allalouf, Hambleton, and Sireci 1999; Ercikan et al. 2010).

1.2 Research questions

This study uses the unique multilingual nature of the ÜGK foreign language assessment to explore how item and text characteristics are related to item difficulty in three different foreign languages and how this relationship differs across languages. It establishes text, item and text-by-item variables which operationalise the textual and assessment-related characteristics of the assessment tasks and studies their contribution to their empirical item difficulty using a combination of psychometric modelling, statistical modelling and exploratory data analysis techniques.

The following Research Questions (RQs) guide the analysis:

RQ 1: Which characteristics of the assessment (its input texts and items) predict empirical item difficulty in each of the four language versions?

RQ 2a: Can the characteristics of the assessment be operationalised across the four languages?

RQ 2b: Are the linguistic characteristics of the translated assessment parts identical across the four language versions?

RQ 2c: Do the characteristics of the assessment predict item difficulty in the same way in the four language versions? If not: what are the major differences?

RQ 3: Do certain characteristics of the test takers interact with the item characteristics for the prediction of item difficulty in the four language versions?

RQ 1 represents the foundation of this analysis. It follows traditional item difficulty research in that suitable variables for this assessment are operationalised and their relationship with item difficulty is modelled statistically, for each language version by itself. RQs 2a and 2b lay the groundwork for the multilingual analysis by establishing to what extent the variables in the analysis lend themselves for crosslinguistic analyses. RQ 2c then asks explicitly whether the prediction of item difficulty is language-dependent or whether the same mechanisms are at play across languages. Finally, RQ 3 attempts to shed some light on the role of the test takers, which has thus far been largely ignored in empirical item difficulty research.

Together, these research questions aim to shed light on the extent to which text, item and test taker characteristics influence reading comprehension difficulty, in languages other than English and across foreign languages at a low proficiency level. They also serve to inform the design and interpretation of foreign language reading assessments more broadly, particularly with respect to construct equivalence, fairness, and comparability across languages.

1.3 Overview

This text consists of four parts: the literature review, the presentation of the study, the results section and a concluding discussion.

The literature review begins with brief definitions of some key concepts. It then discusses reading comprehension and its assessment in some depth, introduces the measurement of item difficulty and then reviews empirical findings from research on multilingual assessment and item difficulty research. First, section 3.1 defines reading (comprehension) and presents four major models of reading comprehension. It is followed by a discussion of possible components of reading and concludes with a consideration of reading in an additional language. The review then turns to the measurement of reading comprehension in chapter 4, which considers the roles of input texts, tasks, and test takers. In chapter 5, the measurement of lexical characteristics of input texts is examined in more detail, and possibilities of operationalising the crosslinguistic distance of words and texts are discussed. This is followed by a discussion of item difficulty and its estimation through psychometric approaches, including both Classical Test Theory (CTT) and Item Response Theory (IRT), in chapter 6. The final two chapters of the literature review

address empirical findings on multilingual assessments and summarise empirical research on item difficulty, with a particular focus on their statistical methods, key variables and main findings.

The second part of the dissertation presents the study. Chapter 9 provides details about the dataset: it describes the context and the test takers and presents the ÜGK assessment instrument, its tasks, texts and items. Section 9.3 details the construction of a comprehensive set of variables, which includes linguistic, text-based and item-level assessment characteristics as well as selected test taker characteristics. This is followed by a somewhat detailed account of how the dataset was prepared and analysed (chapter 10).

The results section is divided into a presentation of various descriptive statistics in chapter 11 and the results from inferential analyses in chapter 12. In section 11.1, all assessment variables are described in some detail, including their behaviour across language versions. Then, in section 11.2, response patterns and the results of the item analysed using CTT and IRT and examined in some depth. The chapter concludes with an overview of the test takers' performance in the ÜGK reading assessment and the other characteristics chosen as covariates in the inferential analyses. Chapter 12 then first discusses interactions between variables before it presents the results of the various approaches to predicting item difficulty. Section 12.3.1 and section 12.3.2 present the correlational analyses, followed by the results of the regression analyses in sections 12.3.3 and 12.3.4. Finally, results from the triangulation efforts using Random Forests analysis and Bayesian statistics are presented in sections 12.3.5 and 12.3.6. A summary concludes this chapter.

The final part of the thesis discusses the main findings in light of the research questions, as well as additional insights that emerged during the study and present opportunities for future research. In section 13.3, limitations of the study are acknowledged. The thesis concludes with a short reflection on the complexity of modelling item difficulty across languages and the need for more research.

Part I

Literature review

Chapter 2

Important concepts and terminology

This study can be classified as causal-comparative or *ex post facto* (Nunes Silva 2012): the study uses some existing consequence (i.e. item difficulty) and searches back through the data for factors (i.e. characteristics of the assessment and the test takers) to explain this consequence. As a result, certain aspects have been established beforehand, particularly the assessment itself, including its construct, design, target group and scoring procedure. Hence, some concepts otherwise central to language assessment research are beyond the author's control. These are defined in the following. Much could be said about each of them—filling entire chapters and books. Here, however, they are only defined within the context of this study.

2.1 Validity and validation

Every book and article on language testing deals to some extent with validity.

Fulcher and F. Davidson 2007, p. 3

Validity may be *the* central issue in (educational) assessment and it is generally recognised as such in the language testing community (Fulcher and F. Davidson 2012, p. 1). It is concerned with whether the conclusions we draw from a test's results can be considered to be justified, plausible or adequate (M. Kane 2006; Messick 1990), or even beneficial (Bachman and Palmer 2010). In order to answer this question, modern theories of validity⁹ purport an argument-based approach: instead of declaring a test to be valid based on some sort of calculation (e.g. a correlation with another test), a chain of arguments is established which connects the purpose of a specific assessment or test to its various components, its administration and its consequences:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. The principles of validity apply not just to interpretive and action inferences derived from test scores as ordinarily conceived, but also to inferences based on any means of observing or documenting consistent behaviors or attributes. [...] the key issues of test validity are the meaning, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of the social consequences of their use. (Messick 1990, pp. 1, 4)

Based on this initial definition, several frameworks were proposed over the years, which help develop and structure such a validity argument (e.g. Bachman and Palmer 1996, 2010; Chapelle

⁹Much has been written about the history of validity. The interested reader is referred to Chapelle 2012; M. Kane 2006; Newton and Shaw 2014; Sireci 2009 for excellent overviews from various perspectives.

2021; M. T. Kane 2013; Mislevy and Yin 2012). Though the various frameworks differ in important ways, they all establish the role of empirical evidence as “backing” for the arguments put forward. In this sense, this thesis can be considered part of a larger validation effort in that it analyses the functioning of low-proficiency foreign language reading tasks in several languages. Following Chapelle (2021, pp. 102 sqq.), knowledge about the functioning of the items can be considered in terms of *domain definition inferences*, as it potentially provides information about the representation of the target domain (i.e. L2 reading) in the test items by revealing which aspects of the reading text influence item difficulty and, as such, are representative of the construct. It can also serve as backing for *evaluation inferences* if the analysis can show that items are influenced by aspects of the assessment in systematic (and intended) ways. For similar reasons, *generalisation inferences* can be drawn, particularly if the differences between the item difficulties in language versions can be shown to have their origin in systematic differences of the input texts, thus allowing for the conclusion that the test scores actually reflect differences in L2 reading across the target groups. Evidence which puts these inferences in question, is, of course, just as relevant. In any case, this type of analysis adds knowledge to the field of foreign language reading assessment and, thus, potentially provides backing for future validation arguments: those concerned with the ÜGK assessment, but also with other language assessments which are similar to the one studied in this thesis.

2.2 Languages

In the context of this research project, the relationship between the different languages is relatively clear-cut: the Swiss public school system defines the *language of schooling* for each school based on the dominant language of its geographical location. School subjects are usually taught in this language and students are expected to be proficient in it. As per the nation-wide HarmoS agreement (Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) 2010), two other languages (English and another national language) are taught as *foreign languages* during compulsory school. By grade 8, the school year in which the ÜGK assessment of 2017 was conducted, all pupils have been learning their first foreign language for at least three years¹⁰. The ÜGK assessment, being a result of the HarmoS agreement, follows this clear distinction of the languages: the assessment of foreign languages targeted the reading and listening comprehension proficiency of the students in the school subject *first foreign language* (English, French or German, depending on the curricular region). The overall test language was the respective language of schooling (German, French or Italian).

This text follows this distinction of the test takers’ individual repertoires into a main language, the L1, and a foreign language, the L2, which is currently being learned. It is obviously understood that this view is a simplification. On a theoretical level, defining what a *language* is, does not come without problems, nor is distinguishing it from other, neighbouring concepts (such as dialects) an entirely straightforward affair. On a societal level, it may be considered problematic to use the term *foreign language*, as it can be understood to imply a *foreignness* that is undesirable in some contexts¹¹. In addition, the individual students who participated in the ÜGK assessment may have very different linguistic profiles from the assumed L1-L2 combination: they may know other languages, have only a limited command of the language of schooling or an above-average command of the taught foreign languages due to prior knowledge, extra-curricular activities or competences in related languages. Also, due to the important role of Swiss German dialects in every-day life, the pupils in the German-speaking part of the country may have a somewhat different connection to their language of schooling (a standard variety)

¹⁰Of course, pupils who have moved across cantonal or national borders during this time, may have had different trajectories.

¹¹In fact, the Ticino educational system does not use the term *foreign languages* (*lingue straniere*), but *second languages* (*lingue seconde*) (DECS 2004).

than the pupils in the French region, and, to a lesser extent, in the Ticino (where the Ticinese dialect is still present in some areas). As for the involved languages themselves, there is reason to believe that the students' English competence, especially their vocabulary knowledge, will have a somewhat different profile due to extramural exposure (e.g. Peters, Noreillie, et al. 2019; Puimège and Peters 2019). Finally, cognitively, it can be assumed that the students' linguistic repertoires do not distinguish between the languages as clearly as the above paragraph may imply—quite the opposite, in fact, as has been amply discussed in the literature (e.g. for an overview Lewis, B. Jones, and C. Baker 2012; Otheguy, O. García, and Reid 2019) and will be shown later (see sections 3.3 and 5.3.2).

2.3 Competences

As the name suggest (*Überprüfung der Grundkompetenzen*), the ÜGK assessment was meant to verify to what extent the Swiss school system has enabled students to acquire basic competences¹² in reading and listening in their first foreign language (Konsortium ÜGK 2019b, p. 5). These basic competences (EDK 2011) were established based on the action-oriented approach to language use described in the Common European Framework of Reference for Languages (CEFR):

Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of **competences**, both **general** and in particular **communicative language competences**. They draw on the competences at their disposal in various contexts under various **conditions** and under various **constraints** to engage in **language activities** involving **language processes** to produce and/or receive **texts** in relation to **themes** in specific **domains**, activating those **strategies** which seem most appropriate for carrying out the **tasks** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences. (Council of Europe 2001, p. 9, *emphasis in original*)

This definition implies a functional-pragmatic concept of competence: instead of theoretical models which try to describe a human's underlying cognitive abilities, competences are defined as the “context-specific dispositions for achievement that can be acquired through learning” (Klieme, Hartig, and Rauch 2008, pp. 7 sqq.) or “the sum of knowledge, skills and characteristics that allow a person to perform actions” (Council of Europe 2001, p. 9). Weinert refers to this understanding as a “knowledge-centered [or task demand-centered] concept of competence” (1999, p. 28), which underlines the fact that an educational assessment should focus on sampling the test takers' ability to solve concrete tasks¹³.

In the context of ÜGK, this means that the main focus of the assessment construct are (slightly) contextualised reading and listening tasks¹⁴ in which students have to demonstrate their understanding of short texts by answering questions. The nature of these texts is defined by the competence descriptors in the relevant documents (Council of Europe 2001; EDK 2011; Lenz and Studer 2007), as well as the respective curricula and textbooks (Konsortium ÜGK 2019a, pp. 70 sqq.). As a result, the construct definition of the assessment focuses on the tasks, e.g. the text types, topics and intended types of reading (Konsortium ÜGK 2019a, pp. 33 sqq.).

¹²There is some inconsistency in the translation of the German word *Kompetenz* into English: while the CEFR uses the term *competence*, some authors prefer the term *competencies* (e.g. Klieme, Hartig, and Rauch 2008), though the two terms are also used alongside each other. This thesis follows the CEFR in preferring the term *competence*. It is generally assumed that both terms mean the same thing.

¹³A similar conceptualisation of language proficiency is undertaken in the context of the “task-based approach” to defining language proficiency as opposed to the “trait-based” and mixed approaches, see for an overview National Academies of Sciences, Engineering, and Medicine 2020, pp. 28 sqq.

¹⁴See below for a definition of the term *task* (section 2.4.2).

The *competence* measured, then, is the ability to solve these tasks, rather than a theoretical construct of cognitive abilities.

This does not mean, of course, that the authors of the CEFR, of the curricula in Switzerland, or of the ÜGK assessment, did not take into account cognitive models of language ability. Instead, theoretical models of language competence provide the link between an action-oriented assessment like the one in the ÜGK study and linguistic or cognitive theories. In the context of language assessment, one well-known model of general language ability was described by Bachman and Palmer (1996; 2010). Their definition of *language use* distinguishes between internal interactions (within the language user) and external interactions (of the language user with the characteristics of the language use situation), as well as reciprocal language use and non-reciprocal language use (Bachman and Palmer 2010, pp. 34 sqq.). Among the individual characteristics of the language user, in addition to language ability, the authors describe personal attributes (e.g. age, gender, nationality), topical knowledge (i.e. the entire real world knowledge a language user has at their disposition), and affective schemata. Language ability, in turn, is divided into language knowledge and strategic competence. While language knowledge subsumes several areas of knowledge about language, ranging from “vocabulary knowledge” to “knowledge of cultural references and figures of speech” (ibid., p. 45), strategic competence is “what makes language use possible”, by allowing an individual to decide what to do (*goal setting*), take stock of the available resources (*appraising*) and create a plan of action (*planning*, see ibid., pp. 48 sqq.).

Another prominent model of language competence in the context of assessment was described by Weir (2005, pp. 43 sqq.). He defined four separate so-called “socio-cognitive frameworks” for the validation of reading, listening, speaking and writing. The part of these frameworks which describes the test takers and their ability, is, in essence, very similar to Bachman and Palmer’s model of language ability: physical, psychological and experiential test taker characteristics influence the internal processes of the test taker, and the external nature of the assessment. The internal processes, described by theory, are divided into executive resources, which include language and content knowledge, and executive processes, which were described for each language skill, but mostly comprise the three elements of strategic competence described by Bachman and Palmer (see above). Weir added a monitoring device to these internal processes, which creates a link between the resources and the processes, and “provides the reader with feedback about the success of [a] particular reading process” (S. Urquhart and Weir 1998, p. 105, cited in Weir 2005, p. 92).

Given such a model, it becomes possible to systematically describe the various aspects that influence a language use situation. Then, using psychometric tools or qualitative analyses, if these various aspects can be operationalised, their theorised interaction can be empirically investigated. This logic underlies many of the empirical studies presented in later chapters and, in fact, this thesis.

Over time, these “aspects” have been formalised in various ways, using different terminologies, but essentially, models of language competence or language ability usually contain a knowledge component and a strategic component. While the knowledge component can often be taught (and measured more easily), the strategic component tends to be partly automatised or unconscious, almost entirely acquired through exposure, and more difficult to measure. This thesis focuses primarily on the linguistic aspects of the knowledge component, as it is interested in the effect of the various languages on item difficulty. The ÜGK assessment, however, also created data on strategic aspects of language use and assessment, by way of questionnaire data collected alongside the language assessment (Erzinger, Hauser, et al. 2019). These are available for research, but they are only used in a very limited way in this thesis.

2.4 Assessment

Assessment instruments are used in many contexts to “collect information” (Bachman and Palmer 2010, pp. 19 sq.), to assess learning of some kind in education, but also, in fields such as psychology, sociology and medicine, to measure a state of mind or a behaviour (Bolaños-Medina and González-Ruiz 2012, p. 718). Throughout those contexts, a variety of terms are in use (e.g. test, measurement, questionnaire, scale, inventory, evaluation), sometimes meaning slightly different aspects or forms of assessment, but also used interchangeably. In the *Standards for Educational Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, p. 2) a *test* is defined as “a device or procedure in which a sample of an examinee’s behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process”. Meanwhile, an *assessment* is considered “a broader term than test, commonly referring to a process that integrates test information with information from other sources (e.g., information from other tests, inventories, and interviews; or the individual’s social, educational, employment, health, or psychological history)”. Despite this distinction, however, “the applicability of the *Standards* to an evaluation device or method is determined by substance and not altered by the label applied to it”, which is perhaps an important observation: no matter how exactly an assessment instrument is designed, and what it is called, many of the principles and considerations apply universally. For this reason, the terms *assessment* and *test* will be preferred throughout this thesis to denote any measurement instrument, including the ÜGK tasks under investigation. It is understood of course, that assessments, tests or measurement instruments may differ in the details, but, as Bachman and Palmer point out, “we believe that precise distinctions among the terms ‘assessment,’ ‘measurement,’ and ‘test’ are unnecessary. What *is* important, [...], is that the test developer clearly and explicitly specifies the *conditions* under which the test taker’s performance will be obtained and the *procedures* that will be followed for recording this performance.” (2010, p. 20)

2.4.1 Educational measurement

Large-scale assessments have become part of the educational landscape in Europe since the introduction of educational standards into many curricula. In an effort to create a more accountable, evidence-based educational system, their aim is to provide empirical data on an educational system (Eurydice 2009; Mons 2009, p. 100). The overall purpose of educational measurement is “to inform [about] educational policy, often under the banner of ‘raising standards’.” (Biesta 2009, p. 34, see also Mons 2009), and it may include more than test results, e.g. statistical data on schools, young people and the larger population (see, for instance, the Swiss Education Report, cf. SCCRE 2023).

Data which stems from this type of assessment must be analysed in specific ways, since it is often used to produce insights into populations: due to the non-random sampling of schools, classrooms and students, statistical analyses interested in the proficiency and characteristics of the *population* (which is represented by the sample of test takers), need to use sampling weights, and individual proficiency estimates need to be based on Plausible Values (PVs) (see Rutkowski et al. 2010 for more details and examples). Item analyses, however, do not suffer from these particularities, as they do not usually intend to draw conclusions for the entire population of reading items. Instead, the item sample of an international assessment is as much drawn from a potential population of items as in other assessments, and the findings from item analyses on this sample are restricted to the specific groups of items or item formats used in the assessment¹⁵.

¹⁵In fact, international assessments are usually subject to rigorous item development processes, pre-operational testing and psychometric analyses, which can be interpreted to mean that the sample of items is potentially of good psychometric quality. This may at least partly be true for the ÜGK foreign language assessment, as well.

The ÜGK assessment is the first Swiss large-scale assessment on a national level, though it is not the first such assessment to be presented to students in Switzerland: the country has participated in well-known international assessments before, particularly in PISA (since 2000¹⁶), and some know-how from these experiences, as well as from other assessment projects was used in the development and administration of the ÜGK assessment. A particularity of this assessment were its low stakes for the pupils: no results on individual test takers, classes and even schools were reported, which supposedly reduced the students' potential test anxiety (D. Putwain 2008; D. W. Putwain 2008, see also Mikami, Leung, and Yoshikawa 2008), though it may also have reduced their motivation to perform well (see e.g. Finn 2015; Penk and Richter 2017).

2.4.2 Tasks and items

The foreign language reading component of the ÜGK assessment conforms to the *Standards'* definition of a test in that it is “standardized”, by which “we understand that instructions provided to examinees, general testing conditions and scoring procedures [...] follow the same strict and specific scheme” (Bolaños-Medina and González-Ruiz 2012, p. 717). The complete test is composed of 26 *tasks* (*Testaufgaben*), each of which contain one or several short input texts (*Aufgabenstimulus*) and three to five *items* (Konsortium ÜGK 2019a, p. 8).

Being a mostly standardised large-scale assessment, the terms *task* and *item* are understood in their technical sense: an item is “each testing point [...] which is given a separate mark” (Association of Language Testers in Europe (ALTE) 1998, p. 149), while a task is “a combination of rubric, input and response” (ibid., p. 165¹⁷). Hence, each individual answer in the assessment stems from one item, and several items form one task.

The term *task* is also loosely understood in the tradition of Task-Based Language Teaching (TBLT), where a task is defined as “a functional activity in a particular setting in which an individual uses language to attain an objective” (Van Gorp and Deygers 2014, p. 588). In a very limited way, this understanding of the term applies for the ÜGK assessment as well: as mentioned above, the basic competences certainly define the construct in terms of real-world language use¹⁸. However, the constraints of the large-scale assessment made it necessary to reduce the *authenticity* of the tasks, which is normally considered an inherent characteristic of tasks as understood in TBLT (ibid., p. 2, see for examples of more “authentic” assessment tasks Norris 2016). As a result, the *tasks* described in this thesis are a group of assessment items linked to each other by their input text and the very basic context in which they are situated (cf. section 9.2 for more information) rather than actual near-real life *tasks* to be solved in the tradition of TBLT.

2.4.3 Question, stem and answer options

Depending on the scientific field and, of course, the item format, the various components of an item are sometimes named differently. Multiple-Choice Questions (MCQs), which are the most common item format in the ÜGK assessment, are usually composed of a stem or question and several answer options. The correct answer option is sometimes called *key* (e.g. Carr 2006; Kostin 2004), or *attractor*, while the incorrect answer options are often referred to as *distractors* (e.g. Neumann 2012). Beinborn et al. (2014, 2015), analysing the difficulty of cloze tests, used the term *candidates* to refer to the available answer options. In this thesis, the term *question* is used to generally describe the item's prompt, and *stem* is only differentiated from it to describe the beginning of a sentence when this distinction is necessary. In a similar vein, the general term *answer options* is employed, which includes both the correct and the incorrect ones.

¹⁶<https://www.oecd.org/pisa/aboutpisa/switzerland-pisa.htm>, last accessed 2023-05-30.

¹⁷The term *rubric* as understood by the Association of Language Testers in Europe (ALTE) glossary refers to the instructions candidates are given to solve a certain task.

¹⁸“Die Schülerinnen und Schüler können einen sehr kurzen, einfachen Text Satz für Satz lesen und dabei klar formulierte Informationen verstehen [...]” (EDK 2011, p. 16).

2.5 Translation

In this thesis, translation is understood in its most common sense—the interlingual transfer of written text from a source language to a target language¹⁹. More specifically, the process of *translating* brings about a product, the *translation*, which is a more or less literal reproduction of a source text (Hatim and Munday 2019, pp. 3 sqq.).

In the context of multilingual assessments, test translation is often linked with test adaptation, both of which are “the procedures being used for creating language versions of a test” (ElAtia 2013, p. 196). Test adaptation is often considered to go beyond the linguistic translation of a test, also taking into account cultural and educational aspects (ibid., p. 197). However, given the sometimes obvious adjustments that need to be made to translated test versions due to different knowledge bases across cultures, the two terms are often used interchangeably.

Strictly speaking, this thesis deals with parallel language versions of test items rather than with translations: there is not *one* source text. Instead, the test tasks were developed first in a language chosen by the individual test developer and then transferred to the other languages as part of the test development process. During the subsequent pre-operational testing phases, any changes were made to all existing language versions (see section 9.2.4 for details on the item development). Hence, it would be nonsensical to identify a source text, even at the item level.

Nevertheless, no matter how the parallel language versions were produced—whether they were translated or adapted from one language version, or developed in parallel—the same principles apply: from a measurement point of view, it is required that the language versions function in the same way, i.e. that they convey the same meaning independently of the different contexts in which they are presented: “[a good test translation] must reflect not only the meaning of the original item, but should also try to maintain the same relevance, intrinsic interest, and familiarity of the item content” (Ercikan 1998, p. 544). As will be discussed in chapter 7, this is no easy task and, in a way, this thesis explores whether it is at all possible to consider the equivalence of translated foreign language assessments.

¹⁹Broader understandings of the term include *interpreting*, the translation of spoken text, *intralingual translation*, i.e. the adaptation of a text into another dialect of the same language, or *intersemiotic translation*, which is the transfer of information from one medium into another, e.g. subtitling (cf. Hatim and Munday 2019, pp. 3 sqq.).

Chapter 3

Reading comprehension

I am not the first person to say that an overview of the study of the nature of reading is impossible.

Alderson 2000, p. 1

Based on a broad review of empirical research spanning several decades and scientific fields, Grabe (2009) describes reading as an inherently complex human action, which involves a multitude of cognitive processes and various fields of knowledge, and which is highly individual and context-dependent. Though the details are a matter of discussion, (skilled) reading is generally described as a purposeful activity involving automatised lower-level processes on the one hand (letter and word recognition, phonological encoding, semantic access, proposition integration, some highly proceduralised compensatory strategies, etc.) and higher-order comprehension processes on the other hand (goal setting, comprehension monitoring, recognizing discourse structures, integrating text content with existing knowledge, application of appropriate comprehension strategies) (cf. Grabe 2009, pp. 21 sqq.; R. A. Mason and Just 2012). The continual interplay of these knowledge sources, processes and strategies creates an ever-evolving internal representation of the text, the *text model*, which is further elaborated by the reader using their background and domain knowledge, current reading goals, personal interpretations of the author's intentions, characteristics etc., thus constructing the *situation model*²⁰ of the text (cf. Grabe 2009, pp. 39 sqq.; Dijk and W. Kintsch 1983)²¹. This definition of reading applies both to L1 reading and reading in one or several additional languages (see section 3.3 for a more detailed discussion).

Looking at the matter from a language assessment perspective, Alderson makes a distinction between the process and the product of reading (2000, pp. 3 sqq.). In many ways, the above description encompasses the *process* of reading, from the recognition of the first word of a text to the resulting situation model. From there, nothing may follow: one may simply read something for pleasure (and retain some memories of that), but not *do* anything in particular with it that would be visible to the outside. Then again, one may react to something that was read (e.g. choose a dish from a menu, or make an excited phone call following a dinner invitation via text

²⁰The choice of the term *situation model* may be explained from the following initial definition brought forth by van Dijk and Kintsch: "A major feature of our model is the assumption that discourse understanding involves not only the representation of a textbase in episodic memory [i.e. the text model], but, at the same time, the activation, updating, and other uses of a so-called situation model in episodic memory: this is the cognitive representation of the events, actions, persons, and in general the *situation*, a text is about." (Dijk and W. Kintsch 1983, pp. 11 sq., *emphasis added*).

²¹Perfetti distinguishes the *meaning* of a text from its reader's *interpretation* (1989, p. 312) and also mentions the concept of mental models, which are opposed to "a mental code that provides a direct linguistic representation of [utterances]" (Johnson-Laird 1981, p. 353), i.e. the text model. *Mental models* or *mental representations* are now often discussed with respect to more general models of comprehension (Johnson-Laird 2001; McNamara and Magliano 2009, p. 302).

message), produce a more or less formal account of the text (e.g. tell a friend about a good book, write a review for a blog, or a summary for school) or, of course, solve some sort of task (e.g. build a wardrobe based on some instructions, or fill in a graphical organiser in a test)²².

These *products* of reading, be they physical reactions, new texts or assessment results, are often the only source of information about someone's reading comprehension. There are, of course, ways to explore the *process* of reading as well, but these are most often indirect: a reader can be observed while reading, e.g. in eye-tracking studies or through think-aloud protocols. In addition, cognitive psychology offers some neuroimaging techniques, e.g. the study of brainwave patterns using Event-Related Potentials (ERPs) (e.g. Perfetti and Stafura 2014, pp. 26 sqq.), Positron Emission Tomography (PET) or Functional Magnetic Resonance Imaging (fMRI) (e.g. R. A. Mason and Just 2006; P. Li and Clariana 2019). All of these techniques require an important amount of interpretation on the part of the researcher. Along with the fact that we are not aware of unconscious processes such as word recognition and, thus, cannot describe them directly, this makes it difficult to establish the exact nature of reading, let alone reading in an additional language. This is where theoretical models come into play: "Models characterise theories of reading, providing ways to represent a theory or part of a theory; they explain what reading involves and, in more detailed versions, how reading works [...]" (Grabe 2009, p. 83, see also Sadoski and Paivio 2007).

Since a study of item difficulty needs to be based on an understanding of the construct in order to choose plausible sources of difficulty, four approaches to modelling reading will be discussed in the next section. The Simple View of Reading (SVR) and the Construction-Integration (C-I) model are perhaps the two most influential approaches to L1 reading, though in very different ways. The Reading Systems Framework (RSF) is a less well-known model which builds on work regarding the importance of lexical knowledge in language use and as such provides a useful conceptualisation of the skills necessary for low-proficiency reading in an L2. Finally, one of the few explicit L2 reading models is presented, which was also used in the ÜGK assessment.

All of these models have one underlying assumption: they consider reading a skill, which exists independently of the individual and the text, and can be described as such. Other theoretical views consider reading a social practice amongst others, which cannot exist or be described in isolation (e.g. Hill and Parry 1992, cited in Alderson 2000, pp. 25 sqq., see also McVee, Dunsmore, and Gavelek 2005). From their point of view, reading always occurs within a social context, alongside other literary practices (e.g. writing) and in complex interactions between reader(s), society, and texts (see for a model of text comprehension which reflects this: Strohner 2006, p. 191). For teaching contexts, this integrative understanding can be advantageous: as Rosebrock (2012) demonstrates, there is no one way of teaching or promoting reading; instead, depending on the readers, their age, and the overall social context, some approaches and texts may be more adequate than others. Educators have to be aware of the social complexity of reading practices in order to choose the most promising teaching approach (ibid., pp. 8 sqq.). In an assessment context, however, it is practical and necessary to model reading as a separate (and, in fact, separable) skill. Otherwise, with nothing tangible to measure, reading assessment simply is not possible (Alderson 2000, pp. 27 sq.)²³.

Once the four models of reading have been presented in section 3.1, the skill *reading* will be deconstructed further into its various components, from word recognition to the establishment of

²²Sadoski and Paivio (2007, p. 341) also suggest a theory of reading which encompasses the process and the product of reading, but they devise three somewhat overlapping phases with different demarcation lines: *decoding* or *recoding* (the conversion of printed language to spoken language), *comprehension* (the meaningful interpretation of the spoken language identified in the first phase, including both the text model and parts of the situation model) and *response* (the inferential/critical interpretation part of the situation model, as well as "affect, appreciation and/or application").

²³Obviously, this does not mean that a reading assessment does not need to take into account the characteristics of the test takers or the social context of the reading construct. It has to accept, however, a certain amount of generalisation over test takers and social circumstances in order to describe a reading construct which can be translated into assessment tasks and/or items.

a situation model and the role of metalinguistic knowledge (section 3.2). Finally, in section 3.3, the particularities of reading in an additional language will be examined further.

3.1 Models of reading

Various models of reading have been proposed over the years, with the C-I model (W. Kintsch 1988) and the SVR (Hoover and Gough 1990) probably the most influential. These two models differ from each other in important ways, representing different scientific traditions, aims, and research evidence. Both models reject the idea of reading as an exclusively bottom-up or top-down process²⁴, but they do so in two very different ways: while the C-I model depicts reading as a highly flexible and adaptive *process*, the SVR aims at identifying “the cognitive *capacities* needed for reading” (Hoover and Tunmer 2022, p. 400, emphasis in original), making it, with its two elements (*word recognition* and *linguistic comprehension*), the most basic of component models of reading (cf. Kendeou, McMaster, and Christ 2016, p. 63).

The C-I model and the SVR will be described in the following sections, as proxies of many other models of reading that have been suggested in the past 40 years (for more thorough accounts cf. Grabe 2009; McNamara and Magliano 2009). Following that, a more general approach to understanding reading is presented (the RSF, cf. Perfetti and Stafura 2014), before the most popular L2 reading model, which also underlies the ÜGK assessment, is discussed (Khalifa and Weir’s Cognitive Model of Reading, 2009). A short conclusion will summarise the findings.

3.1.1 The Construction-Integration Model

In cognitive psychology, comprehension models, including the C-I model, are connectionist in nature: cognitive processes and knowledge are considered small elements (so-called *nodes*) of a large network, which are linked together through association. The strength of each link stands for the probability that two elements will be activated together, which is strengthened by repetition (i.e. learning). According to Grabe, “this notion of massively connected bits of information in dynamic networks” allows connectionist models to explain cognitive processing without requiring strong assumptions about the nature of cognition (2009, p. 86, see also McNamara and Magliano 2009, pp. 304 sqq.).

True to its connectionist foundations, the C-I model describes reading as a process in which the linguistic input from the text base and the reader’s knowledge base work together to *construct* (hence the C) the text model, which is then *integrated* (hence the I) to form a coherent whole (W. Kintsch 1988, p. 164).

Kintsch describes the knowledge base as a massive associative network of propositions which are more or less strongly linked with each other. At any given moment, a small part of this network is activated and can be used while constructing a text base: after the actual linguistic input has been parsed, lexical and world knowledge will help to interpret meaning, e.g. by distinguishing agents from objects, identifying the referents of pronouns or by activating the most probable meanings of homonyms. Once a proposition has thus been formed from the text input, associated concepts and propositions will be activated. If elements are missing or blatantly contradictory, inference processes will try to fill in the gaps. At the end of this construction phase is an associative network of propositions from the text base which are linked to each other more or less

²⁴A bottom-up model understands reading as a process in which the reader perceives letters, decodes them as words, and creates understanding by combining them into sentences and finally, texts. Top-down models go the other way and consider readers to always have an expectation of a text, which is then confirmed or altered by intentionally sampling information from a text (cf. Grabe 2009, p. 89; Koda 2007, p. 9). Both approaches to reading, and even their combination in an interactive model of reading, have been disproven by empirical research (Grabe 2009, p. 90; Sadoski and Paivio 2007, pp. 339 sq.), but any web search of the terms will show that they are still part of the educational landscape as reading processes and strategies (see also S. Urquhart and Weir 1998, p. 37).

strongly and which have also activated associated concepts and elements in the larger knowledge network (ibid., pp. 166 sqq.). Kintsch describes this network as “carelessly constructed and [...] therefore incoherent and inconsistent [because at] all levels of the representation, components associated with the text elements were included without regard to the discourse context, and many of them are inappropriate” (ibid., p. 168). This is where *integration* comes in: the network will be consolidated by “[spreading around activation] until the system stabilises” (ibid., p. 168). Kintsch describes this in more detail, but the gist of the integration process is that probable interpretations are selected while concepts that are less strongly linked to the network are deactivated. This construction-integration process runs in cycles, which “roughly [correspond] to short sentences or phrases” (ibid., p. 168), activating and consolidating propositions continuously to form a situation model across a larger text.

Initially, the C-I model seems to be concerned mostly with comprehending a text *after* words were recognised, which is, admittedly, the more interesting and varied part of reading in an L1 context: word recognition is generally effortless even to a moderately skilled L1 reader, whereas comprehension processes such as inference, disambiguation, building situation models and learning from reading are harder to grasp. However, Kintsch demonstrates how the construction-integration model also applies to the process of identifying and disambiguating a word, driven by the word form and—given enough time—the discourse context (ibid., pp. 171 sq.). This is the great advantage of the C-I model: in line with connectionist modelling, it does not assume any particular knowledge structure or procedural order. Instead, it allows for a reading process which makes use of all available resources to interpret perceptual input as effortlessly as possible (i.e. using automatised behaviours), draws on more conscious knowledge and abilities when comprehension breaks down, and explains how world knowledge and text content are integrated with each other. All the while, the model recognises the highly individual nature of cognition and knowledge, making it possible to describe wildly different reading processes and outcomes, even when readers, texts and even circumstances differ only in details (cf. McNamara and Magliano 2009, p. 353). This flexibility also allows for the addition of another language, making the C-I model applicable to L2 reading comprehension: under its auspices, reading in one or several additional languages merely widens the knowledge source available to the reader, multiplying the possible candidates, especially for word recognition (see also section 3.3). This can even be put forward to explain why multilingual readers may take slightly longer to comprehend an element or a passage or even encounter comprehension problems: they have more words and structures to choose from.

3.1.2 The Simple View of Reading

Following a very different logic, the SVR model defines reading comprehension *simply* as the product of word recognition and linguistic comprehension ($R = D \times L$)²⁵. In this theory, word recognition is the ability to interpret letters (or symbols) as words, while linguistic comprehension then “does the rest” and encompasses understanding language from the word level upwards. Both components are characterised by fluency in that fluent reading comprehension is only reached if word recognition and linguistic comprehension are both highly automatised (Sparks 2021).

This definition, though deliberately general, has important implications for reading instruction and assessment: reading comprehension will fail or be ineffective if either of the two components is very weak or if both of them are (multiplying a number by zero will always give zero). As a result, word recognition and linguistic comprehension are considered the two major predictors of reading comprehension. Both assumptions have been supported by evidence (Hoover and Tunmer 2022, p. 401; Catts 2018, p. 317) and the model in its simplicity has become widely known even outside the scientific community: it reminds stakeholders to consider both reading-as-

²⁵Reading comprehension (R) equals decoding (D) times linguistic comprehension (L) (Hoover and Gough 1990, pp. 132 sq.). Decoding has since been replaced by the broader term *word recognition* (Hoover and Tunmer 2022, p. 400).

decoding and reading-as-comprehension capacities worthy of instruction and assessment (Catts 2018, p. 320). This has practical relevance with respect to learners with reading disabilities, but also for children who learn to read (i.e. decode) in a language they do not (yet) master quite as well as their monolingual peers (i.e. second-language learners).

Despite its popularity, the usefulness of the SVR in modelling *foreign* language reading is debatable at best. Its basic premise may hold for L2 reading (see, however, Grabe 2009, p. 98, who points out that the relationship of decoding and comprehension in L2 learners will be different from L1 readers), but it does not add much in terms of practical applicability. For instance, Sparks (2021, pp. 510 sqq.) summarises research on L2 reading proficiency which shows that a) foreign language listening comprehension is strongly correlated with foreign language reading²⁶ and b) tests of automatised word recognition and vocabulary show that many foreign language learners fall into the *hyperlexic* category of readers according to the SVR model: they possess good word recognition skills, but weak linguistic comprehension skills (see also fig. 3.1). While Sparks

		Decoding	
		Poor	Good
Oral Language (Listening) Comprehension	Good	<u>Dyslexic</u> Poor decoding Good language comprehension	<u>Good</u> Good decoding Good language comprehension
	Poor	<u>Mixed</u> Poor decoding Poor language comprehension	<u>Hyperlexic</u> Good decoding Poor language comprehension

Figure 3.1: Profiles of readers proposed by the SVR model (Sparks 2021, p. 509).

interprets this insight as evidence that the SVR is also applicable to foreign language reading, it seems neither surprising nor particularly helpful: foreign language learners in particular (as opposed to *second* language learners), with a few possible exceptions, will have already built word recognition skills in their L1, which may be easily transferable to the L2. The major task of a foreign language learner, then, according to the model, is building linguistic comprehension. Unfortunately, the SVR's monolithic conceptualization of linguistic comprehension does not allow any further predictions as to how it may be acquired or what may be important to do that.

The concept does remind us, however, that first and foremost, reading comprehension depends on knowledge of the language, which is especially relevant in L2 contexts where the actual act of decoding letters has been acquired long before the meaning of the words.

3.1.3 The Reading Systems Framework

The definition of reading comprehension formulated in this chapter's introduction has emerged as some sort of consensus, or overlap, of the various propositions, as Grabe shows in his review of eleven models of reading (2009, pp. 83 sqq.; see also McNamara and Magliano's (2009) structured account of seven comprehension models). In light of this convergence, it has been suggested that it may be time for a "Unified Theory of Reading" (Sadoski and Paivio 2007) or "comprehensive theory of discourse comprehension" (McNamara and Magliano 2009). McNamara and Magliano formulate a number of aspects that such a comprehensive theory would have to be able to explain beyond what current (reading) comprehension models cover: they would put forward

²⁶The linguistic comprehension component is usually measured through listening comprehension assessments in L1 studies related to the SVR (e.g. Hoover and Gough 1990).

specific characteristics of both the reader and the text (ibid., pp. 370 sq.) but also comprehension processes at the lower and the upper end of the spectrum (ibid., pp. 359 sqq.). Sadoski and Paivio arrive at a similar conclusion, pointing out that established theories of reading often focus either on decoding *or* on comprehension, but not on both to the same extent (2007, pp. 346, 348), and that the reader response—particularly affect and imagery—has thus far been neglected (ibid., pp. 348 sq.).

Perfetti and Stafura went a different way and declared “reading comprehension [...] [to be] too broad a target for precise models”, proposing instead the RSF (Perfetti and Stafura 2014)²⁷, which combines most of the various aspects of reading mentioned above in interconnected components (see Fig. 3.2). Instead of trying to validate the framework in its entirety, the authors suggest that it could serve as a “scaffold for theory development and hypothesis testing”, e.g. to investigate the sources of reading comprehension problems (ibid., p. 26).

The framework contains three knowledge sources—knowledge about one or several languages’ orthographic systems, and knowledge about the linguistic system(s), i.e. phonological, syntactic and morphological knowledge²⁸. Together, these two systems serve word recognition, and knowledge of the linguistic system(s) also supports higher-order comprehension processes. A third general knowledge source contains information on the world, which supports the creation of the situation model and the disambiguation of meaning. The actual reading process is placed in the middle of the framework. It connects lower-level word recognition processes with higher-level comprehension processes by means of the lexicon—the central element of the model. The various components of the RSF (knowledge sources and processes) are bi-directionally connected with each other, which underlines the idea of language comprehension as a network rather than a unidirectional process²⁹.

In addition, Perfetti and Stafura posit that their framework is consistent with Hagoort’s “neurobiological model of language processing” (the Memory, Unification, Control (MUC) model, Hagoort 2005; see also Hagoort 2013). It implies—roughly speaking—an understanding of language processing as a higher cognitive function which is distributed throughout several regions of the brain and dynamically informed by a number of subsystems. Meaning is derived from *memory*³⁰, and established through *unification*, which may be syntactic, semantic and/or phonological. *Control*, then, is the system which chooses the elements that are processed and possibly retained, and, if deemed necessary, initiates the reactions to what was understood (Hagoort 2013, pp. 9 sq., see also Grabe’s discussion of working memory and executive control, 2009, pp. 50 sqq.).

In terms of this thesis, the RSF is interesting in the sense that it conforms with modern cognitive theory, but is concrete enough that it allows the notion of *pressure points*: if an element or a connection in the framework breaks down or is not available, then reading comprehension will fail or has to be achieved through compensation (Perfetti and Stafura 2014, p. 26). As such, the framework is more detailed than the SVR, making it possible to formulate specific hypotheses about what influences reading comprehension. At the same time, while the C-I model is much more detailed in its predictions on how language comprehension is achieved, the RSF includes

²⁷The authors implicitly distinguish between *models*, *theories* and *frameworks* to argue their point, defining models as the most specific unit, from which “precise predictions” can be made (cf. Perfetti and Stafura 2014, p. 23). From this point of view, the two *models* described above should actually be considered theories. However, there does not seem to be a strict consensus on the definitions of the three terms in the field. Hence, this thesis does not follow this terminology. The interested reader is referred to an overview on the topic stemming from the field of health care (Nilsen 2015).

²⁸It is not entirely clear why lexical knowledge is not included with this knowledge source, except for the fact that the lexicon is its own component.

²⁹It needs to be pointed out, however, that the model is clearly not a model of learning: arrows towards the linguistic knowledge sources are not included, which ignores the possibility that readers will also add to their knowledge bases while reading. Especially in the context of reading in an additional language, these arrows may be a fruitful addition, as they would reinforce the significance of “frequency of input” for improving L2 competences, including reading comprehension (cited in Grabe 2009, pp. 60 sqq.).

³⁰The Memory component is “the only language-specific component of the model” and contains “phonological, morphological and syntactic building blocks” (Hagoort 2013, p. 1).

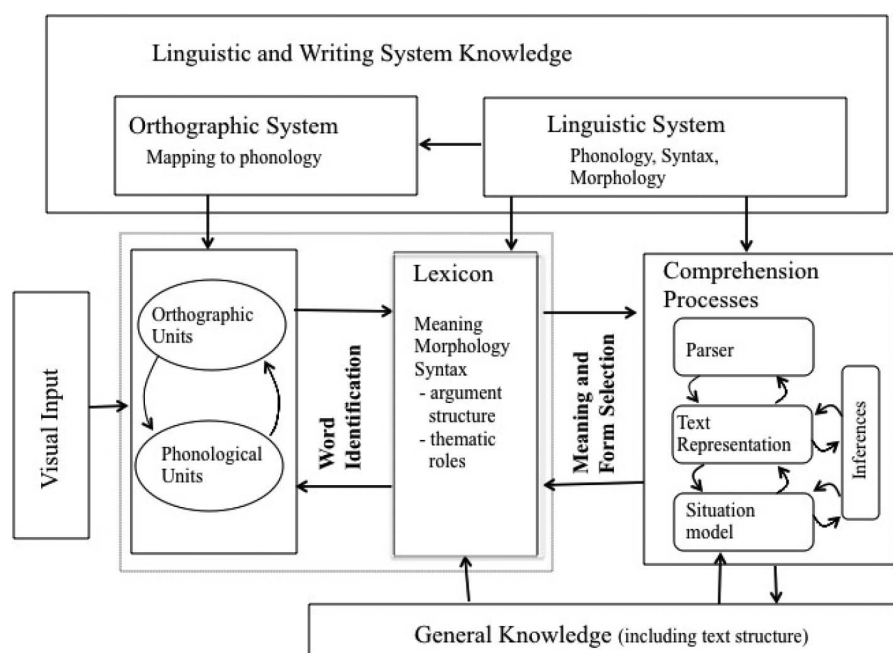


Figure 3.2: Perfetti & Stafura's Reading Systems Framework (2014, p. 24).

dedicated knowledge sources and lower-level word recognition. This makes it possible to model the specifics of L2 reading comprehension more explicitly. As such, it can be considered a component model of reading.

Looking back at the requirements of a comprehensive theory of reading, however, the RSF is clearly missing the *product* part of the reading construct: while Hagoort's MUC model recognises that some sort of action may result from language processing and that understanding meaning also implies understanding "speaker meaning" (2013, pp. 10 sq.), Perfetti and Stafura did not include the result of reading comprehension in their framework (with the possible exception of the integration of new knowledge into the general knowledge system).

3.1.4 The Cognitive Model of Reading for assessment

An approach which *starts* at the product of reading is the definition of a reading construct in the context of (foreign) language assessment. The most notable one has been described by Khalifa and Weir (2009) as part of their socio-cognitive framework for conceptualising reading test validity.

In their text on the validation of an L2 reading assessment, Khalifa and Weir discuss each section of their socio-cognitive framework in a separate chapter. The chapter on cognitive validity reads very similar to a text which establishes a theory of reading, even though the goal is assessment validation. The resulting *cognitive model of reading* (see Fig. 3.3) is based on the same research and establishes elements which are quite similar to the models and theories on L1 reading discussed above. Nevertheless, the model has a strong focus on the observable, or rather, a structure which can be divided into elements that can potentially be assessed through particular tasks. For instance, the path from word recognition to meaning construction is clearly distinguished into steps: the word form, followed by the lemma (the word's meaning), syntactic parsing and the integration into propositions. In the same vein, the building of the text and situation models is detailed into four steps, each of which may be operationalised as a separate task. Knowledge sources are theorised to influence those steps in a manner similar to the RSF.

A notable aspect of Khalifa and Weir's model is the *goal setter* (ibid., pp. 44 sqq., based on S. Urquhart and Weir 1998). The idea of reading as a purposeful activity is not new (Grabe 2009,

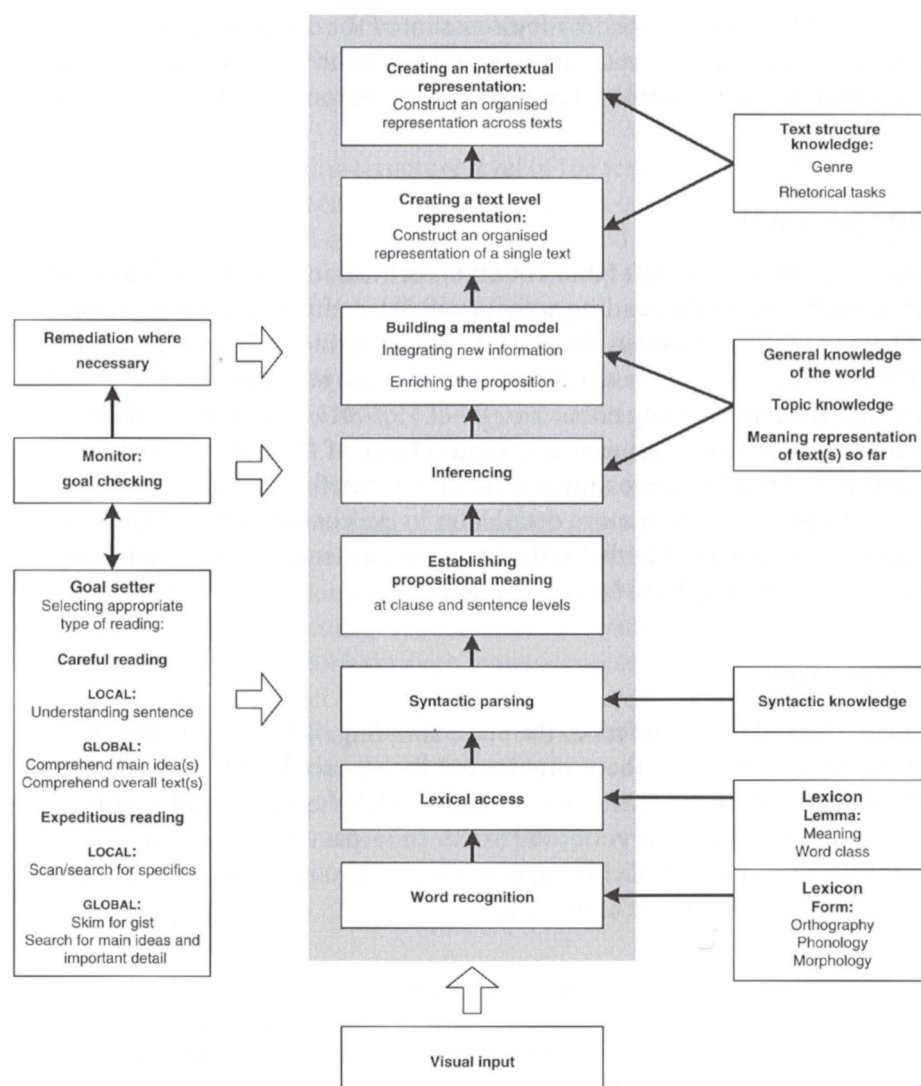


Figure 3.3: Khalifa & Weir's cognitive model of reading (2009, p. 43).

pp. 14 sq., 43 sqq.), though it has often remained implicit in theories of reading and is one of the aspects that theorists have identified to be missing from a comprehensive theory of reading. The function of the goal setter, according to Khalifa and Weir, is the choice of an “appropriate type of reading” in line with what one wants to achieve by reading in a particular moment: *careful* or *expeditious* reading may happen at the *local* or the *global* level (2009, pp. 44 sqq.). Careful local reading, for instance, may be employed when one wants to learn from a text and takes in every detail, while expeditious global reading may serve to establish the overall discourse topic through so-called *skimming* (see also Carver 1992). Of course, as with most things, there are entire continua hidden behind this two-by-two matrix. Assessments which are based on Khalifa and Weir's reading model (among them the ÜGK foreign language assessment) often pay particular attention to these types of reading, and operationalise them during item development in an attempt to ensure a complete representation of the reading construct (Konsortium ÜGK 2019a, pp. 39 sq., see also OECD 2021, pp. 41 sqq.). Finally, the cognitive model of reading makes explicit mention of a *monitor*. It mediates and remediates the entire process continuously, adding another facet of the reader to the model. This monitor is often considered in terms of an executive control device, an inherent part of any cognitive process (McNamara and Magliano 2009, pp. 354 sq.; Grabe 2009, pp. 67 sq.; Hagoort 2013).

3.1.5 Conclusion

Four approaches to thinking about reading have been presented. The Simple View of Reading is meant to be general enough to describe the necessary capacities that need to be developed in order to become a skilled reader: word recognition skills and linguistic comprehension skills. The C-I model, in contrast, describes reading as a cognitive process within a large neuronal network. It is detailed enough to allow computational applications: the reading process can be described mathematically, which makes it possible for it to be modelled by computers (e.g. Sanjose, Vidal-Abarca, and Padilla 2006).

While both approaches are prominent examples of models of reading, and have been widely discussed, neither offers the tools to deconstruct the reading process into elements which can be operationalised to determine the sources of difficulty that test takers may encounter in a reading assessment. For this purpose, two additional approaches to modelling reading have been presented—the Reading Systems Framework and Khalifa and Weir’s cognitive model of L2 reading. These two models of reading differ in the details, but they both presuppose that, at least for measurement purposes, reading comprehension can be distinguished into components. These components will be discussed in more detail in the following sections. After that, the differences and commonalities between L1 reading, which has been the focus up to this point, and L2 reading, which is actually of interest in this thesis, are considered in section 3.3.

3.2 Components of reading

Dividing reading comprehension into separate skills has practical advantages, for both research and assessment purposes (cf. e.g. Koda 2007, pp. 3, 29; Grabe 2009, p. 22), though the division is almost certainly theoretical to some extent: the various knowledge sources described in the following overlap at some points and cannot always be separated entirely (see e.g. Wisniewski 2012, p. 37). In addition, communicative competence encompasses more than just knowledge. Competent users of a language also need to have the ability to use the available knowledge efficiently and fluently and to compensate for communicative difficulties (e.g. lack of knowledge, but also misunderstandings, pragmatic adaptations etc.)³¹.

As mentioned above, some models of reading (including the RSF and the model by Khalifa and Weir) explicitly mention the skills and knowledge sources that are relevant for successful reading comprehension. They differ in the details, but overall, the following linguistic components of reading (i.e. lower-level processes) can be identified and will be discussed in the following (see also Grabe 2009, pp. 39 sqq.; Koda 2007, p. 4): orthographic and phonological knowledge which allow for word recognition (section 3.2.1), semantic and lexical knowledge along with morphological knowledge, which make up word identification (section 3.2.2), and syntactic knowledge, which enables comprehension at the level of the sentence or phrase. Beyond this, higher-order comprehension processes come into play to establish propositional meaning, a text model and, finally, a situation model (section 3.2.3).

In addition to having knowledge *of* the language, knowledge *about* the language facilitates reading comprehension, especially in L2 reading situations (Kieffer and Lesaux 2012, p. 348). This is usually described as a form of metalinguistic awareness about the inner workings of (a) language and will briefly be discussed in section 3.2.4.

Most of the aspects discussed in the following have been established for competent or emerging L1 reading. Nevertheless, at several points, the specificities of reading at a low proficiency level in the L2 are already considered, and the focus lies on the components which will be more relevant in this context.

³¹Wisniewski distinguishes three components in her working model of a lexical competence: *lexikalisches Wissen*, *strategische Kompetenz* and *Zugriff auf das mentale Lexikon* (2012). Though her discussion focusses on lexical competence, the contributions of *strategies* and *access* to the various knowledge sources are well described and can also be applied to other aspects of communicative competence.

3.2.1 Word recognition

Word recognition, or decoding³², combines the knowledge about letters and the ability to very quickly (i.e. within milliseconds) recognise words, analyse their components (e.g. morphological elements) and link them to their sound equivalents. It has been a major focus of reading research for several decades, both as an important aspect of reading in itself (as is most visible in the importance given to it in the SVR) and as a source of reading problems encountered by learners (cf. Jeon and Yamashita 2014, p. 163).

Word recognition is often separated into the stage where letters and/or words are visually perceived and further stages where their meaning is selected and then integrated with the context. During the actual process of reading, however, this distinction may be mostly artificial:

[...] word identification is not simply a matter of lexical access. Rather, it is a complex process that responds to different influences at different stages. These stages, however, are merely convenient verbal labels. In fact, processing is continuous, and there is significant temporal overlap between the different subprocesses defining these stages. In the first stage [...], the number of word candidates consistent with the perceptual input is progressively reduced through perceptual feature analysis. [...] This process rapidly reduces the number of word candidates to some manageable number, but not necessarily to one. At this point (probably at about 50 ms, [...]), the semantic context comes into play. (W. Kintsch 1988, p. 170)

Despite this amalgamation of processes, the part of word recognition that comes before the selection of meaning is often discussed separately from other processes of reading comprehension (Grabe 2009, p. 27, see also Taylor and Perfetti 2016, p. 1070): Jeon and Yamashita “define decoding as a process during which a reader converts letters (graphemes) to sounds (phonemes) and, essentially, to language” (2014, p. 163).

Word recognition is usually considered to require both orthographic and phonological knowledge. Koda defines *orthographic knowledge* as the “effortless extraction of phonological and morphological information from a printed word” (Koda 2007, p. 4). *Phonological knowledge* is the ability to recognise the sounds of a language and decompose them into their phonological constituents. While in an L1 context, this is clearly an ability related to speaking and starts to develop before an individual learns to read, there is strong evidence that phonological knowledge is a good predictor of reading comprehension, no matter the language (cf. *ibid.*, p. 5).

Finally, *fluent* or *automatic* word recognition is a necessary condition for reading comprehension: there is no comprehension beyond a few words if word recognition requires most of a reader’s attentional resources and higher-order reading comprehension may be severely hindered if many words are processed with errors (cf. the Verbal Efficiency Theory, Perfetti 1985, cited in Walczyk et al. 2001, p. 750, see also Jeon and Yamashita 2014, p. 180).

³²In a recent publication on the SVR, the original authors exchanged the term *decoding* for word recognition in order to “[retain] neutrality about how it is achieved” (Hoover and Tunmer 2022, p. 400). This change takes into account that skilled readers do not necessarily decode (as in, recognise letters and sound them out to form words) most of the words they read, but instead recognise their shape.

3.2.2 Word identification

As described by Kintsch (1988, p. 170, see above), once a word³³ has been recognised, its sense must be made available for reading comprehension to proceed. Vital to this is lexical knowledge³⁴, which is “multidimensional, and [...] consists of a complex amalgamation of various types of knowledge” (Jeon and Yamashita 2014, p. 165, see also Kieffer and Lesaux 2012). Without lexical knowledge, there is no reading, even in the presence of word recognition skills (Koda 2007, p. 5; Jeon and Yamashita 2014, p. 165). Consequently, in Perfetti and Stafura’s RSF, the *Lexicon* is accorded a central place, linked to all other processes and knowledge sources involved in reading comprehension (Perfetti and Stafura 2014, p. 24, see also section 3.1.3). This interconnectness points to the fact that understanding words is clearly more than knowing them: lexical knowledge encompasses both broad and deep knowledge of words (Kieffer and Lesaux 2012, pp. 348 sqq., see also Perfetti 2007), which is linked to semantic knowledge. In addition, morphological knowledge is necessary to interpret the precise meaning of a word (e.g. the tense of a verb, but also the meaning of a compound word). These various knowledge sources will be discussed in the following sections.

3.2.2.1 Semantic knowledge

In reading research, semantic knowledge is not usually distinguished from lexical knowledge. In other fields, however, semantic knowledge is defined as the “knowledge about the meanings of words and objects” (L. Chen, Lambon Ralph, and T. T. Rogers 2017, p. 1). This definition goes beyond the knowledge of *words* and, instead, reinforces the idea that meanings (of words, but also concepts, objects, places etc.) are linked to each other and exist in networks (e.g. Thompson-Schill et al. 1997). Considerations of semantics may also include characteristics of languages which could be called morphosyntactic, though, again, with a focus on their *meaning* (see e.g. Odlin 1989, pp. 75 sq.; Levin and Hovav 2017).

In Khalifa and Weir’s cognitive model of reading, semantic knowledge is not mentioned at all, but probably implied within the lemma component of the lexicon (the *meaning*), and possibly the general and topic knowledge sources (Khalifa and Weir 2009, pp. 43 sqq.). Likewise, Perfetti and Stafura do not explicitly mention semantic knowledge, though, with reference to the C-I model and a summary of neuropsychological research (Perfetti, C.-L. Yang, and Schmalhofer 2008, cited in Perfetti and Stafura 2014, pp. 28 sqq.), they underline the importance of not only knowing a word’s meaning, but also the “associations that [it] has with other words in memory and with other words in the text” (ibid., p. 31). Grabe describes this as well, referring to semantic activation processes, which will support comprehension along with syntactic processing through spreading activation: “words that are recognized or activated spread some activation or energy to their semantic neighbors (collocates, similar meanings) in the lexical network when they are

³³The notion of the *word* is somewhat imprecise, even if one were to accept the more or less arbitrary form of the *written* word (i.e. “groups of letters between two blanks” (Bogaards 2001, p. 324)). As Bogaards points out, some words may take on multiple meanings, some only exist in combination with others, and others only have grammatical or pragmatic meaning (ibid., pp. 323 sq.). Instead of *word*, he suggests the term *lexical unit*, i.e. a unit of meaning that is composed of at least one word, but can encompass more (ibid., based on Cruse 1986). This reflects a more precise definition of the *meanings* that a reader has to know and understand. Nevertheless, given that the term *word* is still widely used in the literature, it will continue to be used in this text, with the understanding that there might be better denominations and that multi-word expressions are also a part of the lexicon.

³⁴Some authors (e.g. Hu Hsueh-Chao and Nation 2000; Kieffer and Lesaux 2012) use the term *vocabulary* (*knowledge*). Variations such as lexical or vocabulary *proficiency*, *ability* and *competence* can also be found, underlining that the knowledge of words should not be considered a fixed trait of a language learner, but instead a complex, context-dependent and ever-changing competence (cf. Read and Chapelle 2001; Wisniewski 2012). In this thesis, the term *lexical knowledge* will be used despite its shortcomings: many of the more advanced aspects of a lexical competence will not be observable with low-proficiency L2 learners. Furthermore, the test takers’ lexical knowledge can only be inferred from their hypothesised difficulties with words and expressions in the reading text, leaving little room for the more complex construct of lexical competence. See also chapter 5.

accessed. If a related word is then being accessed, it will be activated by the association with a previously activated word” (Grabe 2009, pp. 25 sq.).

This process supports word-to-text integration in skilled readers, whose more developed semantic knowledge makes the use of words in context for comprehension more effective and more efficient (cf. Perfetti, C.-L. Yang, and Schmalhofer 2008, p. 313).

3.2.2.2 Lexical knowledge

Lexical knowledge, if differentiated from semantic knowledge, represents the link between a word’s form (phonological and/or orthographic) and its meaning. The meaning component of this knowledge—the lemma—contains “information on word class and the syntactic structures in which the item can appear and on the range of possible senses for the word” (Khalifa and Weir 2009, p. 49).

For reading to be efficient, lexical knowledge must be both automatised and of high quality (Taylor and Perfetti 2016; Grabe 2009, pp. 26 sq.). What exactly comprises high quality lexical knowledge depends on the text and on what the reader wants to achieve by reading. Nevertheless, if one is to consider a good overall understanding of most texts as the goal to achieve, then some rules of thumb can be established. Most obvious is the quantitative aspect: a reader has to know most of the words of a text in order to understand it. The exact amount is, of course, impossible to determine, but it is quite high: Hu and Nation suggest that 98% of the words of a text should be known for good comprehension (2000, p. 419), based on reading test results which proved unsatisfactory even at the 95% threshold. For most texts, this translates to 8000 word families that an individual needs to master in order to make (effortless) comprehension a probable outcome (Laufer and Ravenhorst-Kalovski 2010). As a result, since beginning L2 learners cannot be expected to have acquired so many words yet, reading assessments at low proficiency levels must compensate in the form of a limited vocabulary and/or higher redundancy in the reading text, additional context clues such as pictures or graphs, and reading tasks which specifically target text content that test takers are expected to know already (cf. Weir 2005, p. 48).

High quality vocabulary knowledge is also described as requiring deep knowledge of the words. This is most evident with words that may mean more than one thing depending on the context (see for various examples Bogaards 2001), but it is also linked to automaticity: well-developed lexical knowledge leads to more rapid and more precise activation of the intended meaning (Perfetti 2007, p. 359). This is achieved through a thorough representation of the orthographic and phonological form of the word, its grammatical constituents and its meaning(s) (i.e. its semantic component). Only if all of those attributes of a word are well-known, linked to each other and can be activated quickly and simultaneously, will word identification be efficient (ibid., pp. 360 sq., see also Laufer and Nation 2001). This kind of knowledge can only be acquired through repeated exposure, although instruction may support the learning process (Grabe 2009, pp. 28, 60 sq.; Koda 2007, pp. 5 sq., 10 sq.; Graves 1986). As a result, word recognition is more efficient for more frequent words, i.e. words that readers have encountered more often (Khalifa and Weir 2009, p. 49; Grabe 2009, p. 64). Hence, reading texts for readers at lower proficiency levels will need to contain more frequent vocabulary and represent the primary or most frequent meaning of a word or expression³⁵.

3.2.2.3 Morphological knowledge

Morphological knowledge is defined as the ability to analyse the constituting elements of words, the morphemes, and to adapt understanding accordingly (e.g. Grabe 2009, p. 23). In reading comprehension, morphological knowledge (or morphological *processing*, see Nagy, Carlisle, and

³⁵In foreign language contexts, the “primary or most frequent meaning” may not actually correspond to the use of the language in L1 contexts. Hence, foreign language assessments at lower proficiency levels may actually require an analysis of the instruction material in order to determine the vocabulary that learners have had a chance to acquire.

Goodwin 2014, p. 4) may serve grammatical functions, e.g. with respect to verb tenses or noun cases, but it may also be primarily semantic, for instance in analysing compound nouns or understanding meaningful affixes such as *dis-* or *un-* in English or German.

Either tacit or conscious, morphological knowledge is an important correlate of reading comprehension, which does not lose its importance even in advanced readers, and may even be a strategic resource for readers who lack linguistic knowledge in other areas (Nagy, Carlisle, and Goodwin 2014, pp. 3, 6 sq.). Furthermore, morphological knowledge is closely related to lexical knowledge in the sense that it “can help the reader recognize words that are partially known” (Jeon and Yamashita 2014, p. 168), which is especially useful in acquiring academic vocabulary (Nagy, Carlisle, and Goodwin 2014, p. 5).

3.2.3 Sentence-level and discourse comprehension

While the meaning of words is established, the competent reader parses them into larger units to reach comprehension across a text. This is necessary due to the capacity limitations of working memory (W. Kintsch 1995, p. 141) and it is a largely unconscious process. Similar to the division of word reading into a word recognition and a word identification component, the processing of larger units is often separated into two separate strands: syntactic parsing and semantic proposition building (Grabe 2009, pp. 29 sqq.; Dijk and W. Kintsch 1983, pp. 27 sqq., 37 sqq.; Khalifa and Weir 2009, p. 43). In this view, *syntactic parsing* would be the more “structural” aspect of understanding sentences, which involves the detection and identification of phrases and their syntactic function. Then, combining the results of word identification and syntactic analysis, the reader chunks the content from the text base into memorable meaning units to form *semantic propositions*.

3.2.3.1 Building the text model

Syntactic knowledge is mostly concerned with how sentences and their elements are put together, which encompasses knowledge about articles and pronouns, subject-verb agreement, verb tense and aspect, word order etc. (Jeon and Yamashita 2014, p. 165; Shiotsu and Weir 2007, p. 106). As such, syntactic knowledge is necessary to reach sentence-level comprehension and, from there, to ensure ongoing comprehension of a larger discourse (Taşçı and Turan 2021, p. 149; Aryadoust and Baghaei 2016). Grabe lists three major functions of syntactic processing: disambiguation (which includes the activation of more fitting lexical and semantic units), tracking of referents (e.g. pronouns, demonstratives and articles), as well as the selection and support of processing and repair strategies (2009, pp. 201 sqq.). As a result, competent readers may even use grammatical information (i.e. syntactic cues) in a top-down fashion to support lexical access and proposition building, which renders reading comprehension more efficient (Gernsbacher and Foertsch 1999, p. 287). This argument is reinforced by Nagy and colleagues, who illustrate the usefulness of syntactic knowledge when the meaning of a word, particularly a verb, is unknown (1995, p. 3), though in their empirical study, this strategy was most successful when their bilingual participants were already highly proficient in their L2 (ibid., pp. 8 sqq.).

*Semantic propositions*³⁶ are a widely used linguistic expression of meaning units in a text, which can be shown to be retained in memory as whole entities (J. R. Anderson and Bower 1980, cited in Dijk and W. Kintsch 1983, p. 39)³⁷. These meaning units are “generated by the relational properties between nouns, predicates, and modifiers” (André A. Rupp, Ferne, and H. Choi 2006, p. 445). The part of the comprehension process which begins with the building of semantic propositions is the one which has been the focus of most L1 reading models: “[...] all the most

³⁶Note that semantic knowledge as described above (section 3.2.2.1) and semantic proposition building as described in this section are not the same, though they are both related to the construction of meaning: semantic knowledge as a part of word identification functions at the level of words and singular concepts, while semantic propositions span clauses and sentences, connecting concepts with each other.

interesting processes psychologically seem to occur at the level of meaning and language use” (W. Kintsch 1995, p. 139). Accordingly, there is a variety of possible descriptions and models of this process, though many of them now assume a dynamic process which is generally compatible with Kintsch’s C-I model (see e.g. Gernsbacher and Foertsch 1999). The model suggests that propositional units are processed in cycles:

[...] we assume that a few propositions from the current processing cycles, which are considered to be the most important ones, are retained in the buffer to form a connection between the previous text and the current input. If there is no connection between the contents of the buffer and the new input, extra processing is required from the reader to make the textbase cohere. This may either involve a search of previous text which is no longer actively available in working memory, resulting in the reinstatement of a connecting proposition, or it may require a knowledge based inference to form a connection when it is not explicitly expressed in the text [...]. (W. Kintsch 1995, p. 141)

This process may happen at various levels of language: even though Kintsch mentions *propositions*, he underlines the fact that comprehension may be analysed at the word, the sentence or the paragraph level (*ibid.*, p. 141). This reinforces the notion that semantic proposition building is situated somewhat on the boundary between the text model and the situation model, serving comprehension at the local, but ultimately also at the global level.

3.2.3.2 Building the situation model

At this point, other knowledge sources may come into play, which include linguistic aspects such as knowledge about discourse markers and text-structure, but also general knowledge about the world or whichever more specific topic knowledge may be applicable (Koda 2007, pp. 9 sqq.; W. Kintsch 1995, p. 142; Khalifa and Weir 2009, p. 43). Whether or not this knowledge is useful, available and applied depends on the text, the reader and the current context—hence why situation models are highly individual. They may also be, as Kintsch points out in his discussion of understanding narrative discourse, culturally specific: “[...] we can understand, though maybe not ‘really understand’, a story from a different culture where we are not sure about people’s goals and reactions, or we can understand a story without thinking very deeply about it, without fully analyzing its implications” (W. Kintsch 1995, p. 142).

Going a step further, even more resources can be identified, as is well illustrated in neuropsychological models of comprehension (see e.g. Hagoort 2013; R. A. Mason and Just 2006): beyond the “strictly linguistic information”, which is analysed by a semantic processing network, a coherence monitoring network and a text integration network, neuroimaging studies reveal at least two more cognitive networks which support comprehension: one for “interpreting a protagonist’s or agent’s perspective” and a spatial imagery network (R. A. Mason and Just 2006, p. 765). This stresses the importance of *knowledge about the world*, without which reading comprehension will eventually fail.

At the intersection of all of those knowledge sources lies the ability for *inferencing*. This very broad human skill allows us to connect new information with existing knowledge, to fill in blanks or disambiguate vague statements, and to synthesise and evaluate information (cf. Grabe 2009, pp. 68 sqq., see also André A. Rupp, Ferne, and H. Choi 2006, p. 445). Basically, inferences make comprehension possible in the first place; without them, even understanding pronouns seems unlikely, let alone more complex, culturally specific texts loaded with irony or sarcasm (Strohner 2006, pp. 197 sqq.; Schnotz 2006, p. 236). In the context of L2 reading comprehension, Grabe

³⁷In Natural Language Processing (NLP) research, a proposition is more technically defined as “one predicate verb and a set of arguments”, which may be related to other propositions in some logical way (Stanovsky et al. 2014, p. 67). Automatic recognition of propositions is used to identify information in texts for automatic information extraction or summarization (*ibid.*, p. 67).

describes inferencing as an advanced skill which is linked to academic reading³⁸ or as a problem-solving strategy when poor readers are confronted with a difficult text. Though, clearly, the former goes well beyond the demands of a reading assessment at a low proficiency level, its importance for reading comprehension has led to its inclusion in many reading assessment constructs. This includes the ÜGK foreign language assessment (Konsortium ÜGK 2019a, pp. 39 sq.), where it has been applied in a basic form of identifying implied information. Grabe's second definition of inference as a compensation strategy, on the other hand, is very probably applied by many test takers even and especially at lower levels of language proficiency, both consciously and unconsciously. Its impact on assessment scores, however, is difficult to measure without further experimentation (Bachman 1990, pp. 105 sqq.).

3.2.4 Metalinguistic awareness

Nagy defines metalinguistic awareness as “the ability to ‘reflect on and manipulate the structural features of [...] language’” (2007, p. 53, based on Tunmer, Herriman, and Nesdale 1988, p. 136). This encompasses awareness of the inner workings of a language at all levels, ranging from metaphonological to metatextual (Gombert 1992, cited in Nagy 2007, p. 54). Nagy argues that metalinguistic awareness (or *word consciousness*) facilitates both reading comprehension and vocabulary acquisition and, hence, constitutes an essential component within the “vocabulary–comprehension relationship” (ibid., p. 54): both effective reading comprehension and vocabulary learning rely on a profound understanding of the concept of the written word, and on the ability to use word parts or the structure of a sentence for detailed understanding and/or to compensate for partial comprehension³⁹.

In a way, the advantages of metalinguistic awareness put forth by Nagy could also be subsumed under the knowledge sources discussed above—after all, good phonological, lexical or morphological knowledge at least implies the ability to *use* it well (see also Nagy, Carlisle, and Goodwin 2014, p. 4). Based on a similar consideration, Grabe distinguishes metalinguistic-analysis processes (the knowledge of linguistic structures) and metalinguistic-control processes (the ability to use this knowledge) from metalinguistic awareness—the *conscious* use of linguistic knowledge (2009, p. 132). In this sense, metalinguistic awareness may be more of a strategic resource which facilitates both language learning and language use, develops mostly through experience, but may also be taught explicitly.

3.2.5 Conclusion

Dividing reading comprehension into its various components underlines the complexity of the process: at the word recognition level, phonological knowledge and orthographic knowledge interact, followed by an interplay of semantic, lexical and morphological knowledge which allows for word identification and meaning integration at the local and, once syntactic knowledge comes into play, at the sentence and paragraph level. This, in turn, is used to create meaning across one or more texts, bringing in textual knowledge and knowledge about the world to form a large semantic network. All of this happens in real-time and within milliseconds, helped by a host of automatised processes and highly efficient parsing of information in working memory. Hence, experienced readers in particular are mostly unaware of many of those steps.

As Perfetti and Stafura pointed out, a component view of the reading process allows for the study of *pressure points* (2014, p. 26): if a knowledge source fails or a process requires

³⁸It “requires the evaluation of many different linguistic cues, the recognition of sometimes very subtle textual clues, the prioritization of potentially conflicting cues, or the synthesis of evidence from multiple texts to build critical reading comprehension abilities” (Grabe 2009, p. 70).

³⁹Nagy also includes the ability to understand definitions as a facet of metalinguistic awareness (2007, pp. 57 sq.). Arguably, this is somewhat different from, for instance, morphological awareness, and more closely related to strategic or textual competence. Nevertheless, the use of definitions in reading comprehension may be both a good instructional and a good assessment tool within the context of metalinguistic awareness.

more conscious attention than it should, reading comprehension is impaired. Perfetti and Stafura demonstrate how the Lexicon and its various subsystems function as pressure points in L1 reading (*ibid.*, pp. 26 sqq.), but the concept is particularly relevant in L2 reading as well, especially when considering readers whose language proficiency is (still) low: when a new linguistic system is acquired, reading comprehension is limited both by gaps in the various knowledge sources as well as by less effective processing due to limited exposure and practice. This, and the interplay of the various languages in a reader's linguistic repertoire, are the topic of the following section.

3.3 Reading in an additional language

Up to this point, the reading process of L1 and L2 readers has been mentioned interchangeably, as if the two were the same. This is not unusual: Khalifa and Weir's cognitive model of reading was conceived for an L2 reading assessment context, but it is based on L1 reading research (2009, pp. 1, 42 sq.). The authors justify this choice following Field, whose consideration of the cognitive validity of an L2 listening assessment is based on "[...] an empirically attested model of the target skill as employed by expert users under non-test conditions" (2013, p. 84). As such, the competence of an adult, highly competent L1 reader is considered "the goal towards which the L2 reader aspires" (Khalifa and Weir 2009, p. 43), which warrants the use of L1 reading models for considerations of L2 competence, especially when competent academic reading is targeted, as is the case in Khalifa and Weir's analysis.

Grabe goes a step further and argues that L1 and L2 comprehension rely on the same fundamental cognitive processes, "even if local processing details might vary across different L1s and across L1 and L2 contexts" (2009, p. 1)⁴⁰. The main difference, then, between L1 and L2 readers is the number of languages that they have to deal with. This is true with respect to the linguistic system, but also potentially the metalinguistic and strategic knowledge of the reader, the amount of language exposure they can rely on, the types of text they typically encounter, and the expectations towards print and text organisation they may have acquired in the course of their reading in various languages (*ibid.*, pp. 130 sqq.).

This text follows Grabe's conceptualisation of L2 reading as "reading with [several] languages" (*ibid.*, p. 129, see also Koda 2007; Koda and Zehler 2008). This is in line with the definitions of reading laid out above, which consider it a complex procedural skill requiring multiple knowledge sources and sub-skills. When an additional language is learned, these knowledge sources and skills are added to, enhanced, or expanded, similar to how other knowledge, skills and abilities are added to an individual's cognitive structure. Evidence for this view of the multilingual reader stems from research on L1 and L2 transfer, which shows that an individual's various languages are never entirely deactivated, even when they are not needed for communication (e.g. Shen and Del Tufo 2022) or that readers will use other languages in their repertoire to compensate for weaknesses while reading in an L3 (e.g. Peyer, Kaiser, and Berthele 2010). It is also strongly supported by neuropsychological research, which shows that L2 structures are processed by the same areas in the brain which process an L1 (see e.g. Abutalebi 2008; Grant, Fang, and P. Li 2015). Interestingly, this definition of multilingual reading has recently been shown to apply also in the field of NLP: artificial neural networks were successfully trained on multiple languages "without changes to the underlying architecture" (Dhar and Bisazza 2021).

⁴⁰In the past, L1 and L2 reading were considered to be more apart, which engendered a lively discussion on the transferability of L1 reading abilities and strategies. A Linguistic Threshold Hypothesis was postulated, according to which a certain level of L2 linguistic competence had to be achieved before L1 reading skills could be transferred (Alderson 1984). Another view allowed transfer at any time, but admitted that the efficiency of the transfer from the L1 will be limited if there was not enough L2 competence to transfer to (the Short Circuit Hypothesis, Clarke 1980). Neither of these views is necessarily incompatible with the definition of L2 reading as reading with several languages adopted in this thesis, since it allows for cross-linguistic transfer of abilities and strategies to be individual, context-dependent and erratic.

3.3.1 Cross-linguistic transfer in L2 reading

Grabe's definition of L2 reading as "reading with [several] languages" (2009, p. 129) also provides a reasonable explanation for how reading skills are transferred from one language to another while learning to read in an L2, and why research into the relationship between L1 and L2 reading comprehension is largely inconsistent, revealing numerous contextual and individual factors at play (Melby-Lervåg and Lervåg 2014, p. 411). These factors make it impossible to predict to what extent exactly reading skills are transferred from one language to another at any point in time (Grabe 2009, p. 131): individual differences in underlying abilities, such as working memory or phonological awareness, linguistic knowledge at all levels, but also general knowledge, are relevant predictors of reading comprehension and, as such, prone to change the way individuals perceive letters, words, meanings, sentences and texts. This is true within the same language and certainly across languages (Geva and Siegel 2000).

Any consideration of cross-linguistic transfer will have to be understood in this light: transfer is usually possible, but it does not always happen, it works in different ways depending on circumstances, and it is not always effective (Grabe 2009, p. 131, see also Ringbom and Jarvis 2009). Despite this, it is probably safe to assume that readers with several languages will use their multilingual knowledge to some extent, and that, as a result, cross-linguistic transfer will be observed, even if inconsistently. Indeed, Koda and Zehler (2008, p. 2) point out that second-language learning-to-read will be facilitated when the two languages involved are (more) related to each other. Verhoeven and colleagues also cite a number of studies which show that L2 learners will profit from L1 transfer once they have mastered the essentials of the L2, or if the two languages share these essentials (2019, p. 29).

In the following section, with the caveat of individual differences which are always at play, the transferability of the various components of reading comprehension discussed above, or rather, their role in the multilingual reader, will be briefly considered.

3.3.1.1 Linguistic knowledge

The various linguistic systems, especially of related languages, usually share some elements, while others are similar, or differ entirely (Grabe 2009, p. 110; Ringbom and Jarvis 2009, p. 107). Hence, some transfer is usually possible, while some aspects have to be learned anew (Grabe 2009, p. 145). Anyone who has learned another language has experienced this, but it can also be shown in experimental research. For instance, learners can be shown to perceive written words in a second language in different ways, depending on their L1 (Koda 2007, pp. 21 sq., see also Akamatsu 2003; Hamada and Koda 2008; Koda 1999). Findings from neurobiological research support this observation: multilingual individuals who read in two or more languages show brain patterns consistent with all the languages involved, not only their L1 (see for a review Shen and Del Tufo 2022). Lexical transfer is most obvious through cognates (manifested through the *cognate facilitation effect*, cf. Allen 2019a, p. 1122, see also Odlin 1989, pp. 77 sq.), though the lexical entries of an individual's various languages can be shown to form a multilingual lexicon from which readers draw knowledge dynamically, independently of whether the languages are typologically close to each other (Bot 2004). As a result, lexical transfer has a measurable effect on reading comprehension (for a short review, see Hancin-Bhatt and Nagy 1994). Morphological knowledge also benefits from prior knowledge in other languages, though experimental research suggests that learners, and in particular young learners, are less aware of morphological similarities than they are of lexical similarities (i.e. cognates) between languages (Hancin-Bhatt and Nagy 1994, pp. 303 sqq.; Nagy, Diakidoy, and R. C. Anderson 1993). A similar argument can be made for knowledge about text structure and discourse markers: Koda (2007, pp. 8 sq.) points out that these types of knowledge are developmental in nature. They rely on a lot of experience in (advanced) reading, which means that children and young adolescents have a limited knowledge base from which they can draw in L2 reading. Finally, there is ample experimental research

which shows that the syntactic structures of a multilingual speaker influence each other, though, again, with some individual variation, in particular due to L1 background and L2 proficiency (e.g. Hopp 2015, 2019; Jackson and Dussias 2009; Koda 1993).

Across all of these aspects, there is a lively debate on the mechanisms of multilingual transfer once a third language is introduced (see e.g. Cenoz 2013, pp. 7 sqq.). Most particularly, researchers have tried to determine which language serves as the main resource during the acquisition of a third language: an individual's L1, which is considered an individual's default language, or their L2, which is the most recently acquired language (cf. Hopp 2019, pp. 568 sq.). Not surprisingly, the results are inconclusive, though they point to typological effects: languages which are or seem to be similar serve as a model for a new language (the Typological Primacy Model, cf. Rothman 2011, see also Berthele, Lenz, and Peyer 2022). Again, though, the inconclusive evidence points to a highly individual and situation-specific phenomenon that cannot be described by simple cause-and-effect predictions.

3.3.1.2 Higher-order skills

Since it is possible to define more general comprehension skills such as semantic proposition building, inferencing or metalinguistic awareness independently of a specific language, it can be argued that they must be transferable in the sense that once the meaning of words and sentences in an L2 text can be deduced, individuals will build propositions, infer meaning and add new knowledge to their existing knowledge base without having been taught (again) (cf. Grabe 2009, p. 124; Alderson 2000, pp. 59 sq.). Another view would be to consider these higher-order skills as language-independent, or cross-linguistic skills which do not have to be transferred at all.

Of course, there are some aspects new language users have to learn, e.g. cultural details which can be considered semantic knowledge, or knowledge about metaphors, both of which may be necessary for some inferencing processes. Most of these aspects, however, may also be defined as a part of language-independent general knowledge, which is always available and, hence, does not disprove the idea of shared higher-order skills (Odlin 1989, pp. 71 sqq., though see Joh (2004) for the perspective of L2 learners whose L1 is culturally and linguistically distant)⁴¹.

This is further supported by research into the relationship between the two languages in reading: it can systematically be shown that the correlation between L1 and L2 reading proficiency is stronger at higher levels of L2 proficiency. This suggests that higher-order reading skills, which are more relevant in more challenging reading tasks, are shared between an individual's various languages, whether or not their application to either language is counted as transfer (Pae 2018, p. 66).

3.3.2 Reading at a lower proficiency level

Most of the thoughts and findings on reading comprehension laid out above were developed with advanced readers in mind, or, inversely, with regard to poor decoders (e.g. children and adults who have a form of dyslexia). The special circumstances of L2 readers who (still) lack the linguistic knowledge to read fluently, but who are perfectly adequate readers in their L1, have been examined to a lesser extent (though see a few recent studies on the applicability of the SVR to foreign language reading, e.g. Erbeli and Joshi 2022; Kang 2021).

Some inferences on the characteristics of lower-proficiency readers can be drawn from the discussion linked to Alderson's famous article "Reading in a foreign language: a reading problem or a language problem?" (1984). He discusses whether difficulties in L2 reading are more related to an individual's L1 reading abilities or to their L2 proficiency level. At lower levels of foreign

⁴¹This is not to mean that it is not necessary to teach these aspects of reading in foreign language classes, especially to young learners who are still in the process of acquiring them, and particularly with respect to aspects of metalinguistic awareness which are more salient in the presence of more than one language (Grabe 2009, pp. 128, 130 sqq.).

language competence, he hypothesises that it may be more of a language problem, whereas for more advanced readers, L1 reading ability may become a problem in L2 reading if it is limited in the first place (Alderson 1984, p. 24, see also Grabe 2009, pp. 146 sqq.). Linked to this is the idea that L1 reading skills and strategies can only be transferred once a certain level of L2 language proficiency has been reached, which was mentioned above⁴². Indeed, empirical studies show that L2 language proficiency is a better predictor for L2 reading than L1 reading, especially for less proficient groups (see for a review *ibid.*, pp. 146 sqq., see also Alderson, Huhta, and Nieminen 2016). Of course, as with anything else, empirical evidence also suggests that there is some individual variation to this, related to age (Grabe 2009, pp. 143 sq.), general cognitive skills and dispositions (P. Li and Clariana 2019), and the languages an individual has at their disposition (Alderson and A. H. Urquhart 1984; Peyer, Kaiser, and Berthele 2010). This leads Grabe to point out that a single specific linguistic threshold probably does not exist, even for a more or less homogeneous group of learners (2009, p. 148).

Nevertheless, it seems prudent to establish that most low-proficiency readers will be limited in their L2 reading comprehension by their linguistic knowledge, more so than by deficiencies in their L1 reading ability⁴³. Neurocognitive research further indicates that L2 learners are required to engage their language knowledge more actively than L1 readers, whereby their linguistic proficiency becomes less fluent and increasingly reliant on the effectiveness of their individual cognitive control network (Grant, Fang, and P. Li 2015, p. 472). In addition, research surrounding the Shallow Structure Hypothesis, put very simply⁴⁴, suggests that even highly proficient L2 language users have a tendency to use nongrammatical features, i.e. “semantic, pragmatic, probabilistic and surface-level information” (Clahsen and Felser 2018, p. 693), more than their L1 counterparts. This can be taken to mean that language users at lower proficiency levels will rely even more on word-level types of information, because their syntactic and morphological knowledge is not (yet) complete nor fluent enough to be used consistently in real-time processing (e.g. reading).

3.3.3 Conclusion

Overall, it seems clear that reading in an additional language is influenced by knowledge, skills and experiences both in an individual’s L1(s) and their L2(s). Linguistic and discourse comprehension skills are not developed anew and separately in an L2, but instead draw from prior knowledge and skills, both from the first language(s) as well as other second or foreign languages. This is especially true if the language systems are similar, and it is generally supported by good metalinguistic knowledge.

However, one of the main findings of research on reading in an additional language is its variability: although it is presented here as one concept, as in a large part of the literature, there are at least as many ways of reading in an additional language as there are of learning one. For instance, the young ÜGK test takers who have been studying a foreign language in a “drip-feed” language programme at school⁴⁵ have very different prerequisites when reading in their foreign language than, for instance, international students in a graduate school in an English-speaking country who have been reading in one language or another for many years, studied English for

⁴²For further reading, see the Language Threshold Hypothesis (Alderson 2000, p. 39; Alderson, Huhta, and Nieminen 2016) and the Short Circuit Hypothesis (Clarke 1980). See also Grabe 2009, pp. 140 sqq.; Koda 2007, pp. 22 sqq.

⁴³This is not to say that, in some cases, poor L1 reading ability may negatively influence L2 learning outcomes and that, as a consequence, poor readers in their L1 are susceptible to becoming poor readers in their L2, see e.g. Alderson, Huhta, and Nieminen 2016, p. 863.

⁴⁴For a more detailed exposition see Clahsen and Felser 2018.

⁴⁵As opposed to bilingual education, where a second language is the *medium* through which curriculum content is taught, “a ‘drip-feed’ language program [...] makes the language a subject in the curriculum similar to History, Science, Information Technology and Mathematics”, taught for a few hours each week at school in an otherwise majority language environment (C. Baker 2001, p. 200).

a long time and are now surrounded by academic and every-day English on a daily basis. In addition to that, individual characteristics matter: even with very similar linguistic biographies, readers may vary strongly on a cognitive and/or motivational⁴⁶ level, leading to very different reading comprehension performance in their L1 and/or their additional language(s). Finally, in a context where the classroom is the main space for language input, differences between teachers, textbooks and classrooms may lead to large differences between individual learners. All of these factors have to be at least considered when trying to predict reading outcomes in an additional language.

3.4 Summary

It was laid out in this chapter that reading is complex, multilayered and individual, both with one and with several languages. Since it cannot be observed directly, we rely on theoretical models to describe its mechanisms, to explain why it sometimes fails (and why it does not) and how it can be learned.

Perhaps the two most prominent models of reading comprehension are the C-I model (e.g. W. Kintsch 1988; W. Kintsch and Dijk 1978) and the SVR (Hoover and Gough 1990; Sparks 2021). The C-I model, connectionist in nature, describes reading as a cognitive process involving a large interconnected and flexible knowledge network which enables processes at all levels of reading comprehension, from word recognition to the integration of a situation model. In the C-I model, reading comprehension is a highly individual cognitive process embedded in its context. The SVR takes a different approach and describes two components of reading: word recognition and linguistic comprehension. It presupposes that if either of those two components fail, either because they are not available or not automatised, reading comprehension cannot happen. This is particularly helpful in the context of learning to read, where educators are reminded that students must be supported in both areas.

Two additional models of reading comprehension were discussed: the RSF (Perfetti and Stafura 2014) and a cognitive model specific to L2 reading (Khalifa and Weir 2009). Both establish components of reading beyond the SVR, with different emphases. In the RSF, a central function is carried out by the Lexicon, which is connected to lower-level word recognition and higher-level comprehension processes. Perfetti and Stafura (2014, p. 26) posit that if a weakness is introduced at any point in the system, reading will become harder. As such, the RSF is also related to cognitive theories of reading comprehension. The dedicated *L2 model* by Khalifa and Weir (2009), then, describes reading comprehension as a process which can be divided into separate steps. Various knowledge bases, a goal setter and a monitor carry out functions alongside this process. Given that this model has been described as part of a test validation effort, it focusses heavily on the observable, making it possible to describe assessment tasks which operationalise parts of the model.

Irrespective of the model, reading comprehension is usually described as a process which can be roughly categorised into lower-level processes which enable word- and sentence-level comprehension, and higher-level processes, which establish comprehension at and beyond the propositional and textual level. For reading comprehension to succeed, various types of knowledge and the ability to use them are necessary: phonological and orthographical knowledge allow for word recognition, lexical knowledge including semantic knowledge and morphological knowledge enable the identification of lexical units, and syntactical knowledge is necessary to put together semantic propositions. At this point, readers need to invoke other types of knowledge about texts and the world in order to build a well-informed situation model of a text. At all moments, metalinguistic awareness and strategic competence have the potential to facilitate reading comprehension.

⁴⁶Since the ÜGK data does not allow for a more detailed insight into the individual test takers' thought processes and psychological state during test administration, this text mostly eschews motivational and emotional aspects of reading. The interested reader is pointed to Graesser and D'Mello 2012; Zaccoletti, Altoè, and L. Mason 2020.

While in the C-I model, knowledge and skills related to more than one language can be considered part of the cognitive network which enables both reading comprehension irrespective of the language of the text, the component models do not explicitly mention several languages. Instead, it seems to be assumed that each language has “its own” components which interact in the way prescribed by the model. This is, of course, a simplification; it can be shown both theoretically and empirically that reading is not a matter of separate knowledge areas between which skills are transferred. Instead, the conception of “reading with [several] languages” (Grabe 2009, p. 129) is much more in line with what is observed. This view is also helpful in considering low-proficiency L2 reading: if reading comprehension draws on multiple sources of knowledge, skills and abilities, it becomes quite straightforward to model reading comprehension where word recognition skills and higher-order comprehension and inferencing skills are well-developed, but linguistic knowledge is missing or slow to activate.

Chapter 4

Measuring reading

Since reading is a cognitive activity that happens mostly within the human brain and cannot be observed easily (if at all), those who want to assess reading comprehension have to rely on indirect measures. Most often, this is achieved by administering an assessment of some kind, e.g. by asking questions about what was read, or through other ways of eliciting what was understood during reading (e.g. have readers write a summary or a review, do a presentation in class, rewrite the text into another genre). In the ÜGK foreign language reading assessment, multiple-choice questions in the language of schooling were formulated for short target-language input texts. The answers to these questions serve as the actual empirical evidence of the test takers' reading comprehension abilities. This is why, besides the process of reading itself, the process by which test takers reach their answers also needs to be discussed at this point.

In his well-known book on reading assessment, Alderson (2000) discusses three groups of variables which influence the reading process in an assessment context: those which are related to the reader, those which stem from the text (ibid., p. 32), and variables related to the item(s) or task(s) (ibid., pp. 85 sqq.). This is in line with other conceptualisations of language assessment⁴⁷. For instance, Weir's socio-cognitive validation framework for reading (2005, p. 44) separates the elements of an assessment before a response is reached into 1) test taker characteristics, 2) the demands and physical setting of the task and 3) the internal processes that a test taker has to work through. This latter element, which includes aspects such as language or content knowledge, monitoring and goal setting, can be reinterpreted as characteristics of the input text: the knowledge and strategies necessary to solve the task are activated as a function of the input text. Finally, these three groups of variables, conceptualised as *reader factors*, *text factors* and *task factors*, are also identified in the PISA 2018 reading literacy construct (OECD 2019b, p. 31).

The following sections will discuss the effect of each of these variable groups on the difficulty of a test task or item: variables related to the reading text or passage (section 4.1), variables related to the task or item with a focus on MCQs (section 4.2), and a short foray into variables related to the reader (section 4.3). For the sake of clarity, this discussion will allow for each of those groups to be considered on their own, although it is clear that solving a reading assessment task is influenced by all of these variables simultaneously, to varying degrees and in interaction (not to mention by other factors such as the test administration setting, characteristics of the classroom or pure chance).

⁴⁷In the field of cross-cultural psychological assessment, a similar consideration is concerned with the concept of equivalence and bias (for a good overview cf. Van de Vijver and Tanzer 2004). Given the nature of psychological assessments and surveys, this field focuses on the equivalence of constructs, and sources of bias in test methods, test administration and items. As such, some aspects are recognized as possibly introducing bias which are rarely discussed in language assessment, especially communication problems between test administrators or the test itself and the respondents, which may be due to cultural differences, but also to social inequalities, different responding styles, social desirability effects etc. (ibid., p. 124). While these aspects are important and often overlooked, they are not discussed in more detail here: the ÜGK dataset does not really allow for their consideration.

4.1 The input text

As mentioned, Alderson defines reading as an interaction between the reader and the text (2000, p. 32, see also Strohner 2006, pp. 189 sq.), which means that the two cannot easily be considered separately. For instance, whether or not an assessment topic can be considered *familiar*, which is a typical description of suitable input texts in low-proficiency assessments, clearly depends on the test takers. Furthermore, the effect of the text genre (including a distinction between literary and non-literary texts) has proven to be inconclusive, linked to topic familiarity and, again, dependent on the reader's experience (Alderson 2000, pp. 64 sq.). The same argument can be brought forth with respect to the purely linguistic characteristics of a text: there are no syntactic structures, lexical units, or even context factors which can be shown to be systematically more or less difficult to all readers (ibid., pp. 68 sqq.). Overall, Alderson identifies only one aspect of the input text that can be assumed to support comprehension in most cases: transparent text organisation, resulting in greater text coherence (ibid., pp. 67 sq., see also W. Kintsch and Yarbrough 1982; Kobayashi 2002).

4.1.1 Text organisation

Alderson defines *text organisation* as “how the paragraphs relate to each other, how the relationships between ideas are signalled or not signalled” (2000, p. 67). Grabe describes how *discourse structure* “helps readers build coherent representations of texts” (2009, p. 244). This discourse structure is understood in terms of a text hierarchy: at the top level, a text usually belongs to a genre, contains certain main ideas, and follows one or several “patterns of discourse organization” (ibid., pp. 248 sqq.). Below this level, Grabe describes various linguistic systems which support reading comprehension: cohesion, information structuring, lexical signalling, anaphoric signalling, topic continuity systems and text coherence. Together, they represent the multiple ways in which texts convey the information they contain, which is then interpreted by the reader who, in addition to the textbase, uses their knowledge about the topic (and about texts) to form a mental representation of the text content (ibid., pp. 244 sqq.).

In the context of L1 reading, text organisation and its effect on text comprehension and learning have been a major focus of research for several decades (e.g. Beck, McKeown, Sinatra, et al. 1991; W. Kintsch and Yarbrough 1982; McNamara, E. Kintsch, et al. 1996; O'Reilly and McNamara 2007; Slater, Graves, and Piché 1985). Overall, researchers could show that a destruction of the propositional structure of a text unavoidably leads to inferior comprehension (e.g. W. Kintsch and Yarbrough 1982), while the opposite generally improves reading comprehension (e.g. Beck, McKeown, Omanson, et al. 1984; Beck, McKeown, Sinatra, et al. 1991, see also O'Reilly and McNamara 2007, p. 123). At the same time, there seems to be a limit to the amount of text cohesion that readers can deal with: psycholinguistic research has shown that if two sentences were manipulated to have a high degree of relatedness⁴⁸, recall diminished (cf. Todirascu et al. 2013, p. 11). Interestingly, awareness of text structure could be shown to be a characteristic of good readers: if cohesion was changed in experimental studies, good readers tended to show a stronger effect (Grabe 2009, p. 251; Golden, Haslett, and Gauntt 1988, pp. 141 sq.; Kobayashi 2002, p. 206).

Apart from reading skill, prior knowledge was also found to consistently interact with text organisation: McNamara and colleagues found that prior knowledge can compensate for less well-developed reading skills, and thus allows for improved understanding of less coherent texts (though it also interacts with reading skill, cf. McNamara 2001; O'Reilly and McNamara 2007). Also, the inclusion of *explicit* cohesive cues may benefit readers who have less prior knowledge whereas readers with more prior knowledge may even be disadvantaged (McNamara, Louwerse,

⁴⁸E.g. Sentence 1: “Cathy felt very dizzy and fainted at her work.” as opposed to “Cathy worked very hard and became exhausted.” as a cause for Sentence 2: “She was carried unconscious to the hospital” (Myers, Shinjo, and S. A. Duffy 1987, p. 455).

et al. 2010, p. 293). Finally, text structure awareness is probably acquired during adolescence: older readers were found to be more sensitive to a manipulation of the text structure. Similar observations were made for written production (Golden, Haslett, and Gauntt 1988, p. 143).

Text organisation has been studied in at least two ways: either by manipulating reading texts in certain ways (e.g. Beck, McKeown, Omanson, et al. 1984; Beck, McKeown, Sinatra, et al. 1991; McNamara 2001; O'Reilly and McNamara 2007, see also McNamara, Louwerse, et al. 2010, pp. 292 sq.), or by identifying features which distinguished some texts from others (McNamara, Louwerse, et al. 2010; Todirascu et al. 2013). In both cases, researchers found that variables tended to interact with each other, making it impossible to determine precisely which features were related to text comprehensibility (cf. T. M. Duffy et al. 1989, p. 453). As a result, text organisation and cohesion are now operationalised as multivariate constructs: top-level variables include rhetorical structure, text genre, and topic, while lower-level descriptive linguistic variables range from counts such as the number of paragraphs in a text, the average sentence length in words or the incidence of negations per 1000 words⁴⁹ to compound measures such as lexical diversity indices or argument overlap variables produced with the help of Latent Semantic Analysis (LSA) (O'Reilly and McNamara 2007, see also Landauer and Dumais 1997; Landauer, McNamara, et al. 2011). In their validation study of Coh-Metrix, a well-known tool which assesses a multitude of linguistic characteristics of texts in English (Graesser, McNamara, Louwerse, et al. 2004; McNamara, Graesser, et al. 2014), McNamara and colleagues (2010) have shown that high- and low-cohesion texts on the same topic indeed vary along these various linguistic variables, confirming the multivariate nature of cohesion. They also stressed the close relationship between text cohesion and text difficulty (ibid., p. 318), making note of a more recent way of studying the comprehensibility of texts, which will also be adopted in this thesis: *complexity* (see section 4.1.3).

4.1.2 Topic and text genre

When we read a text, the information contained in it is integrated with our existing knowledge and experience (W. Kintsch 1988). How exactly a reader's prior knowledge interacts with the content of a text has been a subject of lively research in various fields (e.g. in relation to schema theory, see for overviews McVee, Dunsmore, and Gavelek 2005; Xie 2005), but the general consensus is that prior knowledge and text content (i.e. topic) necessarily interact with each other. For instance, studies on the effect of text coherence in relation with prior knowledge show that text comprehension is clearly a function of this knowledge, irrespective of, but in interaction with other text and reader variables (e.g. reading skill, cf. McNamara 2001; O'Reilly and McNamara 2007).

In the context of reading assessment, this means that the topic of a text is not without consequence, as readers who are familiar with a topic will read and understand a text in very different ways than those who do not, irrespective of their reading skills (Alderson 2000, pp. 61 sqq.; Weir 2005, pp. 75 sqq., for a recent empirical study see e.g. Shin, Dronjic, and Park 2019). For this reason, assessment developers have long tried to limit the amount of (specialised) topical knowledge that is needed to solve comprehension tasks: theoretically, if test takers are neither advantaged nor disadvantaged by their existing topical knowledge, the test results will reflect their language knowledge more adequately (Bachman and Palmer 2010, p. 216). Usually, this is achieved by choosing topics that all test takers are mostly familiar with, or, inversely, by choosing one that is thought to be equally unfamiliar to most test takers (Alderson 2000, pp. 61 sqq.). In addition, assessments usually cover more than one particular topic in order to further limit the distortion of the results which may arise from some test takers knowing more about certain topics than others (ibid., pp. 63 sqq.)⁵⁰. In a foreign language assessment, the choice of topics is

⁴⁹In some studies, these variables also include typographical or graphic devices such as the typeset or formatting of the text, the use of bullet points, arrows, colours etc. (e.g. Buck, Kumi Tatsuoaka, and Kostin 1997; White 2012).

⁵⁰Bachman and Palmer point out that topical knowledge can also be modelled as part of the construct or

further limited by the language proficiency of the test takers.

Many of the observations related to text organisation made above also apply to the effect of the text genre⁵¹ on reading comprehension. Grabe defines them as “certain conventions for organizing texts in ways which reflect group goals and purposes”, which have “[...]stabilized] into sets of genre conventions” (2009, p. 248). There are various taxonomies of text genre, with a common distinction between narrative and expository texts, which can be further divided into subtypes that all have some elements in common and differ in others (Grabe 2009, pp. 248 sqq.; Graesser, McNamara, and Kulikowich 2011, p. 227). In the context of reading assessment, for instance in PISA, an additional distinction is made between continuous and non-continuous (or linear and non-linear) texts, which reflects a somewhat broader understanding of *text*: non-continuous or non-linear texts are typically organised as lists, and contain incomplete sentences, but also graphic elements such as diagrams, pictures or even audio elements (OECD 2019b, pp. 39 sqq.; White 2012, p. 144). Such a construct of reading literacy includes on-screen reading and hypertexts, but also text types such as phone books or restaurant menus.

In both L1 and L2 reading, it has been shown that it is generally beneficial for readers to have knowledge about text genres and to be able to use strategies to identify and use relevant text structures for text comprehension (Williams 2007, p. 200; Grabe 2009, p. 249; Meyer et al. 2010, pp. 63 sq., see also Foschi Albert and Hepp 2014).

4.1.3 Linguistic complexity

While text organisation, text genre and text topic clearly play their role in reading comprehension, ultimately, they interact with the linguistic complexity of the text, i.e. its words, sentences and paragraphs. This is especially true in L2 reading, when readers tend to have less well developed linguistic knowledge to support their comprehension (Grabe 2009, pp. 129 sqq.). Clearly, then, any consideration of a reading task must take into account the linguistic complexity of the text: if it is beyond the readers’ abilities, even in parts, they will be unable, and, as a result, perhaps unwilling to interact with it.

4.1.3.1 Defining linguistic complexity

Linguistic complexity can be defined as “the observable variedness and elaborateness [...] of a text, which can then be interpreted in relation to the nature of the task for which a text is read or written, or the characteristics of the individuals engaged in reading or writing” (X. Chen and Meurers 2016, p. 113). This reflects the interactive nature of *complexity*, which can be seen as an absolute property (of a linguistic structure, a text or a language user’s linguistic system), but also as a relative characteristic, where the complexity of linguistic units lies in learning or using them (absolute vs. relative complexity, cf. Bulté and Housen 2012, pp. 21 sq.; structural vs. cognitive vs. developmental complexity, cf. Jiménez López and Torrens Urrutia 2018, pp. 484 sq.)⁵².

Bulté and Housen, in a review of L2 (absolute) complexity⁵³, further distinguish a *global* or *system complexity* and a *local* or *structure complexity*. The first describes the linguistic system

as a separate assessment construct next to language ability, depending on the objectives of the assessment and the target language use definition (2010, pp. 217 sq.). See also the Reading for Understanding framework (e.g. Sabatini et al. 2014), an L1 reading construct which considers background knowledge to be an integral part of literacy and, hence, explicitly advocates the assessment of topical knowledge.

⁵¹PISA or the German study *Deutsch Englisch Schülerleistungen International (DESI)* use the term *text type* (German *Textsorte*) instead of *text genre* (OECD 2019b, pp. 46 sqq.; DESI-Konsortium 2008).

⁵²In view of this, Pallotti suggests a “simple view of complexity”, limiting the use of the term *complexity* to its descriptive side (i.e. absolute complexity, see also Bulté and Housen 2012). In turn, he proposes to call the other types of complexity *difficulty* (relative complexity) and perhaps *difficulty of acquisition* (developmental complexity; Pallotti 2015, pp. 119, 123).

⁵³In addition to linguistic complexity, which is the focus of their review, Bulté and Housen also describe discourse-interactional complexity and propositional complexity (2012, pp. 24 sq.). Together, these three constructs make up L2 complexity.

of an L2 learner at large, while local complexity is defined as “a more stable property of the individual linguistic items, structures or rules that make up the learner’s L2 system” (Bulté and Housen 2012, p. 25). Both types of complexity can describe aspects of a learner’s linguistic repertoire in the various linguistic subdomains (the authors mention phonology, lexis, morphology and syntax). Each of these subdomains is further divided into three levels: a theoretical, an observational and an operational level. This stresses the importance of specifying the underlying construct one intends to measure and linking it to the language behaviour manifested in real life (ibid., p. 27). *Lexical* complexity, for instance, encompasses the breadth of an individual’s lexical knowledge as its systemic complexity, while its structural complexity is expressed in terms of the depth of single lexical units (fig. 4.1). On the observational level, these two theoretical constructs are manifested in four measurable subconstructs: lexical density, diversity, compositionality and sophistication. These, in turn, can be measured on the operational level using certain, more or less common, indices (see also section 5.2).

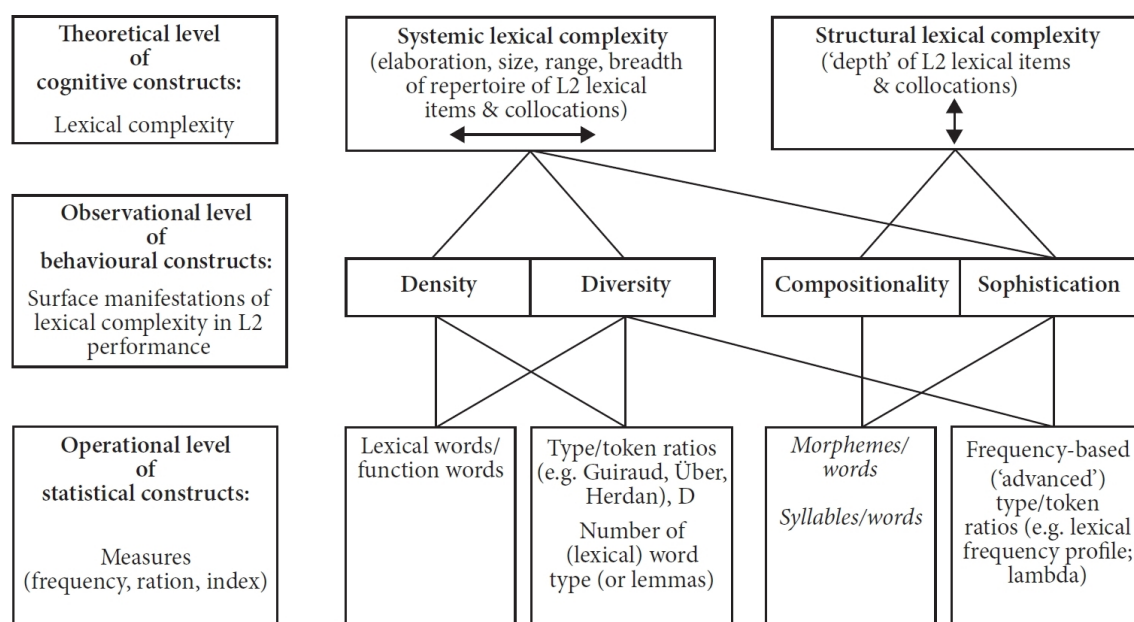


Figure 4.1: Lexical complexity at three levels of construct specification (Bulté and Housen 2012, p. 28).

4.1.3.2 Measuring linguistic complexity

Bulté and Housen’s exemplary construct of lexical complexity underlines the importance of multiple measures: instead of one measure which stands for the entire construct (e.g. a lexical diversity measure or a frequency-based index), the model mentions several existing measures exemplarily, with the clear understanding that the list is non-exhaustive. In light of this, it seems unlikely that the *linguistic* complexity of a text, either read or written by an L2 learner, can be expressed by one or a small number of indices⁵⁴. This is in line with the development of tools which automatically extract a large number of linguistic complexity measures using NLP, e.g. Coh-Metrix (McNamara, Graesser, et al. 2014) or the more recent and multilingual Common Text Analysis Platform (CTAP) (X. Chen and Meurers 2016)⁵⁵. These tools have been replacing simpler means

⁵⁴In some cases, as Jarvis (2013b) points out in his review of lexical diversity, human perception is reasonably efficient at analysing and describing relatively large constructs. Hence, human ratings of complexity may have some validity as single measures of linguistic complexity—though with a considerable error margin.

⁵⁵Apart from the two tools mentioned here, several others have been and are still being developed, e.g. the multilingual MultiAzterTest (Bengoetxea and Gonzalez-Dios 2021) or an unnamed linguistic analyser for French

of establishing text complexity, especially readability measures⁵⁶.

The current version of Coh-Metrix features over 100 linguistic indices at multiple levels of language and discourse: “(1) *words*, (2) *syntax*, (3) the explicit *textbase*, (4) the *situation model*, and (5) the discourse *genre and rhetorical structure*” (McNamara, Graesser, et al. 2014, p. 60, *italics in original*). The original version of the tool is only available for English, though derived tools have been developed for other languages (e.g. Almeida and Aluísio 2009; Leal et al. 2024; Quispesaravia et al. 2016). CTAP, on the other hand, was designed to be extendible in the first place, accepting additional linguistic features and languages (see e.g. Okinina, Frey, and Weiss 2020 for a detailed account of how the Italian language module was added to CTAP). As of December 2022, to the best of my knowledge, the publicly available CTAP contains well over 300 linguistic complexity measures for English and German, over 250 measures for Italian, including about 150 shared between the three languages, as well as almost 200 measures for Chinese, of which 40 are shared with the other three languages (cf. Cui et al. 2022; Okinina, Frey, and Weiss 2020; Weiss, X. Chen, and Meurers 2021). Table 4.1 contains an overview of the categories of available measures in Coh-Metrix and CTAP listed by the respective authors.

Though the terminology varies, both tools assess linguistic complexity in terms of lexical and syntactic complexity, as well as sentence, paragraph and text length. As far as lexical complexity measures are concerned, to keep with the example above, the tools measure the subconstructs depicted by Bulté and Housen (Bulté and Housen 2012, see fig. 4.1). CTAP also includes morphological complexity measures, which have been shown to be relevant in morphologically complex languages (François and Fairon 2012; Hancke, Vajjala, and Meurers 2012). Furthermore, both Coh-Metrix and CTAP contain measures which assess certain features of texts that have been shown to be related to text readability: markers of text and discourse cohesion, incidences of verbs which are used to structure certain text genres (i.e. markers of causality, temporality etc.), and more elaborated measures which reflect the semantic content of the text. The latter is achieved through the use of WordNets (Hamp and Feldweg 1997; Henrich and Hinrichs 2010; Princeton University 2010), and LSA (Landauer and Dumais 1997). CTAP also includes measures related to the Dependency Locality Theory (DLT) (Gibson 2000), which, very simply said, measures the cognitive load of understanding sentences, thus possibly reflecting text complexity as it is perceived by humans. Coh-Metrix even calculates so-called *text easability component scores* which output “information about text at multiple levels of linguistic analysis, including word characteristics, sentence characteristics and the discourse relationships between ideas in text” (McNamara, Graesser, et al. 2014, p. 83). These scores are based on a Principal Component Analysis (PCA) of the original Coh-Metrix indexes conducted with texts for children and young adults.

Studies have shown that the multiple measures included in tools like CTAP or Coh-Metrix are useful in tasks such as distinguishing texts for children from texts for adults (Weiss and Meurers 2018), classifying texts in terms of their “reading level” (Bengoetxea, González-Dios, and Aguirregoitia 2020; François and Fairon 2012; Hamada and Koda 2008; Weiss, X. Chen, and Meurers 2021) or assigning learner texts to a grade level (Weiss and Meurers 2019a). Coh-Metrix

as a Second language (FLE) (François and Fairon 2012). At this moment, it is not clear if one tool is superior to another. Hence, the choice to present CTAP and Coh-Metrix in more detail is mostly practical: Coh-Metrix is probably the most established tool, and is often cited by the authors of the newer tools. CTAP supports the languages of interest in this thesis and, being developed by a team of researchers at the German University of Tübingen, is also geographically and institutionally closest.

⁵⁶In the English-speaking world, especially in the US, readability measures have been developed since the 1920s (Dubay 2004, p. 13). Though widely used, and intensively researched, they have been criticized for representing a too narrow conceptualisation of text complexity (e.g. T. M. Duffy et al. 1989; O’Reilly and McNamara 2007), as even sophisticated readability formulas tend to predominantly reflect sentence and/or text length (Davison and Kantor 1982; Klare 1974; McNamara, Louwerse, et al. 2010). In German, French or Italian, adapted readability formulas have been used in research, but have never known the widespread use that they have in the English-speaking world (see e.g. Hancke, Vajjala, and Meurers 2012, p. 1056; François and Fairon 2012, p. 467). Readability formulas are now often one variable among others in linguistic complexity analyses (e.g. in Coh-Metrix: McNamara, Graesser, et al. 2014, pp. 78 sq.).

Coh-Metrix (McNamara, Graesser, et al. 2014)	CTAP (English & German) (Weiss, X. Chen, and Meurers 2021)	CTAP (Italian) (Okinina, Frey, and Weiss 2020)
Lexical diversity	Lexical complexity	Lexical features
Syntactic complexity, Syntactic pattern density	Syntactic complexity	Syntactic features
	Morphological features	
Descriptive	Surface length	Morpho-syntactic features Text-length features
Readability		Traditional readability indexes
Word information (e.g. POS, word frequency) LSA	Lexical use (i.e. word frequency)	
Referential cohesion, Connectives	Discourse cohesion	
	Human language processing (i.e. cognitive costs of discourse integration)	
Text easability principal components scores		

Table 4.1: Linguistic domains which are assessed in Coh-Metrix and CTAP (references in the table). Categories in one row refer roughly to the same construct, though the measures do not overlap entirely.

has also been used in item difficulty studies to account for the linguistic characteristics of the input texts and/or items (e.g. Aryadoust and C. C. M. Goh 2014; Révész and Brunfaut 2013; Toyama 2018, see also chapter 8).

4.1.4 Conclusion

Perhaps the most obvious conclusion from these reflections is the complexity of the matter: operationalising the characteristics of a reading text in terms of its structure, content and language is by no means a simple undertaking, and it is complicated further by its interactions with other aspects, particularly the characteristics of the readers and, in the context of an assessment, the task. It seems apparent that characteristics of a reading text will systematically influence reading comprehension, though in what way, for which reader and under which circumstances is not always entirely predictable. Any empirical study will therefore have to make a choice of the variables which best fit the texts under investigation, the assessment construct and the research question, all the while keeping in mind practical considerations. Some very general aspects are easily chosen: most texts can be characterised by descriptive measures such as text length, paragraph length and sentence length, the frequencies of certain categories of words (i.e. Parts of Speech (POSs)), which are related to the structure, genre and cohesion of a text (cf. Graesser, McNamara, and Kulikowich 2011, p. 225). Certain co-reference indices, e.g. the proportion of content words which are shared between pairs of sentences (cf. *ibid.*, p. 226), which have to do with the informational density of a text, may also be applicable to many texts. In the context of a reading assessment such as the ÜGK foreign language assessment, categorical

variables which distinguish the various text types may also be useful: the text genre and overall topic, with variable levels dependent on the texts themselves (e.g. a distinction into narrative and expository texts is only useful if a sufficient number of such texts are part of the corpus). Finally, as the various tools described above have shown, linguistic complexity is perhaps best operationalised as a multitude of variables, some of which will be heavily intercorrelated, while others will not distinguish between the texts in a specific study at all. Once applied to the data, they, and most other variables will have to be examined more closely for their meaningfulness, distribution, skewness and intercollinearity.

4.2 Task formats

It almost does not bear mentioning that assessment results are influenced by the choice of the task format⁵⁷ (Weir 2005, pp. 62 sq.).

First, the task format will impact the way a reader processes a text (passage). For instance, if a (perfectly proficient) reader is presented with a question about the general topic of a text, they will aim their reading process towards answering it, ignoring local details. Inversely, if they are asked to fill in blanks in a text, they will concentrate on the local meaning of the context around the blanks, possibly without ever noticing the wider context of the text (W. Kintsch and Yarbrough 1982; Alderson, Clapham, and Wall 1995, p. 44). Second, the task format will determine the type and amount of information that is available to make inferences about the reader's competence: though we may assume as much, a correct answer to an MCQ is not, in fact, irrefutable evidence that the test taker understood the reading text in the way the item developer had intended. Test takers may choose an answer option for a number of reasons, including (lucky) guessing, and these reasons are not reflected in the answer itself. On the other side of the continuum, a detailed excerpt of a reading text will probably contain a lot more information on a reader's understanding, but its successful production may rely on much more than reading comprehension (most notably, writing skills).

Alderson et al. termed this the “method effect”, pointing out that it should be reduced as much as possible:

We are not interested in knowing whether a candidate is good at multiple-choice tests, or can do cloze tests better than other candidates, or finds oral descriptions of a series of pictures particularly difficult. We are interested in finding out about the candidate's grammatical knowledge, or reading ability, or speaking skill. (Alderson, Clapham, and Wall 1995, p. 44)

In their manual on test item construction, Alderson and colleagues describe a number of “problems with particular item types” (ibid., pp. 46 sq.): items may require other skills than they are supposed to, test takers may have difficulties understanding what they are supposed to do, there may be more acceptable answers than intended, or too much room for (educated) guessing, and, finally, scoring an answer may not always be as straightforward as intended. Some of these problems can be reduced through careful item development and, more importantly, pre-operational testing (Kenyon and MacGregor 2012; Kremmel, Eberharter, and Holzknecht 2021). Nevertheless, since the test task provides the readers with a purpose for reading and the test administrator with information on the reading process, the method effect is unavoidable and, thus, knowledge about how it impacts the reading process is necessary.

In the field of (foreign) language assessment, the method effect has been known and researched for several decades (e.g. Cronbach 1971, cited in Bachman 1990, p. 244, see also Alderson 2000, pp. 84 sq.; Weir 2005, p. 62), though the findings remain inconclusive or unsatisfactory

⁵⁷The term *task format* refers to the overall nature of the tasks or items in an assessment (see also section 2.4), i.e. whether there is a question to be answered, a blank to be filled, a summary to be written etc. The same thing is sometimes called *test method*, *item type* or *item format*, depending on the source.

(Bachman, F. Davidson, and Milanovic 1996, pp. 126 sqq.; Alderson, Clapham, and Wall 1995, pp. 44 sq.). As a result, knowledge of the method effect is not always reflected in reading assessment practice (André A. Rupp, Ferne, and H. Choi 2006, pp. 448 sq.), and it has been mostly ignored in empirical research on reading comprehension outside the field of language testing (see e.g. Keenan, Betjemann, and Olson 2008, though see W. Kintsch and Yarbrough 1982).

A major discussion on the method effect, which is relevant to this thesis, is centred around MCQs. Two major lines of criticism can be identified.

One is somewhat technical: experimental studies have shown that MCQs are susceptible to being solved at a higher than chance probability even when the reading text is absent. For instance, using items from the Stanford Achievement Test (SAT), a high-stakes assessment for college admission in the United States (US), Katz and colleagues (1990) could show that even without the reading text available, participants were able to answer a substantial amount of items correctly. Even more strikingly, the authors found important correlations between the test item difficulty in their experiment and the item difficulties of the original test, indicating that answering an MCQ is facilitated by information outside of the reading passage (*ibid.*, p. 125) and that this effect is present whether or not the participants have access to the passage or not. Following this study, Freedle and colleagues did more thorough analyses of the sources of item difficulty in the SAT and TOEFL (Freedle and Fellbaum 1987; Freedle and Kostin 1991, 1993, 1996), and found that variables they considered to be linked to the reading text made up most of the variance explained (e.g. lexical overlap between the item and the input text, or the position of the necessary information in the input text; see also chapter 8). Following this finding, they argued that the process of solving an MCQ without its corresponding reading passage was simply different from solving it with the reading text available, invalidating Katz et al.'s analysis, at least to a point⁵⁸ (Freedle and Kostin 1994). Nevertheless, the development of MCQs clearly has to be done carefully, and the possibility of answering a question without referral to the reading passage always has to be taken into account, along with other technical aspects of MCQs, e.g. the position of the correct answer among the answer options (Hagenmüller 2021; Holzknacht et al. 2021; Lions et al. 2022).

The second type of criticism is concerned with the construct of answering an MCQ as opposed to reading a text: Rupp and colleagues (2006) argue that MCQs on a reading assessment do not elicit the same higher-order comprehension skills as reading in a non-test situation. Instead, respondents use problem-solving strategies, depending on the text and question type they encounter in the test. As such, reading comprehension tests may reflect some variant of reading comprehension, but the authors doubt that the construct is reflected in its entirety⁵⁹. This is also reflected by the strategies that the test takers in their qualitative study described: the MCQs were always part of the reading process, and the search for relevant keywords in the input text seems to have been a major strategy even for these more proficient readers (*ibid.*, p. 462, see also Barras 2021, pp. 183 sqq.). A similar argument is put forth with respect to the distractors: determining that an answer option is false is considered a cognitive activity mostly unrelated to a reading comprehension construct, or it is one that is seldom required outside the assessment context. In addition, distractors may add (erroneous) information, which would otherwise not be there and which may confuse test takers or change their reading experience (Weir 2005, pp. 62 sq.).

Despite these important drawbacks, MCQ is one of the most common test formats in education. It is considered cheap, both in terms of test administration and, even more importantly, scoring. It is easy to implement both on computers and on paper, and it allows the collection of a lot of data points in a short amount of time: in addition to reading the input text, respondents

⁵⁸The observation remains that in Katz et al.'s study, the item difficulties of the items solved without the reading text correlated quite strongly with the original item difficulties. This finding cannot be explained by a different cognitive process.

⁵⁹Even Freedle and Kostin admit that none of their findings show that test takers form a coherent cognitive model of the text by answering MCQs (1994, p. 110).

need only read the item and tick a box. Also, it is usually not necessary to explain to test takers how MCQs work. This saves time during test administration and it represents one fewer source of error compared to other task formats. Once a test has been administered, scoring can either be done by a machine, or it at least does not require trained raters, which is both cheaper and, again, less susceptible to (technical) errors. All of these aspects are related to the actual use of the assessment, and only very marginally to its construct validity. Nevertheless, since it is not easy to justify the use of an assessment which costs more than stakeholders are willing or able to spend (Bachman and Palmer 2010), its cost has to be taken into account and weighed against the drawbacks of the task format⁶⁰.

With respect to item difficulty, it is clear that all components of an MCQ may influence the process of finding the correct answer, in addition to and beyond the characteristics of the reading text. This means that a study of the causes of item difficulty has to include the formal and linguistic characteristics of the stem and the answer options, including the distractors. In addition, qualitative studies of test taker strategies suggest that the overlap between the item and the reading text, the position of the Necessary Information (NI), but also, for instance, the falsifiability of the incorrect answer options, are aspects which will be taken into consideration by test takers (see e.g. Table 15 in Barras 2021, pp. 184 sq.) and thus, will influence item difficulty (e.g. Buck, Kumi Tatsuoka, and Kostin 1997; Embretson and Wetzel 1987; Kostin 2004; Ozuru et al. 2008, see also chapter 8).

4.3 The test taker

Test takers will use their available resources—competences, knowledge, experience, strategies—to the extent that they deem necessary, which influences the way they read a text and solve an assessment item (Alderson 2000, pp. 33 sqq.; Krstić et al. 2018). In language assessment, the linguistic competences of the reader are the most obvious variable: they are the focus of the assessment construct⁶¹, but they are also an important confounding variable, especially in *foreign* or *second* language assessment. Given the reflections on the input texts in section 4.1, it is also clear that characteristics of the test takers such as their background and/or topical knowledge or their knowledge of text structures and text genre will play a role in reading comprehension. All of these aspects are at least partly linked to stable characteristics such as gender, age or personality, as well as socioeconomic status (SES) or educational background (Alderson 2000, p. 33; T. M. Duffy et al. 1989, p. 440).

As has been mentioned above (section 3.3.2), a long discussion in foreign language assessment has turned around the question of whether an individual's L2 reading skills are more related to their L1 reading skills or their L2 linguistic knowledge (Alderson 1984, 2000, p. 39; Alderson, Huhta, and Nieminen 2016; Grabe 2009, pp. 146 sqq.). Since the answer is probably *both*, L1 (reading) skills have to be taken into account when an L2 assessment is considered. Good L1 readers are generally described as having better word representations (i.e. sight word recognition, breadth and depth of vocabulary), making more inferences (i.e. integrating the text with prior knowledge), using more reading strategies and having more intrinsic motivation for reading (cf. O'Reilly and McNamara 2007, pp. 122, 125). Potentially, all of these advantages may transfer to an individual's reading skills in the foreign language, which may lead to better reading comprehension even when the knowledge of the L2 is limited (Koda 2007, pp. 23 sq.). Particularly in a classroom setting, better reading skills may also lead to higher learning gains, which may eventually give good L1 readers a lead over their peers in general L2 competence (for a

⁶⁰Unfortunately, the low cost of test administration and test scoring which results from using MCQs is often confused with a low cost of test development. Well-developed MCQs however, require experienced item writers, a lengthy item development process and possibly multiple rounds of piloting. Thus, the cost argument may not be as strong as it seems.

⁶¹In fact, language competences are an important variable even in assessments which do not focus on language, e.g. in mathematics (cf. Heine et al. 2018).

general introduction into the link between cognitive flexibility and reading competence, see e.g. Cartwright 2009). At a purely linguistic level, knowledge of another language may also directly influence reading comprehension by way of linguistic (dis)similarities, e.g. cognates or similar word order (cf. Grabe 2009, pp. 131 sq.), or, inversely, a different L1 script (e.g. Akamatsu 2003). All of these mechanisms point to a measurable correlation between L1 and L2 (reading) skills (and, possibly, skills in other languages). In reading research, this link has been shown to exist in a number of studies (cf. for references Koda 2007, p. 23; Grabe 2009, pp. 129 sqq.).

Apart from their L1 competence and their general reading skills, readers also need to activate other knowledge sources in order to understand the meaning of a text. Alderson (2000, pp. 34 sqq.) follows Carrell (1983) in distinguishing the knowledge of language and linguistic conventions (formal schemata) from knowledge of the world (content schemata), which he further divides into general background knowledge and topical knowledge relevant to the assessment at hand. In the context of L2 reading, these knowledge sources need to be expanded to include elements of the target language and culture(s) (Alderson 2000, pp. 45 sq.). As the various empirical studies of these issues have shown, an individual's level of knowledge is not easy to measure, and its effect on reading comprehension is always entangled with other factors⁶². As a result, assessment developers usually try to account for the test takers' presumed level of knowledge beforehand, e.g. by determining the text genres and topics that an assessment may cover, either because they are inherent to the construct at hand or because all test takers are expected to master them anyway. In item difficulty research, this usually translates to variables which cover the topic and text type of an input text.

Much like the test takers' knowledge base, their age, gender and, in some cases, educational background, also have to be taken into account during test development: these are usually the broad lines of the description of the test takers for whom a new assessment is intended (Khalifa and Weir 2009, p. 19). For instance, many international language assessments have a separate test version for children and/or adolescents (e.g. TOEFL), while others have clearly defined target populations (e.g. PISA, OECD 2019a, p. 157). Additionally, in an effort to provide equal opportunities for all test takers, test developers try to eliminate effects related to characteristics such as gender, religion or migration background, especially by choosing topics which are (supposed to be) appropriate for all test takers irrespective of their individual characteristics (Alderson 2000, p. 56; Khalifa and Weir 2009, pp. 18 sq.). To some extent, the same would be true for effects created by different educational opportunities, particularly the lack thereof, though mainstream assessments do not usually take this into account (though see Böhme, Heppt, and Haag 2017⁶³). In large-scale assessments, these factors are the target of DIF analyses aimed at identifying unwanted bias (particularly with respect to gender, see e.g. Khorramdel et al. 2020).

4.4 Conclusion

Three sources of variance in item difficulty were identified with respect to a reading assessment: the reading text itself, the task, and the test taker. Characteristics of the *reading text*—its organisation, text type, and topic—and their relationship with reading comprehension have been amply investigated in reading research, revealing clear effects. These effects, however, are always interconnected with other aspects of the reading situation, most notably the knowledge, skills and abilities of the *readers*. In language assessment research, the *method effect* is also recognised as a decisive factor in how reading comprehension is measured, and what conclusions can reasonably

⁶²In fact, as Alderson (2000, p. 40) points out, text comprehension is not only potentially aided by knowledge, but also the readers' ability to apply that knowledge, i.e. metalinguistic knowledge, or, following Cartwright (2009), their cognitive flexibility. Hence, it is not surprising that it is difficult to precisely measure *knowledge*.

⁶³Low literacy is also an assessment topic in itself (e.g. Mellard, Fall, and Mark 2009) and an ongoing discussion in health care (e.g. Mancuso 2009).

be drawn from it.

Once the actual process of administering a reading assessment, scoring it, and reporting on its results, is considered, more sources of variance become obvious: individuals may be tired or distracted by some life event, a test session may be interrupted by an unexpected event or consistently disturbed by noises, mistakes may happen during scoring, and even sophisticated psychometric analyses leave space for measurement error. Diligence and careful consideration can assuage problems related to many of these aspects, but it should nevertheless have become clear that the process of administering a reading comprehension assessment is a complex undertaking, and can never be explained by straightforward cause-and-effect relationships.

Therefore, the analysis conducted in this thesis can only hope to uncover some aspects of the complex ecosystem that is a reading assessment. The choice of variables is limited by the aspects which can be measured at all, the quality of their operationalisation and their variability within the dataset. While some formal characteristics of the input text and the items are relatively straightforward to operationalise, their linguistic complexity in particular is a large field that will be impossible to cover in its entirety; not least because the link between a reader's language skills and the complexity of the text are not fully understood. Hence, in the following chapter, a subconstruct of linguistic complexity will be examined more closely, which is hypothesised to be most relevant to this low proficiency reading assessment: lexical complexity.

Chapter 5

Measuring the lexical characteristics of reading texts

As was pointed out earlier (section 3.2.2), lexical knowledge is a necessary requirement for reading comprehension (cf. Perfetti 2007, pp. 362 sq., see also Aryadoust and Baghaei 2016). Studies of the link between lexical knowledge and reading comprehension have been conducted in various fields, leaving little doubt that the two skills are related, and that reading comprehension is low when lexical knowledge is weak (Jeon and Yamashita 2014, pp. 180 sqq.; Perfetti 2007, pp. 361 sqq.; Taylor and Perfetti 2016, p. 170), or, inversely, that a text which contains a lot of words unknown to the reader will be difficult to understand (Hu Hsueh-Chao and Nation 2000; Laufer 1992; Laufer and Ravenhorst-Kalovski 2010; N. Schmitt, Jiang, and Grabe 2011). Hence, in a low-proficiency context where readers know only a few words and have not had a lot of time to practice reading (i.e. to reach fluency), lexical knowledge may well be one of the main predictors of reading comprehension.

Following a short excursus into empirical research which investigated the link between reading comprehension and lexical knowledge, this chapter discusses the operationalisation of lexical characteristics of texts. In the last section of this chapter, some aspects of measuring crosslinguistic characteristics of texts, i.e. the linguistic distance of words and texts, are considered.

5.1 The role of lexical knowledge in reading comprehension: empirical findings

A major advocate of the importance of vocabulary in language comprehension is Charles Perfetti. Based on his earlier verbal efficiency theory (1985) and subsequent experimental research, he formulated the Lexical Quality Hypothesis (LQH) (2007), a model of (L1) lexical knowledge which deposits that, at any given moment, a reader's vocabulary contains words of widely varying Lexical Quality (LQ). LQ is defined as "the extent to which a mental representation of a word specifies its form and meaning components in a way that is both precise and flexible" (ibid., p. 359). This allows for a multidimensional definition of lexical knowledge, reflecting differences in familiarity with a word's meaning and its form (either or both phonological and orthographic). In addition, the LQH recognises that an individual's vocabulary may contain words which are well-known, not known at all, and anything in between. Experimental studies surrounding this theory (cited in ibid.) have shown that skilled comprehenders perform slightly better in word recognition, word differentiation and even in word learning, which underlines the close, possibly interdependent relationship between reading and vocabulary.

The LQH served as a foundation for the Ontogenesis Model of the L2 Lexical Representation (OM) (Bordag, Gor, and Opitz 2022). This model of L2 (!) lexical acquisition, in addition to the linguistic domains (semantic, phonological and orthographic) and their interconnections, includes

a “dimension of networks”, which describes the connections of lexical representations with others, both in the L2 and the L1 lexicon. Bordag and colleagues stress that vocabulary acquisition (and, interestingly, attrition) evolves dynamically along all of those lines of lexical knowledge, which makes for an L2 lexicon characterised by “fuzziness” (Bordag, Gor, and Opitz 2022, pp. 189 sq.). This is an important caveat to empirical research designs looking into the link between reading comprehension and lexical knowledge: both constructs are naturally multidimensional and fuzzy, which has to be taken into account when developing instruments, scoring answers and interpreting results.

Beyond discussions surrounding theoretical models of lexical proficiency, researchers have investigated the link between lexical knowledge and reading comprehension for a long time. Graves (1986, pp. 58 sqq.) reviewed several studies going back as far as 1938, in which young readers were presented with different versions of reading texts that contained more or less frequent words. Reading comprehension was measured with MCQs. Though the results vary in the details, and the operationalisation of both lexical frequency and reading comprehension may have been questionable in some cases, most of these studies found that the lexical frequency of a text’s words influences reading comprehension.

Studying advanced foreign language learning, Laufer (1992) showed that reading comprehension is dependent on vocabulary knowledge in the sense that if a certain threshold of lexical knowledge is not reached, it will fail. Inversely, knowing enough word families will almost always lead to successful reading⁶⁴. In addition, general academic ability, including L1 reading, may help readers whose vocabulary knowledge is between the two thresholds (*not enough* vs. *entirely sufficient*), but does not play a role otherwise (ibid., pp. 99 sqq.). Schmitt and colleagues (2011) conducted a similar study and found a mostly linear relationship between their reading comprehension measure and the lexical coverage of the texts. Interestingly, they hypothesise that their results may also apply to low-proficiency learners, provided that the texts be less demanding in terms of vocabulary.

More recently, Verhoeven and colleagues (2019) have shown that reading comprehension in early first and second language readers in the Netherlands is largely predicted by the quality of their early Dutch lexicons⁶⁵ in kindergarten. Similar findings have also been reported in empirical research surrounding the SVR, although these studies usually measure some form of general linguistic comprehension instead of lexical knowledge (see for references Verhoeven, Voeten, and Vermeer 2019, pp. 28 sq.; Sparks 2021, pp. 509 sq.).

In a study focussing on younger foreign language learners with little exposure to the target language outside their classroom, Kang (2021) could show that lexical knowledge was strongly related to reading comprehension, along with decoding skills⁶⁶ and reading fluency. A similar finding was reported by Erbeli and Joshi (2022) for Slovenian 7th-graders learning English⁶⁷.

Hacking and Tschirner (2017) investigated the link between reading comprehension and lexical knowledge in the context of learners of *Russian* in the US. Their participants were adult learners in higher education who scored along the entire proficiency continuum in the reading test employed, with a majority situated at the lower end (50% of the test takers scored at the lowest level of the vocabulary test and 40% in the Novice band of the American Council on the

⁶⁴Laufer defined successful reading as reaching at least a certain L2 reading comprehension score in the placement test under investigation (1992, p. 99).

⁶⁵The children’s lexical competence was measured by a standardised receptive vocabulary test and a morphological test which required them to produce plural and past tense forms of nouns and verbs (Verhoeven, Voeten, and Vermeer 2019, p. 31).

⁶⁶Kang investigated the skills and abilities of young Korean learners of English. Given that Korean uses a different alphabet than English, it is possible that the structure of skills with respect to the children’s decoding skills is different than what would be found among English language learners whose L1 uses the Latin alphabet.

⁶⁷Though the authors themselves do not draw this conclusion, Erbeli and Joshi’s results reinforce the importance of lexical knowledge: most of their instruments seem to tap into lexical knowledge to some extent. For instance, their English as a Foreign Language (EFL) decoding instruments, whose results were strongly related to reading comprehension, clearly relied on English word knowledge (Erbeli and Joshi 2022, p. 4).

Teaching of Foreign Languages (ACTFL) proficiency scale). The authors found exceptionally high correlation between the reading assessment and the vocabulary scores ($r = .915$), though this may have been at least partly due to the distribution of the participants, most of which were situated at the lower end of the proficiency scale. In addition, the adaptive nature of the reading assessment used (ACTFL's Reading Proficiency Test (RPT)) may have inflated the correlation.

Looking at the matter from a slightly different perspective, Weiss and Meurers (2019b) tried to predict the CEFR level of 1033 written learner texts⁶⁸ using a broad collection of linguistic complexity measures (see also section 4.1.3). They found that *lexical* complexity measures formed the only linguistically homogeneous model which had at least a chance at correctly identifying learner texts at a level of A1. In fact, lexical complexity was the best-performing group of all linguistic domains at all CEFR levels up to B2, outperformed only by a mixed measure which contained the 150 most predictive measures of their analytic tool (*ibid.*, pp. 428 sq.).

Li and Clariana (2019) discussed L1 and L2 reading comprehension from a neurocognitive perspective, citing research (Grant, Fang, and P. Li 2015) which shows that low-proficiency readers in particular have to concentrate more on single-word access while more proficient readers show signs of automatised semantic access which allows for more capacities in text meaning construction. This, the authors hypothesised, also applies to L2 reading comprehension (P. Li and Clariana 2019, p. 103).

Overall, research on reading comprehension consistently shows that low-proficiency readers with “normal” decoding skills struggle with word access and, consequently, semantic proposition building. This seems to be true for both first and second language reading comprehension, though the foreign language context has been researched much less thoroughly.

5.2 Lexical characteristics of reading texts

In an analysis of reading item difficulty, lacking other sources of information, the lexical knowledge of the test takers can only be inferred by measuring the lexical complexity⁶⁹ of the input texts and the test items. Traditionally, this is done along three major lines of enquiry: lexical diversity, lexical density and lexical sophistication (see also Bulté and Housen's (2012) lexical complexity construct⁷⁰ and section 4.1.3.1 of this thesis). All three constructs contain a qualitative judgment: more proficient language users are considered to know *more* lexical units and to have more *precise* word knowledge, which can be measured by the fact that they produce (and understand) texts which contain many different (i.e. lexical diversity), infrequent (i.e. lexical sophistication) words used with precision (i.e. lexical density). In the following, the three constructs and their most common methods of measurement are presented.

⁶⁸The texts originated in the learner corpus project *Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context (MERLIN)* (Abel et al. 2014; Wisniewski et al. 2013).

⁶⁹For a distinction between *difficulty* and *complexity*, see section 4.1.3.1 and Pallotti 2015.

⁷⁰Bulté and Housen (Bulté and Housen 2012, p. 28) add a fourth subconstruct to lexical complexity: the “*compositionality* of lexical elements, that is, the number of formal and semantic components of lexical items (e.g. phonemes, morphemes, denotations)”. However, they do not elaborate further on this. Since the term arises from a context that seems to be only vaguely related to the study of lexical complexity (cf. e.g. Fukushima 2015; Smolka, Preller, and Eulitz 2014), and it is unclear how it would be operationalised in a reading text, it is not discussed in depth here. At best, based on Smolka et al.'s understanding of compositionality (2014), it could be expressed as the number or proportion of derived and complex word forms (e.g. verbs). This, however, is experimental and requires more in-depth studies of linguistic complexity.

5.2.1 Lexical diversity

As with anything, there is a simple and a more elaborated definition of lexical diversity. The simple version understands lexical diversity as a characteristic of a text which contains many different words. This is most visible in the historic measure Type-Token Ratio (TTR), which simply divides the number of types (i.e. unique words) by the number of tokens (i.e. all words) in a text.

$$TTR = \frac{\text{number of types in a text}}{\text{number of tokens in a text}} \quad (5.1)$$

If a text's TTR is close to 1, then its lexical diversity is considered high, as most of its words are not or rarely repeated. Inversely, a low TTR indicates that many words in a text are repeated. This measure has serious disadvantages, not least its correlation with text length, but it is the foundation of many other measures of lexical diversity that have been developed in the past decades (e.g. *voc-d* and the Measure of Textual Lexical Diversity (MTLD), see for a recent overview of various measures Nasser and Thompson 2021, pp. 3 sqq.).

As Jarvis lays out in some detail, however, there is more to lexical diversity than the (non-)repetition of words (2013b, pp. 18 sqq.). He describes lexical diversity as a matter of human perception composed of various continua along which lexical diversity is judged: variability, volume, evenness, rarity, dispersion, and disparity (ibid., pp. 22 sqq.)⁷¹. The first aspect, *variability*, is closest to the traditional understanding of diversity, meaning the “objective phenomenon” of a more diverse choice of words in a text, the opposite of *lexical repetition* (ibid., pp. 20 sq.). *Volume* essentially describes text length, as the amount of lexical items available also influences the perception of lexical diversity: longer texts can be more diverse than shorter texts, since there is more opportunity to choose from the lexicon. As a result, they are perceived to be more diverse even if they are not according to their TTR (Jarvis 2013a, p. 100). *Evenness* describes whether few words are repeated very often, or whether many words are repeated sometimes, i.e. whether the distribution of more frequent words in a text is more or less even, which would be perceived as more diverse (Jarvis 2013b, pp. 23 sq.). The last two aspects, dispersion and disparity are adopted from the understanding of diversity in the field of ecology: while *dispersion* describes whether tokens appear in clusters of the same type or are dispersed across a text (see also Gries 2008, 2010), *disparity* refers to the differentiation of lexical types in a text, i.e. whether the words in a text are semantically or formally more or less related (Jarvis 2013b, pp. 24 sq.). Finally, Jarvis also proposes to include *rarity*, by which he means lexical sophistication, in his definition of a construct of lexical diversity, even though it has traditionally been discussed separately:

[...] if it turns out that rarity works together with variability, volume, and evenness in determining how lexical diversity is perceived, then this may show that lexical richness is not just a grab bag of loosely related lexical measures after all. Instead, a good number of these measures may be directly linked to properties of a unitary construct. (ibid., p. 24)

To test whether these dimensions are relevant descriptors of a theoretical construct of lexical diversity, Jarvis conducted several empirical experiments (ibid., pp. 31 sqq., see also Jarvis 2017) and proposed initial and tentative operationalisations for all six subconstructs (cf. table 5.1). His research shows that the various measures seem to be related to human perception of lexical diversity, though there may yet be better ways to operationalise the various subconstructs. This problem was later approached by Vanhove and colleagues (2019), who investigated the relationship between human ratings of lexical richness and a large number of quantitative measures of lexical characteristics and compared this approach to a variation of Jarvis's six-variable model of lexical diversity. Using empirically derived measures for the six subconstructs (ibid.,

⁷¹In his three publications on the matter (Jarvis 2013a,b, 2017), Jarvis used somewhat differing terminology to refer to the subconstructs. He sometimes even mentions a seventh subconstruct—*abundance*, the number of types in a text. To avoid confusion, this section is solely based on Jarvis 2013b.

Subconstruct	Suggested operationalisation in Jarvis 2013b	Empirical operationalisation in Vanhove, Bonvin, et al. 2019
Variability	MTLD	MTLD with TTR threshold 0.83 (log-transformed)
Volume	number of words	number of tokens (log-transformed)
Evenness	standard deviation of the number of tokens per type in a text	standard deviation of tokens per lemma (square-root transformed)
Rarity	mean rank for all words in the text, as found in an adequate reference corpus	mean Zipf of unique lemmata
Dispersion	aggregated mean distance between different tokens of the same type	dispersion index for lemmata, $k = 20$ (square-root transformed)
(Semantic) Disparity	mean number of words which share the same semantic sense as indicated by WordNet's semantic sense index (alternative: Coh-Metrix's LSA)	
(Formal) Disparity		mean Levenshtein distance between unique lemmata

Table 5.1: Operationalisation of subconstructs of lexical diversity according to Jarvis (2013b, pp. 28 sqq.) and Vanhove et al. (2019, p. 515).

pp. 508 sq., 515), the authors constructed predictive models of lexical diversity ratings which were only slightly worse than models that included many more predictors⁷². Similar to Jarvis before them, Vanhove and colleagues identified *volume* as the most important predictor, followed by *evenness* and *rarity* (ibid., pp. 515 sq.). Interestingly, this study was conducted on texts in French, German and Portuguese (written by L1 speakers of those languages, including heritage speakers). Overall, it seems that the findings were consistent across the three languages. Kyle and colleagues (2021) also investigated how three of Jarvis's subconstructs related to human judgments of lexical diversity. They found a stronger predictive effect for *abundance*, the number of types in a text, instead of volume, the number of tokens in a text, but otherwise confirmed the findings of Jarvis and Vanhove and colleagues.

Overall, empirical research on lexical diversity suggests that, although the construct is complex, and there is no one way of operationalising it, some well-established measures are reasonably good predictors of human ratings of lexical diversity. In addition, dividing the construct into various lines of evidence has shown that text length (in Jarvis' terms *volume*) is strongly related to human perceptions of lexical diversity, even if other aspects are controlled for. This is an important observation, as it underlines the fact that even though text length seems to be a superficial measure, and is often considered a nuisance, it still contains valuable, if diffuse, information on the quality of a text. This may be particularly true in L2 contexts, where the ability to read and write longer texts implies the knowledge of more words.

5.2.2 Lexical density

Lexical density usually describes the amount of content words in a text: if there is a higher rate of lexical words, particularly nouns, a text is considered to be more complex (Pallotti 2015, p. 126; Nasserri and Thompson 2021, p. 1). Lexical density is either measured as the proportion of content words (or lexical items) as opposed to function words (or grammatical items) in a

⁷²In fact, the authors also investigated how well a model fared which only included one predictor, Guiraud's root TTR, an index of lexical diversity based solely on the number of types and tokens in a text. Even this single predictor, which is strongly correlated to text length, performed reasonably well in comparison to the more complete models (Vanhove, Bonvin, et al. 2019, pp. 516 sq.).

text (To, S. Fan, and Thomas 2013, p. 62), or as a ratio of the number of content words and clauses in a text (Halliday 1985, cited in To, S. Fan, and Thomas 2013, pp. 62 sq.). Because a higher lexical density theoretically accounts for precision in expressing ideas and content, it is often investigated in the context of written or academic language (Nasseri and Thompson 2021, p. 3).

Given the multivariate natures of reading comprehension and linguistic complexity, research on a direct relationship between lexical density and language proficiency has remained largely inconclusive. Interested in reading comprehension, To and colleagues (2013, p. 63) as well as Ağçam and Babanoğlu (2018, pp. 951 sq.) mention several studies which looked for a connection between lexical density and the proficiency levels of reading texts in English textbooks (both L1 and L2), but rarely found it. Indeed, these authors could not find a relationship between lexical density and the proficiency levels of the textbooks they examined in their own studies (To, S. Fan, and Thomas 2013, pp. 65 sqq.; Ağçam and Babanoğlu 2018, p. 956). Nevertheless, Ağçam and Babanoğlu point out that “there is a general consensus from the pedagogical perspective that high lexical density is associated with more propositional content and greater complexity, which can render language more difficult to process particularly for non-native speakers” (2018, p. 949).

More generally speaking, there is evidence that written language is usually more dense than spoken language, and that academic language, too, has a tendency to be denser (To, S. Fan, and Thomas 2013, p. 63). Comparisons between the lexical density of texts produced by native speakers and foreign or second language learners also show that lexical density is higher in texts written by more proficient language users (Nasseri and Thompson 2021, p. 3). Similar findings were reported for the (Swedish) school context (Linnarud and Thoursie 2008), including striking differences between texts written by students in English and German, with the latter being of much lower quality, including lesser lexical density scores. Some researchers, though, have pointed out that while lexical density may rise along with proficiency scores, the variance within proficiency levels is equally high (e.g. Read and Nation 2006). Overall, then, there is reason to believe that within a narrow band of proficiency, as is the case in the ÜGK assessment, the lexical density of the input texts may not vary along with the intended proficiency level, but with other characteristics of the texts (e.g. their text genre or topic). Nevertheless, since lexical density is usually unrelated to text length, and, as some of the aforementioned studies have found, not particularly related to readability scores either, it is still useful to add this source of information to the empirical analysis of reading item difficulty in this thesis.

5.2.3 Lexical sophistication

The concept of lexical sophistication is not easily defined. It stems from an understanding of language use which presupposes that the use of more complicated or difficult words and expressions must be an indication of higher language proficiency (Laufer and Nation 1995, pp. 309 sq.). This is true both in L1 and L2 language use, particularly in (language) learning contexts, where a higher lexical sophistication can be interpreted as evidence of learning (*ibid.*, p. 308). Apart from this very general view, lexical sophistication is often defined by way of its operationalisation (e.g. Kyle and Crossley 2015, pp. 759 sqq.): it is measured as the lexical frequency and range of tokens, types, word families and multiword tokens occurring in natural language, or in terms of the psycholinguistic properties of words (e.g. concreteness, familiarity, usually measured through human judgment). A somewhat related view is the concept of vocabulary difficulty: in instructed language learning contexts in particular, language users may not or very rarely be in contact with naturally occurring language, and they may learn the foreign language with very specific goals in mind. In these cases, the lexical frequency of words and expressions as determined by an L1 corpus may not be particularly meaningful, especially with respect to content words (X. Chen and Meurers 2018, p. 490; D. J. Davidson, Indefrey, and Gullberg 2008, pp. 133 sq.). Hence, either a suitable reference corpus is needed, or the words and expressions of the learners’ language have

to be categorised in different ways to account for lexical sophistication. These two approaches will be briefly discussed in the following.

5.2.3.1 Lexical frequency

A common way of characterising the lexical complexity (or difficulty) of a text is to measure the frequency with which the words it contains appear in natural language. The rationale behind this is based on psycholinguistic research which shows that words which are encountered more frequently are generally processed faster (for a critical review of this assumption see Gries 2010). Since reading comprehension relies on fast, automatic processes of word recognition, this faster processing potentially leads to better reading comprehension (for references cf. X. Chen and Meurers 2018, p. 489; Kyle and Crossley 2015, pp. 759 sq.). In addition, more frequent words are usually “linked to richer bodies of world knowledge”, i.e. readers know more about them and will have less difficulty understanding their meaning in a reading text (Graesser, McNamara, and Kulikowich 2011, p. 225, see also McNamara, Graesser, et al. 2014, p. 43). Hence, texts with lower lexical sophistication are expected to be easier to read, which means that if readers were to understand a text with *higher* lexical sophistication, their language proficiency would be considered to be higher.

Historically, lexical frequency was evaluated based on handmade lists which contained a large number of words ordered by their frequency in a corpus (e.g. Kučera and Francis’s (1967) frequency norms for English, see also Brysbaert, Buchmeier, et al. 2011, pp. 415 sq.; François, Gala, et al. 2014, pp. 3766 sq.). Today, it is possible to use frequency indices taken directly from more or less specialised corpora, which also allows to take into account the needs of different groups of language users (Pintard and François 2020, pp. 86, 90). Analyses of lexical sophistication can be done with specialised software (see e.g. Kyle and Crossley 2015) or even on websites (e.g. Cobb 2023) requiring little specialised knowledge or technical know-how. Most of these tools, however, are only available for English and/or use a specific reference corpus that cannot be changed. More flexibility is gained if more generalised corpus analytic software can be used, e.g. R (R Core Team 2024) along with a specialised package such as koRpus (Michalke 2021).

Beyond the choice of an adequate corpus, it is unclear how exactly the lexical frequency of words in a text can be transformed into a value that is useful for empirical research. One approach to expressing the lexical sophistication of texts, particularly in L2 contexts, is the use of frequency bands (Laufer and Nation 1995): they are derived from a rank-based ordering of words (or multiword expressions) in a corpus, typically in chunks of 1000 units, depending on the overall language proficiency of interest. Added to these is usually an academic word list of some kind to account for words and expressions which may be extremely rare in natural language use, but are relevant in academic language and, hence, in many language assessment contexts. In writing or reading assessment, the mastery of the various frequency bands is often expressed in percentages, e.g. a random English input text of the ÜGK foreign language assessment⁷³ contains 77% of words in the K1-band (the most frequent 1000 words), 8% in the K2-band, and 11% on an academic word list, with the remaining 4% being proper nouns or unknown. Given the high percentage of frequent words, this text can probably be considered easy to comprehend⁷⁴. These percentages do not lend themselves easily to statistical analyses, unless treated as several variables, one for each frequency band (see e.g. Brunfaut and Révész 2015), or by simplifying the matter, e.g. using an index of the proportion of words which are considered to be low-frequency (e.g. Carr 2006, p. 274).

Another approach is the transformation of the frequency indices of the words in a text by some means of aggregation (for a more detailed overview cf. X. Chen and Meurers 2018, pp. 489 sqq.).

⁷³LV03G_A in English analysed on [lexutor.ca](https://www.lexutor.ca) (Cobb 2023) on 2023-04-04.

⁷⁴Note that some experimentation with parts of this chapter and the abstract of Kyle and Crossley (2015) resulted in very similar coverages of the frequency bands, though with some more words being off-list. Some caution is due then, in interpreting these results.

Given a word list or a reference corpus, one could calculate the average word frequency or the mean rank of all words or all content words in a given text (e.g. Embretson and Wetzel 1987). This approach has been used to create readability formulas since the 1920s (X. Chen and Meurers 2018, p. 489) and it has shown to create indices of word frequency which are correlated with reading comprehension and vocabulary difficulty. As Gries (2010, p. 198) points out, however, the frequency of occurrence is only one aspect of word difficulty among others, and can only ever be a rough indicator. In addition, a mean or median of word frequencies or rank orders is easily influenced by extreme values and outliers, leading to even more room for error (X. Chen and Meurers 2018, p. 491). As a result, as with many other linguistic characteristics, researchers have proposed and investigated a variety of approaches to measurement (e.g. *ibid.*). For instance, the analytical tool Coh-Metrix operationalises word frequency as three variables based on word frequencies provided by the CELEX Lexical Database (Baayen, Piepenbrock, and Gulikers 1995): the raw word frequency of all content words in a text (WRDFRQc), the mean log frequency of all words in a text (WRDFRQa), and the mean log minimum frequency of words in sentences (WRDFRQmc) (cf. McNamara, Graesser, et al. 2014, pp. 43, 73; Graesser, McNamara, and Kulikowich 2011, p. 225; *Coh-Metrix version 3.0 indices* 2023).

Finally, instead of, or in addition to the frequency of a word or expression in a reference corpus, some researchers have suggested to analyse its *range*⁷⁵: rather than counting the raw frequency of occurrence of a word or expression across a corpus, its dispersion is also taken into account (Gries 2008, 2010; Kyle and Crossley 2015, pp. 760 sq.; X. Chen and Meurers 2018, p. 491). This is to avoid that certain words or expressions, which occur often in some texts, but not in others, are considered to be frequent when, in fact, their use is limited to specific contexts. The simplest way to do this is to count the number of parts (e.g. documents) in the corpus in which a word or expression appears (Gries 2008, p. 407). There are also ways of adjusting the frequency measure of words or expressions by some indicator of their range (e.g. the distances between successive occurrences, cf. for more details *ibid.*, pp. 409 sq.). The majority of these measures, however, have rarely been researched in terms of their meaning and validity (Gries 2010), and, unlike frequency measures, they are not as readily available in the public domain as frequency information.

5.2.3.2 Lexical difficulty

In the context of second and foreign language learning, researchers and practitioners have been looking for a way to categorise words and multi-word expressions in terms of their difficulty for learning. Beyond the search for reference corpora which are better adapted (e.g. D. J. Davidson, Indefrey, and Gullberg 2008), one way of doing that is to assign a proficiency level to each word (e.g. from the CEFR). Though a careful reading of the CEFR clearly does not purport such a categorisation⁷⁶, the approach is common enough that it merits consideration. Apart from pure expert judgments (e.g. Reference Level Descriptions (RLDs) such as Profile Deutsch (Glaboniat et al. 2005), which are not available in electronic format, cf. Pintard and François 2020, p. 87), there are at least two (open and digital) sources for word lists which are aligned with the CEFR levels: open-source vocabulary learning programmes (e.g. Decaleon) and research surrounding the CEFRlex project. The former, according to its user manual, is originally an automatic translator for Esperanto and has since grown to include many languages (Penzkofer 2023). The publicly

⁷⁵Chen and Meurers refer to the same concept as Contextual Diversity (CD) (2018, p. 491).

⁷⁶The CEFR states that “[no] two users of language [...] have exactly the same competences or develop them in the same way” (Council of Europe 2001, p. 17) and that, as a result, its levels of proficiency are “to some extent arbitrary” (*ibid.*, p. 17). It also stresses that the number of domains in which language users may need to communicate is “indeterminate” (*ibid.*, p. 45) and that the choice of the domains in language learning depends on the objectives, interests and needs of the learners. Hence, since the domain and the resulting communicative situations effectively determine the necessary language knowledge, any categorisation of individual words as belonging to a certain CEFR level would first require a detailed description of the learning context and the prerequisites of the learners.

available database contains words in 19 languages (as of April 2023), classified into the six levels of the CEFR and into 25 topics, though no information is available as to how this classification was done. The software also includes technical vocabulary, information on synonyms, antonyms, homonyms and context, POS, multiword expressions, idioms and cooccurrences, though, again, it is unclear on what basis all this information was collected. The CEFRlex project, on the other hand, is a scientific research project, which aims at building foreign language lexicons aligned with the six CEFR levels (CENTAL 2021). It currently covers the languages French (François, Gala, et al. 2014), English (Dürlich and François 2018) and German (Brysbaert, Buchmeier, et al. 2011), as well as Swedish, Spanish and Dutch. As François and colleagues (Dürlich and François 2018; François, Gala, et al. 2014) explain, the resource is based on a corpus of textbooks, online materials and graded readers in the respective languages which are meant to be consumed by L2 language users, and which were aligned to the CEFR by the publishers. The corpus was POS-tagged and frequency information was extracted using a dispersion-corrected formula (François, Gala, et al. 2014, p. 3769). The result is a wordlist which indicates the lemma, its POS, and its adjusted frequency of occurrence for each CEFR level⁷⁷. In French and English, the underlying frequency list can be downloaded for research purposes. Unfortunately, for German, only a web-based search form is available at the time of writing. The French version of the web-based analyser also features a level attribution method which is based on a machine learning algorithm that integrates word level information from a French RLD (e.g. Beacco et al. 2008; Riba 2016), leading to somewhat better word level classification than a consideration of a word's first occurrence in the corpus (Pintard and François 2020).

NLP offers another approach to determining the difficulty of words. Computer linguists have been working on the development of automated lexical simplification using Complex Word Identification (CWI) (cf. Shardlow, Evans, and Zampieri 2022). Put very simply, based on human rating of complex words or multi-word expressions in sentences or longer texts, algorithms were developed which predict the complexity of words or multi-word expressions in “new” sentences or texts. The results of those algorithms were then compared to human ratings of the same sentences or texts. As Shardlow and colleagues summarise, such approaches were most successful when complexity was considered on a probability continuum, and when multi-word expressions were also included (ibid., p. 7). In a follow-up analysis of the link between traditional psycholinguistic measures of word complexity and human complexity ratings, the authors found aspects such as word familiarity, concreteness and imageability as well as a semantic vector representation to predict word complexity using principal component analysis (ibid., p. 31). Word length and frequency measures were also found to be important predictors: both variables were used in all algorithms described by Shardlow and colleagues (ibid.).

The authors also point out that the construct of *complexity* employed by the human annotators was not actually well-defined, and that the annotators themselves came from diverse backgrounds, including both native and non-native speakers of the target language. The findings from this type of research are interesting because they allow for theory-building based on a large corpus of words or multi-word expressions which were considered complex at least by some human raters. For instance, Shardlow and colleagues identified the following intuitive characteristics of (English) words to be indicators of complexity (ibid., p. 9):

- The word is archaic.
- The word is a borrowing from another language or refers to a concept that is atypical in the culture of the reader.
- The word is uncommon and many people are not generally exposed to it.
- The word refers to a very specialised concept.
- Although the word is common, it is being used with an uncommon meaning in the given context.

⁷⁷For English, the resource only spans the levels A1 to C1.

Interestingly, while these characteristics seem persuasive, when looking for empirical evidence, the authors did not find significant linear correlations between the human complexity ratings and a range of psycholinguistic variables traditionally related to lexical complexity (e.g. concreteness, imageability, word type, word status and others, see Shardlow, Evans, and Zampieri 2022, pp. 11 sq.), with two exceptions: word length and a frequency count (ibid., pp. 10, 13). Hence, even though new approaches have been developed in the past decade, a frequency estimate based on an adequate corpus still seems to be a good proxy for word complexity.

5.3 Measuring cross-linguistic characteristics of reading texts

It has been shown repeatedly that humans tend to access their entire linguistic repertoires, whether or not they are currently in a monolingual or a multilingual setting (Jessner, Allgäuer-Hackl, and Hofer 2016, pp. 159 sq.; Ringbom and Jarvis 2009, p. 106)⁷⁸. Hence, in a study of a foreign language assessment, particularly one which uses items in the common language of schooling of the test takers, its lexical characteristics have to be examined with respect to all the languages present in the test takers' linguistic repertoire. In international foreign language assessments such as TOEFL, this is a difficult task given the diverse backgrounds of the test takers. In the context of the ÜGK assessment, however, a shared language of the test takers can easily be identified, as sufficient knowledge of the language of schooling was a prerequisite for participation (Verner and Helbling 2019, pp. 12 sq.).

In language assessment research or even in L2 reading research, the L1 of the readers or test takers is seldom considered as more than a property of the test takers (e.g. Alderman and Holland 1981), if at all. Hence, in the following, after a short definition of cross-linguistic transfer, the particular topic of cognates is discussed, as well as a way to operationalise the lexical distance of words and texts. The insights from these deliberations will be used in sections 9.3.1.4 and 9.3.3.4 to develop experimental variables which try to account for the role of the language of schooling in solving the foreign language reading tasks in the ÜGK assessment.

5.3.1 Cross-linguistic transfer

Language use can be shown to be influenced by the transfer of patterns and matter⁷⁹ from one language to another (see an overview of research in Odlin 2005 or a more thorough account by Matras 2011). This is true in both reception and production (Berthele, Lenz, and Peyer 2022), and can be helpful, a hindrance, or, of course, neither (Davis, Bowman, and Kaushanskaya 2018; Schluer 2017).

Ringbom and Jarvis (2009, p. 109) define three distinct types “in the continuum of cross-linguistic similarity relationships”: “*similarity*, *contrast*, and *zero relations*”. Whereas similarity occurs when “an item or pattern in the TL [target language] is perceived as formally and/or functionally similar to a form in the L1 (or another previously learned language)” (ibid., p. 109), contrast happens when “the learner perceives a TL pattern as in important ways differing from an L1 form or pattern, though there is also an underlying similarity between them” (ibid., p. 109). The authors suggest that, despite the differences, this underlying similarity allows for more effortless learning than a “zero relation, [where] an average learner cannot easily notice

⁷⁸In fact, the activation and processing of language may work along similar lines whether or not an individual has access to more than one language. For instance, Mulder et al. (2015, p. 2) mention a number of studies with monolinguals which showed that words from larger morphological families are activated quicker. Obviously similar to the faster activation rates of cognates (see below, 5.3.2), these findings were the starting point for their study on bilingual word processing.

⁷⁹Very generally speaking, matter can be understood in terms of the transfer of words, while pattern transfer refers to linguistic structures, e.g. grammatical constructions. Ringbom and Jarvis (2009, p. 110) speak of *item transfer*, *system* or *procedural transfer* and *overall transfer*, though this terminology is not used here to avoid confusion related to the word *item*.

features that a totally different TL has in common with the L1” (ibid., p. 110). Since learning and reading are intrinsically connected, it is reasonable to assume from these findings that L2 reading assessment results will be influenced by the transfer of both patterns and matter from other languages. In the context of the ÜGK assessment, this is reinforced by the fact that the reading items have been formulated in the language of schooling (cf. section 9.2.2.4).

5.3.2 Cognates

Cognates are an object of research in at least three linguistic fields: in *historical linguistics*, or etymology, cognates are defined as “word pairs, across languages, which have a common etymological origin” (Kanojia et al. 2021, p. 1). In this field, cognates are historical evidence: they inform on the relatedness of language families and their evolution, and they allow a tentative reconstruction of ancestor languages, of which no written, let alone spoken, documents remain (Grzegorz Kondrak 2001, p. 1). This strict diachronic definition has been broadened with the advent of Automatic Cognate Detection (ACD) in : “with the understanding that the primary distinction is whether words are etymologically related, [...] the decision whether they are related by inheritance or borrowing will be made at a later refinement stage” (Dellert 2018, p. 3123).

ACD is widely used in fields such as machine translation, lexicography and related applications (e.g. Kanojia et al. 2021; Grzegorz Kondrak 2001; Melamed 1999; W. Wu and Yarowsky 2018). In these fields, the term cognate usually loses its diachronic perspective entirely, “denoting words in different languages that are similar in form and meaning” (Grzegorz Kondrak 2001, p. 1). After all, identifying translational equivalence in words does not depend on them being historically related. The same “synchronic viewpoint” (Daulton 2008, p. 46) is adopted in *psycholinguistics* and *Second Language Acquisition (SLA)* research, which deals with cognates to tap into language processing and acquisition (e.g. Lemhöfer, Dijkstra, and Michel 2004; Mulder, Dijkstra, and Baayen 2015). Using evidence such as reaction times or learning effects, researchers differentiate between identical and non-identical cognates, false cognates and control words to experimentally investigate the structure and development of the mental lexicon.

The main finding from these experiments is the “cognate facilitation effect”: there is strong evidence that cognates are processed faster and learned better (e.g. De Groot and Keijzer 2000; Lemhöfer, Dijkstra, and Michel 2004, p. 587, though see Davis, Bowman, and Kaushanskaya 2018 for their inhibitory effect on reading fluency). This effect is more pronounced when the cognates are more similar (Mulder, Dijkstra, and Baayen 2015; Vanhove and Berthele 2017a) and when more than two languages are involved (Lemhöfer, Dijkstra, and Michel 2004). In foreign language learning, cognates have been shown to facilitate both incidental vocabulary learning (e.g. through reading or watching TV, cf. Vidal 2011; Peters and Webb 2018; De Wilde, Brysbaert, and Eyckmans 2020) and lexical instruction in the classroom (Otwinowska et al. 2020, p. 688). In combination with other facilitation effects stemming from linguistic similarities between (related) languages (e.g. the suggestions made by the EuroCom programmes: Stegmann 2005), it has been suggested that cognates should be taught explicitly in the classroom (though see Otwinowska et al. 2020). In Switzerland, cognates are currently part of the foreign language teaching curriculum, along with other aspects related to the transfer of knowledge and competences from one language (including the L1) to others (e.g. D-EDK 2016, p. 8; DECS 2004, p. 35). This is particularly evident in the foreign language textbooks *Mille feuilles/Clin d’œil* and *New World*, which were developed for the *Passepartout* cantons (where French is taught as the first foreign language to German-speaking pupils) and contain various texts and tasks that emphasise the use of cognates for comprehension (cf. e.g. Kofler, Peyer, and Barras 2020).

5.3.3 Operationalising transfer in a reading assessment

The linguistic similarity (or distance) between two words is often operationalised using the Levenshtein distance (Schepens, Dijkstra, and Grootjen 2012, p. 159; Vanhove 2014, pp. 106 sqq.)⁸⁰. This metric counts the number of insertions, deletions and substitutions necessary to arrive from one string to another (see table 5.2 for an example of the orthographic and phonetic Levenshtein distances of the German word *Geister* [gaɪstɐ] and the English word *ghosts* [gousts]⁸¹). Traditionally, each edit operation is assigned the weight 1, as in the example. If more than one way of transforming one word into another can be thought of, the minimum number of insertions, substitutions and deletions is chosen. Across multiple words, the Levenshtein distance is usually

1	2	3	4	5	6	7	8
G		e	i	s	t	e	r
g	h	o		s	t	s	
	I	S	D			S	D = 5
g	a	ɪ	s	t	ɐ		
g	o	ʊ	s	t	s		
	S	S			S		= 3

Table 5.2: Example of an orthographic and a phonetic Levenshtein distance. The German word *Geister* and its English counterpart *ghosts* are transformed into each other by an insertion (I), two substitutions (S) and two deletions (D), when their orthographic form is considered. Their phonological forms are separated by three substitutions.

corrected for word length (eq. (5.2), cf. Schepens, Dijkstra, and Grootjen 2012, p. 159), which leads to a score between 0 (every letter underwent some change, i.e. the words do not overlap at all) and 1 (no changes, i.e. the words are identical).

$$\text{Levenshtein score} = 1 - \frac{\text{minimum distance}}{\text{length of longest string}} \quad (5.2)$$

Hence, the words *Geister* and *ghosts* feature a Levenshtein score of $1 - \frac{5}{8} = .375$ if their orthographic distance is considered (and the German capital letter ignored), and $1 - \frac{3}{6} = .5$, if their phonological distance is of interest.

The Levenshtein distance has been used in research to measure the linguistic distance of words used in experimental research (e.g. Beijering, Gooskens, and Heeringa 2008; Vanhove and Berthele 2017b), or as a measure of the distance between languages in general (e.g. Berthele, Lenz, and Peyer 2022; Shatz, Alexopoulou, and Murakami 2024). Sometimes (e.g. Ptáčková and Vitásková 2020), the Damerau-Levenshtein distance is used instead, which is a variant of the Levenshtein distance that includes transpositions of adjacent letters (Damerau 1964, see also Yarkoni, Balota, and Yap 2008). A word like the German *Meter* and the French *mètre* would be considered to have a Damerau-Levenshtein distance of 1 instead of a Levenshtein distance

⁸⁰Originally, the Levenshtein distance is a metric established in computer science to measure the difference between two sequences of code (Levenshtein 1966). It is also used in other fields, e.g. in biology (Berger, Waterman, and Yu 2021).

⁸¹Note that instead of counting the *-h-* as an insertion, it could also be counted as a substitution for the *-e-*, lowering the distance to 4, which would probably be more correct. The solution in the example has been chosen for illustrative reasons. The pronunciations of the two words are Standard German and US American respectively, as indicated on [wiktionary.org](https://www.wiktionary.org) on 2023-04-25.

of 2 to indicate that the letters *-e-* and *-r-* were merely reversed instead of substituted⁸². This accounts for the fact that a word which only contains two reversed letters is clearly more similar to another than a word in which two letters were changed for others (e.g. the French word *métal*, which has both a Levenshtein *and* a Damerau-Levenshtein distance of 2 (or 4) to the German *Meter*).

There is not a lot of research which focuses on the effect of linguistic distance on reading comprehension, particularly in foreign language learning (Berthele, Lenz, and Peyer 2022, p. 422). When linguistic distance is considered in reading research, it is often operationalised as the linguistic distance between the languages at play, rather than the actual linguistic distance of the texts under investigation (e.g. *ibid.* or Shatz, Alexopoulou, and Murakami 2024 for a study concerned with foreign language writing). An exception is a study of the intelligibility of reading texts in Frisian and Afrikaans for adult speakers of Dutch (Bezooijen and Gooskens 2007). The authors rated all words⁸³ in their reading texts for their cognateness and their transparency for Dutch readers, and calculated the Levenshtein distance for all cognates⁸⁴. They found that their participants understood the Afrikaans texts significantly better, and that these were also closer to Dutch linguistically: there were more cognates, particularly content words, and the spelling diverged less from Dutch than the Frisian spelling, which was indicated both by human transparency ratings and the Levenshtein distance. This indicates that the Levenshtein distance may be useful to model the effect that the linguistic proximity of two specific texts has on reading comprehension.

5.4 Conclusion

In reading comprehension and language assessment research, the lexical characteristics of the reading text have usually been measured in terms of lexical diversity, lexical density and/or lexical sophistication, though in all three cases, the concrete operationalisation is often up for debate. The actual effect of these three constructs on reading comprehension (and, as a result, item difficulty) is similarly unclear, though the consensus is that more lexically diverse, dense and sophisticated texts are harder to understand. Once another language is introduced to this complex situation, simple cause-and-effect explanations cease to be available, as even the lexical distance of texts in very related languages is hard to measure and its influence on reading comprehension difficult to extract from other mechanisms. This is where test results like the one from the ÜGK foreign language assessment may allow for a unique insight: since the tasks were largely identical, and the assessment was administered under comparable conditions, it is mainly the languages themselves which differ between the four versions. If this difference could be measured adequately, it would be possible to discern the extent to which the language of schooling influences reading comprehension in a (low proficiency) foreign language.

⁸²This supposes that capitalisation and accents are not taken into account when calculating the distances. If they were, the Damerau-Levenshtein distance would be 3, and the Levenshtein distance 4.

⁸³Due to the relative linguistic proximity of these three languages, the authors were able to align most of the words in the texts with their counterparts in the other languages, which would be much more challenging with more distant languages such as German, French and English.

⁸⁴Well over 90% of all words in this study were rated as cognates.

Chapter 6

(Item) Difficulty: definition and measurement

6.1 A critical look at item difficulty

In his discussion and ultimate rejection of the notion of difficulty, Bachman (2002) leads with the following words:

The notion that test items or tasks themselves differ in difficulty is deeply ingrained in both the way we conceptualize the difficulty of assessment tasks and in how we operationalize difficulty in most current measurement models.

In fact, any test taker, when asked, will be able to say with some certainty which assessment task they find difficult and which not. The same is true for teachers: as part of their job description, they will regularly predict what task will be doable for their students and what they will find too difficult (although whether or not these predictions are accurate in either group is another question, see e.g. Impara and Plake 1998; Watering and Rijt 2006). In psychometric measurement (in particular CTT and IRT), *item difficulty* is also quite clearly defined, although it is rarely meaningful without additional measures and/or a consideration of the test takers.

Nevertheless, and with good reason, Bachman goes on to explain that the “conceptualization of difficulty as a characteristic that resides in test tasks” (2002, p. 462) is in many ways misleading because it is dependent on test takers’ characteristics: any ability, action or skill that has to be learned is easier for those who have already mastered it whereas it may be impossibly difficult for those who have not (yet) learned it. As a result, *item difficulty* cannot exist as a static property of a task; instead, it is essentially an artefact of the interaction between the characteristics of the test takers and those of the test task or item. This is also true in psychometric conceptualisations of the term, which leads to a dilemma: if *item difficulty* is the result of an interaction which *includes* task characteristics, any relationship between these task characteristics and item difficulty will be at least partly due to autocorrelation. Therefore, it will remain unclear whether a task characteristic actually *influences* difficulty or whether it is simply part of the interaction which created the item difficulty value in the first place (ibid., p. 468).

Bachman ultimately states that more “research into task effects” is crucial and that there will be various approaches to doing that. Having rejected *difficulty* as a task characteristic in itself, he suggests that the variables which are considered to be influencing task performance be distinguished into a) “characteristics inherent in the task itself” (i.e. which can be determined by examining only the task itself without taking into consideration the test takers), b) “attributes of test-takers” and c) their interactions (ibid., p. 469). He also points out that research into task performance should consider the processes and strategies which test takers use to solve assessment tasks, using appropriate qualitative or quantitative methodology (ibid., p. 470).

While Bachman’s argumentation is persuasive and his main point—that difficulty is not actually an inherent quality of a task, but instead the product of an interaction between the task and the person solving it—should be kept in mind, item difficulty has not lost its central role in the type of research he advocated. There are practical reasons for this: Bachman does not actually give a feasible alternative to the use of item difficulty as a dependent variable in item difficulty research (nor, in fact, to its use in actual psychometric analyses). One could perhaps suggest the raw test answers of an assessment, though it will become clear in the following that item difficulty is, in many ways, merely a mathematical expression of exactly that. A theoretical argument in favour of item difficulty is validation: a test takers’ performance is usually expressed in terms of a score, either a mere sum of points, or a more sophisticated psychometric measure. In both cases, the interaction between the task and the test taker that Bachman argues against has already taken place, it is expressed in the final score and all consequences of the assessment are based on it. Hence, as long as educational measurement uses the concept of item difficulty, it seems prudent to analyse its nature using psychometric analyses (Zumbo 1999, p. 12). Furthermore, even if it is imperfect, item difficulty is linked to other statistical measures of item quality, which makes a deeper study of the characteristics of items possible.

6.2 Item analyses in psychometrics

As was mentioned above, there is a psychometric understanding of item difficulty which has been refined over the years and has not lost its importance in educational measurement or assessment research. Its conceptualisation in CTT and IRT will be presented in the following along with item discrimination, and a short excursus will consider other ways of estimating item difficulty (section 6.2.1.3). In a second section, the notion of item fit and DIF, two ways of examining item quality, will be explained in more depth.

For this discussion, it is important to keep in mind that statistical theories of measurement are necessarily simplified representations of reality. At their most basic, both CTT and IRT consider tests to be essentially one-dimensional: all items are supposed to measure the same trait or ability and only that. There are, of course, ways to investigate tests and items which do not fit this assumption, and multidimensional IRT models have been developed to describe tests which measure more than one construct, but for the sake of clarity, it will be assumed in the following that all items do indeed measure the intended ability of the test takers and nothing else. Section 6.3 will then describe a form of item response modelling which takes characteristics of the items and/or the test takers into account before estimating item parameters.

6.2.1 Item difficulty and item discrimination

In psychological and educational measurement, item difficulty and item discrimination are empirical measures derived in some way from the performance of a group of test takers or respondents. Very generally speaking, an item’s *difficulty* measure should predict in some way if a test taker with a certain ability⁸⁵ will be able to solve this item. Item *discrimination*, on the other hand, should give an indication as to how well one item can distinguish between respondents at different ability levels: if an item has a good item discrimination at a cut-off point between two ability levels, two respondents can be quite close to either side of this cut-off point and still not obtain the same result.

⁸⁵The concept of a test taker’s *ability estimate* is an entire topic of its own in psychometric research, perhaps even more well-researched than item difficulty. See e.g. Lechner et al. 2021 for an accessible overview and critical discussion.

6.2.1.1 Classical Test Theory

Originally, CTT was not concerned with the characteristics of individual test items, but with the meaning and interpretation of test scores (i.e. the number of correct answers given by individuals on the entire test; cf. Hambleton and R. W. Jones 1993 for an overview, Novick 1966 for the mathematical foundations, Crocker and Algina 2008 and Moosbrugger and Kelava 2008 for more thorough textbook accounts). Nevertheless, conscious of the fact that test scores are composed of item scores, and that certain conditions have to be met at item level to actually draw conclusions from test scores, psychometricians developed methods of item analyses early on⁸⁶. In the CTT framework, two approaches to item analysis are now usually mentioned: *item difficulty* (usually denoted p) and *item discrimination* (an index denoted D or several correlation coefficients r , cf. Crocker & Algina 2008, pp. 314 sqq.).

Item difficulty At its most basic, item difficulty is defined as the proportion (P or p)⁸⁷ of test takers who answered an item i correctly (ibid., p. 90):

$$p_i = \frac{\text{number of persons with a score of 1 on item } i}{\text{number of all examinees}} \quad (6.1)$$

The logic of this definition is rather straightforward: an item is considered difficult when only a few respondents solved it correctly (p is small), whereas an item that many test takers solved correctly (p is high) is considered to be easy. Several mutations of this approach exist, which take into account aspects like items with a maximum score of more than 1, guessing in multiple-choice items, missing data or different probabilities of choosing certain answer options (e.g. Lienert and Raatz 1998, pp. 74 sqq.; Crocker and Algina 2008, pp. 312 sq.; Moosbrugger and Kelava 2008, p. 75). The general concept of item difficulty, however, remains the same.

An important drawback of this definition is its dependence on the sample: if the same item (e.g. a reading comprehension item in English) is solved by two groups of very different test takers (e.g. students of English after several *weeks* vs. several *years* of study) and the resulting item difficulty is calculated for each group, the p -values for the same item will differ greatly. In the original definition of P , this fact is clear: Lord (1953, p. 523) defines $P_{i,c}$ and $P_{j,c}$ as “the proportion of correct answers given by examinees *at a fixed ability level* (c) to items i and j respectively” (emphasis added)⁸⁸.

Since this mathematical requirement is difficult to observe in real testing situations, Lord and other psychometricians later tried to transform p in some way to make it comparable across ability groups (cf. for somewhat accessible summaries Ali and M. E. Walker 2014; Guo, Rudner, and Talento-Miller 2009). A prominent example of such a transformation is *Delta* (Δ), a value of item difficulty used by the Educational Testing Service (ETS) in the past (Holland and Thayer 1985), which was shown to be approximately linearly related to person ability. Since p and Δ could be converted into one another, it was possible to administer new items along with known anchor items to a group of examinees distributed on the ability continuum. By then equating the observed item difficulty of the new items in a reference group to the known difficulty of the anchor items, item difficulties at various levels of ability could be calculated (Ali and M. E.

⁸⁶It is not easy to determine who developed which measure at what point and for what purpose. In fact, it seems that a lot of psychometric concepts and measurements have existed for several decades, alongside each other, in a variety of forms and denominations. Since it is not the purpose of this text to portray the history of psychometrics, the following sections will not contain detailed information on the origins of the various concepts. If interested, the reader is invited to consult the cited references to retrace the history of psychometric theory and discussion. It would certainly be a fascinating endeavour.

⁸⁷Both notations can be found in the literature, with the capital letter having been originally used by Lord (1953, p. 523) to denote a probability. To distinguish this concept of item difficulty from the discussion of $P(\theta)$ below, p will be used in this text. It is not to be confused with the p -value of statistical significance.

⁸⁸Interestingly, in his article, Lord does not in any way propagate P as a fixed measure in item analysis to be used in the future. Instead, he uses it as a simple mathematical assumption for his discussion of the relationship between true score and ability level in tests of differing difficulty and discrimination.

Walker 2014, p. 2). Obviously, this procedure demanded a lot of respondents and items to work properly, which may be the reason that p in its simplest form is still the most prominent measure of item difficulty in accounts of CTT, where it is used alongside measures of item discrimination to check item quality (see e.g. Crocker and Algina 2008; Lienert and Raatz 1998; Moosbrugger and Kelava 2008).

Item discrimination As mentioned above, item discrimination in the CTT framework can either be a simple index D or one of several correlation coefficients, amongst which the point-biserial correlation ρ_{pbis} is probably the most well-known (Crocker and Algina 2008, p. 317).

To compute D , the proportion of correct answers to an item in a high-performing group (p_u for *upper* group) is subtracted from the proportion of correct answers to the same item in a low-performing group (p_l of *lower* group).

$$D = p_u - p_l \quad (6.2)$$

The idea of this conceptualisation is that a well-functioning item should be solved correctly by more test takers at a high level of ability than by those who perform less well. Thus, values of D well above 0 are expected: Crocker & Algina (ibid., p. 315) mention a threshold of $D = .30$ above which little or no revision of an item is required. Items with an item discrimination below 0 would actually show *negative discrimination*: such an item would be solved correctly more often by less able test takers than by the high-performing group. Of course, much like the concept of item difficulty described above, D is highly dependent on the ability level and performance of the test takers whose results are analysed. As a result, it cannot be compared across respondents or tests, especially since the two groups of low and high-performing test takers are not clearly defined.

A somewhat more elaborate operationalisation of item discrimination was developed in the form of item-test correlations. There are several ways of calculating these, but the principle is always the same: the score of one item is correlated with the total score of all or all *other* items on the test across examinees (the latter being the point-biserial correlation ρ_{pbis} mentioned above). This correlation will be high when an item is mostly solved correctly by examinees who scored high on the test and not correctly by examinees who scored low on the test. Since it is not necessary to determine groups of low and high achievers, but instead the entire data set is used, item-test correlations contain somewhat more information than D .

All approaches to calculating item discrimination in CTT are static in the sense that they are calculated from the raw scores of a group of respondents without taking into account differences between their abilities or other characteristics. As a result, they are more or less related to CTT's item difficulty p and just as sample dependent. In fact, with the exception of the point-biserial correlation coefficient, item discrimination statistics are most meaningful for items of medium difficulty (ibid., p. 319) and lose much of their predictive power at the extremes of the ability range.

Conclusion CTT did not originally comprise individual item analyses. The tools that *were* eventually developed are sample dependent: they are only meaningful for a group of examinees whose characteristics conform to the instrument's intended audience as much as possible. In the CTT framework, this is not necessarily a problem, since its models rely on the idea of a carefully selected examinee sample anyway. If this condition is met, item analysis can probably fulfil its specific purpose: the identification of items that do not function as expected by the measurement model (Hambleton and R. W. Jones 1993, p. 40). For an analysis of item functioning across groups of test takers and items, however, a more sophisticated approach is necessary.

6.2.1.2 Item Response Theory

As the name suggests, IRT is a measurement theory based on individual item responses rather than raw test scores (as was the case in CTT). In fact, at its core, IRT is concerned with “how performance relates to the abilities that are measured by the items in the test” (Hambleton and R. W. Jones 1993, p. 40): it estimates an examinee’s ability on the same scale as item difficulty based on their item responses and overall performance. Thus, item difficulty and item discrimination are no longer fixed numbers, but parameters in a logistic function. Item difficulty is now understood as a probability, denoted $P(\theta)$: the probability that an examinee with a certain ability θ ⁸⁹ will answer an item correctly. $P(\theta)$ can be plotted as a function of ability (cf. F. B. Baker and S.-H. Kim 2017), resulting in the Item Characteristic Curve (ICC), an example of which is depicted in fig. 6.1.

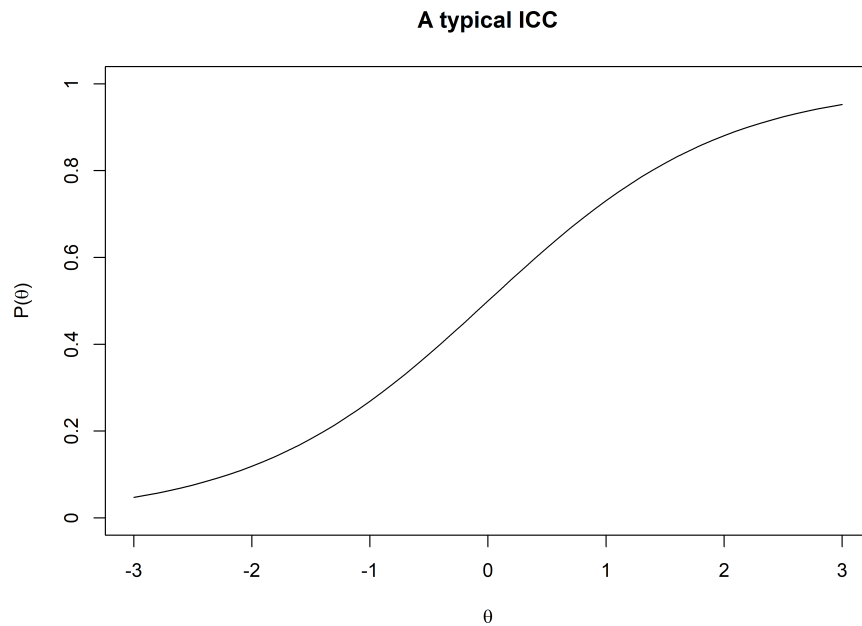


Figure 6.1: A typical ICC under the Rasch model, where $b = 0$ and $a = 1$

This curve is characterised by the following equation:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta-b)}} \quad (6.3)$$

where e is the base of the natural logarithm (a constant: 2.718) and θ is a given ability level. b , then, is the item difficulty parameter; it is defined here as the point on the ability scale where $P(\theta) = 0.5$. a is the item discrimination parameter, which is responsible for the slope of the curve at its steepest point (where, by definition, $\theta = b$). For any given ability level θ (on the x-axis of the ICC graph in fig. 6.1), this function will yield the probability of solving an item correctly (represented by the y-axis). For instance, a student with a low ability level of $\theta = -2$ will only have a 12% chance of solving this item correctly ($P(-2) = 0.119$), a student in the middle of the ability range (at $\theta = 0$) has a 50% chance and a student with a relatively high ability ($\theta = 1.5$) will have an 82% probability of giving the correct answer ($P(1.5) = 0.818$). Similarly, the steepness of the slope at any given θ represents the discriminating power of the

⁸⁹ θ can be estimated in various ways statistically. As a result, person ability estimates are usually spoken of as e.g. Expected A Posterioris (EAPs) or Weighted Likelihood Estimates (WLEs). Technically speaking, these are still θ , but estimated in different ways, with different forms of bias and distortions. See for a somewhat accessible explanation of these, and other person ability estimates, Lechner et al. 2021.

item at that ability: if it is steep, a slight change in ability will noticeably increase or decrease the chances of solving an item correctly. If the slope is rather moderate instead, the item will not differentiate much between test takers in that ability range.

In fig. 6.1, the ICC represents an item of moderate difficulty which fits a one-parameter logistic model (often called a 1-PL model). In this subgroup of IRT models, a is fixed at a certain value for all items (1 in the case of the prominent Rasch model) and only b is estimated, which is computationally light and allows for a reasonably straightforward interpretation of b . Of course, fixing a at 1 also limits the use of the model to a group of items which all discriminate in a similar way (or reduces model fit if this is not the case). Therefore, eq. (6.3) describes an ICC under the two-parameter model (2-PL), where item discrimination can be estimated for each item separately.

Both of these models imply that respondents at a given ability level will (probably) answer an item correctly if they know the answer, whereas if they do not, their answer will (probably) not be correct. Since human nature also allows for lucky guesses and false responses due to distraction, carelessness or other factors, psychometricians have modified the 2-PL model to include one or two more parameters: a guessing parameter c which is added to the probability of answering an item correctly (accounting for guessing) and a fourth parameter d , which “allows a high-ability student to miss an easy item without his or her ability being drastically underestimated” (Liao et al. 2012, p. 1680). These so-called 3-PL and especially 4-PL models are rarely used because their estimation requires a large amount of items and test taker responses to even approach accuracy. Nevertheless, eq. (6.3) expressed under the 4-PL model (where c and d are not fixed at 0 and 1 respectively) would be as follows:

$$P(\theta) = c + (d - c) \frac{1}{1 + e^{-a(\theta - b)}} \quad (6.4)$$

Graphically, c and d are the lower and the upper asymptote of the logistic curve respectively, as can be seen in fig. 6.2.

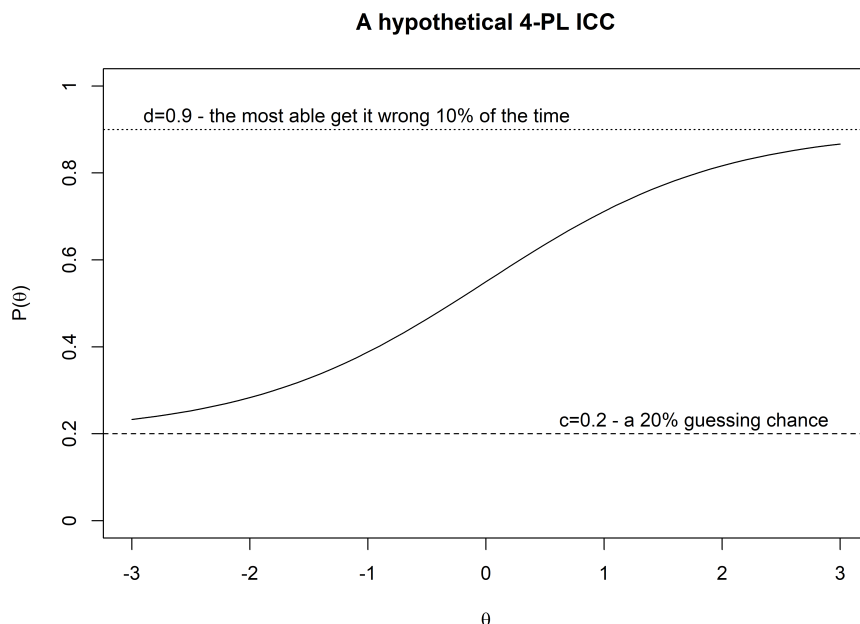


Figure 6.2: A hypothetical 4-PL item characteristic curve, annotated, where $b = 0$, $a = 1$, $c = 0.2$ and $d = 0.9$.

Conclusion In the IRT framework, item difficulty and item discrimination are the two main parameters of a logistic function which links test taker ability and performance. Item difficulty is expressed as the probability of solving an item correctly at a certain level of ability, rising or falling with test taker ability along the logistic curve. This also implies that an item is considered to discriminate best between test takers of a certain ability range (where the curve is steepest), whereas below and above that range, the item will have gradually less discriminating power (as is evident by the asymptotic nature of the ICC). Since ability and the probability of answering correctly are on the same scale, item statistics under the IRT framework are essentially sample independent: as long as the logistic model is assumed to fit the data, the item parameters are meaningful across the ability range and can be determined with an examinee sample that does not have to represent or be limited to any specific ability range—any large enough group of respondents composed of individuals who are able to solve the item and others who are not will be sufficient (cf. Hambleton and R. W. Jones 1993, F. B. Baker and S.-H. Kim 2017).

6.2.1.3 Other methods of estimating item difficulty

In contexts where it is not possible, cost-effective or desirable to collect the amount of test results necessary for a statistical determination of item difficulty, other approaches have been suggested to estimate item difficulty. Among them are ratings by learners or experts (e.g. by teachers, test developers or other individuals considered to have some sort of special knowledge about the subject matter), sometimes using standard setting methods (Watering and Rijt 2006). It is also possible to determine item difficulties *on the fly*, either by letting human raters judge the difficulty of a new item in comparison to other items (which have known item difficulties) or by continually updating difficulty and ability statistics based on new responses (Wauters, Desmet, and Van Den Noortgate 2012). However, as Wauters et al. point out, these approaches are most suitable when it is not (yet) possible to estimate item difficulty using IRT: in their study, as sample sizes increased, the difficulty estimates from all methods under investigation tended to converge towards the item difficulty value obtained by IRT. This lends some credibility to the way item difficulty is determined in IRT models.

6.2.2 Item fit

In the CTT framework, item difficulty and item discrimination are essentially measures of item fit: they *are* the tools used to identify items which do not add information to the test score as expected. Thus, unusually difficult or easy items (which have a low or high p respectively) or items which discriminate badly between more and less able respondents will be revised or eliminated entirely from the test⁹⁰. In IRT models, extreme or unexpected values of b and a may also raise flags, but there are additional ways of determining item quality: item fit and, more relevant to this thesis, DIF. These will be explained in some more details in the following.

⁹⁰In fact, initial item analyses are still conducted with the help of CTT before more sophisticated IRT models are used, e.g. in PISA (OECD 2019c, chapter 9).

6.2.2.1 Item fit statistics in IRT

The most used item fit statistics used in the IRT context is based on an analysis of the residuals⁹¹, i.e. of the difference between the actual response X_{pi} of person p to item i and their expected response E_{pi} :

$$R_{pi} = X_{pi} - E_{pi} \quad (6.5)$$

Since eq. (6.5) can create positive and negative values, the sum of the residuals over the entire test is squared and divided by the number of test takers N to create the *unweighted fit mean square* or *outfit* O_i :

$$O_i = \frac{\sum_{n=1}^N R_{pi}^2}{N} \quad (6.6)$$

The outfit statistic is sensitive to unexpected answering patterns: if high-ability test takers score low on an easy item, or inversely, low-ability test takers solve an otherwise hard item, then the outfit will be high. Since this statistic reacts to outliers, however, a *weighted fit mean square*, the *infit* statistic is often preferred, which includes the variance of the expected responses $W_{pi} = E_{pi}(1 - E_{pi})$:

$$I_i = \frac{\sum_{n=1}^N W_{pi} R_{pi}^2}{\sum_{n=1}^N W_{pi}} \quad (6.7)$$

Both fit statistics are centred around 1. If the value is below 1, then the data shows less variability than expected (so-called *overfit*). *Underfit* occurs if the value is well above 1. How much these values may divert from 1 depends, as Wu demonstrates, on the number of test takers: whereas 95% of the fit indices vary between 0.8 and 1.2 in a simulated data set with 500 respondents, they vary only between 0.92 and 1.08 when 2000 respondents are simulated (M. Wu 2022, ch. 10.2). Hence, instead of using a standard threshold such as 0.8 and 1.2 used in the PISA 2012 study (OECD 2014, p. 151), Wu suggests that the acceptable limits of infit and outfit statistics be determined based on the number of respondents (which was, very simply speaking, done in the most recent PISA study as well; see OECD 2019b, ch. 9, p. 14).

6.2.2.2 Differential item functioning

The concept of DIF has to do with subgroups in the population of test takers who have the same ability on the measured construct but still perform differently on certain items: “Differential item functioning is present when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest.” (Clauser and Mazor 1998, p. 31). Typically, DIF studies are concerned with gender or ethnic group, and related to test fairness: test items should be developed so as to not advantage one group of test takers over another if their measured overall ability is the same. In the context of translated assessments and questionnaires, DIF can be used to ensure the equivalence of the translation at the item level (Zumbo 2003), i.e. the groups which are compared are the test takers of the (two) language versions (see also chapter 7).

DIF is easily understood by plotting the ICCs of the same item for two separate groups of test takers (cf. Zumbo 1999, pp. 18 sqq.). If these are identical or very similar, then there is no DIF. However, if they do differ visibly, then the item differs in difficulty and, possibly, item discrimination for the test takers who belong to the different groups. Zumbo distinguishes further

⁹¹ Apart from *infit* and *outfit*, there are mathematically more sophisticated approaches to both model and item fit in the IRT family. These, however are beyond the abilities of the author, and will not be taken into account in this thesis. The interested reader is pointed to the quite informative technical report of the PISA 2018 study (OECD 2019b), which also contains references for further reading.

between uniform and non-uniform DIF: uniform DIF means that an item is systematically more difficult for one group than for another across the entire ability continuum (fig. 6.3). If an item

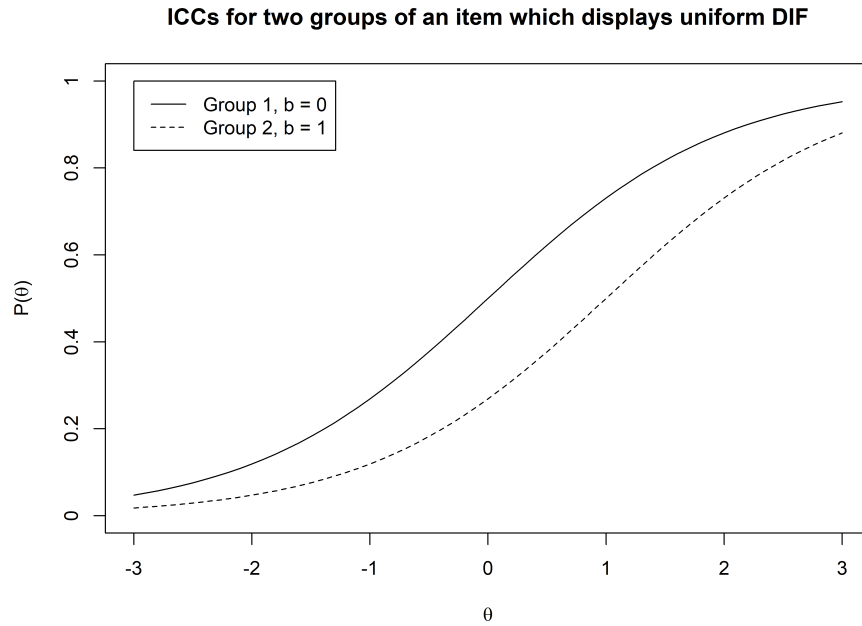


Figure 6.3: A hypothetical item that displays uniform DIF, where $b_1 = 0$ and $b_2 = 1$. b represents the difficulty of the item at $P(\theta) = 0.5$.

shows non-uniform DIF, on the other hand, it advantages one group over another up to a certain level of ability, whereas beyond that, the advantage lies with the other group (fig. 6.4, where respondents from Group 1 are disadvantaged compared to respondents from Group 2 below the ability level of $\theta \approx -0.5$, whereas more able respondents of Group 1 are actually advantaged).

Though implied by the graphs in figs. 6.3 and 6.4, DIF analyses are not necessarily limited to IRT. For instance, the Mantel Haenszel (MH) approach, a well-known DIF test, is based on CTT. Very simply said, the MH approach is a χ^2 test of whether the number of test takers who solved an item correctly is the same in two otherwise comparable groups (Dorans and Holland 1992, pp. 6 sq., see for other techniques Clauser and Mazor 1998). Within the IRT framework, item parameters (i.e. usually item difficulty and item discrimination) are estimated separately for two groups of test takers and then placed on the same scale. To detect and quantify the differences between the item parameters, the area between the ICCs can be estimated, or model fit with and without separate group parameter estimates can be compared (ibid., pp. 33 sq.). One of the most popular approaches to DIF in IRT has been proposed by Lord (1980, cited in Stubbe 2011, p. 468): “To decide if an item functions differently in two specific [...] [groups], the difference between the item difficulties is divided by the standard error of the difference [...] to receive a t value with $n - 2$ degrees of freedom.”

$$t = \frac{b_1 - b_2}{\sqrt{SE_{(b_1)}^2 + SE_{(b_2)}^2}} \quad (6.8)$$

According to Lord (1980, p. 219), “[the] same can be done with the $\hat{a}_i$ ” (i.e. the estimated value of item discrimination in 2-PL models).

Since the various DIF techniques function in different ways and sometimes give wildly different results, this type of analysis is only meaningful when a qualitative evaluation of the items in

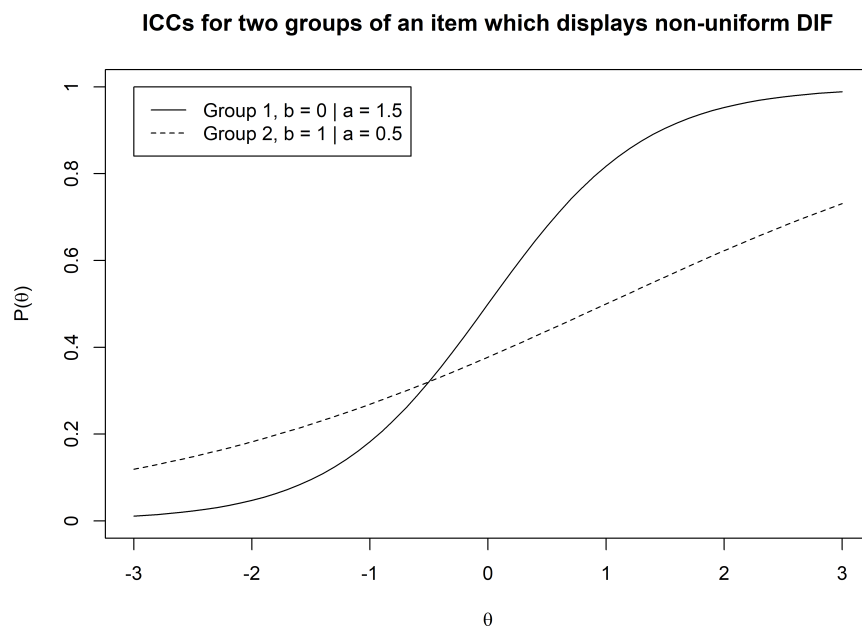


Figure 6.4: A hypothetical item that displays non-uniform DIF, where $b_1 = 0$ and $a_1 = 1.5$, while $b_2 = 1$ and $a_2 = .5$. b represents the difficulty of the item at $P(\theta) = 0.5$, while a stands for the discrimination of the item across the ability continuum.

question is also conducted. It is usually needed to produce an informed judgment on whether the statistically identified DIF has a theoretical basis, i.e. whether a reason can be found for why one group of test takers performs less well than another. This is not a simple matter: the interaction between a test item and any test taker is a complex process, and DIF may happen for any number of reasons:

Potential DIF comes in many forms. [...] An item may be considered to exhibit content DIF if it utilizes knowledge that is not uniformly available to all subgroups within the population of examinees [...] if it uses terms that are not commonly used throughout the examinee population, or which have different connotations within different subgroups. [...] Also, an item may be considered to function differentially in terms of structure and/or format if the structure or format is constructed in such a way as to favor or hinder a particular subgroup of examinees. (Hambleton and R. W. Jones 1992, pp. 4 sq.)

In addition, DIF studies have been conducted in various ways, on differing types of assessment instruments and on the basis of several, sometimes contradictory methods of DIF detection (cf. Ercikan 2002, pp. 201 sq.). As a result, there is no consensus on how DIF should be identified exactly, what counts as substantial DIF and how the sources of DIF can be identified. Nevertheless, the method is a powerful way to investigate the functioning of test items, especially if it is done in concert with a consideration of the test construct (cf. Clauser and Mazor 1998, p. 31; Sireci, Y. Yang, et al. 2006, pp. 555 sq.).

6.3 Explanatory item response modelling

Another way of dealing with group differences in assessment results is the use of explanatory item response models⁹² (Wilson, De Boeck, and Carstensen 2008): instead of a *post hoc* determination of group differences, explanatory item response models use person attributes during the estimation of the item response model. In the same vein, characteristics of the items can be entered into the model beforehand (cf. fig. 6.5). Wilson and colleagues explain that, in addition to and beyond IRT, other statistical techniques are available for this type of analysis. They put forth Generalized Linear Mixed Models (GLMMs) (see for more details De Boeck and Wilson 2004), but they also suggest that nonlinear mixed models and other item response models can be used for explanatory item response modelling (Wilson, De Boeck, and Carstensen 2008, p. 108).

Model	$\eta_{pi} =$			Model Type
	Person part	Item part	Random Effect	
Rasch model	θ_p	$-\beta_i$	$\theta_p \sim N(0, \sigma_\theta^2)$	Doubly descriptive
Latent regression Rasch model	$\sum_j \vartheta_j Z_{pj} + \theta_p$	$-\beta_i$	$\theta_p \sim N(0, \sigma_\theta^2)$	Person explanatory
LLTM	θ_p	$-\sum_k \beta_k X_{ik}$	$\theta_p \sim N(0, \sigma_\theta^2)$	Item explanatory
Latent regression LLTM	$\sum_j \vartheta_j Z_{pj} + \theta_p$	$-\sum_k \beta_k X_{ik}$	$\theta_p \sim N(0, \sigma_\theta^2)$	Doubly explanatory

Figure 6.5: Overview of four explanatory item response models (Wilson, De Boeck, and Carstensen 2008, p. 96).

Put very simply, explanatory item response models estimate item difficulty and person ability parameters *after* they have taken into account systematic variation in person attributes and item characteristics (for a more technical, but reasonably accessible description cf. Petscher et al. 2020, pp. 161 sqq.). In other words, instead of assuming that the interaction between an item and a person is only determined by the item's difficulty β_i and the person's ability θ_p , one or several additional effects related to an item's characteristic k or a person's attribute j as well as, if needed, their interactions, are now added to the two parameters (cf. eq. (6.9)). The authors also present two models in which either item or person characteristics are modelled (cf. fig. 6.5). As such, the object of measurement for explanatory item response models is, as the name suggests, more precise information on how items and/or person characteristics influence test results, whereas descriptive item response models are normally sufficient when item difficulties or, particularly, person abilities are sought (De Boeck and Wilson 2004, p. 7, cited in Stanke and Bulut 2019, pp. 260 sq.).

$$\eta_{pi} = \sum_{j=1}^J \vartheta_j Z_{pj} + \theta_p - \sum_{k=1}^K \beta_k X_{ik} \quad (6.9)$$

In the example put forth by Wilson and colleagues (2008), a sub sample of German test results from the PISA 2003 mathematical literacy assessment was analysed using explanatory item response modelling. The person characteristics under consideration were gender, school program (i.e. the German school forms ranging from *Hauptschule* to *Gymnasium*) and SES. All three characteristics proved to be significant predictors of person ability (ibid., p. 102). As for the

⁹²Despite the name, these models are not necessarily IRT models, or, rather, IRT models are a subset of explanatory item response models (cf. Wilson, De Boeck, and Carstensen 2008, p. 93; Stanke and Bulut 2019, p. 260).

items, topic area (e.g. arithmetic) and modelling type (e.g. numerical modelling) were used, which were derived from the construct description. Though the item-explanatory Linear Logistic Trait Model (LLTM) had a better fit than the descriptive Rasch model, the amount of information added by these characteristics was not substantial (*ibid.*, p. 105), which reinforces the notion that descriptive item response models may be sufficient if the purpose of the analysis is information on the test taker or item parameters.

Explanatory item response modelling has since been used to model item response data in more detail (see Toyama 2019, p. 22; Brizuela and Montero Rojas 2013, p. 2 for references). *Item* explanatory models, i.e. LLTMs, were used in a series of item difficulty studies following Embretson and Wetzel (1987, see e.g. Brizuela and Montero Rojas 2013; Gorin and Embretson 2006; Sonnleitner 2008). Recently, Toyama (2018, 2019) applied a doubly explanatory Latent Regression-Linear Logistic Trait Model (LR-LLTM) to an item difficulty analysis of an L2 English reading assessment (see also chapter 8), using the Rasch model as the link function underlying all models.

Overall, explanatory item response modelling may be a promising approach to item difficulty research as it takes into account not only characteristics of the item but also of the test takers. This is in line with Bachman's problematisation of the concept of item difficulty presented at the beginning of this chapter. If applied through Bayesian IRT, interactions between item and test taker characteristics can be modelled as well, which is another step towards a more precise understanding of item functioning.

6.4 Conclusion

In this chapter, the concept of item difficulty was discussed, first from a conceptual point of view, and then in terms of its definition in CTT and IRT, the two major measurement theories in use in educational research. Alongside, item discrimination was explained, and ways to assess item quality in the IRT context were presented (i.e. item fit statistics and DIF). In both CTT and IRT, all of these item parameters are based on the responses provided by the test takers and on several assumptions regarding the nature of item difficulty. In the IRT framework, for instance, all items in a test are considered to be unidimensional (i.e. they measure one construct only), locally independent (i.e. the probability of solving one item does not depend on another), monotonous (i.e. the probability of solving an item increases with the competence level of the test taker) and invariant (i.e. they function in the same way for all test taker groups) (see e.g. Nguyen et al. 2014). Clearly, considering the complexity of the reading comprehension process and the various factors which influence its assessment (cf. chapters 3 and 4), these assumptions can only ever be approximated, and some measurement error will necessarily remain.

In light of this, the final section considered explanatory item response modelling, which allows for an estimation of item parameters based on known item or test taker characteristics. This family of measurement models retains more information about the test situation than mere test responses, but it also presupposes that the characteristics which influence item difficulty are known. Since this is not yet a given, this issue will be investigated more in depth in the following. Two lines of empirical research can be identified for this purpose: analyses of multilingual assessments, to account for the language versions of the ÜGK foreign language assessment (chapter 7), and item difficulty research, which provides a framework for the analysis of the actual functioning of language assessment items (chapter 8).

Chapter 7

Research on multilingual assessments

This thesis investigates the properties of a *translated* foreign language assessment. As was laid out in section 2.5 and will be described in more detail in section 9.2.4, the four language versions of the ÜGK reading (and listening) assessment were created alongside each other in an iterative item development process. With some minor exceptions, they are identical in terms of content and structure, differing only with respect to the language in which they were presented to the test takers. Nevertheless, psychometric analyses revealed substantial differences between the language versions, which led to their being considered separate tests for the actual ÜGK study (Angelone and Keller 2019b, p. 13). In this thesis, the four language versions of the ÜGK foreign language reading assessment are examined together, in the hopes of shedding some light on why these differences occur.

The practice of translating an assessment is not unusual, though it is rare in the second or foreign language assessment context. In this chapter, empirical investigations of the translation and/or cross-cultural use of assessment instruments are presented, along with some theoretical considerations of what makes a “good” translation, and how the effect of doing a translation can be measured. Most of these discussions regard educational assessments for reading in the language of schooling, mathematics or science, or they revolve around assessments and questionnaires outside of education proper, i.e. psychological, sociological or medical assessment instruments. First, however, some instances of foreign language assessments are explored which were actually employed across linguistic, educational or cultural borders.

7.1 Translated foreign language assessments

The translation of foreign language assessments is either extremely uncommon or there is almost no published research on the subject (or, most likely, a combination of both). A thorough exploration of the literature yielded only three publications which explicitly discuss translated foreign language tests, of which only one (J. L. D. Clark 1965) actually analyses the effect of the translation itself.

Manuel (1953) compared the language ability of college students in an L1 reading assessment and its L2 counterpart (using parallel editions of the *Cooperative Inter-American Tests*). He found that the students ($N = 232$, L1 English) who sat both versions of the same test performed much better in the English version than in the Spanish version. Comparing percentiles, Manuel concludes that the “average student [in their second year of college] reads Spanish less efficiently than the average seventh-grade pupil [...] reads English” (drawing from test data of another standardised reading test used in schools, *ibid.*, p. 434). He also reports a reasonably high correlation of .59 between the total scores in the Spanish and the English version of the reading test, which he interprets as evidence that students’ success in foreign language learning is related to the “mastery of their own language”, though it is “sufficiently low to show that achievement in reading English is but one of the factors associated with achievement in reading Spanish as

a second language” (ibid., pp. 435 sq.). Unfortunately, the author does not discuss the actual comparability of the two test versions beyond stating that “it will be assumed that the same score on the two editions represents equal ability in English and in Spanish” (ibid., p. 431).

A decade later, Clark (1965) *did* investigate the consequences of translating a foreign language assessment. He reports on several reading and listening tests which were translated on an item-by-item basis from Spanish into Brazilian Portuguese, staying as close as possible to the original wording but allowing for changes to make the Portuguese version “read naturally” (ibid., p. 78). Both language versions were administered to a varying number of test takers in language learning programs. Clark analysed the test results using CTT after equating examinees from both language versions on their raw score ($N = 2 \times 42$ up to $N = 2 \times 73$) and found medium to strong correlations between the measures of item difficulty on each item (proportion correct, $r = .52$ up to $r = .79$). He also calculated point-biserial correlation coefficients for each item (measuring item discrimination) and arrived at medium correlations between the two language versions ($r = .30$ up to $r = .45$). Looking more closely, he investigated the answering patterns on the multiple choice response options ($N = 274$) and identified about 30% of all items which significantly differed between the two language versions (chance probability below 10%, $p < 0.10$). Of those, according to the author, “several cases [contained] previously unnoticed flaws in item content” (ibid., p. 82), either related to the frequency or idiomaticness of words or phrases in the item or to textual overlap between an incorrect option and the correct option present in the Spanish version but not its translation. Clark concludes his analysis by stating that item-level test translation seems to be possible as long as the two languages are as closely related as Spanish and Portuguese. This, of course, is often not the case.

Both Manuel and Clark, using the accepted measurement theory of their time, consider the raw test results on the translated language versions to represent equal ability, in their case reading and/or listening in the test language. Obviously, this assumption is hardly tenable from a modern measurement perspective. Nevertheless, Clark’s findings especially are interesting, even though they are visibly based more on the author’s intuition than on any hard criterion. They anticipate the results of modern studies on item difficulty, as will be shown in chapter 8: variables related to word choice and lexical overlap between all parts of a test are often among the better predictors of item difficulty.

Considerably more recent is Zumbo’s (2003) discussion of a translated foreign language assessment. In a simulation study using item parameter estimates from a selection of TOEFL items, the author shows that common scale-level analyses of measurement invariance (e.g. coefficient alpha) are not powerful enough to detect even considerable DIF. This study is based on simulated test response data from an imagined assessment “that has been translated for use in a second language and culture” (ibid., p. 137). Unfortunately, this is where the discussion of the pretend-translation ends: no mention is made of the fact that it is very uncommon to translate foreign language tests such as TOEFL. This is all the more remarkable in light of the test section the author chose to investigate: “a test of knowledge of structural and grammatical elements of standard written English” (ibid., p. 142)—a construct one may assume to be much more difficult to “translate” to another foreign language than, for instance, reading comprehension. Thus, while interesting from a measurement perspective and a strong argument for the use of DIF methods, Zumbo’s study does not add much to the examination of translated foreign language assessments *per se*.

7.2 Cross-cultural foreign language assessment

DIF is also the focus of two publications on the European Bank of Anchor Items for Foreign Language Skills (EBAFLS), a since abandoned European project, which “investigated the possibility of producing banks of calibrated [foreign language assessment] items [...] with a known level as defined in the [CEFR]” (Noijons and Gille 2010, p. 49, see also Tardieu 2010). Though

the project did not investigate a translated foreign language assessment, it generated interesting findings with respect to the cross-cultural use of foreign language assessment items.

The project investigated the feasibility of using the same foreign language items across teaching contexts (i.e. countries), assessing French, English and German listening and reading comprehension in at least four European countries each. Information on this project is limited, but the two publications cited above make it clear that teachers and students in the participating countries employed a wider range of foreign language assessment methods than anticipated, with notable differences in how test items were designed. Thus, a pilot study was undertaken in which items collected from the participating countries were administered to students who were expected to have reached a proficiency level of B1 in their respective school systems (Tardieu 2010, p. 5). To assure the meaningfulness of the international assessment, the researchers studied both item fit and DIF in detail, finding “that the vast majority of the items [showed] a substantial amount of DIF” (Noijons and Gille 2010, p. 51), both in item difficulty and item discrimination⁹³. Noijons and Gille (ibid., p. 58) advance five possible reasons for these differences, though they were not empirically investigated:

- students are not familiar with the item format⁹⁴;
- items focus on a sub-skill (such as grammar), rather than on communicative skills;
- items contain vocabulary that is unknown;
- items require knowledge of the world;
- the vocabulary in items may contain false friends in a particular Mother Tongue.

In addition to these aspects, the authors point out that the different educational cultures, including curricula, teaching practice and “l’identité culturelle des tests propre à chaque pays” (Tardieu 2010, p. 7) may have caused the large amount of DIF that was found in the project. Overall, it seems that the differences were such that the effort of creating a European test item bank for foreign languages was abandoned.

Around the same time that EBAFLS was conducted, the English *Asset Languages* project aimed at providing “voluntary ‘low stakes’ tests in 25 languages”, which were supposed to be comparable in the sense that learners of different languages who scored at the same level of the “Languages Ladder”⁹⁵ had comparable functional proficiency of the language (Ashton 2014, pp. 105 sq., see also Ashton, N. Jones, and T. Walker 2010). The project was scaled back soon (Ashton 2014, fn. 1) and seems to have disappeared since, but from the information available, it is clear that, perhaps rightfully so, the focus was on the comparability of can-do statements, rather than on any psychometric equating. The reading and listening assessments in the various languages seem to have been developed separately, and then undergone standard settings to align them with the Language Ladder’s proficiency scales (see Ashton, N. Jones, and T. Walker 2010). The few published findings from this project make note of the difficulty of describing and assessing A1 language proficiency (“the level is predominantly defined by what learners cannot do” (ibid., p. 234)) and of assessing reading and writing proficiency in languages which do not use the Latin script (Ashton 2014; Ashton, N. Jones, and T. Walker 2010).

⁹³Tardieu (2010, p. 3) also reports on the effect that translating the instructions and items into the target language (which had become necessary because some countries formulated the items in the local language of schooling and others did not) had on the test results: “Les résultats montrent en premier que le fait de traduire tous les items dans la langue cible induit des réponses qui n’avaient pas été prévues par les concepteurs du test original dans les consignes de correction ou leur traduction. Autrement dit, le fait de changer la langue des consignes et des items a modifié le caractère même des items d’une manière imprévisible”.

⁹⁴See also ibid., p. 6.

⁹⁵According to Ashton, this is a series of “can-do statements for reading, writing, listening and speaking at six levels (broken down to 17 grades)” (Ashton 2014, p. 106), perhaps similar to the Lingualevel descriptors (Lenz and Studer 2007).

Another, somewhat more successful project which investigated foreign language proficiency across languages, was *SurveyLang* (European Commission: Directorate-General for Education, Youth, Sport and Culture 2012; N. Jones et al. 2012). In this European project, approximately 54.000 students in 16 different educational systems across Europe were assessed in a foreign language⁹⁶. The assessment targeted listening, reading and writing in English, French, German, Italian or Spanish. Depending on the educational system's language teaching policies, the two languages were assessed at the end of lower secondary school (ISCED 2) and/or in the second year of upper secondary school (ISCED 3) (N. Jones et al. 2012, pp. 9 sq.). Interestingly, context questionnaires were given to students, language teachers and school principals.

Though the SurveyLang tasks were created by language-specific teams, the project used cross-language vetting and some common (i.e. adapted) tasks to ensure the comparability of the test results (European Commission: Directorate-General for Education, Youth, Sport and Culture 2012, pp. 37 sq.). The common tasks were selected based on test results from an extensive pilot (ibid., p. 38). They were subsequently used as anchors during the standard settings which linked the test results to the CEFR levels. A major finding of this study is that there are large differences between the various educational systems in terms of the foreign language proficiency achieved by their students. While in some countries only a small minority of students scored beyond A1, in others, relatively large groups of students reached at least B1 in both foreign languages tested in SurveyLang (ibid., pp. 40 sqq.). Through the results of the contextual questionnaire, it is also clear that the language learning environment in these educational systems differs in many respects. Some relationships between aspects of this environment and learning achievement have also been examined. Most importantly, a larger amount of lessons, informal exposure to the language outside the classroom and the students' perception that learning the language is useful generally resulted in higher test scores (ibid., pp. 100 sqq.). The comparability of the language tasks across languages, however, was considered a given at this point, following the careful selection of common tasks and a standard setting. It is unclear whether the project also dealt with DIF due to the educational system or language version of the tasks. Much like in the original ÜGK project, of course, that was not the aim of the project.

7.3 Translated assessments and surveys

Outside of foreign language assessment, test translation⁹⁷ is much more common and has been practised for a long time (ElAtia 2013, pp. 193 sq.), though it exclusively refers to translations of tests or surveys into another group's main language⁹⁸. As such, test translation is either a form of test accommodation (e.g. to include a minority language group into a medical survey) or a tool in surveys and assessments that are administered across two or more linguistic groups (Maneesriwongul and Dixon 2004; Sireci and Allalouf 2003). Also, translating an existing instrument is sometimes considered to be less expensive than creating a new one (Bolaños-Medina and González-Ruiz 2012, p. 716).

In sociological and psychological research, in fact, test translation is frequent enough that the International Test Commission (ITC), an association of national psychological associations and related bodies, has published the *ITC Guidelines for Translating and Adapting Tests* (Intern-

⁹⁶Switzerland did not participate in this study.

⁹⁷Some authors prefer the term *test adaptation* (or even transadaptation, cf. Cohen, Gafni, and Hanani 2007), which underlines the fact that a "test translation" is usually more than the literal translation of the words (cf. Sireci, Y. Yang, et al. 2006, note 2; Van de Vijver and Tanzer 2004, p. 122; ElAtia 2013, pp. 196 sqq.). In some cases, a test adaptation may not even include translation, but only a change of content, scoring procedures or test administration to account for cultural differences between target groups (cf. e.g. the Warsaw example in Van de Vijver and Tanzer 2004, pp. 119 sq., see also Cohen, Gafni, and Hanani 2007).

⁹⁸The literature discussed in this section covers both educational assessments, e.g. PISA, and psychological, sociological or medical surveys, which are often questionnaires. The term *test translation* (or *adaptation*) is used in both contexts, see e.g. ElAtia 2013; Sireci, Y. Yang, et al. 2006.

tional Test Commission 2017), which have their origins in 1996 (Van de Vijver and Hambleton 1996). These guidelines cover a broad spectrum of aspects related to the cross-linguistic and/or cross-cultural⁹⁹ adaptation of tests, including copyright issues, the choice of language experts and item formats, linguistic aspects to consider as well as a strong appeal to include small-scale try-outs, medium-scale piloting sessions and large (enough) psychometric calibration studies. Aside from these more technical aspects, the Guidelines also address the need for construct validation:

Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores.

Explanation. This guideline requires that what is assessed should be understood in the same way across language and cultural groups [...] (International Test Commission 2017, p. 9, *bold in original*).

Following this fundamental stipulation, the guidelines call for specific consideration of the underlying construct (e.g. sources of construct-irrelevant variance) and a thorough documentation of the translation and adaptation process. All of this is quite clearly evocative of a modern view of validity in educational (and psychological) measurement (cf. Messick 1990; M. Kane 2006; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014, pp. 60 sq.).

Researchers in various fields have investigated the effect of test translations, which included the search for suitable translation methods and for adequate ways of operationalising translations and translation errors. These two topics will be reviewed briefly in the following sections, along with a short overview of empirical studies in the field. Section 7.4 will then report on empirical studies which looked at the translation of *educational* assessments, more closely related to the topic at hand.

7.3.1 Methods of test translation

Conscious of the changes to scores brought about by the translation of tests and surveys (Ercikan 1998; Hambleton 1993; Sireci, Y. Yang, et al. 2006; Van de Vijver and Tanzer 2004), researchers have explored and discussed methods of test translation for a while (e.g. Brislin 1970, though see Bolaños-Medina and González-Ruiz 2012, pp. 728 sqq. for a critical discussion from the point of view of the translation sciences). Sireci and colleagues identified two main consensus guidelines for the translation or adaptation of tests and surveys: several instances of quality-control checks by trained translators during test development, and statistical analyses focused on differences between language versions after the test has been administered (Sireci, Y. Yang, et al. 2006, p. 558). In addition, the field has developed and trialled a variety of translation methods which are supposed to identify problematic items already during test development. Van de Vijver and Tanzer (2004, pp. 122 sq.) describe the widely known *translation-backtranslation procedure*, where a text is translated into another language, and then translated back into the original language, without referral to the original text (see also Sireci, Y. Yang, et al. 2006, p. 560). A comparison of the two texts in the original language may then reveal certain types of errors, especially linguistic differences between two language versions of an item (Hulin 1987, pp. 117 sq.). This, in turn, can be a point of departure for an item review, though the method has its flaws (Behr 2017; Ozolins et al. 2020). Back translation is in wide use in the medical professions (Kuliš et al. 2017; Ozolins et al. 2020), but has largely been abolished in the translation of sociological and educational assessments (Zhao and Solano-Flores 2021, p. 679; Behr 2017, p. 575).

⁹⁹Note that the term *culture*, here mentioned as an entity *across* which adaptations are made, has its own substantial body of literature. Much like *language*, it is not easily defined and has even been denied conceptual validity altogether (e.g. Altmayer 2002). An accessible discussion of the term and its neighbouring concepts in the context of assessment translation can be found in Solano-Flores (2021, pp. 3 sq.). While undoubtedly relevant and of interest, this discussion lies beyond the scope of the present thesis.

Another method is the *committee approach*, where experts from various fields collaborate to create and/or review the translation (Van de Vijver and Tanzer 2004, p. 123). This is increasingly the preferred method of creating a translation (Behr 2017, pp. 575, 582), because it allows for identifying specific problems with the translated language version that the back translation method cannot account for. On the one hand, these are formal problems of the target version, ranging from spelling mistakes to unidiomatic language. On the other hand, the committee approach also allows an evaluation of the translated version by experts of various fields, which may lead to the identification of more “subtle” problems such as potential sources of misunderstandings or items which will probably function differently, due to linguistic or cultural differences between the source and the target culture.

Once a translated assessment has been created, pilot studies can be used to determine whether the two language versions can be considered equal. Apart from qualitative examinations of the cognitive processes that respondents undergo, researchers use statistical methods which can identify group differences and item bias, particularly DIF statistics (Sireci and Allalouf 2003) and IRT (Hambleton and R. W. Jones 1992; Hulin 1987; Sireci, Y. Yang, et al. 2006, see also section 6.2.1.2)¹⁰⁰.

Overall, a combination of these approaches is often the best choice, especially when little is known about the construct, the method or the target population(s). This is underlined by empirical research which shows that even a careful translation process cannot account for all sources of DIF (Sireci, Y. Yang, et al. 2006, pp. 555 sq.; Cohen, Gafni, and Hanani 2007). One such combined approach has been suggested by Vallerand (1989, see also Bolaños-Medina and González-Ruiz 2012, p. 727) based on a literature review. He describes seven phases “pour la validation trans-culturelle de questionnaires psychologiques” (Vallerand 1989, pp. 663 sq.): the creation and evaluation of a preliminary translated version through back translation (1) and the committee approach (2), a small-scale pretest which focuses on the comprehensibility of the preliminary translation (3), a “validation” using the bilingual method (4), a reliability analysis (5) and a further “validation” using factorial analyses and interitem correlations, as well as a construct validation based on theory (6). At all points, if the quality criteria are not met, Vallerand suggests that one turn back to a revision of the translated test using the committee approach. Once the instrument is deemed adequate, the seventh step is the establishment of norms. Having been published before Messick’s reinterpretation of validity (1990), and visibly based on the principles of CTT, Vallerand’s suggestion seems somewhat outdated. Nevertheless, being clearly structured and as such directly applicable, it is still in use, especially in the medical sciences, though often in a somewhat condensed form (e.g. Chaniaud, Sagnier, and Loup-Escande 2022).

7.3.2 Measuring the effect of test translation

When a text is translated into another language, one might envision the existence of a measurement tool that quantifies the disparities between the translated version and its original, with the goal of identifying a suitable translation for a specific purpose. Regrettably, the reality is far more intricate than initial impressions suggest. Consequently, there is no straightforward method to measure the impact of translation tests. Instead, different disciplines employing translations approach the problem in different ways.

In *translation science*, defining a *good* or *appropriate* translation, which is produced by a *competent translator* (Hague, Melby, and Zheng 2011, pp. 244 sq.), has proved just as elusive

¹⁰⁰To achieve a direct comparison of the scores, the *bilingual method* is sometimes suggested, where respondents who know both languages answer items from both versions of the test (Maneesriwongul and Dixon 2004, pp. 176, 181; Sireci 2005, pp. 125 sq.). The results from both language versions are then compared with the expectation that, if there are differences, they are due to translation problems. This, however, implies assumptions about bilinguals that are now generally recognised to be presumptuous. For instance, bilinguals seldom have the exact same level of proficiency in their two languages and, more importantly, they do not necessarily adequately represent monolingual respondents (Sireci 2005, pp. 123 sq.; Bolaños-Medina and González-Ruiz 2012, p. 726).

as defining a *good text* or a *good reader* in reading research¹⁰¹. Propositions have been made, e.g. a good translation has been produced when it reaches “optimum equivalence” of both the source and the translated text (the equivalence-based approach, Lauscher 2000, pp. 151 sqq.), or when the “target reader [reaches] a coherent interpretation of the target text” (the functional approach, *ibid.*, pp. 156 sq., see also Hague, Melby, and Zheng 2011). However, as Lauscher points out, a text is only very rarely just the sum of its words. Instead, the context of its production and reception always has an influence and cannot be entirely separated from a text’s content. When it is translated, that context is added to the matter, making an evaluation of translation quality depend on the text in both languages, the personality and interpretation of the author(s) and translator(s), and the contexts of the text’s production and reception, with the latter evolving over time (Lauscher 2000, pp. 158 sqq.). As a result, Lauscher suggests that both translation critics and translators themselves need to be more aware of the complexity of the process: “Translation quality assessment and the judgement of translations are a matter of communication, co-operation and consent” (*ibid.*, p. 164). Unfortunately, this cooperative, discursive approach does not lend itself to a quantitative study of translation quality.

Translations of psychological or sociological instruments are most often considered in terms of *bias* and *equivalence* (cf. in particular Bolaños-Medina and González-Ruiz 2012; Van de Vijver and Tanzer 2004). The terms refer to the interpretation of the measurement scores: to what extent will two measurement instruments used under different conditions yield measures of the same construct? Both concepts, bias and equivalence, are understood hierarchically, from a state where two instruments do not even measure the same construct (i.e. construct bias occurs, and construct equivalence is absent), to two instruments being comparable in the sense that a score *X* on the one instrument means the same as a score *X* on the other (i.e. there is scalar equivalence and, necessarily, no bias) (Van de Vijver and Tanzer 2004, pp. 120 sqq.). In this logic, translations and adaptations need to avoid or minimise the creation of bias, and validation studies of a translation need to produce evidence of the supposed level of equivalence. This conceptualisation is useful as it allows for (translated) assessment instruments to be connected in various ways: they may “only” measure the same construct, have the same measurement unit, but different (often unknown) offsets, or they may share a scale. In addition, showing evidence of the absence of bias requires researchers to consider a number of potential sources of *inequivalence*: the construct, the method (which includes the selection of the sample, the choice of the method including the format and the test administration itself) and the items, in the form of DIF (*ibid.*, pp. 120 sq., 124 sqq.). This view leads to the understanding that translating an instrument across languages and cultures requires a great deal of care and multiple lines of evidence, both qualitative and quantitative (*ibid.*, pp. 128 sqq.).

To account for the effect of the test translations themselves (while assuming that there should be little construct or method bias), Solano-Flores and colleagues have proposed the Theory of Test Translation Error (TTTE) (2009)¹⁰². With translation errors defined as “the lack of equivalence between the source language version and the target language version of test items”, this framework can, according to the authors, be used to identify items which present “fatal” translation errors, i.e. errors which are likely to introduce measurement error (*ibid.*, pp. 83 sq.). The framework is constructed of test translation error dimensions, which are broad categories of possible sources of translation error that have been identified to be relevant in a given test translation (e.g. dimensions related to the design, language and content of an item). Each dimension contains a variety of error types, which are essentially descriptions of possible problems of the new language version (e.g. Item Design–Conventions: “*grammatical inconsistency between stem and options in multiple choice items*” or Content–Curriculum: “*the target knowledge or skill is not taught at the*

¹⁰¹Note that in the references cited here, the focus lies more on long texts, particularly literary and expository productions.

¹⁰²In fact, the concept of translation errors is also known in translation science, where it stems mostly from a consideration of how to assess student translators at universities or candidates for translation posts in institutions, cf. Waddington 2001.

corresponding grade level”, cf. *ibid.*, p. 82). The dimensions and error types are established for a specific test translation by a team of reviewers. An important aspect of TTTE is the idea that translation errors tend to be multidimensional, e.g. a stylistic change may very well also represent a change in the item’s information (Solano-Flores, Backhoff, and Contreras-Niño 2009, p. 83; Zhao, Solano-Flores, and Qian 2018, pp. 19 sq.). This also implies that some translation errors are serious and need to be remediated while others cannot be avoided (Solano-Flores, Contreras-Niño, and Backhoff 2013, pp. 72 sq.).

Solano-Flores and colleagues have shown in several studies that review teams can successfully identify problematic items using this method, and that the various dimensions are at least partly transferable to other contexts and language combinations (Solano-Flores, Backhoff, and Contreras-Niño 2009; Solano-Flores, Contreras-Niño, and Backhoff 2013; Zhao and Solano-Flores 2021; Zhao, Solano-Flores, and Qian 2018)¹⁰³. TTTE is meant to introduce a more systematic approach to test translations (Solano-Flores, Contreras-Niño, and Backhoff 2013, p. 83) and it may also allow for more rigorous investigations of the sources of test translation error. Nevertheless, the approach is still judgmental in the sense that the dimensions established depend on the review team’s expertise and that their validity is still up for empirical review.

Overall, the various scientific fields have come up with only one promising way of identifying problems in test translations: human review, if possible by a multiprofessional team which unites expertise from several fields. The review may be fruitfully supported by psychometric analyses, particularly the identification of DIF items. Apparently though, the sources for differences between translated items are so varied, reaching from “pure” translation errors to subtle cultural differences between groups of test takers, that it is as of yet impossible to adequately measure the effect of a test’s translation. This is where studies of actual test translations as well as item difficulty research are relevant: if we know more about the functioning of assessments tasks, within *and* across cultures, we might be able to better predict it, both for test development and test translation. Both types of research will be discussed in the following.

7.3.3 Empirical research on translated assessments and surveys outside education

Outside education, the translation or adaptation of surveys and tests into another language is so frequent that the studies discussed below can only be illustrative. They have been chosen mainly to give an insight into the variety of approaches and applications, though they do not necessarily represent particularly good or well-known examples.

Medical questionnaires are presented to patients and caregivers for a variety of reasons, e.g. to establish a measure of their Quality of Life (QOL) before, during or after some kind of treatment. In this field, it seems to be generally assumed that it is better to use a translation of an “established” measure (often meaning an instrument which has been demonstrated to have adequate psychometric properties) than to develop one’s own (e.g. Swami and Barron 2019). A recent example is the translation of an eHealth questionnaire into French (Chaniaud, Sagnier, and Loup-Escande 2022). The questionnaire consists of eight items using a 5-point Likert scale to indicate agreement, and the study includes a detailed description of both the translation and adaptation process as well as results of a validation effort. After a twofold forward- and backward translation, the adapted language version was discussed in a committee, and presented to a small group of native speakers who evaluated and commented on the clarity and comprehensibility of the items. The final version from these three steps was then administered to members of a chosen target population (university students whose L1 is French), along with a

¹⁰³Unfortunately, it seems that the authors did not have access to the DIF values of the items they investigated. Instead, they correlated their results with item difficulty (using the *p*-value), finding mostly low negative correlations. Having studied science and mathematics items, the authors considered a measurable effect of translation errors on item difficulty to be evidence of a problem (Solano-Flores, Contreras-Niño, and Backhoff 2013, pp. 80 sq.).

background questionnaire and two other instruments which measured a neighbouring construct (health literacy). A subgroup of this sample repeated the assessment one month later. The authors reported on the internal consistency of the instrument (Cronbach's α and interitem correlations), a factor analysis, and correlations of the test-retest scores, and the instrument with the other variables collected during their study, concluding that their translation presented a valid measurement of eHealth literacy. As such, the authors clearly followed Vallerand's seven phases of test translation (1989), though, apparently, they skipped the theoretical aspects of construct validation. Interestingly, in this field, translations are "validated" in their own right, without any referral back to the original instrument once the translation has been produced¹⁰⁴. Hence, while translating an existing questionnaire cut down on production costs for the French questionnaire, the authors did not establish measurement equivalence between the two questionnaires.

Inversely, Elosua and López-Jáuregui (2008) reported on the psychometric equating of two language versions of a measure of self concept, a questionnaire with 31 ordered response items that load onto four dimensions (the Autoconcepto Forma A (AFA), Musitu, F. García, and Gutiérrez 1997). Their goal was to create a common scale, which could then be used to compare the scores of respondents to either version with each other. Situated in the Basque Country in Spain, where Basque and Spanish coexist, the authors assumed that there would be no cultural difference between the speakers of the two language groups and that as a result, the underlying construct would be comparable (see also F. Garcia et al. 2018 for a similar study involving Brazil and Spain). Hence, using an analysis of factorial invariance, they compared the factorial structure of both language versions, identified items which functioned similarly in both versions, and used those as anchor items to determine a common scale (Elosua and López-Jáuregui 2008, pp. 390 sqq.). According to the authors, this equating method, requiring only a subsample of items of a test, allows for a common interpretation of test scores even in the absence of theoretical equivalence (i.e. the construct is considered to be equivalent but there is no evidence to claim the equivalence of the tests).

Another study which is also methodological in nature, is an investigation of three national versions of an employee attitude survey by Sireci and colleagues (2003). Studying responses on 50 likert-scale items, the authors investigated construct equivalence and measurement equivalence between the English, French and Spanish versions of the questionnaire, using Weighted Multidimensional Scaling Analysis (MDS) and two DIF detection methods. MDS is described as an alternative to Structural Equation Modeling (SEM) for explorative dimensionality analyses (ibid., pp. 134 sqq.). Overall, the authors found that the dimensions corresponded well to the instrument's intended dimensions across all three language groups. The DIF analysis identified a small numbers of items related to "interpersonal relationships and emotions" (ibid., pp. 145 sqq.), which shows the difficulties of cross-cultural translations. Overall, the authors concluded that their statistical methods were well-suited for their purpose (see, for a similar study on an educational assessment, Sireci and Khaliq 2003), and that the instrument under investigation performed equally well across the three language versions.

¹⁰⁴For instance, although the sample populations differed somewhat, the factor analysis of the French instrument revealed a different loading structure than the validation study of the English original (cf. Norman and Skinner 2006). This, however, did not lead Chaniaud and colleagues to doubt their translated version. Interestingly, the inverse strategy is also applied: Juul and colleagues, for instance, used the "validation" of the original language version as a benchmark for their own study (2015, p. 910).

7.4 Empirical research on translated educational assessments

Large international assessments such as the Trends in International Mathematics and Science Study (TIMSS), PISA or the Progress in International Reading Literacy Study (PIRLS) rely on the notion that their items are equivalent across the various language versions: if they were not, it would not be possible to establish rankings of the participating countries, i.e. without measurement equivalence, the scores should not be compared with each other (Kankaraš and Moors 2014). As noted above, measurement equivalence between language versions of an assessment is not a purely linguistic matter: differences between language versions of an educational measurement may be due to “poorly translated items” (Ercikan 1998, p. 544), but they may also be created by other forms of bias. According to Ercikan (*ibid.*, p. 544), educational measurements typically evaluate multidimensional abilities that can be influenced by “cultural, language, and curriculum differences”. For instance, students from diverse curricular regions may not have covered the same subjects, or they may have had different amounts of exposure to certain domains (see also Stubbe 2011, p. 466). Variances in instructional cultures can also lead to varying levels of familiarity with task formats, text types, technologies, or even divergent strategic approaches to test items. Clearly, it is crucial to consider these underlying variations when examining the impact of test translations.

There are numerous studies which examined the effect that a translation had on the test results in order to determine whether test taker abilities could reasonably be compared based on the test results. These studies usually investigate whether the assessment contains items that show DIF, i.e. whether test takers from the different language groups solve certain items in different ways even though they had the same level of ability (see also section 6.2.2.2). More often than not, those studies reveal a large quantity of items which show DIF. For instance, in an influential study of the PISA 2000 Reading results, Grisay and Monseur (2007) compared item difficulties among and across the national test versions in English, French, German, and Spanish. Within the same language but across countries, if the same language version was used with only minor adaptations, they found very little DIF (less than 3% of all items on average), whereas the five Spanish versions, which had been translated independently of each other, showed visibly more DIF (17% of all items, see also Stubbe (2011) for a similar analysis of three German versions of the 2006 PIRLS reading assessment). This finding, the authors concluded, is evidence that test translations “always [have] at least a basic cost in terms of loss of equivalence” (Grisay and Monseur 2007, p. 82). Furthermore, comparisons *across* languages showed DIF in about 30% of all items, increasing to almost 50% when the test results from Asian countries were considered (see fig. 7.1). In further analyses of this data, the authors established a measure of commonality between the various language version, and found that Asian countries and countries with low gross domestic products (GDPs) differed the most from all the other language versions (*ibid.*, pp. 74 sqq., see for similar results on the PISA 2006 Science items: Grisay, Jong, et al. 2007). A part of this difference might have been explained by the item format (Grisay and Monseur 2007, pp. 77 sq.), but, the authors still stressed, the distorting effect of test translation remained and should be taken into account during the translation process.

Kankaraš and Moors (2014) analysed the PISA 2009 results from 64 countries and found moderate DIF in the majority of the items. They also estimated the effect this had on the country scores, finding decreases or increases of up to 60 points on the PISA scale ($\mu = 500$, $sd = 100$) (*ibid.*, pp. 391 sqq.). Among the three PISA assessments, the strongest nonequivalence was found in the reading scale, which led the authors to conclude that “language specific characteristics are difficult to translate in an universal test” (*ibid.*, p. 396). They also pointed out that countries in the same general region (e.g. Southeast Asian, Baltic or Scandinavian countries) are more likely to have equivalent results, i.e. that the results of countries in the same cultural group are more comparable to each other. Finally, their analysis revealed that the majority of DIF items is uniform, i.e. the b -parameter of 2-PL item response models show the most amount of DIF whereas the a -parameter, item discrimination, is less susceptible to DIF.

	...the group of English versions	...the group of French versions	...the group of German versions	...the group of Spanish versions	...all participating countries
English versions					
IRL.ENG	4.8%	28.0%	28.0%	31.2%	14.4%
CAN.ENG	0.0%	27.2%	30.4%	26.4%	15.2%
AUS.ENG	0.0%	27.2%	23.2%	28.8%	16.0%
NZL.ENG	0.8%	28.0%	24.8%	32.8%	16.8%
QSC.ENG	1.6%	29.6%	31.2%	32.0%	19.2%
QUK.ENG	3.2%	29.6%	25.6%	34.4%	20.8%
USA.ENG	6.4%	36.0%	35.2%	33.6%	22.4%
Mean ENG	2.4%	29.3%	28.3%	31.3%	17.8%
French versions					
QBR.FRE	24.8%	0.0%	25.6%	23.2%	16.0%
CHE.FRE	32.0%	2.4%	24.0%	28.8%	17.6%
CAN.FRE	21.6%	5.6%	21.6%	28.8%	18.4%
FRA.FRE	36.0%	2.4%	31.2%	31.2%	24.0%
Mean FRE	28.6%	2.6%	25.6%	28.0%	19.0%
German versions					
LUX.GER	22.4%	19.2%	1.6%	24.8%	15.2%
DEU.GER	25.6%	22.4%	0.0%	31.2%	16.0%
CHE.GER	27.2%	23.2%	5.6%	33.6%	18.4%
AUT.GER	28.8%	29.6%	1.6%	37.6%	20.0%
Mean GER	26.0%	23.6%	2.2%	31.8%	17.4%
Spanish versions					
ESP.SPA	28.8%	22.4%	27.2%	17.6%	17.6%
ARG.SPA	28.2%	25.8%	36.3%	9.7%	20.2%
CHL.SPA	38.7%	40.3%	41.9%	12.1%	29.0%
MEX.SPA	44.0%	44.8%	46.4%	22.4%	32.8%
PER.SPA	53.2%	47.6%	46.8%	23.4%	38.7%
Mean SPA	38.6%	36.2%	39.7%	17.0%	27.6%
Asian versions					
HGK.CHI	36.8%	44.8%	43.2%	41.6%	32.8%
JPN.JAP	43.2%	38.4%	36.8%	45.6%	35.2%
THA.THA	48.4%	46.0%	47.6%	37.9%	36.3%
KOR.KOR	45.6%	45.6%	45.6%	53.6%	36.8%
IDN.IND	48.4%	55.6%	50.0%	45.2%	42.7%
Mean Asian	44.5%	46.1%	44.6%	44.8%	36.7%

Figure 7.1: Percent of items with at least moderate DIF in the PISA 2000 Reading data. Average item difficulties compared between groups. Source: Grisay and Monseur 2007, p. 72.

Only a few studies investigated possible reasons for the existence of DIF (or other effects of a test translation). In a few studies, scientists identified “translation problems” with some, but not all DIF items, sometimes attributing the rest to possible cultural or curricular differences (e.g. Ercikan 1998). Approaching the matter more systematically by presenting DIF and control items to groups of translators, Allalouf et al. (1999, see also Sireci and Allalouf 2003) identified four possible sources of DIF in a multilingual Israeli university admission test (the Psychometric Entrance Test (PET)): 1) changes in word difficulty, 2) changes in item content (due to translation errors or multiple meanings of words), 3) changes to the format of the item (longer sentence completions, different configurations to the stem and answer options in a MCQ item) and 4) aspects of cultural relevance. Interestingly, the sources for DIF in seven reading comprehension items were identified as either *cultural relevance* or *unknown* (Allalouf, Hambleton, and Sireci 1999, p. 194), while other item groups (analogies, logic) were, according to the review team, more clearly related to specific problems of the translation. It is unclear how “cultural relevance” was defined exactly, and whether there was any empirical evidence of cultural differences beyond the judgment of the reviewers. On a side note, the authors mention that the higher word count in Russian could not be identified as a source for DIF.

Elosua and López-Jauregui (2007) also identified four possible sources of DIF in a similar study to the one by Allalouf et al. though focussing on a general cognitive test: 1) genuine translation errors which could have been avoided prior to testing, 2) grammatical (i.e. syntactical and morphological) differences between the languages involved, which led to changes in the form of the item, 3) a lack of semantic equivalence, which refers to differences between translated words in terms of “familiarity, concretion, or emotional valence” (ibid., p. 47) and to polysemous words, and 4) cultural relevance. In this study, some items were considered to be influenced by more than one of these sources, which, though fairly obvious, is an important change compared to the two earlier studies discussed above. Also, interestingly, the authors did not identify DIF items which were thought to be affected by cultural aspects, possibly given the bilingual context of their study.

Though often plausible, the post-hoc identification of possible sources for DIF by expert committees does not constitute empirical evidence that students were actually influenced by these aspects in their answering processes. Ercikan and colleagues (2010) tackled this problem by conducting think-aloud interviews with 7th-graders in British Columbia using 20 science and mathematics items which had been identified as showing DIF due to linguistic reason related to the translation of the assessment into French or English¹⁰⁵. Their analyses confirmed the sources of DIF identified by expert committees in half the items under investigation. In one additional item, the interviews clearly revealed a source of DIF that the experts had not identified, while the experts’ explanation for the remaining 9 items’ DIF could not be confirmed. Considering the fact that expert committees usually cannot determine the sources of DIF for a substantial number of items in the first place, this study shows the low reliability of expert committees identifying sources of DIF post-hoc (see also e.g. Gierl and Khaliq 2001), and it reinforces the importance of more (qualitative) research on how test takers perceive and solve assessment tasks (e.g. Barras 2018, 2021).

In a rare mixed-method study, García Duncan and colleagues (2005) were able to show the influence that different learning experiences had on test scores: though they also identified some minor translation problems with their Spanish translations of mathematics items, their quantitative analysis of test scores, focus interviews and think-aloud sessions revealed that maths skill and English language proficiency were clearly correlated. They could also show that the Hispanic heritage speakers employed construct-irrelevant test wiseness strategies more often than their English-speaking peers, which implies that 1) they did not profit from the English-instruction maths class in the same way and 2) that the assessment was not able to measure the underlying cause of this problem sufficiently well¹⁰⁶ (see also Oliveri and Ercikan 2011, p. 351).

Finally, an unusual study interested in translated mathematics items used in PISA 2012 was conducted by Bergqvist and colleagues (2018). Methodologically, this study is closer to the item difficulty studies presented in chapter 8, and, in fact, quite close to the study presented in this thesis: the authors analysed the relationship between the difficulty of items (using the *p* value) and seven linguistic variables in three language version (Swedish, German and English) of the PISA 2012 maths assessment. They measured word length (in four different ways as suggested by the literature), sentence length (number of words and other elements, including formulae), task length (number of words) and information density (the ratio between the number of nouns and verbs in a sentence). In addition, using a PCA of both the mathematics and the reading tasks administered in PISA 2012, the authors first identified the difficulty in *reading* associated with each mathematics item (ibid., pp. 47 sq.) and then analysed the link between this portion of the items’ difficulty and the linguistic variables. Both the correlation analyses and their regression

¹⁰⁵The items were part of a Canadian national assessment, the School Achievement Indicators Program (SAIP), which assessed competences of 13-year-olds in mathematics and sciences and had been developed in either French or English and then translated into the respective other language.

¹⁰⁶The same context was investigated in a thorough statistical analysis by Sireci and Khaliq (2003). Their results suggest that weak maths students in the English-only assessment performed similarly to the Spanish-speaking test takers. This might also hint at an underlying problem of the assessment itself.

analyses, conducted separately for each language, suggest a stronger relationship between the linguistic characteristics of the item in German than in English and Swedish, particularly regarding word length and information density (though see Bergqvist, Theens, and Österholm 2018, p. 52 for a critical discussion of this result, especially the hypothesis that linguistic variables may not mean the same in different languages).

7.5 Conclusion

The translation of measurement instruments has been practised for decades. Over time, sophisticated methods of test translation have been developed, but it has become evident that there can be no translation that does not need any evaluation of quality. Instead, researchers mostly agree that the use of a translated version of an assessment should be validated much in the same way as any other new instrument (ElAtia 2013, pp. 199 sq.). Overall, though there has been a substantive amount of research into the translation of measurement instruments in a variety of fields, it has become clear that little is known about the detailed functioning of test items. Analyses of translated educational assessments have repeatedly shown that verbal items in particular show DIF, though it is not always easily identifiable why. This is consistent across test taker groups, languages and cultures, and it is where research into item difficulty, quantitative, but also qualitative, comes into play: if the inner workings of test items can be described in more detail, perhaps the reasons for unexpected effects of translation may also become more evident.

Chapter 8

Research on item difficulty

Most linguists will not be surprised to hear that language processing is complex!

Buck and Kikumi Tatsuoka 1998, p. 149

Explaining item difficulty has been an objective in assessment research for several decades, often linked to achieving a better understanding of a test (or test format), its underlying construct or both (Neumann 2012, pp. 27 sq., 257). In an early publication on the subject, Drum, Calfee and Cook (1981) examined “surface variables common to many types of reading tests that are useful in identifying important components of comprehension” (ibid., p. 487). Their objective was to better explain how students make their choices in multiple-choice comprehension tests, looking for more insight into both L1 reading comprehension and the functioning of standardised tests. A similar study was conducted by Davey (1988), though this author separated the test takers into “successful and unsuccessful readers” to investigate whether reading ability influences the way readers perceive test items. A few years later, researchers at ETS studied the structure of item difficulty as part of their validation efforts for several tests developed at their institution, including its English L2 assessment TOEFL (Freedle and Fellbaum 1987; Freedle and Kostin 1991, 1993, 1999; Kostin 2004; Nissan, DeVincenzi, and Tang 1996).

More recently, Carr (2006) also studied TOEFL items, though he focussed on the “relationships among the characteristics of language tasks” (ibid., p. 270) instead of trying to predict empirical item difficulty. His study was meant to devise new ways of characterising test items which would include empirical predictions of test taker performance. In a similar vein, Beinborn et al. (2014, 2015) approached foreign language cloze and C-tests from an NLP perspective and tried to automatically predict the difficulty of these tests based purely on their linguistic characteristics.

Looking at the matter from a cognitive psychology perspective, Embretson and Wetzel (1987) explicitly tried to model item difficulty in order to select and eventually develop items according to their cognitive characteristics. Over the years, several studies followed in their footsteps, replicating parts of their findings and adding new methodological and theoretical approaches (Gorin and Embretson 2006; Ozuru et al. 2008; Svetina, Gorin, and K. K. Tatsuoka 2011; Toyama 2018). Buck and colleagues (1998; 1997) also studied language assessments from this perspective, examining the knowledge, skills, abilities and strategies involved in second language listening and reading comprehension. Their aim was to identify groups of test takers who use similar sets of sub-skills defined in the construct (using a method called Cognitive Diagnostic Assessment (CDA), K. K. Tatsuoka 2009).

Similarly, researchers interested in reading comprehension have investigated the effect of test items in order to determine what exactly their test results were telling them about the cognitive skills underlying reading (e.g. Keenan, Betjemann, and Olson 2008, see also Melby-Lervåg and Lervåg 2014, pp. 413 sq.).

Explicitly interested in aspects of language assessments which could be manipulated during test development, Brindley and Slatyer (2002) reported on a rare experimental study which investigated the effect of five specific variables on item difficulty which were manipulated *before* test administration.

Working on a German L1 assessment, Neumann (2012) as well as Böhme et al. (2010) added analyses of L1 listening and reading comprehension items used in a school context. Neumann also investigated to what extent stakeholders (item developers and teachers) were able to predict empirical item difficulty (see also I.-C. Choi and Moon 2019), while Böhme and Robitzsch (2009) proceeded to investigate more refined analytical methods useful for this line of research. The same is true for several recent studies undertaken by Aryadoust and colleagues (Aryadoust 2015, 2018; Aryadoust and C. C. M. Goh 2014), who have used and compared several modern approaches to map the underlying structure of both test taker and task characteristics.

Finally, test specialists from other fields (e.g. the natural sciences) also study item difficulty in their content exams, regularly finding language to be an important aspect and trying to limit its impact through manipulations of the tests' linguistic complexity (e.g. Abedi and C. Lord 2001; Stiller et al. 2016, cf. also an overview by Heine et al. 2018).

In the following, a selection of these studies will be discussed in more detail, focusing on the research designs and statistical methods, the variables and the results¹⁰⁷.

8.1 Research designs and statistical methods

Given the various disciplines interested in item difficulty that were mentioned above, particularly cognitive psychology and language assessment research, it is not surprising that there are various methodological approaches to the problem. Apart from qualitative approaches to the problem, which will not be discussed here, three main quantitative approaches can be identified: correlations and regressions, CDA and more exploratory methods, ranging from factor analyses to machine learning. These will be discussed in more detail in the following sections.

8.1.1 Correlation, regression modelling and related methods

The simplest approach to studying item difficulty statistically is to establish a direct link between individual task or item characteristics and empirical item difficulty. An example of this is the study by Nissan, DeVincenzi and Tang (1996) on a TOEFL listening comprehension format. Trying to identify task characteristics which task developers can use to influence item difficulty, Nissan et al. analysed the impact of 17 individual variables on *Delta*¹⁰⁸ using *linear regression* for two continuous variables and *t-tests* or *ANOVAs* for categorical variables. They also studied the interactions of three individually significant variables using ANOVAs, but did not create any more complex models (ibid., p. 19). Similarly, Brunfaut and Révész (2015), also interested in L2 listening but using a small data sample, analysed the impact of various task and test taker variables on item difficulty using *Spearman rank-order correlations* for each individual variable (see also Révész and Brunfaut 2013, who used simple linear regression in a way reminiscent of Pearson correlations). In both studies, the authors reported on significant and insignificant variables and discussed their findings in some detail.

Brindley and Slatyer (2002) went a different way, but also assumed a direct relationship between item difficulty and various task characteristics: in their experimental study, they created different versions of the same two tasks and administered them to test takers along with a common unchanged task. They then *compared the item difficulties* of the tasks in their various conditions

¹⁰⁷Note that not all of the studies mentioned in this introduction will be discussed in all following sections, in order to avoid unnecessary repetitions and because they are not always relevant to the discussion at hand.

¹⁰⁸*Delta* or Δ is a sample-independent measure of item difficulty derived from *proportion correct p* used by ETS in the past (Holland and Thayer 1985, cf. also section 6.2.1.1).

using one common Rasch model over all items¹⁰⁹. Where item difficulties differed visibly, the authors assumed an effect of their task variation.

The large majority of studies on item difficulty used some form of *regression modelling* to account for the impact of multiple variables and their interactions on empirical item difficulties. Drum et al. (1981) identified 52 initial variables based on the literature and then selected 16 of those based on collinearity checks and preliminary regression analyses. Using the *p*-values (proportion correct) of 210 items from seven different standardised reading assessments used in various grades (18 data sets), they conducted *multiple stepwise regression* analyses and reported the variance explained (R^2) and the beta value (β)¹¹⁰ based on 10 variables which had proved to be significant in most data sets. Davey (1988) studied the impact of 20 variables on reading item difficulty for two different groups of test takers: successful and unsuccessful readers (using *p*, which is sample dependent and differed slightly between the two groups). She described a *three-fold conditional regression approach*, where she first conducted separate stepwise regression analyses for each variable group (passage, question type or item format) to select effective predictors and then used those variables to model item difficulty for each group of readers, identifying the same two variables in the final model. After re-admitting all other variables into these two models and finding no further effective predictors, she compared whether the regression parameters differed between the groups.

Very similar designs were chosen by Freedle and Kostin (1993, 1996, 1999) in their various studies of TOEFL reading and listening items, though they considered a much larger number of variables: their analyses contained up to 62 (reading) or 60 (listening) variables, which were retained from even longer theory-based lists after checking for collinearity and interactions with the predefined item type (using MANOVAs, cf. Freedle and Kostin 1993, pp. 23 sq., 1996, p. 19). The authors conducted correlation analyses as well as stepwise regression on both nested and non-nested item samples¹¹¹ with *Delta* as the dependent variable. In the regression models, only predictor variables which showed a significant *t* value were retained. The authors reported on all variables which correlated significantly with item difficulty and presented the final regression models along with their R^2 . They also cross-validated their analyses by using different (or sub)sets of items to run new regression analyses and compare the variables retained or by predicting the item difficulties of one set of items with the model estimates identified by another. This often yielded mixed results, although they were usually better when non-nested samples of items were used.

Using a new set of predictor variables to better account for the specificities of a TOEFL listening comprehension format, Kostin (2004) proceeded in much the same way, though she used a slightly different stepwise regression procedure and added a *jackknife procedure* to cross-validate her findings: the 365 items used in the analysis were divided into 10 groups of (approximately) equal size and equal difficulty. For each group, the item difficulties were predicted using regression parameters based on an analysis of the nine other groups. The predicted item difficulties from each run were then correlated with the observed item difficulties to see how well regression could

¹⁰⁹This procedure makes a direct comparison of item difficulties possible (cf. Andrich and Marais 2019, p. 92), but it does not take into account that certain items are more closely related to each other (i.e. the *same* item in the various conditions) than others. This violates the condition of local independence in Rasch modelling (cf. *ibid.*, p. 93), although it is somewhat unclear if this actually invalidates the general tendencies found by the authors. After all, Rasch modelling is often used on items from the same testlet in violation of the same condition and whether or not two items, which elicit the same information under different conditions, are more closely related than items from a testlet is open to discussion.

¹¹⁰ β is the standardised regression coefficient. Its value indicates how much the outcome variable (e.g. item difficulty) changes in the regression model when a predictor variable (e.g. length of input text) is changed by one standard deviation. Thus, a β value which differs strongly from 0 is an indication of a strong relationship between a predictor variable and the outcome variable.

¹¹¹In a nested sample, several items may refer to the same input text whereas a non-nested sample would be composed of as many input texts as items with each input text related to only one item. Since text variables are identical for each item related to the same text, using a non-nested sample makes it more likely to identify text variables which significantly predict item difficulty (Freedle and Kostin 1996, p. 12).

be used to predict the difficulty of “new” items (cf. Kostin 2004, p. 25).

Neumann, interested in the listening part of a German large-scale assessment of receptive competences in the language of schooling (the so-called IQB-tasks, cf. Neumann 2012, pp. 133 sqq. or Institut zur Qualitätsentwicklung im Bildungswesen 2020), reports on correlations of various task and item characteristics¹¹² with empirical item difficulties, means across several items or raw scores. In addition to that, she conducted *factor analyses* on the various groups of task or item characteristics in order to describe the way in which they appear together (or not) (cf. also Carr 2006), and studied partial correlations of several inter-related characteristics with item difficulty (Neumann 2012, pp. 197 sqq.). In her regression analyses, she used a three-fold approach: she estimated separate multiple stepwise regression models for a) all variables which correlated most strongly with item difficulty, b) significant variables which had loaded onto common factors in her factor analyses and c) significant variables which were thematically related. In all cases, she eliminated all variables which did not add significantly to the variance explained. Those variables which remained in all three approaches were then used to create two final regression models—one on the task level and one on the item level.

Working on the same large-scale assessment, Böhme, Robitzsch and Busè (2010) used *multiple linear regression* to investigate the impact of 20 predictor variables on the empirical difficulty of reading and listening items (obtained from two Rasch models, one for each skill). To account for collinearity, instead of eliminating strongly related variables through stepwise regression, they calculated *lmq* (ibid., pp. 93 sqq.), a measure of “the proportion of the R^2 contributed by each individual predictor” (Zhang and Wang 2017, p. 110). Rather unusual, but apparently promising, is the use of a so-called *commonality analysis*, which identifies the *unique* contribution of certain predictors to R^2 (Böhme, Robitzsch, and Busè 2010, pp. 94 sq.; see also Nimon et al. 2008).

Regression modelling, though, no matter how it is conducted, has certain drawbacks. First, as Bachman points out (2002, p. 468, cf. also 6.1), using empirical item difficulty as the independent variable is problematic because it is mathematically derived from test results, which are, in turn, influenced by item difficulty. Drawing on Bachman’s argument, Carr adds to it from the opposite direction: “it seems likely that task characteristics will be able to predict overall examinee performance in some cases [...] but not in others, until such time as the characteristics of individual examinees can be included in the modeling” (2006, p. 285). Thus, regression modelling which investigates the link between empirical item difficulty and task characteristics is necessarily imprecise: the two aspects are possibly autocorrelated and there are other variables at play. Second, Buck and Tatsuoaka argue that empirical item difficulty is not actually the variable of interest, since “we are generally far more interested in understanding person performance than item difficulty” (1998, p. 126). Although this point is certainly arguable from the perspective of a test developer, relating task characteristics to a derivative of individual test takers’ task performance (i.e. item difficulty) can certainly be considered a contradiction. Third, the regression analyses described above often resulted in R^2 values below .50, leaving more than half of the variance in the empirical item difficulties unexplained¹¹³. Buck and Tatsuoaka point out that this may be due to the fact that many predictors are established to predict the difficulty of a very limited number of items, which leaves “researchers constantly in danger of finding chance results, and [thus] restricted in the analyses they can perform” (ibid., p. 126). Another reason may be that all of the above studies had to assume a linear relationship between the predictor variable and item difficulty and a normal distribution of both, neither of which is necessarily a given (cf. Aryadoust and C. C. M. Goh 2014, p. 7; Aryadoust and C. Goh 2019, p. 210).

¹¹²Neumann conducted most of her analyses separately on the level of the stimuli (the input text which is related to several items) and on the level of the individual items.

¹¹³Drum et al. (1981, pp. 501 sq.) are a notable exception, reporting R^2 values between .55 and .94 depending on the test and grade (mean over all samples: .71). Freedle and Kostin (1993, p. 135) point out, however, that in this early study, too many predictors were fitted to too few items, inflating R^2 out of proportion.

8.1.2 Factor analysis

As a result of these shortcomings, some researchers have adopted other approaches to study the underlying structure of test tasks and items. For instance, Carr (2006) eliminated the role of empirical item difficulty (almost) completely and looked at the configuration of a large number of task and item characteristics instead. First, he coded 15 reading passages, their related items, as well as key sentences for each item according to a 33-page-list of characteristics. He then performed *exploratory and confirmatory factor analyses* on these characteristics. This method aims to describe the associations of the text and item characteristics *with each other* and has also been used as a data exploration tool by Neumann (Neumann 2012, pp. 169 sqq.), as mentioned above. In a second phase, Carr added empirical item parameters (difficulty, discrimination and a guessing parameter from a 3PL model) to the model, which is akin to predicting these parameters solely by the task characteristics.

8.1.3 Cognitive Diagnostic Assessment

Researchers with their conceptual roots in cognitive psychology have also analysed the inner workings of test items, using CDA models. Not driven by the need to establish a one-dimensional hierarchy of test takers, they have been working on analysing test performance to determine the sub-skills needed to solve the various test items. This is, in many ways, similar to the search for item characteristics which can predict item difficulty, but considered from the perspective of the *test takers'* knowledge, skills and abilities. At the heart of most CDA models is the so-called *Q-matrix*, which lists the items of a test and the sub-skills defined in the test construct, and determines for each item whether each sub-skill is targeted (1) or not (0) (cf. for more technical information Aryadoust 2012, p. 85).

In the past several years, a variety of different CDA models have been applied to language assessment (Aryadoust 2021, p. 33). The “earliest statistical model in CDA” (Jang 2009, p. 35) is LLTM (Fischer 1973), which was prominently used by Embretson and Wetzel (1987). They combined linear regression and IRT, investigating various *LLTMs*. Starting from a “simple” Rasch model, Embretson and Wetzel added variables to the linear prediction of b , the model’s item difficulty¹¹⁴. They then examined the goodness of fit of each model using three indicators: log likelihood, the chi-square goodness of fit as compared to a null model and an incremental fit index which indicates the amount of information predicted by the model (ibid., p. 181). Some models were based on the findings of earlier research, including the study by Drum et al. (1981) described above, others were theoretically derived. Combining these several models led the authors to a proposed final model, which contained two groups of variables: “six text representation variables and seven decision processing variables” (Embretson and Wetzel 1987, p. 188). The advantage of using LLTMs, according to the authors, may be lying in the interpretation: the interrelation of the two groups of variables of the final model and their influence on item difficulty can be studied for individual items, making it, theoretically, possible to devise items with a specific cognitive difficulty according to quantifiable rules.

More recently, but in the same tradition, Ozuru and colleagues (2008) used Hierarchical Linear Modelling (HLM) to account for the fact that studies of item difficulty usually combine item-level and passage-level variables in the same analysis even though they are essentially nested within each other: item-level variables influence an individual item’s difficulty along with passage variables which are common to more than one item. Based on their findings, Toyama (2019) applied an LR-LLTM to an analysis of item difficulty of an English L2 reading comprehension assessment and a related vocabulary assessment, including both item and person characteristics in the model.

Another early statistical model in CDA is the *Rule Space Method* (K. K. Tatsuoka 2009), which aims to identify the knowledge, skills and abilities needed to solve individual test items and

¹¹⁴See section 6.3 on explanatory item response modelling, which is akin to these approaches.

then classify each test taker according to whether or not they master each of those. This is, very broadly speaking, achieved by establishing the Q -matrix and putting it in relation with the test takers' total scores (in the form of IRT's θ) and their item response patterns mathematically. The result is a two-dimensional rule-space in which test takers with similar response patterns will be clustered (e.g. K. K. Tatsuoka 2009, p. 33, for somewhat more understandable descriptions of the process cf. Svetina, Gorin, and K. K. Tatsuoka 2011, pp. 4 sqq. or Davier, DiBello, and Yamamoto 2006, pp. 2 sq.). In their application of the *Rule Space Method* to language assessment, Buck and colleagues studied the cognitive structure of a listening (1998) and a reading assessment (1997) for EFL learners, while Svetina and colleagues (2011) studied an English L1 reading assessment. The authors chose item characteristics (which they called *attributes*) based on the literature and then eliminated those which did not seem to influence the item response patterns of the examinees. The results was a number of significant attributes as well as a large majority of the test takers classified in their "respective latent knowledge state" (Buck and Kikumi Tatsuoka 1998, p. 138).

A more recent approach, which builds on Tatsuoka's work on the rule space method is the *fusion model* (also known as the Reparameterized Unified Model (RUM), cf. Aryadoust 2021, p. 35). It is a form of IRT in the Bayesian framework (Aryadoust 2012; Davier, DiBello, and Yamamoto 2006). In the fusion model, the (non-)mastery of sub-skills in the Q -matrix as well as a non- Q competence is estimated as a probability for each test taker in a way similar to θ in IRT. The model also provides more information on item functioning, which can be used to examine the appropriateness of the Q -matrix and, thus, of the item variables used during coding (Sawaki, H.-J. Kim, and Gentile 2009). Sawaki et al. used the item information obtained by the fusion model to refine the definition and application of reading and listening sub-skills in TOEFL iBT reading and listening items (ibid., pp. 10 sq.). Aryadoust (2012), on the other hand, used the model to compare the appropriateness of three different Q -matrices and to narrow down the list of sub-skills which could be meaningfully applied to differentiate between ability groups.

A major drawback of CDA is the establishment of the Q -matrix: as the authors of the aforementioned studies all point out at some point, the interpretability of CDA modelling in the field of language assessment remains uncertain, especially concerning "coding procedures and attribute selection as well as attribute interpretation" (Svetina, Gorin, and K. K. Tatsuoka 2011, p. 20). As a result, a plethora of new CDA approaches have been developed lately, considering aspects such as the relationship between the various subskills under investigation. The applicability, validity and comparability of these models is currently under research (cf. a concise overview in Aryadoust 2021).

8.1.4 Regression trees and artificial neural networks

Beyond CDA, recent studies surrounding item difficulty or test taker sub-skills have generally focused on finding more appropriate methodological approaches in the hopes of overcoming the various issues mentioned above. Of particular interest is the study by Aryadoust and Goh (2014) who compared the results of three very different statistical approaches: linear regression, Classification and Regression Tree (CART) and two types of Artificial Neural Networks (ANNs). Item difficulty was a dichotomous variable of low- and high-difficulty items based on a Rasch analysis and split at the median (ibid., pp. 10 sqq.).

CART is a non-parametric data mining technique, where the data is split along a number of input variables one after another to achieve homogeneous subgroups of observations (creating a *decision tree*). This is done mathematically using one of several partitioning techniques, so-called tree-growing rules that are used to determine which variable splits the data best at any given moment during tree-growing. Since it is usually possible to envision more than one possible tree and since non-parametric statistical methods cannot check for statistical significance, CART also comprises methods of cross-validation and is based on checks of model fit (cf. for more concise information on the various techniques Aryadoust and C. Goh 2019). In their CART

study, Aryadoust and Goh (2019; 2014) used 12 input variables for tree-growing, all of which ended up in their final tree. The authors reported on various performance indices of the tree as well as the so-called *variable importance index* (*VII*), which indicates the importance of each input variable in the model.

Rupp et al. (2001) also examined the performance of the CART method as compared to a multiple regression, studying the impact of 12 variables on proportion correct (p). Other than Aryadoust and Goh (2014, p. 10), who studied linguistic characteristics of the items and input texts, Rupp et al. included variables related to the items and text-by-item interactions. Based on the structure of six trees and the best-fitting regression model, they indicated for each variable whether it was found to be, on the one hand, significant in the regression model and, on the other hand, “important” according to the relative ordering of the variables in the various trees (Andre A. Rupp, P. Garcia, and Jamieson 2001, p. 210). Gao and Rogers (2011) also used tree-based regression in their analysis of two separate forms of the Michigan English Language Assessment Battery (MELAB) reading test, using the item difficulty parameter obtained from a 3-PL IRT testlet model. More closely related to CDA, Gao and Rogers first established and validated a cognitive model of test taker processes, which led them to a detailed description of the cognitive processes each item required according to the established tree (*ibid.*, pp. 93 sqq.). A simpler approach was adopted by Sheehan and colleagues (2006), who used tree-based regression to examine whether their supposed comprehension processes and the semantic relatedness of the items and the passage could be used to explain item difficulty.

ANNs are the result of machine learning¹¹⁵: instead of linking the values of independent variables to the values of a dependent variable through one single equation, neural networks process information from input variables through a series of processing units. These so-called neurons are interconnected with each other by means of activation functions, mathematical rules which transform the data from one neuron to another. In addition, neurons have an associated weight, which determines whether and to what extent their information is passed on in the network (Perkins, Gupta, and Tammana 1995, p. 36). The weight, distribution and number of neurons, the activation functions and the overall structure of the network are determined through “trial and error” using one part of a data set for training and the rest for testing. ANNs do not presuppose any kind of distribution of the variables and they are rather flexible concerning the type of mathematical relationships between the variables. Hence, they allow for highly complex relationships between variables and can theoretically describe any type of decision given enough input variables, layers, training data and computing power.

In language assessment, ANNs have not been used much, although Perkins et al. (*ibid.*) demonstrated their predictive power more than two decades ago. Based on the test results of 70 students from various linguistic backgrounds (*ibid.*, p. 39), their use of ANN modelling achieved a reasonably good prediction of the items’ empirical difficulties, with correlations between the predicted and the empirical values above $r = .83$ (*ibid.*, p. 44). At the same time, a linear regression analysis of the same data failed because the data set was too small (*ibid.*, p. 40).

As mentioned above, Aryadoust and Goh (2014) also examined and compared the performance and results of linear regression, CART and ANNs. They studied the relationship between test taker performances and the linguistic characteristics (*ibid.*, p. 7, see also below, section 8.2.4) of 242 items from the listening section of the Michigan English Test (MET) test, an English L2 test for “work and academic situations” (Michigan Language Assessment 2020). Of the techniques chosen in their study, one ANN model (a Multilayer Perceptron (MLP) ANN model with one hidden layer) proved to be most in line with the underlying theory and best in predicting item difficulty¹¹⁶. The CART technique, although less precise in determining the variables which contributed to item difficulty than the ANN model, also turned out to have promising features,

¹¹⁵The actual technical details of machine learning far exceed the purposes of this text. Reasonably accessible descriptions of ANNs, their specific terminology and possible applications can be found on-line. For more technical information cf. the references in Aryadoust and C. C. M. Goh 2014; Perkins, Gupta, and Tammana 1995.

especially with regards to the “validation method of the accuracy of the model” (Aryadoust and C. C. M. Goh 2014, p. 26). The *worst* modelling technique in the study turned out to be, consistent with the criticism mentioned above, linear regression. An important reason in this study seems to be the “less than ideal [data]” (ibid., p. 25), especially the relatively short input texts: since linear regression makes strict assumptions about the distribution and nature of the data, violations of this type may influence its results more heavily than in ANN or CART modelling (see also Aryadoust 2013).

8.1.5 Summary

Over the years, researchers have tried a variety of approaches to explore the inner workings of item difficulty. Some used mere correlations, others investigated how various statistical methods described the same data. In the latter, regression in particular has often proven to be somewhat limited in the way it can describe small data sets. Nevertheless, the large majority of item difficulty studies *has* used regression, or a variant of it, and produced interpretable results, often not dissimilar to the results of studies which applied other statistical techniques (see section 8.3). In all studies, however, an important aspect beyond the choice of the method is the selection and operationalisation of the variables under investigation: after all, any statistical method can only reveal effects that have been measured in the first place. In the following section, this topic is examined in more detail.

8.2 Variables in item difficulty research

Taken together, researchers interested in item difficulty used a large number of different item characteristics, operationalised in various ways, ranging from automatically determined corpus-linguistic measures to manual ratings done by members of the research team or other stakeholders (e.g. teachers, Neumann 2012). Based on theoretical models of text comprehension and decision-making, prior experimental research in several fields, and the findings of earlier similar studies, the authors chose anything from five (Brindley and Slatyer 2002) to 515 (Carr 2006) variables to analyse their impact on item difficulty or, in CDA studies, to explore the skills, knowledge and abilities test takers employ to solve tasks.

8.2.1 Sources of variables

Early studies such as Drum et al. (1981) or Embretson and Wetzel (1987), but also, to some extent, Freedle and colleagues (Freedle and Fellbaum 1987; Freedle and Kostin 1991, 1993) chose their variables based on two main sources. On the one hand, they used theoretical models of reading comprehension and/or decision-making. Most prominent is the C-I model (W. Kintsch and Dijk 1978, see also section 3.1.1), which laid the groundwork for propositional analysis (Embretson and Wetzel 1987, pp. 177 sq.). On the other hand, they drew from results of experimental research on comprehension, which often dealt with the comprehensibility of specific grammatical structures such as negations (e.g. Carpenter and Just 1975), or more content-oriented aspects, e.g. how the main idea of an expository text is established (e.g. Kieras 1985).

Research on linguistic complexity, readability and, more recently, NLP, has also had an impact on the selection of probable item and text variables. When computer-assisted linguistic analysis was still in its infancy, researchers used variables such as the “number [of] words with more than two syllables among [the] first 100 words” (Freedle and Kostin 1999, p. 27) as a measure of lexical complexity proposed in readability research. In more recent studies, corpus-linguistic measures

¹¹⁶The authors also used a second form of ANN modelling, Adaptive Neuro-Fuzzy Inference System (ANFIS), which had performed well in an earlier study by Aryadoust (2013). In the comparative study, however, ANFIS performed worse than the CART model, which underlines the uncertainty still inherent with this type of statistical modelling.

such as lexical diversity indices, proportions of more or less frequent vocabulary (e.g. Brunfaut and Révész 2015), but also, for instance, complexity measures derived from analytical tools like Coh-Metrix (Graesser, McNamara, Louwerse, et al. 2004, e.g. Aryadoust and C. C. M. Goh 2014) are much more wide-spread.

Many researchers also chose task or item characteristics based on the specific characteristics of the assessments they were studying. Brindley and Slatyer (2002), for instance, manipulated five task characteristics in their quasi-experimental study because they were considered important in the literature but did not figure in the test specifications of the assessment they studied. Nissan et al. (1996) used the opposite approach and investigated the impact of several task characteristics prescribed in the ETS and TOEFL item writer guidelines relevant for their study. Freedle and Kostin (1999) or Böhme et al. (2010) explicitly included a number of variables related to *listening* comprehension to study the comparability of reading and listening assessments.

A different approach to determining salient sources of item difficulty is to ask the stakeholders: Aryadoust (2021, pp. 38 sqq.) refined an initial *Q*-matrix based on observations from retrospective think-aloud interviews used alongside an eye-tracking study. Much earlier, Buck (1991) laid the groundwork for the item difficulty analyses in his and his colleagues' later work (1998; 1997) using verbal reports with six Japanese students of English. Gao and Rogers (2011) did the opposite: they first established a theoretical cognitive model based on a review of the literature to explain the reading performance of Mandarin-speaking test takers on the MELAB. This model was then presented to “judges familiar with L2 reading and Mandarin-speaking students who take the MELAB tests” (ibid., p. 82). In a final step, it was further revised based on verbal reports (think-aloud and retrospective) collected from 16 university students who shared the L1 of the test taker group studied.

Finally, researchers have worked on replicating the findings of others by choosing variables which showed promise in earlier, similar studies.

8.2.2 Text, item and text-by-item variables

As much as reading comprehension hinges on the nature of the text, it has always been clear that solving a reading test item implies challenges beyond the mere lecture of the text. In terms of test validity, this poses a problem: theoretically, given that a reading comprehension assessment is supposed to determine whether a test taker is able to read a certain type of text, the difficulty of the assessment should only be influenced by characteristics of the input text¹¹⁷. Freedle and Kostin (1993, p. 134) describe this issue as follows:

Multiple-choice tests of reading comprehension [are] intended to measure a student's ability to comprehend a particular reading passage. But, [...] in order to assess passage comprehension, one must first pose a problem for the student to solve in terms of a reading item [...]. The solution to the problem typically involves the selection of one option from a set of response options. The question is whether the complexity of posing the problem to be solved has itself become more important than the passage to be comprehended. That is, is the content and structure of the item itself a better predictor of reading item difficulty than the content and structure of the passage? If so, then one cannot claim that the test is construct valid since the intended purpose of such tests is to evaluate whether the passage itself has been comprehended.

Freedle and Kostin then go on to argue that in order to show the construct validity of a multiple-choice test, there must be a relationship between the input text's characteristics and item dif-

¹¹⁷It is also possible to conceive an assessment where the process of finding a certain information is part of the construct (e.g. in an assessment of academic language proficiency where extracting certain information from a text is a main purpose of reading). In this case, however, a closed-answer item format such as multiple-choice is probably not the preferred choice. Instead, note-taking or some sort of summarization would be more closely related to the intended construct (cf. for an example Sabatini et al. 2014 and related publications). If a closed-answer format was used, the variables described in the following would probably still apply.

difficulty, whereas the influence of purely item-related characteristics must be negligible (1993, pp. 134, 142).

The authors also postulate the existence of so-called text-by-item variables, which can only be determined by considering *both* the item and the input text (ibid., p. 143). Examples of this category are variables which describe the linguistic characteristics of the key sentence (i.e. the passage of the input text which contains an item's solution), and overlap variables, which measure the linguistic material shared by the items and the input text. Interestingly, the authors did not consider this third category of variables a danger to construct validity, arguing that they imply comprehension of the input text¹¹⁸. In a way, this invalidates parts of the argument put forth above: if text-by-item variables are construct-relevant, then solving the problem must be part of the construct. If, in addition, those variables are good predictors of item difficulty, then the way an item is phrased *does* influence test results.

Despite this questionable definition of construct-relevance, the distinction of text, item and text-by-item variables for item difficulty research has its advantages. It is conceptually convincing and reinforces the importance of the item or task in an assessment situation. The variables chosen for the empirical analysis in this thesis will follow this structure. For the following review, however, a simpler distinction was chosen: first, the variables which operationalise the *language* of the input text and/or the items are discussed. Many of these may be either text or item variables, or both, depending on where they are applied. Second, variables which are unique to reading assessments are examined. Following Freedle and Kostin, these are often text-by-item variables, though some item variables also fall into this second category, and some, of course, were not considered by Freedle and colleagues (1987; 1991, 1993, 1996, 1999).

Before these groups of variables are presented, however, the concept of the NI needs to be introduced. It replaces the *text* in Freedle and colleagues' original taxonomy to some extent, and its characteristics have sometimes proven to be better predictors of item difficulty than those of the entire input text. Hence, in this thesis, it is used alongside the input text to investigate the sources of item difficulty.

8.2.3 The Necessary Information

The NI, also called *relevant information* or *key (text) sentence*, has been used in various studies to define the section of the input text which needs to be understood in order to be certain of the correct answer (Buck, Kumi Tatsuoka, and Kostin 1997, p. 437; Carr 2006, p. 271; Freedle and Kostin 1993). In some studies, the NI is considered to be one sentence (Freedle and Kostin 1993), in others, it is allowed to be more than one, though adjacent sentences (Carr n.d., p. 18). Without specifically referring to the term, Aryadoust used a variable he called "word count", which is, in fact, the length of the NI: "the number of words in the oral text that are necessary for answering the test item" (2013, p. 45). Embretson and Wetzel used a similar variable, though they counted the number of propositions relevant to finding the correct answer (1987, p. 180). What exactly amounts as the NI seems to be a matter of manual coding (Buck, Kumi Tatsuoka, and Kostin 1997, p. 438; Aryadoust 2013, p. 45).

Some researchers operationalised the linguistic features of the NI in a similar manner as the input text itself (e.g. Carr 2006, n.d., see below, section 8.2.4). Some also investigated the position of the NI in the input text (see below, section 8.2.5.2), its distribution and its redundancy: Freedle and Kostin (1996, p. 56) assessed whether a relevant information could be found more than once in an input text, and if yes, whether it was sufficient to process one instance, or if two (or more) were needed. Carr used a 5-point scale which required raters to indicate whether the item could be solved without reading the passage at all up to understanding the entire passage (Carr n.d., p. 22). Buck and colleagues (1998; 1997) rated whether the NI was easy to locate (i.e. explicitly

¹¹⁸In light of the results of their study, this is not surprising: about half of the variables which correlate with item difficulty in their study are text-by-item variables, including those which correlate the strongest (Freedle and Kostin 1993, p. 157). The same is true for the results of the regression models (ibid., p. 162).

stated or marked, see also Davey 1988, p. 75), scattered in the input text, not in the order of the items or how the NIs of adjacent items related to each other (Buck and Kikumi Tatsuoka 1998, p. 135). Similar variables were chosen by Böhme and colleagues in their analysis of listening vs. reading comprehension item difficulty (2010).

Variables with regard to the NI have been shown to predict item difficulty better than variables which include the entire input text (*ibid.*, p. 82). Buck and Tatsuoka added a consideration of the text surrounding the NI and found that this, too, had a strong link with item difficulty in a listening assessment (1998, p. 134, see also Brindley and Slatyer 2002). The concept of a necessary information is also supported by the results of eye-tracking studies such as Krstić and colleagues (2018): their data of 8th graders solving PISA items in their L1 clearly show that students rarely read the entire text and instead focus only on the part of the text which is relevant to answering the question. According to the authors, this is particularly true for high-proficiency readers.

8.2.4 Linguistic variables

Drum and colleagues (1981) conducted a very early study of item difficulty in seven standardised English L1 reading tests used in the US. Their linguistic variables were mostly counts: of words and words with more than one syllable, of syllables, of content words¹¹⁹, of sentences, of medium-long and long sentences, of sentences with few, some or many content words, and of subordinate clauses (*ibid.*, pp. 513 sq.). The authors also created three variables which represented the amount of information and redundancy in a text: in essence, these were log-transformations of the number of content words (*ibid.*, pp. 513, 494 sq.). To assess the “vocabulary familiarity” of the texts and items, words which did not appear on the Dale-Chall list of 3,000 most frequent words were counted (*ibid.*, p. 495). The input text’s overall content (academic or not), and the text genre were also considered as categorical variables¹²⁰.

Davey (1988) used linguistic variables similar to Drum et al., in particular measures of the length of the input text, stem and answer options, the number and proportion of multisyllabic words as an index for lexical complexity and sentence length and mean number of clauses per sentence as proxies for syntactic complexity. She also added a measure of coherence, a dichotomous variable of whether “explicit connectives and clear pronoun referents” were present in the input text.

Freedle and colleagues (Freedle and Fellbaum 1987; Freedle and Kostin 1991, 1993, 1996) analysed several standardised comprehension tests (reading and listening, English L1 and L2), and established a considerable number of detailed linguistic variables. In addition to the various counts (e.g. sentence length, number of paragraphs, number of words in stem, correct and incorrect option), they considered specific linguistic structures for their analyses: based on experimental research, they counted the occurrence of “different text ‘frontings’”¹²¹, rhetorical text questions, special punctuations, referentials and negations (Freedle and Kostin 1993, pp. 148 sq., 1996, pp. 3 sq.). In their studies, the text genre was operationalised as a type of rhetorical organisation based on Grimes (1975) and Meyer (1985) (argument, list/describe, cause, compare,

¹¹⁹ *Content words* (German: *Inhaltswort*) generally include nouns, verbs, adjectives and adverbs, sometimes with exceptions such as *to be* or auxiliaries. Other parts of speech, such as articles, prepositions, conjunctions and pronouns, which combine the content words to generate meaning, are usually called *function words*. In addition, based on Clark (1976), Drum et al. (1981, p. 494) define *content-function words*: “[they] have the properties of both content and function words; like the function words, they are important in syntax. The italicized words in the following passage sentence are examples of content-function words. ‘She was *so* tall and strong that she was *almost* a giant’ (CTBS/Q2, Passage 4)”.

¹²⁰ Note that the actual item difficulty analyses in this study were conducted only with a selection of those variables, identified via collinearity analyses (*ibid.*, pp. 495 sq.).

¹²¹ “‘fronted’ structures at sentence beginnings and paragraph beginnings” (Freedle and Kostin 1993, pp. 138 sq.) are rhetorical structures which are put in front of the main subject and predicate of a sentence, e.g. cleft structures (“*It is true* that [...]”), co-ordinators (“*But*,” “*And*,”) or structures which mark a topic (e.g. “*Quickly*, *near the lodge*, the boat overturned.”) (*ibid.*, pp. 138 sq., 148 sq.). See also Carr n.d., pp. 4 sqq.

problem/solution), whereas the topic of the input text was measured in terms of the scientific area (physical science, life science, social science etc.) or, in the listening items, a selection of nonacademic topics (e.g. extracurricular activities, university procedures). An additional variable represented a judgment of the abstractness of the topic (1 = abstract to 5 = concrete) (Freedle and Kostin 1993, pp. 146 sqq.).

Interestingly, Freedle and colleagues barely assessed vocabulary, using only the number of syllables as a rough estimate of lexical complexity: “number of words with three or more syllables for the first one hundred words of the passage”¹²². In a later study of TOEFL item difficulty, however, Kostin (2004) expanded on this, and added lexical variables based on work by Nissan, DeVincenzi and Tang (1996), taking into account whether understanding of an infrequent word was necessary to find the correct answer or not. She also used a rating variable to assess “whether or not comprehension of an idiom [...] was relevant to responding correctly to the item” (Kostin 2004, pp. 32 sqq.).

Perhaps the most in-depth list of linguistic variables in item difficulty research has been collected by Carr (n.d., see also Carr 2006) in his factor analysis of text and item features in reading comprehension sections of TOEFL. His “rating instrument” distinguishes between passage variables, key sentence variables and item variables (further separated into stem, correct option and incorrect options). For all three elements, Carr elaborated on the Freedle and Kostin studies and included fronted structures of various types, referentials (resulting in cohesion variables) and negations as well as a long list of “other language-related variables” which assess syntax, morphology and length in detail (e.g. counting words and sentences, but also participles, gerunds, noun, adjective or subjunctive clauses, predicates, propositional clauses and other structures, per passage and per 100 words). In addition to those relatively objective counted variables, he added variables which were rated by humans to operationalise the (perceived) grammatical and rhetorical complexity of the passage and key sentence, the concreteness and content specificity of the passage, and the range of different speech acts, language functions and registers used in the passage. Passage variables also assessed the text genre (expository, narrative, argumentative, literary, entertainment), rhetorical features, and subject matter. Finally, vocabulary variables included measures of lexical variety and ratios of frequent vs. infrequent vocabulary based on word lists of the most frequent 1000, 2000 and academic words (ibid., p. 274). In addition, the instrument includes human ratings of the amount of specialised vocabulary in the passage, key text sentence and item (“such as slang or technical terms” (Carr n.d., pp. 11, 19, 22)).

A different approach was chosen by Embretson and Wetzel (1987): based on research related to Kintsch and van Dijk’s work on the C-I reading comprehension model (Dijk and W. Kintsch 1983; W. Kintsch and Dijk 1978), they expected that item difficulty would be more related to the number and density of propositions in the input text rather than the number of words (Embretson and Wetzel 1987, p. 178). Hence, they first applied propositional analysis to all input texts and then counted the number of propositions (including types of propositions, e.g. predicates). To contrast their results, they also used more “typical” linguistic variables such as the number of words, content words, sentence length as well as Flesch’s reading grade level, a readability measure. Vocabulary was measured using the Kučera and Francis word frequency counts, for all words and for content words only (ibid., p. 180).

A mixture of the linguistic variables used first by Drum and colleagues *and* the propositional approach by Embretson and Wetzel was adopted by Perkins and colleagues (Perkins, Gupta, and Tammana 1995): they incorporated both “simple” counts (words, content words, lines, sentences, paragraphs per passage, number of test items per passage etc.) and counts of propositions (number and density of arguments, predicates and modifiers etc.) in their ANN study of item difficulty. A dichotomous variable of whether the input text content belonged to the humanities

¹²²In a DIF study of verbal analogy items solved by black and white test takers (Freedle and Kostin 1997), the authors explored a vocabulary predictor based on the Kučera/Francis word frequency count (the log of the less frequent word in the item stem), but could not successfully predict DIF using this variable.

or not was also included.

In an item difficulty study of a German L1 listening assessment, Neumann (2012, pp. 140 sqq.) chose a broad approach and established five large groups of variables ranging from linguistic aspects of the passage and items to a global assessment of task difficulty. Her first group, “Komplexität des Wortschatzes und sprachliche Merkmale”, measures lexical complexity using categorical variables of the length of words, the proportions of monosyllabic and polysyllabic words, content words, and a frequency measure which indicates whether more or less than 60% of the words in the input text overlap with the Grundwortschatz Leipzig¹²³. The language-related variables in this group include the existence of particular structural, rhetorical and deictic features, retakes, referentials and negations as well as the proportion of nouns and verbs. In a group called “Inhaltlich-thematische Merkmale”, Neumann operationalises the content of the input texts, categorising whether a text is literary or not, as well as its text genre, communicative function and topic. A third group, “Struktur der Stimuli und propositionale Dichte”, is based on propositional analysis. The variables in this group are counts of propositions overall, their length, and the existence of specific types of propositions (*Relationstypen*, cf. *ibid.*, p. 145). All variables are operationalised as categorical variables which either measure the mere existence of a linguistic structure, or, more often, stand for a certain amount (e.g. the variable VER is dichotomous: if there are more than 15% of verbs in a stimulus, the variable changes).

In their experimental study of item difficulty, Brindley and Slatyer (2002) varied the text genre of their listening assessment tasks, using either a monologue or a dialogue to investigate whether this changes item difficulty. They also varied three aspects relevant to listening comprehension: the speech rate, whether or not the test takers heard a recording or a live rendition of the input text, as well as the number of times the test takers could hear the input text.

With the proliferation of computer-assisted linguistic analysis, researchers have added variables to their analyses which are determined by machines. Carr (2006, p. 274), for instance, used vocabulary measures provided by the RANGE program (Heatley, Nation, and Coxhead 2022), as did Brunfaut and Révész (2015, pp. 151 sq., see also Révész and Brunfaut 2013, pp. 45 sq., a different study with similar variables) who used WebVocabProfile (Cobb 2023), a relative of RANGE. In both cases, the programs allowed the researchers to operationalise the lexical structure of the input texts in more detail, providing the proportions of words among the most 1000, 2000 etc. frequent words, as well as lexical diversity measures such as TTR. In addition, as has been described earlier (see section 4.1.3), linguistic complexity analysers such as Coh-Metrix (Graesser, McNamara, Louwerse, et al. 2004) offer a selection of more complex indices of linguistic complexity. From these, Brunfaut and Révész (2015, p. 152) selected indices of syntactic and discourse complexity (e.g. incidence scores of negations or (adjacent) argument overlap) as well as a score of the concreteness of the content words, applied to the input text of their listening assessment. Révész and Brunfaut (2013, pp. 46 sq.) also applied these variables to the NI. To these, they added “the corpus-based frequency of multiword expressions in the texts” in both studies, using a list of phrasal expressions (Martinez and N. Schmitt 2012), and variables specific to spoken language (e.g. speech rate, or frequency of contractions). Ozuru and colleagues used the analytic power of Coh-Metrix, though they chose only three variables they considered to be related to reading comprehension: number of propositions, number of words per sentence and a variable which measures lexical frequency (2008, p. 1005). Aryadoust and Goh (2014) even went as far as using Coh-Metrix indices exclusively for their methodological study of item difficulty. Their choice of variables was first guided by prior research and the construct, and then limited to 12 variables that correlated significantly with item difficulty. This led to an eclectic choice of variables ranging from a simple word count to the principal component score reflecting temporality specific to Coh-Metrix (*ibid.*, p. 11). Due to the limits of applying complex linguistic indices to small amounts of text, the authors analysed the NI, the item stem and the answer

¹²³Unfortunately, the author does not explain how or why she used this reference corpus. She merely cites the website of the Leipzig Corpora Collection (Goldhahn, Eckart, and Quasthoff 2012).

options together. Finally, Toyama (2018, 2019) investigated the extent to which Coh-Metrix's eight principal component scores (instead of specifically chosen individual indices) could predict item difficulty. In her study, she compared Coh-Metrix's predictive power to three other sources of linguistic variables which may explain item difficulty: average sentence length and mean log word frequency as used by the readability measure Lexile (D. R. Smith et al. 1989), three passage features established by Gorin and Embretson (2006), and a group of complex variables stemming from TextEvaluator (K. M. Sheehan et al. 2014), a tool for text analysis similar to Coh-Metrix.

A different set of linguistic variables was chosen by Sheehan and colleagues (2006): in their study of a standardised L1 reading assessment, the linguistic variables focused on “the degree of semantic relatedness between the propositions presented in an item's option list and corresponding sets of propositions in the referenced reading passage” (ibid., p. 8). Their analysis used the techniques of LSA, creating cosines of semantic vectors between the NI and the four answer options (ibid., p. 9). This provided the authors with measures of the semantic overlap between the relevant passage of the reading item and the MCQ.

Finally, Beinborn and colleagues (2014, 2015) analysed the difficulty of C-tests, a Multiple-Choice (MC) cloze test and a prefix-deletion test from the perspective of research on NLP, using word-level and passage-level variables to predict the difficulty of the gaps in the three test formats (see for a similar study of cloze item difficulty Trace et al. 2017). Trying to predict the difficulty of individual gaps in an English C-test in particular, they used the normalised word frequencies of a large web-based corpus (Brants and Franz 2006) including the probabilities of n-grams, identified gaps which would be filled with compound words, calculated the number of senses a word may have, and took into account the POS of the words in the gaps and of their direct neighbours. The authors also used measures of cognateness based on string similarity in a parallel corpus and word lists containing e.g. English words with Latin roots. Using machine learning, the authors included measures of the spelling difficulty of English words, checked whether a word was in its lemma form or a derivate and whether it appeared elsewhere in the same test. Arguing that gaps are more difficult if they are situated in a more difficult passage, some traditional passage variables were also added: average word and sentence length, TTR, verb variation and pronoun ratio as well as syntactic readability features (the number of certain POS types, phrase patterns and entity mentions) (Beinborn, Zesch, and Gurevych 2014, pp. 521 sqq.). Most of these variables were recycled in their multilingual study featuring tests in English, German and French. Unfortunately, they did not adapt their cognate measure, using instead the one they had created for their analysis of the English C-test, a measure of the cognateness between English and German and of the Latin roots of English words (Beinborn, Zesch, and Gurevych 2015, p. 5).

8.2.5 Variables related to the assessment of reading comprehension

As has been mentioned above, when reading comprehension is measured, the measurement instrument itself must be taken into account in addition to the characteristics of the reading text. Over time, based on empirical observation, theoretical models and plain common sense, a variety of variables have been created which reflect the various sources of difficulty which may arise from taking a reading comprehension test.

8.2.5.1 Task format

Most of the assessments studied in item difficulty research contained MCQs (e.g. Carr 2006; Embretson and Wetzel 1987; Freedle and Kostin 1993, 1999; Andre A. Rupp, P. Garcia, and Jamieson 2001), while the assessment studied by Buck and Tatsuoka (1998) seems to be the only one reviewed in this section which was made up of Short Answer Questions (SAQs) exclusively. In some studies, the assessment used varied task formats. In these, an item's task format was usually a variable, sometimes called item type (Aryadoust 2013, p. 45; Böhme, Robitzsch, and Busè 2010, p. 90) or item format (e.g. Neumann 2012, p. 148, who used two separate variables to

capture the degree of “openness” of each item). Brindley and Slatyer (2002) explicitly varied the task format in their experimental study, interested in whether this would influence item difficulty. While keeping other aspects stable, they used either a sentence completion task, a table filling task or SAQs.

8.2.5.2 Position effects

Two major position effects have been operationalised in the various studies: the position of the correct answer in MCQs (e.g. Freedle and Kostin 1999; Neumann 2012), and the position of the relevant information in the input text. The position of the critical information (or the NI, depending on the study) has been measured either as a count of the words before the critical information is reached (e.g. Freedle and Kostin 1993) or as a categorical variable specifying where the NI is situated (e.g. in the first or last sentence, or in the middle of the passage, cf. Freedle and Kostin 1993, pp. 149 sqq.; Buck and Kikumi Tatsuoka 1998, p. 135). Freedle and Kostin (1993, pp. 138 sqq.) included this variable based on empirical research of reading comprehension, which indicated that readers expect information on the main idea of a text in its beginning sentences (e.g. Hare, Rabinowitz, and Schieble 1989). In her analysis of TOEFL listening comprehension items, Kostin also added a variable to assess whether the relevant information was in the *last* clause, citing Gernsbacher (1990, cited in Kostin 2004, pp. 5 sq.).

8.2.5.3 Overlap between input text and item

Based on findings by Freedle and Fellbaum (1987, cited in Freedle and Kostin 1993, p. 136), Freedle and colleagues added variables to their analyses which operationalised the lexical and semantic overlap between the input text and the item. The premise of these variables is that test takers will recognise words and expressions in an MC item which are either identical or related to words and expressions in the input text. If these are in the item stem or in the correct answer option, the item is predicted to be easier. Inversely, if incorrect answer options present overlap with the input text, the item may be more difficult. Since this effect is somewhat variable (the overlap may be helpful or not, it may be subtle or obvious, etc.) it has been operationalised in several ways over time. Drum et al. (1981, p. 508) calculated the percentage of new content words in the stem, the correct and the incorrect choices and found that “the inclusion of new content words in the correct choice [...] negatively biases the responses”, as did Davey (1988, p. 73). Freedle and Kostin (1993, pp. 150 sq.) considered the matter from the other side and used variables such as the “ordinal position of the earliest word on the first line that overlaps with a content word in the correct answer of a main idea item”, the “percent words in correct answer that overlap with words in key text sentence [in inference items]” or a variable which indicates that the “key word in stem corresponds to unique word in text and relevant information is in the same sentence [in supporting idea items]”. In their analyses, they usually found those variables to impact item difficulty once lexically related words were included (Freedle and Kostin 1993, pp. 150 sq., 1999, p. 16).

Influenced by these findings, both Kostin (2004) and Carr (n.d.) expanded on these variables in their subsequent studies of TOEFL item difficulty: Carr (ibid., pp. 24 sq.) defined as many as 18 variables to capture the lexical relatedness of the stem and the key sentence(s) alone, differentiating between words appearing in the stem *and only* the key sentence(s), the stem and the passage *excluding* the key sentence(s), lexical repetitions, lexically related substitutions and lexically unrelated substitutions and considering the number and percentage of words both in the stem and the key sentence(s). In all cases, the variable was operationalised as the number and proportion of the overlapping words identified. Kostin (2004, pp. 44 sqq.) chose a different method, using mostly dichotomous variables that indicate aspects such as whether the correct answer has more or less lexical overlap with the input text than the incorrect options, or whether the overlap is with the last sentence of the input text.

Unfortunately, it is quite uncommon to use a different language for the test items, at least in the assessments examined in the majority of item difficulty studies. Hence, no precedent for the operationalisation of overlap between items and input texts *in different languages* can be described.

8.2.5.4 MCQ: plausibility and falsifiability of the answer options

An important finding of the Drum et al. study (1981) is the predictive power of the *plausibility* of the incorrect answer options: in most of the assessments they analysed, this variable was the best predictor of item difficulty (ibid., p. 508). Plausibility was a dichotomous variable rated by humans once per item and defined as “a condition in which one or more of the incorrect choices was reasonable and was not denied by the text”.

Based on cognitive research interested in decision processes, Embretson and Wetzel (1987, pp. 178 sq.) operationalised the process of choosing one answer option over another in their “Information-Processing Model for Evaluating the Response Alternatives” (cf. fig. 8.1).

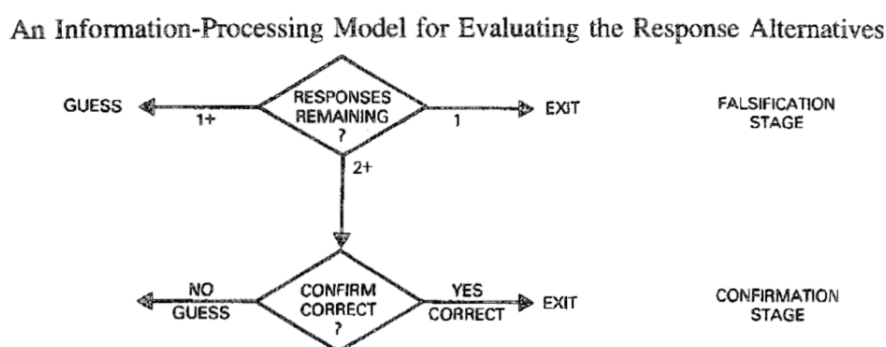


Figure 8.1: Embretson & Wetzel’s Information-Processing Model for Evaluating the Response Alternatives (1987, p. 179).

According to this model, test takers will first try to falsify as many answer options as possible and then either guess or confirm the remaining one(s). If many answer options can be falsified, then an item is considered to be easier. Inversely, if the correct answer cannot be confirmed by the input text, then the item is thought to be more difficult. In their study, falsifiability was operationalised as “1.0 minus the probability that the correct answer could be selected by random guessing among the remaining non-falsified set” (ibid., p. 179), while a dichotomous variable expressed whether or not the correct answer could be confirmed by reading the text. The authors also included a plausibility variable defined by Drum and colleagues earlier, but found that their more detailed model fit the data better (ibid., p. 190). In fact, Drum et al.’s definition of plausibility seems to condense two types of incorrect answer options: those which can be discarded by the attentive test taker without referring to the input text (which is, interestingly, described as a desirable trait of good readers, see Drum, Calfee, and Cook 1981, p. 493), and those which can only be falsified if the input text is processed correctly.

Other studies also included either the plausibility or the falsifiability of the distractors in their analyses, though they did not always explicitly define what they meant by either term. Buck and colleagues included the plausibility of the answer options in their item difficulty analyses of a listening assessment, citing qualitative findings by Alderson (1990a) as backing for this choice. Gao and Rogers (2011) counted the number of distractors which were plausible or had lexical overlap with the input text, while Ozuru and colleagues used a dichotomous variable indicating whether an item had at least one falsifiable distractor (defined as such “if the passage provided explicit textual evidence that the distractor was incorrect”, 2008, pp. 1004 sq.). Rupp and colleagues chose a more empirical operationalisation, using propositional overlap between

the distractors and the input text to determine whether an answer option would be considered *plausible* (2001, p. 194)¹²⁴.

8.2.5.5 Comprehension processes

While many researchers focused on linguistic aspects of the input text and items, some, particularly cognitive psychologists, chose to study the influence of the reading and decision-making processes involved in solving a reading assessment task. Embretson and Wetzel (1987) laid the foundation for this type of study: along with linguistic variables, variables based on propositional analyses, the falsifiability and plausibility of the answer options and the length of the NI (as mentioned above), the authors used Anderson's (1972) comprehension levels to describe the link between the input text and the items. These range from questions which are verbatim reproductions of statements in the input text (level 1) to inductive questions (level 6) which require the test taker to induce an answer from various pieces of information in the text (see for an example *ibid.*, p. 159). Much later, Ozuru and colleagues (2008, p. 1004) substituted this taxonomy by a four-level passage/questions relation coding system they deemed more closely related to the C-I model of reading comprehension: the first category (*text-based questions*) represents shallow word-level comprehension, the second (*restructuring/rewording*) processing at the sentence level, while the third and fourth item type (*integration or bridging questions* and *knowledge-based inference questions*) operate across sentences, representing, respectively, the construction of the text and situation model. To this, they added a five-level abstractness variable and falsifiability measure based on Embretson and Wetzel (1987, see also above, section 8.2.5.4). Toyama (2019) adopted these three coding schemes in her study of an online reading assessment.

In a small study focused on explaining comprehension processes for diagnostic purposes, Sheehan and colleagues (2006) classified L1 reading items into one of three categories: *Right There*, *Think and Search* and *On My Own* (the terminology is based on Raphael and Gavelek 1984 (cited in K. Sheehan, Kostin, and Persky 2006, p. 11) who developed the categories for instructional purposes). The *Think and Search* category was further specified based on how the relevant information was distributed in the input text (in multiple locations, concentrated, dispersed, intertextual, see *ibid.*, pp. 17 sq.). A similar approach was chosen by Sawaki and colleagues (2009), who identified four comprehension skills each for reading¹²⁵ and listening items through expert judgments, applied these to a selection of TOEFL items using a Q-matrix and checked whether their categorisation fit the test taker responses (*ibid.*, pp. 196 sq.).

Rupp and colleagues (2001) established several variables to operationalise the complexity of the comprehension process: *item type* (detail, main idea/gist, prediction, understanding relations), *type of information* (abstract or concrete), *type of match* (a measure of the cognitive complexity of finding the correct answer, ranging from 0 to 21) and the *directness of the information* (explicit or implicit), all variables requiring human rating. Davey (1988, p. 75) also used a categorical variable she called *Inference Type*, which ranges from 0 – no inference required, to 4 – background knowledge inference required, while Perkins and colleagues (1995, pp. 38 sq.) condensed Anderson's (1972) comprehension levels into a dichotomous variable of whether the information given in the stem was explicitly stated in the input text.

Finally, Gao and Rogers (2011) established seven reading processes, four test management processes and three test wiseness processes through a literature review. In a subsequent expert review and based on cognitive interviews with advanced test takers, the reading processes and test

¹²⁴A similar approach was chosen by Sheehan and colleagues, who, based on LSA measures of semantic relatedness between four answer options and the NI, created a measure (Lambda) of whether the correct answer options was more or less related to the passage than the incorrect options (K. Sheehan, Kostin, and Persky 2006, p. 20, see also Svetina, Gorin, and K. K. Tatsuoka 2011). This measure, however, could not provide more information about item difficulty than the semantic relatedness of the correct answer option with the NI itself, from which it was mathematically derived.

¹²⁵1) Understanding word meaning, 2) Understanding specific information, 3) Connecting information, 4) Synthesizing & organizing information (Sawaki, H.-J. Kim, and Gentile 2009, p. 199).

management processes were retained for a tree-based regression. The reading processes are: 1) word recognition, 2) using vocabulary knowledge, 3) using syntactic knowledge, 4) using knowledge of discourse structure, 5) synthesis¹²⁶, 6) drawing inferences, 7) using pragmatic knowledge. Each of these variables was coded categorically, usually ranging from *not required* to *critical* (Gao and W. T. Rogers 2011, pp. 84 sqq.). Their test management processes include the position of the NI, the number of plausible distractors, but also the need for background knowledge and a variable which assesses the explicitness of the match between item and input text (*literal, synonymous, no match*). Each of these variables is, again, categorical (ibid., pp. 86 sq.).

Of course, the more linguistic studies presented above also included variables which operationalised the comprehension process in some way. Neumann, for instance, employed two variables, almost hidden amongst others: one measures the amount of words in an input text for which an inference is needed (part of her group IV), the other counts all words and expressions which require background knowledge (Neumann 2012, pp. 144 sq.), though it is somewhat unclear how the latter is operationalised with respect to the input text. Carr (2006, n.d.) used rating variables that establish the link between the input text, the NI and the item, assessing the extent to which the key sentence and the item are ambiguous, how well unfamiliar information in the key sentence is explained, how much of the input text has to be processed and how much topical knowledge is needed in addition to the input text in order to identify the correct answer. Freedle and Kostin included a five-level abstractness measure of the input text, as well as item variables dependent on the predefined item type (main idea, supporting ideas, inference), which implicitly considered the comprehension processes intended by the test developers (1993, pp. 146 sq., 1999, pp. 28 sqq.).

Similar variables are present in the CDA studies by Buck and colleagues (1998; 1997): though not explicitly separated from the more linguistic variables, they coded the items in their analyses for aspects such as the need for background knowledge, the possibility to solve an item using background knowledge, and the item type (main idea, inference). Aryadoust (2013), studying listening comprehension, also included a variable he called *information type*, which is similar to what others have called *item type* (covering items which require test takers to understand explicitly stated input and to make paraphrases).

8.2.6 Test-taker variables

In item difficulty research, given its focus on the test and its items, the test taker is rarely taken into account. Among the studies discussed in this chapter, there are only three notable exceptions: Brunfaut and Révész (2015) administered, along with an English L2 listening assessment (a practice test of the Pearson Test of English (PTE) Academic), two working memory tests and a questionnaire of foreign language listening anxiety. Neumann (2012) also reports on data from a working memory test of a subset of students, as well as on the results of a test taker questionnaire. Toyama's data (2019) allowed her to add vocabulary knowledge to her analysis: in addition to five sets of reading comprehension items, the test takers had been presented with a vocabulary assessment. Toyama then analysed the item difficulty of one of those testlets while controlling for the test takers' vocabulary knowledge.

8.2.7 Summary

Though the general objective of all of these studies is very similar—to explain in some way how comprehension test items differ from each other—the actual operationalisation of test, task and item characteristics comes in many forms. Almost all authors tried to describe the language of the input text and the items, and many created measures which attempted to capture the interplay of input text and item. Some researchers focused instead on the cognitive processes

¹²⁶The need to “integrate, relate or summarize” across various pieces of information in the input text (Gao and W. T. Rogers 2011, p. 85).

that test takers are expected or hypothesised to go through while reading or listening to a text and/or solving comprehension items. Often, these two approaches were combined. In some cases, attempts were made to describe the test(s) in every detail using many variables, while in others, only a handful of variables was established to explain their unique way of contributing to item difficulty. Some variables are “objective” counts or results of computerised linguistic analyses, or they are characteristics described by test developers. Others are results of human ratings, sometimes done by two or three raters. Overall, only a few studies use the exact same variables, or study the same tests in order to examine the replicability of the results. This may be a major drawback of item difficulty research, and it may be one of the reasons why, as will be shown in the next section, a lot of open questions remain as to the sources of item difficulty.

8.3 Results of item difficulty research

Perhaps the most important finding of most of the item difficulty studies reviewed in this chapter is the often relatively low predictive power of the variables under investigation and the amount of variance that remains once all analyses have been run. This is supported by the uncertainty with respect to the possible sources of DIF that have been identified in studies of multilingual assessments (see section 7.4).

8.3.1 Primarily linguistic modelling

In Freedle and Kostin’s (1993) study of TOEFL reading items, though 31 out of 71 variables were correlated significantly with item difficulty, most of these correlations were low (16 below $r = .20$, 14 between $r = .20$, and $r = .30$, the highest $r = -.35$ ¹²⁷, *ibid.*, pp. 157 sqq.). The linear regression analysis of the same data was conducted on a nested and a non-nested sample of the items¹²⁸, yielding two final models: the nested sample was described in a model that included eight significant predictors and had an R^2 of .329, and the non-nested sample could be predicted by 11 predictors ($R^2 = .57$). Among the two models, only the following three variables were useful predictors in both (*ibid.*, p. 162):

- percentage of words in the correct answer that overlap with content words in the key text sentence, including lexically related words, for supporting idea items,
- the number of words in the key text sentence for supporting idea items,
- the ordinal position of the earliest word in the first text line that overlaps with the content word in the correct answer for main idea items.

To test the findings from the latter model, the authors applied the model to another non-nested selection of items, and could explain 29.8% of the variance in this set. Overall, Freedle and Kostin found that lexical overlap, sentence length, the location of the NI, rhetorical organisers and subject matter could predict item difficulty across their two samples in some way (*ibid.*, pp. 163 sq.).

Similar results were observed in Freedle and Kostin’s study of TOEFL listening items (1996, 1999): of 85 variables that were entered in the analysis, 39 significantly correlated with item difficulty for the nested, and 42 for the non-nested sample. In the regression analyses, 14 variables were predictors of item difficulty in the nested sample ($R^2 = .350$), and 12 in the non-nested sample ($R^2 = .332$). Of those, seven variables were found in both models:

¹²⁷The “percentage of words in the correct answer that overlap with content words in the key text sentence, including lexically related words, for supporting idea items” (Freedle and Kostin 1993, p. 162).

¹²⁸The non-nested sample was a subset of the nested sample. The latter contains input texts which refer to more than one item, while the non-nested sample is a selection of those items where each item refers to its own, separate input text.

- attractiveness of the correct option compared to the attractiveness of the most popular incorrect (*number of overlapping words with the input text in correct answer option is higher than in the correct answer options + the most frequently overlapping word is in the correct answer option*),
- the topic is nonacademic (topic is one of miscellaneous, sightseeing, extracurricular, university procedure, academic procedure),
- the item type is *main idea*,
- the topic is social sciences (anthropology, economics, sociology, political science),
- first 20 words, lexical match for incorrects, including lexically related (*in the first 20 words of the input text, there is lexical overlap (or semantic overlap) with an incorrect option, but not the correct option, or the “incorrect overlap” comes first*),
- the input text is organised as a list or a description,
- the main idea information is in the second sentence.

The authors applied several additional regression analyses to sub-samples of this data, getting similar results (Freedle and Kostin 1996, pp. 35 sqq.). In general, if the selected items were more similar, i.e. if only one item type was analysed, the variance explained was higher, though the number of variables which were predictors of item difficulty remained relatively low and the selection unstable.

Additional studies of the TOEFL listening comprehension assessment were conducted by Nissan and colleagues (1996), as well as, some years later, Kostin (2004). Nissan and colleagues, much like Freedle and Kostin, found only five of their 19 variables to have a significant relationship with item difficulty (*Delta*). One of these was *infrequent vocabulary*, one was the presence of two or more negatives in the input text, and the other three were related to the kind of information assessed in the item (most notably, implicit vs. explicit) (Nissan, DeVincenzi, and Tang 1996, pp. 25 sqq.). Kostin elaborated on this study, and also found several lexical variables (infrequent vocabulary, understanding of idiom) as well as the presence of negatives to be significant predictors of item difficulty (*Delta*). In addition, she identified variables related to the overlap of the correct answer with the input text, the distribution of the NI, two syntactic variables, and, again, the topic of the input text. Overall, 14 out of 49 original variables remained in the final regression model ($R^2 = .398$), of which 7 operationalised the characteristics of the correct answer (Kostin 2004, p. 26). A similar result was reached by Davey (1988) in her study of item difficulty features in an English L1 standardised reading assessment. Though she had set out to model item difficulty separately for lower- and higher-proficiency readers, Davey found only two significant predictors, independently of reader group: a variable she called Response Location, which is an operationalisation of whether the answer can be found verbatim in the input text, and stem length (ibid., p. 72). In her analysis, she also describes a number of interactions between her various variables, emphasising the impact of the concrete nature of the reading assessment under investigation (ibid., pp. 72 sqq.).

Carr (2006), who approached the matter in a different way, using factor analysis to explore the relationships between the many task and item variables he had established¹²⁹, made two important observations: first, a large majority of the variables which “emerged as constituting a key dimension along which reading items may vary” (ibid., p. 285) were operationalisations of the *key sentence* (especially lexical and syntactic characteristics, with the input text only adding relevant variance by way of its content and lexical variety). This, he concludes, underlines the

¹²⁹Note that in his 2006 article, Carr only reported on the results related to the key sentence and passage characteristics. Variables related to the items were not part of the analysis (Carr 2006, p. 275). Unfortunately, the author does not seem to have published any more research on this study.

need to take into account the characteristics of the NI rather than of the input text in its entirety. Second, of the a , b and c -parameter of a 3-PL IRT model he added to his factor analytic model, only item discrimination (a) added explanatory power via the “passage specificity factor” which measures the topic of the input text. In Carr’s own words: “No significant relationship was found between item difficulty or guessing [...] and empirically salient characteristics of passages and key sentences” (ibid., p. 282). This finding can be interpreted as evidence that item difficulty as measured by IRT modelling does not hinge primarily on the characteristics of the input text or the relevant passage. Instead, Carr notes, based on Bachman (2002), test taker characteristics and item characteristics, as well as their interactions need to be taken into consideration in order to predict item difficulty in a more meaningful way (Carr 2006, pp. 285 sq., see also Davey 1988, p. 74).

Similar evidence can be inferred from the regression analyses of 18 different reading assessment by Drum and colleagues (1981). Though, as has been pointed out by other researchers (Freedle and Kostin 1993, p. 135; Davey 1988, p. 74), the results may be overinflated given the low number of items and the high number of characteristics in this study: a closer look at the regression coefficients per variable and per assessment (Drum, Calfee, and Cook 1981, pp. 504 sqq.) reveals that none of the 16 variables is a stable predictor in all 18 assessments under investigation. Instead, a predictor like “Actual Information”¹³⁰ explains as much as 44% of the variance of p in one test, but is not among the ten strongest predictors in six other assessments. Nevertheless, Drum and colleagues (ibid., pp. 509 sq.) draw some general conclusions, particularly with respect to the age of the readers, having studied reading assessments for grades 1 through 10: younger readers, they argue, encounter difficulties with word recognition and word meaning, while upper elementary students struggle when the information presented in input texts is densely packed. With respect to the MCQs, a consistent effect was observed when infrequent words appeared in the answer options: such answers were often avoided. Again with respect to younger children, the authors also hypothesised a debilitating effect of question words, which are acquired relatively late. Finally, as has been mentioned above, one of the most consistent predictors for test takers above grade 3 was the plausibility of the answer options, though this effect was particularly strong in some assessments, and less so in others (ibid., p. 507). This may point to a combination of developmental effects (i.e. very young readers do not yet base their decisions on this type of evidence) and aspects of test development (i.e. how MCQs were constructed).

Using only correlational analyses, mainly due to a small task sample size ($n=30$), Brunfaut and Révész (2015, pp. 156 sq.) identified eight out of approximately 50 variables to be significantly correlated with item difficulty (estimated by a Rasch model). Four of these were measures of the frequency of multiword expressions, which has rarely been operationalised in this type of research. Three variables are argument overlap variables created by Coh-Metrix, indicating that input texts which have “a more uniform thematic [focus] of (parts of) passages” (ibid., p. 160), and, thus, more cohesion, may be easier to understand. A final variable was related to listening: input texts with more contractions, indicating more speech-oriented (academic) discourse, proved to be easier to understand. In a similar study, the same authors found only lexical variables to be related to item difficulty, with the exception of causal content, which also predicted a substantial amount of variance, when regressed on its own (Révész and Brunfaut 2013, p. 52). In both studies, the assessments were aimed at second language listening comprehension in an academic context.

8.3.2 Modelling response processes

Somewhat different predictions of item difficulty were reached by studies which explicitly modelled comprehension processes in addition to the linguistic characteristics of the input texts and items. For instance, the best model reached by Embretson and Wetzel (Embretson and Wetzel

¹³⁰Measured as “ $\log_2$ of the number of content words that were mentioned only once in the passage” (Drum, Calfee, and Cook 1981, p. 494).

1987, p. 188) in a study of adult L1 reading comprehension contained six variables related to understanding the text, and seven variables related to the decision process, explaining, according to the authors, roughly 37% of the variance (a multiple correlation of .61). Interestingly, more variance was explained by the decision process, operationalised as the word frequency in the answer options, Anderson's decision processes, the falsifiability of the incorrect, and the "confirmability" of the correct answer option (Embretson and Wetzel 1987, pp. 189 sq.). The comprehension of the input text was, as the authors had expected, influenced by the propositional characteristics of the input text: more explicit cohesion made input texts easier to understand, while denser propositional content did the opposite (*ibid.*, pp. 188 sq.). Overall, the authors pointed out that the two processes are not related, and that reading comprehension items may be more difficult on either or both dimensions. Hence, depending on which dimension is more important in the items of a test, it may measure primarily reading comprehension *or* reasoning.

Ozuru and colleagues (2008) also modelled passage variables and item variables as separate, interactive predictors for item difficulty (proportion correct p) using HLM. With respect to a standardised commercial reading assessment for 7th to 9th-graders, the authors found that passage characteristics were better predictors of item difficulty, in particular word frequency (measured as the mean minimum frequency per sentence), whereas almost no systematic explanation for item difficulty was found in the 10th-to-12th-grade assessment. This suggests, perhaps, that the variables chosen in this study are not the source of the variation in difficulty, or that the items do not vary systematically with respect to the theory-based aspects of reading comprehension (*ibid.*, p. 1014).

Using similar predictors, but LR-LLTM as her statistical approach, including a control for vocabulary knowledge, Toyama (2019, pp. 41 sq.) compared the fit of a variety of models. Modelling only features of the input text, she found mean sentence length, mean log word frequency, syntactic simplicity, connectivity and temporality to be significant predictors of item difficulty, the latter three being Coh-Metrix text easability principal components. In a model which only contained item features, most of the suspected variables turned out to be significant, in particular the vocabulary demand of the answer options, but also the item type (i.e. Ozuru et al.'s passage/question relations) and the abstractness of the information requested. If both aspects—text representation and response decision—were taken together, the same variables remained significant, and the combined model fit the data better (pseudo- $R^2 = .588$, including vocabulary knowledge). This, much like the results of Embretson and Wetzel (1987, p. 186), points to the fact that finding the correct answer is a separate process from reading comprehension. Toyama, however, went one step further and analysed interactions of the significant variables and reader vocabulary knowledge, finding complex mechanisms between item types, text features and vocabulary knowledge (2019, pp. 47 sq.).

Predicting test taker skills instead of item difficulty, Buck and colleagues usually reached quite high classification rates: in an analysis of the reading section of the Test of English for International Communication (TOEIC) involving 5,000 test takers, 91% of the test takers were classified "into their latent knowledge states" (Buck, Kumi Tatsuoka, and Kostin 1997). This means that, based on item characteristics defined in a Q -matrix and the raw test results, for most of the test takers, the model could give an estimate of mastery for 24 attributes¹³¹. This data—to what extent a test taker was estimated to master certain test taking skills—was used on the same item set, a related set of reading items, and a set of listening items, estimating total test scores with reasonable accuracy (adj. R^2 for the same test: .97, different reading items: .80, listening items: .54, cf. *ibid.*, pp. 446 sq.). Overall, the authors identified five sets of variables that influence item difficulty in their chosen reading assessment: vocabulary knowledge, inferencing including the need for background knowledge, the location of the NI, lexical overlap between the

¹³¹ Attributes are descriptions of test solving skills based on item characteristics (e.g. "The ability to synthesize scattered information", operationalised by "The necessary information is [...] scattered across more than one place."). They may also be interactions of these (Buck, Kumi Tatsuoka, and Kostin 1997, pp. 441 sq.).

input text and the answer options as well as the plausibility of incorrect answer options, and, finally, a number of interactions between those variables. The latter in particular are considered an important finding, as they provide empirical support for the multidimensionality of test solving processes: “when two attributes occurred together, something extra appeared or some additional difficulty arose that wasn’t there when they occurred separately” (ibid., p. 451). An earlier study of listening comprehension items by the same authors revealed similar aspects of item difficulty (inferences and background knowledge, the distribution of the NI, aspects related to the task format) in addition to aspects specific to listening comprehension (Buck and Kikumi Tatsuoka 1998).

8.3.3 Methodological studies

Aryadoust (2013), as well as Aryadoust and Goh (2014) were more interested in methodological aspects of item difficulty research, comparing the results of various statistical approaches. In his comparison of a type of ANN (ANFIS) and a path analysis, Aryadoust (2013, pp. 47 sq.) could show that the ANN model fitted best when all seven of his supposed variables were taken into account. The path analysis revealed a mediating effect of the test section of the International English Language Testing System (IELTS) listening practice test used in the study (which, according to the author, represents differences in the text genre, speech rate and other text effects). In addition, it also showed that, once test section was taken into account, only the propositional density of the input text and of the items, as well as the item format were significant predictors of item difficulty, while information type (i.e. listening skills) and the presence of modal verbs (as well as, again, the propositional density of the items) were mediated through the section of the test. In the study by Aryadoust and Goh (2014), in which the authors compared the results of linear multiple regression, CART and two types of ANN, only linguistic variables created by Coh-Metrix were part of the analysis, while item difficulty was discretised into low-difficulty and high-difficulty items. The regression analysis showed only two significant predictors of item difficulty (*words before the main verb* and *noun and verb hypernymy*), while the CART analysis identified nine out of 12 variables, which added information to item difficulty. Of those, the temporality principal component was the most important variable (which had also been identified by Toyama 2019, p. 41), followed by the average givenness of each sentence¹³², word count, content word overlap and the Flesch-Kincaid grade level. In one of the ANN analyses, the best model was made up of similar variables as in the two other analyses: word count, content word overlap, logical connectives incidence, verb incidence and, again, the Flesch-Kincaid grade. This modelling approach, however, did not achieve a particularly high prediction, displaying moderate fit both in the learning as in the testing sample. The other ANN method, MLP, achieved a much better predictive power, with classification accuracies above 80%. Interestingly, this model included all 12 variables. Despite the still somewhat unclear functioning of ANN, Aryadoust’s results in both studies point to the multidimensionality of item difficulty, and underline Carr’s conclusion that item difficulty is made up of multiple aspects, above and beyond the linguistic features of the input text or items.

8.3.4 Test taker characteristics

In DIF studies, characteristics such as gender (e.g. Zenisky, Hambleton, and Robin 2004; Zhu and Aryadoust 2019), age (e.g. F. Garcia et al. 2018; Khorramdel et al. 2020) or SES (e.g. Abedi and C. Lord 2001) are known to influence test results (see H. Li, Hunter, and Bialo 2022 for a review of DIF research in language testing and further references), but it is not

¹³²“LSA Given-New” is a measure of how much *new* information is provided in each sentence compared with the prior information given by the text (McNamara, Graesser, et al. 2014, p. 66). Since all measures in this study were applied to a combined text object made up of the NI, the item stem and the answer options, Aryadoust and Goh interpret this measure as an operationalisation of overlap between the NI and the item (2014, p. 11).

always clear whether this is due to the characteristics of the test, or a result of differing learning opportunities. Hence, empirical research of the interaction between these two aspects, as Bachman (2002, p. 469) pointed out more than two decades ago, is long overdue. Unfortunately, given the small number of studies which took both task *and* test taker characteristics into account, not a lot is known about this interaction. Brunfaut and Révész (2015), who had administered a listening assessment, two working memory assessments and a listening anxiety questionnaire, correlated the scores between the various instruments. They did not find a significant relationship between the 30 MCQ items they were studying in particular (a section of the listening assessment) and working memory, but they did find a low to moderate relationship between the overall listening scores and the two working memory scores (*ibid.*, p. 158). Interestingly, they found even stronger correlations between the listening assessments and the scores from the listening anxiety questionnaire: “Higher levels of listening anxiety were associated with lower L2 listening performance” (*ibid.*, p. 159). Similar findings were reported by Neumann (2012, pp. 186 sq.), who found moderate correlations between the results of a working memory test and item difficulty, in particular when she eliminated very easy tasks from the analysis. She also investigated the linguistic background of the test takers, and found low to moderate correlations between the languages spoken at home, the first language, and the test takers’ performance on the listening assessment. Not surprisingly, native speakers of German and students who spoke German at home performed somewhat better, especially on harder tasks (though note that the correlations are rarely above .25; cf. *ibid.*, pp. 187 sq.).

In a statistically more sophisticated study, Toyama identified a generally large predictive power of the test takers’ vocabulary test results on their results of the reading comprehension test, as evidenced by the consistently high fixed effect of vocabulary knowledge in all models (e.g. Toyama 2019): not surprisingly, test takers who scored well on the vocabulary test, also scored well on the reading comprehension assessment. More interestingly, Toyama also calculated models which contained interactions between certain text passage variables, item variables and readers’ vocabulary knowledge. Among other things, these analyses revealed an interaction effect for mean sentence length, mean log word frequency and syntactic simplicity, indicating that test takers who scored higher on the vocabulary test could take more advantage of the facilitating effects of these variables than test takers whose vocabulary was measured to be smaller (*ibid.*, pp. 48 sq.).

8.3.5 Languages other than English

Of particular interest in this thesis are the results of item difficulty studies on languages other than English, and whether they differ notably in their results. Neumann (2012), as described above, studied a German L1 listening assessment, using a variety of approaches. Her correlational analyses of stimulus and item characteristics with task¹³³ and item difficulty showed significant¹³⁴ effects for the amount of long words (> 6 letters) and monosyllabic words, the presence of negations and a variable “Hörkontext”, which operationalised a form of text genre, though the latter might be an effect of low occurrences (*ibid.*, pp. 170 sqq.). As for item characteristics, Neumann identified item format, the position of the item within the task and, most importantly, the plausibility of the incorrect answer option (*ibid.*, pp. 178 sq.) as relevant predictors of item difficulty. She also pointed out that comprehension processes, e.g. the cognitive operation demanded by the item, were much less decisive than more concrete characteristics of the items. Using varying groups of variables, Neumann also conducted regression analyses on item and task difficulty. In her analysis of *task* difficulty, based on 30 tasks, a model with 12 variables explained 87.9% of the variance, though this result may be overinflated in the same way as Drum et al.’s study (see Freedle and Kostin 1993, p. 135; Davey 1988, p. 74). When stimulus characteristics were used to

¹³³Task difficulty was defined as the average of the item difficulties in the task (2012, pp. 169 sqq.).

¹³⁴Neumann actually named a few more variables, though the significance of these correlations is above $p = .05$. Among these variables are two more operationalisations of word length, several syntactical structures, text genres, average propositional length, and some variables specific to listening comprehension (*ibid.*, p. 170).

predict *item* difficulty, much less variance could be explained (22.6% for the best model). In both cases, the variables were similar to the ones mentioned above: the amount of long words, word frequency, the need for background knowledge, some syntactical variables and some variables related to listening comprehension. Neumann's analyses of *item* difficulty using *item* characteristics are more similar to other regression analyses, explaining 45% of the variance in the best model. Among all her analyses, the plausibility of the incorrect answer options, the comprehension process ("Anforderungsbereich" according to the underlying construct) and aspects of the item format seemed to be the most relevant (see for more details Neumann 2012, pp. 250 sqq.).

Böhme and colleagues (2010), studying the same assessment as Neumann, used regression analyses to examine the contribution of task and item characteristics to item difficulty in reading and listening. Though they did not publish a full list of variables, their choice was clearly inspired by other empirical studies on item difficulty (e.g. Buck and Kikumi Tatsuoka 1998). Their analysis of the listening tasks reveals a strong effect of the distribution of the NI and a variable they called "Lebensweltbezug", possibly measuring the concreteness of the input text. In reading, the item format seems to have been a major predictor, along with, again, the distribution and the explicitness of the NI (Böhme 2011, pp. 90 sq.).

Studying two Dutch L1 reading comprehension assessments for kindergarten, Spelberg and colleagues were interested in the comparability of two different tests, and the relationship between the test results and some linguistic features. Correlational and regression analysis showed word knowledge (based on a list of words which gives the percentage of teachers who expect six-year-olds to know each word) and number of words to be the best predictors of item difficulty, regardless of the instrument, resulting in $R^2 = .45$.

None of these authors discussed the fact that most of their variables were based on item difficulty studies in English, or whether this had any influence on the results of their studies.

8.3.6 Summary

Taken together, the results of these various studies (and others which have not been presented here, cf. e.g. a recent metastudy by Ferrara, Steedle, and Frantz 2022) allow some tentative conclusions. First, despite the large number of possible variables described in section 8.2, researchers have been unable to conclusively identify specific sources of item difficulty in comprehension assessment. Instead, not really surprisingly, the results of these various studies can be interpreted as evidence of the multidimensionality of (reading) comprehension and of test item solving: depending on the assessment under investigation, its specific item formats and input texts, the test takers, and, of course, the variables and the specific way in which characteristics were operationalised, the results are different, sometimes contradictory. This was particularly obvious if the same study included more than one test (e.g. Drum, Calfee, and Cook 1981; Ozuru et al. 2008; K. Sheehan, Kostin, and Persky 2006), i.e. if the same set of variables was applied to different tests. Also, as Buck and Tatsuoka pointed out, some variables may be important in some constellations of test and test takers, but not in others (1998, p. 148), again leading to differing results. The multidimensionality of reading items is also underlined by Toyama's analysis of interactions between text, item and test taker characteristics (2019, pp. 47 sqq.), which shows that the link between item difficulty and these characteristics is much more complex than a one-on-one relationship. Second, as Carr (2006) hypothesised, text and item characteristics are almost certainly not sufficient to model item difficulty. Though there is very limited research on the influence of both task and test taker characteristics on item difficulty, the findings that do exist are evidence of an undeniable relationship between test taker characteristics and item difficulty (Brunfaut and Révész 2015; Toyama 2019). Third, there is, as of yet, no statistical approach to this problem that yields unequivocally good results. Instead, anything from simple correlations to sophisticated IRT modelling and machine learning has produced interpretable results to some extent. However, some researchers (e.g. Aryadoust 2013; Aryadoust and C. C. M. Goh 2014; Perkins, Gupta, and Tammana 1995) have shown that if different modelling techniques are applied to the

same data, the results do not necessarily compare. This may be linked to the other two problems mentioned here—we have not yet been able to establish enough variables to fully describe the functioning of test items—but it is also further evidence that the matter is complicated.

Variable group	References
Lexical overlap between item and input text	Aryadoust and C. C. M. Goh 2014; Buck, Kumi Tatsuoka, and Kostin 1997; Davey 1988; Freedle and Kostin 1993, 1996; Kostin 2004; K. Sheehan, Kostin, and Persky 2006
Lexical complexity, frequency or diversity (of input text, NI, item elements)	Brunfaut and Révész 2015; Buck, Kumi Tatsuoka, and Kostin 1997; Carr 2006; Drum, Calfee, and Cook 1981; Embretson and Wetzel 1987; Kostin 2004; Neumann 2012; Nissan, DeVincenzi, and Tang 1996; Ozuru et al. 2008; Révész and Brunfaut 2013; Toyama 2019
Word count, sentence length and some other syntactic variables	Aryadoust 2013; Aryadoust and C. C. M. Goh 2014; Carr 2006; Davey 1988; Freedle and Kostin 1993; Kostin 2004; Neumann 2012; Nissan, DeVincenzi, and Tang 1996; Toyama 2019
Location and distribution of the NI	Böhme, Robitzsch, and Busè 2010; Buck and Kikumi Tatsuoka 1998; Buck, Kumi Tatsuoka, and Kostin 1997; Freedle and Kostin 1993, 1996; Kostin 2004
Subject matter or topic of the input text	Böhme, Robitzsch, and Busè 2010; Carr 2006; Freedle and Kostin 1993, 1996; Kostin 2004
Rhetorical organisers or text genre	Aryadoust 2013; Freedle and Kostin 1993, 1996; Neumann 2012
Item type, comprehension process or need for inference	Aryadoust 2013; Buck and Kikumi Tatsuoka 1998; Buck, Kumi Tatsuoka, and Kostin 1997; Davey 1988; Embretson and Wetzel 1987; Freedle and Kostin 1996; Neumann 2012; Nissan, DeVincenzi, and Tang 1996; K. Sheehan, Kostin, and Persky 2006; Toyama 2019
Cohesion or information density of input text	Aryadoust 2013; Aryadoust and C. C. M. Goh 2014; Brunfaut and Révész 2015; Drum, Calfee, and Cook 1981; Embretson and Wetzel 1987; Révész and Brunfaut 2013; Toyama 2019
Plausibility, falsifiability or confirmability of answer options	Buck, Kumi Tatsuoka, and Kostin 1997; Davey 1988; Drum, Calfee, and Cook 1981; Embretson and Wetzel 1987; Neumann 2012

Table 8.1: The most relevant predictors of item difficulty, according to empirical findings cited in the literature. *Note that the cited studies vary in their operationalisations of these predictor groups and that they did not all measure variables from each predictor group.*

Nevertheless, a tentative selection of good predictor variables may still be possible, as some aspects tend to be among the relevant predictors in several studies, even though they are operationalised in different ways and applied to a variety of comprehension tests, from L1 reading assessments used in the lower primary grades, to advanced L2 listening assessments for graduate students. Table 8.1 contains a non-exhaustive overview of these predictor groups.

Another important finding from studies which considered response decision processes (e.g. Embretson and Wetzel 1987; Toyama 2019) is the probable existence of (at least) two separate processes involved in test taking: the actual reading comprehension, and the response decision. Three conclusions may be drawn from this: first, it seems untenable that a reading comprehension construct be solely focused on the reading comprehension part of the test taking process. Instead, test developers need to take into account the decision processes related to test item solving—either to make sure that the measurement of reading comprehension is not unnecessarily obstructed by the decision process, or by defining a more comprehensive reading comprehension construct which includes decision and evaluation processes (e.g. Sabatini et al. 2014). Second, no

matter how the construct is defined, the characteristics of item and of text-by-item variables need to be taken into account when studying item difficulty. Third, considering the different ways of operationalising comprehension and response processes in the studies described above, it seems that more research is needed into the actual mechanisms involved in reading comprehension and test item solving (e.g. qualitative studies, eye-tracking studies).

Finally, the three studies of assessments in languages other than English do not point to obvious differences between the languages, though the sample is quite small, the languages are strongly related, and, most notably, the question had not been asked by these authors. The study by Bergqvist and colleagues (2018, see section 7.4) may be more relevant in this respect, as it identified differences in item difficulty prediction *across (native) languages*. According to the authors, several explanations for this finding are possible, including structural differences between the languages, properties of the mathematics tasks under investigation, or the operationalisation of the variables across languages (ibid., pp. 51 sq.). One of the objectives of this thesis is to further investigate this issue.

8.4 The role of other languages in foreign language assessment

As has been mentioned, item difficulty research does not usually take into account the fact that in an assessment of second or foreign language proficiency, more than one language is at play, if only in the test takers' minds. In fact, in most of the studies cited above, the status of the language under investigation, i.e. whether it is an assessment used for native speakers or whether it is an international assessment mainly solved by speakers of other languages, does not seem to matter. This is even true for the few studies which did not examine an English language assessment (e.g. Böhme, Robitzsch, and Busè 2010; Neumann 2012; Spelberg, Boer, and Bos 2000).

In other lines of (foreign) language assessment research, of course, the role of other languages has been investigated. A major difficulty in researching this effect is how to distinguish between construct-relevant differences in language ability and construct-irrelevant effects produced by the assessment. For instance, group differences in test results may also be due to different instructional experiences by learners from different language backgrounds, leading to true differences in language proficiency (Elder 1996, p. 236). Also, though studies show that linguistic distance plays a role in how learners from different backgrounds solve test items, it remains unclear whether this should be considered construct-irrelevant. After all, as Brown and Iwashita point out, “the focus of language testing is generally on where the learners stand in relation to the target language rather than on the path by which they got there” (1998, cited in Elder 1996, p. 237). Nevertheless, the various empirical studies conducted on the subject still point to some effect of the other languages in a test taker's repertoire. Whether or not it is detrimental to the validity of a specific assessment may be difficult to determine, but since it can be shown that *something* is going on, the matter still needs some consideration.

A major early study on the topic was an analysis of TOEFL test results collected in the 1970s, conducted by Swinton and Powers (1980, pp. 17, 25). They found evidence that the three parts of the assessment (reading and vocabulary, structures and writing, listening comprehension) were solved in different ways by test takers from seven different language groups (African, Arabic, Chinese, Farsi, Germanic, Japanese and Spanish). Most importantly, they found that test takers from non-Indo-European language groups were more influenced by their vocabulary knowledge than by their reading proficiency when solving the writing and reading assessment parts, while the inverse was true for the Germanic and Spanish-language groups. This may be interpreted as evidence that vocabulary plays a particular role in language tests, dependent on whether the languages in an individual's linguistic repertoire are related to the test language or not. Alderman and Holland (1981) studied the same test data in an early DIF study. Not surprisingly, across the seven language groups, they identified seven eighths of all items as showing significant DIF. They also found overall larger differences between the results of the language groups with respect to

the reading and vocabulary section (Alderman and Holland 1981, pp. 21, 27). Interestingly, based on an analysis of data from a second test round, the authors cautioned against over-interpreting this result and those of Swinton and Powers: in the second test administration, the correlations between the item difficulties of the various language groups on the reading and vocabulary test section were higher than in the first test administration, which the authors attributed to “variation in test form” (ibid., p. 18). As in other DIF studies, a judgmental review of the reasons for the DIF by expert reviewers revealed some obvious effects related to cognates, but otherwise offered little insights (ibid., pp. 27 sqq.).

Elder (1996) analysed DIF in foreign language reading and listening assessments across L1 groups of pupils (actually comparing foreign language learners with heritage learners). Based on a detailed analysis of the DIF items found in her study, she advanced several lines of explanation to account for the differences between the linguistic groups: some lexical topics proved easier for heritage learners (e.g. family, household objects, expressions of time), while others advantaged the other learners (e.g. vocabulary related to school). The same was true for some text types and items which required precise understanding of a grammatical structure. In some cases, Elder also identified the need for background cultural or sociolinguistic knowledge, which was easier for heritage speakers. Finally, she also mentioned differences between the test results of heritage speakers and other students with respect to whether an item was written in English, the language of schooling, or in the foreign language, with the latter proving generally easier for heritage learners. Overall, the author concluded that most of the DIF she found could be linked to the higher language proficiency of the heritage learners, i.e. to true differences in ability and to systematic differences in how those learners approached the test items, rather than bias towards certain learner groups. This is strengthened by the observation that the Chinese assessment, which contained many strong DIF items, was clearly unsuited for the Chinese heritage learners of whom many achieved perfect scores. Overall, Elder’s results put into question the use of a language proficiency test for diverse groups of learners, in her case, second- and foreign-language learners.

Focusing on the use of strategic competences in reading assessment, Abbott (2007) looked into the results of two L1 groups of test takers (adult speakers of Mandarin and Arabic respectively) on an English as a Second language (ESL) reading assessment. She compared those to an expert rating of the main strategies that test takers would need to apply to all items: either items would primarily be solved using top-down strategies (i.e. local reading), or they would be solved using bottom-up strategies (i.e. global reading). Her findings indicate that the two groups generally (though not consistently) performed differently on top-down and bottom-up items, which may mean that readers from different backgrounds apply reading strategies in different ways. Hence, if a certain type of item were to be preponderant in a test, some groups of readers would be disadvantaged.

Outside of reading assessment, Brown and Iwashita (1998) focussed on the effect of the L1 in a Computer-Adaptive Test (CAT) of Japanese grammar, comparing the performance of Australian, Chinese and Korean learners. They found that test takers whose L1 lacked a specific grammatical structure were systematically disadvantaged compared to those whose L1 included it. The authors concluded that the use of a CAT with different L1 groups does not meet the criteria for a valid test use as long as the underlying IRT analyses do not take the differing item functioning into account.

Finally, an interesting study of the effect of *cognates* in a lexical proficiency test was conducted by Allen (2019a). In his analysis of the Vocabulary Levels Test (VLT) (N. Schmitt, D. Schmitt, and Clapham 2001) given to Japanese learners of English, he found that cognates were solved more correctly than non-cognates across the entire proficiency continuum, but especially by low-proficiency learners. He also analysed the incidence of English-to-Japanese cognates on the VLT, concluding that there were more cognates on the test than the two languages shared overall (see also Allen 2019b). In terms of validity, this means that the results of the test will probably over-

estimate the lexical proficiency of the Japanese-speaking test takers, given that a disproportionate amount of the words can be solved with the help of their L1. This is true even though Japanese and English do not share a script, so that the test takers could rely only on the phonological similarity of the words (Allen 2019a, p. 1122).

Overall, empirical research on various types of second and foreign language assessment strongly indicates that test takers use the knowledge, skills and abilities related to other languages in their multilingual repertoire to solve test items and assessment tasks. This is independent of whether the assessment situation in some way activates these other languages or not. Where learners from different L1 groups are compared, DIF is often found, or some other indicator of differing approaches to solving the test items. It is not always clear, however, whether these differences can be attributed to inadequate test designs or to heterogeneous make-ups of the test takers' language proficiencies. Given the complexity of the matter, thoroughly discussed in prior chapters, this is not surprising, and validity considerations will have to be made anew for each individual test and group of test takers. For this thesis, however, the influence of other languages on foreign language test results can be considered a given, and will be taken into account to the extent that it is possible.

8.5 Conclusion

This chapter examined quantitative research on item difficulty in reading and listening comprehension assessments, highlighting the diversity of methods and variables used and the uncertainty which remains even when sophisticated statistical analyses are conducted. The results across studies highlight the multidimensional nature of comprehension and item solving, making it clear that beyond linguistics characteristics of input texts, aspects such as the position and redundancy of the NI, overlap between items and the input text, and, particularly, cross-linguistic features play important roles in item solving. In addition, test taker characteristics and cognitive processes involved in test taking need to be investigated further. These findings are in line with the insights on reading comprehension and multilingual assessments mentioned in earlier chapters: the comprehension process is multivariate, complex, and impossible to describe with simple models or cause-and-effect descriptions. The addition of another language adds a layer of complexity which further challenges narrow definitions of comprehension constructs and invites more research, both quantitative and qualitative.

This thesis brings together insights from research on item difficulty and multilingual assessment by examining item difficulty across the four language versions of the ÜGK foreign language assessment in a quantitative study. Particular attention is paid to the role of the different languages involved. The study is outlined in detail in the parts that now follow. The assessment context and the ÜGK assessment itself are introduced in sections 9.1 and 9.2. The selection and operationalisation of the variables used in the analysis are described in section 9.3. Chapter 10 outlines the data analysis process, including dataset preparation, variable extraction, and the statistical procedures applied. The results are presented in part III and discussed in depth in part IV.

Part II

Presentation of the study

Chapter 9

Data

The theoretical and empirical findings reviewed above have underlined the complexity of the cognitive process of reading and its assessment, particularly if more than one language is involved. In many fields, its multilingual nature has been largely ignored or is at least underresearched. This is particularly puzzling in the context of foreign language assessment, which is by definition multilingual. This thesis attempts to shed some light onto this issue: the ÜGK foreign language reading assessment has been administered in four language versions to students whose linguistic repertoires are at least partly known. Their responses and the characteristics of the assessment are analysed in an item difficulty study, with a particular focus on the languages involved in the assessment context.

In the following, this context is presented in more detail in section 9.1. Section 9.2 then describes the foreign language assessment with a focus on the reading part and briefly presents the other two components of the 2017 ÜGK assessment: the reading and orthography test in the language of schooling, and the background questionnaire. This study uses results from both to enrich the analyses.

The variables selected for the item difficulty study are then presented in some depths in section 9.3. Most importantly, table 9.4 (page 150) provides a succinct overview of all variables, their operationalisation and their hypothesised relationship with item difficulty. Chapter 10 will then provide more technical information on how the variables were extracted from the raw assessment data and the test responses.

9.1 Context

Given the complexity of reading comprehension, and the various factors which may influence its assessment, it seems obvious that the context in which the ÜGK reading assessment was administered needs to be taken into account. Any conclusions which can be drawn from the analysis of item difficulty in this thesis have to be considered in this context, and the application of any findings to other assessments can only be sensible to the extent that these share at least some characteristics with the ÜGK foreign language reading assessment. The following sections provide an overview of key aspects of the foreign language learning context, and some important features of the ÜGK assessment. They also include background information about the students who took part in the testing.

9.1.1 Foreign language learning in Switzerland

As has been briefly described in the introduction, the data analysed in this thesis has its origins in Switzerland, a multilingual country which lies at the crossroads between three major linguistic regions in Europe: the majority of the Swiss cantons use German as their official language,

while the southern regions of Switzerland are primarily French- and Italian-speaking¹³⁵. The educational system, along with the language it uses, lies in the jurisdiction of the cantons, although HarmoS represents an effort to establish a tentative national standard which all school systems in Switzerland adhere to. The HarmoS agreement between Swiss cantons (signed 14 June 2007, in force since 1 August 2009) defines certain common objectives and structures of the public school systems in the cantons. Among other implications, HarmoS stipulates the choice and minimal duration of foreign language teaching, the development of curricula which span the language regions, as well as common educational objectives and, most relevant for this thesis, the implementation of educational monitoring (Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) 2010; SPLASH-db.eu 2014).

A Swiss school's *language of schooling* usually corresponds with the official language of the canton or, in the bilingual cantons, the dominant language of its geographical location¹³⁶. Subjects are usually taught in this language and it is assumed that most students have a reasonably good command of it¹³⁷. In addition to this, as per the HarmoS agreement, two other languages (English and another national language) are taught as *foreign languages* during compulsory school. These two languages cannot be chosen by the students; the decision lies with the cantonal school authorities.

Following the implementation of the HarmoS agreement, Switzerland now has four curricular regions with respect to the foreign language policies in the cantons¹³⁸: in Western Switzerland (the *Romandie*), the students in the French-speaking cantons learn German as their first foreign language. The German-speaking cantons which share a border with the French-speaking world (the *Passepartout* region, so called after the foreign language curriculum in effect), also chose to teach their neighbours' language French first, while students in the rest of German-speaking Switzerland learn English instead. Finally, the students in the Ticino canton also study French as their first foreign language (EDK 2018, pp. 93 sqq., see also fig. 9.1). With one exception, all cantons start to teach the first foreign language in grade 5, in the third year of primary school (Zurich: grade 4).

For their second foreign language, the students in the *Romandie* and the *Passepartout* region study English, whereas the students in German-speaking Switzerland who have already started learning English, now study French. The Ticino canton is special in that its students learn both German and English in secondary school. In the context of this thesis, the second foreign language is relevant as most of the students who participated in the ÜGK assessment had already started learning it: during the school year 2016/17, only the students in the cantons Aargau, Appenzell Innerrhoden, Ticino and Uri had not yet started their second foreign language at the time of the assessment (i.e. by grade 8, cf. *ibid.*, p. 108).

9.1.2 ÜGK

The ÜGK foreign language reading assessment was administered as part of the broader language assessment which consisted of two assessments in the language of schooling (reading and spelling), the reading and a listening assessment in the foreign language, and a questionnaire (Konsortium ÜGK 2019a, pp. 51 sq.). The assessment was conducted on normal school days in the classroom, led by a qualified test supervisor under the responsibility of the "Koordinationsstelle Überprüfung

¹³⁵The Grisons canton, which has German-, Romansh- and Italian-speaking municipalities, is a special case both linguistically and administratively, and did not participate in the ÜGK foreign language assessment of 2017. Therefore, its specificities will not be presented here.

¹³⁶The so-called *Territorialitätsprinzip* is, of course, more complicated than this simplified definition. Cf. for a more detailed, albeit somewhat outdated discussion Thürer and Burri (2006).

¹³⁷In fact, in the ÜGK project, students with insufficient knowledge of the test language (i.e. the language of schooling) were exempt from the test if they met certain conditions (cf. Verner and Helbling 2019, pp. 12 sq.).

¹³⁸The situation described here applies to the school year 2016/17, in which the ÜGK assessment was administered. At the time of writing, it has not changed significantly. As before, the Grisons canton, essentially a fifth curricular region, is not included in this description.

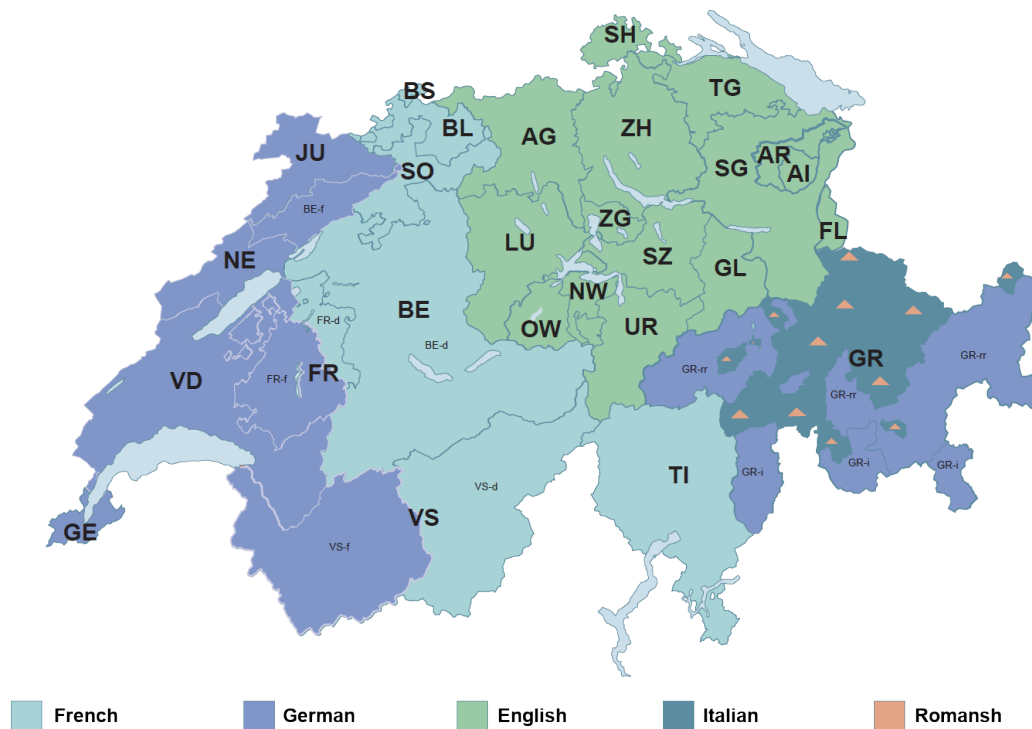


Figure 9.1: Map of Switzerland and Liechtenstein indicating the first foreign language taught at public schools as of the school year 2016/17. Source: EDK 2018, p. 106. *Legend translated.*

des Erreichens der Grundkompetenzen” of the General secretariat of the Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) between April and June of 2017. Including pauses, it took 2 hours and 55 minutes to complete (see table 9.1). All tasks were solved on specifically prepared tablets (with the exception of part of the questionnaire, cf. *ibid.*, p. 51). During the introduction, students were informed about the objectives of the assessment, and instructed on how to use the tablets, including demonstrations of the handling of all item types. Once a test had started, navigation was manual, i.e. test takers went from one item to the next, and from one task to the next by pressing the button *Weiter/Suivant/Avanti*. Between assessments, passcodes had to be entered so that overall, all test takers worked on the same subtests at all times. The test supervisors were present in the room at all times, and helped with technical problems.

<i>Introduction</i>	10 min
Language of schooling: reading	35 min
Language of schooling: spelling	15 min
<i>Break</i>	5 min
Foreign language: reading	25 min
<i>Break</i>	15 min
Foreign language: listening	25 min
<i>Break</i>	5 min
Questionnaire	40 min
Total	175 min

Table 9.1: Schedule of the ÜGK language assessment (2017) as indicated in Konsortium ÜGK 2019a, p. 51.

All four language assessments were developed using the specialised software CBA Item Builder (DIPF and Nagarro 2018b). This software essentially allows the creation of interactive websites which are presented to the test takers in a web browser. During test administration, the related server application EE4CBA (DIPF and Nagarro 2018a) logs all interactions of the individual test takers with the websites according to rules specified during item development. As the logs are stored in a machine-readable XML format, item responses and additional information (e.g., timestamps) can be retrieved using software such as R (R Core Team 2024). For the ÜGK assessment, this processing was carried out by the responsible institutions. The data set analysed in this thesis is based on the output of that extraction.

9.1.3 The test takers

The main objective of the ÜGK assessment was to determine to what extent the pupils at the end of grade 8 have reached the basic competences established by the HarmoS agreement. To achieve this, the assessment was administered to a representative sample of pupils in each canton or, to be more precise, in each of 29 cantonal curricular regions¹³⁹. With the exception of students in a small number of English- or Japanese-speaking private schools, as well as pupils in special education (approximately 1.6% of the entire population, cf. Verner and Helbling 2019, p. 31), all pupils in Switzerland who were in grade 8 in spring 2017 were considered for participation (*ibid.*, pp. 11 sqq.). In smaller cantons, a representative sample of pupils was selected from all schools of the canton, following certain criteria (for more information cf. *ibid.*, pp. 18 sqq.). In larger cantons, a representative sample of schools was selected, from which a sample of students was chosen (i.e. a two-stage sampling procedure was applied, see *ibid.*, pp. 8 sqq.). All psychometric analyses of the results in the ÜGK project were then conducted using weights, to attenuate distorting effects of the sampling procedure and of missing data due to technical problems. The concrete techniques for these analyses is described in more detail in the technical report (*ibid.*), but it is of less importance to this thesis, as the focus of the analyses lies on the items, and not on the test takers and their abilities.

Overall, the data set analysed here is based on an assessment which was administered to 20,177 pupils in Swiss schools, with males and females evenly distributed (10,027 girls). The mean age of all students was 12 years 6 months when the test was administered. Given the varying sizes of the four linguistic regions, the number of participants in each language version differed (cf. table 9.2).

Language version	Curricular region	Number of test takers
DEfr	Romandie	5,484
ENde	ZH, BKZ, East	11,845
FRde	Passepartout	2,025
FRit	Ticino	823
All	Switzerland	20,177

Table 9.2: Number of test takers per language version

¹³⁹The three bilingual cantons Berne, Fribourg and Valais were separated into a German- and a French-speaking part each, which were considered much like cantons of their own in the ÜGK analysis (Verner and Helbling 2019, p. 6).

9.2 The foreign language assessment

As part of the ÜGK assessment, as mentioned before, students encountered a reading and a listening assessment of their first foreign language. The two assessments were developed simultaneously by the same team, following mostly the same procedures and guidelines. The following sections present the assessment tasks in more depth, discussing aspects of their construct, development and administration which are deemed necessary to understand the conditions and results of the empirical study presented in this thesis. An overview is given of the construct and the tasks, with an emphasis on the input texts and items. The section finishes with a description of the item development process. For more information on the ÜGK assessment in its entirety and insights into the reasoning behind certain decisions the reader is referred to the published documentation on the project (e.g. Angelone and Keller 2019a; Konsortium ÜGK 2019a).

9.2.1 The construct

The national educational objectives for the foreign languages defined in the HarmoS project (EDK 2011), and with them the ÜGK foreign language assessment construct, are explicitly based on the action-oriented understanding of communicative competence formulated in the CEFR (Council of Europe 2001, 2018):

Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of **competences**, both **general** and in particular **communicative language competences**. They draw on the competences at their disposal in various contexts under various **conditions** and under various **constraints** to engage in **language activities** involving **language processes** to produce and/or receive **texts** in relation to **themes** in specific **domains**, activating those **strategies** which seem most appropriate for carrying out the **tasks** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences. (Council of Europe 2001, p. 9)

As such, foreign language competence is defined in terms of *what* one is able to do with language, rather than *how* one uses their language skills. Following the CEFR, the HarmoS basic competences are formulated as “Can Do” statements in non-specialist language, which gives an immediate but broad impression of what students ought to be able to do with their foreign languages at the various points of their schooling (Konsortium ÜGK 2019a, p. 36). In reading, at the end of primary school, the expected basic competences in both foreign languages are as follows:

Die Schülerinnen und Schüler können einen sehr kurzen, einfachen Text Satz für Satz lesen und dabei klar formulierte Informationen verstehen – vorausgesetzt es ist möglich, den Text mehrmals zu lesen, und vorausgesetzt, Lexik und Grammatik sind sehr einfach sowie Thema und Textsorte sehr vertraut. Sie können sich bei einfachem Informationsmaterial und kurzen, einfachen Beschreibungen eine Vorstellung vom Inhalt machen. (EDK 2011, p. 16)

The document goes on to give some examples of text genres (e.g. short personal messages, comics, short argumentations) and topics (e.g. school, hobbies, every-day life) and adds some recommendations about language learning strategies and metalinguistic activities related to the multilingual education promoted by the CEFR and the HarmoS project (ibid., p. 16).

This guideline, along with more precise descriptors of the competence levels A1.2 and A2 in other relevant documents (e.g. Bersinger et al. 2005; Lenz and Studer 2007), the specifications in the regional foreign language curricula (Bertschy, Cuenat, and Stotz 2015; Conférence intercantonale de l’instruction publique de la Suisse romande et du Tessin 2010; D-EDK 2016;

DECS 2004) and the contents of the textbooks in use across the various cantons (for references see Konsortium ÜGK 2019a, pp. 71 sqq.), served as the framework for the reading assessment construct: reading comprehension is understood as a problem-solving process defined by elementary, often insufficient linguistic knowledge which requires careful, sometimes repeated reading of carefully selected and adapted texts. All tasks were designed to account for this type of limited understanding: additional information about the reading situation was provided in the language of schooling, and the items requested only partial or very superficial information which students were expected to be able to extract from the reading texts in due time. All input texts are very short (less than 250 words) and there are multiple text genres and topics so as to not disadvantage students who lack language knowledge in specific fields. In addition, the assessment only contains text genres and topics which appear in the majority of the textbooks in use in the various cantons. Finally, in an effort to represent a broad understanding of reading comprehension, i.e. to assure that students did not only look for specific information, but varied their reading style according to the task at hand, the reading tasks were designed to match all four of Weir's types of reading (2005, pp. 89 sq.): test takers were expected to adopt an expeditious or careful reading style, sometimes at the global level of the text and sometimes at a local level, i.e. a single phrase or a few sentences (Konsortium ÜGK 2019a, p. 39, see also section 3.1.4).

9.2.2 The tasks

Each of the 13 reading tasks used in the final ÜGK assessment contains two subtasks which are loosely linked together by a scenario, but otherwise separate (i.e. no understanding of the first task was required for the completion of the second)¹⁴⁰. Each subtask consists of at least one input text, and several items. Most of the tasks use MCQ items with three answer options, but in some cases, more answer options are offered. Some items are in a Matching format, where the input text(s) represent the answer options, while the “item” itself consists only of the stem. A few items use pictures as answer options or in the stem. Table 9.3 lists all tasks as they were used in the main study¹⁴¹.

Each subtask consists of two or more computer pages which are presented to the test takers in order. Each page consists of two major sections (cf. fig. 9.2): the content section on the left, which usually contains the input text, and the item section on the right. In the upper part of each page, a slim section contains a short description of the situation, and the official ÜGK item identifier. The item and the description of the situation are always written in the local language of schooling, while the input text is, obviously, in the foreign language of interest. Colours are used to help distinguish the sections: the content is placed on a yellow background, while the items and the situation are on a blue background. To highlight the turn of page, the items change colours: the first item is always on a darker blue background with the item options written in a lighter box, while for the second item, the hues are inverted (cf. fig. 9.3). The third item uses the same colours as the first, and so on. For the items and all “administrative” text, the font Arial was chosen (question or stem: size 15 bold, answer options: size 15, all other text: size 12), while the input text is printed in Verdana (size 13). All design choices are based on intensive try-outs and

¹⁴⁰During item development, the team had intended to create a somewhat easier first (i.e. A1.2) and a more difficult (i.e. A2.1) second task, but this endeavour was soon abandoned: a prediction of difficulty at this level of precision is challenging in any case (cf. e.g. Urhahne and Wijnia 2021), and it was rendered impossible by the numerous revisions that each item underwent during its development.

¹⁴¹In preparation of the national pilot study, the ADB, the responsible institution, created new identifiers for all items, which encompass both the foreign language assessment and the assessment in the language of schooling. For technical reasons, however, the original IDs remained in the data set and were used to identify the items in the exported results. For this reason, and because the focus is on the test itself rather than on the entire assessment, this thesis uses the original identifiers which consist of the letters *LV* (for German *Leseverstehen*), a two-digit identifier of the task, and the letters *G* or *E* for a subtask expected to assess the expected proficiency level A1.2 (*Grundkompetenz*) or the more challenging level A2.1 (*Erweiterte Kompetenz*). These identifiers are used in table 9.3.

Task ID	Task format	Number of input texts	Number of items
LV02G	Matching	1	5
LV02E	MCQ	1	3
LV03G	MCQ	1	3
LV03E	MCQ	1	3
LV04G	Matching	1	5
LV04E	MCQ	1	3
LV05G	Matching	1	4
LV05E	MCQ & MCQ-pic	2	4
LV06G	Matching	1	3
LV06E	MCQ	1	3
LV07G	MCQ	1	3
LV07E	MCQ	4	4
LV08G	Matching & Matching-pic	1	4
LV08E	MCQ	4	4
LV09G	MCQ-pic	3	3
LV09E	MCQ	4	4
LV10G	MCQ	1	4
LV10E	MCQ	1	4
LV11G	Matching	1	4
LV11E	MCQ	1	3
LV12G	MCQ	1	3
LV12E	MCQ	1	4
LV14G	MCQ	1	3
LV14E	MCQ	1	4
LV15G	MCQ	1	3
LV15E	MCQ	1	3

Table 9.3: Overview of reading tasks in the ÜGK foreign language reading assessment as used in the main study (2017). The supplement *-pic* describes items which contain one of more pictures to convey meaning.

observations of computerised test item development for this and similar target groups in research projects at the Institute of Multilingualism (IoM) (Karges, Barras, and Lenz 2021; Karges and Lenz 2014; Lenz, Karges, Kull, et al. 2019), and it is expected that they were suitable for the large majority of students in the public school system. However, the assessment does not contain specialised fonts for test takers with a dyslexic disorder, and it does not use a colour palette adjusted to the needs of colour-blind test takers. Neither was there an accessible version of the assessment for hearing or visually impaired students.

The ÜGK foreign language reading assessment was developed to be handled without any particular instruction: test takers only needed to be able to tap a radio button or its associated text, and use the *Weiter/Suivant/Avanti*-button to advance to the next page. It was not necessary to choose an answer to go forward in the test. However, once the next page was loaded, it was no more possible to go back.

Nach der Schule chattet Tom mit Marc, seinem englischen Gastbruder. F1L8_dE_006_001

Marc 14:29
Hi Tom :-). Finished school for today? All good?

Tom 14:45
Hi! Yes, finished :-). And we don't have any homework.
I am going back home with your dad now. Where are you?

Marc 14:54
I'm in the park with some friends.

Tom 15:00
What else are you doing today?

Marc 15:07
We're playing football on the sports ground at 4 pm.
Are you coming? It takes 15 minutes to get there from home.

Tom 15:17
Sure! Football, oh yeah :-)! Can I walk there?

Marc 15:23
Great :-)! No, you have to take the bus. It's too far to walk.
You can't take my bike, it's broken.

Tom 15:25
Okay, I'll be on my way in a few minutes.

Marc 15:31
Fantastic... I'm going to win!

Tom 15:40
We'll see about that :-)!

Tom und Marc ...

- ☐ machen zusammen Hausaufgaben.
- ☐ nehmen denselben Bus.
- ☐ spielen um 16 Uhr Fussball.

Content section Item section Weiter

Figure 9.2: Reading item LV03E_1: the content and item section are highlighted.

Nach der Schule chattet Tom mit Marc, seinem englischen Gastbruder. F1L8_dE_006_002

Marc 14:29
Hi Tom :-). Finished school for today? All good?

Tom 14:45
Hi! Yes, finished :-). And we don't have any homework.
I am going back home with your dad now. Where are you?

Marc 14:54
I'm in the park with some friends.

Tom 15:00
What else are you doing today?

Marc 15:07
We're playing football on the sports ground at 4 pm.
Are you coming? It takes 15 minutes to get there from home.

Tom 15:17
Sure! Football, oh yeah :-)! Can I walk there?

Marc 15:23
Great :-)! No, you have to take the bus. It's too far to walk.
You can't take my bike, it's broken.

Tom 15:25
Okay, I'll be on my way in a few minutes.

Marc 15:31
Fantastic... I'm going to win!

Tom 15:40
We'll see about that :-)!

Nach der Schule geht Tom als ...

- ☐ ...
- ☐ ...
- ☐ ...

Weiter

Figure 9.3: Reading item LV03E_2: the second item uses inverted colours for the item. *Note:* The item text has been blurred because the item has not been published.

9.2.2.1 The items

As is indicated in table 9.3, the assessment contains two item formats (MCQ and Matching). The most common task format is an input text accompanied by an MC item with three answer options (e.g. all items of the task LV03E, as exemplified in fig. 9.2). In some cases, more answer options are provided, e.g. when test takers have to choose between a number of newspaper headlines to indicate which one best fits one of four short news texts¹⁴². In most cases, words, expressions or short sentences are given as answer options, though sometimes, pictures are used instead. In the Matching items, the answer options are not part of the item themselves, but sections of the input text. Figure 9.4 shows an example of a Matching item: the item itself does not have answer options, though a picture is used to “ask the question”. To give their answer, test takers had to select a section of the input text, in this case one of several short movie summaries.

Deine Schulklasse besucht das internationale Zeichentrickfilmfestival AniMov in Yverdon. Ihr schaut das Programm an, um einen Film auszuwählen.

F1L8_dE_015_001

The Long Search | Monday, 2:45 p.m.
Alexa's grandfather was an explorer. He has never returned from his last journey, but he wrote a letter to Alexa. Now, Alexa goes looking for him. She travels to the North Pole to look for her grandfather. On her journey through the land of snow and ice, she finds good friends.

Léonie and the Wrong World | Tuesday, 10:15 a.m.
Paris, 1951: Léonie's parents are inventors. An evil scientist has kidnapped them. Léonie has to travel around the world to save them. Go Léonie! Science is counting on you.

Before Judie left | Tuesday, 2:30 p.m.
During her summer holidays, Annie meets Judie. They come from very different families, but the two girls have a lot in common. The movie is based on a book by Joan Reece.

Ghost Soul | Wednesday, 10:30 a.m.
Eleven-year-old Lucy has a special power: she can leave her body and float through the air. When she has to go to the hospital, she meets detective Daniel. Together, they try to catch a dangerous criminal.

Answer options

Content section

Du siehst ein Poster. Zu welchem Film passt es?

Klicke auf den passenden Film.

Item section

Weiter

Figure 9.4: Reading item LV08G_1, annotated: a matching item where the answer options are sections of the input text. In addition, a picture is used to illustrate the question.

All reading items were presented along with their input text. Test takers could supply their answer in their own time, advancing to the next item by themselves. The reading test itself had a time limit (25 minutes), at which the test was cancelled and the next test was automatically loaded. Response analyses indicate that the time limit was amply sufficient: less than 1% of all items were categorised as *not reached*. Test takers were not required to provide an answer to the items, though they were asked to guess rather than not reply at all. As a result, at least 89% of all answers provided for any item are actual answers by test takers.

9.2.2.2 The input texts

To account for the low language proficiency under assessment, all input texts were written specifically for the assessment, though they were usually based on authentic (i.e. existing) texts of the

¹⁴²Here and henceforth, task, text and item identifiers are only used when the item is available to the public. Otherwise, descriptions of tasks and items are kept deliberately vague.

same genre. During item revision, reviewers paid attention to the use of authentic language in all parts of the assessment, though considerations of the test takers language proficiency and the multilingual nature of the assessment were sometimes deemed more important.

As mentioned above, an essential aspect of the assessment construct is the choice of input texts with respect to their genre and topic. To cover a wide range of possible reading situations, a variety of text genres was chosen, ranging from an index or a form to a smartphone chat or short news flashes, and to informative letters, a blog post or a short story. The text genre is indicated in the description of the situation and it is visible to some extent in the layout of the text. With the exception of the pictures used in some items (cf. fig. 9.4), no graphical support was chosen. Only the positioning of the text, very simple formatting option such as text colour, bold type or capital letters were used to distinguish text types¹⁴³. In some tasks, parts of the input text are printed in gray to indicate that they do not need to be read to find the answer to the current item. All input texts fit on one page (average text length: 115 tokens, range: 23-257 across all three languages), some are continuous (e.g. letters, stories, reviews), while others represent non-continuous text genres such as forms, or lists of contents. In some cases, e.g. in a task which contains tweets, or in a smartphone chat, the distinction is not quite as clear: short sections of these texts can probably be considered continuous, but the input text as a whole does not present as a unit.

9.2.2.3 The language versions

Each input text exists in English, German and French, while the description of the situation and the item are formulated in the three languages of schooling German, French and Italian. Given the four curricular regions in the ÜGK assessment, there are four language versions: two tests of French with German and Italian items, a test of German with French items, and a test of English with German items (table 1.1, page 20). Since the four language versions were developed alongside each other, and continually edited throughout the development process, there is no original language version. Instead, efforts were made to ensure that the language versions remain as similar to each other as possible, while making sure that all input texts would still be considered linguistically correct in their respective languages.

Two concessions were made with respect to the input texts: first, in some cases, names and situations were adapted to match the combination *language of schooling–foreign language*. For instance, in task LV03E (see fig. 9.3), the names of the two chat participants were changed to match their supposed origin: in the German test conducted in the French-speaking *Romandie*, Alain is visiting Germany, and talking to Marko, while in the English test for the German-speaking Eastern Switzerland depicted in fig. 9.3, Tom is visiting England, and talking to Marc. With the exception of these names, the texts are identical in their content. Second, minor differences between the input texts were allowed when it made individual language versions either more idiomatic, more authentic or, inversely, when it was deemed that a particular formulation would be easier to understand for foreign language learners. Also, after the pre-operational testing of all items in 2016, some changes were made to only one language version when items were particularly easy or difficult in this language version, but not in the others. This was particularly true for some English tasks, where certain cognates and internationalisms were replaced by less well-known terms. Similar changes were not made to the other language versions.

The three language versions of the items were kept comparable in the same way, though more leeway was left to allow for idiomatic language since it was assumed that the test takers were more familiar with the language of schooling. Also, in some cases, differences between language versions were accepted to observe the basic requirements of good item writing (e.g. all answer options should have roughly the same length, all should have the same grammatical structure, or each a different one).

¹⁴³There is one exception.

Despite these particularities, the four language versions are quite similar to each other, their formatting is virtually identical, and notable elements such as pictures, words in other languages, brand names and numbers were deliberately kept stable to ensure a maximum of comparability beyond the languages themselves. Hence, differences in item difficulty between the language versions are most probably due to linguistic characteristics of the texts and items in the four language versions rather than divergent formatting or obvious translation errors.

9.2.2.4 The language of schooling in foreign language assessment

The items used in the ÜGK assessment were presented to the students in their language of schooling. Although this is somewhat unusual in the context, it was a conscious design decision based on results from prior research project involving the target group (the Task Lab study, cf. Karges, Barras, and Lenz 2021). In that study, roughly 630 students solved computer-based reading comprehension tasks in their first foreign language French. These tasks were developed in four versions: in an MCQ and a SAQ format, each with the items in the language of schooling and in the target language. Each student solved a subset of these items (see for analyses of the task format Lenz, Karges, and Barras 2019).

In the subsequent analysis, and supported by a review of related research studies (cited in Karges, Barras, and Lenz 2022, pp. 85 sqq.), a clear preference for the language of schooling could be identified: not only was this preferred by a large majority of the students, but an analysis of the answer patterns and observations from stimulated recall interviews showed that the items in the target language were more vulnerable to answering processes which were only indirectly related to reading comprehension (e.g. guessing or the use of test-wiseness strategies, cf. *ibid.*, pp. 105 sqq.). In addition, during the initial ÜGK field tests conducted by the item development team, the issue was investigated again, by administering a short questionnaire in class. Here, too, an overwhelming majority of students in all curricular regions reported to be in favour of using the language of schooling, or did not care either way (*ibid.*, pp. 100 sq.).

9.2.3 Other components of the ÜGK 2017 assessment

As has been mentioned before, the foreign language assessment was administered alongside two assessments of the language of schooling (reading and spelling) and a questionnaire (section 9.1.2, see also Konsortium ÜGK 2019a, pp. 51 sqq.). All three were developed as part of the ÜGK project by various educational institutions across Switzerland (*ibid.*, pp. 133 sq.), with a small number of the questionnaire items relevant to foreign language learning provided by the item development team of the IoM. The reading assessment in the language of schooling is similar to the foreign language reading assessment in that it was presented in the same general format on computer screens and also used a MCQ format: a short input text is followed by two to four questions with four answer options each. The input texts were categorised as narrative, informative or argumentative, and the items as eliciting global, explicit or implicit information, with incorrect answer options clearly contradicted by the input text (Angelone and Keller 2019b, p. 5; Konsortium ÜGK 2019a, pp. 20 sq.). The so-called context questionnaire, administered to all test takers at the end of the testing session, also in a computer-based format, elicited factors which had been identified to be influential for school success in prior studies (see for references and more detailed information Erzinger, Hauser, et al. 2019, pp. 6 sqq.): contextual aspects (e.g. family, school), characteristics of the learning situation, and individual prerequisites for learning (e.g. motivation, self-efficacy). The objective of the questionnaire was to “möglichst genau erklären [...], weshalb einige Schülerinnen und Schüler die Grundkompetenzen erreichen und andere nicht” (*ibid.*, p. 7).

The results of the entire test administration are available via the Swiss research data base SWISSUbase¹⁴⁴. In this data set, the results of the language assessments are in the form of

¹⁴⁴As of 2025-08-09: <https://www.swissubase.ch/de/catalogue/studies/12954/18253/overview>.

imputed PVs which represent the competence of each test taker on the respective assessment, while the questionnaire data is available as raw answers. This thesis makes use of the PVs for the assessments in the language of schooling, the foreign language listening assessment as well as some responses from the context questionnaire to create a few selected test taker variables. The item characteristics and concrete test item responses from the foreign language reading assessment are not available to the greater public and were provided separately, for the purpose of this thesis.

9.2.4 Item development

Item development¹⁴⁵ was conducted at the IoM in 2015 and 2016 by a small multilingual team under the leadership of Peter Lenz. The team included myself and three other team members. Three people, including myself, were native speakers of German or Swiss German, while one person was a bilingual speaker of French and English who grew up in Western Switzerland. All team members had at least a B2 proficiency in German, English and French and B1 receptive knowledge of Italian. Between all team members, there was at least one who had a C-level or native language competence in all four languages involved in the project. All team members had completed linguistic studies at a university in Switzerland. The team leader and two team members had prior experience in the development of large-scale foreign language assessment items, while a third team member had already collaborated as a research assistant in an assessment research project before joining the ÜGK development project.

The four language versions were developed alongside each other in an iterative process¹⁴⁶. After an analysis of the textbooks that were in use in the various curricular regions at the time, the team identified topics and text types which had probably been seen by all or most of the students. All team members then chose topics and text genres based on this list to draft a number of initial tasks. To avoid linguistic bias, some tasks were first developed in German, others in French, and some in English, independently of the native language of the team members. Once a draft was finished, a second member of the team reviewed it thoroughly. The original author then reworked the first draft, often substantially. This version was again reviewed by a member of the team. After that, a second language version was created, which often led to an adaptation of the original version. All available tasks were discussed and reviewed at various points in sessions involving the entire team. Once a task was deemed to be of good quality, it was implemented as a computer-based task in one or two language versions using the software CBA ItemBuilder (DIPF and Nagarro 2018b). This task was then available for initial piloting. For this phase, the input texts of the listening tasks were recorded as drafts using a voice recorder and some rough audio processing in Audacity (Audacity Team 2014).

Due to limited resources and because members of the team had just conducted a research project on foreign language reading comprehension with the target group (Karges, Barras, and Lenz 2021), only the listening comprehension tasks were piloted in classroom settings during task development. The piloting sessions were conducted in three of the four curricular regions: the *Passepartout* region (language of schooling: German, first foreign language: French), the Ticino (language of schooling: Italian, first foreign language: French), and Western Switzerland (language of schooling: French, first foreign language: German). During these sessions, interviews were conducted with individual students in so-called “coglabs”: while or just after solving the tasks, they were asked to explain their reasoning to an interviewer. Meanwhile, the rest of the class worked on the tasks under conditions which were similar to the actual test administration:

¹⁴⁵Unfortunately, there is no “official” account of the ÜGK 2017 item development process. Hence, this description was created by myself based on memory and remaining documents. It was reviewed by Peter Lenz prior to its publication.

¹⁴⁶Hence, the assessment was not translated in the traditional sense, but instead created in multiple source versions at once. This has been described as perhaps the best way to create a multilingual assessment instrument (e.g. Bolaños-Medina and González-Ruiz 2012, p. 733).

after a small introduction, each student worked on the tasks on their own on an individual laptop. Overall, 522 students participated in these sessions. Based on the results from the classroom pilots and observations from the audiotaped interviews, additional changes were made to the tasks. All tasks were then implemented as computer-based tasks with CBA ItemBuilder (DIPF and Nagarro 2018b) in the four language versions. At this point, the listening comprehension texts were recorded in a professional recording studio at the University of Teacher Education Fribourg.

After several quality control checks, 15 reading and 15 listening comprehension tasks in four language versions, each featuring between three to five items, were delivered to the responsible institution (then the Aufgabendatenbank der EDK (ADB)) for further review. Of these, the ADB eventually piloted 13 tasks each in all four curricular regions along with the other language assessment tasks (L1 reading and orthography). This pilot test was conducted under near-test conditions and on a representative national sample. Based on psychometric analyses of the results (particularly regarding item fit and DIF), some minor adjustments were made, individual items deleted, and in some listening tasks, the second run of the audio file was eliminated. During spring 2017, the tasks were administered to representative samples of students in all Swiss cantons. The data collected in this main data collection has been made available for the item difficulty analysis in this thesis.

9.3 The variables

This section describes all variables that were extracted from the various data sets. Table 9.4 lists all variables and gives some rough information on their operationalisation as well as some indication of what their relation to the reading comprehension construct may be. The table as well as the following sections separate the variables into the following groups: 1) the *dependent variable* item difficulty which was estimated from the test responses, 2) so-called *linguistic variables* which operationalise the linguistic characteristics of the various texts in question, i.e. the input text and the items, 3) text variables, which describe some characteristics of the input texts themselves, 4) item variables which represent specific characteristics of the items, 5) text-by-item variables which intend to describe aspects of the interaction between input text and item, and 6) test taker variables, extractions from the ÜGK questionnaire and PVs of the other assessments.

The variables mentioned in table 9.4 were initially determined as described in this section. Where they failed to describe the data in meaningful ways, or where important collinearity was found, the variables were eliminated prior to the statistical analyses. They are mentioned here for completeness.

In the following, the rationale for and operationalisation of each variable is discussed in more depth. Additional information on how exactly the variables were created can be found in the respective R scripts. See section 10.1 for more information.

Variable ID	Variable	Description	Hypothesised impact: <i>Item difficulty is higher if ...</i>	Relation to construct
<i>Dependent variable</i>				
ItemDiff	Item difficulty	Item difficulty from a language-by-item many-facet Rasch model		
<i>Linguistic variables</i> ¹⁴⁷				
TokenCount	Token count	Length of the input text and item parts in number of tokens	there are more tokens in the input text or item.	Longer texts are more difficult to process.
SentenceLength	Mean sentence length	Mean length of the sentences in the input text in number of tokens	the sentences in the input text are longer.	Longer sentences are more difficult to process.
POS_N	Incidence of nouns	Relative frequency of nouns in the input text and item parts (per 100 tokens)	the incidence of nouns is higher.	Texts which contain more nouns are more difficult to process.
POS_V	Incidence of verbs	Relative frequency of verbs in the input text and item parts (per 100 tokens)	the incidence of verbs is lower.	Texts which contain more verbs are easier to process because they are potentially of a more narrative type.
POS_Ad	Incidence of adjectives and adverbs	Relative frequency of adjectives and adverbs in the input text and item parts (per 100 tokens)	the incidence of adjectives and adverbs option is higher.	Texts which contain more adjectives and adverbs are more difficult to process.
POS_Pro	Incidence of pronouns	Relative frequency of pronouns in the input text and item parts (per 100 tokens)	the incidence of pronouns option is lower.	Texts which contain more pronouns are easier to process because they will also contain more verbs.
POS_NAM	Incidence of names	Relative frequency of names in the input text and item parts (per 100 tokens)	the incidence of names option is lower.	Texts which contain many names are easier to understand because they are potentially of a more narrative type.
LingCompl_Neg	Incidence of negations	Relative frequency of negations in the input text and item parts (per 100 tokens)	the incidence of negations option is higher.	Texts which contain more negations are more difficult to process.

¹⁴⁷If not indicated otherwise, all variables apply to the input text (_T), the Necessary Information (_NI), the item stem (_Q) and each answer option (_AO).

Variable ID	Variable	Description	Hypothesised impact: <i>Item difficulty is higher if ...</i>	Relation to construct
LexDiv_Var	Lexical variability	MTLD with TTR threshold of 0.72 (input texts only)	lexical variability (=MTLD) is higher.	Texts with a higher lexical variability are more difficult to process.
LexDiv_Eve	Lexical evenness	Standard deviation of number of tokens per lemma	<i>(no hypothesis)</i>	<i>Since this measure has not been used in item difficulty research yet, a prediction is difficult.</i>
LexDispersion	Lexical dispersion	Average amount of repeated tokens and lemmata within 20 words (Vanhove, Bonvin, et al. 2019, p. 509)	<i>(no hypothesis)</i>	<i>Since this measure has not been used in item difficulty research yet, a prediction is difficult.</i>
LexDisparity_form	Formal lexical disparity	Mean Damerau-Levenshtein distance between unique lemmata	<i>(no hypothesis)</i>	<i>Since this measure has not been used in item difficulty research yet, a prediction is difficult.</i>
LexDisparity_sem	Semantic lexical disparity	Mean number of words which share the same semantic sense as indicated by WordNet's semantic sense index	<i>(no hypothesis)</i>	<i>Since this measure has not been used in item difficulty research yet, a prediction is difficult.</i>
LexDens	Lexical density	Relative frequency of content words (nouns, verbs, adjectives, adverbs) per 100 words	the proportion of content words is higher.	Texts with a higher lexical density are more difficult to process.
LexSoph_tokens	Lexical sophistication: all words	Mean log word frequency of all words based on SUBTLEX	the frequency of the words is lower.	Texts with many low-frequency words are more difficult to process.
LexSoph_content	Lexical sophistication: content words	Mean raw word frequency of all content words based on SUBTLEX	the frequency of the words is lower.	Texts with many low-frequency words are more difficult to process.
LexSoph_low_freq	Lexical sophistication: low-frequency words	Mean log minimum word frequency of words in sentences based on SUBTLEX	the frequency of the words is lower.	Texts with many low-frequency words are more difficult to process.
LexDiff_Deca_A1	Number of A1 content words in the input text	Raw number of content words that were classified as A1 in the Decaleon data base.	there are less A1 content words.	Texts with more unknown words are more difficult to process.

Variable ID	Variable	Description	Hypothesised impact: <i>Item difficulty is higher if ...</i>	Relation to construct
LexDiff_CEFRlex_A1	Number of A1 content words	Raw number of content words in the input text that were classified as A1 in CEFRlex.	there are less A1 content words.	Texts with more unknown words are more difficult to process.
LexDiff_Deca_A2	Number of A2 content words	Raw number of content words in the input text that were classified as A2 in the Decaleon data base.	there are more A2 content words.	Texts with more unknown words are more difficult to process.
LexDiff_CEFRlex_A2	Number of A2 content words	Raw number of content words in the input text that were classified as A2 in CEFRlex.	there are more A2 content words.	Texts with more unknown words are more difficult to process.
CL_Cogn	Proportion of cognates	Proportion of transparent cognates with language of schooling/-foreign language among content words (human rating)	there are less cognates.	Texts which contain more cognates with known languages are easier to understand. Items which contain cognates with the foreign language are easier to solve.
CL_Lev	Mean Damerau-Levenshtein distance	Mean Damerau-Levenshtein distance of all content words with their direct translation equivalents in the other language.	the linguistic distance between the content words is higher.	Texts which contain more closely related words are easier to understand. Items which contain cognates with the foreign language are easier to solve.
<i>Text variables</i>				
ContinuousText	Text continuity	Relative amount of continuous input text (human rating)	the text is more continuous.	Non-continuous texts are easier to process.
TextFunction	Text genre	Primary text function (argumentative, expository, narrative, communicative, based on Neumann 2012, p. 144) (human rating)	the text is less expository in nature.	Less familiar text genres are more difficult to process.
Topic	Closeness to real-world situations	Degree to which the input text relates to a real-world situation	the text is more distant from the real world.	Familiar, every-day situations are easier to understand.

Variable ID	Variable	Description	Hypothesised impact: <i>Item difficulty is higher if ...</i>	Relation to construct
<i>Item variables</i>				
ItemType	Item type	Item type (MCQ with 3 answer options, pictures as answer options, or <i>other</i>)	the item is a matching item or an MCQ with more than 3 answer options.	More answer options are more difficult to process. Answer options with pictures rely on a different type of processing.
StemType	Type of the item stem	Is the item stem a question or an incomplete sentence?	the item is a stem.	Questions are easier to process.
CorrPos	Position of the correct answer option	Position of the correct answer option in MC or Matching item	the correct answer is not the first.	If the correct answer is in the first position, it is selected more often.
<i>Text-by-item variables</i>				
RespProcess	Response process: passage/question relation	Categorical variable which indicates the passage/question relation of the item (Ozuru et al. 2008) (human rating)	more inferencing is needed.	If the passage/question relation is less direct, the answer is more difficult to find.
BackgroundKn	Need for background knowledge	Degree to which background knowledge is necessary to answer the item correctly (human rating)	the answer requires more background knowledge.	An item is more challenging if it requires background knowledge in addition to reading comprehension skills.
Fals	Falsifiability of the incorrect answer options	Number of incorrect answer options which can be falsified using the text (human rating)	less incorrect answer options can be falsified using the text.	If incorrect answer options are falsifiable, the correct answer option can be found more easily.
NI_Dist	Distribution of the NI	Categorical variable which indicates how the NI is distributed in the input text (human rating)	the NI is scattered over the input text.	If more text has to be processed and/or an inference is necessary, reading comprehension is more difficult.
Overl_sem_NI	Semantic overlap between item and NI	The proportion of content words in the item which are direct translations of words in the NI	there is less semantic overlap between the NI and the item.	Overlap between the item and the NI has been shown to lead to easier items.
Overl_sem_T	Semantic overlap between item and input text	The proportion of content words in the item which are direct translations of words in the input text	there is more semantic overlap between the input text and the item.	Overlap between the item and the input text may divert attention from the NI.

Variable ID	Variable	Description	Hypothesised impact: <i>Item difficulty is higher if . . .</i>	Relation to construct
Overl_form_NI	Formal overlap between item and NI	The proportion of content words in the item which are orthographic neighbours of words in the NI	there is less formal overlap between the NI and the item.	Overlap between the item and the NI has been shown to lead to easier items.
Overl_form_T	Formal overlap between item and input text	The proportion of content words in the item which are similar to words in the input text	there is more formal overlap between the input text and the item.	Overlap between the item and the input text may divert attention from the NI.
Overl_dist_NI	Linguistic distance between item and NI	The mean Damerau-Levenshtein distance of content words in the NI and the item	the linguistic distance between the NI and the item is small.	Texts which contain more similar words are easier to comprehend.
Overl_dist_T	Linguistic distance between item and input text	The mean Damerau-Levenshtein distance of content words in the input text and the item	the linguistic distance between the input text and the item is large.	Overlap between the item and the input text may divert attention from the NI.
<i>Test taker variables</i>				
ImmSt	Immigration status	Migratory background of the test taker.	test takers have a migration history.	Items in the language of schooling are more difficult to process for students who have a history of migration.
L2Home	Use of the foreign language at home	The test taker uses the foreign language outside of school.	test takers are less familiar with the foreign language.	Test takers who use the foreign language outside of school will score higher.
L2List	Listening comprehension ability in the L2	PV of the ÜGK foreign language listening assessment	test takers score lower on the listening assessment.	Test takers who score higher on the listening assessment will also score higher on the reading assessment.
L1Read	Reading comprehension ability in the L1	PV of the ÜGK L1 reading assessment	test takers score lower on the L1 reading assessment.	L1 literacy skills influence L2 reading comprehension.
L1Orth	Orthography skills in the L1	PV of the ÜGK L1 orthography assessment	test takers score lower on the L1 orthography assessment.	L1 literacy skills influence L2 reading comprehension.

Table 9.4: Overview of predictor variables, hypothesised impact of item difficulty and relation to construct

9.3.1 Linguistic variables

As has been laid out in section 4.1, the characteristics of a reading text have been shown to influence reading comprehension and, as a result, reading item difficulty. These characteristics can be operationalised in terms of their topic and overall structure, and with respect to the complexity of the language used in the text. The latter can be operationalised in a variety of ways, ranging from human ratings (e.g. Carr n.d., p. 12) to applying a multitude of automatically retrieved NLP variables which capture a specific linguistic feature each (e.g. McNamara, Graesser, et al. 2014; Weiss, X. Chen, and Meurers 2021; Weiss and Meurers 2018). In this thesis, given the low proficiency level of the reading assessment and its multilingual nature, a compromise was chosen: the focus lies on operationalisations of lexical and cross-linguistic characteristics of the input texts and items to account for the assumed importance of lexical knowledge in low-proficiency foreign language reading. In addition, some variables which have proven to be particularly important predictors in item difficulty research were selected. In the following, these variables are discussed in more detail, and the concrete operationalisation of these variables is explained. In most cases, the same variables were applied to the input text, the NI and to the various parts of the item, though they are only described once. Variables which could not be meaningfully applied to extremely short pieces of text (e.g. answer options consisting of parts of sentences or even single words) were only used with respect to the input text.

Most of the variables described below were extracted from the input texts and items by way of POS-tagged texts created by TreeTagger (Schmid 1994) via the `koRpus` package (Michalke 2021) in R (R Core Team 2024). TreeTagger automatically parses language into words and sentences, distinguishes punctuation and other symbols, and tags each word with a POS, i.e. whether it is a noun, verb, conjunction etc. TreeTagger uses specialised tag sets for each language¹⁴⁸ and tags language according to probabilistic principles using decision trees (Schmid 1994). The `koRpus` package calls the underlying TreeTagger application, and makes use of its output to count and calculate a variety of linguistic measures. It also supplies a function to change individual POS-tags manually after TreeTagger has finished. This was used to correct instances where proper names were not recognised as such, specific tokens were falsely tagged in certain instances (e.g. French adjectives which were sometimes tagged as past participles of verbs), and to tag foreign words and non-words (e.g. web addresses) correctly. A more detailed correction of the tagged texts, i.e. of individual words in individual texts, was not conducted due to time constraints. Nevertheless, based on prior empirical evaluations, it can be assumed that the large majority of all tags is correct (see e.g. Allauzen and Bonneau-Maynard 2008, p. 3770; Giesbrecht and Evert 2009).

9.3.1.1 Descriptive variables

In various item difficulty studies (e.g. Freedle and Kostin 1993; Kostin 2004; Nissan, DeVincenzi, and Tang 1996), some purely descriptive characteristics of the input text and items have been found to be related to item difficulty. Four of these have been included in these analyses: text length (`TokenCount`) and sentence length (`SentenceLength`¹⁴⁹), the incidence of nouns, verbs, pronouns and adjectives/adverbs¹⁵⁰ (`POS_N`, `POS_V`, `POS_Pro` and `POS_Ad`), and the incidence of

¹⁴⁸The TreeTagger homepage provides parameter files for each language used in this thesis (<https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>). All parameter files were downloaded last on 2023-05-07. At that time, the German parameter files were based on the Stuttgart Tübingen Tag Set (STTS), the French and the Italian parameter files had been specifically developed for TreeTagger, and the English parameter files used a British National Corpus (BNC) tag set. Please note that the parameter files provided on the TreeTagger homepage are sometimes updated without notice, which may lead to different tagging results if parameter files are downloaded at another point in time.

¹⁴⁹Since the items rarely consist of more than one sentence, the average sentence length mostly coincides with the token count. For that reason, this variable is only applied to the input text.

¹⁵⁰Adjectives and adverbs have been combined because both word classes were expected to be rare in the low proficiency input texts.

negations (`LingComp1_Neg`). These variables all operationalise aspects of the complexity of texts: if test takers have to process longer texts, if sentences are generally longer, if there are more negations, and if texts contain more adjectives and adverbs as opposed to nouns and verbs, text comprehension can be expected to be more demanding (see also sections 3.2.3 and 4.1). In addition, the incidence of proper names, which covers mostly names of people or places, was measured (`POS_NAM`): a notable characteristic of reading at A1 or A2 of the CEFR is understanding very concrete and highly familiar information, and recognising names and “shared international vocabulary items” (Council of Europe 2018, p. 54). Hence, an input text which contains a high proportion of proper names is hypothesised to be easier to understand.

Token count and *mean sentence length* were extracted automatically from the POS-tagged text. The tagging automatically distinguishes words from punctuation symbols. Abbreviated forms such as *j’ai* or *don’t* are separated into their components and counted as two tokens. Compound words are considered to be one token if they contain a dash (e.g. *Fussball-Weltmeisterschaft* or *Etats-Unis*), but are otherwise not recognised (e.g. the lexical unit¹⁵¹ *flamant rose* is counted as two words). Numbers are counted as words as well, and usually, the tagging recognises dates and decimal numbers as such, i.e. the dates *30.11.2022* or *07/03/1922* are counted as one token only.

Based on the POS-tagging provided by TreeTagger, the `koRpus` package outputs information on the number of various *parts of speech* in a tagged text, including nouns, verbs, pronouns, adjectives and adverbs, and proper names. After some manual corrections to the POS-tagging, these numbers were extracted and normalised to provide a relative frequency per 100 words in order to allow a comparison of these values across texts and language versions.

Finally, *negations* were coded manually, mostly following Carr (n.d., p. 16). He identified negations as explicit negating words (*not*, *nicht*, *(ne) ... pas* etc.), negating adverbs and negative determiners, as well as negative prefixes and suffixes¹⁵². Here, the two human raters additionally used the following principles: in French, the adverb *ne* was not counted separately, as it only occurs along with another word of negation, e.g. *pas* or *jamais*. In Italian, the same logic was applied where *non* appears in conjunction with another word of negation (e.g. *mai*). Inter-rater agreement in all four languages was generally high, disagreements were mostly due to oversights rather than actual differences in rating. Once identified, the incidence of negations was normalised to 100 words.

9.3.1.2 Lexical diversity and lexical density

Lexical diversity was operationalised following Jarvis (2013a,b, see also section 5.2.1) for several reasons. First, his reasoning in favour of a multivariate construct is persuasive considering the complexity of many of the concepts discussed in this thesis. It is also in line with the concept of lexical complexity brought forth by Bulté and Housen (2012). Second, empirical studies have shown the various lexical diversity measures suggested by Jarvis are related to human judgments of lexical richness (Jarvis 2017; Kyle, Crossley, and Jarvis 2021; Vanhove, Bonvin, et al. 2019). Third, while individual lexical diversity indices have been found to predict item difficulty to some extent (e.g. TTR: Andre A. Rupp, P. Garcia, and Jamieson 2001, p. 199; Aryadoust and C. C. M. Goh 2014, pp. 15 sqq., *D-value*: Révész and Brunfaut 2013, pp. 52 sq.), the more complex construct definition suggested by Jarvis has, to the best of my knowledge, not yet been investigated in the context of item difficulty. Hence, this study presents a good opportunity to further investigate both the operationalisation of lexical diversity and its relationship with item difficulty in reading assessment.

Apart from lexical volume, which is expressed by the descriptive variable `TokenCount`, and

¹⁵¹See footnote 33, page 42.

¹⁵²Carr (n.d., p. 16) actually distinguished these various types of negations in his rating instrument. This is not replicated here, since the overall number of negations across all texts is small, and with one or two exceptions, only negating words were used.

lexical rarity, which is operationalised as lexical sophistication and lexical difficulty (see below, section 9.3.1.3), four of Jarvis’s subconstructs were included in this analysis: lexical variability (*LexDiv_Var*), expressed as the MTL¹⁵³, lexical evenness (*LexDiv_Eve*), lexical dispersion (*LexDispersion*) and formal lexical disparity (*LexDisparity_form*¹⁵⁴). The MTL¹⁵³ was calculated by the respective native *koRpus* function, while lexical evenness, dispersion and formal disparity were determined as suggested by Vanhove and colleagues (2019). In all cases, the corrected POS-tagged text provided by TreeTagger through the *koRpus* package underlay the calculations. The lexical disparity index excluded the five most frequent tokens of each language, based on the reference corpora described below, in section 9.3.1.3. Table 9.5 describes the operationalisation of all lexical diversity indices in some detail, excluding the measures of lexical sophistication, which can be found in table 9.6.

In addition, a typical measure of lexical density (the percentage of content words in the text, *LexDens*) was added to the analyses since studies of reading comprehension have shown that texts which are informationally more dense tend to be more difficult to comprehend (see sections 4.1 and 5.2.2 for references).

9.3.1.3 Lexical sophistication and difficulty

Though few item difficulty studies focused on lexical aspects, those which did usually revealed a relationship between item difficulty and lexical sophistication or difficulty (e.g. Révész and Brunfaut 2013, pp. 50 sqq.; Brunfaut and Révész 2015, pp. 156 sq.; Nissan, DeVincenzi, and Tang 1996, pp. 26 sq.; Kostin 2004, pp. 15 sq.). Révész and Brunfaut (2013, p. 54) also note that in their stimulated recall interviews, “[all] nine participants reported lexis as a source of difficulty. [...] except for one learner, this category was the most frequently mentioned by the participants.”, which underlines the importance of understanding the *words* in (listening) comprehension, even at the high proficiency level under investigation in this particular study (C1, *ibid.*, p. 42).

The *lexical sophistication* of texts can be measured in a variety of ways, as was laid out in section 5.2.3.1. For this thesis, the frequency of the words in the reading texts was measured as proposed by Coh-Metrix (*Coh-Metrix version 3.0 indices* 2023): the mean log word frequency of all words (*LexSoph_tokens*), the raw word frequency of all content words (*LexSoph_content*), and a mean log frequency of the least frequent word in each sentence of a text (*LexSoph_low_freq*), see also table 9.6. These operationalisations have been chosen since Coh-Metrix has been widely used and studied extensively, at least in English-language contexts. In all cases, a normalised raw frequency of tokens per million words was used. However, differing from Coh-Metrix’ suggestion, the frequency lists on which these measures are based stem from the SUBTLEX family (Brysbaert, Buchmeier, et al. 2011; Brysbaert and New 2009; New et al. 2007)¹⁵⁵. As the name suggests, these frequency lists are based on subtitles, and have been shown to be more strongly related to behavioural data than lists based on written texts (Brysbaert, Buchmeier, et al. 2011, p. 421, see also D. J. Davidson, Indefrey, and Gullberg 2008, p. 137). This implies that they may also be more adequate for lexical sophistication analyses in foreign language contexts, especially given that foreign language learning at the lower proficiency levels usually focuses on every-day

¹⁵³The MTL¹⁵³ is a measure of lexical variability which has been shown to be relatively independent of text length, even for short texts (McCarthy and Jarvis 2010). This is a useful characteristic for this study given the short texts used in the low proficiency ÜGK assessment.

¹⁵⁴*Semantic* lexical disparity was not measured, as the necessary reference data, e.g. WordNet (Fellbaum 2005; Princeton University 2010) in at least the three foreign languages (English, French, German), was not freely available at the time of writing. Also, it is not entirely clear to what extent this type of data would be comparable between languages.

¹⁵⁵Though French was not listed on the website which provides access to the SUBTLEX frequency lists (<http://crr.ugent.be/programs-data/subtitle-frequencies>, accessed on 2023-04-16), the French subtitle frequency list used in the *Lexique* database is clearly related: some of the authors are the same as for other SUBTLEX lists, and the *Lexique* corpus has been used in comparisons of frequency lists which otherwise focused on SUBTLEX frequency lists (e.g. Brysbaert, Buchmeier, et al. 2011, p. 414).

Lexical Diversity Variable	Variable ID	Operationalisation
Lexical volume	TokenCount	The raw number of tokens in a text is counted. <i>Interpretation:</i> raw count
Lexical variability	LexDiv_Var	The MTLTD value of the text, with a default factor size of 0.72, is calculated. See McCarthy and Jarvis (2010, pp. 384 sq.) for more details. <i>Interpretation:</i> a number ranging primarily between 50 to 150 (with outliers below and above). Higher values mean more lexical variability.
Lexical evenness	LexDiv_Eve	For each lemma, the number of tokens in the text is counted. Then, the standard deviation of these counts is calculated. <i>Interpretation:</i> a number ranging primarily between 0.5 and 2.5. Higher numbers should mean less diversity: some lemmas are repeated more often while others are not.
Lexical dispersion	LexDispersion	For each token, the number of repetitions within ranges of 20 words is counted. The sum of all repetitions is divided by the total length of words. The five most frequent tokens in each language are not taken into consideration. <i>Interpretation:</i> a number ranging primarily between 0 and 0.3 with a mean around 0.1. Higher numbers should mean less diversity.
Formal lexical disparity	LexDisparity_form	For each lemma in a text, its Damerau-Levenshtein distance with all other lemmas is calculated. Then, the mean of these distances is calculated. Each lemma is considered only once. <i>Interpretation:</i> a number ranging from 0 to 1. A higher number means lemmas are more similar to each other, i.e. less lexical diversity.
Semantic lexical disparity	(none)	<i>This subconstruct was not measured.</i>

Table 9.5: Operationalisation of the lexical diversity measures in this thesis. Note that the term *text* may refer to the input text, the NI or parts of the items.

topics and interpersonal interactions, which are probably more prominently represented in language drawn from television and movies than from newspapers and literary works. Also, from a practical viewpoint, the SUBTLEX frequency lists are publicly available in the four languages of interest.

In all cases, the individual tokens in all texts and items were matched to the respective SUBTLEX lists, controlling for word class. In French¹⁵⁶, the lists contained frequency information on the lemma, which was used to identify the relevant tokens more precisely. In some rare cases, this necessitated a choice between two identical tokens from different lemmas, which was done manually after the word lists were joined to the tagged texts, but before the sophistication measures were calculated (see also section 10.1.1).

As a measure of *lexical difficulty*, the number of content words identified as belonging to

¹⁵⁶While the German frequency list did not contain any information concerning the lemma, the English and Italian SUBTLEX frequency lists also featured frequency information for various lemmas in cases where a token could refer to more than one lemma (or POS). In both cases, however, the frequency information for the token was added up across all lemmas, so that a further differentiation was not possible in the same way that it was done for the French texts.

Lexical Diversity Variable	Variable ID	Operationalisation
Lexical rarity I	LexSoph_tokens	The mean log word frequency of all tokens in a text is calculated. <i>Interpretation:</i> a number ranging primarily from 5 to 7, peaking above 6. Higher numbers mean more frequent words on average, i.e. less lexical sophistication.
Lexical rarity II	LexSoph_content	The mean raw word frequency of all content words in a text is calculated. <i>Interpretation:</i> a number ranging between 0 and 4000+. Higher numbers mean more frequent words on average, i.e. less lexical sophistication.
Lexical rarity III	LexSoph_low_freq	For each sentence, the token with the lowest raw frequency is chosen. Then, the mean log frequency of all these words is calculated for each text. <i>Interpretation:</i> a number ranging between -1 and 5. Higher numbers mean more frequent words on average, i.e. less lexical sophistication.

Table 9.6: Operationalisation of the lexical rarity measures in this thesis. Note that the term *text* may refer to the input text, the NI or parts of the items.

the CEFR levels A1 and A2 was extracted from the input texts¹⁵⁷. As was described in section 5.2.3.2, an operationalisation of *vocabulary difficulty* is based on the idea that in the foreign language learning context, words can be categorised as belonging to a certain language proficiency, e.g. a CEFR level. This is theorised to be better suited to operationalise the difficulty (in the sense of learnability) of certain words than their frequency in L1 language use, i.e. lexical sophistication. Since this approach has never been used in item difficulty research, two data sources were implemented: the Decaleon multilingual vocabulary list (Penzkofer 2023) and the CEFRlex vocabulary lists for German, English and French (Dürlich and François 2018; François, Gala, et al. 2014; Pintard and François 2020¹⁵⁸). In both cases, the number of both A1 and A2 content words in each input text and NI was counted.

The Decaleon list, originally an XML data structure, is composed of uninflected translation equivalents in multiple languages which are identified as belonging to one specific CEFR level, though it is unknown how these levels were determined. This list was joined to the tagged texts and items in all four languages by way of their lemmas. Since Decaleon often provides multiple translations for words in the four languages, the joined lists were manually disambiguated¹⁵⁹. Once all words were either matched to exactly one translation equivalent, or to none at all, two variables were extracted: **LexDiff_Deca_A1**, which is the number of content words categorised as A1 according to Decaleon, and **LexDiff_Deca_A2**, which is a count of the content words categorised as A2.

The CEFRlex lists, based on corpora of language learning materials, were joined to the input texts in French and English and manually tagged in German. No analysis was done in Italian

¹⁵⁷Note that measures of lexical difficulty were not applied to the items, which were formulated in the language of schooling. Since the large majority of the test takers is highly proficient in their respective language of schooling, a categorisation of lexical difficulty with respect to the CEFR's lower levels does not apply.

¹⁵⁸According to the website, a reference article on the DAFlex resource is still forthcoming (<https://cental.uclouvain.be/cefrlex/>, last accessed 2025-06-02).

¹⁵⁹For example, the French lemma *découvrir* was matched to the German words *entdecken*, which was categorised as belonging to A2, and *herausfinden*, a B1 word according to the Decaleon list. Though both words are translated as *scoprire* in Italian, the first was chosen in both language versions based on the meaning of the sentence, which invited visitors of a zoo to discover its monkey park. In some cases, where it was clear that the matched word was not intended by the source, no match was retained (e.g. the verb *être* was matched only to the German word *lauten* and categorised as B1, which is clearly not appropriate).

since this language was only used in items as the language of schooling. The French CEFRlex list was joined by way of the texts' lemmas, whereas the English CEFRlex list contains plurals and conjugated verbs, which made it possible to merge it onto the tokens of the texts. Both lists indicate the number of occurrences of each lemma or token per CEFR level in their reference corpus. Following the CEFRlex website, the level of the first occurrence of each lemma was operationalised as a token's "CEFR level". The French list additionally provided a singular level based on machine learning and expert ratings of word difficulty (for more details see Pintard and François 2020). Finally, the German CEFRlex list was not open to the public at the time of writing. Hence, all texts were entered in a dedicated web page¹⁶⁰, and the CEFR levels, automatically shown as first occurrence, were entered manually in a list of tokens.

Similar to the Decaleon difficulty analysis, two variables were extracted: `LexDiff_CEFRlex_A1` and `LexDiff_CEFRlex_A2`. In addition, the variables `LexDiff_CEFRlex_Beacco_A1` and `LexDiff_CEFRlex_Beacco_A2` were evaluated for the French texts. These are based on a somewhat more sophisticated categorisation of the words in the FLElex corpus which includes information from an "official reference level description" for the French language (ibid.).

No measure of *lexical range* as suggested by Gries (2008, 2010) was chosen since the SUBTLEX frequency lists in the four languages did not all contain information on the distribution of the tokens and lemmata across the various source documents in the corpora (i.e. films and episodes). Also, since the majority of the vocabulary used in these texts was chosen to be highly frequent given the low proficiency of the readers, the effort to create a new reference corpus, though theoretically possible, was deemed too great to outweigh the probable advantage of including such a measure in this analysis.

9.3.1.4 Cross-linguistic variables

It is clear from cross-linguistic research that multilingual reading will be influenced by the presence of cognates¹⁶¹ and that this will, in turn, affect test results (cf. section 5.3). It is further hypothesised here that the influence of cognates will be particularly prevalent in a low-proficiency reading assessment where readers will by definition be unable to know each word in the input text. Therefore, the proportion of cognates in all input texts and items was extracted, similar to the operationalisation of lexical difficulty. This was done in two ways: first, all content words in the input texts were rated for their cognateness with the language of schooling, and all content words in the items were rated for their cognateness with the foreign language. A human rating was deemed necessary because reliable lists of cognates between languages, which would be useful for the educational context, are hard to find: publicly available lists of cognates (e.g. Greg Kondrak 2009; Wiktionary 2022) usually contain etymologically related words, some of which, though they may share a common ancestor, are not similar enough to be recognisable by young students. Instead, the human raters tried to anticipate what the young students who participated in the study would realistically perceive as a cognate. To account for personality effects, double ratings of the cognateness of all words were conducted, the results were discussed in detail and a consensus reached¹⁶². This resulted in the variable `CL_Cogn`, the proportion of all content words in the input text, NI or item parts which were tagged as a cognate with the respective "other" language of the assessment. Second, an experimental approach based on the multilingual Decaleon lists was chosen: the Damerau-Levenshtein distance of all content words and their direct translations in the respective language of schooling (in the case of the input texts) or foreign language (for the items) was calculated. To ensure the quality of this variable, all automatically matched translations were checked and disambiguated by a human rater be-

¹⁶⁰<https://cental.uclouvain.be/cefrlex/dafllex/analyse/>, accessed on 2023-07-20.

¹⁶¹Note that false cognates could also influence reading comprehension. Since these are even more difficult to identify, however, they were not included in this analysis.

¹⁶²Great thanks go to Nina Hicks, whose input was invaluable for this task. The quality of these ratings would not have been what it is without her precious help.

fore the linguistic distance was calculated. The resulting variable **CL_Lev** reflects the mean of all Damerau-Levenshtein distances calculated for each input text, NI, and the various parts of the items.

9.3.2 Text variables

Apart from the linguistic characteristics of a reading text, research has shown that certain textual attributes also play a role in how effective reading comprehension will probably be (see sections 4.1.1 and 4.1.2). Though their impact on reading comprehension is often compounded with other aspects of the reading situation, the topic and genre of a text, and its very general structure are decisive enough that they need to be at least included in this analysis. Item difficulty studies have also found this type of variable to have some predictive power (see section 8.3, see also for references table 8.1, page 128). Unfortunately, the low proficiency level of the assessment tasks, and the relatively small number of texts only cover a small sample out of the possible continuum of text genres and reading topics. Hence, a detailed operationalisation of these aspects has proved impossible and only two variables were ultimately retained, which capture the structure of the input texts (see table 9.7).

Variable and variable ID	Operationalisation
Text continuity ContinuousText	1: The input text is a list. It contains, at the most, unrelated short sentences. 2: The input text is composed of several pieces of text that are only loosely connected. Some pieces may be complete sentences, but they are no longer than two short sentences. 3: The input text is composed of several pieces of text that share a topic. Most pieces of text are at least two sentences long and coherent in themselves. 4: The input text is one continuous text that can stand alone.
Text genre TextFunction	argu: argumentative or persuasive statements, orders or instructions, opinions, arguments on a topic expo: brief definitions, summaries, descriptions, announcements narr: stories, jokes, documentaries, blog posts comm: chats, exchange of personal information, on-line messaging

Table 9.7: Operationalisation of the text variables used in this thesis. All text variables were coded by at least one human rater.

The variable **ContinuousText** describes on a scale of 1 through 4 to what extent a reading text can be perceived as continuous (cf. table 9.7 for a more detailed description of the four levels). This variable was chosen given that the input texts in the ÜGK assessment were developed along a continuum of continuous and less-continuous text types. The categorical variable **TextFunction** represents a rough distinction of text genres or rhetorical functions: all input texts were identified as belonging (primarily) to an argumentative, expository, narrative or communicative text genre. The operationalisation of this variable is loosely based on Neumann (2012, p. 144), who analysed listening items in a German language of schooling assessment and, thus, texts which were similar to the ones used in the ÜGK assessment. Both variables were rated by the author, at two separate moments in time for two language versions. Since internal agreement on those two ratings was high, the categories were then attributed to the input texts in all four language versions.

A categorisation of *text topic* was also attempted, but it proved unproductive: it was not possible to distinguish the low-proficiency reading texts into (academic) subject matters as was done e.g. by Freedle and Kostin (1993, p. 147), or to establish a list of topics which would cover more than one or two texts at a time. Instead, an operationalisation of how much a text “appears indicative of real-world situations” was considered, based on a variable suggested by Carr (n.d., p. 11). This, however, turned out to be barely distinctive in the sense that most topics

could be understood as being related mostly or entirely to the real world. Therefore, after some deliberation, a variable *Topic* was not retained.

9.3.3 Item and text-by-item variables

It has been amply laid out above (see section 4.2) that the task format and the items' characteristics influence the way readers will approach a reading text and, as a result, how they will perform on a reading assessment. As can be inferred from the various item difficulty studies discussed in section 8.2, the impact of items and tasks on reading test results is quite important: at least four of the nine variable groups identified in table 8.1 (page 128) refer to the items in some way, which seems quite a lot for what is essentially “only” the measurement instrument.

Freedle and Kostin postulate the existence of two types of variables which relate to the items of an assessment: pure “item variables” (1993, p. 146) and “special text-by-item interactions” (ibid., pp. 149 sqq., see also section 8.2.2). Among the former are variables which describe the item format or response type and linguistic variables such as the length or incidence of negations in the item's stem or answer options. The latter contain linguistic descriptors of the NI (called “key overlapping sentence”), lexical overlap between input text, NI, item stem and answer options. This thesis follows Freedle and Kostin's categorisation in that there are *text-by-item variables* concerned with the NI and with processes related to answering the item¹⁶³, and *item variables*, which can be determined without referral to the input text. Differently from Freedle and Kostin, linguistic descriptors of the items, such as their length, are subsumed under linguistic variables in this thesis (see above, section 9.3.1).

9.3.3.1 Item variables

As explained above (section 9.2.2), the ÜGK foreign language reading assessment used two different closed-answer formats: MCQs and Matching. Most MC items are composed of a question or stem and three answer options, ranging from one word to a short sentence at most. In two tasks (LV05E and LV09G), pictures were used instead of words. Three tasks (LV08G (item 4 only), LV09G, LV09E) also contain items with more than three answer options, making them more similar to the matching tasks in all but appearance. Items which, visually, consist only of a question and have to be answered by selecting parts of the input text (see e.g. fig. 9.4), are classified as Matching tasks. To account for the low number of non-MCQs, the variable *ItemType* only differentiates between MCQs with three answer options [*mcq*], items which used pictures as answer options [*pic*], and others [*oth*].

A second variable which has been included in item difficulty studies before, is the form of the question: in the ÜGK assessment, this may be an actual question, or it is the beginning of a sentence for which test takers have to choose an appropriate ending from the answer options. The variable *StemType* is used to explore whether the form of the item question changes item difficulty in a particular way¹⁶⁴.

Finally, empirical research on the position of the correct answer in MCQs has shown an effect on item difficulty, often a preference for earlier options (e.g. Holzknicht et al. 2021), though this finding is inconsistent and possibly interrelated with other characteristics of answer options, e.g. its strength or plausibility (Kiat, Ong, and Ganesan 2018). Nevertheless, since this variable has been found to be related to item difficulty in some cases, it is included here (*CorrPos*). The position of the correct answer option is expressed as a number ranging from 1 to 13.

¹⁶³Interestingly, though Freedle and Kostin explicitly argue that the plausibility of the incorrect options should be considered a text-by-item variable (Freedle and Kostin 1993, p. 143), they do not actually include it in their 1993 study.

¹⁶⁴Note that this variable was only applied to MCQs.

9.3.3.2 Text-by-item variables

More interestingly, text-by-item variables operationalise the interrelationship between the item and the input text, targeting in some way the process of solving an assessment item. In this thesis, four variables describe aspects of the process itself: the variable **RespProcess** tries to operationalise the cognitive process that each item is supposed to elicit, the variable **BackgroundKn** is meant to indicate whether responding to the item needs or would benefit from any additional knowledge which is not made available to the test taker in the assessment itself. A third variable **Fals** represents the number of incorrect answer options which can be falsified, i.e. explicitly be identified as not being the correct answer, if the text is understood correctly. Finally, the variable **NI_Dist** is an indicator of how much of the input text the test taker needs to understand in order to identify the correct answer. Table 9.8 contains more precise information on the operationalisation of the variables and their categories.

Variable and variable ID	Operationalisation	Source
Response process RespProcess	text: <i>text-based questions</i> – The answer to the question was explicitly stated within a single sentence in almost verbatim fashion. rewo: <i>restructuring/rewording</i> – The target information for the question was reworded or restructured. brid: <i>integration or bridging questions</i> – The target information is located across multiple sentences from the source passage and must be integrated. infe: <i>knowledge-based inference questions</i> – The information required to answer a question was not explicitly stated in the source passage.	Ozuru et al. 2008, p. 1004
Background knowledge BackgroundKn	0: No background knowledge is necessary at all. 1: A little background knowledge would be helpful. 2: Some background knowledge is necessary. 3: Background knowledge is definitely necessary.	
Falsifiability of answer options Fals	0-4: The number of (incorrect) answer options which can be falsified by reading (and correctly understanding) the text.	
Distribution of the NI Dist_NI	1: No relationship to the passage; items can be answered without reference to the passage, OR relationship of item to passage is not clear. 2: Item relates to one specific part of the passage, and requires only localized understanding of that part. 3: Item relates to more than one specific part of the passage, or relates to one specific but major portion of the passage, or requires the test taker to relate one specific part to one or more others. 4: Item relates to a number of specific parts of the passage, and/or requires an understanding of most of the entire passage. 5: Item relates to the entire passage, and requires an understanding of the entire passage.	Carr n.d., p. 22

Table 9.8: Operationalisation of some text-by-item variables used in this thesis. These variables were qualified by at least one human rater. Note that not all categories of these variables were actually found in the data.

All four variables were human-rated, partly during the process of identifying the NI (see section 9.3.3.3): while identifying the NI, for each item, the raters were also asked to indicate whether they thought an item response required any background knowledge, and whether each

incorrect answer options was falsifiable by the input text (see fig. 9.5). This information was used to record the variables **BackgroundKn** and **Fals** respectively. In both cases, the ratings were aggregated: if at least three of the five raters agreed on a category, it was adapted for the item in all language versions. The number of falsifiable incorrect answer options was then counted for each item.

Do you think a sixth-grader will need background knowledge to answer this question correctly?

Background knowledge: Information which is not provided by the text and cannot be considered a given at this age. This does **not** include language knowledge!

Choose one of the following answers

☐ No background knowledge is necessary at all.

☐ A little background knowledge would be helpful.

☐ Some background knowledge is necessary.

☐ Background knowledge is definitely necessary.

☒ No answer

Please enter your comment here:

Check the incorrect answers below which you consider to be falsifiable if the text is understood correctly.

A falsifiable answer option is clearly negated by the text, not just because it was not mentioned or does not seem plausible, but because there is information in the text which contradicts the information in the answer option.

Check all that apply

☐ machen zusammen Hausaufgaben.

☐ nehmen denselben Bus.

Figure 9.5: Screenshot of additional questions about reading item LV03E_dE_1 that raters encountered while identifying NIs. The questions concern the necessary background knowledge and the falsifiability of the answer options.

The other two variables, **RespProcess** and **Dist_NI** were rated by the author herself, at two separate moments in time for two language versions. Since overall agreement between these two ratings was high, no second rater was tasked. Also, though some deliberation was required, the variable **Dist_NI** was only entered once the NIs had been identified in the data, so that this variable is partly based on the multiple ratings of the NI described in the following.

9.3.3.3 Identification of the NI

The NI was identified by five human raters¹⁶⁵ using a LimeSurvey questionnaire¹⁶⁶.

The five raters all had advanced knowledge of the foreign language in question and were native speakers of the language of schooling¹⁶⁷. The content of the LimeSurvey questionnaires was constructed during the data preparation phase (see section 10.1) using the text data from all items and a LimeSurvey import function.

Each rater identified the NIs of all items in one language version by copying the relevant text passage(s) from the reading text into a text input field (see for the exact instruction and an

¹⁶⁵ All raters volunteered to participate and were informed of their personal rights and of the confidentiality of the test tasks prior to starting the questionnaire. They could work on the questionnaire at their own pace, stop, and continue at a later point as they wished. Their answers were treated anonymously.

¹⁶⁶ The LimeSurvey application was hosted by the Centre NTE of the University of Fribourg. It was used between June and September 2023. Access to the questionnaires was restricted to the raters at all times. To assure the confidentiality of the ÜGK items, the surveys were deleted from the servers once all answers had been recorded.

¹⁶⁷ The rating of the FRit language version was conducted by a native speaker of German who has excellent knowledge of both Italian and French.

example fig. 9.6). For each item, the five answers were then compared and the sentences of the input text which at least three¹⁶⁸ of the five raters had included in their answer were marked as the item's NI in all four language versions.

Ça prend 15 minutes pour venir depuis la maison.
Tom 15:17
Bien sûr! Du foot, ouaaais :-)! Je peux venir à pied?
Marc 15:23
Super :-)! Non, tu dois prendre le bus. À pied, c'est trop loin. Et tu ne peux pas prendre mon vélo, il est cassé.
Tom 15:25
Okay, je pars dans quelques minutes.
Marc 15:31
Très bien... C'est moi qui vais gagner!
Tom 15:40
On va voir... :-)

Item:
Tom und Marc ...

- machen zusammen Hausaufgaben.
- nehmen denselben Bus.
- spielen um vier Uhr Fussball.

🔗 Please copy the passage(s) of the input text which you think are relevant to answer the question into the field below. If there is more than one, copy all, leaving a blank line in between. Copy only parts of sentences or word groups if you feel that understanding (only!) them will be sufficient.

Figure 9.6: Partial screenshot of NI identification question of reading item LV03E_dF_1. The raters could read the input text in its entirety and copy and paste the passage(s) they considered relevant to find the correct answer for each item.

The NIs are recorded in the file `2_a_taggedTexts_dataCons.xlsx`, which contains all POS-tagged input texts in the four language versions. An item's NI was marked by entering the item number in the column NI in the rows containing all words and punctuation symbols of the one or several sentences which were considered to be the NI for a specific item. If a passage contained the information for more than one item, both item numbers were entered, divided by a pipe symbol (|). Based on this column, it was possible to extract all linguistic variables from the NI much in the same way that it had been done for the input text.

9.3.3.4 Overlap variables: operationalising multilingual test items

Finally, the remaining three variables are operationalisations of the overlap between the NI and the item, and between the input text and the item. As has been shown in section 8.2.5.3, lexical overlap between the various parts of an assessment has often been identified as a useful predictor of item difficulty and it is hypothesised that it will also be relevant in the multilingual ÜGK assessment items for the following reasons:

- The four languages are related to each other and share both lexical and syntactical similarities. Thus, it is possible to identify sections of the input text and the items which resemble each other more than others.
- The assessment contains a number of items which can be solved by matching individual words with each other, given its low level of expected language proficiency. Since this

¹⁶⁸Note that usually, agreement was higher and the five raters often agreed quite strongly on what constituted the NI. They merely differed in how much exactly of the text they decided to include, whether they chose entire sentences or focused on key words, and how much additional information they allowed.

process is at the heart of most of the overlap variables, it would not be correct to ignore them in this analysis.

- The lexical and syntactical similarities differ between the language pairs, which may be one source of the differences between the items in this multilingual assessment.

Unfortunately, it was not possible to directly use overlap variables suggested in the literature (e.g. in Carr (n.d., p. 24) or Freedle and Kostin (1993, p. 150)), as they had all been developed for input texts and items in the same language (and, in fact, a morphologically less complex language). Instead, three approaches to operationalising similarities between items and the NI or input text were chosen: first, *semantic overlap* was defined as the proportion of content words which are direct translations of each other. To identify these, the POS-tagged texts matched with the Decaleon database were used (see section 9.3.1.3 for more information on this match). The variable `Overl_sem_NI`¹⁶⁹ represents the proportion of content words in the various parts of the item which are direct translation of content words in the corresponding NI. In the same vein, the variable `Overl_sem_T` operationalises the overlap between the item and the entire input text. Second, *formal overlap* captures orthographic similarities between the item and the NI or the input text, irrespective of individual word meanings. For each item part, the proportion of tokens which are formally very similar to words in the NI and input text was counted¹⁷⁰. This resulted in the variables `Overl_form_T` and `Overl_form_NI`. Finally, the variable `Overl_dist` represents the overall string distance between each item part and their corresponding NI or input text: it was measured as the mean of the Damerau-Levenshtein distances between all content words in the two texts.

For all three variables, only content words were taken into account in order to enhance the comparability between the language versions: since function words are used in different ways in the languages, particularly in questions, these multilingual measures had the potential to differ strongly simply for structural reasons. Additionally, as mentioned in footnote 169, all variables were extracted separately for the various item parts (questions or stems, and all available answer options) to account for effects produced, for instance, by overlap with the correct answer option.

9.3.4 Test taker variables

Four variables were extracted from the much larger ÜGK questionnaire: the migratory background of the test takers, whether or not they use the foreign language at home, and the results of the L2 listening and the L1 reading orthography assessment the students solved at the same time as the ÜGK foreign language assessment.

Test takers who are immigrant students were identified following PISA (OECD 2019a, p. 179) based on information provided by the students themselves in the ÜGK questionnaire: non-immigrant students (`ImmSt: 1`) have at least one parent who was born in Switzerland, irrespective of where the student was born themselves. First-generation immigrant students (`ImmSt: 3`) were born outside Switzerland to parents who were also born outside Switzerland. Finally, second-generation immigrant students (`ImmSt: 2`) were born in Switzerland, but to parents who were not born in Switzerland themselves.

In the questionnaire, the students indicated the first and, if applicable, the second language spoken at home. Where either of these languages is the foreign language of the assessment, the

¹⁶⁹In the original data set, this variable and the formal overlap variable described below are divided up into the variables `Overl_sem_Q`, `Overl_sem_A01`, `Overl_sem_A02` and so forth to measure the overlap of each item part.

¹⁷⁰Originally, this variable was supposed to identify orthographic neighbours, i.e. content words which are of the exact same length and have a Damerau-Levenshtein distance of 1 or lower. Given the orthographic distances between the languages however, it was necessary to broaden this definition to allow for words like *éléphant* and *Elefant* to register as similar words. Hence, the definition was broadened to encompass words of the same length plus or minus one letter, and a Damerau-Levenshtein distance of 2 or lower. Also, accented and capital letter were eliminated before the measure was calculated.

variable `L2Home` is 1. If the child did not indicate a language for some reason, the variable is `NA` to avoid distortions.

Finally, the data set provided by SWISSUbase (formerly FORSbase) contains PVs for the test results on all sections of the ÜGK assessment: L1 reading, L1 orthography, L2 reading and L2 listening. These PVs are probabilistic estimations of the students' "true" competence in the assessed construct, based on their test results and various individual characteristics (e.g. their school and canton, gender, migration status, languages, cf. Konsortium ÜGK 2019a, p. 51; Angelone and Keller 2019b, pp. 18 sq.). For the purposes of this thesis, the mean of these PVs was calculated, resulting in the variables `L2List`, `L1read` and `L1Orth` to represent the test takers' competences in L2 listening, L1 reading and L1 orthography respectively. In addition to the mean, the Standard Deviation (SD) of the PVs was output.

9.3.5 The dependent variable

Item difficulty in this thesis has been estimated using IRT modelling. Both separate Rasch models of the four language versions and a Many-facet Rasch (MFR) model of the entire data set were estimated and the usual item and model fit analyses were conducted (see section 10.1.2). MFR models allow for an estimation of separate versions of an assessment, estimating all parameters on the same scale. This makes it possible, in this concrete study, to compare item difficulty estimates across language versions. Rasch modelling was chosen for practical reasons: first, though an IRT model with more parameters, e.g. a 2-PL model, would have been possible to estimate given the large number of students and would probably have a better fit due to its more precise estimation, it is unclear how the additional parameters (e.g. item discrimination) could be meaningfully included in an analysis of item difficulty. Second, MFR modelling typically requires a 1-PL item response model. If multiple facets and multiple parameters need to be modelled, a GLMM or similar approaches are required, which is computationally more demanding.

The item difficulty estimated in Rasch models is a parameter b that indicates the ability level at which the probability of a correct response reaches a certain value, e.g. $P(\theta) = 0.5$, i.e. 50% (see also section 6.2.1.2). This parameter is on a logit scale and usually a number between -4 and +4 (though theoretically $-\infty$ to $+\infty$). A higher number indicates a more difficult item, and values below 0 describe items which are relatively easy for the majority of the test takers. The MFR model provides estimations of item difficulty for the various facets, but these can be summed up to an aggregated item difficulty value which is essentially interpreted in the same way as that of a more "traditional" Rasch model.

In the final steps of this analysis, a Bayesian logistic modelling approach is chosen which combines raw test taker responses instead of item difficulty estimates, test taker and item characteristics. The resulting models estimate a response probability for individual test takers or items, and the effect of the various predictors is an estimate with confidence intervals rather than a linear relationship between an estimate of item difficulty and item characteristics. In this type of modelling, item difficulty can be inferred from all predictors of a model which are related to the items, though of course with a margin of error that becomes larger with each predictor variable added to it.

Chapter 10

Analysis

All analyses for this thesis were conducted in R (R Core Team 2024) and documented using R Markdown (Allaire et al. 2023) and the `tidyverse` package family (Wickham et al. 2019) for more readable code. All files, including the scripts, have been placed in an R project managed by the R package `renv` (Ushey 2023). Within the project, the source files, R scripts and all resulting documents and plots were named 1) according to the step in the analysis they belong to and 2) according to the data source that was under consideration (cf. table 10.1).

Filename starts with ...	Part of the analysis
1	Data preparation
2	Variable extraction
3	Descriptive analyses
4	Analytical statistics
Filename continues with ...	Data source
a	Item data
b	Item response data
c	Test taker data

Table 10.1: Filename conventions for data analysis

The R scripts and all dependent files have been placed into two electronic appendices: a complete electronic appendix is available only to the members of the dissertation committee and was issued directly upon the submission of this thesis. A public appendix¹⁷¹ contains the results of step 2, as well as all descriptive (step 3) and analytical statistics (step 4, except for Bayesian modelling), but not the original data files. This division was necessary since the original data on the ÜGK assessment (i.e. the items themselves and the test results) is not part of the public domain and was kindly provided by the ADB and SWISSUbase for the purpose of this thesis. Section C.2 contains an annotated list of all files in the two appendices.

10.1 Data preparation and variable extraction

First, all original data files were imported into R and, if necessary, modified to provide the data structure necessary for the subsequent analyses. Once the item data, item response data and the test taker data had been thus treated and saved, the variables described in table 9.4 were

¹⁷¹https://osf.io/s25xf/?view_only=f6783d0c09a44a0280166b3411bfa68a. Refer to chapter C (page 291) for more information.

extracted. In some cases, particularly for the text, item and text-by-item variables, this needed several steps and intermittent quality checks. A very general account of this process is given below. More details can be found in the respective R scripts. For further rationales on the variables and descriptions of the tasks, refer to sections 9.2.2 and 9.3 and their various subsections. To increase readability, the paragraphs below do not contain cross-references to these sections.

10.1.1 Item data

To extract the characteristics of the input texts and items, the original raw texts were copied manually from the original CBA Item Builder files into a machine-readable table¹⁷². In this table, each item in each language version represents one line, identified by its technical ID and the ÜGK ID. The various parts of each item, the description of the situation, the instruction and the input text were entered into separate columns. This resulted in the file `1_a_ItemData.xlsx`, which also contains some readily available meta data, e.g. the answer format and the correct answer as well as some of the human-rated text, item and text-by-item variables. Table C.2 (page 301) contains a description of all columns in this table.

In the initial data preparation phase, this file was used to create the import files for the LimeSurvey questionnaire that was subsequently presented to five human raters who identified the NIs of all items and supported the rating of the variables `BackgroundKn` and `Fals`. This export of LimeSurvey files was done in the first part of the R script `1_DataPreparation`¹⁷³, which also prepares the item response and test taker data.

In the following variable extraction phase, the original assessment data was transformed into the data object `itdat` in the R script `2_a_VariableExtraction_itdat`. All input and item texts were extracted from this data object and then POS-tagged using the R package `koRpus` (Michalke 2021) and `TreeTagger` (Schmid 1994)¹⁷⁴. Variables which could be determined directly from these POS-tagged data objects were added to the data object `itdat` (e.g. `TokenCount`, the various POS incidences, some lexical diversity measures) as well. The raw POS-tagged texts in all language versions were written to file and served as the input for the human cognate rating and the identification of negations.

In a second step, the reference corpora, the multilingual Decaleon matrix and the CEFRlex lists were added to the POS-tagged input texts and, again, written to file. These files then underwent a thorough quality check: most notably, multiple translation equivalents provided by Decaleon were marked for deletion, as well as some duplicated matches to entries in the reference corpora and CEFRlex. In addition, the NI for each item was identified in the POS-tagged versions of the input texts. Once these files were consolidated, they were read by the R script, and the remaining variables, e.g. lexical sophistication and lexical difficulty measures, were extracted and added to the overall data object `itdat`. At this point, the NI for each item was known, so all variables mentioned up to this point were extracted for the NI as well and added to `itdat`.

Much the same process was then applied to the item texts, though all variables were, if at all possible, applied to the various item parts (Q: question or stem; A01, A02 . . . : answer option 1, 2 and so forth). Finally, the input text and the item text data were combined to create the overlap variables, which relied on knowing the NI and having the Decaleon translation equivalents available for all texts.

The resulting data file is `2_a_itdat`¹⁷⁵. It does not contain the actual texts any more, and all item and text IDs have been transformed into mere numbers so as to reduce the recognisability

¹⁷²The author sincerely thanks Malgorzata Barras for her help and support with this arduous task.

¹⁷³All R scripts exist in two file formats: the original `.Rmd` file can be opened in R to rerun the analysis. The `.html` file is a rendered version of the original, optimised for human reading. To review the analysis, this file is recommended. It can be opened with any available browser application.

¹⁷⁴Individual tags were semi-automatically corrected in this step using the `correct.tag()` function provided by the `koRpus` package.

¹⁷⁵To preserve data structures in R, all data files were saved in an R-specific `.rds` format. For compatibility and perpetuity, they were also saved to the `.csv` format. The contents of files with the same name are identical.

of unpublished items in the public data set. Apart from this change, the file contains all text, item, and text-by-item variables for all items in all language versions.

10.1.2 Item response data

The actual item responses by all test takers were kindly provided by the ADB in a `.sav` file, native to the statistical software SPSS (IBM Corp. 2020). This file was imported into R and transformed into the data object `rsdat`. The resulting table contains 20,177 rows (the test takers) and 189 columns (the test taker ID column, and two columns for each item, one with raw answers and the other with a dichotomous correct/false coding). Non-answers, which had been marked as *missing by design*, *not reached* or *omitted*, were recoded into `NA`, `7` and `9` respectively purely for aesthetic reasons. Also, the original ADB data set uses item IDs which contain a name for each item in addition to its identifier, e.g. “FussballChat” for task LV03E. This name was eliminated for further treatment.

At this point, the data set does not identify the language version that the individual test takers responded to. This information is only available in the test taker data, where the language version can be inferred from a combination of the test taker’s canton and their language region. Since the test taker IDs in the two data sets refer to the same individuals (see below, section 10.1.3,) the information on the language versions available in the test taker data was added to the item response data. The entirety of these preparatory steps was documented in the file `1_DataPreparation`.

The variable extraction phase concerned with the item responses then focused on extracting item difficulty values for all items. To this end, a new table `ridat` was created in the R script `2_b_VariableExtraction_rsdats`. It would eventually contain all item statistics, listed by item and language version: the proportion of all answer options, not reached and omitted answers, CTT’s proportion correct and point-biserial correlation, and the results of the IRT analyses described in the following.

With the help of the *TAM* package (Robitzsch, Kiefer, and M. Wu 2022) in R, the test responses were analysed using IRT modelling: first, for each separate language version, a simple Rasch model using Marginal Maximum Likelihood (MML) estimation¹⁷⁶ was created and item fit analyses were conducted. Second, an MFR estimation using the language versions and items as facets was added. Modelling the test response data as an MFR model allows for an estimation of DIF. Instead of one item difficulty per item, three separate item difficulties are estimated in this model: 1) for each item across the language versions, 2) for each language version across all items, 3) for each combination of item and language version. As a result, the difference between the same items in the four language versions can be expressed numerically, and item difficulties across language versions are comparable. To demonstrate the suitability of this approach, the function `tam.threshold()` was used, which calculates an aggregated item difficulty value for each item¹⁷⁷. These single item difficulties were compared to the item difficulty values estimated in the separate Rasch models mentioned above. Estimates of person abilities, the WLEs, were also extracted from both modelling types, and compared to each other.

The item fit analyses conducted on the Rasch models indicate that the estimated item difficulty values only change marginally if misfitting items are eliminated. Also, the estimated item difficulties and person abilities from the MFR model are virtually identical to the parameters estimated in the separate Rasch models for each language version. Hence, the following information from IRT modelling was written to `ridat` and is, thus, available for the analyses in step 3 and

¹⁷⁶IRT modelling can be done using a variety of estimation methods (Robitzsch 2021). Here, MML estimation was chosen because it is generally the most common, and because MFR modelling only works with MML in the *TAM* package. As is shown in Robitzsch (ibid.) however, the various estimation methods do not differ substantially for larger data sets.

¹⁷⁷In fact, the function merely adds the three item difficulties together, as can be easily checked by performing the same calculation manually.

4: the three item difficulty values from the MFR model, the aggregated item difficulty value for each item, and the DIF values for each item. The latter was added as a z -score, which indicates how far each item's difficulty is from the mean difficulty, expressed in terms of the number of standard errors from the mean item difficulty. It was calculated by dividing the proportion of the item difficulty of each item in each language version by its own standard error (as suggested in Kiefer, Robitzsch, and M. Wu 2013¹⁷⁸). This value can also be interpreted in terms of its significance, which was added to `ridat` as well.

`ridat` was written to the file `2_b_ridat`. All item IDs were replaced by numbers in the same way that `itdat` had been “anonymised” before. From this moment on, analyses related to the test response data use the identifier `ridat` instead of the original `rsdat` to account for the fact that the original test responses are no longer available. The script also outputs Wright maps for each language version (using the R package `WrightMap`, cf. Torres Irribarra and Freund 2014) as well as summaries for the various Rasch models that were estimated in the script. Finally, the WLEs estimated by the MFR model were also written to file, to be used in the descriptive analyses of the test taker data.

10.1.3 Test taker data

Information on the test takers was extracted from the complete ÜGK data set provided by SWISSUBase. From this data set, the R script `1_DataPreparation` reads a `.csv` file and transforms it into the data object `ttat`. This contains structural information about each test taker (e.g. which school and class they belong to, their date of birth and gender), information on the test session they participated in (e.g. its date and time, but also ratings on how motivated and concentrated the students appeared and whether there were any disturbances), PVs on the various test parts, statistical information on the canton's school system (e.g. on educational costs, education, gender and age of the teachers) as well as every student's raw answers to the ÜGK questionnaire. Test takers are identified by a number. In the data preparation phase, this identifier was compared to the item response data, which also contains student IDs, though in a different order. Given that the number of correct answers on the reading comprehension test clearly correlates with the PVs for the same test, it was assumed that the person identifiers refer to the same individuals¹⁷⁹. As a result, the test taker data was used to identify the language version which was solved by each student in the item response data. Otherwise, the test taker data was preserved as is, a table with 20,177 rows (the test takers) and 821 columns.

In the variable extraction phase, this large data set was reduced to five variables only: the test takers' immigration status, their home language, and their mean PVs on the two L1 assessments and the L2 listening assessment¹⁸⁰ (see table 9.4, section 9.3.4). For all variables, the relevant information had already been present in the test taker data and was merely converted. The resulting data object, a table with 20,177 rows and 12 columns, was written to the file `2_c_ttdat`.

10.2 Statistical analyses

Once all test, item and test-taker information had been prepared, the actual analyses could be conducted. At this point, all variables operationalising the characteristics of the assessment were available in the file `2_a_itdat`, with items and input texts identified by numbers instead of their original ÜGK IDs. The performances on the assessment were expressed as item difficulties in the file `2_b_ridat` and a few test taker characteristics had been extracted from the questionnaire data into the file `2_c_ttdat`.

¹⁷⁸Though not identical, this approach is similar to the one suggested by Lord (1980, cited in Stubbe 2011, p. 468), mentioned in section 6.2.2.2.

¹⁷⁹This assumption was also confirmed by Florian Keller, who provided the raw item response data (*personal communication*, 2020).

¹⁸⁰In addition to the means, the SDs were also extracted.

These files provided the basis for the initial descriptive analyses of all variables. The actual inferential analyses were then conducted on merged data sets: initial multicollinearity analyses on the input text and item variables were conducted, including factor and cluster analyses on closely related variables. Then, the relationship of all text, item and text-by-item variables with item difficulty was studied in correlational and regression analyses, both within and across language versions. Finally, the findings from these analyses were used to conduct some more exploratory analyses: a random forest analysis was conducted to triangulate the findings from the regression analyses, and Bayesian logistic models were estimated in which promising assessment variables (i.e. variables which had been identified as suitable predictors before) and test taker variables were combined and their interaction studied.

Graphical outputs were created using primarily the R package `ggplot2` (Wickham 2016). All correlational analyses were conducted with the help of either the `cor()` or `cor.test()` functions from R's native `stats` package (R Core Team 2024) or their `tidyverse` equivalent `cor_test()` from the `rstatix` package (Kassambara 2023). For the regression analyses, the `lm()` function of the `stats` packages and the `lmer()` function of the `lme4` package (Bates et al. 2015) were used depending on the model to be estimated. The random forest analysis was mainly supported by the `randomForest` package (Liaw and Wiener 2002) and the Bayesian models were estimated by the `brms` package (Bürkner 2021). Of course, many other packages contributed to the analysis, particularly for diagnostic functions (see the R scripts for details).

10.2.1 Descriptive statistics

Given the amount of data under consideration in this thesis, and the complexity of the matter, thorough descriptive analyses of all variables were established before any inferential statistics were applied. These analyses were conducted in the scripts `3_a_DescriptiveStatistics_itdat`, `3_b_DescriptiveStatistics_ridat` and `3_c_DescriptiveStatistics_ttdat`. In many cases, these analyses primarily follow the data rather than a prescribed path of analysis, particularly with respect to the crosslinguistic characteristics of the input texts and items, and the patterns of item difficulties across language versions. This strategy was chosen to account for the fact that little is known about the operationalisation of linguistic characteristics of texts *across* languages.

First, all text, item, and text-by-item variables were considered by themselves. The distribution of each variable was examined in each language version, both graphically and numerically. Categorical variables were immediately recoded where certain levels were clearly under-represented (i.e. no or only a small minority of cases were identified). Many linguistic characteristics of the input texts and items were examined in terms of how they differ across the language versions. As part of these analyses, some variables which operationalise similar constructs were also considered for how they relate to each other. This was sometimes helpful to ascertain the meaningfulness of these purely numerical values.

Next, the test response and item difficulty data was examined without reference to the various text and item characteristics. The analysis focussed on items which were particular in some way (i.e. very hard or easy, or featuring an unexpected response pattern), or which functioned very differently across language versions. To this end, the R script `3_b_DescriptiveStatistics_ridat` first outputs graphs which show the number of test taker responses to each answer option as well as non-answers, i.e. the number of test takers who skipped an item or did not reach it due to time constraints. Second, the results from the CTT and IRT analysis were examined in more detail, and, again, particular items were highlighted. For both types of modelling, the R script outputs some descriptive statistics and a few graphs showing distributions and relationships. The results section also includes a consideration of item difficulties across the language versions, both in terms of cross-linguistic correlations and DIF (sections 11.2.4 and 11.2.5).

Finally, the test takers themselves are briefly described in section 11.3, and their performance in the L2 reading assessment and the other ÜGK assessments is examined.

10.2.2 Inferential statistics

The inferential analyses in this thesis were conducted in three major steps: first, interactions between the variables of each data set were examined in light of potential multicollinearities and to establish patterns in how the characteristics of the assessment interact with each other. Second, the linear and additive relationships between item difficulty and the characteristics of the assessment were investigated. Third, two additional statistical approaches were employed to analyse the data: random forests and Bayesian logistic modelling.

10.2.2.1 Interactions between variables

As indicated in the literature review (e.g. section 8.3.6) and inferable from the operationalisation of certain variables, multiple variables in this data set are bound to be related to each other. Some variables naturally measure aspects of the same construct (e.g. the three lexical sophistication measures suggested in Coh-Metrix), others contain similar information due to the way they were operationalised. These relationships and interactions between the variables were analysed using correlations as well as factor and cluster analyses. In order to facilitate these analyses, all variables relating to the incorrect answer options were reduced to their mean value. This reduced the number of possible combinations substantially and at the same time consolidated the characteristics of the incorrect answer options into one single value per variable for easier interpretability.

First, correlation matrices for all variables were established using Spearman's correlations in each language version. For this analysis, categorical and ordinal variables were considered continuous, as this allowed for the identification of strong relationships with other variables. Item difficulty was not included, as it serves as the dependent variable in the inferential analyses. The correlation matrices were also examined visually, with the help of heat maps created in R using the package `heatmaply` (Galili and O'Callaghan 2024), but also using the conditional formatting function in Excel (Microsoft 2024). Based on these large correlation matrices, interesting relationships between variables were examined in more depth, often by creating more specific graphs.

In an effort to reduce the number of variables for the subsequent analyses, factor and cluster analyses were conducted on the linguistic characteristics of the assessment. The linguistic characteristics of the input texts, all clearly related to each other in complex ways, underwent factor analyses to determine underlying commonalities of the texts. For the shorter text parts in the NIs and items, cluster analyses on the POS variables were conducted in order to determine groups of texts which are similar to each other.

The factor analyses were first conducted on the linguistic variables of each language separately, given that the linguistic properties of the input texts were not always identical. Prior to establishing the actual model, a Kaiser-Meyer-Olkin (KMO) test was conducted to identify variables which were not suitable for the factor analysis. The result of this was rechecked using Bartlett's Test of Sphericity, and the number of factors was established using a scree plot. Based on the findings from these preliminary analyses, a factor model was estimated. Then, based on the fit values, factor loadings and factor commonalities, changes were made to a) the number of factors, b) the included variables, c) the rotation or estimation method. The final model was used to extract scores for each of the 39 input texts. Based on the factor loadings, each factor was named with respect to what the scores mostly reflect in terms of the linguistic characteristics of the texts. All analyses were made with the help of the R package `psych` (Revelle 2024) using the functions `fa()`, `KMO()` and `cortest.bartlett()` in particular. More detailed information and comments on the analyses can be found in the R script `4_AnalyticalStatistics_Multicollinearity`.

Due to the lower density of data¹⁸¹, the linguistic characteristics of the NIs and item parts were

¹⁸¹Since non-MC items could not be analysed for all item parts in the same way as the MC items, not all items contain information in all variables. Also, very short texts did not allow for certain linguistic analyses, which

subjected to a k-means cluster analysis. The objective was to create one variable which identified similar NIs or item parts in terms of their POS structure. Using the R package `NbClust` (Charrad et al. 2014), the best number of clusters was determined, and the clusters were estimated using the function `kmeans()` from the `stats` package (R Core Team 2024). The resulting clusters were added to the data set for further analysis.

10.2.2.2 Item difficulty prediction: correlational analyses

First, the correlations of the item difficulty values from the MFR model with all text and item variables were estimated using Spearman’s correlations and the results written to file. Variables which correlated significantly with item difficulty in at least one language were identified and a list of them and the respective correlations was written to file. In addition, to gain an understanding of how much these bivariate relationships are actually monotonic and thus represented by the correlation, scatterplots were exported for all of these variables in all language versions.

Second, all item difficulties related to the same input text were averaged and a new “mean task difficulty” was calculated, similar to Neumann (2012, pp. 169 sq.). These were correlated with all variables related to the input texts and significant correlations were, again, identified and plotted.

10.2.2.3 Item difficulty prediction: regression analyses

The complex relationships between the variables in this study, and the often meagre findings from previous regression studies suggest that the selection of variables for a regression model is a demanding and complex process. This proved to be the case in this study, too: trials involving the inclusion of variables based on some sort of objective criterion were unhelpful. Neither the inclusion of only variables which correlate significantly with item difficulty, nor automatised variable selection produced lists of variables which predicted item difficulty to a satisfactory amount.

Instead, stepwise hierarchical regression was conducted based on the findings from the literature review as well as the prior analyses of the variables and their interactions. Using the ENde version as the first baseline model, a mixed effects model with `TextID` as a random effect and the aggregated MFR item difficulty as the dependent variable was estimated, and predictors were added to it successively, starting with `TokenCount_T`. Text length was chosen as the first predictor due to its high correlations with item difficulty and its prominence in other item difficulty studies. In a stepwise forward procedure, predictors were added in groups of theoretically or empirically related variables and remained only if their contribution to the model was significant. At each step, possible interactions among the variables in the model were taken into consideration. In cases where several variables were identified as possible candidates, the better fitting model was chosen based on Analysis of variance (ANOVA) tests and the fit indices Bayesian information criterion (BIC) and Akaike information criterion (AIC) or, in rare cases, on which model reached a better R^2 value. The important steps of this procedure can be found in the script `4_AnalyticalStatistics_Regression`, though for the sake of brevity, not every attempted combination of variables has been documented.

Once a final model for the ENde data set was found, further regression models were established for the other three language versions. In all cases, the final models were inspected in more detail: model fit was evaluated using the coefficient of determination (R^2) and the more conservative *adjusted* R^2 . The estimated regression coefficients and their 95% confidence intervals were examined to assess the contribution of individual predictors. In addition, relative importance metrics were computed to further assess the relative contribution of each predictor using the Lindeman-Merenda-Gold (LMG) measure from the R package `relaimpo` (Grömping 2007, based on Lindeman, Merenda, and Gold 1980). Predicted effects were visualized and compared to the further reduced the number of meaningful variables for some items.

observed values to evaluate model adequacy. In addition, Variance Inflation Factors (VIFs) were computed to assess multicollinearity, and residuals were visually inspected to check for deviations from homoscedasticity¹⁸² and normality.

To evaluate the transferability of the findings across language versions, each model was used to predict the item difficulty of the other language versions and the Root Mean Squared Error (RMSE) was calculated (so-called *cross-validation*). RMSE provides an indication of how far the predictions diverge from the observed item difficulties.

Finally, an overarching model based on the entire data set was estimated, with the language version and the input text as random effects in a mixed-effects linear model. It was constructed in the same way as the language-specific models and underwent the same quality checks.

10.2.2.4 Item difficulty prediction: further exploratory statistics

To triangulate the findings from the regression modelling, an exploratory random forest analysis and Bayesian logistic modelling were applied to the entire data set. For the random forest analysis, all variables were initially allowed to be predictors. Based on their mean squared error (MSE), which indicates the importance of a predictor in a random forests analysis, variables which did not add much to the model were eliminated. Following this step, a VIF analysis was applied to a linear regression model containing the remaining variables. Since at this point, the structure of the data was well-known, one variable among those which interact strongly with each other was chosen for a final random forest. From this, several fit indices were extracted, and a single regression tree was estimated based on the final data set to explore the structure of the data.

Finally, Bayesian modelling was explored both for triangulation purposes and to investigate the effects of some test taker characteristics. Three Bayesian logistic regression models using the `brms` package in R were estimated: the first model was designed to be comparable to the frequentist MFR model estimated for the regression analyses. The second model combined the results of the regression modelling with the test taker variables retained for this analysis. Finally, informed by exploratory modelling and variable importance estimates, selected interaction effects between item properties and person variables were introduced in the third model to explore differential item functioning. The dependent variable in all cases was the binary outcome of item responses (correct/incorrect), and all models included random intercepts for students to account for differences in overall ability. Items and the language version were also included either as random slopes or random intercepts depending on the model. Limited model refinement was guided by posterior summaries and model comparison criteria (especially leave-one-out (LOO) cross-validation, see the R script `4_AnalyticalStatistics_LLTM`). Since Bayesian modelling requires a lot of computing power, all analyses were conducted on a random sample of the data set (about 20% of the test takers in each language version).

¹⁸²Homoscedasticity is the assumption that the variance of a variable is the same across different groups or values. In inferential statistics, this assumption usually has to be met to assume accurate model prediction. It can be checked by plotting the predicted values with their residuals: if they do not scatter randomly, then a predictor functions differently within the model and homoscedasticity is violated.

Part III

Results

Chapter 11

Descriptive statistics

In sum, when cross-linguistic [...] comparisons are based on individual linguistic features, the [...] analyst is faced with tantalizingly interesting findings coupled with fundamental doubts concerning the appropriate methodologies to use.

Biber 1995, p. 82

Over the entire data set, 239 text, item and text-by-item variables¹⁸³ were extracted for 39 input texts and 94 items in each of the four language versions. The item difficulties of these 4×94 items were calculated according to CTT and estimated in an MFR model based on the item responses from 20,177 test takers (see table 9.2, page 136, for a breakdown of the number of test takers per language version). For these 20,177 test takers, two individual variables were extracted from a background questionnaire and three performance indicators were added, based on their test results in the three language assessments that were administered along with the L2 reading assessment under investigation in this thesis.

In the following sections, prior to any inferential analyses, the statistical characteristics of the variables themselves are described. First, the variables operationalising the characteristics of the input texts and items are discussed. Next, some observations on the item responses and the estimated item parameters from CTT and IRT analyses are presented. Finally, in section 11.3, a short overview of the test taker groups is given.

11.1 Text, item and text-by-item variables

The various variables described in section 9.3.1 can be roughly divided into nominal or ordinal, i.e. *categorical* variables (which were rated by humans) and *continuous* (i.e. interval or ratio) variables, which are counts (e.g. of the tokens in a text), proportions or were otherwise calculated based on the language data. Most of the former are the same across the language versions, while the latter are often language-specific.

First, the various categorical variables will be discussed. Second, the more numerous continuous variables are separated into thematic groups: Section 11.1.2 focusses on the text length and POS variables, section 11.1.3 presents observations on the various lexical variables (lexical diversity, sophistication and difficulty), section 11.1.4 is concerned with the two cross-linguistic variables, and section 11.1.5 with the six overlap variables. Though some details will be mentioned, the number of variables and language versions makes it necessary to choose the most salient findings. Detailed statistics and graphs can be found in the electronic appendix: the R script `3_DescriptiveStatistics_itdat` contains descriptive statistics produced by the `describe()` function of the `psych`-package as well as boxplots, barplots or density plots of all variables in all

¹⁸³Note that this number includes the variables extracted for up to eight answer options, which means that there are many empty table cells. Excluding all variables beyond the usual three answer options, 138 variables remain.

language versions. All correlational analyses and plots of multivariate and cross-linguistic relationships mentioned in this section are also documented there (see chapter C of the appendix for an overview of the electronic appendix).

11.1.1 Categorical variables

The human-rated variables in the data set operationalise characteristics of the input texts, items or their interaction with each other. They usually have three or four levels (or categories), of which one is most prominent. These variables show that most items are MCQs, and refer to explicitly stated information in short passages of the input text. Also, most input texts are expository, roughly continuous texts (see detailed information on the distributions of these variables in table 11.1).

Due to the low number of occurrences in one or two of the levels, six variables were recoded: **BackgroundKn**, **ItemType** and **NI_Dist** were dichotomised, so that they distinguish between 1) items which can be linked to background knowledge in some way and those which cannot, 2) MCQ items with three written answer options and other item types, and 3) items which rely on one short, specific passage of the input text for the correct answer, and those which refer to more than that. Furthermore, the variables **ContinuousText** and **RespProcess** had two neighbouring categories condensed into one so that the remaining three options are more evenly distributed across all items. These two ordinal variables were further recoded so that the lowest level is still represented by 1. The variable **CorrPos** newly only distinguishes between items which have the correct answer in position 1, 2, 3 or beyond.

Finally, the variable **StemType** will not be analysed further given the very low incidence of stems which are not questions. With so few occurrences, any observations about item difficulty have a high potential to be made by chance.

11.1.2 Linguistic variables: descriptives

The large majority of the variables which describe the linguistic characteristics of the input text, NI and item parts can be considered interval or ratio variables, though in some cases, the number of occurrences of the various counts is low enough that they may as well be considered ordinal (e.g. the variable **LingCompl_Neg**, which represents the raw number of negations in the relevant piece of text, has a maximum range of [0,8] with only eight levels across all languages and assessment text parts). In the following, an overview of the distributions and characteristics of the linguistic variables is given.

11.1.2.1 Input texts and NIs

Text length was measured for all input texts, NIs and item parts in all languages, along with average sentence lengths¹⁸⁴, and incidences of nouns, verbs, adverbs and adjectives, pronouns, names and human-rated negations. Across the three foreign languages, input text length proved to be bimodal: some input texts are rather short, centred around 50 tokens, while another group of input texts is between 100 and 200 tokens long, with some texts longer than 200 tokens. The short texts originate, with one exception (LV06G), in tasks which contain more than one input text, i.e. where not all items of the task refer to the same input text (see section 9.2.2.2). If the input texts of these tasks are summed up, they also fall in the range between 98 (LV09G in German) and 238 tokens (LV08E in French)¹⁸⁵. In general, German texts are the shortest, and French texts the longest, though the differences are minor (see table 11.2), and the variable correlates very strongly across the three foreign languages ($r = .99$ for all three combinations). Mean sentence

¹⁸⁴This measure was only applied when TreeTagger recognised at least one sentence in the text section of interest.

¹⁸⁵The task LV05E is somewhat different in that it contains two longer input texts that are each linked to two items. Summed up, this task would have the longest input text (German: 245, English: 260, French: 275).

Categorical variable	Distribution	Comment
CorrPos $N = 94$	1: 22 2: 26 3: 25 4, 5 and 6: 4 each 7, 8 and 10: 2 each 9, 11 and 12: 1 each	Correct answers in 1st, 2nd and 3rd position are reasonably well distributed, but items with more answer options are rare, and the position of the correct answer options declines. Answer positions 4 to 12 ($N = 21$) are recoded into “3+”.
ContinuousText $N = 94$	1 (list): 8 2 (loose pieces of text): 23 3 (short texts): 31 4 (one text): 32	About two thirds of all input texts are mostly continuous. Categories 1 and 2 are merged into one to represent lists and loose pieces of text.
TextFunction $N = 94$	argu: 18 expo: 38 narr: 17 comm: 21	Expository texts are present about twice as often as any other text type. All other text types are evenly distributed.
ItemType $N = 94$	mcq: 60 pic: 5 oth: 29	The majority (64%) of all items are MCQ and only very few use pictures. A third of all items are not MCQ. The items using pictures are added to the “oth” category.
StemType $N = 62$	q(uestion): 55 s(tem): 7	The number of stems used is quite low, particularly given that not all items are MCQ. This variable is eliminated from the data set.
RespProcess $N = 94$	text (-based): 44 rewo (reword/restructure): 32 brid (integration/bridging): 15 infe (inference): 3	Three quarters of all items elicit either verbatim or explicitly stated information in the input text. 15 items require text integration, while only 3 items are inference. The latter two categories were merged into one.
BackgroundKn $N = 94$	0 (none required): 85 1 (a little may be helpful): 7 2 (some is necessary): 2	Less than 10% of all items were rated as possibly requiring a little or some background knowledge. No item was rated as definitely needing background knowledge. The two categories 1 and 2 will be condensed into one, leaving a dichotomous variable.
NI_Dist $N = 94$	2 (one specific passage): 75 3 (major passage): 17 4 (most of the text): 2	Most of the items refer to one specific, short passage in the text. Only about 20% of all items require an understanding of a larger passage of the text, or several parts of it. No items were rated as unrelated to the input text or requiring understanding of its entirety. The categories 3 and 4 will be condensed to one, leaving a dichotomous variable.
Fals $N = 65$	0 (no falsifiable answer option): 21 1 (one falsifiable answer option): 13 2 (two falsifiable answer options): 31	About half of all MCQ items feature two falsifiable answer options while the other half have either one or no falsifiable answer option.

Table 11.1: Descriptive statistics of categorical item variables before recoding. *Note:* All information is presented for one language version only. It is identical in all language versions. Where $N < 94$, the variable was not applicable to all items. Refer to tables 9.7 and 9.8 for more information on the variables’ categories.

length is between 5 and 10 tokens in all three languages, with German sentences generally shorter. Some outliers in mean sentence length are due to TreeTagger having been unable to properly

determine sentence length in short or non-continuous texts, and where sentence punctuation was unusual (e.g. in task LV09E, ellipses, i.e. “...”, were used several times instead of a period; these were not recognised as sentence punctuation). For the two tasks in which this was the case, all `SentenceLength_T` values were manually recoded to NA.

Language version	Foreign language	Input text: mean length & range	NI: mean length & range
DEfr	German	105.1 [23, 229]	13.9 [1, 34 122]
ENde	English	113 [24, 251]	13.9 [1, 34 130]
FRde & FRit	French	119.4 [31, 257]	15 [1, 38 152]

Table 11.2: Text length in tokens of the input texts and NIs in the three foreign languages.

Note: The range of the text lengths of the NIs is represented by the smallest, the third-to-largest and the largest value to account for the fact that most NIs are very short.

When only the NIs are considered, i.e. the part of the input text in which the answer to each item can be found, text length is considerably shorter: with the exception of two items which require understanding of the entire input text ([45], an inference item, and [68], a gist item), all NIs are shorter than 40 tokens, with a median around 11.5 tokens in all three foreign languages (German: 11.0, English: 11.5, French: 12.0, see also table 11.2).

Normalised incidences of POSs in the input texts are similar in the three foreign languages. Nouns make up roughly 20% of each text on average, though some input texts contain up to 40%. Verbs in the French and German texts are generally fewer, averaging around 15% of all tokens. In English, their proportion is slightly higher. Adjectives and adverbs also reach 15% on average in all languages, though some input texts contain considerably less or more. Pronouns make up more than 10% of the German and French texts, while the English texts contain less ($\approx 7\%$). Names are generally rare, with some texts not containing any names at all. Two texts contain a lot of names: a biography in the form of a list and a messenger chat. In both cases, the same names are repeated several times. Finally, negations are generally rare, with most texts containing one negation and four texts containing more than five.

In all three foreign languages, though to varying extents, the proportion of nouns with all other POSs correlates negatively, while the proportion of verbs and pronouns in particular correlate positively with each other (see fig. B.1, page 276). Neither the incidence of names nor of negations are related as strongly to the incidence of the other POSs, though negations occur more often when there are more verbs and pronouns, while names occur more often when there are less adjectives. Nevertheless, both clearly also depend on other aspects of the input texts, most probably their content or topic (see section 12.1.2). All of these observations hold across the three foreign languages: the correlations between the various POSs are all in the same ranges, and all POS-variables correlate strongly across the language versions (e.g. `POS_N_T` in English and German: $r = .95$, $p < .001$, or `POS_Ad_T` in English and French: $r = .86$, $p < .001$).

In the NIs, the above findings are repeated, though to a somewhat weaker extent given their short lengths. All POS-variables in the input texts and their related NIs show a positive, moderate correlation with each other in all language versions (e.g. `POS_N_T` and `POS_N_NI`: $r = .69$ in English; `POS_NAM_T` and `POS_NAM_NI`: $r = .41$ in French). The POS-variables related to the NIs are considerably more skewed towards the left, with some short NIs not containing any adjectives/adverbs or pronouns. Names and negations are also less frequent or not present at all.

11.1.2.2 Questions and answer options

The stems and questions of the 4×94 items range in length from 2 to 16 tokens with a median of 7 or 8, depending on the language of schooling. Given these low numbers, the proportions of

nouns, verbs and the other POSs seem to be distributed mostly at random, depending on how the question or stem was phrased exactly¹⁸⁶: a handful of questions in each language version do not contain either nouns or verbs and about a third does not contain any pronouns. Meanwhile, only a small number of questions contains an adverb or adjective¹⁸⁷ or one or two names.

Out of all items, 65 used an item format which made it possible to POS-tag answer options. Of these, only five items feature more than three answer options. All answer options, irrespective of their position, range in length between 1 and 14 tokens, though most are shorter than 8 tokens. In all cases, shorter answer options (of only 1, 2 or 3 tokens) are the most frequent. As was observed with respect to the item questions, the proportion of POSs does not follow any particular pattern, though the incidence of nouns is much higher in all answer options than all other POSs, even verbs.

Across the language versions, the questions and answer options continue to follow similar patterns across the language versions: token length often varies by only 1 or 2 tokens for the same item, and the order of the POS-variables is usually the same, though differences in linguistic structure, formulation and POS-tagging lead to different raw numbers in some cases and radically different proportions in others.

11.1.3 Lexical variables

The lexical variables used in this study are concerned with lexical diversity, sophistication and difficulty (see for more details on their operationalisation: section 9.3.1.2 and following). Their scales differ widely: for instance, the variable **LexDispersion**, meant to measure how much similar tokens are clustered in a text¹⁸⁸, ranges between 0 and 0.32 and is centred around 0.12 (German; French: $\mu = .11$, English: $\mu = .15$), while the variable **LexSoph_content** ranges from 36.90 up to 4228.95 (in French texts) with a mean of 1166.89 (French; German: $\mu = 1517$, English: $\mu = 784.32$)¹⁸⁹.

11.1.3.1 Input texts

For the 39 input texts in each foreign language, these variables follow a unimodal or bimodal distribution, often skewed. Unsurprisingly, given the low number of input texts, most of these distributions cannot be considered normal. Table B.1 (page 278) lists descriptive statistics for each variable¹⁹⁰.

Most of these variables are derived or calculated in some way, often based on operationalisations established for English. All of them rely on the accuracy of the POS-tagger, and the sophistication and difficulty variables are also linked to reference lists that vary between the languages, and in quality. Hence, as is evident in table B.1, the range and distributions of these variables tend to vary across the language versions. To ascertain whether this observation is actually reflected in the data, the differences between the means of the lexical variables extracted

¹⁸⁶Note that due to the low number of tokens in the various item parts, the error rate in TreeTagger's POS-tagging is bound to be larger. Furthermore, an erroneous tagging of only one word may change the calculated proportions of the POSs significantly. Therefore, the numbers and distributions of these variables should not be over-interpreted.

¹⁸⁷There are two questions which actually feature an adverb *and* an adjective: item [2] contains the string "loud enough", and item [33] "(not) yet well". Consequently, in all three languages of schooling, these two items register a large "proportion" of adverbs/adjectives.

¹⁸⁸Roughly speaking, this indicates the percentage of words in a text that are repeated within 20 words (with the five most frequent words of the language eliminated first).

¹⁸⁹This value indicates the mean raw frequency per million tokens of the content words in an input text. Since individual words may, obviously, appear quite often in a language, this number is sometimes quite high. A high value is indicative of lower lexical sophistication.

¹⁹⁰The data set contains the variable **LexDens_contentW**, which is merely a count of the content words of the passage under consideration, and correlates strongly with token length. This variable was added as a helper variable to allow for normalisations of other variables based on the number of content words in a text, and it is not considered in the following descriptions.

from the input texts across the three foreign languages were checked for statistical significance (cf. table 11.3¹⁹¹). Interestingly, the means of the *lexical diversity* variables differ significantly only between English and the other two languages, but not between the German and French texts. The means of the *lexical evenness* and *lexical density* measures do not differ significantly across the language versions at all. Meanwhile, the means of all *lexical sophistication* measures differ significantly from each other in the language versions, whereas the means of the *lexical difficulty* measures only ever differ between two languages, but not the third (though no one language combination is systematically similar).

Variable	English-German	English-French	German-French	Test
LexDiv_Var_T	*** $p < 0.001$	** $p = 0.002$	n.s.	Pairwise t-test with Bonferroni corr.
LexDiv_Eve_T		n.s.		Friedman rank sum test
LexDispersion_T	n.s.	** $p = 0.007$	n.s.	Pairwise t-test with Bonferroni corr.
LexDisparity_form_T	*** $p < 0.001$	*** $p < 0.001$	n.s.	Pairwise t-test with Bonferroni corr.
LexDens_T		n.s.		Friedman rank sum test
LexSoph_tokens_T	*** $p < 0.001$	*** $p < 0.001$	*** $p < 0.001$	Pairwise t-test with Bonferroni corr.
LexSoph_content_T	*** $p < 0.001$	* $p = 0.02$	** $p < 0.01$	Pairwise t-test with Bonferroni corr.
LexSoph_low_freq_T	*** $p < 0.001$	*** $p < 0.001$	*** $p < 0.001$	Pairwise t-test with Bonferroni corr.
LexDiff_Deca_A1_T	*** $p < 0.001$	n.s.	*** $p < 0.001$	Pairwise t-test with Bonferroni corr.
LexDiff_Deca_A2_T		n.s.		Friedman rank sum test
LexDiff_CEFRlex_A1_T	*** $p < 0.001$	** $p = 0.003$	*** $p < 0.001$	Pairwise t-test with Bonferroni corr.
LexDiff_CEFRlex_A2_T	n.s.	* $p = 0.03$	n.s.	Pairwise t-test with Bonferroni corr.

Table 11.3: Significance of difference between means of lexical variables in input text across foreign languages

To complete this analysis, the language-specific values of all of these variables were plotted against each other¹⁹² with some interesting results: the lexical *variability* measure also varies with regard to the individual data points, though a correlation of around $r = 0.6$ in all three language combinations still shows a general tendency of matching low and high values. The lexical *evenness* variables in the three languages are strongly and visibly correlated ($r > 0.84$, $p < 0.001$), as are the lexical *density* measures ($r > 0.82$, $p < 0.001$), though it should be pointed out that both lexical evenness and lexical density are highly correlated with text length as well (see also section 12.1.2.1). The cross-linguistic scatterplots of lexical *dispersion* seem to suggest that the English values are different from the German and the French values in that

¹⁹¹This analysis was not repeated for the NIs or the item parts given their generally low token counts.

¹⁹²See the R script `3_DescriptiveStatistics_itdat` for all statistics and graphs.

they are systematically smaller than the German and higher than the French values, though still correlated, of course (German-English: $r = 0.47$, $p < 0.01$; French-English: $r = 0.57$, $p < 0.001$, French-German: $r = 0.62$, $p < 0.001$). The lexical *disparity* values, already significantly different in terms of their means, also have a low-to-moderate correlation across the foreign languages, though, again, the German and French values are more similar to each other ($r = 0.4$, $p = 0.01$).

The two log-transformed lexical *sophistication* measures `LexSoph_tokens_T` and `LexSoph_low_freq_T`, despite their differences in mean, are moderately (the token variable, $0.63 < r < 0.7$, $p < 0.001$) and strongly correlated (the low-frequency variable, with $0.73 < r < 0.85$, $p < 0.001$), which is an indication that the log-transformation is a good choice when measuring lexical sophistication, particularly across languages¹⁹³. Inversely, the variable `LexSoph_content_T` which uses raw frequency scores and relies on the identification of content words based on TreeTagger's POS-tagging, differs much more between the language groups (English-French: $r = 0.51$, $p < 0.001$; English-German: $r = 0.49$, $p < 0.01$, French-German: $r = 0.39$, $p < 0.1$).

Finally, the lexical *difficulty* variables differ across the languages as a function of their reference list: while the identification of A1- and A2-level lemmas based on the multilingual Decaleon database led to mostly similar proportions of these words in the three languages, the use of the language-specific CEFRlex lists¹⁹⁴ created lexical difficulty values which are not correlated across language versions (see fig. 11.1). In fact, a close look at the axes in fig. 11.1(b) or at fig. B.3 (page 280) reveals that the proportion of A1 content words in the German texts was measured as being systematically higher than in the other two languages. This is despite the fact that these texts should be translations of each other (see section 13.2.2 for further discussion of this point).

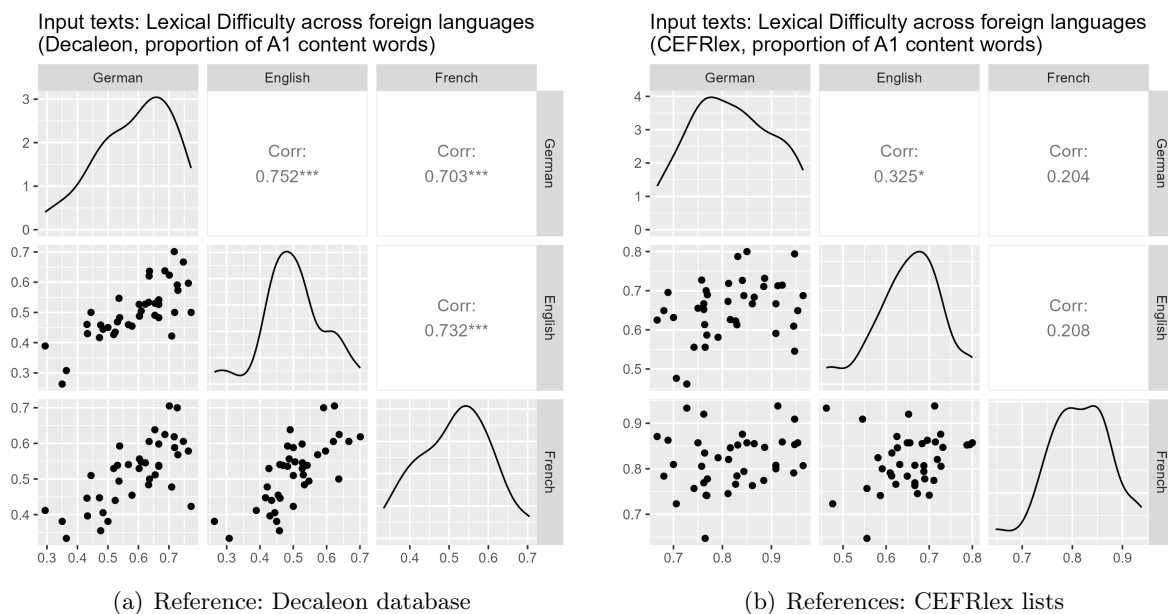


Figure 11.1: Lexical difficulty variables (proportion of A1 content words in input texts) across languages. Note the differing scales of the individual graphs.

¹⁹³This conclusion, of course, only applies if it is assumed that lexical sophistication in the three languages should be comparable to a large extent. See the discussion on this in section 13.2.3.

¹⁹⁴Note that for the French input texts, the CEFRlex Beacco measures (Pintard and François 2020, see also section 9.3.1.3) were used given that they differentiated more strongly between texts. Also, the A1 and A2 Beacco measures are more clearly connected with each other than the CEFRlex estimates based on a token's first occurrence in the corpus.

11.1.3.2 NIs, questions and answer options

Among the NIs and item parts, the lexical variables are often difficult to interpret: lexical variability, evenness and dispersion could only be calculated for longer pieces of text, of at least 12 or more tokens¹⁹⁵. Hence, these three variables were only extracted for up to half of all NIs and not at all for the item parts. The values that *were* extracted, show clear relationships with text length or with each other, which at the very least puts their meaningfulness in question (see fig. B.2, page 279). Lexical disparity and lexical density as well as the lexical sophistication variables were calculated even for shorter texts, with less obvious effects of the short text length: though there are some more extreme values for all variables in the NI and item part statistics, their overall distribution is similar to that of the input texts.

A closer investigation of the variable `LexSoph_content`, which is a measure of only the content words in a text, puts its informative value in question: across the foreign languages, the value for what is essentially the same piece of text often varies wildly (e.g. the NI of item [35] in German: 205.6, English: 171.95, French: 0.23; the question/stem of item [48] in French: 1738.9, German: 193.4, Italian: 121.2). This indicates that it may be prudent to not take a variable in consideration which is based on as little information as a handful of words and a non-exhaustive, language-specific reference corpus. A similar conclusion may be necessary with regards to the lexical difficulty values recorded for the NIs¹⁹⁶: these variables were only based on the content words in an NI known to the respective reference lists, and proved to be 0 more often than not. In these cases, it is just as likely that the few content words found in the NI were not found in the reference list as that they were classified as more difficult than A1 or A2.

Given these uncertainties, the lexical diversity, lexical sophistication and lexical difficulty values extracted for the NIs and the items parts were excluded from all further analyses with one exception (`LexSoph_tokens_NI`, which may be the most unproblematic variable given its distribution and the way it is operationalised). These variables are either difficult to interpret or simply missing due to the low number of (content) words of the piece of text on which they are based, or their correlation with the respective values of the input texts is strong enough to not provide much additional information.

11.1.4 Cross-linguistic variables

For each input text, NI and item part, two variables were used to operationalise the distance between the foreign language and the language of schooling: first, human raters decided for each content word in the input texts and items whether it could be considered a cognate recognisable by 6th-graders. Second, the mean Damerau-Levenshtein distance of multilingual matches of all content words in the respective languages was calculated and subtracted from 1 (see section 9.3.1.4). In both cases, a number between 0 and 1 represents the strength of the connection between the words in the text or item and the same words in the respective other language, but the variables measure this connection in different ways: while `CL_Cogn` is the proportion of content words which were considered to be transparent cognates by the human raters, the variable `CL_Lev` operationalises the mean string distance between some, but not all content words in a text or item and its translation equivalent according to Decaleon, with higher values representing smaller string distances, and, hence, more similarity.

11.1.4.1 Input texts and NIs

Depending on the language version, the proportion of cognates in the input texts averages between 20% (DEfr) and 40% (FRit), while the mean Damerau-Levenshtein distance of the content

¹⁹⁵To be precise, lexical evenness and lexical dispersion are often 0 when the analysed texts are short, whereas the MTLTD value could not be calculated in these cases.

¹⁹⁶Lexical difficulty was not extracted for the items since these were presented to the students in their language of schooling.

words known to Decaleon is between .295 (DEfr) and .520 (FRit). In both cases, the German-French and French-German versions have a lower, but similar amount of “cognateness”, while the English-German and French-Italian pairs show a stronger relationship between the two respective languages. For the NIs, these two groups are also observed: in DEfr and FRde, the number of NIs rated as containing a large proportion of cognates in their content words is quite small, and about a third of all NIs in both language versions do not contain any cognates at all¹⁹⁷. Inversely, the NIs in the ENde and FRit language versions have generally larger proportions of cognate words, and only a small number is entirely devoid of cognates. For the Damerau-Levenshtein variable, the numbers are different, and the distributions more normal, but there, too, the ENde and the FRit language versions have generally higher values, and the distributions of the FRde and DEfr versions are skewed towards the left, which means that their average string distance is higher (i.e. the words are less similar).

A comparison of the language versions reveals an important difference between the two variables: while the mean string distances of the content words calculated from the Decaleon multilingual database are positively correlated between all language versions (though to differing extents), the proportions of human-rated cognates among the content words are not (see fig. 11.2). The FRit language version in particular has a different pattern of cognateness than the other three language versions¹⁹⁸.

An examination of the Damerau-Levenshtein distances across the language versions, particularly of the plotted relationships (fig. 11.2(b)), reveals similar patterns, though there is also a visible connection between FRde and FRit that is not observed in the human-rated variable. Given that the language of the input texts is the same, it is possible that this connection stems more from the vocabulary available in the Decaleon database than from actual relationships between the cognateness of the input texts in the two language versions.

11.1.4.2 Items

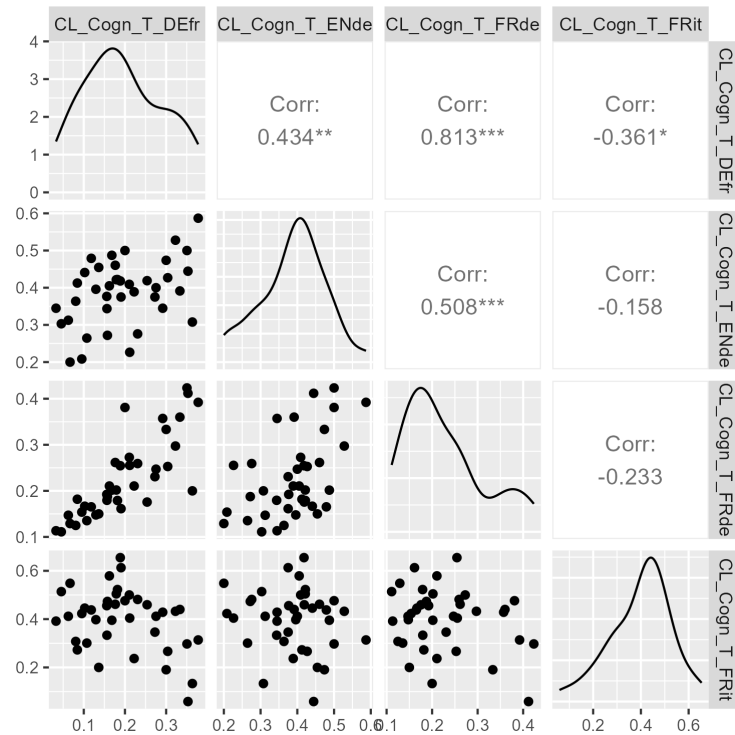
The pattern indicating that the English-German and the French-Italian language versions contain significantly more cognates than their French-and-German counterparts continues to be observed in the items. However, given the shortness of the questions and, more particularly, the answer options, the distributions of the two variables are more extreme. While a majority of questions and answer options in the FRde and DEfr language versions have no or few cognates, and low string distance values, a good proportion of the content words in the ENde and FRit questions and answer options were rated as cognates. Interestingly, while a number of questions and answer options in the English language version were rated to be composed entirely of cognate content words, the average string distance of the French-Italian language version is visibly higher, indicating a generally closer relationship between the (written) French and the Italian content words. This effect is less obvious with regards to the input texts and NIs.

11.1.5 Overlap variables

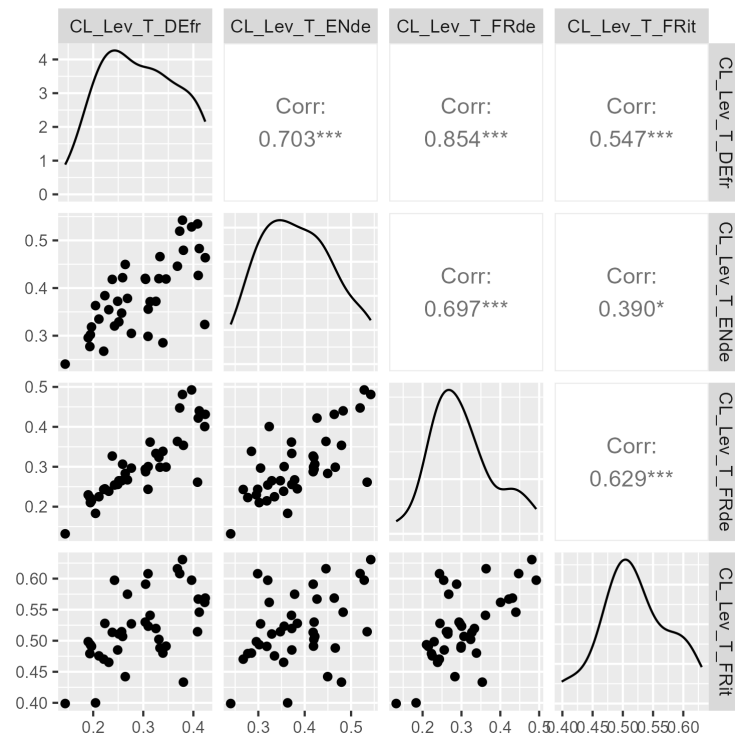
As was explained in section 9.3.3.4, lexical overlap was operationalised as an interaction between the NIs or the input text and the items, separated into the various item parts. Since the items were formulated in the language of schooling, the variables are approximations: translation equivalents according to the Decaleon multilingual database, somewhat loosely related orthographic

¹⁹⁷A bit less than a third of all NIs were rated to not contain any cognates (DEfr: 27 of 91; FRde: 24 of 90), while for most of the others, 40% or less of all content words were rated to be cognates (DEfr: 54 of 91; FRde: 46 of 90), and only a small group of NIs contains more than 50% of cognates among their content words (DEfr: 4 of 91; FRde: 10 of 90).

¹⁹⁸Though fig. 11.2(a) indicates negative correlations between FRit and ENde as well as FRde, these are not significant. The significant, but low positive correlation between FRit and DEfr also seems more of a mathematical artifact, given the corresponding plot.



(a) Proportion of human-rated cognates



(b) Mean Damerau-Levenshtein distance

Figure 11.2: Relationships of cognate variables across language versions: human ratings and mean Levenshtein-Damerau distance of content words in input texts.

neighbours, and the Damerau-Levenshtein distance of all content words were tentative operationalisations of what would otherwise be a relatively “simple” measure of how many words in the input text or NI are repeated in the item.

Overall, these measures seem to indeed capture at least part of the overlap between the NIs and their respective items: in the four language versions, about half of all items were found to contain at least one content word in the stem which can also be found (in its translation) in the NI. Formal overlap, i.e. the existence of formal neighbours in the two assessment parts, is usually lower but still non-negligible¹⁹⁹. Overlap with the input text is most often even higher.

Examining the overlap variables of the answer options with the NI and the input text reveals one most remarkable observation: a majority of all items in the four language versions were found to have semantic overlap between the *correct* answer options and the NI, whereas the same observation cannot be made for the mean overlap of the *incorrect* answer options. Furthermore, a comparison of the two variables reveals that there are very few items which have as much or even more semantic overlap between the *incorrect* answer option and the NI as between the *correct* answer option and the NI (see the scatterplot at the lower left of fig. 11.3).

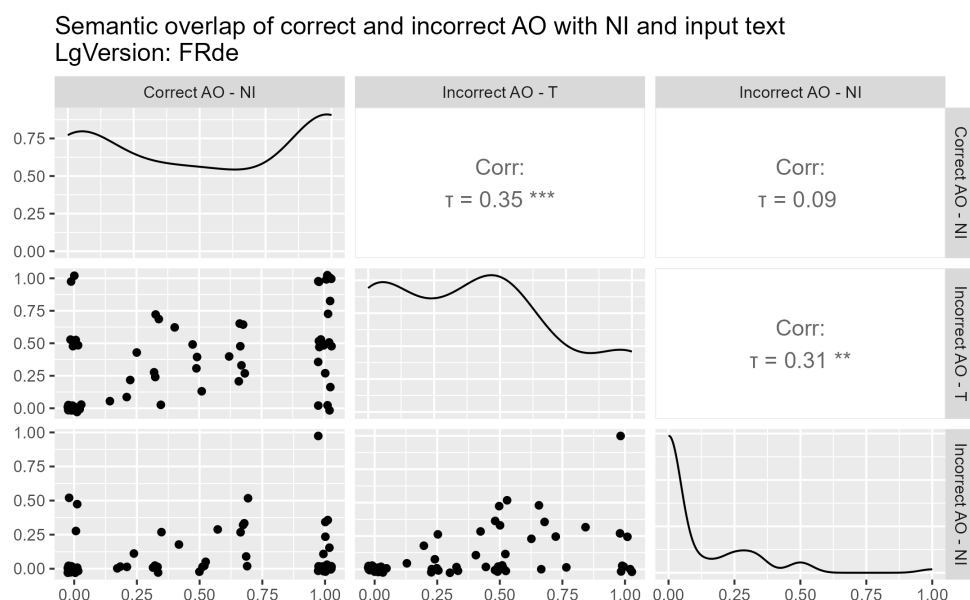


Figure 11.3: Semantic overlap of correct answer option with NI, compared to semantic overlap of incorrect answer option with NI and input text. Language version: FRde. All other language versions are similar.

This latter observation, of course, is not entirely surprising: since the correct answer can be found in the NI, it should be expected that these two text sections share certain content words. Accordingly, when the overlap of the incorrect answer options and the *input text* is considered, the effect disappears: there are visibly fewer items with more semantic overlap between the correct answer option and the NI than between all the other answer options and the input text (see the second row of the scatterplot in fig. 11.3). This finding also confirms that the semantic overlap variable actually measures overlap between item and input text or NI.

Finally, the variable `Overl_dist`, independently of which overlap it operationalises, is centred around .13²⁰⁰, and shows very little meaningful pattern: the correlation with the other two variables is very low and most often non-significant, the values of individual items show no

¹⁹⁹Semantic overlap (`Overl_sem_NI_Q`) of more than 0: DEfr 45 of 90 questions/stems, ENde 38/90, FRde 40/89, FRit 50/89; formal overlap (`Overl_form_NI_Q`) of more than 0: DEfr 33 of 90 questions/stems, ENde 36/90, FRde 29/89, FRit 42/89.

²⁰⁰The standard deviation of the overlap variables including the longer input texts is .02 across all languages, whereas the overlap variables including the shorter NIs is larger (.05).

relationship among each other, and not even the language version seems to influence this value in any particular way. Apparently, a mean string distance between the content words of texts in different languages is not sensitive enough to detect differences in the amount of cognates.

Given all of these observations, only the semantic overlap variable has been retained for the statistical analyses.

11.1.6 Changes to the variables

Based on the various observations made in this phase, several columns were eliminated from the data set or were recoded so as to facilitate subsequent analyses. First, several categorical variables were recoded as explained in section 11.1.1. Second, the lexical difficulty variables, which represented the number of A1 or A2 words in the input texts and NIs, were replaced by proportions. Third, all variables which had been extracted for each individual answer option were recoded to either reflect the value of the correct answer option, or the mean of the values of all incorrect answer options. This decision was made in view of the somewhat difficult interpretability of the individual values. Fourth, the variable `StemType`, which does not have enough variability, two of the three overlap variables and the lexical diversity, sophistication and difficulty variables for the NIs and item parts were eliminated due to their low interpretability given the short texts on which they were based. Finally, some administrative columns which had been in the data set from the beginning, but were not used in any way, were eliminated at this point as well.

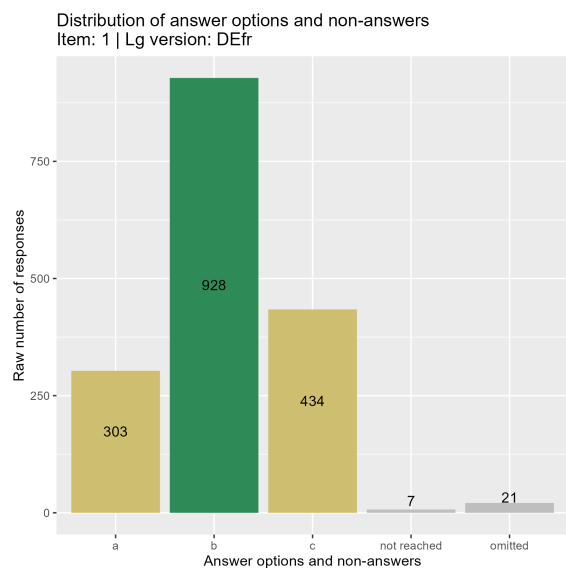
11.2 Item parameters

As described in section 10.2.1, the test responses and item parameters were examined with a focus on identifying items which stand out in some way. First, the response patterns of all items were examined graphically, and anomalies will be highlighted in the following. Next, items which are either very easy or very difficult or which have an unusual point-biserial correlation according to CTT will be discussed. Finally, the estimated item difficulties and fit indices from the MFR model will also be investigated. In many cases, some qualitative interpretation of why the item may have behaved in this particular way is offered. This is, of course, merely tentative, and it is based on personal experience and impressions from more systematic analyses of qualitative interviews on test solving processes (e.g. Barras 2021). It seemed important to offer these thoughts at this point, however, as it proved impossible to retrace them in the inferential analyses.

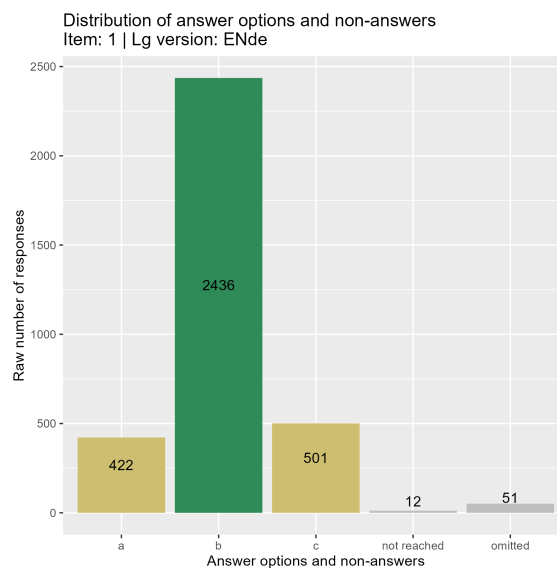
11.2.1 Response patterns

In general, items in the ÜGK L2 reading assessment were solved correctly by a majority of students in all curricular regions. Often, the FRde and the DEfr version show higher proportions of incorrect answer options than the ENde and the FRit version, and in all cases, a small group of students chose to not answer the item, or did not reach the item in the allotted time frame. Only very few answer options that were available were not chosen by any students, though in some cases, incidences were extremely low. It should also be pointed out that among the MC items with three answer options, the two incorrect answer options are, with exceptions, often both similarly attractive, i.e. students who answered incorrectly were not overwhelmingly drawn to one answer option. This is generally considered a characteristic of “good” item writing. Figure 11.4 shows a typical example of the test response patterns of this type.

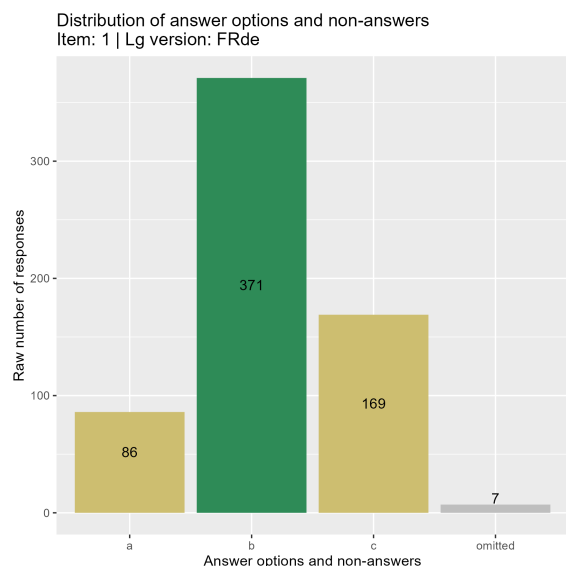
Naturally, among the 4×94 items, there are those which deviate from this “typical” pattern. For instance, students in the *Passepourtout* region, who were assessed in French (language version FRde), chose the distractor *c* of item [2] more often than the correct answer *b* (a: 56, b: 273, c: 291, *omitted*: 13), while students in the other curricular regions generally answered correctly, and those who did not are distributed more equally across the two incorrect answer options (e.g. DEfr: a: 249, b: 936, c: 442, *omitted*: 59, *not reached*: 7). It is not entirely clear why this group



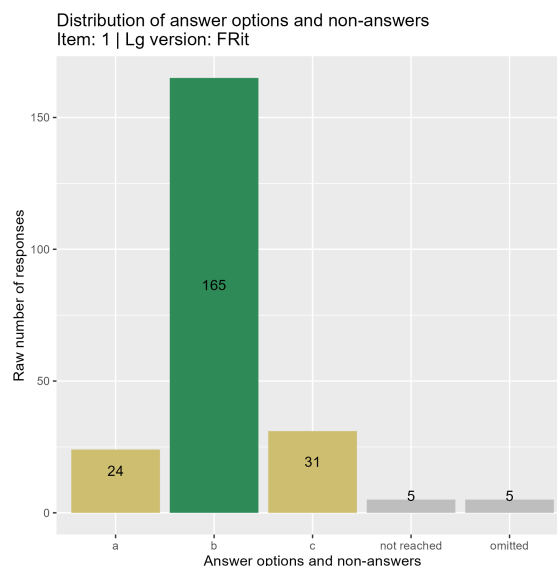
(a) Foreign language: German | Language of schooling: French



(b) Foreign language: English | Language of schooling: German



(c) Foreign language: French | Language of schooling: German



(d) Foreign language: French | Language of schooling: Italian

Figure 11.4: Response pattern of item [1], which is typical for a multiple-choice item in this assessment. Most test takers chose the correct answer, the incorrect answer options are similar in their attractiveness. Visibly, both the ENde and the FRit language versions were solved correctly more often.

of students made such a different choice than the others: it may have something to do with understanding the word *flamant rose* in the input text, but the difficulty may also be linked to correctly identifying the French word *tu chantes* instead of *tu cries* to match to the item question's *wer singt*. In any case, the students in the Ticino region, who read the same text, did not seem to encounter the same problem. Item [62] may be a similar case, where the phrase *écouter de la musique* in the input text was much more attractive than the correct *dormir*. In this case, however, all items of the FRde language version related to this input text are notably more difficult than in the other language versions, perhaps because the topic (and, with it, the

vocabulary) is unfamiliar to the students of this curricular region.

Another item which shows unexpected response patterns is item [15], to which a majority of students in all but the English language version answered **b** instead of the correct **c**, particularly students from the Ticino region. This may simply be quite a difficult item given that the incorrect answer **b** is mentioned in the NI, but, for the Ticino students, perhaps the French word *parents*, which is quite unrelated to the Italian word *genitori*, may also have played a role. A similar phenomenon may explain the response pattern of the DEfr version of item [17]: if we consider the German word *Sommer* to be somewhat more transparent to French-speaking learners of German and English than the French word *été* to German speakers who are learning French and English, then the incorrect answer option involving summer may simply have been more attractive to the students in Western Switzerland. The Ticinese students' choice of answer option **c** for the item [66] may also be related to its attractiveness: it contains the cognate *tempeste* for the word *tempête* in the input text, and is by far the longest answer option (though only in this language version). A final example of this may also be the quite difficult item [76], where in the language versions involving French, German and English most test takers chose the incorrect answer **c**, which involves the word *offen* - *ouvert* - *open* in the item and the input text, while in the FRit language version, **b** was the most attractive distractor. This answer option contains the word *coraggiosa* and can be matched to *courageuse* in the input text.

An overwhelmingly attractive distractor is seen in item [37] in all four language versions. In this item, students were asked to decide whether a reviewer found a restaurant good or merely mediocre. Since many students chose “good”, this particular review was clearly not obviously enough “mediocre”, no matter in which language. A similar observation can be made for [47], which is also a mediocre review that was considered “good” by many students, particularly in the DEfr and the ENde language version.

Finally, a perhaps “classically” difficult item is [69]: each answer option cites a number from the input text, but the correct answer can only be found if the test taker understands the sentence surrounding the correct number. Both in the ENde and the FRit version, most students succeeded at this, but the option most chosen by the FRde and DEfr test takers is the more attractive wrong answer option in all four language versions—perhaps because it can be easily matched to the input text across all language versions due to its wording. The next item [70] seems to follow a similar logic, though here, the Ticinese students could match the word *handicapés* in the input text to *handicappati* in the item stem. This probably helped them solve this item at a much higher rate than in the other language versions, in which this word was not a cognate.

Items which are not MCQ with 3 answer options, but either MCQs with more than three answer options or Matching items, usually follow the “typical” pattern described above: the FRde and the DEfr language version have higher incidences of incorrectly chosen answer options, but the response patterns are otherwise similar to the other two language versions. In some cases, one incorrect answer option is much more attractive than the others (e.g. [55], [29], [5]), or the item is generally very easy (e.g. [27], [18], [34]). Among these non-MC items, item [35] has a somewhat particular response pattern in the DEfr version in that students found the correct answer much less often than in the other language versions, perhaps due to the somewhat formal key word *Ort* in the correct answer option. It may be less known or transparent than the French *localité* or the English *town*.

One final interesting aspect of item response patterns are non-responses. In the ÜGK data, three types of non-responses were recorded: “missing by design”, which means that students did not see the task at all, “omitted”, which means that students saw the item, but chose not to answer it, going on to the next item instead, and “not reached”, which means that due to time constraints, the student should have seen the item but did not. The latter always concerns entire tasks and happened quite rarely: less than 2% of all response sets fall in this category. Since the order in which students encountered “their” four tasks is not known, this type of non-answers, and missing-by-design-items, do not contain any relevant information for this analysis.

Hence, only items which were deliberately²⁰¹ omitted by students are considered here, particularly those which a larger number of students than normal chose to skip. On average, only 3% of test takers chose not to answer any item across all language versions ($\mu = 0.036$, $sd = 0.026$), and students tended to omit items at the end of a task more often than the first item. Across the four language versions, 33 items were not answered by more than 7.5% of all test takers, and 11 items by more than 10%. The highest incidence of omitted answers is 14.7% ([21]_FRit, a Matching item). Incidentally, the majority of items which were skipped more often, belong to non-MC items: the mean proportion of omitted answers of all but the first item of task LV02G is .083, and of task LV04G .099. The items of the tasks LV05G, LV06G and LV11G were also skipped more often than others, though to a lesser extent. All of these tasks feature Matching items with more than three answer options. Quite clearly, if there are more answer options, more test takers of this target group (pupils in 6th grade) will refrain from answering instead of choosing an answer by guessing or at random. An omitted answer, then, at least in this assessment, is probably the result of an item's format rather than of its difficulty.

11.2.2 CTT item parameters

In all language versions, item difficulty according to CTT²⁰² is reasonably well distributed, with a majority of all items situated at a medium difficulty level between $p_i = .5$ and $p_i = .7$ in all language versions. The easiest items were solved correctly by approximately 90% of the test takers, while the hardest items were only solved correctly by around 20%, though both groups of items are quite small. The ENde language version is an exception in that even the hardest item has a p-value of .35 (item [76], see also table 11.4). On average, this language version also has higher values of *proportion correct*, which implies that a larger proportion of the students who took this test performed well compared to the other curricular regions. The same is true, though to a lesser extent, for the FRit language version.

Language version	mean p_i	sd	min	max
DEfr	.57	.17	.17	.89
ENde	.69	.13	.35	.93
FRde	.58	.16	.19	.90
FRit	.68	.15	.23	.94

Table 11.4: Descriptive statistics of CTT's proportion correct by language version

The point-biserial correlation²⁰³ of these items is generally acceptable, averaging around .35. The lowest point-biserial correlation is actually negative, though there is only one such item ([66]_DEfr, $p_{[66]} = .49$, $\rho_{pbis[66]} = -.048$). Another twelve items, eight (!) of them from the DEfr test version, have point-biserial correlations between 0 and .1. These would, in a CTT paradigm,

²⁰¹Technically, it is possible that some of these omitted items were not skipped deliberately but that students either clicked the *Next*-button twice by accident or skipped items due to boredom, nervousness or refusal to participate. Given the low overall occurrence of omitted items, these cases are probably negligible.

²⁰²In CTT, item difficulty is operationalised as the proportion of correct answers given by the test takers who provided an answer to the item (see also section 6.2.1.1). Given the rotation design and the high number of items missing by design, this means that these values are not really comparable, though the large number of test takers who solved each item partly makes up for this. High proportion-correct values indicate easy items, low values describe difficult items.

²⁰³This value represents the correlation between the responses to an item and the total score of all other items. A value close to +1 would indicate that students who solved an item correctly also scored high, whereas a value close to -1 would suggest that only students who scored low solved an item correctly (or, inversely, students who did not solve the item correctly performed otherwise well on the test). Values close to 0 indicate that there is no such relationship. According to Moosbrugger and Kelava (2008, p. 84), point-biserial correlations between .4 and .7 are considered "good" item discrimination values. See also section 6.2.1.1.

be examined closer in order to identify potential misfitting items. Perhaps the most interesting observation among these is that the items [64] to [66], all with low point-biserial correlations in the DEfr version²⁰⁴, refer to the same input text. The item response patterns of this subtask, and of the second subtask in this task (items [60] and up), have already been found to be unexpected in several ways (see above). Apparently, there is something about these two tasks which makes them behave in unexpected ways in the four curricular regions.

11.2.3 IRT item parameters

As was described in section 10.1.2, an MFR model was estimated using the language versions and all items as a facet. Item difficulty values extracted from this model are on the same scale, and, therefore, comparable across language versions. These values are actually composed of three separate “types” of item difficulty: each item has one single “multilingual” item difficulty, to which the item difficulties for the language version facet and the item facet are added.

Not taking into account these corrections for language version and item, item difficulties range from -2.65 logits to 1.41 logits ($M = -.74$, $SD = 0.77$, $N = 94$). The item difficulty value for the language facet, which is common to all items of one language version, imply that ENde and FRit items are generally about 0.3 logits easier, whereas FRde and DEfr items are generally about 0.3 logits more difficult than the average (ENde: -0.35, FRit: -0.30, DEfr: 0.36, FRde: 0.29). Between these two groups, then, there is a difference between 0.6 and 0.7 logits. Finally, each individual item has an additional, language-version specific item difficulty component. These are centred around 0, with most item corrections ranging between -0.8 and +1²⁰⁵. The DEfr language version contains one item that has an additional item difficulty of +1.95 logits (item [35]), and the corrections of FRit items range from -1.41 to 1.17. Taking all of this together, table 11.5 shows the statistics of the aggregated item difficulties in each language version.

Language version	mean	sd	min	max
DEfr	-0.38	0.89	-2.32	1.84
ENde	-1.09	0.79	-3.09	0.78
FRde	-0.45	0.84	-2.46	1.69
FRit	-1.04	0.93	-3.32	1.50

Table 11.5: Descriptive statistics of item difficulties by language version as estimated in an MFR model. All numbers are expressed on the same logit scale.

The aggregated item difficulties give the impression that most items are on the easier side: mean item difficulty is below zero in all language versions, and the hardest item is always closer to zero than the easiest one. *Zero*, in this model, represents the mean test taker ability across all language versions: the EAPs²⁰⁶ estimated by the model are centred around zero.

This observation is reinforced by the Wright maps depicted in section B.1.7 (page 283). These show that, overall, the language-version specific item difficulties in this assessment are well distributed, with some very easy and some very hard items, but a lot of items with medium difficulty values in all languages. In fact, most items are situated in the lower-medium difficulty range, which aligns with the assessment’s objective: to check whether students reach the *basic* competence level in their first foreign language (Konsortium ÜGK 2019a).

²⁰⁴[64]_FRde and [66]_FRit also register low ρ_{bis} values.

²⁰⁵Note that these individual deviations are larger on average than the correction of the language facet. This implies that item difficulty is more dependent of the item itself than of its language version.

²⁰⁶EAPs are one way of estimating person abilities in an IRT model. They are an alternative to the more well-known WLEs, though in a large study such as this, the two ability estimation methods lead to very similar results (cf. for an overview Lechner et al. 2021).

Across all four language versions, the following items are particularly easy: [25], [34], [38]. All three items can be solved correctly by relying on individual words which are cognates across the four languages (*polar*, *super*, *date*). Also, in the case of [34] and [38], the answer can be found at several points in the input text. Item [58] is the easiest item of the FRit language version, and among the easiest in the ENde version, but it is only among the easier half of all items in the FRde and DEfr language versions. This item, along with its predecessor [57], also much easier in ENde and FRit, relies on understanding words like *hiver/inverno* and *winter/Winter*, *vent/vento* and *wind/Wind* or *neige/neve* and *snow/Schnee*. Clearly, these words are much more transparent across the Italian-French and German-English language combinations than between German and French. Nevertheless, overall, these items are quite easy, perhaps because the input texts contain redundant information leading to the correct answer.

On the other side of the continuum, [15] and [76] are (among) the most difficult items in all four language versions. This is interesting in that in both cases, the incorrect answer options are mentioned in the same or the directly adjacent sentence to where the correct answer can be found in the input text. Here, very precise reading and understanding of a passage are necessary to decide on the correct answer, in any language. Similar observations can be made for the equally difficult items [53] and [93]. However, there are additional difficult items in each language version which are less so in the others. For instance, item [35] is the most difficult item in the DEfr language version, while it is of medium difficulty in the ENde and FRde version and even quite easy in the FRit version. This item, requiring knowledge of the German word *Ort*, has already been observed in the analysis of the response patterns above. Item [66] has also been mentioned above: it is among the most difficult items in the FRit version, perhaps due to a strong distractor that is linked to the cognate *tempeste/tempête*.

The results of the fit analyses from the Rasch models, as explained in more detail in the R script `2_VariableExtraction_rsd`, identify a small number of items which have a significant positive infit value²⁰⁷ in all four language versions ([12], [41], [47], [66], [79] and [81]). Most of these items have at least one incorrect answer option that was chosen by a lot of students—given the misfit, possibly for reasons that do not coincide with the construct measured otherwise. However, eliminating them does not change much in terms of the overall item difficulty values extracted from the Rasch models, so they were not deleted from the data set. For the MFR model, a more comprehensive fit analysis was conducted (see section B.1.6, page 282). For the purposes of this study, both item and person fit were deemed acceptable.

11.2.4 DIF across language versions

In the above considerations of item response patterns and IRT item difficulties, some cross-linguistic aspects have already been mentioned. However, as was mentioned in section 10.1.2, the MFR model allows for a more detailed account of how much individual items differ across the language versions: overall, 304 out of 376 items are significantly different from their mean item difficulty even though the language version itself is controlled for²⁰⁸. This implies that the items in the four language versions differ from each other quite significantly in terms of item difficulty, and not systematically along the lines of the language versions. Considering the cross-linguistic variance in the item characteristics described in section 11.1.2, and some of the more qualitative observations on item response patterns and item difficulty estimates made above, this seems hardly surprising.

²⁰⁷See section 6.2.2 for a short explanation of fit values in IRT.

²⁰⁸Very simply put, any item's difficulty is composed of three elements: its difficulty across the language versions (b_i), the difficulty of the language version as a facet (b_l^{language}) which is the same value for all items of a language version, and the difficulty of each item in each language version as a facet ($b_{i,l}^{\text{item}}$). This latter "remaining" item difficulty is what is of interest here. It is divided by its standard error to create a standardised Z-score. This contains information about the significance of the distance to the mean item difficulty of an item in a language version.

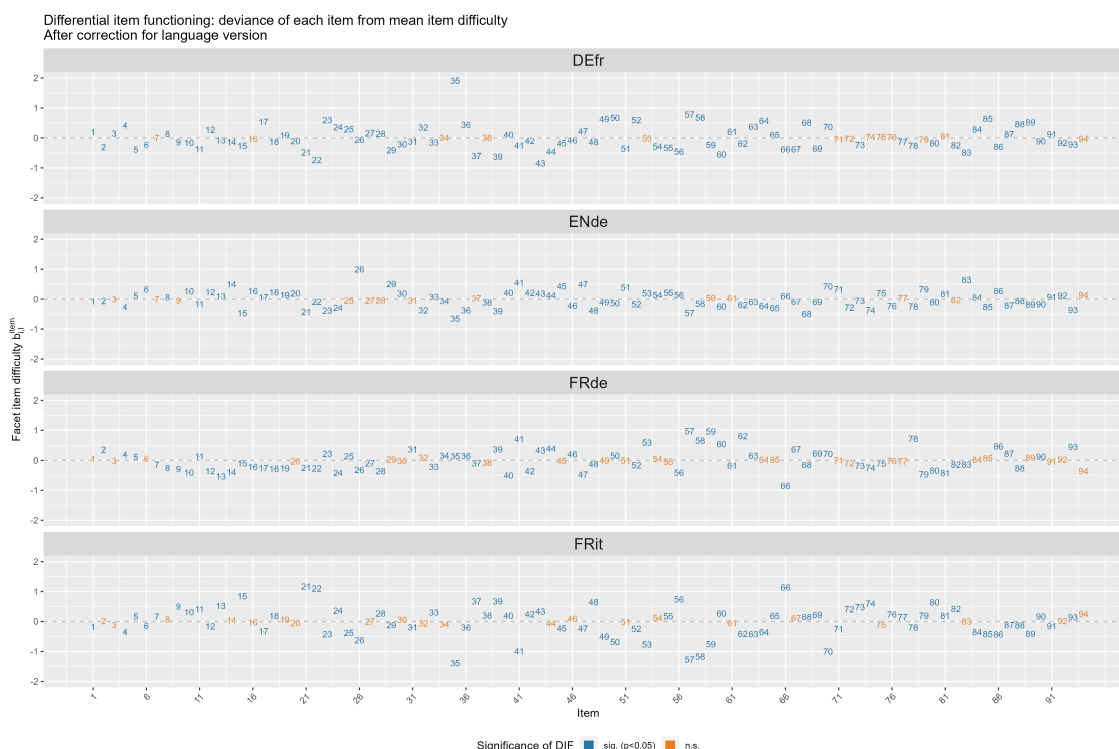


Figure 11.5: Graphical depiction of DIF across languages: facet item difficulties $b_{i,l}^{item}$ per language version. The more an individual item diverts from the zero line, the more its item difficulty is different from what can be considered its common multilingual item difficulty, corrected for the language version.

In fig. 11.5, a graphical representation of these item-by-language-specific difficulty values, essentially their amount of DIF, is attempted, including an indicator of their statistical significance. In addition to showing that *some* amount of item difficulty has to be added or subtracted from *most* items, the figure allows for a comparison between language versions: it is obvious that each language version has its own set of items which deviate more from the mean than others. The graph also shows that there is no discernible pattern to the items showing DIF in the four language versions: no two versions are similar, not even the FRit and the FRde version, and, with a few exceptions, there is no visible task effect (i.e. adjacent items do not behave more similarly than others).

Most of the extreme items in these graphs have been discussed above. Item [35] is, by far, the most divergent item in the DEfr language version: it is the most difficult item there, easier in FRde and ENde, and among the easiest in the FRit language version. The items [57] to [59] also tend to diverge visibly in all language versions, though, of course, in opposite directions. As explained above, these rely on knowledge of vocabulary related to the weather which is shared by French and Italian and by English and German, but not across the two groups. Item [66] is visibly different for the two groups who sat the French reading assessment. This is the item with the (incorrect) *tempête/tempeste* answer option that may have been overly attractive to the Italian-speaking test takers in the Ticino. Inversely, items [41] or [70] may have been much easier to these students due to relevant vocabulary being shared by the French and Italian language. Finally, the two items [21] and [22] were, again, unexpectedly difficult to the Ticinese test takers, perhaps because they were less familiar with the required vocabulary than their peers in the other curricular regions.

11.2.5 Item difficulty across language versions

Despite the variety in item difficulties across languages described in the above section, it is, of course, still interesting to consider how the item difficulties in the four language versions correlate with each other. This is best depicted in fig. 11.6. The conservative Kendall rank correlation coefficient for each language combination ranges from $\tau = .42$ (FRit and DEfr) to $\tau = .62$ (FRde and DEfr), which implies a medium to strong relationship between all language versions. The four scatterplots in fig. 11.6 support this: in general, the difficulty of items is in the same general zone in all language versions, but the relatively large band in which the items are situated shows that at any difficulty level, the item's difficulty in other languages may well vary by up to 1.5 logits, which is, of course, in line with the facet difficulties described above. The graphs also show that the FRit version differs more from the other three, and especially from the DEfr version, which was already implied by the larger item-specific difficulties visible in fig. 11.5 and is probably due to the lower number of test takers who solved the items in this language version.

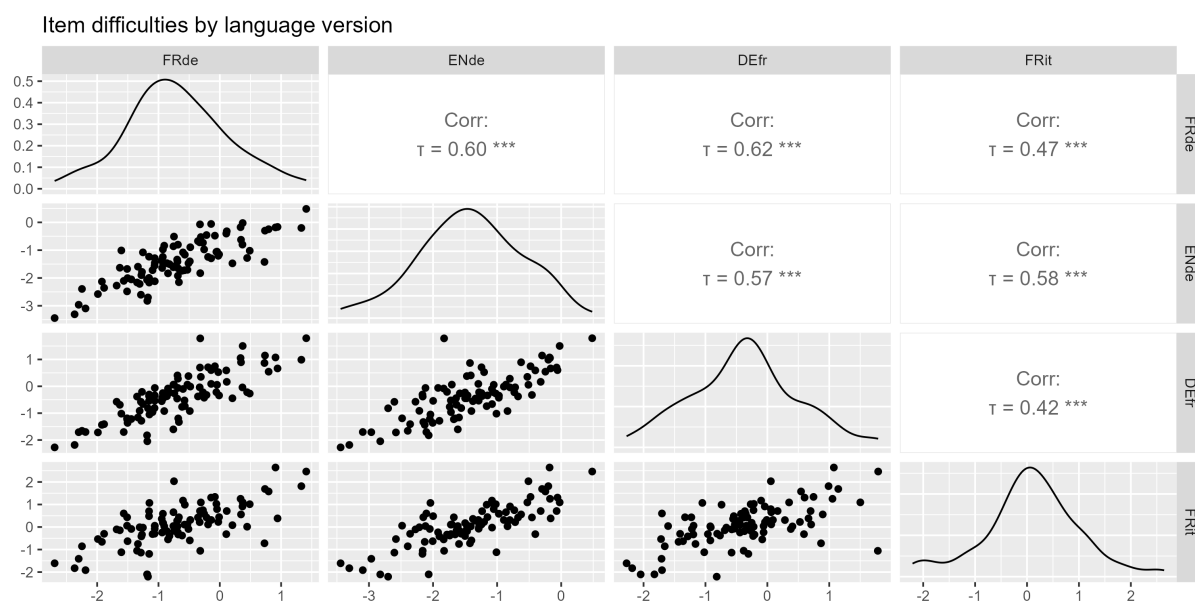


Figure 11.6: Estimated item difficulty in each language version in comparison (WLEs extracted from MFR model)

11.2.6 Conclusion

All item analyses confirm the non-systematic divergence of item difficulties across the language versions, but they also underline the generally good psychometric quality of the overall assessment. Items which are somehow particular in one or several language versions have a tendency to be identifiable by various means, and they often feature cognates of some kind. The results of the MFR modelling will be used in all subsequent analyses given that they are on the same scale across the language versions.

11.3 Person attributes

The ÜGK foreign language assessment was solved by 20,177 test takers as described further above. In the following, their performance at this assessment is briefly described and some information on the additional test taker variables is given. Since this study focusses on the items and their characteristics, the following section is very brief. Some more detailed analyses of the test takers

and their performance in the ÜGK assessments can be found, for example, in Konsortium ÜGK 2019a and related publications.

11.3.1 Performance at the assessment

The Wright maps depicted in section B.1.7 (page 283) contain information on the person ability estimates output by the MFR model²⁰⁹. They show that in all language versions, the majority of test takers performed above the difficulty value of the majority of the items. Also, there is a more or less small group of students whose performance surpassed even the most difficult items. The latter effect is particularly strong in the ENde language version and perhaps the least visible in the DEfr language version. Inversely, there are only very few students who performed *below* the difficulty level of even the easiest items in all language versions. All of this is in line with the objective of the assessment, which was to assess a minimal basic competence level that all students are supposed to have reached.

Table 11.6 reports the descriptive statistics of the WLEs by language version. For the entire MFR model, the person ability estimates range from -4.26 logits to 4.34 logits ($M = 0.01$, $SD = 1.17$, $N = 20,177$). The WLE reliability of this model is $.788$, which is reasonably good²¹⁰.

Language version	mean θ	sd	min	max
DEfr	0.00	0.94	-3.16	4.34
ENde	0.02	1.29	-4.26	3.55
FRde	0.01	1.07	-3.68	4.33
FRit	0.00	1.19	-3.18	3.63

Table 11.6: Descriptive statistics of person parameters by language version as estimated in an MFR model. All numbers are expressed on the same logit scale.

11.3.2 Test taker variables

As was mentioned earlier, the 20,177 test takers in the data set were on average 12.5 years old on the day of testing, and there is an even distribution of boys and girls. Since the language groups are unevenly distributed in Switzerland, and the test takers are a representative sample of the population on the cantonal level, the numbers of test takers for each language version differ quite significantly (see table 9.2, page 136).

In all curricular regions, roughly 30% of the test takers have an immigration history, either because the students themselves and/or both their parents were born outside of Switzerland²¹¹. Among these, there are about twice as many students who were born in Switzerland themselves, but not their parents. A χ^2 test indicates that the distribution of non-immigrant, 1st- and 2nd-generation immigrant students is significantly different in the four curricular regions, though Cramer's V indicate that this difference is not large ($\chi^2 = 197.5$, $df = 6$, $p < 0.001$, with Cramer's $V = 0.07$). A look at the observed and expected values from the χ^2 test shows that there are

²⁰⁹As was mentioned in section 10.1.2 and demonstrated at length in the R script `2_b_VariableExtraction_rsd`, the person ability estimates produced by the MFR model are virtually identical to separate Rasch models. This is despite the fact that one could argue that the four language versions measure four different constructs and therefore should not be estimated in one IRT model. Mathematically, at least, an MFR model with the language version and all items as facets will perform no differently than four separate Rasch models, but has the advantage of using the same scale for the item and person parameters in all four language versions. For this reason, the MFR estimates are reported here.

²¹⁰WLE reliability can be interpreted in the same way as traditional reliability measures in CTT (R. J. Adams 2005). Hence, a reliability above $.7$ is considered acceptable, above $.8$ is good, and above $.9$ would be excellent.

²¹¹The specific numbers and calculations can be found in the R script `3_c_DescriptiveStatistics_ttd`.

slightly more than expected students *with* an immigration status in Western Switzerland (35%), and more than expected students *without* an immigration status in the *Passepartout* region (81%), though, again, while significant, the differences are not large.

A similar observation can be made for the number of students who use the foreign language they were being assessed on as a home language. The proportion varies significantly, but not strongly, among the curricular regions ($\chi^2 = 50.32$, $df = 3$, $p < 0.001$, with Cramer's $V = 0.05$). In the ENde and FRit test version, about 5% of all test takers indicated that they used English or French respectively at home. In the FRde language version, this proportion amounts to 9.5%, and in Western Switzerland, about 6% of the test takers use German or Swiss German at home.

Finally, given the way PVs are created, not much can be said about the student's test results on the other ÜGK assessments they sat that day (L2 listening, L1 reading, L1 orthography). PVs based on WLEs are naturally centred around 0, and range from -2 to 2 . Since they were estimated separately for each language, the distributions in the various language versions do not allow for any particular interpretations.

11.3.3 Test taker variables in the IRT model

In an exploratory analysis, four test taker variables were added to the MFR model as covariates (ImmSt, L2Home, L1Read, L2List). The resulting model still shows good WLE reliability (.76), though it has not improved compared to the original MFR model. The covariates explain 86% of the variance in latent L2 reading ability. By far the strongest predictor of L2 reading ability in this assessment is the mean PV from the L2 listening assessment (standardised $\beta = 0.77$), followed by the estimated L1 reading ability (standardised $\beta = 0.37$) and a binary variable of whether the target language is spoken at home (standardised $\beta = 0.36$). The migratory background of the students is a minor predictor (standardised $\beta = 0.07$). In all cases, person ability goes up when the variable has a higher value.

Unsurprisingly, the model has a significantly better fit than the MFR without covariates ($\chi^2(1) = 128424.1$, $p < .0001$). The item difficulties extracted from both models are virtually identical²¹². The WLEs estimations in the model including covariates have not changed within-group either, but they have been shifted on the logit scale: the estimations for the FRit group in particular are shifted up by half a logit, while the ability estimates of the participants from the other three curricular regions are slightly lower in the covariate model, by up to 0.4 logits, with the FRde person ability estimates the least changed. These shifts are probably due to group differences in the covariates, though it is unclear whether these are “true” differences or whether they are caused, e.g. by the separate estimations of the PVs in the L2 listening assessment or the L1 reading assessment. In any case, since the item difficulty estimations are identical even after the inclusion of the test taker variables, all inferential analyses will be conducted on the results of the simpler MFR model described above.

11.4 Summary

The ÜGK foreign language assessment is composed of 94 items linked to 39 input texts in four language versions. Two thirds are MCQs, while 34 items in each language version are in a somewhat different format. The correct answer is distributed over all answer options available across the assessment. The majority of items require answers which can be drawn directly from the text, either verbatim or by rewording or restructuring the information provided. Inferencing is needed much less often. In most cases, the answer can be found in one specific passage of the input text, and background knowledge is almost never required.

²¹²Even though common sense may sometimes suggest that more information about the test takers should influence the estimation of item difficulty, this is not modelled in IRT estimation.

The input texts are most often expository, though there are a bit less than 20% of primarily argumentative, narrative and communicative texts each. About one third of the input texts are either lists or loose pieces of text, short texts of only a few sentences or else longer, continuous texts. Text length ranges from 23 tokens to 257 tokens (mean: ≈ 110) across the three foreign languages. The incidence of the various parts-of-speech (nouns, verbs etc.) varies across input texts along text genres and text continuousness. The length of the NI and of the various item parts is often quite short, which lessens the meaningfulness of many of the linguistic variables related to them to some extent. Across the three foreign languages and the three languages of schooling, most of these indicators are moderately or strongly correlated with each other, though when the piece of text is very short, the crosslinguistic correlations tend to be lower.

The lexical variables in this study are supposed to operationalise the lexical diversity, sophistication and difficulty of the input texts and other assessment parts. They usually show a large range of values across the input texts and a skewed distribution, and often vary significantly across the language versions. In many cases, the French and German input texts seem to be more similar to each other in terms of their lexical characteristics than the English texts. For the shorter NIs and item parts, these variables are often less meaningful since their calculation is subject to large fluctuations when there are only a few tokens or content words to draw from. As a result, these variables were eliminated from further analyses.

The crosslinguistic variables clearly reveal that the FRit and the ENde language versions contain more cognates in all assessment parts. This finding confirms prior assumptions about the lower difficulty of the reading assessment in English and in the Ticino version given the closer relationship between the two language pairs. The two variables in use, a human-rated and a semi-automatic operationalisation, also show some differences in the way they are operationalised, putting the crosslinguistic comparability of the semi-automatic CL_Lev variable into question. A similar observation can be made for the overlap variables which operationalise the proportion of content words in the item parts that can be found in the input text and NI as a translation equivalent. Out of three variables, only the one measuring actual translations has been retained for the analyses, as formal overlaps or string distance did not produce reliable results. Overall, overlap between item parts and the NI was found in about half of all items, with more overlap identified between the correct answer options and their respective NI.

Analyses of the response patterns, the CTT item parameters and the IRT parameters all confirm the generally good psychometric quality of the assessment. All items were answered correctly by at least some students and most of them were neither extremely hard nor extremely easy, though on average on the easier side. Item fit is generally good. The ENde and FRit language versions have lower item difficulties, both according to their raw response statistics and as estimated in the multilingual MFR model. Easy items across all language versions are often those which ask very specific questions found easily or even several times in the input text. Hard items seem to require detailed understanding of pieces of the input text. Across the language versions, item difficulties are correlated, but a large majority of all items shows significant DIF in at least one, often several language versions. Some items are particularly hard or easy in one language version, or their response patterns diverge notably. Often, impressionistic observations of these items point to words or expressions which are related or similar in the two languages involved in the assessment. If these words are related to the correct answer, then the item is easier in a particular language version. Inversely, some unexpectedly hard items contain prominent cognates in incorrect answer options.

The roughly 20,000 test takers generally performed well, which is expected given the lower-than-average item difficulties. There are, of course, low and high performers in all language versions. About 30% of all students have a history of migration in their family and between 5 and 10% of the learners in each curricular region use the foreign language at home. The data set also contains information on their performance on the foreign language *listening* assessment and the reading and orthography assessment in the language of schooling.

Chapter 12

Inferential statistics

The results of the inferential statistical analyses are presented in two stages. First, in section 12.1, the relationships among item, text, and text-by-item variables are examined to better understand patterns in the data and to ensure that multicollinearity is systematically addressed in the subsequent regression analyses. This section also includes the results of exploratory factor and cluster analyses, which, though ultimately unsuccessful, were intended to reduce the complexity of the data set. Second, several approaches to predicting item difficulty are presented in section 12.3. These include correlations between item difficulty and individual variables, followed by regression analyses conducted both on the separate language versions and on the combined multilingual data set. Finally, results from a complementary random forests analysis and exploratory Bayesian regression modelling are discussed. The latter incorporates test taker variables, offering additional insight into the complexity of predicting item difficulty.

12.1 Interactions between variables

As described in section 10.2.2, correlation matrices of all variables were established and visualised in heat maps. Given that the correlation matrices in the four language versions differ significantly from each other²¹³, this analysis was done separately for each language.

Figure 12.1 shows the heat map for the ENde language version²¹⁴. Very light yellow-greenish patches represent moderate and strong correlations between variables, while very dark blue patches show moderate negative correlations. The variables were ordered so that those which behave similarly are closer together. It is immediately visible in this heat map that linguistic variables which refer to the input text, the NI, the question or the answer options respectively cluster together. Where they do not, variables of the same type tend to cluster: the POS_NAM variables of all test parts can be found close together. The same is true for most of the various lexical sophistication indices and the two cross-linguistic variables CL_Cogn and CL_Lev. These patterns can be observed in all four language versions. A similar visualisation, though with more emphasis on the actual correlation coefficients can be found in the file `4_tables/CorrM_vars.xlsx`. This Excel file contains the correlation matrices of the four language versions in four separate sheets, formatted so that high correlation coefficients, both negative and positive, are emphasised. In some cases, this allows for a better detection of unexpected relationships between variables.

Overall, the correlation matrices indicate that there are groups of variables which correlate quite strongly with each other. These groups are discussed in more detail in the following sections.

²¹³The only two combinations of language versions which do not differ significantly from each other were the ones which shared a language, and, thus, have partly identical correlation values (FRit and FRde, and ENde and FRde respectively). This was determined using a χ^2 -test. For more details refer to the R-script.

²¹⁴The heat maps created by the `heatmaply` package are actually HTML widgets which can be examined in more detail by dragging the cursor over specific sections. Therefore, please refer to the R script `4_AnalyticalStatistics` for a more detailed look and for the heat maps of the other three language versions.

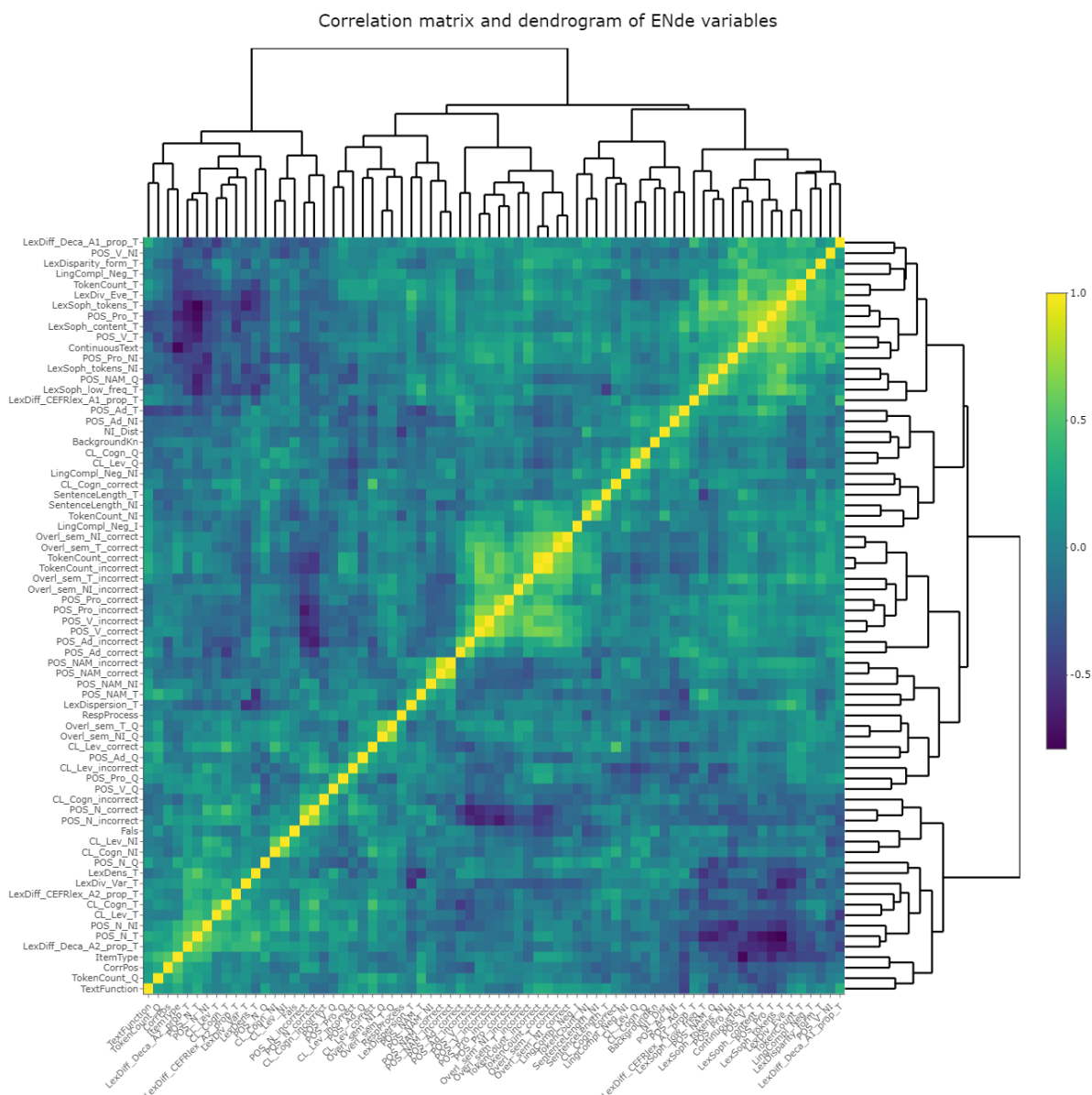


Figure 12.1: Heat map of the correlations of all text, text-by-item and item variables of the ENde language version. Created by the R package `heatmaply` (Galili and O’Callaghan 2024). The heat maps of the other language versions can be consulted in the electronic appendix (see also footnote 214).

To enhance readability, the following sections sometimes combine the presentation of key findings with a discussion of possible explanations. In some cases, this includes interpretations of results that are interesting in their own right, but not directly relevant to the prediction of item difficulty discussed later.

12.1.1 Categorical variables

The variables `ContinuousText` and `ItemType`, here interpreted as numeric variables, correlate strongly negatively with each other ($\rho = -.77$). They also correlate, though to a lesser extent, with `CorrPos` (`ContinuousText`: $\rho = -.45$ and `ItemType`: $\rho = .52$) and with `TokenCount_T`

(ContinuousText: $\rho \approx .61$ and ItemType: $\rho \approx -.40$ ²¹⁵) as well as with various other variables which describe linguistic characteristics of the input text. These are, themselves, generally related to TokenCount_T and to each other (e.g. POS incidences, lexical evenness and disparity, lexical difficulty, see below, section 12.1.2). The reason for these high correlations probably lies, incidentally, in the way the assessment was created: as is visible in fig. 12.2, almost all items which are not MCQs are connected to short input texts which were rated as non-continuous. In addition, the highest “continuousness” rating was almost exclusively given to the longest texts.

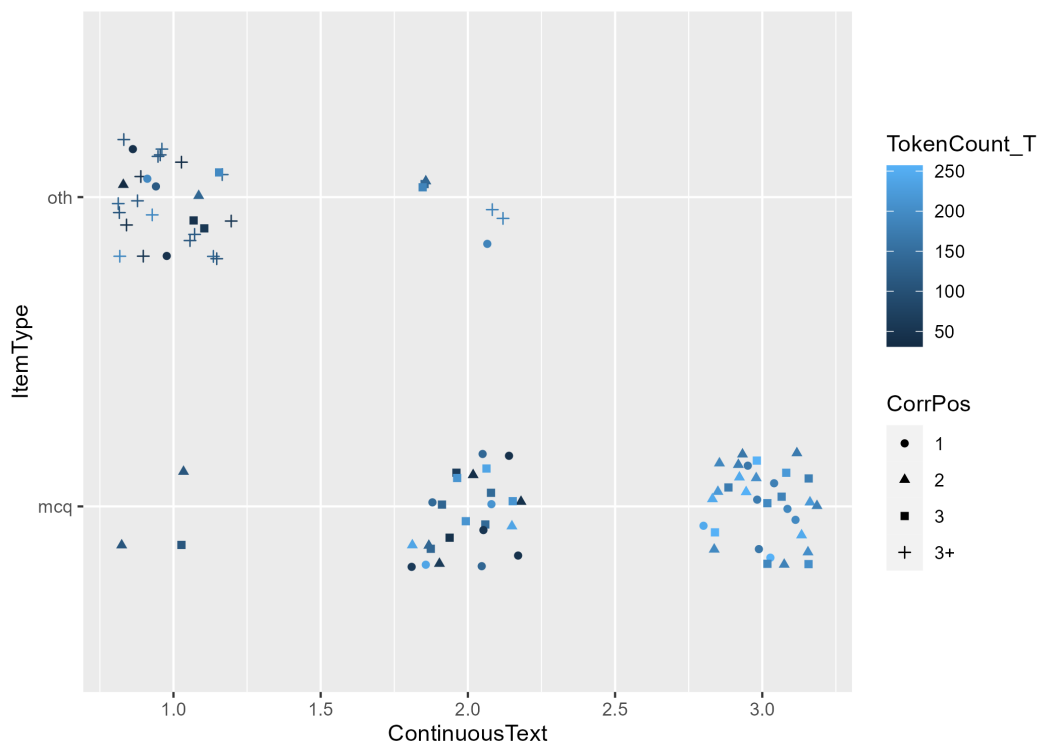


Figure 12.2: Scatterplot of variables ContinuousText, ItemType and TokenCount of input texts, plot uses data from FRde version, but all language versions are similar.

The figure also suggests a possible reason for the correlation of ItemType and CorrPos, the position of the correct answer: items which can have more than three answer options were categorised as *other* item types in this sample. These, obviously, are the only item types that can have an answer option *beyond* 3 (CorrPos: 3+). This connection is probably the reason for this correlation. For the items coded as *mcq*, the CorrPos variable seems to be reasonably well distributed and independent of both text length and text continuousness.

Though not depicted here²¹⁶, these patterns also explain the moderate-to-low correlation of ItemType and TextFunction: most of the tasks which use non-MC item types used expository texts, while the other three text functions (narrative, communicative and argumentative) were almost exclusively used with MC items. Incidentally, most of the short texts were expository (and some argumentative), which may also explain the correlations between TextFunction and several POS variables.

The remaining categorical variables RespProcess, BackgroundKn, NI_Dist and Fals are less conspicuous in terms of their interactions with other variables. Most notable is an interaction between RespProcess and NI_Dist which is easily explained: the two “lower” response processes (text-based answers and reword/restructure) are linked to NIs which are *not* distributed across

²¹⁵Here and in the following, the correlations given in the text are an approximate value of the correlations found in all four language versions. Where the correlation in a language version deviates visibly, this is indicated.

²¹⁶See the R-script for a graphical analysis of this relationship.

the text whereas the more complex bridging and inferencing response processes can also rely on NIs which are scattered over the input text. `NI_Dist` also correlates somewhat with the length of the NI: naturally, if an NI is coded as being scattered across the input text, chances are that it is longer. Other correlations of those categorical variables are negligible.

12.1.2 Linguistic characteristics of the input text

The correlation matrix reveals complex relationships among the linguistic variables related to the input texts. While some variables are correlated, these associations are not consistent within categories or across the four language versions.

The variable `TokenCount_T` correlates moderately (between $\rho \pm .30$ and $\rho \pm .40$) with all POS variables except `POS_Ad` even though all of these variables were normalised to indicate the proportion of nouns, verbs, pronouns or names per 100 words. A possible explanation for these correlations may be related to the continuousness of the input texts mentioned above: shorter texts in this sample were usually rated as non-continuous *and* they have a larger proportion of nouns (thus fewer verbs and pronouns) (see fig. 12.3). Hence, the correlation between text length and the POS variables may be due to a text genre effect, expressed by the continuousness of the text. After all, research has shown that different text genres are associated with different linguistic characteristics (see e.g. Biber 1995, pp. 113 sqq. on the linguistic characteristics of text genres).

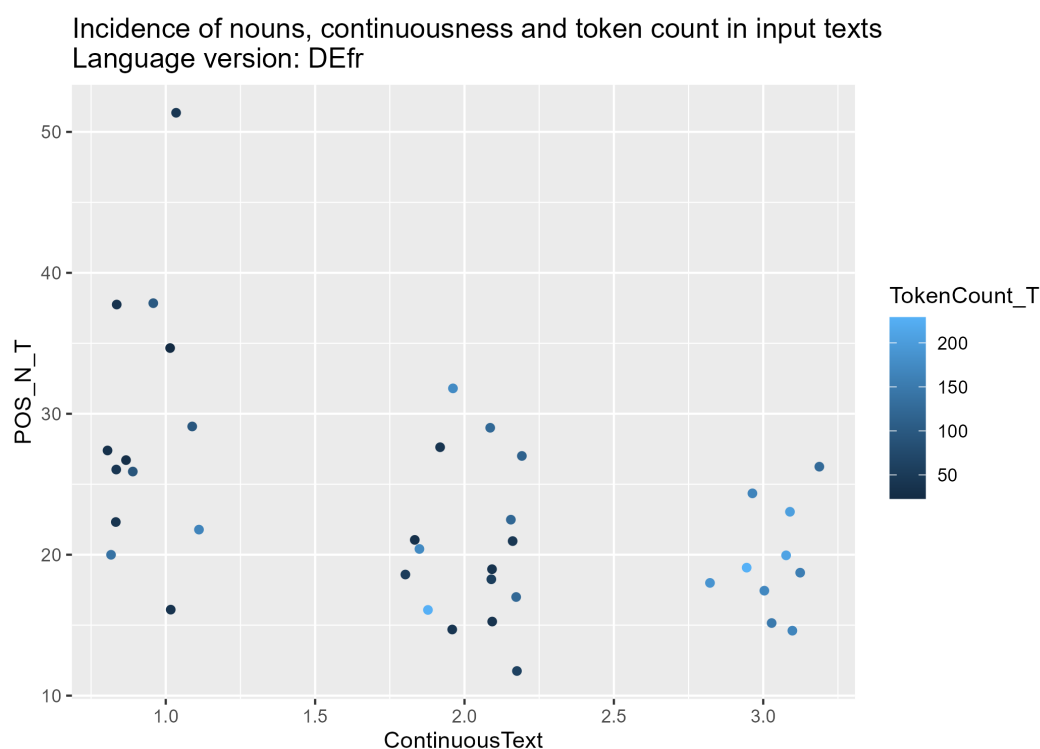


Figure 12.3: Scatterplot of the variables `ContinuousText`, `POS_N` and `TokenCount_T`. Short texts with more nouns are non-continuous in this sample. The plot uses data from the DEfr version, but all language versions are similar.

Aside from text length, the various normalised POS indices are also strongly correlated with each other in all language versions (see fig. B.1, page 276): the proportion of nouns with all other POSs correlates negatively, while the proportions of verbs and pronouns in particular correlate positively with each other. Neither the incidence of names nor of negations are related as strongly to the incidence of the other POSs, though negations occur more often when there are more verbs

and pronouns, while names occur more often when there are less adjectives. Nevertheless, both clearly also depend on other aspects of the input texts, most probably their content or topic²¹⁷. All of these observations hold across the three foreign languages: the correlations between the various POSs are all in the same ranges, and all POS variables correlate strongly across the language versions (e.g. POS_N_T in English and German: $r = .95$, $p < .001$, or POS_Ad_T in English and French: $r = .86$, $p < .001$).

12.1.2.1 Lexical diversity

Lexical diversity is measured using several different, partly experimental variables, following Jarvis (2013a,b, 2017) and Vanhove and colleagues (2019). Among them, the most obvious finding is that LexDiv_Eve²¹⁸ is always quite strongly related to text length ($\rho = .67$ in French, $\rho = .70$ in German and $\rho = .87$ in English), which puts into question whether it measures much else besides the number of words in a text. Apart from that, the lexical diversity variables in this data set behave somewhat differently depending on the language version: in English, all five variables are correlated at least moderately with each other and with most POS variables. In the German and French input texts, this is only true for lexical variety, lexical evenness and lexical dispersion, while lexical disparity and lexical density are most strongly related to the incidence of names, but not to the other lexical diversity or POS variables. In the case of lexical density, this finding is probably due to three extreme values, especially text [2], in which names are frequent, but which has the lowest lexical density of all texts. Unfortunately, an in-depth analysis of these patterns cannot be part of this thesis, though it would be a worthwhile undertaking.

12.1.2.2 Lexical sophistication

Perhaps the most interesting observation in this correlation analysis is with respect to the three lexical sophistication measures: LexSoph_tokens, LexSoph_content and LexSoph_low_freq (cf. section 9.3.1.3 for more details on these variables). The tokens variable related to the input texts is strongly negatively correlated with the proportion of nouns in the same texts ($\rho \approx -.79$). The same is true, though to a lower extent, for the other two lexical sophistication variables²¹⁹. These correlations indicate that if there are relatively fewer nouns in an input text, it will have a higher overall lexical sophistication value, i.e. the tokens or content words are *more* frequent, and lexical sophistication is actually *lower*. A similar, but much weaker effect is observed with respect to the same variables in the NIs. In addition, the variable LexSoph_low_freq also correlates with sentence length ($\rho \approx -.58$ in all language versions). These observations have consequences for the observed effect of lexical sophistication on item difficulty, which are discussed in section 13.1.1.

12.1.2.3 Lexical difficulty

The four lexical difficulty variables capture the proportion of A1 and A2 content words in the input texts, based on the Decaleon and CEFRlex databases. Their correlations with other linguistic variables are consistently inverted: when the A1 variables correlate positively with another variable, the corresponding A2 variables correlate negatively, and vice versa. This is unsurprising, as the A1 and A2 variables are themselves negatively correlated with each other, though to varying

²¹⁷Chats, e-mails and internet reviews consistently have a smaller incidence of nouns (less than 20 nouns per 100 words), and, inversely, a higher incidence of verbs (15-25) and pronouns (around 20). Adverbs and adjectives, too, tend to be more frequent in texts which have less nouns, though there are exceptions. Texts which have an incidence of nouns above 25, on the other hand, tend to have considerably smaller proportions of verbs (15 and less), adjectives/adverbs (less than 10), and pronouns (around 10 and less). These texts are the less continuous ones in the assessment, e.g. a registration form, short news or a webpage.

²¹⁸For each lemma, the number of tokens in the text is counted. Then, the standard deviation of these counts is calculated.

²¹⁹LexSoph_content_T: $\rho = -.40$ for the French texts, $\rho = -.62$ for the German texts and $\rho = -.70$ for the English texts. LexSoph_low_freq_T: $\rho \approx -.60$ in French and English, and $\rho = -.43$ in German.

degrees (ranging from $\rho = -0.72$ for the CEFRlex variables in the French texts to $\rho = -0.23$ for the English CEFRlex variables). These patterns may reflect either a systematic structure in the lexical difficulty of the input texts or a flaw in how it has been operationalised. A graphical analysis (see section B.1.4, page 280) examining relationships among the A1 and A2 variables from each database, the proportion of nouns in the texts, and the variable **ContinuousText** (as a proxy for text types) does little to clear up the picture. It does suggest, however, that the few texts with a high proportion of nouns tend to contain fewer A1 words and somewhat more A2 words, similar to what was described above for the lexical sophistication measures. The variable **ContinuousText**, by contrast, spans the full lexical difficulty spectrum, with only a few exceptions.

An interesting observation is the sometimes quite high correlation of lexical difficulty and lexical sophistication²²⁰. Graphical output of selected correlations between those two groups of variables (see the R script **4_AnalyticalStatistics_Multicollinearity**) suggests that some combinations in some language versions do indeed show a connection in that texts with a generally high lexical sophistication value (i.e. more highly frequent words) have higher proportions of A1 words and lower proportions of A2 words, as would be expected. The connection is particularly visible in German and French, and more visible with the CEFRlex variables and the **LexDiff_Decaleon_A2** variable. It suggests that the CEFRlex reference lists in particular follow similar patterns as the SUBTLEX corpora used for the operationalisation of the lexical sophistication measure (cf. for a discussion section 13.2.3).

12.1.2.4 Cross-linguistic variables

The two variables **CL_Cogn_T** and **CL_Lev_T** measure whether there are content words in the foreign language input texts that are similar to words in the language of schooling. It is hypothesised that such words make texts easier to understand for speakers of the language of schooling irrespective of their receptive *foreign* language competence.

Both variables correlate highly with each other in the DEfr and FRde language version (DEfr: $\rho = .94$; FRde: $\rho = 0.88$), but less so in the ENde language version ($\rho = .63$) and only very slightly in the FRit version ($\rho = .18$). Interestingly, in the FRde and DEfr versions, the two variables are also moderately correlated with the POS variables, **TokenCount_T** and the lexical sophistication variables, which is much less the case in the ENde and even less in the FRit version. A plot of the relationship between the two CL variables and the proportion of verbs in the input texts (fig. 12.4) contains a possible explanation: if the human rating variable is considered valid in that it measures the proportion of transparent cognates among the content words of a text, then in the French-and-German language versions, the Levenshtein-Damerau variable also seems to be able to reflect this construct. Higher and lower values of the two variables correspond in those two language versions. In the ENde and the FRit version, the closer relationship between the two languages makes for generally higher Levenshtein-Damerau values. These correspond visibly less with the **CL_Cogn_T** variable, particularly in the FRit language version. This indicates that the human rating may be better suited to identify the number of transparent cognates in more related language pairs.

Figure 12.4 also shows that, in the French-and-German language versions, texts with a higher proportion of verbs tend to contain fewer cognates. One possible explanation is that if a narrower range of verbs is used, compared to nouns, the likelihood of cognates is reduced overall. Moreover, highly frequent verbs may be less likely to be cognates between these two languages than between English and German or French and Italian. This could also account for the marked negative correlation of **CL_Cogn_T** with **ContinuousText** in these two language versions (DEfr: $\rho \approx -.56$, FRde: $\rho \approx -.62$), as verb use in this assessment appears to be related to text genre.

²²⁰E.g. **LexDiff_CEFRlex_A1_prop_T** and **LexSoph_tokens_T**: $\rho = .68$ (German texts), $\rho = .61$ (French texts); **LexDiff_Deca_A2_prop_T** and **LexSoph_low_freq_T**: $\rho = -.64$ (French texts); **LexDiff_Deca_A2_prop_T** and **LexSoph_content_T**: $\rho = -.56$ (English texts).

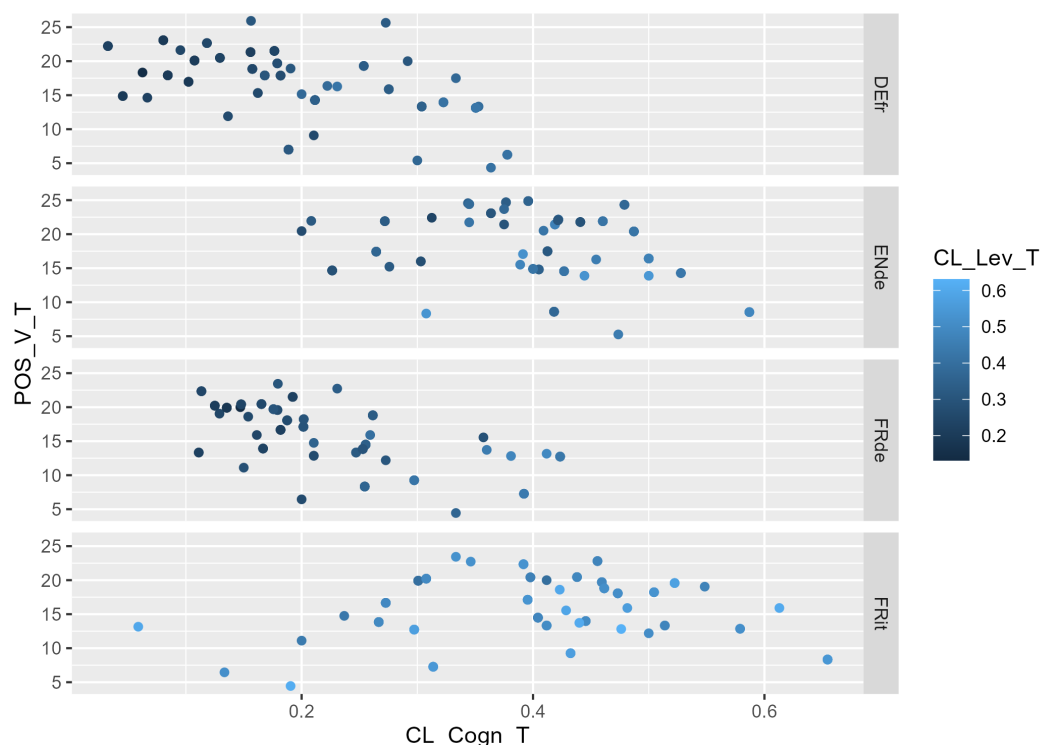


Figure 12.4: Scatterplots of cross-linguistic variables and the proportion of verbs in the input texts, separated by language version.

12.1.3 Linguistic characteristics of the NIs

The correlations between the various linguistic variables related to the NIs usually mirror the observations made for the input texts, though often to a lesser degree. In all language versions, the highest correlations within the NI variables are between text length and average sentence length, which is probably due to the shortness of most of the NIs. The two cross-linguistic variables also correlate highly with each other in all but the FRit language version. The POS variables of the input texts all correlate positively with their respective counterparts of the NIs (e.g. POS_N_T and POS_N_NI: $r = .69$ in English; POS_NAM_T and POS_NAM_NI: $r = .41$ in French), which is hardly surprising given that the NIs are subsets of the input texts. The one remaining lexical sophistication variable for the NI, LexSoph_tokens_NI also correlates quite considerably with the same variable for the input texts (DE: $\rho = .46$, EN: $\rho = .66$, FR: $\rho = .64$) and the input text's POS variables, which perhaps suggests a certain stability of the lexical sophistication within the input texts.

12.1.4 Linguistic characteristics of the items

One of the most striking observations among the linguistic characteristics of the *items* is the clear correlation between the presence of names in item stems or questions and various features of the input texts and NIs (see table 12.1 for illustrative examples). This pattern does not appear to be incidental: when a question contains a name, the corresponding input text typically features people and everyday events. As a result, such texts tend to include more verbs and fewer nouns, which in turn leads to higher lexical sophistication values (i.e. the use of more frequent words). While some extreme cases may influence the correlation coefficients (e.g., very short stems like “Alain et Marko...” or noun-heavy texts that do not elicit person-related questions), there is nonetheless a clear relationship between the content of a question and the type of text it accompanies.

Variable [LgV]	TokenCount	POS_N	LexDiv_Var	LexSoph_tokens	CL_Cogn
POS_NAM_Q [DEfr]	.02	-.50	-.37	.25	-.03
POS_NAM_Q [ENde]	.13	-.64	-.35	.55	-.37
POS_NAM_Q [FRde]	.13	-.64	-.15	.74	-.13
POS_NAM_Q [FRit]	.18	-.53	-.26	.61	.10

Table 12.1: Correlations of POS_NAM_Q with linguistic characteristics of input texts. Note that all variables in the columns refer to the input texts in the language version which is indicated in the respective rows.

Otherwise, the linguistic characteristics of the questions do not show any particular patterns except for the relationships of the POS variables among each other and correlations between the two cross-linguistic variables which are similar to those described above (CL_Cogn and CL_Lev; DEfr: $\rho = .76$, FRde: $\rho = .798$, ENde: $\rho = .57$, FRit: $\rho = .33$).

In contrast, the linguistic characteristics of the correct and the mean linguistic characteristic of the incorrect answer options²²¹ correlate with each other, sometimes quite strongly: the TokenCount of the two assessment parts correlates strongly (de: $\rho = .95$, fr: $\rho = .93$, it: $\rho = .92$), and their respective POS variables also have medium to strong correlations between $\rho \approx .60$ and $\rho \approx .90$, with the Italian items generally in the lower and the German items in the higher ranges of this continuum. Given that it is a sign of good item writing that answer options are similar to each other within the same item, this finding is positive. The POS variables also correlate quite strongly with the TokenCount of the answer options: if an answer option is short, there are proportionally more nouns (sometimes not much more than that), but if the answer options are longer, the other POSs become proportionally more frequent. This is also true for the variable LingComp1_Neg_I, which operationalises the incidence of negations in the item as a whole. The proportion of names in the answer options, in contrast, follows other patterns, similar to those in questions described above.

Finally, the cross-linguistic characteristics of the correct and incorrect answer options respectively correlate with each other, as has been observed for the other assessment parts. Across the answer options, however, there are no major correlations between the cross-linguistic variables: it seems that a higher proportion of cognates in the correct answer is not automatically linked to a higher mean proportion in the incorrect answers and vice-versa: the highest correlation between those variables is $\rho = .38$ between the variables CL_Cogn_correct and CL_Cogn_incorrect in the DEfr language version. Other combinations of the same variables or the CL_Lev variables are closer to 0 or even slightly negative, though, indicating a random occurrence.

12.1.5 Overlap variables

The semantic overlap variable was operationalised as the proportion of translation equivalents among the content words of the item parts and their respective NI and input text. It does not show high correlations with any other variable except their own _NI or _T counterparts: both for the question/stem and the correct answer, the overlap variables related to the NI and the input text correlate strongly with each other (for the questions, $\rho \approx .75$, for the correct answer $\rho \approx .90$). The correlations of these two variables related to the incorrect answers are lower, but still moderate (between $\rho = .40$ in FRde and $\rho = .52$ in FRit). The latter also have a tendency to correlate somewhat with the POS variables of the incorrect answer option as well as the overlap variables of the correct answer, but these correlations are in the low ranges, and are not observed in all language versions. All of this indicates that the semantic overlap variables may

²²¹As described in section 10.2.2.1, the linguistic characteristics of the incorrect answer options were reduced to their mean in an effort to reduce data complexity.

indeed capture an aspect of the items that is conceptually different from the other linguistic characteristics of the assessment parts.

12.2 Data reduction

Data reduction was attempted both on language-specific data sets and on the entire multilingual data set using factor and cluster analysis. The various factors and clusters extracted from this step were entered into the regression analyses but did not perform better or add any significant information to the models beyond the individual variables. Nevertheless, the results of these analyses are presented in the following since they add to the overall impression of how the various linguistic variables relate to each other.

12.2.1 Factor analyses of linguistic characteristics of input texts

All factor analyses started out with all linguistic variables as candidates: `TokenCount_T` and `SentenceLength_T`, all POS variables including the incidence of negations, all lexical diversity, sophistication and difficulty indices as well as the two cross-linguistic variables `CL_Cogn_T` and `CL_Lev_T`. In some cases, variables were eliminated even before the actual factor analysis based on their low KMO values. Other variables were eliminated during model selection to achieve better fit. In the following, the final model for the ENde language version is described in detail including its factor loadings and fit indices. The final models of the other three language versions are summarised more succinctly and in section B.1.8 (page 286). Detailed comments can also be found in the R script `4_AnalyticalStatistics_Multicollinearity`.

The final factor model of the linguistic characteristics of the input texts in the ENde language version was estimated by Exploratory Factor Analysis (EFA) with quartimin rotation. The overall KMO Measure of Sampling Adequacy was 0.77, indicating a moderate level of factorability. Individual KMO variables ranged between .40 (`LexDiv_Var_T`) and .90 (`POS_Pro_T`). Bartlett's test of sphericity confirmed that the data set was suitable for factor analysis ($\chi^2(45) = 224.74$, $p < .001$, indicating that the data is significantly different from an identity matrix²²²).

Three factors were extracted based on the analysis, accounting for a cumulative variance of 64%. The first factor, defined by POS and lexical sophistication variables explained 40% of the variance, the second factor (lexical variety and a negative loading of lexical sophistication) explained 13%, and the third factor (characterised by text length and the incidence of names) explained 12% (see table 12.2²²³).

The correlation matrix reveals moderate relationships between the factors: the “syntactic and lexical sophistication” factor (`POSLexSoph`) and the “lexical diversity” factor (`LexVar`) are negatively correlated ($r = -0.42$), while the third factor (`TextType`) is only weakly correlated with the other two ($r = 0.16$ with factor 1; $r = -0.07$ with factor 2).

Overall, the model demonstrated a good fit with a Tucker-Lewis Index (TLI) of 0.958 and a Root Mean Squared Error of Approximation (RMSEA) of 0.057 (90% CI: [0, 0.164]). The Likelihood Chi-Square was 20.79 ($df = 18$, $p = 0.29$), suggesting that the three-factor solution adequately represents the data. Factor score adequacy was assessed by the correlation of scores with factors (`POSLexSoph` = 0.97, `LexVar` = 0.96, `TextType` = 0.87), multiple R-squared values (`POSLexSoph` = 0.94, `LexVar` = 0.93, `TextType` = 0.75), and minimum correlations of possible factor scores (`POSLexSoph` = 0.88, `LexVar` = 0.86, `TextType` = 0.51), indicating strong factor score reliability, particularly for the first two factors.

Factor analyses were conducted for the other three language versions as well, with similar but not identical results (see for factor loadings and model information section B.1.8, page 286).

²²² An identity matrix is a $n \times n$ square matrix with ones in the diagonal and zeros elsewhere. In the context of factor analysis, such a correlation matrix would mean that no two variables are correlated with each other and that, thus, a factor analysis is not possible.

²²³ In the actual data set, the scores from these factors were named `LingFA_FactorLabel_T`.

Variable	POSlexSoph	LexVar	TextType
TokenCount_T	0.372		0.483
POS_N_T	-0.654		
POS_V_T	0.886		
POS_Pro_T	0.860		
POS_NAM_T			0.819
LingCompl_Neg_T	0.535		
LexDiv_Var_T		0.813	
LexSoph_tokens_T	0.593	-0.579	
LexSoph_content_T	0.866		
LexSoph_low_freq_T	0.519		0.447

Table 12.2: Factor loadings from exploratory factor analysis on linguistic characteristics of ENde input texts. Factor loadings below $|\cdot 30|$ are not depicted. Variables which are not listed in the table were eliminated at some point during model selection.

For the FRde data, three factors were extracted: a POS factor including nouns, verbs and pronouns, a lexical sophistication factor which includes all three lexical sophistication measures, with **LexSoph_tokens** slightly dominating the other two, and a crosslinguistic factor, onto which both CL variables load to similar extents. Overall, the fit of this model is adequate, but not as good as for the English data. It explains 68% of the variance, but with a TLI of only .866 and an RMSEA index of 0.122. In return, factor score adequacy is quite good. For the FRit data, five factors were needed to reach good fit. The factor with the highest explained variance is similar to the **TextType** factor in English. The common POS and lexical sophistication factor from the English model is separated into three separate factors in the FRit data, and there is a fifth factor of lexical difficulty that is not modelled in the English data because these variables had been eliminated. The fit values of the FRit factor model are excellent and factor score adequacy indicates strong factor reliability. The variance explained by this model is 76%. Finally, the DEfr factor model is perhaps the most complete, with most variables adding to the model. Four factors were extracted: a POS and lexical sophistication factor similar to the one in English, a CL factor with information from POS variables, a text length factor and a lexical diversity factor. This model fits the data quite well, and explains 71% of the variance. However, several variables in this model crossload on various factors, which makes it harder to interpret (see table B.4, page 286).

12.2.2 Cluster analyses of POS variables of NIs and item parts

Given the systematic and cross-linguistic patterns of the item parts' linguistic characteristics, an attempt was made to condense their POS variables into one variable using k-means cluster analyses for all 376 items in the same analysis: questions were clustered into three groups, characterised by a larger proportion of names, verbs and pronouns, and nouns respectively. The POS variables of the correct and incorrect answer options were clustered into two groups, those with more nouns and low proportions of other POSs and those with less nouns, but more of the other POSs. For incorrect answers, there is a third, small cluster which is characterised by a large proportion of adjectives and adverbs. The **TokenCount** and **LingCompl_Neg** variables were not added to the cluster analyses to retain their informational value. The same is true for the two cross-linguistic variables, which were kept in order to account for their differences in operationalising the linguistic closeness of the four languages. Ultimately, given the results of preliminary correlational analyses conducted in preparation of the analyses described in the following, the clusters were not retained.

12.3 Prediction of item difficulty

The main focus of this thesis is the prediction of item difficulty by the variables extracted from the items and input texts. The results from the various approaches to this question adopted in this thesis will be presented in the following. First, “simple” bivariate correlations are discussed, followed by language-specific linear regression models. From section 12.3.4 onwards, the entire data set is considered: first, a mixed-effects linear regression model using the language version and the input texts as random effects is discussed. Then, the results from the random forests analysis and the Bayesian regression modelling add more depth to the analysis.

12.3.1 Correlations with item difficulty

In the literature, correlations between individual item characteristics and item difficulty are one way of investigating the sources of item difficulty. In the following, the results from separate correlation analyses in all four language versions are presented alongside each other.

Table 12.3 contains all item and text variables²²⁴ which correlate significantly with item difficulty in at least one language version of the ÜGK foreign language reading assessment²²⁵. The variables are listed in alphabetical order so that the same types of variables, even if they refer to different parts of the assessment, can be examined together. The following observations can be made:

- Out of 72 variables, 37 correlate significantly ($p < .05$) with item difficulty in at least one language version.
- If the threshold of $p < .1$ is deemed acceptable, 12 variables correlate significantly with item difficulty in all four language versions²²⁶.
- The strongest correlation over all variables is $\rho = .40$ (the correlation of `LexDiv_Eve_T` and item difficulty in the ENde language version).
- There are 24 out of 81 significant correlations across all language versions which are stronger than $\rho = |.30|$.
- As long as they are significant, correlations of any variable with item difficulty always point in the same direction in the language versions. In fact, they are usually in the same range as well.
- Nine of the 37 variables in this selection operationalise the linguistic characteristics of the item, i.e. its question, correct or incorrect answer option. None of them are significantly correlated in all language versions.
- Four of the nine categorical variables in the data set correlate significantly with item difficulty in some language versions. These were meant to operationalise aspects of the *content* of the input texts and items rather than their linguistic characteristics (though see section 12.1.1).

²²⁴As mentioned above (section 12.2), the various factors and clusters do not add any helpful information to this analysis beyond what can be interpreted from the individual variables. Therefore, their correlations with item difficulty are not reported here, though they were considered at some point in the analytic process.

²²⁵Given the decidedly not normal distribution of most variables, the Spearman’s rank correlation coefficient ρ was estimated in all analyses.

²²⁶`ContinuousText`, `ItemType`, `LexDiff_Deca_A2_prop_T`, `LexDiv_Eve_T`, `LexSoph_tokens_NI`, `LexSoph_tokens_T`, `POS_N_NI`, `POS_N_T`, `POS_Pro_NI`, `POS_Pro_T`, `POS_V_T` and `TokenCount_T`. Note that there are no item-specific variables in this list.

Variable	ρ (DEfr)	ρ (ENde)	ρ (FRde)	ρ (FRit)
BackgroundKn	-0.26*	-0.19.	-0.15	-0.11
CL-Cogn-NI	-0.27**	-0.03	-0.23*	0.03
CL-Cogn-Q	-0.05	-0.04	-0.27**	0
CL-Cogn-T	-0.21*	-0.22*	-0.2.	-0.01
CL-Lev-NI	-0.24*	-0.21.	-0.21*	0.04
CL-Lev-Q	-0.11	-0.18.	-0.2.	-0.26*
CL-Lev-T	-0.21*	-0.14	-0.23*	-0.09
ContinuousText	0.37***	0.33**	0.34***	0.27**
ItemType	-0.17.	-0.28**	-0.23*	-0.23*
LexDiff-CEFRlex-A1-prop-T	0.24*	0.01	0.24*	0.15
LexDiff-Deca-A1-prop-T	0.25*	0.11	0.13	0.12
LexDiff-Deca-A2-prop-T	-0.34***	-0.26*	-0.34***	-0.3**
LexDisparity-form-T	0.24*	0.29**	0	0
LexDispersion-T	0.09	0.13	0.17.	0.22*
LexDiv-Eve-T	0.19.	0.4***	0.23*	0.32**
LexDiv-Var-T	-0.24*	-0.31**	-0.17	-0.19.
LexSoph-content-T	0.22*	0.2*	0.21*	0.05
LexSoph-low-freq-T	0.12	0.24*	0.19.	0.28**
LexSoph-tokens-NI	0.21*	0.31**	0.35***	0.19.
LexSoph-tokens-T	0.31**	0.37***	0.39***	0.31**
LingCompl-Neg-I	0.12	0.13	0.14	0.22*
LingCompl-Neg-T	0.13	0.1	0.21*	0.07
Overl-sem-T-correct	-0.26*	-0.31*	-0.09	-0.1
POS-Ad-Q	0.09	0.28**	0.12	0.21*
POS-N-NI	-0.2.	-0.26*	-0.26*	-0.24*
POS-N-Q	-0.15	-0.21*	-0.24*	-0.34***
POS-N-T	-0.31**	-0.28**	-0.38***	-0.39***
POS-NAM-Q	0.04	0.19.	0.25*	0.25*
POS-NAM-T	0.2.	0.13	0.18.	0.27**
POS-Pro-NI	0.27**	0.28**	0.24*	0.18.
POS-Pro-Q	0.23*	0.07	0.12	-0.11
POS-Pro-T	0.2.	0.25*	0.23*	0.17.
POS-Pro-correct	0.04	0.09	0.17	-0.28*
POS-V-T	0.33**	0.3**	0.3**	0.33**
RespProcess	0.08	0.08	0.25*	0.07
SentenceLength-NI	0.27*	0.18.	0.15	0.05
TokenCount-T	0.39***	0.35***	0.35***	0.3**

Table 12.3: Correlations of item and text variables with item difficulty (Spearman's correlation coefficient). Only variables which correlate significantly in at least one language version are listed. Indicators of significance: . < 0.1, * < 0.05, ** < 0.01, *** < 0.001.

Since a correlation coefficient by itself does not indicate how well the data fits the assumption of a monotonic relationship between two variables, all significant correlations were also graphically examined: for each language version, item difficulty and the respective variable were plotted in a scatterplot. These generally show that even when moderate correlations were observed in all language versions, the relationship between the two variables is visibly less prominent than one might assume. For instance, the relationships between item difficulty and lexical sophistication, apparently significant and relatively high in all four language versions, are depicted in fig. 12.5²²⁷. The upwards direction of the regression line, which is reflected in the correlation coefficient, is quite clearly determined by a few data points on the outer edges: in all language versions, there are only a few very easy items (the dots on the left-hand side of the plots) and these are generally related to texts which have a lower lexical sophistication value²²⁸. The inverse is true for the few dots representing the most difficult items in all language versions. These extreme items seem to determine most of the correlations, whereas the majority of items, those of medium difficulty, figure on the entire spectrum of lexical sophistication. Clearly, as has been mentioned in various parts of this thesis, simple one-on-one relationships cannot be assumed in this context.

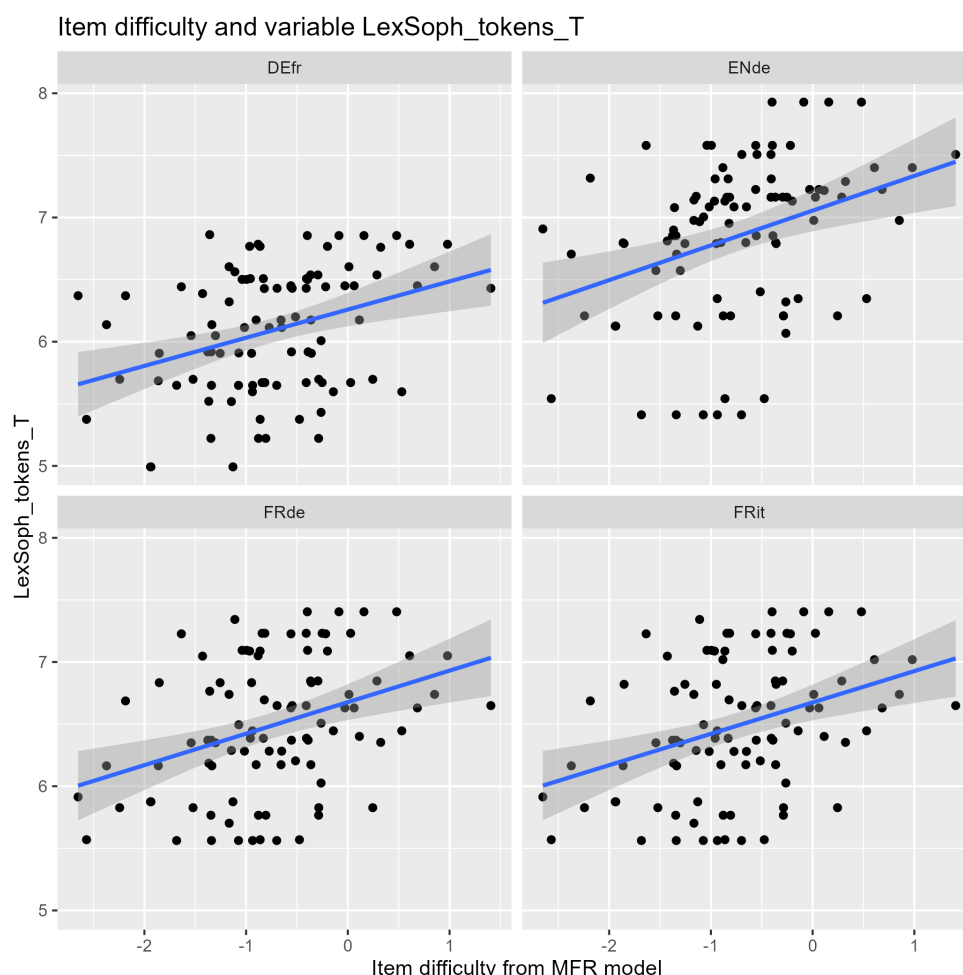


Figure 12.5: Scatterplots of item difficulty and mean (log-transformed) lexical sophistication of tokens in input texts, separated by language version.

²²⁷This is a good example of scatterplots where the correlation is actually visible, even if only just. In several other scatterplots, it is clear that outliers or the way the variable was operationalised are responsible for the correlation. See for all graphs the files output by the R script `4_AnalyticalStatistics`.

²²⁸Note that a lower lexical sophistication value actually means that the tokens in the input texts were less frequent, i.e. more sophisticated. Input texts with less frequent tokens related to easier items seem to be a counter-intuitive finding. See section 13.1.1 for a discussion of this.

Nevertheless, in summary, the correlations of all variables with item difficulty point to a moderate but stable effect of the structural and lexical characteristics of the input texts and NIs. Some structural aspects of the questions and answer options also seem to play a role in item difficulty, though to a lesser extent and never in all language versions. A similar observation can be made for the text-by-item variables **BackgroundKn** and **RespProcess**. In the language versions including German and French, the crosslinguistic variables, an operationalisation of the proportion of cognates in the input texts and NIs also show a low effect on item difficulty. All of these correlations, even when significant, are in the lower to moderate ranges (around $\rho \approx |.30|$).

12.3.2 Correlations with mean task difficulty

In her analysis of a German L1 listening assessment, Neumann (2012, pp. 169 sqq.) correlated characteristics of the input text (which she called *Stimulus*) with item difficulties and with mean task difficulties, i.e. the means of the difficulties of all items related to one input text. She found higher correlations between the input text characteristics and the mean task difficulties as compared to their correlations with individual item characteristics (ibid., p. 170). This analysis was repeated with the ÜGK data, but the effect was not generally observed. Instead, given the lower number of input texts, the p-values of the correlations were often higher by several degrees. Where a correlation remained significant, it was quite comparable to the correlation with the individual item difficulties, often within $\pm .1$. Overall, only the following seven variables correlate significantly with mean task difficulty in at least two languages, though the same caveat concerning the linearity of the relationships mentioned above remains: **ContinuousText**, **TokenCount_T**, **POS_N_T** and **POS_V_T**, **LexDiv_Eve_T**, **LexSoph_tokens_T** and **LexDiff_Deca_A2_prop_T**. Interestingly, this list is quite similar to the one in footnote 226, which lists all variables that correlate significantly with item difficulty in all four language versions.

12.3.3 Regression modelling

Regression modelling was conducted for each language version separately in a stepwise forward procedure where predictors were added to the model in thematic groups as described in section 10.2.2.3. The initial model was always a mixed-effects linear model with **TextID** as a random factor, item difficulty as the dependent variable (**xsi.mfr**) and **TokenCount_T** as the first predictor. The final models, however, are fixed-effects models since the inclusion of **TextID** proved to be unnecessary. In the following, the final models based on the ENde and the DEfr data sets are presented in more detail. The regression models for the French tests are summarised, since they do not differ much from the findings in the other language versions.

12.3.3.1 Item difficulty prediction in the English test

The final model based on the ENde data set is a linear fixed-effects model with three predictors (see table 12.4): lexical variety (**LexDiv_Var_T**), the lexical overlap between input text and correct answer option (**Overl_sem_T_correct**) and the Damerau-Levenshtein distance of the content words in the question or stem with their translation equivalents in the target language (**CL_Lev_Q**). These three predictors account for 20% of the variance in item difficulty ($R^2 = .243$, adjusted $R^2 = .199$), with each predictor contributing about a third of the information²²⁹. The model is statistically significant overall, $F(3, 52) = 5.55$, $p = .002$. The residual standard error is 0.636. Visual inspection of residual plots indicated no major deviations from assumptions of the linear model (homoscedasticity and normality). Multicollinearity in the three predictors was checked, but is unproblematic (all VIFs ≈ 1). Due to **NAs** related to the item types, the model is based on the characteristics of only 56 items, all of them with an item difficulty below 0.

²²⁹The relative importance summed to 100% is .32 for the overlap variable and $\approx .34$ for the other two variables according to the LMG metric.

Predictor	Estimate	95% CI Lower	95% CI Upper	<i>p</i>
Intercept	0.340	-0.346	1.025	.325
LexDiv_Var_T	-0.0096	-0.0176	-0.0016	.020*
Overl_sem_T_correct	-0.487	-0.934	-0.040	.033*
CL_Lev_Q	-0.760	-1.506	-0.013	.046*

Table 12.4: Linear regression model predicting ENde item difficulty. Note: * indicates $p < .05$. CI = Confidence Interval.

The model indicates that item difficulty will be *lower* if the MTLT of the input text is higher, if there is more semantic overlap between the input text and the correct answer, or if the Levenshtein distance of the question or stem goes up (i.e. the content words in the German question or stem are more similar to their translation equivalent in English). The estimates in table 12.4 also indicate that the effect of lexical variety in particular is not very strong: a difference of 10 points on the MTLT scale is linked to a difference of $-.096$ logits in item difficulty. The effects of the other two variables is stronger, though the confidence intervals of both estimates show a large uncertainty of the estimation: between items linked to an overlap of 0 (i.e. no overlap at all) and items which have an overlap of 1 (i.e. 100% of the content words in the correct answer appear in the input text), item difficulty may barely change ($-.04$) or go down by an entire logit ($-.93$). A similar statement can be made for the question's Damerau-Levenshtein distance. The uncertainty of this model is also illustrated in fig. 12.6, which shows predicted item difficulties at different values of the three variables (right-hand panel) and the actual values of the items. The two graphs are visibly quite distinct.

12.3.3.2 Item difficulty prediction in the German test

The final model based on the DEfr data is different from the ENde model described above. It is highly significant ($F(6, 40) = 7.98$, $p < 0.001$) and explains a variance of 48% ($R^2 = .545$, adjusted $R^2 = .477$), but it is only based on 47 observations due to missing data. The residual standard error is 0.59, and quality checks confirmed model fit. This model contains six predictors: text length (TokenCount_T), the proportion of human-rated cognates and the average sentence length of the NI (CL_Cogn_NI, SentenceLength_NI), the mean semantic overlap between the NI and the *incorrect* answer options (Overl_sem_NI_incorrect), as well as two POS variables related to the items (POS_Pro_Q and POS_Ad_correct)²³⁰.

Similar to the ENde regression model above, the predictive power of this model is low, despite the relatively high R^2 (see table 12.5): it predicts that with every 10 tokens in the input text, item difficulty will be .05 logits higher, i.e. a difference of 100 tokens is related to half a logit in item difficulty. This is not a lot considering that most texts in this assessment are short. A higher proportion of cognates in the NI and more overlap between the NI and the *incorrect* answer option are linked to a quite considerably lower item difficulty, though the 95% confidence intervals are quite large. A higher mean sentence length in the NI, the proportion of pronouns in the question and of adverbs and adjectives in the correct answer option are related to higher item difficulty, though the strength of the effect, again, is uncertain to a degree of 10.

12.3.3.3 Item difficulty prediction in the French test

Given that the FRit and FRde item difficulties do not correlate better with each other than with the other two language versions, two separate regression models were constructed. This is

²³⁰ According to the LMG metric, the NI's sentence length and the overlap variable contribute 9% and 13% to the model's R^2 respectively, while the other variables contribute about 20% each.

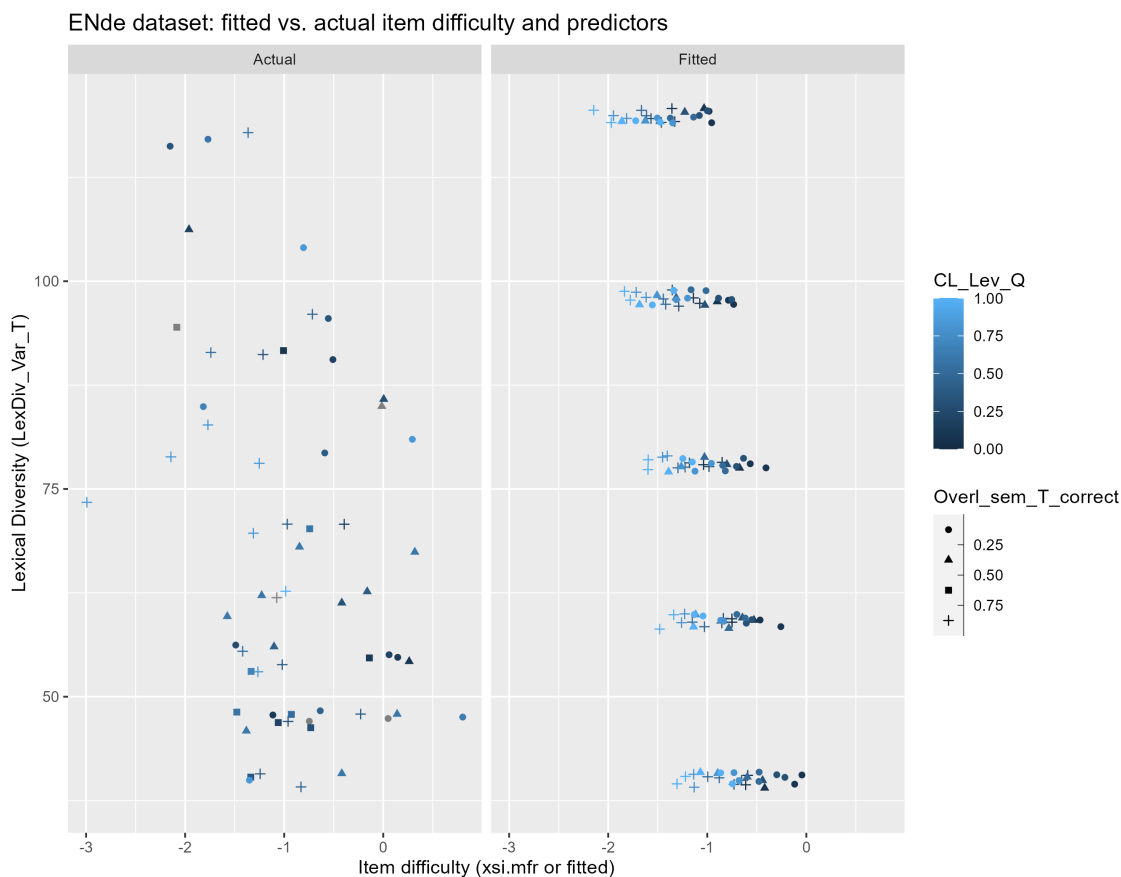


Figure 12.6: Comparison of fitted and actual effects of linear regression on ENde data set.

Predictor	Estimate	95% CI Lower	95% CI Upper	<i>p</i>
Intercept	-1.622	-2.483	-0.762	.0005***
TokenCount_T	0.0046	0.0009	0.0083	.016*
CL_Cogn_NI	-1.389	-2.58	-0.197	.023*
SentenceLength_NI	0.071	0.021	0.12	.006**
Overl_sem_NI_incorrect	-1.299	-2.182	-0.418	.005**
POS_Pro_Q	0.017	0.002	0.032	.024*
POS_Ad_correct	0.013	0.004	0.022	.008**

Table 12.5: Linear regression model predicting DEfr item difficulty. Note: * indicates $p < .05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, CI = Confidence Interval.

despite the fact that the linguistic characteristics of the input texts are identical in these two language versions. The final models for these two language versions can be found in section B.2.1 (page 287). The FRit model is similar to the models described above in that the predictors contain a variable for each assessment part. Overall, it is dominated by the input text's token count and the NI's POS structure. The FRde model turned out different in that a significant interaction between the proportion of nouns and text length of the input text was modelled initially, which led to the inclusion of only one other predictor (CL_Cogn_NI).

12.3.3.4 Conclusion

Overall, the regression models confirm the role of the input text's linguistic characteristics in predicting item difficulty: in all cases, either `TokenCount_T` or a lexical variable related to the input text and/or the NI are among the relevant variables. The models also underline the importance of the language of schooling: in each model, there is one predictor which operationalises the linguistic closeness of the two languages at play. Finally, in the ENde and the DEfr data set, there is also an overlap variable at play, which confirms earlier findings that the wording of the item in relation to the input text is linked to item difficulty (see table 8.1, page 128 for references).

The final models presented above also confirm that linear regression modelling is only partially suited for item difficulty prediction: the models leave large proportions of variance unexplained, and due to the nature of the data set, they often predict the difficulty of only a subset of items. Clearly, the mechanisms at play when test takers solve test items are much more complex than can be modelled by a few linear relationships.

12.3.4 Multilingual prediction of item difficulty

While the language-specific analyses offered somewhat underwhelming results, the central question remains: can item difficulty be predicted in a similar fashion across all language versions? The following sections explore this question in more detail. First, the applicability of the language-specific regression modelling to the other language versions is checked. Then, the results of a stepwise hierarchical mixed-effects regression modelling are presented, which included data from all language versions.

12.3.4.1 Crosslinguistic validation

The final regression models developed for each language version were used to predict the item difficulties in the other language versions. Table 12.6 presents the resulting RMSEs, which indicate the average deviation between predicted and observed item difficulties. These values are expressed on the logit scale used in the underlying MFR model, thus higher values express stronger deviation. As shown in the table, the DEfr model is overall best at predicting item difficulty in the other language versions, since it produced the lowest RMSEs across the other language versions. In contrast, the ENde model yields the least accurate predictions for item difficulties in the remaining language versions.

<i>Training data</i>		ENde	DEfr	FRit	FRde
<i>Test data</i>	ENde	<i>0.61</i>	0.69	0.87	0.89
	DEfr	1.26	<i>0.55</i>	0.99	0.79
	FRit	0.9	0.97	<i>0.64</i>	0.98
	FRde	1.08	0.71	0.89	<i>0.7</i>

Table 12.6: RMSEs from model testing across languages. Rows represent the RMSEs for the item difficulties of the four language versions (test data) predicted by the models based on the data set indicated in each column (training data). Italic numbers are RMSEs of each model with its own original data. They indicate overall model fit.

12.3.4.2 Mixed-effects modelling

In addition to testing the separate regression models on the data from the other language versions, a mixed-effects linear regression model was fitted to the entire data set with the language version

(**LgVersion**) serving as a random effect. As before, **TextID** was also included as a random effect. This time, however, it proved to be an important addition to the model, and was not eliminated. For the final model, the marginal R^2 (R^2m), which reflects the variance explained by the fixed effects alone, was 0.23. The conditional R^2 (R^2c), which accounts for both fixed and random effects, was 0.72. This suggests that while the fixed effects contribute to a moderate proportion of the variance in item difficulty, the inclusion of the language version and the input text as random effects considerably improve the predictive power of this model.

The final model contains similar fixed effects as the language-specific models described above: the text length of the input texts, the proportions of verbs in the NIs and of adverbs/adjectives in the questions or stems, the proportion of cognates in the NIs and the semantic overlap between the correct answers and the input texts and the incorrect answers and the NIs respectively have all featured in the separate models described above. Only the variable operationalising the response process at play (**RespProcess**) appears for the first time in a final model (see table 12.7).

Random Effect	Variance	Std. Dev.
TextID	0.354	0.595
LgVersion	0.064	0.252
Residual	0.242	0.492

Predictor	Estimate	95% CI	t-value	Sig.
Intercept	-0.717	[-1.27, -0.17]	-2.56	*
TokenCount_T	0.003	[-0.0003, 0.006]	1.77	.
Overl_sem_T_correct	-0.269	[-0.47, -0.07]	-2.70	**
POS_V_NI	-0.019	[-0.027, -0.012]	-4.80	***
Overl_sem_NI_incorrect	-0.745	[-1.11, -0.38]	-4.04	***
CL_Cogn_NI	-0.732	[-1.07, -0.39]	-4.25	***
RespProcessrewo	0.263	[0.01, 0.52]	2.03	*
RespProcesstext	0.637	[0.40, 0.88]	5.20	***
POS_Ad_Q	0.016	[0.008, 0.024]	3.90	***

Table 12.7: Summary of mixed-effects linear regression model on multilingual data predicting item difficulty (**xsi.mfr**). Note: . indicates $p < .1$, * indicates $p < .05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, CI = Confidence Interval.

The model suggests that item difficulty is *lower* when the input text is longer, the proportion of verbs in the NI is higher, when there are more cognates in the NI and if the semantic overlap between the input text and the correct answer or the NI and the incorrect answer options is larger. Inversely, *more* difficult items tend to have questions or stems which contain more adjectives or adverbs, and require rewording or text-based answers as opposed to inferencing or bridging²³¹. Most of these fixed effects have relatively broad confidence intervals, suggesting weak effects. In fact, the confidence interval of **TokenCount_T** is on both sides of 0, indicating that it may or may not play a role.

The random effects in this model have larger predictive power than the fixed effects. Text-level variance in particular is substantial ($SD = 0.59$), which suggests that item difficulty varies depending on the input text, beyond what is captured in the text length and the overlap variable including the input text. Interestingly, during modelling, no other text-related variable proved to be a significant predictor. The variance of the language version is smaller ($SD = 0.25$), which indicates that the variation between language versions is smaller on average than between input

²³¹See for a discussion of this counter-intuitive finding section 13.1.1.

texts, but still there. In fact, the residual standard deviation of the language version in the mixed-effects model is a bit lower than that of the language facet in the MFR model which supplied the item difficulty values²³². This suggests that a small part of the variance originally attributed to language differences in the MFR is now being explained by the item-level predictors in the mixed model, while the remaining variance ($SD = 0.25$) reflects what is not accounted for by the fixed effects and thus still varies by language version.

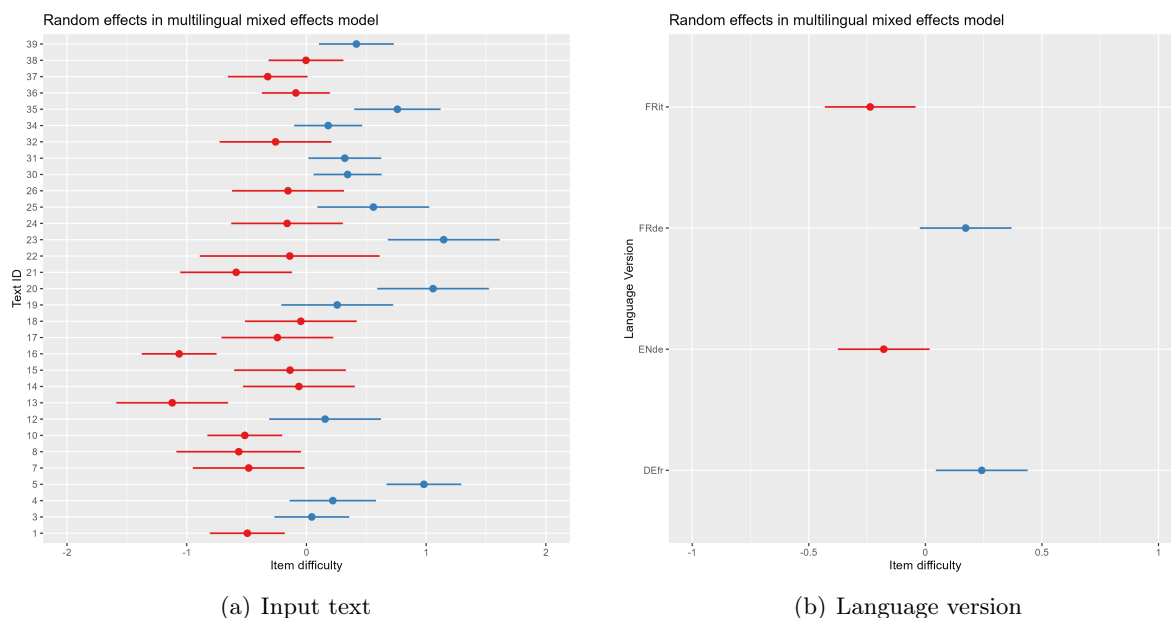


Figure 12.7: Plot of random effects of mixed-effects linear regression model based on multilingual data set

Figure 12.7 shows the group-level differences of `TextID` and language version beyond the fixed effects (by plotting the so-called *conditional modes* or *Best Linear Unbiased Predictors (BLUPs)*). The right-hand graph shows, much like the MFR model, that the `ENde` and the `FRIt` language version are linked to lower item difficulties, while the `DEfr` and the `FRde` language version feature more difficult items. The left-hand graph indicates how much the item difficulties related to each input text diverge from the mean. Though most texts lie within each other's confidence intervals, some input texts are clearly linked to very easy (texts 13 and 16) or to quite hard items (texts 5, 20 and 23)²³³.

12.3.5 Triangulation: exploratory random forest modelling

An exploratory random forest analysis was applied to the data for triangulation purposes. Two random forest regression models were trained with 500 trees and default `mtry` setting, each on the entire data set: the first model includes `LgVersion`, `ItemID` and `TextID` as possible predictor variables, allowing the outcome variable (`xsi.mfr`) to cluster within those categories. The second model only retains `LgVersion` as a predictor to allow the text, item and text-by-item variables to predict item difficulty as well.

In both cases, the random forests confirmed the general results of the regression models described above. If included, `TextID` is by far the most important predictor of item difficulty,

²³²The values of the language facet in the MFR model are: `ENde`: -0.35 , `FRIt`: -0.30 , `DEfr`: 0.36 , `FRde`: 0.29 . The standard deviation of these values is 0.325 .

²³³Note that among these five texts, three (13, 20 and 23) are short input texts which are associated with one item only.

responsible for most of the splits in a tree extracted from the forest. This model achieves an out-of-bag R^2 of 48.59% and an RMSE of .66. As such, it is comparable in predictive power to the mixed-effects regression model described above. Important predictors (according to the %IncMSE measure²³⁴) are, in addition to **LgVersion**, **ItemID** and **TextID**, the variables **RespProcess**, **POS_N_NI**, **CL_Cogn_NI** and **POS_V_incorrect**. The model also includes the input text's token count, its cognateness and lexical difficulty as well as linguistic variables related to the NI and the question or stem, but it only considers five variables for each tree estimation.

To examine the contribution of the assessment-related and linguistic variables beyond the information that is clustered in an item's link to its input text, a second random forest regression model was trained on all variables except for **TextID** and **ItemID**. This model explains considerably less variance (35.4%) and has a higher RMSE than the full model, but it contains more information on how text and item characteristics influence item difficulty. In this model, text length, lexical difficulty and the proportion of cognates in the input text, the NI and the question are the most important predictors (see fig. 12.8).

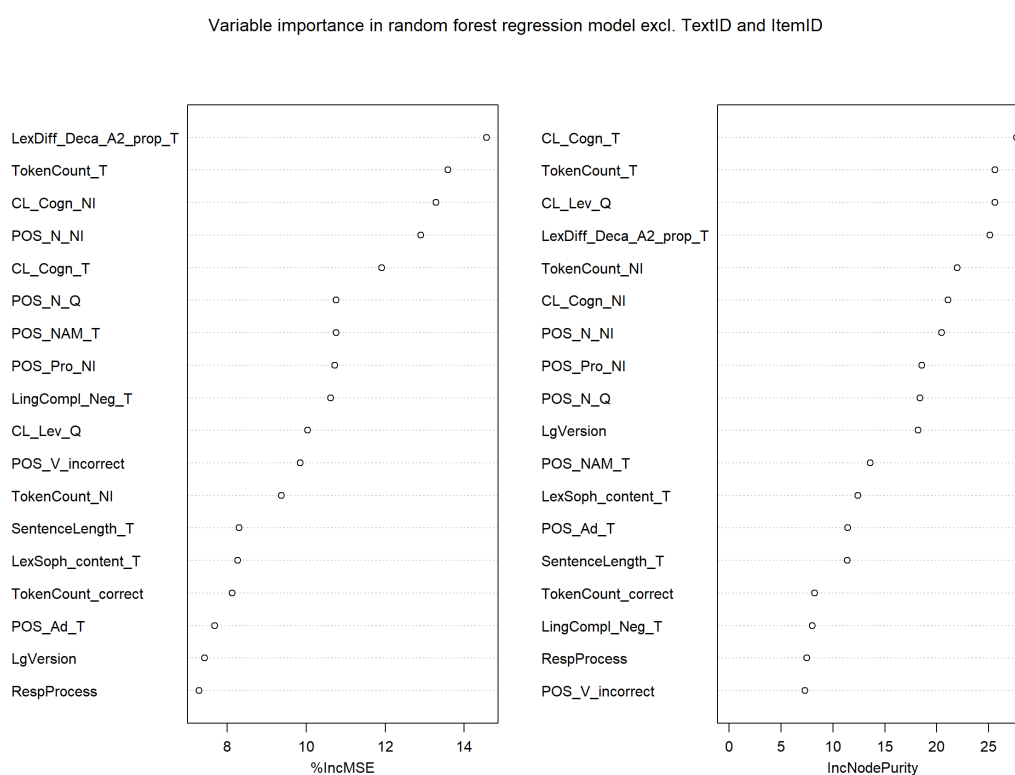


Figure 12.8: Variable importance plot of random forest regression model. The two plots represent different measures of importance.

The effect of a single variable within a random forest regression model is notoriously difficult to interpret. Hence, a single tree was estimated using CART (see fig. B.11, page 289). It is a good example of how the language version serves as a first predictor, indicating that the ENde and the FRit language version are linked to easier items. Among the two language groups, indicators of text length and POS variables are prevalent as splits, which, on closer examination, links easier items with shorter, often non-continuous texts about facts, and harder items with questions about actions that are described in more detailed, often longer texts on people²³⁵.

²³⁴This indicates the percent increase in mean squared error if a variable is permuted, i.e. its values scrambled, which effectively cancels its predictive power. Particularly high values indicate a high importance of said variable in the random forest model.

²³⁵Compare, e.g., the items and input texts in the first leaf from the left with those in leaves 9 through 11.

12.3.6 Beyond assessment characteristics: Bayesian regression

Given the shortcomings of regression modelling and random forests, a third statistical approach was chosen to investigate the patterns in the data: Bayesian logistic regression modelling. Compared to frequentist IRT, Bayesian statistics allow for more flexibility in modelling the responses and item characteristics, including covariates. They are also more informative in their output. However, even a relatively simple model comparable to the frequentist MFR model described above (see section 11.2.3) is computing-power intensive and takes several hours to converge on a standard computer. For this reason, in this study, only three models were estimated using 20% of the data set²³⁶. These are reported in the following.

To establish a baseline, a Bayesian multilevel logistic regression model was constructed first (following Bürkner 2021, pp. 14 sq.)²³⁷. It allows each item to vary within each language version, making it comparable to the parametric MFR model that was used in this study. Weakly informative priors (i.e. half-normal (0, 3)) were used for the fixed effects and the language version random effect (again following *ibid.*, pp. 10, 23). Item and person parameters were modelled via partial pooling, which forced their estimates towards a common mean (cf. *ibid.*, pp. 14 sq.). The model converged well ($\hat{R} \approx 1$) and showed moderate to high correlations among item parameters across language versions. Predictive performance was strong according to LOO cross-validation diagnostics²³⁸. The estimated item difficulties of this model correlate highly with those of the MFR model ($\rho = -.91$ ²³⁹, $p < 0.001$, see also fig. B.12, page 290). This is despite the fact that it is based only on 20% of the responses which were used to estimate the parametric MFR model.

A second model estimated response probability based on assessment and test taker characteristics (see table 12.8). The random effect of the language version on items from the above model was replaced by a random intercept for test takers, items, and texts under language versions for computing reasons. In addition, the predictors from the mixed-effects regression model and the test taker variables prepared for this analysis were added to the model. More informative priors were used in this analysis based on the results from the MFR-like model and additional preliminary exploratory modelling²⁴⁰. This model also converged well, and LOO cross-validation diagnostics are good. Due to missing data in the NIs and item parts, this model is based on 61 items and data from 4797 test takers, approximately 15% of the data.

The Bayesian version of R^2 for this model suggests that it explains 22.3% of the variance in the data, which is similar to the results of the regression analyses. This is reflected in the still substantial variation which remains in the random effects. The SDs related to the interaction between texts and language versions in particular show that the input text still carries a lot of information beyond the linguistic covariates, and that the texts differ substantially between the language versions. Among the covariates related to the characteristics of the assessment, four do not include 0: `TokenCount_T`, `LexSoph_tokens_T`, `CL_Cogn_NI` and `Overl_sem_T_correct` seem to add some information to the response probability estimated in the model, beyond the

²³⁶20% of all test takers in each language version were sampled randomly in order to retain the ratio of responses from the curricular regions.

²³⁷The dependent variable in this and the other two models is response probability. High values mean more chances at a correct response. The predictors, both random and fixed effects, are added to this response probability. If random effects have large SDs, this can be interpreted as a lot of variance not explained by the fixed effects. If the confidence interval of the fixed effects is around 0, they potentially do not add reliable information to the estimation of the dependent variable. Significance testing is not done in Bayesian statistics. In all models, test taker ability is modelled as a random effect.

²³⁸LOO cross-validation is a technique where a model is trained on the entire data set except one observation. It is then tested on this one observation. In the context of Bayesian statistics, this exact technique would require immense computing power, but the `loo()` function of the `brms` package uses an approximation which has been shown to be a good estimation of predictive accuracy. See for more technical information Vehtari, Gelman, and Gabry 2017.

²³⁹The model estimates item difficulty as item easiness, which is on an inverted scale compared to most estimations in frequentist IRT (Bürkner 2021, p. 6).

²⁴⁰The R script only contains the final models to reduce complexity. Due to the relative novelty of this statistical approach, some experimental modelling, which was done beforehand, is not reported.

sd(Intercept) of	Estimate	lower 95% CI	upper 95% CI
Respondent	0.33	0.30	0.36
Item	0.58	0.47	0.71
LgVersion	0.50	0.24	0.95
LgVersion:TextID	0.32	0.29	0.42

Predictor	Estimate	lower 95% CI	upper 95% CI
<i>Intercept</i>	2.85	0.77	5.02
TokenCount_T	-0.00	-0.01	-0.00
LexDiv_Var_T	0.00	-0.00	0.00
LexSoph_tokens_T	-0.31	-0.58	-0.01
LexDiff_CEFRlex_A1_prop_T	-0.33	-1.11	0.43
Overl_sem_T_correct	0.24	0.13	0.34
POS_V_NI	0.00	-0.00	0.01
CL_Cogn_NI	0.25	0.06	0.43
POS_Ad_Q	-0.01	-0.01	-0.00
RespProcessrewo	0.06	-0.35	0.45
RespProcesstext	-0.01	-0.33	0.32
Overl_sem_NI_incorrect	-0.01	-0.33	0.32
<i>Immigration Status</i>	0.05	0.02	0.08
<i>L2 at Home</i>	0.33	0.23	0.42
<i>L2 Listening</i>	0.72	0.68	0.75
<i>L1 Reading</i>	0.27	0.23	0.31
<i>L1 Orthography</i>	0.04	0.01	0.08

Table 12.8: Summary of Bayesian multilevel logistic regression with covariates. *Note:* CI = Confidence Interval. All $\hat{R} = 1.00 - 1.01$.

item, input text and language version themselves. Both longer texts²⁴¹ and a higher lexical sophistication value (i.e. on average more frequent tokens in the input text) lead to a decrease in response probability, while a higher proportion of cognates in the NI or more overlap between the correct answer and the input text will lead to higher scores. Inversely, the test taker variables are highly informative: higher scores in the L2 listening assessment are associated with much higher response probabilities. Students who indicated that they use the L2 at home also score higher. The L1 reading assessment results can also be considered a good predictor of L2 reading, and even the L1 orthography assessment results add to the model, though the effect is weak. Interestingly, students who have a migration history also have a slightly higher response probability.

A final, quite exploratory Bayesian GLMM modelled interactions between some item and person characteristics. Two types of interactions were hypothesised: a) item difficulty may function differentially if students have a migration history and/or if they use the foreign language at home; b) the item characteristics involving the language of schooling may function differently for students who speak the foreign language at home. The model is a mixture of the MFR-like model

²⁴¹Though the estimate is given as -0.00 , visual inspection of the conditional effect of this variable shows that if respondents encountered long texts, their response probability is substantially lower, just outside the confidence interval of the shortest texts. The low estimate is due to this variable not having been scaled.

and the second model described above: item difficulty is modelled separately for each language version, and the covariates were chosen based on the results from the second model. It converges nicely, and LOO cross-validation shows good fit, though this model still explains only 24% of the variance. See table 12.9 for a summary of the model estimates.

sd(Intercept) of	Estimate	lower 95% CI	upper 95% CI
<i>Test takers</i>	0.37	0.33	0.40
<i>Items in DEfr</i>	0.72	0.60	0.87
<i>Items in ENde</i>	0.73	0.60	0.88
<i>Items in FRde</i>	0.78	0.64	0.96
<i>Items in FRit</i>	0.79	0.62	0.98
LgVersion	0.44	0.20	0.88
LgVersion:TextID	0.13	0.01	0.29

Predictor	Estimate	lower 95% CI	upper 95% CI
<i>Intercept</i>	0.36	-0.30	0.97
Overl_sem_T_correct	0.39	0.16	0.62
CL_Cogn_NI	0.44	0.06	0.82
Overl_sem_NI_incorrect	0.26	-0.24	0.74
<i>Immigration Status</i>	0.04	-0.03	0.11
<i>L2 at Home</i>	0.33	0.05	0.61
<i>L2 Listening</i>	0.74	0.70	0.79
<i>L1 Reading</i>	0.33	0.29	0.37
Overl_sem_T_correct:L2Home	0.07	-0.17	0.31
CL_Cogn_NI:L2Home	-0.24	-0.61	0.12
Overl_sem_NI_incorrect:L2Home	-0.14	-0.54	0.27
item	-0.00	-0.01	0.00
ImmSt:item	0.00	-0.00	0.00
L2Home:item	0.00	-0.00	0.00

Table 12.9: Bayesian multilevel logistic regression with interactions. *Note:* CI = Confidence Interval. All $\hat{R} = 1.00$.

Table 12.9 confirms the finding from above that the cognate and the overlap variable and the test taker variables add some information to the estimation of the response probability. The effect is even more pronounced when item difficulty is modelled to vary across the language versions. The assumed interactions, however, do not appear to make a large difference, though the confidence intervals of the interactions including **L2Home** lean towards a negative effect: a higher proportion of cognates in the NI or more overlap between the NI and the incorrect answer, paired with the L2 being spoken at home, may lead to a lower response probability beyond the effect that those variables already have on their own. No interaction was found between the item intercepts and the two test taker variables **ImmSt** and **L2Home**.

12.4 Summary

Multicollinearity analyses showed moderate or even strong correlations between several groups of variables. Most importantly, there is an unexpected effect of the way the input texts were designed in this particular assessment which links operationalisations of theoretically distinct aspects with each other: item type is related to a continuousness measure of the input texts, which is then connected to text length and incidences of POSs. These, in turn, have connections to all measures of lexical aspects of the input texts, especially to the lexical sophistication measures. The crosslinguistic variables are also linked to this complex structure through the incidence of verbs. Similar patterns can be observed for the NIs and the item parts. Some characteristics of the items also feature particularities that may be related to the design of the assessment: some questions or stems have a high incidence of names that occurs only in connection with certain input text types. Also, some features of the question or stem, the correct and the incorrect answer options vary together along with certain question types. Overall, only the semantic overlap variables are less connected to other aspects of the assessment.

Factor and cluster analyses of these variables were conducted in an attempt to reduce data complexity. Factor analyses were applied to the linguistic characteristics of the input texts in each language version, while the characteristics of the NIs and item parts were used in a crosslinguistic cluster analysis. Though the results from these analyses proved to not be helpful for the subsequent analyses, they confirmed the patterns identified in the multicollinearity analyses: the POS variables are related to each other with the exception of POS_NAM_T. The lexical sophistication measures are naturally linked with each other, but they also vary along with the POS measures and with text length. The lexical diversity measures are also connected to the POS variables, but they do not load onto the same dimension among themselves. Finally, in the DEfr and the FRde language versions, the two crosslinguistic variables load onto a common factor with minor contributions from POS variables. The ENde and FRit version do not include these two variables. Overall, the factor analyses underline the impression that the linguistic characteristics of the language versions are similar in many cases, but diverge quite clearly in others, particularly with respect to the lexical diversity and the crosslinguistic variables. Finally, the cluster analyses identified groups characterised by more or less nouns, verbs or names respectively.

Correlational analyses identified several, exclusively text-related variables which are moderately related with item difficulty in all language versions (Spearman's $\rho \approx |.30|$). The strongest variables are text length and the continuousness measure (positive correlations), one lexical sophistication measure (positive correlations²⁴²), the lexical difficulty of the input texts operationalised as the proportion of A2 words identified by Decaleon (negative correlations), and the incidences of nouns (negative correlations) and verbs (positive correlations) in the input texts. Some additional variables are correlated with item difficulty in most language versions, but these are usually related to the six variables already mentioned. The variables operationalising cognates in the input text, question and NIs also feature significant correlations in at least one language version, though not always the same. When significant, their correlation with item difficulty is around $\rho = -.20$. Variables referring to characteristics of the answer options are rarely correlated with item difficulty, often only in one language version. The need for background knowledge and the response processes were found to correlate significantly (positively) with one language version, possibly pointing to a random occurrence.

In all cases, graphical examinations of these correlations show that even the highest correlations leave a lot of uncertainty: often, a few data points on the edges of the item difficulty spectrum determine the slope of the regression line, but this accounts for very little information in the middle of the item difficulty spectrum.

²⁴²As mentioned before, a positive correlation actually indicates that easier items are linked to input texts with lower lexical sophistication values, i.e. *less* sophisticated texts in terms of word frequency. See section 13.1.1.

Four language-specific regression models were estimated. In English, lexical variety of the input text, overlap between the input text and the correct answer, and the Damerau-Levenshtein distance of the question make up the predictors of a model which explains 20% of the variance based on 56 items. In the German test, text length of the input text, cognateness and sentence length of the NI, another overlap variable and a POS variable of the question and the correct answer explain 48% of the variance, though based only on 47 items. The model for the FRde data set is similar to those two models: the lexical sophistication and cognateness of the input text and three POS variables, one each of the NI, the question and the correct answer predict about 34% of the variance in item difficulty. The FRit regression model is the only model which is structured differently, with an interaction of text length and the incidence of nouns in the input text, joined by the NI's cognateness predicting 24% of the variance in item difficulty, though based on 90 items instead of the lower number of items in the other models. All of these models have large remaining proportions of variance, and their performance on the data from the other language versions is mediocre at best: RMSEs are often close to 1 logit.

To assess the transferability of item difficulty prediction, an additional multilingual regression model was fitted to the entire data set. The final mixed-effects regression model has two random effects: the language version and the **TextID**. The fixed effects are similar to the language-specific regression models, but include the categorical text-by-item variables **RespProcess**. The entire mixed-effects model explains 72% of the variance, though most of it is dominated by the input text and, to a lesser extent, the language version. The fixed effects only explain 23% of the variance, comparable to what was found in the language-specific modelling. The model confirms that the ENde and the FRit language version have lower item difficulties, all else being equal.

A random forest analysis of the entire multilingual data set confirms the general findings of the regression analyses: when the **TextID** is included in the model, it is by far the best predictor of item difficulty. If it is eliminated, linguistic characteristics of the input texts, and the cognate variables related to both the NI and the input text become the best predictors. Since random forests can usually take into account more variables than linear regression, the variance explained by these analyses is a bit higher than of the fixed-effects regression models (35%), but it still leaves a lot of uncertainty.

Finally, Bayesian modelling mostly confirms the findings mentioned above. An overlap variable and the cognateness of the NI are the strongest assessment-related predictors in the models. The role of the input text and its linguistic characteristics also still figure prominently. The effect of the language versions is confirmed. This type of modelling also underlines the predictive power of the information on the test takers used in this study: if their performance at the L2 listening assessment and the L1 reading assessment are known, and if their use of the foreign language at home is taken into account, most of the assessment characteristics become of lesser importance to the prediction of the response. Nevertheless, a large amount of the variance remains unexplained: even these relatively complex models only reach R^2 values up to 0.24.

Part IV

Discussion and conclusion

Chapter 13

Discussion

This study analysed the interaction between empirical item difficulty and characteristics of the assessment and test takers of a low-proficiency foreign language reading assessment in four language versions: an English test administered to speakers of (Swiss) German (ENde, $N = 11,845$), a French test administered to speakers of (Swiss) German (FRde, $N = 2,025$) and to speakers of Italian (FRit, $N = 823$), and a German test administered to speakers of French (DEfr, $N = 5,484$). The tests consisted of 4×39 input texts and 4×94 closed-response items which were translations of each other. The assessment was part of an educational monitoring conducted in Switzerland in 2017 (ÜGK, cf. Konsortium ÜGK 2019a). The test takers were 12-year-old students at the end of primary school, selected as a representative sample at the cantonal level. The relationship of 42 selected variables with empirical item difficulty was studied using both inferential and non-parametric statistics. The variables represent assessment-related and linguistic characteristics of the input texts, NIs and items as well as attributes of the test takers. To study how item difficulty varies with the *language* of the assessment, analyses were conducted both on separate language versions and on the entire dataset.

It can be shown that item difficulty in this foreign language assessment is influenced by the test takers' knowledge of the language of schooling: variables which operationalise the existence of cognates (i.e. words which resemble a translation equivalent in the language of schooling) and overlap between input texts in the foreign language and item parts in the language of schooling often feature among the better predictors of item difficulty. This finding is supported by qualitative examinations of items which show DIF due to the language versions. Overall, the results confirm the complexity and multivariate nature of item difficulty found in earlier studies: no single variable has an identifiable, immutable effect on item difficulty. Instead, characteristics of the input texts and items interact with each other, cancel each other out or create unexpected side-effects. Crosslinguistic comparisons underline the difficulty in operationalising the linguistic characteristics of short texts in several languages. The addition of test taker attributes underlines the importance of L2 knowledge and L1 reading ability in L2 reading comprehension.

13.1 Summary of the main findings

The three major RQs in this thesis aim at a deeper understanding of the reasons for item difficulty on three levels: RQ 1 is interested in the prediction of item difficulty in each of the four language versions of the ÜGK foreign language reading assessment. RQ 2c, supported by the findings for RQs 2a and 2b, aims at predicting item difficulty *across* the language versions. Finally, RQ 3 examines the role of the test takers in predicting test results. In the following, these questions are answered and discussed in some detail based on the results presented above.

13.1.1 RQ 1: Prediction of item difficulty in foreign language reading assessment

As a first step in this analysis, item difficulty was investigated as if the ÜGK foreign language reading assessment was four different tests. For each of the language versions, good predictors of item difficulty were chosen among the characteristics of the assessment. As such, these analyses are in line with other item difficulty studies (see chapter 8), though the study of assessments of languages other than English is rare.

RQ 1: Which characteristics of the assessment (its input texts and items) predict empirical item difficulty in each of the four language versions?

Correlational and regression analyses identified a variety of item, text and text-by-item variables as potential predictors for empirical item difficulty. In all language versions, several characteristics of the input text correlated significantly with item difficulty (see table 12.3, page 206): text length, a lexical sophistication measure, the proportion of A2 words, the proportions of nouns and verbs, and a categorical variable which expresses whether a text is continuous or not, correlate moderately with item difficulty in all language versions. At least one cognateness variable shows a significant correlation with item difficulty in each language version. In addition, each version has its own set of significant predictors: in the ENde test version, most lexical diversity and sophistication variables related to the input text emerge as strong predictors. In the DEfr version, various operationalisations of the NI's characteristics correlate significantly with item difficulty. In both test versions involving French and German (DEfr and FRde), both the proportion of cognates and the Damerau-Levenshtein distance of the input text and the NI figure prominently among the significant predictors. Additionally, the number of negations in the input text is correlated significantly with item difficulty in the FRde language version. The FRit language version has several POS variables, particularly of names and pronouns, which are significant predictors of item difficulty. The same is true for the number of negations in the items. The text-by-item variables are less obviously related to item difficulty in all language versions, though the semantic overlap between the input text and the correct answer features at least in the ENde and the DEfr language version. The need for background knowledge and the variable operationalising response processes are also on the list, though they are significantly correlated with item difficulty in only one language version.

Many of the variables identified in the correlational analyses were also found to play some role in the language-specific regression models. In each language version, one text-related predictor, either text length, lexical diversity or lexical sophistication explains a portion of the variance in item difficulty, joined by a cognateness variable of the input text or the NI. In the DEfr and ENde tests, an overlap variable is included in the model, while in the FRde language version, items are characterised by their proportion of pronouns. The FRit language version is the only one in which the characteristics of the items were not selected for the final regression model. Instead, an interaction between text length and the proportion of nouns is modelled, which can be understood as an operationalisation of text genres (see section 12.1.4). Since it could be shown that text types and the characteristics of the item are related to each other in this assessment, this interaction can be interpreted in a similar way as the POS variable in the FRde language version.

In all cases, the regression models confirm that the overall linguistic characteristics of the input text and the NI play a role in determining item difficulty, but they also underline the importance of the language of schooling in this foreign language assessment, expressed as a proportion of transparent cognates. There is also evidence that the way items are constructed, what they ask and how they refer to the input text, is linked to empirical item difficulty.

The direction of the predicted effects of all of these variables on item difficulty is often as expected: more cognates in all parts of the assessment are related to lower item difficulty, as is more semantic overlap between the input text and the correct answer option. Longer input texts

are related to more difficult items. In this assessment, these texts are more continuous and feature relatively more verbs and less nouns, which may explain the counter-intuitive finding that texts with a *lower* mean lexical sophistication are related to items with *higher* difficulty: the verbs in these low-proficiency foreign language reading texts are often quite frequent in the reference corpora, whereas there are a number of nouns which do not appear as often in everyday language as measured by SUBTLEX. As a result, longer texts in this assessment, which are related to more difficult items but usually have a higher proportion of verbs, are operationalised as having a lower mean lexical sophistication than several short texts which have a high proportion of sometimes quite infrequent nouns. The same mechanism could be the reason that a *higher* proportion of A2 words, which should make texts more difficult to understand, is also linked to *lower* item difficulties.

The language-specific correlation of **BackgroundKn** and the predicted effect of **RespProcess** in the multilingual regression model also point in the “wrong”, i.e. unexpected direction: both the need for (some) background knowledge and response processes that are expected to be more challenging are correlated with *lower* item difficulties in one language version each. Graphical examinations of these connections, however, suggest that they are random artifacts: in both cases, the items which were coded as needing background knowledge or as bridging and inference items, are not frequent and they are, possibly by chance, not among the most difficult ones²⁴³. Similar observations can be made for other variables which have low occurrences in some categories, e.g. the number of negations in the item or POS variables of the item parts. In general, visual inspections of the correlations suggest that the link between item difficulty and most variables is tenuous. Even where correlations are higher, they are often quite clearly driven by items at the outer ends of the item difficulty continuum. The relatively low R^2 in the regression models and the wide confidence intervals for the predicted effects point in the same direction.

13.1.2 RQ 2: Prediction of item difficulty in foreign language reading *across* languages

At the heart of this study, and at its origin, is the question of how the differences in item difficulty in the ÜGK foreign language reading assessment can be explained. Is reading item difficulty at a certain CEFR level essentially the same across languages, with differences resulting from how the languages work? Or is item difficulty language-specific and not predictable in general terms? The three RQs 2a to 2c are an attempt to shed light on this.

RQ 2a: Can the characteristics of the assessment be operationalised across the four languages?

The answer to this question is either technical or conceptual. Technically, it could be shown that, if these reasonably related languages can be POS tagged and if the necessary reference corpora are available, then it is possible to extract a variety of linguistic variables from texts in all languages. This assertion is limited by two aspects, however. First, the number of tokens: only longer pieces of text, perhaps beyond 50 tokens, can be reliably tagged in the semi-automatic way used in this thesis. If only a few tokens or lemmas are available, even small differences between the language versions can lead to large differences in their operationalisation. In addition, since traditional POS taggers rely on sentence boundaries, even longer texts need to be as continuous as possible. This means that the characteristics of NIs and items are difficult to determine with traditional means of linguistic analysis and, hence, hard to compare. Second, POS taggers are not (yet) multilingual. The TreeTagger instance used in this thesis tagged the languages based on separate tag sets and parsers, which led to different interpretation of word groups and lemmas in some cases. This has probably changed the way some variables were calculated, even before the

²⁴³See the plots `Corrs_xsi_BackgroundKn.png` and `Corrs-xsi_RespProcess.png` output into the folder `4_plots` by the R script `4_AnalyticalStatistics_Regression`.

differences in reference corpora and operationalisations are taken into account. This is particularly true when only content words are considered.

Conceptually, the answer to RQ 2b (see below) becomes relevant: quite a few variables are significantly different across the language versions, despite their being operationalised in the same way. For most of them, however, well-founded validity studies are rare, especially across languages. It remains open, then, whether they simply measure different things in the various languages, or whether these texts actually differ in these aspects even though they are translations of each other. Given the evidence presented in this thesis, the answer is probably: both.

RQ 2b: Are the linguistic characteristics of the translated assessment parts identical across the four language versions?

The descriptive and multicollinearity analyses described above clearly show that despite their being translations of each other, the linguistic characteristics of the reading texts are not identical across the three foreign languages of the assessment. The surface-level characteristics of these texts match, text length is highly correlated, and the proportions of POSs are, too. Yet, the lexical variables in particular differ, sometimes wildly, across the language versions. This is especially true for the measures of lexical diversity, which are somewhat experimental and have not been researched much in terms of their validity in languages other than English. For these measures, a remarkable finding is that the French and German texts are often more similar to each other than to the English ones. This may be due to them both being more inflected languages, but that would have to be investigated more closely and lies outside the purpose of this thesis. The lexical sophistication measures rely on language-specific reference corpora, which makes it unsurprising that their scales differ, but also remarkable that the log-transformed variables still correlate strongly with each other. Given their connection to other linguistic characteristics of the input texts, it remains open whether lexical sophistication can actually be meaningfully compared across languages if the reference corpora are big and similar enough or whether the fact that these texts are similar in other aspects (e.g. their distributions of POSs) is at the origin of this shared variance. Finally, with regards to lexical difficulty, the use of language-specific CEFRlex corpora results in proportions of A1 and A2 words that are clearly not comparable across the language versions. The cause of this discrepancy remains uncertain. If we assume that tokens can, in principle, be reliably assigned to CEFR levels, is it plausible that the input texts truly differ so significantly across the three foreign languages that they contain markedly different proportions of A1 and A2 words? Or, alternatively, might the CEFRlex corpora themselves be too inconsistent in their construction to allow for meaningful cross-linguistic comparisons of lexical difficulty? This question remains unresolved, but would certainly be an interesting endeavour to investigate.

RQ 2c: Do the characteristics of the assessment predict item difficulty in the same way in the four language versions? If not: what are the major differences?

When the entire dataset is modelled together, the input text usually becomes a major predictor, more productive even than the language version. This suggests that items which belong to the same input text behave in similar ways, especially across language versions, beyond what is explained by the linguistic characteristics of these texts. In the ÜGK foreign language assessment, then, item difficulty is at least partly created by the input text, certain characteristics of the NI, and their interaction with the items, irrespective of the languages at play. A smaller, but still essential part of the variance is attributed to the language version: the ENde and the FRit version are systematically linked to easier items, while the items of the FRde and the DEfr language versions are more difficult. These language-specific effects act on item difficulty *beyond* the variance predicted by the input texts and a few linguistic variables, including cognateness and overlap variables. Impressionistic evidence from individual items suggests that particularly

easy or hard items feature specific words in the NI which are transparent to test takers in one language version but not in another. Given the fact that English and German are related, as are French and Italian, it does not seem surprising that items in these two language versions are more often on the easier side than in the other two.

It has to be pointed out, though, that the results underline that only a smaller part of the variance in item difficulty can actually be explained by the variables chosen in this study, both in language-specific and in multilingual analyses. This may, of course, be due to the selection of variables or their inadequate operationalisation (or both). The above-mentioned impressionistic item analyses point to more complex processes of item solving however, which cannot be modelled by relating a variety of characteristics with item difficulty in more or less linear models. Instead, each item has its own group of mechanisms which determine whether students will solve it correctly or not. In the ÜGK foreign language assessment, this is particularly visible in items which function differently in the various language versions: often, one or several specific features can be identified which have a potential to change the way test takers will read an input text, interpret an item and/or find the correct answer. The answer to RQ 2c, then, is *probably yes* in the sense that features which make items easier (or harder) were identified in all language versions, though their effect is individual to each item and hence, impossible to predict unequivocally.

Nevertheless, very general mechanisms, like a connection between the longer, more continuous input texts and harder items, or the facilitating effect of transparent cognates in input texts and NIs can be observed in all language versions in similar ways.

13.1.3 RQ 3: Test taker variables in test response prediction

A final matter of interest in this thesis, given the availability of the data, is the role of certain test taker attributes in predicting test responses. Since neither CTT nor IRT actually model information about test takers in their estimation of item difficulty, Bayesian statistics were employed to study possible interactions between test taker attributes and characteristics of the assessment.

RQ 3: Do certain characteristics of the test takers interact with the item characteristics for the prediction of item difficulty in the four language versions?

In modelling test taker ability, both parametric and Bayesian statistics identified the results of the L2 listening assessment as the strongest predictor for L2 reading ability, followed by the results of the L1 reading assessment. Test takers who indicated that they used the foreign language at home also usually reached higher ability estimates. None of these findings are surprising, but they can be considered as validity evidence: L2 listening and L1 reading are both expected to tap into similar skills as L2 reading, and the fact that potentially more proficient users of the L2 scored higher on this assessment is at least a good indication that L2 comprehension is measured.

Beyond the prediction of test taker ability, Bayesian statistics actually made it possible to model interactions between test takers and items: first, it was tested whether students with a migration history or students who use the L2 at home interpreted certain items differently, i.e. if there is an interaction between items and these two test taker variables beyond the influence of the fixed variables. This could not be confirmed. Interactions between certain crosslinguistic variables which were known to predict item difficulty, and the L2Home variable was also included in the model, but did not prove to be productive either. They do, however, show a weak tendency which implies that the effect of crosslinguistic characteristics of assessment items may be stronger for students who do not use the foreign language at home.

All findings based on Bayesian statistics have to be taken with a grain of salt, since this statistical approach does not lend itself to exploratory analyses: even a limited amount of data takes many hours to compute on a home computer, which makes these kinds of analyses bothersome to conduct. Under these conditions, an exhaustive exploration of combinations of variables which are good predictors of test responses is a difficult endeavour.

13.2 Additional findings

The multitude of variables and descriptive analyses, as well as the experimental nature of some variables lend themselves to an investigation of several additional lines of questioning beyond the nature of item difficulty. In the following, some findings related to the operationalisation of textual characteristics, particularly across languages, are discussed. The study also allows some observations about the nature of L2 reading and assessment, which are briefly considered in sections 13.2.4 and 13.2.6. A short final section summarises further notions which could be examined in this dataset.

13.2.1 Transferability of findings from research on English to other languages

Research on item difficulty and on linguistic characteristics of reading (or listening) texts has been focusing on English, both as an L1 and an L2, and it is unclear whether or to what extent the findings from these lines of research can be transferred to languages other than English. In fact, even in research which did not (exclusively) study English, the transferability of methodological approaches or of earlier findings from English to another language is not usually discussed. Instead, operationalisations are adapted more or less directly, though sometimes supplemented by findings from empirical research on linguistic complexity in the target language (e.g. Neumann 2012, p. 259).

Linguistic research in other fields, especially NLP, offers more insights into the operationalisation of textual characteristics across languages. For instance, Bengoetxea and Gonzalez-Dios (2021) described the creation of a vast multilingual text analyser for Basque, Spanish and English (the *MultiAzterTest*). They analysed the efficiency of the tool when only parallel reference material is used in all three languages, and concluded that though the multilingual analyses performed well, they were not better than monolingual analyses using all available resources of a language. Given the field, the article focuses on technical aspects of the multilingual operationalisation of textual characteristics rather than on their constructs, but the authors do point out “that linguistic feature design should not only be based on mainstream and majority languages but also in [sic!] languages with other typologies” (ibid., p. 25).

In the present study, despite the open questions regarding the validity of English-based textual characteristics, identical operationalisations were used in the analysis of English, French and German texts and they were treated as the same construct across languages during data analysis. At least two problematic aspects result from this: first, most of the textual variables used in this thesis relied on results of the POS tagging. Since TreeTagger used language-specific resources, it had the potential to introduce error into the operationalisation of the linguistic variables. The often observed cross-linguistic variation of some variables confirms this expectation, but in some cases, the language-specific POS tagging alone does not seem to be a sufficient explanation (see section 13.2.2 below for the example of the MTLT). Second, in the multilingual data analyses, the linguistic characteristics of the input texts were not modelled to vary across the language versions for computational reasons. Instead, they were assumed to measure the same construct, e.g. lexical variability, which, as has been discussed, is not actually a given. It is possible, then, that the models might have reached a better explanation of variance if certain characteristics of the assessment had been allowed to interact with the language version as well²⁴⁴.

Thus, the transferability of the operationalisation of linguistic characteristics from English to other languages remains mostly unclear, though the dataset contains some potential for deeper analyses of the matter. Tools like the MultiAzterTest (ibid., see above) or CTAP (X. Chen and Meurers 2016, see also section 4.1.3.2), with their ability to extract multiple measures of linguistic complexity in several languages, may be a helpful resource in this endeavour. Nevertheless, these

²⁴⁴Note, however, that the results of the monolingual item difficulty analyses were overall similar to those based on the multilingual dataset. As a result, it is equally possible that an interaction of the predictive variables with the language version would not have changed much.

cannot replace a careful study of construct representation for many of the linguistic measures and reference materials used in linguistic research.

13.2.2 Measuring lexical diversity

In section 5.2.1, it was explained how lexical diversity can be defined as a multi-dimensional construct composed of *variability*, *volume*, *evenness*, *rarity*, *dispersion* and *disparity* (Jarvis 2013b, pp. 22 sqq.). Jarvis and others (Jarvis 2013a, 2017; Kyle, Crossley, and Jarvis 2021; Vanhove, Bonvin, et al. 2019) suggested operationalisations of these subconstructs, some of which were adopted in this thesis to capture a broad construct of lexical diversity (see table 9.5, page 154 for their operationalisations). Since these measures are still largely experimental and little is known about their meaningfulness across languages, this section highlights some observations related to them. As was mentioned above, very short texts (i.e. the item parts and most of the NIs) could not be meaningfully analysed for their lexical diversity. The various lexical diversity measures of the *input texts* in the three foreign languages, however, were investigated with respect to their crosslinguistic comparability, their relationship with other linguistic characteristics of the input texts, and their power in predicting item difficulty.

The variable `LexDiv_Eve` had been adapted from Vanhove and colleagues (2019) and has proved to be highly correlated with text length in all language versions. Since it also systematically loaded highly on the same factor as text length in the factor analyses, it will probably have to be re-operationalised to measure something (more) unrelated to text length²⁴⁵. A similar caveat has to be formulated for the lexical density and, possibly, the lexical dispersion measures.

With respect to the applicability of these measures across languages, it is interesting to note that though they usually correlate moderately across the languages, the means of the several lexical diversity measures are significantly different in English than in the other two languages (see also section 11.1.3). For instance, the MTLDs of the English texts are lower on average and have a visibly lower variance than those of the French and German texts (see fig. B.5, page 281): most particularly, the MTLDs of the 39 English input texts range between 39 and 117. In German and French, the upper margin is much higher (German: 287, French: 247).

This is remarkable since these texts are translations of each other, which implies that their lexical variability should be at least similar. A closer look at the MTLDs of the various input texts shows that usually, the MTLD of the English text is the lowest, but not by much (see fig. B.6, page 282). In some cases, however, either the French or the German input text reach an unexpectedly high MTLD, which is usually an artifact of the way the measure is calculated: the MTLD is a count of how many segments in an input text reach a certain TTR; if it is not reached, either due to the number of tokens or because the text has very little repetition, a “partial” segment is used to calculate the final MTLD value. This can lead to values which have little actual meaning²⁴⁶. A possible reason for this difference in calculating the MTLD may be the more inflected nature of German and French: the TTR falls more slowly because there are more diverse tokens in the languages due to inflections. A possible solution would be to adjust the TTR threshold for this measure, though it is unclear what this would mean for its cross-linguistic comparability²⁴⁷.

Another way of examining the lexical diversity measures suggested by Jarvis and others are the results of the factor analyses described in section 12.2.1. These include the lexical sophistication measures, which Jarvis includes in his construct of lexical diversity as “rarity” (see section 5.2.1 for references). The factor analyses in the four languages did not specifically focus

²⁴⁵Or, alternatively, it is possible that lexical evenness is, in fact, strongly correlated with text length. This would have to be investigated further.

²⁴⁶For instance, text 16 in German reaches a TTR of 0.91 after 96 tokens. This translated to an MTLD of 286.72. Whether this should really be interpreted as an *extremely* diverse text is doubtful.

²⁴⁷Admittedly, even without this adjustment, it is unclear how to judge the validity of a comparison of the MTLD across languages.

on the lexical diversity indices, but it has to be pointed out that most of these measures were eliminated when they were considered along with the other linguistic characteristics of the input texts. Only lexical evenness and text length, which always loaded together on the same factor, as well as the lexical sophistication values usually remained in the models. The latter had a tendency to form a common factor along with the POS variables (especially nouns) or at least the incidence of negations in the input texts²⁴⁸. Where the lexical variability measure MTLTD remained in the factor structure, it showed an opposing loading pattern on the factor which also included lexical sophistication. Three possible conclusions can be drawn from this: first, the various values do not form an obvious factor, at least if other linguistic characteristics are included in the analysis. This does, of course, not mean much, but it is an indication that these variables do not (yet) form a clear, separate construct, particularly given the role played by the POS variables in the factor analyses. Second, there is some evidence that lexical sophistication and lexical variability measure aspects of the input text which have something in common, while text length (i.e. lexical volume) and perhaps lexical evenness follow a separate mechanism. Third, the quite substantial differences between the factor structures in the four language versions suggest again that there are differences in the way the linguistic characteristics are operationalised.

Finally, three aspects of the input texts' lexical diversity were found to play a role in the prediction of item difficulty: lexical volume (i.e. text length), lexical variability (the MTLTD) and lexical rarity (i.e. lexical sophistication). Interestingly, these variables never appear together in the same model which may suggest that they ultimately describe the input texts in similar ways. It may, though, also mean that their power in predicting item difficulty is unstable. After all, the sample of 39 input texts is not large. In any case, there is clearly a lot more to be learned about lexical diversity, its operationalisation and its link with item difficulty in reading comprehension assessment, particularly across languages.

13.2.3 Using reference corpora: measuring lexical sophistication and difficulty

The lexical sophistication and one type of the lexical difficulty measures used in this thesis are based on reference corpora. For lexical sophistication, the SUBTLEX frequency lists were employed, which are derived from L1 television subtitle corpora and are supposed to reflect word frequency in (scripted) spoken language. Lexical difficulty was estimated using either the CEFRlex corpora or the multilingual Decaleon wordlist (see section 9.3.1.3 for details and references). While the Decaleon list is consistent across the four languages assessed in the ÜGK foreign language assessment (and not actually based on a corpus), both SUBTLEX and CEFRlex were constructed separately for each language, albeit using comparable methods. The differences in construction, especially across languages, add to the uncertainty related to the interpretation of the measures derived from them.

For instance, the SUBTLEX lists contain different amounts of words overall and do not all differentiate between word types and lemmas in their calculation of token frequency²⁴⁹. Given the languages, they also differ, among other things, in how very frequent tokens like articles (e.g. English has only one definite article, German has six), pronouns (e.g. Italian uses pronouns much less than the other languages) or other function words (e.g. Italian *di*, English *to*, French *ne* and *pas*) are counted, and, as a result, how often nouns and verbs appear in a million tokens. Hence, it seems hardly surprising that the lexical sophistication of the input texts and item parts differ across the languages, even if the measures have been log-transformed (see section 12.1.2.2). It

²⁴⁸Note that this finding also supports the hypothesis that lexical sophistication is dependent on the proportion of nouns in the input texts. This may have led to the unexpected effect that texts with higher lexical frequencies (i.e. lower lexical sophistication) are linked to higher item difficulties, as was explained in section 13.1.1.

²⁴⁹In fact, as can be inferred from the data preparation section in the R script `2_a_VariableExtraction_itdat`, the reference lists are more different from each other in terms of their structure than what one may assume given that they stem from related corpus projects.

is unclear how this should be interpreted: how much of these differences should be attributed to actual differences in lexical frequency in those translated texts, and how much is due to the way the reference lists were constructed? Unfortunately, very little research has, to date, been dedicated to the comparability of lexical sophistication measures across languages (cf. e.g. Palfreyman and Karaki 2019), so the question will have to remain unanswered at this point.

Similar problems can be identified with respect to the CEFRlex resource. The variables starting with `LexDiff_CEFRlex` represent the proportion of content words in the input texts which are attributed to the A1 and A2 CEFR level respectively. Since both the English and the French CEFRlex datasets are actually lists of how many tokens of a word appear at each CEFR level, the variables operationalise the level of first appearance as “the” CEFR level of a token. In German, no downloadable resource was available, so a web interface was used to extract the CEFR level of all tokens. It used the same logic of first appearance²⁵⁰. This is, of course, a somewhat rough simplification of the matter: even if a word can be assumed to belong to a certain CEFR level²⁵¹, the mere fact that it has appeared at least once in an A1 textbook or simplified reader should not actually be enough to consider it being part of the “natural” lexicon of A1 users. At the very least, it seems that the determination of a CEFR level for a lexical unit would need to include a consideration of the target group and the Target Language Use (TLU) domain²⁵² under consideration. With respect to the analyses conducted in this thesis, it should be pointed out that in the three foreign languages, the majority of the content words in an input text were attributed to the A1 level (see e.g. fig. B.4). This is in line with the general expected reading level of the assessment. A2 words were identified much less, representing often less than 15% of all content words. Interestingly, the German texts contain visibly more A1 words according to this measure. Their proportion ranges roughly between .55 and .75 in French and English, but between .60 and .90 in German. It is doubtful, however, that the German input texts are this much less difficult or, inversely the French and English texts so much harder. In fact, a short quality check of the CEFR labels reveals that certain tokens were probably miscategorised: in English, the verb forms *will* and *can* are not included in the A1 vocabulary (whereas the German *werden* and *kann* are). This is probably a misrepresentation, since both verbs are part of highly frequent chunks which are taught early in ordinary English classes, and, in addition, quite transparent to speakers of German. Similar observations can be made with respect to the Decaleon database (e.g. the French verb *être* only exists as a translation of the German verb *lauten*, and is thus categorised as B1, which is certainly not adequate, cf. section 9.3.1.3). In this case, however, obvious mistakes such as the *être/lauten* one could be manually corrected before the lexical difficulty variables were calculated.

Despite these numerous incongruities and uncertainties, both the lexical sophistication and lexical difficulty measures emerge as relevant predictors of item difficulty. They also show medium to strong correlations with other linguistic measures and, notably, with each other. This latter finding provides at least partial evidence that the two groups of variables tap into closely related constructs, which is plausible: very frequent lexical units, particularly high-frequency verbs, as well as conjunctions, pronouns and other function words, are likely to appear in A1 textbooks and easy readers. There are, of course, exceptions²⁵³. Still, a general overlap between high-frequency and beginner-level vocabulary, especially at the A1 level, can be expected. The negative association between both sets of measures and the proportion of nouns in the input texts further support the idea that some form of lexical frequency is captured by both constructs. Moreover,

²⁵⁰As indicated in the web interface, cf. <https://cental.uclouvain.be/cefrlex/daflex/analyse/>, last accessed 2025-06-17.

²⁵¹In the literature, it is often discussed *how* this level can be determined (see e.g. Kusseling and Lonsdale 2013), but not whether it is conceptually sound to do this at all. As was mentioned before, the CEFR’s definition of language use does not really support the definition of levels for individual lexical units either (cf. footnote 76).

²⁵²See Bachman and Palmer 1996, pp. 18, 45 sq. for more details on TLU.

²⁵³For instance, many language classes start by teaching students how to greet each other and exchange personal information, which involves vocabulary that is not necessarily frequent in everyday language.

their (albeit weak) link to item difficulty suggests they contribute additional information to the assessment of reading comprehension. Nevertheless, future research should investigate the constructs underlying lexical sophistication and lexical difficulty more closely, and explore in more depth how, and whether, they can be meaningfully distinguished.

13.2.4 L2 reading as *reading with several languages*

In section 3.3, L2 reading was defined as “reading with [several] languages” (Grabe 2009, p. 129), a skill that draws on multiple knowledge sources and sub-skills, including more than one language. Under this framework, the languages involved interact in complex ways, shaped by context, the individual reader, and the text. In this study, the interaction between the foreign language and the language of schooling was deliberately operationalised as a characteristic of the assessment, given that all test takers were expected to have competences in both. Several of the variables reflecting this interaction emerged as recurring predictors of item difficulty across all language versions. Most notably, the CL_Cogn variables, representing the proportion of transparent cognates in the input text, NI, and item parts, consistently predicted item difficulty. This supports the notion that L2 reading is facilitated when related languages are part of a learner’s multilingual repertoire (e.g. Koda and Zehler 2008; Verhoeven, Voeten, and Vermeer 2019).

The relevance of overlap variables further reinforces this finding. As established in previous item difficulty studies (e.g. Freedle and Kostin 1993, 1996; K. Sheehan, Kostin, and Persky 2006), lexical overlap between the input text and the item is a significant predictor of item difficulty. That this effect is replicated in the present study even though input texts and item are presented in different languages underscores the idea that reading comprehension draws on similar mechanisms, whether one or several languages are involved. At the same time, the overall low variance explained by the models serves as a reminder that reading comprehension is highly individual and does not follow a single, universal pattern at all moments and for every person.

13.2.5 Multilingual assessment

In a similar vein, the multilingual (or rather: plurilingual) nature of this assessment provides evidence that the various languages in the test takers’ repertoires influence test solving processes: in addition to the quantitative findings mentioned above, the qualitative review of item difficulties and DIFs identified possible influences from the language of schooling in several items. In fact, it may well have been the most evident reason for DIF in this particular assessment. It seems hard to deny, then, that students use their knowledge of the language of schooling (and possibly other languages) while solving an L2 assessment, and that item difficulty is influenced by this. Any study of foreign language item difficulty and DIF needs to take this into account. The matter also puts into question the interpretability of foreign language test results: if test takers solve items differently depending on their linguistic repertoires, then how can we be sure of our judgment of their foreign language ability? Do we over- or underestimate competences of certain individuals due to their knowledge of other languages? On the other hand: language use is influenced by an individual’s linguistic repertoire in any case. If test results are affected in the same way, then they do, in fact, reflect an individual’s “true” competence in the foreign language, even if part of it stems from knowledge of other languages. If, however, test results are skewed due to additional language knowledge in a way that does not represent every day language use, then their interpretation becomes problematic. It is not a simple matter, unfortunately, to control for crosslinguistic effects in L2 assessments, as is evident in the present analysis of the ÜGK foreign language assessment.

13.2.6 Is L2 reading a language problem or a reading problem?

In section 3.3.2, the relationship between L1 and L2 reading was examined through the lens of Alderson’s influential article (1984), which poses a fundamental question in second language

research and assessment: is L2 reading primarily dependent on the learners' command of the target language, or on their general reading ability in their first language? The analyses presented in sections 11.3.3 and 12.3.6 support the hypothesis formulated in that section: at this relatively low level of proficiency, performance on the L2 reading assessment is more strongly associated with L2 listening comprehension than with L1 reading proficiency. This suggests that for beginning learners, good knowledge of the foreign language, which is captured in both L2 assessments, contributes more to success in L2 reading than do reading skills developed in the first language. That said, L1 reading still evidently plays a role, not least because it is connected to overall academic achievement.

Whether this balance shifts as learners become more proficient remains an open question. It is plausible that with increasing mastery of the L2, general reading strategies and abilities developed in the first language may exert greater influence on L2 reading outcomes. To explore this possibility, data from the ÜGK 2023 assessment (Erzinger, Angelone, et al. 2025) may represent a promising dataset: the assessment had a similar design but targeted students at the end of secondary school, thus capturing a later stage of language development and potentially offering insight into the ever evolving interplay between language proficiency and reading competence.

13.2.7 Additional open questions

Apart from the more prominent lines of inquiry described above, further open questions remain unexplored. In some cases, the dataset may not be suitable to investigate these points, in others, further analyses seem possible.

One area involves the role of the NI, i.e. the passage in the input text required to answer the item correctly. Previous research has shown that features of the NI can sometimes be more predictive of item difficulty than the properties of the entire input text (e.g. Böhme, Robitzsch, and Busè 2010; Brindley and Slatyer 2002; Buck and Kikumi Tatsuoka 1998). In the present study, this could not be confirmed conclusively. A likely explanation is that many NIs in this particular assessment are extremely short, which is common in assessments at lower proficiency levels but limits the potential for detailed linguistic analysis. The task of locating the NI itself may also have been harder for the low-proficiency readers involved in this assessment, amplified further by the fact that items and input text were in different languages. All of this may potentially have increased the relative importance of the full input text. These are merely hypotheses though, which would have to be investigated in another assessment or using more qualitative analyses.

Another open question concerns the cognitive processes involved in answering comprehension questions. In the present study, variables designed to capture qualitative aspects of answering an item correctly (e.g. response processes or the falsifiability of incorrect answers) did not emerge as strong predictors. This, again, contrasts with earlier work and could at least partly be due to developmental factors: younger learners may not yet apply the kinds of test-taking strategies that shape item difficulty in assessments solved by older test takers. Alternatively, or additionally, the operationalisation of these variables may have lacked precision or coverage, or the effects may only emerge in interaction with other variables.

Another as of yet underexplored dimension concerns the multilingual repertoires of test takers. The present study focused on the language of schooling and the tested foreign language, but many students in the ÜGK assessment were already learning a second foreign language, typically French or English, both of which are linguistically related to the other assessed languages. It is plausible that transparent cognates from these additional languages supported reading comprehension, especially in the case of English, which is omnipresent in learners' everyday lives. Future studies could investigate whether such cross-linguistic transfer can be measured. These would further benefit from comprehensive data on students' full linguistic repertoires, including informal language exposure through family, peers or media. This type of data is, unfortunately, not available in the ÜGK background information.

This brings into focus a final point: the need to explore interactions between assessment

characteristics and learner attributes. The current study focused on features of the assessment items and the input texts, but test takers themselves, especially their language backgrounds, reading strategies, and metacognitive abilities, may interact in complex ways with assessment design. Bayesian statistical approaches offer promising avenues for modelling such interactions more flexibly. In this case, the background data in the ÜGK dataset provides some interesting avenues, particularly regarding language learning experiences and strategies, which could be used in future analyses to better understand how individual learners approach comprehension tasks under different conditions.

Together, these extensions would help capture more fully the complexity of multilingual reading comprehension and contribute to a more nuanced understanding of item difficulty in L2 assessment contexts.

13.3 Limitations of the study

While the analyses presented here offer insights into the relationship between assessment and test taker characteristics and item difficulty in L2 reading comprehension, several limitations need to be kept in mind. These concern both the design and implementation of the study, as well as broader considerations about the nature of the constructs investigated. Specifically, the scope of the item sample and the operationalisation of the linguistic and cognitive variables pose some problems. Also, the unpredictable nature of the test-taking process, especially with young learners, is difficult to account for. Finally, methodological limitations, especially with regard to the statistical approaches used, and the potential of artificial intelligence in this type of research raise questions that cannot be answered here. The following sections outline these limitations in more detail.

13.3.1 Limited item sample

In this thesis, all analyses are based on the characteristics of 94 items in four language versions which were developed for the purpose of this specific assessment. This means that not all combinations of relevant linguistic features, such as part-of-speech categories and lexical diversity, are represented systematically. Similarly, text genres, predicted response processes, NIs of varying length and most other item and text-by-item characteristics were not evenly distributed across the items²⁵⁴. As a result, the coverage of some, if not most variables is uneven, affecting the robustness of conclusions drawn about their effects on item difficulty.

In addition, the adaptation of the assessment into the various language versions was primarily informed by the needs of the assessment rather than any consideration of a study of item difficulty. Though an effort was made to keep the language versions as similar as possible, and major changes were always applied to all versions, when other considerations weighed more heavily than the comparability of the language versions, the former were preferred. As a result, not all differences between the language versions are systematic. This, too, limits the power of the analysis.

To reduce the problem of the low coverage of some variables, at least two strategies are possible: obviously, a larger dataset would have the potential to provide more evidence of some aspects of item difficulty. Additional item difficulty studies could focus on similar assessments and/or use the same operationalisations of assessment characteristics to reach some level of comparability. Since foreign language assessments are not often translated, multilingual analyses may prove to be elusive, though the ÜGK 2023 assessment (Erzinger, Angelone, et al. 2025), similar in design if not in target group, is a promising candidate. Another strategy is a more targeted assessment design: certain aspects of an assessment (or its translation) could be manipulated experimentally to investigate their effect on item difficulty. This is harder to put into practice, as it presupposes resources for the creation, piloting and administration of assessment items. In addition, though

²⁵⁴Wilson and colleagues describe this as “unplanned variation” (2008, pp. 89 sq.).

experimental manipulation of assessment items has been attempted in the past (e.g. Gorin 2005; Gorin and Embretson 2006), the results are not much better than in item difficulty studies based on conventional assessments. For instance, Gorin (2005) identified significant changes to item difficulty only for items which had negative wording and passive voice introduced to the reading passage of an English L1 reading assessment. Other theory-based manipulations of the assessment did not significantly influence item difficulty. She theorises that this may be due to an unusual target group or to limitations imposed by the choice of the assessment that was manipulated (*ibid.*, p. 368). Given the results from the present study, it could also be argued that changes to one single aspect may not affect all items in the same way given the multivariate nature of test item processing and reading comprehension (see also Brindley and Slatyer 2002, pp. 387 sqq.).

13.3.2 Variables

As has been amply discussed further above, it is somewhat unclear whether the linguistic characteristics extracted from the input texts can be considered valid across the three foreign languages. The use of language-specific POS tagging, limited options for correcting tagging errors, an only semi-automatic matching of content words across languages and the use of experimental operationalisations further added to the uncertainty inherent in most of the (cross-)linguistic variables used in this thesis. In addition, most of these variables are prone to large variations when extracted from very short texts. Hence, the linguistic characteristics of the NIs and item parts had to be eliminated entirely or their effect on item difficulty has to be considered with caution.

Caveats can also be formulated for some of the categorical variables used in this thesis: while the need for background knowledge and the falsifiability of the answer options was rated by several individuals and can be considered more robust, the operationalisation of response processes, text continuity and text functions had to be adapted to the nature of the assessment and could not be subjected to interrater reliability analyses. Though the insufficient distribution of the categories of most of these variables in this thesis renders the problem moot, it would be preferable to use well-established operationalisation for these aspects, along with a more targeted selection of items.

Finally, the selection of the variables, be they linguistic or more assessment-related in nature, is incomplete and to some extent biased. Syntactic characteristics of the input texts in particular remain largely unexplored. This is partly due to methodological constraints: syntactic features are more difficult to operationalise consistently across languages and are less easily captured through automated methods. Moreover, at the low proficiency level targeted in this study, syntactic complexity may play a smaller role in reading comprehension than lexical factors, as was discussed in chapter 5. Additional linguistic aspects which could be investigated more thoroughly include features related to cohesion and discourse, such as the use and type of connectives. Referential clarity, i.e. the distance between referring expressions and their antecedents, or the ease with which references can be resolved, but also morpheme complexity and variability may affect reading comprehension, and they may differ across the language versions, making them potentially interesting to investigate.

13.3.3 Young language learners

An important unknown aspect of this thesis is the role of the test takers' age and experience: item difficulty studies are usually conducted on existing assessments, and interpreted irrespective of the characteristics of the test takers. In the present case, this may be particularly problematic given the young age of the test takers (mean age: 12.5 years). Certain aspects of the assessment may not predict item difficulty as strongly as in other studies simply because the test takers are still developing their reasoning abilities (see for references e.g. Metallidou et al. 2012), making their comprehension of texts and their thinking about multiple-choice items difficult to predict. For example, it is possible that the (im)plausibility of incorrect answer options does not register

as strongly with this group of young test takers as it would with older adolescents or adults (cf. e.g. Koslowski and Masnick 2010, pp. 384 sq.). Also, items which require background knowledge may be easier to solve for some students who have the required knowledge, but not for others²⁵⁵. This, of course, is not easy to measure since the same students may have had difficulty forming an adequate understanding of the input texts in the first place (cf. e.g. R. Smith et al. 2021). Similar arguments can be put forward for other variables, including those which have clear relationships with item difficulty, e.g. the existence of cognates in the input text.

13.3.4 Sociocultural aspects

The somewhat technical nature of this study, which is based on the premise that item difficulty can be estimated and systematically predicted by characteristics of the assessment, has mostly excluded considerations related to social aspects of reading and assessment: neither the role of reading as a social practice nor possible sources of cultural or social bias in the reading assessment are reflected in the choice or operationalisation of the variables. In fact, the entire study rests on the assumption that characteristics of the assessment and test takers influence item difficulty in similar ways, irrespective of underlying social or cultural differences. This remains true even where test taker attributes were included in the modelling: students who have an immigration history were coded as such but not differentiated further, even though it is obviously recognised that these students do not form a homogeneous group. Similarly, students who use the target language at home may do so in very different ways and for very different reasons, but this was not included in the modelling (and, in fact, not known).

Finally, the study does not account for the multiple and shifting identities that learners bring to the assessment situation, including factors such as test anxiety, classroom experiences or the perceived stakes of the test, all of which may influence performance in ways that are not captured in psychometric modelling. Nor does it consider how the design and presentation of test items might interact with students' prior experiences, interests or cultural backgrounds. These omissions reflect a broader methodological limitation in quantitative approaches to item difficulty research: while they offer valuable insights into measurable patterns, they risk obscuring the individual, contextual and experiential dimensions of reading, as will be further discussed in the following section.

13.3.5 Methodological shortcomings

At several points in this thesis, the findings of the impressionistic analyses of items showing regional DIF were brought forward to lay out possible reasons for the low amount of explained variance in the quantitative analyses. This underlines the main weakness of the statistical approach: the multivariate and mutable nature of individual human beings solving reading items in a foreign language does not conform with the structured nature of the (inferential) statistics used in this thesis. It is appealing, of course, to envisage a mechanical explanation of how a test item is solved given certain attributes of the input text and items. Some very broad mechanisms, which are often at work and have a strong impact on item difficulty, can be confirmed in this manner. However, the results of the above analyses clearly indicate that this type of explanation will always lack precision: an item may be linked to a long text with few cognates and no detectable overlap between the correct answer option and the NI and still be easy to solve for test takers of one curricular region, perhaps because the answer is dependent on a lexical unit that students have seen a lot in their textbooks. In another case, an item may be predicted to be quite easy given its characteristics, but an incorrect answer option may be linked to an information that is particularly transparent in an input text to speakers of a certain language of schooling, which renders the item unexpectedly difficult. These interactions between attributes of the assessment,

²⁵⁵In addition, this variable was rated by adult raters who may not always have an adequate understanding of what children know and what they do not know.

or non-linear effects of certain variables cannot be modelled in LLTMs and only in a limited way in Bayesian LLTMs since these assume linearity and additivity of the predictors. This is without taking into account individual characteristics and the actual thought processes of the test takers, which may well be impossible to model in their entirety (see also Alderson 1990b, p. 477).

Response processes, clearly, are too complex to be modelled entirely *post ex facto* by regressing selected assessment and test taker attributes on item difficulty. This is true even if one were to assume that it is possible to operationalise these attributes and estimate item difficulty to the degree of precision required for this type of analysis *and* take into account the individual characteristics of the various test taker groups (e.g. younger test takers, test takers with larger multilingual repertoires). Other methodological approaches, both quantitative (e.g. eye-tracking) and qualitative (e.g. stimulated recall) are better suited to understand the cognitive processes of test item solving and their interactions with characteristics of the assessment, but also to identify and examine additional contextual and affective factors that may play a role. Qualitative studies of this kind are, of course, harder to implement and accordingly quite rare (see for an overview Barras 2021, pp. 52 sqq.). Nevertheless, this type of study is quite important, not least because it has the potential to cast serious doubt on the assumed predictability of item difficulty (e.g. Alderson 1990b).

13.3.6 Artificial intelligence

During the years this thesis was written, Artificial Intelligence (AI) technology made leaps in both its possibilities and its availability to the larger public. As of 2025, when this thesis was finished, it is not unlikely that parts of this study could be conducted more efficiently and perhaps with less room for error with the support of AI. This is particularly true for the multilingual POS tagging and the extraction of certain lexical variables. As of yet, research of this kind is still quite rare, and it often has a clear focus on the technical aspects of item difficulty prediction instead of a better understanding of the test construct (e.g. AlKhuyaey et al. 2024; Razavi and S. J. Powers 2025). Given the rapidity of this development, however, knowledge about these tools and experience with their use will probably increase rapidly. In the same vein, it would be interesting to investigate how AIs would predict item difficulty and what predictors they would choose.

13.4 Implications

Despite the uncertainty inherent in analysing the sources of item difficulty, the results contain several implications for assessment and future research. Most notably, they highlight the role of the multilingual repertoire in language assessment: the findings clearly show that in foreign language assessment a (closely) related language of schooling has a measurable effect on item difficulty, at least at the lower proficiency levels. This implies that the presence of shared vocabulary in a foreign language assessment influences the estimated proficiency of the test takers. If certain groups of learners encounter primarily items that are easier due to the presence of cognates or otherwise transparent language, their foreign language proficiency may be overestimated²⁵⁶. This raises the question of whether such features should be systematically controlled during item development, or at least monitored during test assembly.

²⁵⁶One could, of course, argue that intercomprehension occurs in real-life foreign language use and is, as such, part of the foreign language comprehension construct. After all, it is technically irrelevant whether, for example, a proficient user of German understands the English word “wind” because they speak German or because they learned it as an English word. In order to adequately measure knowledge of English, however, an assessment must be balanced: all test takers should have the *same* opportunities to draw on their multilingual repertoires. Otherwise, proficiency estimates may be biased towards those who, by chance, encounter more shared vocabulary in their assessment.

This consideration can be extended further: since foreign language competence is inevitably shaped by an individual's knowledge of other languages, it is unlikely that one can be measured independently of the others. Any assessment taken by learners with differing multilingual repertoires therefore carries the potential to measure, at least to some extent, not only the intended target language proficiency but also abilities deriving from other, possibly closely related languages. For example, international English assessments may be somewhat easier for proficient speakers of other Germanic (and, to a lesser extent, some Romance) languages than for speakers of most other language families. For the same reason, however, these learners may also find it easier to acquire English in the first place. If an assessment reflects this real-world advantage, rather than artificially amplifying or suppressing it, the integrity of the measurement does not necessarily seem compromised. The challenge, then, is to determine whether an assessment captures language ability as it occurs in every-day language use, or whether it inadvertently privileges certain linguistic backgrounds beyond what would occur outside the testing situation. More fundamentally, this suggests that we could consider whether we should conceptualise foreign language reading more as an inherently multilingual construct rather than a strictly language-specific ability. This applies not only to language testing, but also to SLA and psycholinguistic research: whenever we measure an individual's linguistic abilities, the results may reflect knowledge of multiple languages. Depending on the research question, the measurement instrument, and the population under study, this may substantially affect the conclusions that can be drawn.

In a similar vein, the findings related to text type appear particularly relevant (see sections 12.1.2 and 13.1.1). Clearly, there is no straightforward cause-and-effect relationship between the comprehensibility of a text and its individual characteristics. Rather, features such as the distribution of word classes (i.e. POSs), lexical diversity, sophistication, and difficulty interact in complex ways. The analyses presented in this thesis suggest that these interactions may align with text type and, consequently, with text length. This may help explain why the relatively simple measure of text length often accounts so well for item difficulty. While the abundance of text complexity measures in reading research shows that such observations are not new, a more detailed understanding of how these aspects interact may still prove beneficial for test development and, more broadly, for research that relies on text comprehension as an indicator of language competence.

Finally, a major implication of this thesis concerns multilingual assessment. It set out to investigate why the four language versions did not appear to be directly comparable, despite their close similarity. In many respects, the results suggest that the languages involved in the ÜGK assessment do, in fact, function in similar ways with regard to text comprehension and item difficulty. Across all language versions, text length and related indicators emerge as key predictors, alongside the presence of transparent cognates and measurable overlap between texts and items. The observations described in section 11.2 point in a similar direction: possible explanations for particularly easy or difficult items, or for marked deviations across language versions, are rarely language-specific. Instead, they can often be traced to the way a text must be processed to arrive at the correct answer, or to individual lexical elements that likely shape how test takers perceive the item. Thus, the underlying processes and mechanisms appear to be largely comparable, at least for these closely related European languages.

At the same time, the results underline the importance of individual lexical items, as reflected in the various lexical and cross-linguistic variables included in the analysis. Since these differ necessarily in the various languages and across the multilingual repertoires of the test takers, it is not surprising that individual items vary in difficulty, despite similar underlying mechanisms²⁵⁷. Multilingual assessment should therefore carefully consider whether results across languages can meaningfully be compared. The fact that test items are translations of one another is clearly not sufficient to warrant such comparisons. Instead, robust standard setting, carefully calibrated

²⁵⁷It would be interesting to explore whether this is particularly true for low-proficiency assessments, or whether more advanced tests (e.g. at B2 level or above) exhibit similar patterns.

items, and a detailed understanding of the factors that influence item difficulty across language combinations may be better suited to arrive at adequate judgments about comparability. Finally, methodologically, the results also highlight the value of combining different modelling and validation approaches, as no single method fully captures the complexity of reading assessment and item difficulty. Instead, if we want to better understand language competence, its acquisition and its use, both within and across languages, we will need to accept its complexity and account for it in the theories we formulate and the empirical methods we use.

Chapter 14

Conclusion

[By] doing testing, provided that we research what we design, we can contribute to a growing understanding of the construct.

Alderson 2000, p. 2

This thesis investigates the functioning of a low-proficiency foreign language reading assessment by relating characteristics of the input texts and assessment items, particularly lexical aspects, to empirical item difficulty. The assessment was administered in three foreign languages to learners in four curricular regions. As such, it is situated on the crossroads between monolingual item difficulty research and the often non-linguistic study of multilingual assessments and produces unique insights into the functioning of a foreign language reading assessment across languages. The study also includes a limited number of attributes of the test takers and examines both their predictive power for item difficulty and their potential interaction with assessment characteristics.

The findings underline the complexity of reading comprehension and its assessment: Though some very general effects can be shown to be strong enough to reach predictive power, e.g. certain long text types are related to more difficult items, and a higher proportion of cognates in the input text seems to be linked to easier items, a lot of variance remains unexplained. Impressionistic analyses of certain items with regional DIF suggest that this is linked to the multivariate nature of reading comprehension and test item processing. In some cases, a single expression is decisive for a correct answer. If it is a cognate, knowledge of the language of schooling comes into play. In other cases, elements from incorrect answer options can be found near the NI, which forces test takers to read more precisely, rendering items more difficult. In both cases, most of the other characteristics of the assessment have become irrelevant. On the level of the variables, similar confounding effects are at play: In this particular assessment, texts with a higher lexical sophistication are often related to items with lower item difficulty, but this counterintuitive finding probably has to do with the text genres used in the assessment rather than an actual facilitating effect of more sophisticated vocabulary in foreign language reading.

The inclusion of test taker attributes confirms the importance of both L1 reading competence and L2 knowledge for L2 reading comprehension and underlines the relatively higher importance of L2 knowledge at this low level of reading: Results from the L2 listening assessment contribute about twice as much to the prediction of the L2 reading responses as the L1 reading results.

What remains, then, is to keep researching, as pointed out by Alderson (*ibid.*, p. 2) in the above epigraph. This thesis contributes a small piece to our understanding of the L2 reading construct, if only by revealing how much we are still uncertain about. The analysis confirms that the languages of an individual interact in complex ways and that this influences foreign language reading scores. The study further underlines the complexity of operationalising characteristics of texts and response processes, particularly across several languages. Finally, it emphasises the

importance of individual words and lexical units in reading assessment items at lower levels of reading competence.

In many ways, these results are not surprising. Much of what has emerged will probably be obvious to experienced assessment developers and language teachers. Instinctively, then, we already know a great deal about L2 reading assessment, from the construction of items to the analysis and interpretation of test results. Yet there is a lot in the details that we mostly assume or at least cannot predict in its complexity. More research, then, is certainly needed, on all levels: We need to know more about the inner workings of large-scale assessments, about languages other than English, but we also need to better understand in detail how and why individuals solve test items and what this tells us about the construct we intend to measure. We will need to combine qualitative and quantitative research, and we will need to take into account findings from other fields of research, other types of assessment and in different languages. Yet, even then, we will probably have to remember that item difficulty is an estimation, and that assessment results cannot be explained fully. Too much is going on in the minds of test takers at any given moment that we cannot predict. This, of course, to finish with a phrase from *Star Trek*, as requested by a friend, is a good thing: the beauty of human cognition, including reading in a foreign language, is that it thrives on “infinite diversity in infinite combinations”²⁵⁸.

²⁵⁸Phrase from the fictional Vulcan philosophy in *Star Trek*, first introduced in the original series episode “Is There in Truth No Beauty?” (1968).

References

- Abbott, Marilyn L. (2007). “A confirmatory approach to Differential Item Functioning on an ESL reading assessment”. In: *Language Testing* 24.1, pp. 7–36.
- Abedi, Jamal and Carol Lord (2001). “The language factor in mathematics tests”. In: *Applied Measurement in Education* 14.3, pp. 219–234.
- Abel, Andrea et al. (2014). “A trilingual learner corpus illustrating European reference levels”. In: *Ricognizioni. Rivista di Lingue e Letterature e Culture moderne* 1.2, pp. 111–126.
- Abutalebi, Jubin (2008). “Neural aspects of second language representation and language control”. In: *Acta Psychologica* 128.3, pp. 466–478.
- Adams, Ray and Margaret Wu, eds. (2002). *PISA 2000 Technical Report*. URL: <https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-report/PISA%202000%20Technical%20report.pdf> (last accessed on 08/06/2025).
- Adams, Raymond J. (2005). “Reliability as a measurement design effect”. In: *Studies in Educational Evaluation* 31.2, pp. 162–172.
- Ağçam, Reyhan and M. Babanoğlu (2018). “A comparative study on EFL textbooks in Turkish and German secondary public schools”. In: *International Journal Of Eurasia Social Sciences* 9.32, pp. 948–959.
- Akamatsu, Nobuhiko (2003). “The effects of first language orthographic features on second language reading in text”. In: *Language Learning* 53.2, pp. 207–231.
- Alderman, Donald L. and Paul W. Holland (1981). *Item Performance Across Native Language Groups on the Test of English as a Foreign Language*. TOEFL Research Report RR-81-16. Princeton: ETS.
- Alderson, J. Charles (1984). “Reading in a foreign language: A reading problem or a language problem?” In: *Reading in a Foreign Language*. Ed. by J. Charles Alderson and A. H. Urquhart. London: Longman, pp. 1–24.
- (1990a). “Testing reading comprehension skills (part one)”. In: *Reading in a Foreign Language* 6.2, pp. 425–438.
- (1990b). “Testing reading comprehension skills (part two): Getting students to talk about taking a reading test (A pilot study).” In: *Reading in a Foreign Language* 7.1, pp. 465–503.
- (2000). *Assessing Reading*. Cambridge: Cambridge University Press.
- Alderson, J. Charles, Caroline Clapham, and Dianne Wall (1995). *Language Test Construction and Evaluation*. Cambridge, New York: Cambridge University Press.
- Alderson, J. Charles, Ari Huhta, and Lea Nieminen (2016). “Characteristics of weak and strong readers in a foreign language”. In: *The Modern Language Journal* 100.4, pp. 853–879.
- Alderson, J. Charles and A. H. Urquhart (1984). “Postscript on Alderson”. In: *Reading in a Foreign Language*. Ed. by J. Charles Alderson and A. H. Urquhart. London: Longman, pp. 25–27.
- Ali, Usama S. and Michael E. Walker (2014). *Enhancing the Equating of Item Difficulty Metrics: Estimation of Reference Distribution*. ETS Research Report RR-14-07. Princeton: ETS.
- AlKhuyaey, Samah et al. (2024). “Text-based question difficulty prediction: A systematic review of automatic approaches”. In: *International Journal of Artificial Intelligence in Education* 34.3, pp. 862–914.

- Allaire, JJ et al. (2023). *rmarkdown: Dynamic Documents for R*. <https://github.com/rstudio/rmarkdown>. URL: <https://github.com/rstudio/rmarkdown> (last accessed on 04/16/2023).
- Allalouf, Avi, Ronald K. Hambleton, and Stephen G. Sireci (1999). "Identifying the causes of DIF in translated verbal items". In: *Journal of Educational Measurement* 36.3, pp. 185–198.
- Allauzen, Alexandre and Helene Bonneau-Maynard (2008). "Training and evaluation of POS taggers on the French MULTITAG corpus". In: *Proceedings of the International Conference on Language Resources and Evaluation*. Marrakech.
- Allen, David (2019a). "Cognate frequency and assessment of second language lexical knowledge". In: *International Journal of Bilingualism* 23.5, pp. 1121–1136.
- (2019b). "The prevalence and frequency of Japanese-English cognates: Recommendations for future research in applied linguistics". In: *International Review of Applied Linguistics in Language Teaching* 57.3, pp. 355–376.
- Almeida, Daniel Machado de and Sandra Maria Aluísio (2009). "Text readability analysis with Natural Language Processing tools: The adaptation of Coh-Metrix metrics for Portuguese". In: *2009 Seventh Brazilian Symposium in Information and Human Language Technology*, pp. 53–62.
- Altmayer, Claus (2002). "Kulturelle Deutungsmuster in Texten. Prinzipien und Verfahren einer kulturwissenschaftlichen Textanalyse im Fach Deutsch als Fremdsprache". In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 6.3.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington D.C.: AERA Publications.
- Anderson, John R. and Gordon H. Bower (1980). *Human Associative Memory: A Brief Edition*. Hillsdale: Lawrence Erlbaum.
- Anderson, Richard C. (1972). "How to construct achievement tests to assess comprehension". In: *Review of Educational Research* 42.2, pp. 145–170.
- Andrich, David and Ida Marais (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Singapore: Springer.
- Angelone, Domenico and Florian Keller (2019a). *Überprüfung des Erreichens der Grundkompetenzen (ÜGK) im Fach Mathematik im 11. Schuljahr. Technische Dokumentation zur Testentwicklung und Skalierung*. URL: https://www.vecof-svizzera.ch/wp-content/uploads/2019/06/%C3%9CGK2017_Technischer-Bericht_ADB.pdf (last accessed on 08/06/2025).
- (2019b). *Überprüfung des Erreichens der Grundkompetenzen (ÜGK) in den Fächern Schulsprache und erste Fremdsprache im 8. Schuljahr. Technische Dokumentation zur Testentwicklung und Skalierung*. URL: https://www.vecof-svizzera.ch/wp-content/uploads/2019/06/%C3%9CGK2017_Technischer-Bericht_ADB.pdf (last accessed on 08/06/2025).
- Aryadoust, Vahid (2012). "Using Cognitive Diagnostic Assessment to model the underlying structure of a listening test: a sub-skill-based approach". In: *Asian EFL Journal* 14.4, pp. 81–106.
- (2013). "Predicting item difficulty in a language test with an adaptive neuro fuzzy inference system". In: *2013 IEEE Workshop on Hybrid Intelligent Models and Applications (HIMA)*. Singapore: IEEE, pp. 43–50.
- (2015). "Fitting a mixture Rasch model to English as a foreign language listening tests: The role of cognitive and background variables in explaining latent Differential Item Functioning". In: *International Journal of Testing* 15.3, pp. 216–238.
- (2018). "Using recursive partitioning Rasch trees to investigate Differential Item Functioning in second language reading tests". In: *Studies in Educational Evaluation* 56, pp. 197–204.
- (2021). "A Cognitive Diagnostic Assessment study of the listening test of the Singapore-Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM". In: *International Journal of Listening* 35.1, pp. 29–52.

- Aryadoust, Vahid and Purya Baghaei (2016). “Does EFL readers’ lexical and grammatical knowledge predict their reading ability? Insights from a perceptron artificial neural network study”. In: *Educational Assessment* 21.2, pp. 135–156.
- Aryadoust, Vahid and Christine Goh (2019). “Classification and regression trees in predicting listening item difficulty”. In: *Quantitative Data Analysis for Language Assessment Volume II: Advanced Methods*. Ed. by Vahid Aryadoust and Michelle Raquel. London, New York: Routledge, pp. 193–214.
- Aryadoust, Vahid and Christine C. M. Goh (2014). *Predicting Listening Item Difficulty with Language Complexity Measures: A Comparative Data Mining Study*. CaMLA Working Papers 2014-2. Ann Arbor: CaMLA.
- Ashton, Karen (2014). “Using self-assessment to compare learners’ reading proficiency in a multilingual assessment framework”. In: *System* 42, pp. 105–119.
- Ashton, Karen, Neil Jones, and Tamsin Walker (2010). “Asset languages: A case study of piloting the CEFR Manual”. In: *Aligning Tests With the CEFR: Reflections on Using the Council of Europe’s Draft Manual*. Ed. by Waldemar Martyniuk. Cambridge: Cambridge University Press, pp. 227–246.
- Association of Language Testers in Europe (ALTE), ed. (1998). *Multilingual Glossary of Language Testing Terms*. Cambridge: Cambridge University Press.
- Audacity Team (2014). *Audacity: Free Audio Editor and Recorder*. URL: <http://audacity.sourceforge.net/?lang=de> (last accessed on 06/08/2025).
- Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers (1995). *CELEX2*. URL: <https://catalog.ldc.upenn.edu/LDC96L14> (last accessed on 04/04/2023).
- Bachman, Lyle F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- (2002). “Some reflections on task-based language performance assessment”. In: *Language Testing* 19.4, pp. 453–476.
- Bachman, Lyle F., Fred Davidson, and Michael Milanovic (1996). “The use of test method characteristics in the content analysis and design of EFL proficiency tests”. In: *Language Testing* 13.2, pp. 125–150.
- Bachman, Lyle F. and Adrian S. Palmer (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford, New York: Oxford University Press.
- (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford, New York: Oxford University Press.
- Baker, Colin (2001). *Foundations of Bilingual Education and Bilingualism*. 3rd ed. Clevedon: Multilingual Matters.
- Baker, Frank B. and Seock-Ho Kim (2017). *The Basics of Item Response Theory Using R*. Cham: Springer International.
- Barras, Malgorzata (2018). “‘Stimulated Recall’ in der Sprachtestforschung. Ein praktisches Beispiel aus der Erprobung eines computerbasierten Leseverstehenstests”. In: *Lautes Denken, "Stimulated Recall" und Dokumentarische Methode. Rekonstruktive Verfahren in der Fremdsprachenlehr- und -lehrforschung*. Ed. by Karin Aguado, Claudia Finkbeiner, and Bernd Tesch. Frankfurt: Peter Lang, pp. 69–86.
- (2021). “Strategien beim Lösen rezeptiver Sprachtestaufgaben”. Doctoral thesis. University of Fribourg/Switzerland.
- Bates, Douglas et al. (2015). “Fitting linear mixed-effects models using lme4”. In: *Journal of Statistical Software* 67, pp. 1–48.
- Beacco, Jean-Claude et al. (2008). *Niveau A2 pour le français - Un Référentiel*. Paris: Didier.
- Beck, Isabel L., Margaret G. McKeown, Richard C. Omanson, et al. (1984). “Improving the comprehensibility of stories: The effects of revisions that improve coherence”. In: *Reading Research Quarterly* 19.3, pp. 263–277.

- Beck, Isabel L., Margaret G. McKeown, Gale M. Sinatra, et al. (1991). “Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility”. In: *Reading Research Quarterly* 26.3, pp. 251–276.
- Behr, Dorothée (2017). “Assessing the use of back translation: the shortcomings of back translation as a quality testing method”. In: *International Journal of Social Research Methodology* 20.6, pp. 573–584.
- Beijering, Karin, Charlotte Gooskens, and Wilbert Heeringa (2008). “Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm”. In: *Linguistics in the Netherlands* 25.1, pp. 13–24.
- Beinborn, Lisa, Torsten Zesch, and Iryna Gurevych (2014). “Predicting the difficulty of language proficiency tests”. In: *Transactions of the Association for Computational Linguistics* 2, pp. 517–530.
- (2015). “Candidate evaluation strategies for improved difficulty prediction of language tests”. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, pp. 1–11.
- Bengoetxea, Kepa and Itziar Gonzalez-Dios (2021). *MultiAzterTest: a multilingual analyzer on multiple levels of language for readability assessment*. URL: <https://arxiv.org/abs/2109.04870> (last accessed on 08/06/2025).
- Bengoetxea, Kepa, Itziar González-Dios, and Amaia Aguirregoitia (2020). “AzterTest: Open source linguistic and stylistic analysis tool”. In: *Procesamiento del Lenguaje Natural* 64, pp. 61–68.
- Berger, Bonnie, Michael S. Waterman, and Yun William Yu (2021). “Levenshtein distance, sequence comparison and biological database search”. In: *IEEE Transactions on Information Theory* 67.6, pp. 3287–3294.
- Bergqvist, Ewa, Frithjof Theens, and Magnus Österholm (2018). “The role of linguistic features when reading and solving mathematics tasks in different languages”. In: *The Journal of Mathematical Behavior* 51, pp. 41–55.
- Bersinger, Simone et al. (2005). *Europäisches Sprachenportfolio ESP II: Version für Kinder und Jugendliche von 11 bis 15 Jahren*. Bern: Schulverlag plus.
- Berthele, Raphael, Peter Lenz, and Elisabeth Peyer (2022). “Predicting foreign language skills based on first languages: The role of lexical distance and relative morphological complexity”. In: *Poznan Studies in Contemporary Linguistics* 58.3, pp. 419–448.
- Bertschy, Ida, Mirjam Egli Cuenat, and Daniel Stotz (2015). *Lehrplan Französisch und Englisch*. URL: https://be.lehrplan.ch/passepartout/Lehrplan_Passepartout.pdf (last accessed on 09/21/2019).
- Bezooijen, Renée van and Charlotte Gooskens (2007). “Interlingual text comprehension: linguistic and extralinguistic determinants”. In: *Receptive Multilingualism and Intercultural Communication: Linguistic Analyses, Language Policies and Didactic Concepts*. Ed. by Jan D. ten Thije and Ludger Zeevaert. Amsterdam: John Benjamins, pp. 249–264.
- Biber, Douglas (1995). *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.
- Biesta, Gert (2009). “Good education in an age of measurement: On the need to reconnect with the question of purpose in education”. In: *Educational Assessment, Evaluation and Accountability* 21.1, pp. 33–46.
- Bogaards, Paul (2001). “Lexical units and the learning of foreign language vocabulary”. In: *Studies in Second Language Acquisition* 23.3, pp. 321–343.
- Böhme, Katrin (2011). “Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen. Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich”. Doctoral thesis. Humboldt-Universität zu Berlin.
- Böhme, Katrin, Birgit Heppt, and Nicole Haag (2017). “Inclusive literacy education and reading assessment for language-minority students and students with special educational needs in

- German elementary schools". In: *Inclusive Principles and Practices in Literacy Education*. Ed. by Marion Milton. Emerald Publishing, pp. 69–86.
- Böhme, Katrin and Alexander Robitzsch (2009). "Methodische Aspekte der Erfassung der Lesekompetenz". In: *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule*. Ed. by Dietlinde Granzer et al. Weinheim: Beltz, pp. 250–289.
- Böhme, Katrin, Alexander Robitzsch, and Anne-Kathrin Busè (2010). "Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben". In: *Zuhörkompetenz in Unterricht und Schule: Beiträge aus Wissenschaft und Praxis*. Ed. by Volker Bernius and Margarete Imhof. Göttingen: Vandenhoeck & Ruprecht, pp. 81–104.
- Bolaños-Medina, Alicia and Víctor González-Ruiz (2012). "Deconstructing the translation of psychological tests". In: *Meta : journal des traducteurs / Meta: Translators' Journal* 57.3, pp. 715–739.
- Bordag, Denisa, Kira Gor, and Andreas Opitz (2022). "Ontogenesis Model of the L2 Lexical Representation". In: *Bilingualism: Language and Cognition* 25.2, pp. 185–201.
- Bot, Kees de (2004). "The multilingual lexicon: Modelling selection and control". In: *International Journal of Multilingualism* 1.1, pp. 17–32.
- Brants, Thorsten and Alex Franz (2006). *Web 1T 5-gram Version 1*. URL: <https://catalog.ldc.upenn.edu/LDC2006T13> (last accessed on 02/21/2023).
- Brindley, Geoff and Helen Slatyer (2002). "Exploring task difficulty in ESL listening assessment". In: *Language Testing* 19.4, pp. 369–394.
- Brislin, Richard W. (1970). "Back-translation for cross-cultural research". In: *Journal of Cross-Cultural Psychology* 1.3, pp. 185–216.
- Brizuela, Armel and Eiliana Montero Rojas (2013). "Prediction of the difficulty level in a standardized reading comprehension test: contributions from cognitive psychology and psychometrics". In: *RELIEVE - E-Journal of Educational Research, Assessment and Evaluation* 19.2.
- Brown, Annie and Noriko Iwashita (1998). "The role of language background in the validation of a computer-adaptive test". In: *Validation in Language Assessment*. Mahwah: Lawrence Erlbaum, pp. 195–207.
- Brunfaut, Tineke and Andrea Révész (2015). "The role of task and listener characteristics in second language listening". In: *TESOL Quarterly* 49.1, pp. 141–168.
- Brysbaert, Marc, Matthias Buchmeier, et al. (2011). "The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German". In: *Experimental Psychology* 58.5, pp. 412–424.
- Brysbaert, Marc and Boris New (2009). "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English". In: *Behavior Research Methods* 41.4, pp. 977–990.
- Buck, Gary (1991). "The testing of listening comprehension: an introspective study". In: *Language Testing* 8.1, pp. 67–91.
- Buck, Gary and Kikumi Tatsuoka (1998). "Application of the rule-space procedure to language testing: Examining attributes of a free response listening test". In: *Language Testing* 15.2, pp. 119–157.
- Buck, Gary, Kumi Tatsuoka, and Irene Kostin (1997). "The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension". In: *Language Learning* 47.3, pp. 423–466.
- Bulté, Bram and Alex Housen (2012). "Defining and operationalising L2 complexity". In: *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*. Ed. by Alex Housen, Folkert Kuiken, and Ineke Vedder. Amsterdam: John Benjamins, pp. 21–46.
- Bürkner, Paul-Christian (2021). "Bayesian Item Response Modeling in R with brms and Stan". In: *Journal of Statistical Software* 100.5.

- Carpenter, Patricia Ann and Marcel Adam Just (1975). "Sentence comprehension: A psycholinguistic processing model of verification". In: *Psychological Review* 82.1, pp. 45–73.
- Carr, Nathan T. (2006). "The factor structure of test task characteristics and examinee performance". In: *Language Testing* 23.3, pp. 269–289.
- (n.d.). *Rating Instrument for Judging the Text Characteristics of and Reader Attributes Engaged by a Test of Reading: Master Version*.
- Carrell, Patricia L. (1983). "Some issues in studying the role of schemata, or background knowledge, in second language comprehension". In: *Reading in a Foreign Language* 1.2, pp. 81–92.
- Cartwright, Kelly B. (2009). "The role of cognitive flexibility in reading comprehension: Past, present, and future". In: *Handbook of Research on Reading Comprehension*. Ed. by Susan E. Israel and Gerald G. Duffy. New York: Routledge, pp. 115–139.
- Carver, Ronald P. (1992). "Reading rate: Theory, research, and practical implications". In: *Journal of Reading* 36.2, pp. 84–95.
- Catts, Hugh W. (2018). "The Simple View of Reading: advancements and false impressions". In: *Remedial and Special Education* 39.5, pp. 317–323.
- Cenoz, Jasone (2013). "Defining multilingualism". In: *Annual Review of Applied Linguistics* 33, pp. 3–18.
- CENTAL (2021). *CEFRLex. Lexical Resources For Foreign Language (L2) Learning, Teaching and Research Graded Following the Common European Framework of Reference for Languages (CEFR)*. URL: <https://cental.uclouvain.be/cefrlex/> (last accessed on 06/03/2023).
- Chaniaud, Noémie, Camille Sagnier, and Emilie Loup-Escande (2022). "Translation and validation study of the French version of the eHealth literacy scale: Web-based survey on a student population". In: *JMIR Formative Research* 6.8.
- Chapelle, Carol A. (2012). "Conceptions of validity". In: *The Routledge Handbook of Language Testing*. Ed. by Glenn Fulcher and Fred Davidson. London, New York: Routledge, pp. 22–33.
- (2021). *Argument-Based Validation in Testing and Assessment*. Thousand Oaks: SAGE Publications.
- Charrad, Malika et al. (2014). "NbClust: An R Package for determining the relevant number of clusters in a data set". In: *Journal of Statistical Software* 61, pp. 1–36.
- Chen, Lang, Matthew A. Lambon Ralph, and Timothy T. Rogers (2017). "A unified model of human semantic knowledge and its disorders". In: *Nature Human Behaviour* 1.3, pp. 1–10.
- Chen, Xiaobin and Detmar Meurers (2016). "CTAP: A web-based tool supporting automatic complexity analysis". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 113–119.
- (2018). "Word frequency and readability: Predicting the text-level readability with a lexical-level attribute". In: *Journal of Research in Reading* 41.3, pp. 486–510.
- Choi, Inn-Chull and Youngsun Moon (2019). "Predicting the difficulty of EFL tests based on corpus linguistic features and expert judgment". In: *Language Assessment Quarterly* 17.1, pp. 18–42.
- Clahsen, Harald and Claudia Felser (2018). "Some notes on the Shallow Structure Hypothesis". In: *Studies in Second Language Acquisition* 40.3, pp. 693–706.
- Clark, Herbert H. (1976). *Semantics and Comprehension*. The Hague: Mouton.
- Clark, John L. D. (1965). "Item-by-item parallel achievement tests for different foreign languages". In: *Journal of Educational Measurement* 2.1, pp. 77–83.
- Clarke, Mark A. (1980). "The Short Circuit Hypothesis of ESL reading - Or when language competence interferes with reading performance". In: *The Modern Language Journal* 64.2, pp. 203–209.
- Clauser, Brian E. and Kathleen M. Mazor (1998). "Using statistical procedures to identify differentially functioning test items". In: *Educational Measurement: Issues and Practice* 17.1, pp. 31–44.

- Cobb, Tom (2023). *VocabProfilers*. URL: <https://www.lex tutor . ca / vp /> (last accessed on 02/16/2023).
- Coh-Metrix version 3.0 indices* (2023). URL: http://cohmetrix.memphis.edu/cohmetrixhome/documentation_indices.html (last accessed on 04/03/2023).
- Cohen, Yoav, Naomi Gafni, and Pnina Hanani (2007). “Translating and adapting a test, yet another source of variance; the Standard Error of Translation”. In: *The Annual Meeting of the IAEA*. Baku.
- Conférence intercantonale de l’instruction publique de la Suisse romande et du Tessin, ed. (2010). *Plan d’études romand (PER)*. URL: <https://portail.ciip.ch/per/domains> (last accessed on 08/06/2025).
- Council of Europe, ed. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- ed. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe.
- Crocker, Linda and James Algina (2008). *Introduction to Classical and Modern Test Theory*. Ohio: Cengage Learning.
- Cronbach, Lee J. (1971). “Test validation”. In: *Educational Measurement*. Ed. by Robert L. Thorndike. 2nd ed. Washington D.C.: American Council on Education, pp. 443–507.
- Cruse, D. A. (1986). *Lexical semantics*. Cambridge, New York: Cambridge University Press.
- Cui, Yue et al. (2022). “CTAP for Chinese: A linguistic complexity feature automatic calculation platform”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 5525–5538.
- D-EDK (2016). *Lehrplan 21. Sprachen*. URL: <https://v-fe.lehrplan.ch/index.php?code=b%7C1%7C0&1a=yes> (last accessed on 08/05/2025).
- Damerau, Fred J. (1964). “A technique for computer detection and correction of spelling errors”. In: *Communications of the ACM* 7.3, pp. 171–176.
- Daulton, Frank E. (2008). *Japan’s Built-in Lexicon of English-based Loanwords*. Clevedon: Multilingual Matters.
- Davey, Beth (1988). “Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers”. In: *The Journal of Experimental Education* 56.2, pp. 67–76.
- Davidson, Douglas J., Peter Indefrey, and Marianne Gullberg (2008). “Words that second language learners are likely to hear, read, and use”. In: *Bilingualism: Language and Cognition* 11.1, pp. 133–146.
- Davies, Matthias von, Lou DiBello, and Kentaro Yamamoto (2006). *Reporting Test Outcomes Using Models for Cognitive Diagnosis*. ETS Research Report RR-06-28. Princeton: ETS.
- Davis, Andréa K., Natalie Bowman, and Margarita Kaushanskaya (2018). “Bilingualism, cognates and reading fluency in children”. In: *Journal of Research in Reading* 41.S1, S12–S29.
- Davison, Alice and Robert N. Kantor (1982). “On the failure of readability formulas to define readable texts: A case study from adaptations”. In: *Reading Research Quarterly* 17.2, pp. 187–209.
- De Boeck, Paul and Mark Wilson, eds. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.
- De Groot, Annette M. B. and Rineke Keijzer (2000). “What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting”. In: *Language Learning* 50.1, pp. 1–56.
- De Wilde, Vanessa, Marc Brysbaert, and June Eyckmans (2020). “Learning English through out-of-school exposure: How do word-related variables and proficiency influence receptive vocabulary learning?” In: *Language Learning* 70.2, pp. 349–381.
- DECS (2004). *Piano di formazione della scuola media*. URL: https://www4.ti.ch/fileadmin/DECS/DS/SIM/Cosa_facciamo/PF_SM.pdf (last accessed on 08/06/2025).

- Dellert, Johannes (2018). “Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe: Association for Computational Linguistics, pp. 3123–3133.
- DESI-Konsortium, ed. (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie*. Weinheim et al.: Beltz.
- Dhar, Prajit and Arianna Bisazza (2021). *Understanding cross-lingual syntactic transfer in multilingual recurrent neural networks*. URL: <http://arxiv.org/abs/2003.14056> (last accessed on 06/14/2022).
- Dijk, Teun A. van and Walter Kintsch (1983). *Strategies of Discourse Comprehension*. New York, London: Academic Press.
- DIPF and Nagarro (2018a). *CBA Execution Environment*. URL: <https://tba.dipf.de/de/infrastruktur/softwareentwicklung/cba-execution-environment> (last accessed on 04/10/2023).
- (2018b). *CBA ItemBuilder*. URL: <https://tba.dipf.de/de/infrastruktur/softwareentwicklung/cba-itembuilder> (last accessed on 04/10/2023).
- Dorans, Neil J. and Paul W. Holland (1992). *DIF Detection and Description: Mantel Haenszel and Standardization*. ETS Research Report RR-92-10. Princeton: ETS.
- Drum, Priscilla A., Robert C. Calfee, and Linda K. Cook (1981). “The effects of surface structure variables on performance in reading comprehension tests”. In: *Reading Research Quarterly* 16.4, pp. 486–514.
- Dubay, William (2004). *The Principles of Readability*. Costa Mesa: Impact Information.
- Duffy, Thomas M. et al. (1989). “Models for the design of instructional text”. In: *Reading Research Quarterly* 24.4, pp. 434–457.
- Dürlich, Luise and Thomas François (2018). “EFLLex: A graded lexical resource for learners of English as a Foreign Language”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA).
- EDK (2011). *Grundkompetenzen für die Fremdsprachen. Nationale Bildungsstandards*. URL: http://edudoc.ch/record/96780/files/grundkomp_fremdsprachen_d.pdf (last accessed on 08/06/2025).
- (2018). *Kantonsumfrage 2016/2017. Grundlegende Informationen zu den kantonalen Bildungssystemen: Strukturdaten*. URL: <https://edudoc.ch/record/133445?ln=de> (last accessed on 10/23/2022).
- ElAtia, Samira (2013). “Test adaptation and translation: The language dimension”. In: *Translation in Language Teaching and Assessment*. Ed. by Dina Tsagari and Georgios Floros. Newcastle upon Tyne: Cambridge Scholars, pp. 193–213.
- Elder, Catherine (1996). “The effect of language background on ‘foreign’ language test performance: the case of Chinese, Italian, and modern Greek”. In: *Language Learning* 46.2, pp. 233–282.
- Elosua, Paula and Alicia López-Jauregui (2007). “Potential sources of Differential Item Functioning in the adaptation of tests”. In: *International Journal of Testing* 7.1, pp. 39–52.
- Elosua, Paula and Alicia López-Jauregui (2008). “Equating between linguistically different tests: Consequences for assessment”. In: *The Journal of Experimental Education* 76.4, pp. 387–402.
- Embretson, Susan E. and C. Douglas Wetzel (1987). “Component latent trait models for paragraph comprehension tests”. In: *Applied Psychological Measurement* 11.2, pp. 175–193.
- Erbeli, Florina and R. Malatesha Joshi (2022). “Simple View of Reading among Slovenian English foreign language learners: A latent interaction modeling approach”. In: *Learning and Individual Differences* 93, p. 101958.
- Ercikan, Kadriye (1998). “Translation effects in international assessments”. In: *International Journal of Educational Research* 29.6, pp. 543–553.

- Ercikan, Kadriye (2002). “Disentangling sources of Differential Item Functioning in multilanguage assessments”. In: *International Journal of Testing* 2.3-4, pp. 199–215.
- Ercikan, Kadriye et al. (2010). “Application of Think Aloud protocols for examining and confirming sources of Differential Item Functioning identified by expert reviews”. In: *Educational Measurement: Issues and Practice* 29.2, pp. 24–35.
- Erzinger, Andrea B., Domenico Angelone, et al., eds. (2025). *Nationaler Bericht zu der Überprüfung des Erreichens der Grundkompetenzen (ÜGK) 2023, Sprachen 11. Schuljahr: ein Beitrag zum Schweizer Bildungsmonitoring*. URL: <https://boris-portal.unibe.ch/handle/20.500.12422/204751> (last accessed on 06/18/2025).
- Erzinger, Andrea B., Manuela Hauser, et al. (2019). *Erläuterungen zu den Skalen des Kontextfragebogens der ÜGK Sprachen 2017: Theoretischer Hintergrund, Inhalte und Konstrukte*. URL: https://www.uegk-schweiz.ch/wp-content/uploads/2019/09/%C3%9CGK2017_Sprachen_KFB_def.pdf (last accessed on 08/06/2025).
- European Commission: Directorate-General for Education, Youth, Sport and Culture (2012). *First European Survey on Language Competences: Final Report*. Luxemburg: Publications Office.
- Eurydice, ed. (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. URL: <https://data.europa.eu/doi/10.2797/18294> (last accessed on 07/13/2022).
- Fan, Jason and Trevor Bond (2019). “Applying Rasch measurement in language assessment: Unidimensionality and local independence”. In: *Quantitative Data Analysis for Language Assessment*. Ed. by Vahid Aryadoust and Michelle Raquel. Vol. 1. London, New York: Routledge, pp. 83–102.
- Fellbaum, Christiane (2005). “WordNet and wordnets”. In: *Encyclopedia of Language and Linguistics*. Ed. by Keith Brown. 2nd ed. Oxford: Elsevier, pp. 665–670.
- Ferrara, Steve, Jeffrey T. Steedle, and Roger S. Frantz (2022). “Response demands of reading comprehension test items: A review of item difficulty modeling studies”. In: *Applied Measurement in Education* 35.3, pp. 237–253.
- Field, John (2013). “Cognitive validity”. In: *Examining Listening: Research and Practice in Assessing Second Language Listening*. Ed. by Ardeshir Geranpayeh and Lynda Taylor. Cambridge: Cambridge University Press, pp. 77–151.
- Finn, Bridgid (2015). *Measuring Motivation in Low-Stakes Assessments*. ETS Research Report RR-15-19. Princeton: Educational Testing Service.
- Fischer, Gerhard H. (1973). “The linear logistic test model as an instrument in educational research”. In: *Acta Psychologica* 37.6, pp. 359–374.
- Foschi Albert, Marina and Marianne Hepp (2014). “Drei Passwörter für das einsetzende fremdsprachige Leseverstehen: Textwissen, Textmuster, transkulturelle Textsortenzüge”. In: *Glottodidaktik: früher, heute und morgen*. Ed. by Luiza Ciepielewska-Kaczmarek and Gabriela Gorąca-Sawczyk. Poznań: Wydawnictwo Naukowe UAM, pp. 47–62.
- François, Thomas and Cédric Fairon (2012). “An ‘AI readability’ formula for French as a Foreign Language”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island: Association for Computational Linguistics, pp. 466–477.
- François, Thomas, Nùria Gala, et al. (2014). “FLELex: a graded lexical resource for French foreign learners”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. Reykjavik: European Language Resources Association (ELRA), pp. 3766–3773.
- Freedle, Roy and Christiane Fellbaum (1987). “An exploratory study of the relative difficulty of TOEFL’s listening comprehension items”. In: *Cognitive and Linguistic Analyses of Test Performance*. Ed. by Roy O. Freedle and Richard P. Durán. Norwood: Ablex, pp. 162–192.

- Freedle, Roy and Irene Kostin (1991). *The Prediction of SAT Reading Comprehension Item Difficulty for Expository Prose Passages*. ETS Research Report RR-91-29. Princeton: Educational Testing Service.
- (1993). “The prediction of TOEFL reading item difficulty: implications for construct validity”. In: *Language Testing* 10.2, pp. 133–170.
- (1994). “Can multiple-choice reading tests be construct-valid? A reply to Katz, Lautenschlager, Blackburn, and Harris”. In: *Psychological Science* 5.2, pp. 107–110.
- (1996). *The Prediction of TOEFL Listening Comprehension Item Difficulty for Minitalk Passages: Implications for Construct Validity*. TOEFL Research Report RR-96-29. Princeton: ETS.
- (1997). “Predicting black and white Differential Item Functioning in verbal analogy performance”. In: *Intelligence* 24.3, pp. 417–444.
- (1999). “Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL’s minitalks”. In: *Language Testing* 16.1, pp. 2–32.
- Fukushima, Kazuhiko (2015). “Compositionality, lexical integrity, and agglutinative morphology”. In: *Language Sciences* 51, pp. 67–85.
- Fulcher, Glenn and Fred Davidson (2007). *Language Testing and Assessment: An Advanced Resource Book*. Routledge Applied Linguistics. London, New York: Routledge.
- (2012). “Introduction”. In: *The Routledge Handbook of Language Testing*. Ed. by Glenn Fulcher and Fred Davidson. London, New York: Routledge, pp. 1–17.
- Galili, Tal and Alan O’Callaghan (2024). *Package: heatmaply*. URL: <https://talgalili.github.io/heatmaply/> (last accessed on 06/27/2024).
- Gao, Lingyun and W. Todd Rogers (2011). “Use of tree-based regression in the analyses of L2 reading test items”. In: *Language Testing* 28.1, pp. 77–104.
- Garcia, Fernando et al. (2018). “Validation of the five-factor self-concept questionnaire AF5 in Brazil: Testing factor structure and measurement invariance across language (Brazilian and Spanish), gender, and age”. In: *Frontiers in Psychology* 9.
- García Duncan, Teresa et al. (2005). “Study of a dual-language test booklet in eighth-grade mathematics”. In: *Applied Measurement in Education* 18.2, pp. 129–161.
- Gernsbacher, Morton Ann (1990). *Language Comprehension As Structure Building*. Hillsdale: Lawrence Erlbaum.
- Gernsbacher, Morton Ann and Julie A. Foertsch (1999). “Three models of discourse comprehension”. In: *Language Processing*. Ed. by Simon Garrod and Martin J. Pickering. New York: Psychology Press, p. 283.
- Geva, Esther and Linda S. Siegel (2000). “Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages”. In: *Reading and Writing* 12, pp. 1–30.
- Gibson, Edward (2000). “The dependency locality theory: A distance-based theory of linguistic complexity”. In: *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*. Ed. by Alec Marantz, Yasushi Miyashita, and Wayne O’Neil. Cambridge: The MIT Press, pp. 94–126.
- Gierl, Mark J. and Shameem Nyla Khaliq (2001). “Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis”. In: *Journal of Educational Measurement* 38, pp. 164–187.
- Giesbrecht, Eugenie and Stefan Evert (2009). “Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German Web as Corpus”. In: *Proceedings of the Fifth Web as Corpus Workshop (WAC5)*. Ed. by Iñaki Alegria, Igor Leturia, and Serge Sharoff. San Sebastián: Elhuyar Fundazioa, pp. 27–35.
- Glaboniat, Manuela et al. (2005). *Profile deutsch. Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel. Niveau A1-C2*. Neue Fassung. Berlin: Langenscheidt.
- Golden, Joanne, Beth Haslett, and Helen Gauntt (1988). “Structure and content in eighth-graders’ summary essays”. In: *Discourse Processes* 11.2, pp. 139–162.

- Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff (2012). “Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 759–765.
- Gombert, Jean Émile (1992). *Metalinguistic Development*. Chicago: University of Chicago Press.
- Gorin, Joanna S. (2005). “Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation”. In: *Journal of Educational Measurement* 42.4, pp. 351–373.
- Gorin, Joanna S. and Susan E. Embretson (2006). “Item difficulty modeling of paragraph comprehension items”. In: *Applied Psychological Measurement* 30.5, pp. 394–411.
- Grabe, William (2009). *Reading in a Second Language. Moving From Theory to Practice*. New York: Cambridge University Press.
- Graesser, Arthur C. and Sidney D’Mello (2012). “Moment-to-moment emotions during reading”. In: *The Reading Teacher* 66.3, pp. 238–242.
- Graesser, Arthur C., Danielle S. McNamara, and Jonna M. Kulikowich (2011). “Coh-Metrix: Providing multilevel analyses of text characteristics”. In: *Educational Researcher* 40.5, pp. 223–234.
- Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, et al. (2004). “Coh-Metrix: Analysis of text on cohesion and language”. In: *Behavior Research Methods, Instruments, & Computers* 36.2, pp. 193–202.
- Grant, Angela M., Shin-Yi Fang, and Ping Li (2015). “Second language lexical development and cognitive control: A longitudinal fMRI study”. In: *Brain and Language* 144, pp. 35–47.
- Graves, Michael F. (1986). “Vocabulary learning and instruction”. In: *Review of Research in Education* 13.1, pp. 49–89.
- Gries, Stefan Th. (2008). “Dispersions and adjusted frequencies in corpora”. In: *International Journal of Corpus Linguistics* 13.4, pp. 403–437.
- (2010). “Dispersions and adjusted frequencies in corpora: further explorations”. In: *Corpus-Linguistic Applications: Current Studies, New Directions*. Ed. by Stefan Th. Gries, Stefanie Wulff, and Mark Davies. Amsterdam, New York: Brill, pp. 197–212.
- Grimes, Joseph Evans (1975). *The Thread of Discourse*. Janua linguarum 207. The Hague: Mouton.
- Grisay, Aletta, John H. A. L. de Jong, et al. (2007). “Translation equivalence across PISA countries”. In: *Journal of Applied Measurement* 8.3, pp. 249–266.
- Grisay, Aletta and Christian Monseur (2007). “Measuring the equivalence of item difficulty in the various versions of an international test”. In: *Studies in Educational Evaluation* 33.1, pp. 69–86.
- Grömping, Ulrike (2007). “Relative importance for linear regression in R: The package relaimpo”. In: *Journal of Statistical Software* 17, pp. 1–27.
- Guo, Fammin, Lawrence Rudner, and Eileen Talento-Miller (2009). *Scaling Item Difficulty Estimates from Nonequivalent Groups*. GMAC Research Report RR-09-03. McLean: Graduate Management Admission Council.
- Hacking, Jane F. and Erwin Tschirner (2017). “The contribution of vocabulary knowledge to reading proficiency: The case of College Russian”. In: *Foreign Language Annals* 50.3, pp. 500–518.
- Hagenmüller, Bettina (2021). “On the impact of the response options’ position on item difficulty in multiple-choice-items”. In: *European Journal of Psychological Assessment* 37.4, pp. 290–299.
- Hagoort, Peter (2005). “On Broca, brain, and binding: a new framework”. In: *Trends in Cognitive Sciences* 9.9, pp. 416–423.
- (2013). “MUC (Memory, Unification, Control) and beyond”. In: *Frontiers in Psychology* 4, Article 416, 1–13.

- Hague, Daryl, Alan Melby, and Wang Zheng (2011). “Surveying translation quality assessment”. In: *The Interpreter and Translator Trainer* 5.2, pp. 243–267.
- Halliday, Michael A. K. (1985). *Spoken and Written Language*. Oxford: Oxford University Press.
- Hamada, Megumi and Keiko Koda (2008). “Influence of first language orthographic experience on second language decoding and word learning”. In: *Language Learning* 58.1, pp. 1–31.
- Hambleton, Ronald K. (1993). *Translating Achievement Tests for Use in Cross-National Studies*. Tech. rep. New York, Washington D.C.: International Association for the Evaluation of Educational Achievement, National Center for Education Statistics.
- Hambleton, Ronald K. and Russell W. Jones (1992). “Comparison of empirical and judgmental methods for detecting Differential Item Functioning”. In: *Laboratory of Psychometric and Evaluative Research Report*. San Francisco: University of Massachusetts at Amherst.
- (1993). “Comparison of Classical Test Theory and Item Response Theory and their applications to test development”. In: *Educational Measurement: Issues and Practice* 12.3, pp. 38–47.
- Hamp, Birgit and Helmut Feldweg (1997). “GermaNet - a lexical-semantic net for German”. In: *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 9–15.
- Hancin-Bhatt, Barbara and William E. Nagy (1994). “Lexical transfer and second language morphological development”. In: *Applied Psycholinguistics* 15.3, pp. 289–310.
- Hancke, Julia, Sowmya Vajjala, and Detmar Meurers (2012). “Readability classification for German using lexical, syntactic, and morphological features”. In: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, pp. 1063–1080.
- Hare, Victoria Chou, Mitchell Rabinowitz, and Karen Magnus Schieble (1989). “Text effects on main idea comprehension”. In: *Reading Research Quarterly* 24.1, pp. 72–88.
- Hatim, Basil and Jeremy Munday (2019). *Translation: An Advanced Resource Book for Students*. 2nd ed. Milton Park: Routledge.
- Heatley, Alex, Paul Nation, and Averil Coxhead (2022). *RANGE*. URL: <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs> (last accessed on 08/06/2025).
- Heine, Lena et al. (2018). “Modellierung sprachlicher Anforderungen in Testaufgaben verschiedener Unterrichtsfächer: Theoretische und empirische Grundlagen”. In: *Zeitschrift für Angewandte Linguistik* 2018.69, pp. 69–96.
- Henrich, Verena and Erhard Hinrichs (2010). “GernEdiT - The GermaNet Editing Tool”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta: European Language Resources Association.
- Hill, Clifford and Kate Parry (1992). “The test at the gate: Models of literacy in reading assessment”. In: *TESOL Quarterly* 26.3, pp. 433–461.
- Holland, Paul W. and Dorothy T. Thayer (1985). *An Alternate Definition of the ETS Delta Scale of Item Difficulty*. Program Statistics Research Technical Report RR-85-43. Princeton: Educational Testing Service.
- Holzknicht, Franz et al. (2021). “The effect of response order on candidate viewing behaviour and item difficulty in a multiple-choice listening test”. In: *Language Testing* 38.1, pp. 41–61.
- Hoover, Wesley A. and Philip B. Gough (1990). “The Simple View of Reading”. In: *Reading and Writing* 2.2, pp. 127–160.
- Hoover, Wesley A. and William E. Tunmer (2022). “The primacy of science in communicating advances in the science of reading”. In: *Reading Research Quarterly* 57.2, pp. 399–408.
- Hopp, Holger (2015). “Individual differences in the second language processing of object–subject ambiguities”. In: *Applied Psycholinguistics* 36.2, pp. 129–173.
- (2019). “Cross-linguistic influence in the child third language acquisition of grammar: Sentence comprehension and production among Turkish-German and German learners of English”. In: *International Journal of Bilingualism* 23.2, pp. 567–583.

- Hu Hsueh-Chao, Marcella and Paul Nation (2000). “Unknown vocabulary density and reading comprehension”. In: *Reading in a Foreign Language* 13.1, pp. 403–30.
- Hulin, Charles L. (1987). “A psychometric theory of evaluations of item and scale translations: Fidelity across languages”. In: *Journal of Cross-Cultural Psychology* 18.2, pp. 115–142.
- IBM Corp. (2020). *IBM SPSS Statistics for Windows*. URL: <https://www.ibm.com/de-de/products/spss-statistics> (last accessed on 08/06/2025).
- Impara, James C. and Barbara S. Plake (1998). “Teachers’ ability to estimate item difficulty: A test of the assumptions in the Angoff Standard Setting Method”. In: *Journal of Educational Measurement* 35.1, pp. 69–81.
- Institut zur Qualitätsentwicklung im Bildungswesen (2020). *IQB - Überprüfen von Bildungsstandards - IQB-Bildungstrends*. URL: <https://www.iqb.hu-berlin.de/bista/control> (last accessed on 07/27/2020).
- International Test Commission (2017). *ITC Guidelines for Translating and Adapting Tests (Second edition)*. URL: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf (last accessed on 05/25/2020).
- Jackson, Carrie N. and Paola E. Dussias (2009). “Cross-linguistic differences and their impact on L2 sentence processing”. In: *Bilingualism: Language and Cognition* 12.1, pp. 65–82.
- Jang, Eunice Eunhee (2009). “Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment”. In: *Language Testing* 26.1, pp. 31–73.
- Jarvis, Scott (2013a). “Capturing the diversity in lexical diversity”. In: *Language Learning* 63.s1, pp. 87–106.
- (2013b). “Defining and measuring lexical diversity”. In: *Vocabulary Knowledge: Human Ratings and Automated Measures*. Ed. by Michael Daller and Scott Jarvis. Amsterdam, Philadelphia: John Benjamins.
- (2017). “Grounding lexical diversity in human judgments”. In: *Language Testing* 34.4, pp. 537–553.
- Jeon, Eun Hee and Junko Yamashita (2014). “L2 reading comprehension and its correlates: A meta-analysis”. In: *Language Learning* 64.1, pp. 160–212.
- Jessner, Ulrike, Elisabeth Allgäuer-Hackl, and Barbara Hofer (2016). “Emerging multilingual awareness in educational contexts: from theory to practice”. In: *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 72.2, pp. 157–182.
- Jiménez López, María Dolores and Adrià Torrens Urrutia (2018). “Sobre el concepto de complejidad en Lingüística”. In: *Actas do XIII Congresso Internacional de Lingüística Xeral*. Vigo, pp. 483–489.
- Joh, Jeongsoon (2004). “Which prior knowledge? Relative contribution of prior knowledge components to EFL reading comprehension”. In: *Language Research* 40.4, pp. 1035–1057.
- Johnson-Laird, Philip N. (1981). “Comprehension as the construction of mental models”. In: *Philosophical Transactions of the Royal Society of London* 295.1077, pp. 353–374.
- (2001). “Mental models and deduction”. In: *Trends in Cognitive Sciences* 5.10, pp. 434–442.
- Jones, Neil et al. (2012). *First European Survey on Language Competences*. Technical report. Brussels: European Commission.
- Juul, T. et al. (2015). “Validation of the English translation of the low anterior resection syndrome score”. In: *Colorectal Disease* 17.10, pp. 908–916.
- Kane, Michael (2006). “Validation”. In: *Educational Measurement*. Ed. by R.L. Brennan, National Council on Measurement in Education, and American Council on Education. 4th ed. Phoenix: Greenwood, pp. 17–64.
- Kane, Michael T. (2013). “Validating the interpretations and uses of test scores”. In: *Journal of Educational Measurement* 50.1, pp. 1–73.
- Kang, Yusun (2021). “Does the Simple View of Reading explain Korean elementary EFL learners’ reading comprehension?” In: *English Teaching* 76.1, pp. 57–78.

- Kankaraš, Miloš and Guy Moors (2014). “Analysis of cross-cultural comparability of PISA 2009 scores”. In: *Journal of Cross-Cultural Psychology* 45.3, pp. 381–399.
- Kanojia, Diptesh et al. (2021). *Cognition-aware cognate detection*. arXiv: 2112.08087. URL: <http://arxiv.org/abs/2112.08087> (last accessed on 02/16/2022).
- Karges, Katharina, Malgorzata Barras, and Peter Lenz (2021). *Leseverstehen in einer Fremdsprache testen. Task Lab – eine empirische Studie zu computerbasierten Aufgaben in Französisch*. Bericht des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit 2021.001. Fribourg/Freiburg: Institut für Mehrsprachigkeit.
- (2022). “Assessing young language learners’ receptive skills: Should we ask the questions in the language of schooling?” In: *Current Research into Young Foreign Language Learners’ Literacy Skills*. Ed. by Stefanie Frisch and Jutta Rymarczyk. Berlin: Peter Lang.
- Karges, Katharina and Peter Lenz (2014). “Challenges in computer-based testing of young learners’ oral competences in French”. In: *Proceedings of EDULEARN14 Conference*. Barcelona: IATED, pp. 7009–7018.
- Kassambara, Alboukadel (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. URL: <https://rpkgs.datanovia.com/rstatix/> (last accessed on 03/20/2025).
- Katz, Stuart et al. (1990). “Answering reading comprehension items without passages on the SAT”. In: *Psychological Science* 1.2, pp. 122–127.
- Keenan, Janice M., Rebecca S. Betjemann, and Richard K. Olson (2008). “Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension”. In: *Scientific Studies of Reading* 12.3, pp. 281–300.
- Kendeou, Panayiota, Kristen L. McMaster, and Theodore J. Christ (2016). “Reading comprehension: Core components and processes”. In: *Policy Insights from the Behavioral and Brain Sciences* 3.1, pp. 62–69.
- Kenyon, Dorry M. and David MacGregor (2012). “Pre-operational testing”. In: *The Routledge Handbook of Language Testing*. Ed. by Glenn Fulcher and Fred Davidson. London, New York: Routledge, pp. 295–306.
- Khalifa, Hanan and Cyril J. Weir (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge, New York: Cambridge University Press.
- Khorramdel, Lale et al. (2020). “Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach”. In: *Psychological Test and Assessment Modeling* 62.2, pp. 179–231.
- Kiat, John Emmanuel, Ai Rene Ong, and Asha Ganesan (2018). “The influence of distractor strength and response order on MCQ responding”. In: *Educational Psychology* 38.3, pp. 368–380.
- Kiefer, Thomas, Alexander Robitzsch, and Margaret Wu (2013). *Tutorial 7: Differential Item Functioning*. URL: <https://www.edmeasurementsurveys.com/TAM/Tutorials/7DIF.htm> (last accessed on 07/18/2024).
- Kieffer, Michael J. and Nonie K. Lesaux (2012). “Knowledge of words, knowledge about words: Dimensions of vocabulary in first and second language learners in sixth grade”. In: *Reading and Writing* 25.2, pp. 347–373.
- Kieras, David (1985). “Thematic processes in the comprehension of technical prose”. In: *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text*. Ed. by Bruce K. Britton and John B. Black. Hillsdale: Lawrence Erlbaum, pp. 89–107.
- Kintsch, Walter (1988). “The role of knowledge in discourse comprehension: A construction-integration model”. In: *Psychological Review* 95.2, pp. 163–182.
- (1995). “The role of syntactic cues and causal inferences”. In: *Coherence in Spontaneous Text*. Ed. by Morton Ann Gernsbacher and T. Givón. Amsterdam, Philadelphia: John Benjamins, pp. 139–160.
- Kintsch, Walter and Teun A. van Dijk (1978). “Toward a model of text comprehension and production”. In: *Psychological Review* 85.5, pp. 363–394.

- Kintsch, Walter and J. Craig Yarbrough (1982). "Role of rhetorical structure in text comprehension". In: *Journal of Educational Psychology* 74.6, pp. 828–834.
- Klare, George R. (1974). "Assessing readability". In: *Reading Research Quarterly* 10.1, pp. 62–102.
- Klieme, Eckhard, Johannes Hartig, and Dominique Rauch (2008). "The concept of competence in educational contexts". In: *Assessment of Competencies in Educational Contexts*. Ed. by Johannes Hartig, Eckhard Klieme, and Detlev Leutner. Göttingen: Hogrefe & Huber, pp. 3–22.
- Kobayashi, Miyoko (2002). "Method effects on reading comprehension test performance: text organization and response format". In: *Language Testing* 19.2, pp. 193–220.
- Koda, Keiko (1993). "Transferred L1 strategies and L2 syntactic structure in L2 sentence comprehension". In: *The Modern Language Journal* 77.4, pp. 490–500.
- (1999). "Development of L2 intraword orthographic sensitivity and decoding skills". In: *The Modern Language Journal* 83.1, pp. 51–64.
- (2007). "Reading and language learning: Crosslinguistic constraints on second language reading development". In: *Language Learning* 57, pp. 1–44.
- Koda, Keiko and Annette Marie Zehler (2008). "Conceptualizing reading universals, cross-linguistic variations, and second language literacy development". In: *Learning to Read Across Languages: Cross-Linguistic Relationships in First- and Second-Language Literacy Development*. Ed. by Keiko Koda and Annette Marie Zehler. New York, London: Routledge, pp. 1–8.
- Kofler, Karolina, Elisabeth Peyer, and Malgorzata Barras (2020). "Wie wird die Mehrsprachigkeitsdidaktik in Lehrwerken umgesetzt? Eine Analyse der mehrsprachigkeitsdidaktischen Ansätze in Englisch- und Französisch-Lehrwerken der Deutschschweiz". In: *Perspektiven auf Mehrsprachigkeit im Fremdsprachenunterricht – Regards croisés sur le plurilinguisme et l'apprentissage des langues*. Ed. by Birgit Schädlich. Berlin, Heidelberg: Springer, pp. 101–120.
- Kondrak, Greg (2009). *True French-English cognates*. URL: <http://webdocs.cs.ualberta.ca/~kondrak/cognatesEF.html> (last accessed on 06/26/2023).
- Kondrak, Grzegorz (2001). "Identifying cognates by phonetic and semantic similarity". In: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Konsortium ÜGK (2019a). *Überprüfung der Grundkompetenzen: Nationaler Bericht der ÜGK 2017: Sprachen 8. Schuljahr*. URL: https://uegk-schweiz.ch/wp-content/uploads/2019/05/UGK_2017_DE.pdf (last accessed on 02/28/2022).
- (2019b). *Überprüfung der Grundkompetenzen. Nationaler Bericht der ÜGK 2016: Mathematik 11. Schuljahr*. URL: <https://phsg.contentdm.oclc.org/digital/collection/p15782coll13/id/386> (last accessed on 03/14/2022).
- Koslowski, Barbara and Amy Masnick (2010). "Causal reasoning and explanation". In: *The Wiley-Blackwell Handbook of Childhood Cognitive Development*. Ed. by Ushna Goswami. 2nd ed. Chichester: Wiley-Blackwell, pp. 377–398.
- Kostin, Irene (2004). *Exploring Item Characteristics That Are Related to the Difficulty of TOEFL Dialogue Items*. ETS Research Report RR-04-11. Princeton: ETS.
- Kremmel, Benjamin, Kathrin Eberharter, and Franz Holzknecht (2021). "Pre-operational testing". In: *The Routledge Handbook of Language Testing*. Ed. by Glenn Fulcher and Luke Harding. 2nd ed. London: Routledge.
- Krstić, Ksenija et al. (2018). "All good readers are the same, but every low-skilled reader is different: An eye-tracking study using PISA data". In: *European Journal of Psychology of Education* 33.3, pp. 521–541.
- Kučera, Henry and W. Nelson Francis (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.
- Kuliš, Dagmara et al. (2017). "Reviewing back translation reports of questionnaires: the EORTC conceptual framework and experience". In: *Expert Review of Pharmacoeconomics & Outcomes Research* 17.6, pp. 523–530.

- Kusseling, Françoise and Deryle Lonsdale (2013). “A corpus-based assessment of French CEFR lexical content”. In: *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 69.4, pp. 436–461.
- Kyle, Kristopher and Scott A. Crossley (2015). “Automatically assessing lexical sophistication: Indices, tools, findings, and application”. In: *TESOL Quarterly* 49.4, pp. 757–786.
- Kyle, Kristopher, Scott A. Crossley, and Scott Jarvis (2021). “Assessing the validity of lexical diversity indices using direct judgements”. In: *Language Assessment Quarterly* 18.2, pp. 154–170.
- Landauer, Thomas K. and Susan T. Dumais (1997). “A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.” In: *Psychological Review* 104.2, p. 211.
- Landauer, Thomas K., Danielle S. McNamara, et al., eds. (2011). *Handbook of Latent Semantic Analysis*. New York: Routledge.
- Laufer, Batia (1992). “Reading in a foreign language: How does L2 lexical knowledge interact with the reader’s general academic ability?”. In: *Journal of Research in Reading* 15.2, pp. 95–103.
- Laufer, Batia and Paul Nation (1995). “Vocabulary size and use: Lexical richness in L2 written production”. In: *Applied Linguistics* 16.3, pp. 307–322.
- (2001). “Passive vocabulary size and speed of meaning recognition: Are they related?” In: *EUROSLA Yearbook* 1.1, pp. 7–28.
- Laufer, Batia and Geke C. Ravenhorst-Kalovski (2010). “Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension”. In: *Reading in a Foreign Language* 22.1, pp. 15–30.
- Lauscher, Susanne (2000). “Translation quality assessment”. In: *The Translator* 6.2, pp. 149–168.
- Leal, Sidney Evaldo et al. (2024). “NILC-Matrix: Assessing the complexity of written and spoken language in Brazilian Portuguese”. In: *Language Resources and Evaluation* 58.1, pp. 73–110.
- Lechner, Clemens M. et al. (2021). “Why ability point estimates can be pointless: A primer on using skill measures from large-scale assessments in secondary analyses”. In: *Measurement Instruments for the Social Sciences* 3.2.
- Lemhöfer, Kristin, Ton Dijkstra, and Marije C. Michel (2004). “Three languages, one ECHO: Cognate effects in trilingual word recognition”. In: *Language and Cognitive Processes* 19.5, pp. 585–611.
- Lenz, Peter, Katharina Karges, and Malgorzata Barras (2019). “Investigating test method effects in French L2 reading items for young learners”. In: *Developments in Language Education: A Memorial Volume in Honour of Sauli Takala*. Ed. by Ari Huhta, Gudrun Erickson, and Neus Figueras. Jyväskylä: University of Jyväskylä & EALTA, pp. 182–202.
- Lenz, Peter, Katharina Karges, Anna Kull, et al. (2019). *Computerbasierte Erfassung von Fremdsprachenkompetenzen im Unterricht der Primarstufe. Bericht zum Projekt „Erfassung von Entwicklungsprofilen im frühen schulischen Fremdsprachenlernen“*. Bericht des Wissenschaftlichen Kompetenzzentrums für Mehrsprachigkeit. Fribourg/Freiburg: Institut für Mehrsprachigkeit.
- Lenz, Peter and Thomas Studer (2007). *lingualevel: Französisch und Englisch. Instrumente zur Evaluation von Fremdsprachenkompetenzen*. Bern: Schulverlag.
- Levenshtein, Vladimir Iosifovich (1966). “Binary codes capable of correcting deletions, insertions and reversals”. In: *Soviet Physics - Doklady* 10.8, pp. 707–710.
- Levin, Beth and Malka Rappaport Hovav (2017). “Morphology and lexical semantics”. In: *The Handbook of Morphology*. Ed. by Andrew Spencer and Arnold M. Zwicky. Hoboken: Wiley, pp. 248–271.
- Lewis, Gwyn, Bryn Jones, and Colin Baker (2012). “Translanguaging: origins and development from school to street and beyond”. In: *Educational Research and Evaluation* 18.7, pp. 641–654.

- Li, Hongli, Charles Vincent Hunter, and Jacquelyn Anne Bialo (2022). “A revisit of Zumbo’s third generation DIF: How are we doing in language testing?” In: *Language Assessment Quarterly* 19.1, pp. 27–53.
- Li, Ping and Roy B. Clariana (2019). “Reading comprehension in L1 and L2: An integrative approach”. In: *Journal of Neurolinguistics* 50, pp. 94–105.
- Liao, Wen-Wei et al. (2012). “The four-parameter logistic Item Response Theory model as a robust method of estimating ability despite aberrant responses”. In: *Social Behavior and Personality* 40.10, pp. 1679–1694.
- Liaw, Andy and Matthew Wiener (2002). “Classification and regression by randomForest”. In: *R News* 2.3, pp. 18–22.
- Lienert, Gustav A. and Ulrich Raatz (1998). *Testaufbau und Testanalyse*. 6th ed. Weinheim: Beltz, Psychologie Verlags Union.
- Lindeman, Richard H., Peter F. Merenda, and Ruth Z. Gold (1980). *Introduction to Bivariate and Multivariate Analysis*. Glenview: Scott, Foresman and Company.
- Linnarud, Moira and Stig Thoursie (2008). “English and German in Swedish classrooms: Writing in the two languages compared”. In: *Nordic Journal of English Studies* 7.2, pp. 75–98.
- Lions, Séverin et al. (2022). “Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review”. In: *Applied Measurement in Education* 35.2, pp. 133–152.
- Lord, Frederic M. (1953). “The relation of test score to the trait underlying the test”. In: *Educational and Psychological Measurement* 13.4.
- (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale: Erlbaum.
- Mancuso, Josephine M. (2009). “Assessment and measurement of health literacy: An integrative review of the literature”. In: *Nursing & Health Sciences* 11.1, pp. 77–89.
- Maneesriwongul, Wantana and Jane K. Dixon (2004). “Instrument translation process: a methods review”. In: *Journal of Advanced Nursing* 48.2, pp. 175–186.
- Manuel, Herschel T. (1953). “The use of parallel tests in the study of foreign language teaching”. In: *Educational and Psychological Measurement* 13.3, pp. 431–436.
- Martinez, Ron and Norbert Schmitt (2012). “A phrasal expressions list”. In: *Applied Linguistics* 33.3, pp. 299–320.
- Mason, Robert A. and Marcel Adam Just (2006). “Neuroimaging contributions to the understanding of discourse processes”. In: *Handbook of psycholinguistics*. Ed. by Matthew J. Traxler and Morton Ann Gernsbacher. 2nd ed. Boston: Elsevier, pp. 756–799.
- (2012). “Identifying component discourse processes from their fMRI time course signatures”. In: *Reading—from Words to Multiple Texts*. Ed. by M. Anne Britt, Susan R. Goldman, and Jean-François Rouet. New York, London: Routledge, pp. 147–159.
- Matras, Yaron (2011). *Language Contact*. Cambridge: Cambridge University Press.
- McCarthy, Philip M. and Scott Jarvis (2010). “MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment”. In: *Behavior Research Methods* 42.2, pp. 381–392.
- McNamara, Danielle S. (2001). “Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge”. In: *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale* 55, pp. 51–62.
- McNamara, Danielle S., Arthur C. Graesser, et al. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press.
- McNamara, Danielle S., Eileen Kintsch, et al. (1996). “Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text”. In: *Cognition and Instruction* 14.1, pp. 1–43.
- McNamara, Danielle S., Max M. Louwerse, et al. (2010). “Coh-Metrix: Capturing linguistic features of cohesion”. In: *Discourse Processes* 47.4, pp. 292–330.

- McNamara, Danielle S. and Joe Magliano (2009). “Toward a comprehensive model of comprehension”. In: *Psychology of Learning and Motivation* 51, pp. 297–384.
- McVee, Mary B., Kailonnie Dunsmore, and James R. Gavelek (2005). “Schema theory revisited”. In: *Review of Educational Research* 75.4, pp. 531–566.
- Melamed, I. Dan (1999). “Bitext maps and alignment via pattern recognition”. In: *Computational Linguistics* 25.1, pp. 107–130.
- Melby-Lervåg, Monica and Arne Lervåg (2014). “Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners.” In: *Psychological Bulletin* 140.2, pp. 409–433.
- Mellard, Daryl F., Emily Fall, and Caroline Mark (2009). “Reading profiles for adults with low-literacy: cluster analysis with power and speeded measures”. In: *Reading and Writing* 22.8, pp. 975–992.
- Messick, Samuel (1990). *Validity of Test Interpretation and Use*. ETS Research Report RR-90-11. Princeton: Educational Testing Service.
- Metallidou, Panayiota et al. (2012). “Changes in children’s beliefs about everyday reasoning: Evidence from Greek primary students”. In: *Australian Journal of Educational & Developmental Psychology* 12, pp. 83–92.
- Meyer, Bonnie J. F. (1985). “Prose analysis: purposes, procedures, and problems”. In: *Understanding Expository Text*. London: Routledge.
- Meyer, Bonnie J. F. et al. (2010). “Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth- and seventh-grade readers”. In: *Reading Research Quarterly* 45.1, pp. 62–92.
- Michalke, Meik (2021). *koRpus: Text Analysis with Emphasis on POS Tagging, Readability, and Lexical Diversity*. URL: <http://reaktanz.de/?c=hacking&s=koRpus> (last accessed on 04/04/2023).
- Michigan Language Assessment (2020). *MET Test*. URL: <https://michigan-test.com/met/> (last accessed on 12/10/2020).
- Microsoft (2024). *Microsoft Excel LTSC MSO*. URL: <https://learn.microsoft.com/en-us/office/ltsc/2021/overview> (last accessed on 08/07/2025).
- Mikami, Hitoshi, Chi Yui Leung, and Lisa Yoshikawa (2008). “The threshold of anxiety in low-stakes testing for foreign language reading”. In: *Reading in a Foreign Language* 30.1, pp. 92–107.
- Mislevy, Robert J. and Chengbin Yin (2012). “Evidence-Centered Design in language testing”. In: *The Routledge Handbook of Language Testing*. Ed. by Glenn Fulcher and Fred Davidson. London, New York: Routledge, pp. 208–222.
- Mons, Nathalie (2009). “Effets théoriques et réels des politiques d’évaluation standardisée”. In: *Revue française de pédagogie* 169, pp. 99–140.
- Moosbrugger, Helfried and Augustin Kelava, eds. (2008). *Testtheorie und Fragebogenkonstruktion*. ger. Springer-Lehrbuch. Berlin: Springer.
- Mulder, Kimberley, Ton Dijkstra, and R. Harald Baayen (2015). “Cross-language activation of morphological relatives in cognates: The role of orthographic overlap and task-related processing”. In: *Frontiers in Human Neuroscience* 9.
- Musitu, Gonzalo, Fernando García, and Melchor Gutiérrez (1997). *AFA: Autoconcepto forma A*. 3rd ed. Madrid: TEA.
- Myers, Jerome L., Makiko Shinjo, and Susan A. Duffy (1987). “Degree of causal relatedness and memory”. In: *Journal of Memory and Language* 26, pp. 453–465.
- Nagy, William E. (2007). “Metalinguistic awareness and the vocabulary-comprehension connection”. In: *Vocabulary Acquisition: Implications for Reading Comprehension*. Ed. by Richard K. Wagner, Andrea E. Muse, and Kendra R. Tannenbaum. New York, London: Guilford, pp. 52–77.

- Nagy, William E., Joanne F. Carlisle, and Amanda P. Goodwin (2014). “Morphological knowledge and literacy acquisition”. In: *Journal of Learning Disabilities* 47.1, pp. 3–12.
- Nagy, William E., Irene-Anna N. Diakidoy, and Richard C. Anderson (1993). “The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives”. In: *Journal of Reading Behavior* 25.2, pp. 155–170.
- Nagy, William E., Erica F. McClure, et al. (1995). *Linguistic Transfer and the Use of Context by Spanish-English Bilinguals*. Technical report 616. Champaign: Center for the Study of Reading.
- Nasseri, Maryam and Paul Thompson (2021). “Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences”. In: *Assessing Writing* 47, p. 100511.
- National Academies of Sciences, Engineering, and Medicine (2020). *A Principled Approach to Language Assessment: Considerations for the U.S. Foreign Service Institute*. Washington, D.C.: The National Academies Press.
- Neumann, Daniela (2012). *Schwierigkeitsbeeinflussende Merkmale bei Aufgaben zum Hörverstehen im Fach Deutsch in der Sekundarstufe I*. Duisburg: Gilles & Francke.
- New, Boris et al. (2007). “The use of film subtitles to estimate word frequencies”. In: *Applied Psycholinguistics* 28.4, pp. 661–677.
- Newton, Paul E. and Stuart D. Shaw (2014). *Validity in Educational and Psychological Assessment*. London, Cambridge: SAGE & Cambridge Assessment.
- Nguyen, Tam H. et al. (2014). “An introduction to Item Response Theory for patient-reported outcome measurement”. In: *The Patient - Patient-Centered Outcomes Research* 7.1, pp. 23–35.
- Nilsen, Per (2015). “Making sense of implementation theories, models and frameworks”. In: *Implementation Science* 10.1, p. 53.
- Nimon, Kim et al. (2008). “An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example”. en. In: *Behavior Research Methods* 40.2, pp. 457–466.
- Nissan, Susan, Felicia DeVincenzi, and K. Linda Tang (1996). *An Analysis of Factors Affecting the Difficulty of Dialogue Items in TOEFL Listening Comprehension*. ETS Research Report RR-95-37. Princeton: Educational Testing Service.
- Noijons, José and Erna Gille (2010). “Issues in constructing an international item bank: Differences in item behaviour across countries”. In: *Revue française de linguistique appliquée* XV.1, pp. 49–59.
- Norman, Cameron D. and Harvey A. Skinner (2006). “eHEALS: The eHealth Literacy Scale”. In: *Journal of Medical Internet Research* 8.4, e507.
- Norris, John M. (2016). “Current uses for task-based language assessment”. In: *Annual Review of Applied Linguistics* 36, pp. 230–244.
- Novick, Melvin R. (1966). “The axioms and principal results of Classical Test Theory”. In: *Journal of Mathematical Psychology* 3.1, pp. 1–18.
- Nunes Silva, Carlos (2012). “Ex post facto study”. In: *Encyclopedia of Research Design*. Ed. by Neil Salkind. Thousand Oaks: SAGE, p. 466.
- O’Reilly, Tenaha and Danielle S. McNamara (2007). “Reversing the Reverse Cohesion Effect: Good texts can be better for strategic, high-knowledge readers”. In: *Discourse Processes* 43.2, pp. 121–152.
- Odlin, Terence (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press.
- (2005). “Crosslinguistic influence and conceptual transfer: What are the concepts?” In: *Annual Review of Applied Linguistics* 25, pp. 3–25.
- OECD (2014). *PISA 2012. Technical Report*. Paris: OECD Publishing.
- (2019a). *PISA 2018 assessment and analytical framework*. Paris: OECD Publishing.

- (2019b). *PISA 2018 Results (Volume I): What Students Know and Can Do*. Vol. 1. PISA 2018 Results. Paris: OECD Publishing.
- (2019c). *PISA 2018 Results (Volume II): Where All Students Can Succeed*. Vol. 2. PISA 2018 Results. Paris: OECD Publishing.
- (2021). *PISA 2025 Foreign Language Assessment Framework*. OECD Publishing.
- Okinina, Nadezda, Jennifer-Carmen Frey, and Zarah Weiss (2020). “CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 7123–7131.
- Oliveri, Maria E. and Kadriye Ercikan (2011). “Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions?” In: *Applied Measurement in Education* 24.4, pp. 349–366.
- Otheguy, Ricardo, Ofelia García, and Wallis Reid (2019). “A translanguaging view of the linguistic system of bilinguals”. In: *Applied Linguistics Review* 10.4, pp. 625–651.
- Otwinowska, Agnieszka et al. (2020). “Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross-linguistic similarity matter?” In: *Language Learning* 70.3, pp. 685–731.
- Ozolins, Uldis et al. (2020). “Translation and back-translation methodology in health research – a critique”. In: *Expert Review of Pharmacoeconomics & Outcomes Research* 20.1, pp. 69–77.
- Ozuru, Yasuhiro et al. (2008). “Where’s the difficulty in standardized reading tests: The passage or the question?” In: *Behavior Research Methods* 40.4, pp. 1001–1015.
- Pae, Tae-Il (2018). “Effects of task type and L2 proficiency on the relationship between L1 and L2 in reading and writing: An SEM approach”. In: *Studies in Second Language Acquisition* 40.1, pp. 63–90.
- Palfreyman, David M. and Suha Karaki (2019). “Lexical sophistication across languages: A preliminary study of undergraduate writing in Arabic (L1) and English (L2)”. In: *International Journal of Bilingual Education and Bilingualism* 22.8, pp. 992–1015.
- Pallotti, Gabriele (2015). “A simple view of linguistic complexity”. In: *Second Language Research* 31.1, pp. 117–134.
- Penk, Christiane and Dirk Richter (2017). “Change in test-taking motivation and its relationship to test performance in low-stakes assessments”. In: *Educational Assessment, Evaluation and Accountability* 29.1, pp. 55–79.
- Penzkofer, Markus (2023). *Decaleon*. URL: <https://sourceforge.net/projects/pentaleon2/files/DatabasePlus/> (last accessed on 04/17/2023).
- Perfetti, Charles A. (1985). *Reading Ability*. New York: Oxford University Press.
- (1989). “There are generalized abilities and one of them is reading”. In: *Knowing, Learning, and instruction: Essays in Honor of Robert Glaser*. Ed. by Lauren B. Resnick. Lawrence Erlbaum Associates, pp. 307–336.
- (2007). “Reading ability: Lexical quality to comprehension”. In: *Scientific Studies of Reading* 11.4, pp. 357–383.
- Perfetti, Charles A. and Joseph Stafura (2014). “Word knowledge in a theory of reading comprehension”. In: *Scientific Studies of Reading* 18.1, pp. 22–37.
- Perfetti, Charles A., Chin-Lung Yang, and Franz Schmalhofer (2008). “Comprehension skill and word-to-text integration processes”. In: *Applied Cognitive Psychology* 22.3, pp. 303–318.
- Perkins, Kyle, Lalit Gupta, and Ravi Tammana (1995). “Predicting item difficulty in a reading comprehension test with an artificial neural network”. In: *Language Testing* 12.1.
- Peters, Elke, Ann-Sophie Noreillie, et al. (2019). “The impact of instruction and out-of-school exposure to foreign language input on learners’ vocabulary knowledge in two languages”. In: *Language Learning* 69.3, pp. 747–782.

- Peters, Elke and Stuart Webb (2018). “Incidental vocabulary acquisition through viewing L2 television and factors that affect learning”. In: *Studies in Second Language Acquisition* 40.3, pp. 551–577.
- Petscher, Yaacov et al. (2020). “Past perspectives and new opportunities for the explanatory item response model”. In: *Annals of Dyslexia* 70.2, pp. 160–179.
- Peyer, Elisabeth, Irmtraud Kaiser, and Raphael Berthele (2010). “The multilingual reader: Advantages in understanding and decoding German sentence structure when reading German as an L3”. In: *International Journal of Multilingualism* 7.3, pp. 225–239.
- Pintard, Alice and Thomas François (2020). “Combining expert knowledge with frequency information to infer CEFR levels for words”. In: *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding Difficulties (READI)*. Marseille: European Language Resources Association, pp. 85–92.
- Princeton University (2010). *About WordNet*. URL: <https://wordnet.princeton.edu/citing-wordnet> (last accessed on 09/27/2022).
- Ptáčková, Monika and Kateřina Vitásková (2020). “Orthographic neighbourhood in Czech language and its use in research of reading impairment”. In: *Linguistic Frontiers* 3.1, pp. 54–59.
- Puimège, Eva and Elke Peters (2019). “Learners’ English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables”. In: *Language Learning* 69.4, pp. 943–977.
- Putwain, David (2008). “Do examinations stakes moderate the test anxiety–examination performance relationship?” In: *Educational Psychology* 28.2, pp. 109–118.
- Putwain, David William (2008). “Deconstructing test anxiety”. In: *Emotional and Behavioural Difficulties* 13.2, pp. 141–155.
- Quispesaravia, Andre et al. (2016). “Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC ’16)*. Ed. by Nicoletta Calzolari et al. Portorož: European Language Resources Association (ELRA), pp. 4694–4698.
- R Core Team (2024). *R: A Language and Environment for Statistical Computing*. URL: <https://www.R-project.org> (last accessed on 08/07/2025).
- Raphael, Taffy E. and James R. Gavelek (1984). “Question-related activities and their relationship to reading comprehension: Some instructional implications”. In: *Comprehension Instruction: Perspectives and Suggestions*. Ed. by Gerald G. Duffy, Laura R. Roehler, and Jana Mason. New York: Longman, pp. 234–250.
- Razavi, Pooya and Sonya J. Powers (2025). *Estimating item difficulty using large language models and tree-based machine learning algorithms*. URL: <http://arxiv.org/abs/2504.08804> (last accessed on 05/04/2025).
- Read, John and Carol A. Chapelle (2001). “A framework for second language vocabulary assessment”. In: *Language Testing* 18.1, pp. 1–32.
- Read, John and Paul Nation (2006). *An Investigation of the Lexical Dimension of the IELTS Speaking Test*. IELTS Research Reports 8. IELTS Australia and British Council.
- Revelle, William (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. URL: <https://CRAN.R-project.org/package=psych> (last accessed on 03/13/2025).
- Révész, Andrea and Tineke Brunfaut (2013). “Text characteristics of task input and difficulty in second language listening comprehension”. In: *Studies in Second Language Acquisition* 35.1, pp. 31–65.
- Riba, Patrick (2016). *Niveaux C1/C2 pour le français. Éléments pour un référentiel*. Paris: Didier.
- Ringbom, Håkan and Scott Jarvis (2009). “The importance of cross-linguistic similarity in foreign language learning”. In: *The Handbook of Language Teaching*. Chichester: John Wiley & Sons, pp. 106–118.

- Robitzsch, Alexander (2021). “A comparison of estimation methods for the Rasch model”. In: *Book of Short Papers: SIS 2021*. Ed. by Cira Perna, Nicola Salvati, and Francesco Schirripa Spagnolo. Pearson, pp. 157–162.
- Robitzsch, Alexander, Thomas Kiefer, and Margaret Wu (2022). *TAM: Test Analysis Modules*. URL: <https://CRAN.R-project.org/package=TAM> (last accessed on 07/19/2024).
- Rosebrock, Cornelia (2012). “Was ist Lesekompetenz, und wie kann sie gefördert werden?” In: *leseforum.ch* 2012.3.
- Rothman, Jason (2011). “L3 syntactic transfer selectivity and typological determinacy: The typological primacy model”. In: *Second Language Research* 27.1, pp. 107–127.
- Rupp, Andre A., Paula Garcia, and Joan Jamieson (2001). “Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items”. In: *International Journal of Testing* 1.3-4, pp. 185–216.
- Rupp, André A., Tracy Ferne, and Hyeran Choi (2006). “How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective”. In: *Language Testing* 23.4, pp. 441–474.
- Rutkowski, Leslie et al. (2010). “International large-scale assessment data: Issues in secondary analysis and reporting”. In: *Educational Researcher* 39.2, pp. 142–151.
- Sabatini, John P. et al. (2014). “Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges”. In: *L'Année psychologique* 114.4, pp. 693–723.
- Sadoski, Mark and Allan Paivio (2007). “Toward a unified theory of reading”. In: *Scientific Studies of Reading* 11.4, pp. 337–356.
- Sanjose, Vicente, Eduardo Vidal-Abarca, and Olga M. Padilla (2006). “A connectionist extension to Kintsch’s Construction–Integration model”. In: *Discourse Processes* 42.1, pp. 1–35.
- Sawaki, Yasuyo, Hae-Jin Kim, and Claudia Gentile (2009). “Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments”. In: *Language Assessment Quarterly* 6.3, pp. 190–209.
- SCCRE (2023). *Education Report Switzerland 2023*. URL: https://www.skbf-csre.ch/fileadmin/files/pdf/bildungsberichte/2023/BiBer_2023_E.pdf (last accessed on 05/01/2023).
- Schepens, Job, Ton Dijkstra, and Franc Grootjen (2012). “Distributions of cognates in Europe as based on Levenshtein distance”. In: *Bilingualism: Language and Cognition* 15.1, pp. 157–166.
- Schluer, Jennifer (2017). “Wortbewusstheit und Konzeptbewusstheit beim Lesen fremdsprachiger Texte: Theoretische Überlegungen und empirische Erkenntnisse”. In: *Zeitschrift für Fremdsprachenforschung* 28.2, pp. 181–208.
- Schmid, Helmut (1994). “Probabilistic part-of-speech tagging using decision trees”. In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, pp. 44–49.
- Schmitt, Norbert, Xiangying Jiang, and William Grabe (2011). “The percentage of words known in a text and reading comprehension”. In: *The Modern Language Journal* 95.1, pp. 26–43.
- Schmitt, Norbert, Diane Schmitt, and Caroline Clapham (2001). “Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test”. In: *Language Testing* 18.1, pp. 55–88.
- Schnotz, Wolfgang (2006). “Was geschieht im Kopf des Lesers? Mentale Konstruktionsprozesse beim Textverstehen aus der Sicht der Psychologie und der kognitiven Linguistik”. In: *Text - Verstehen. Grammatik und darüber hinaus*. Ed. by Hardarik Blühdorn, Eva Breindl, and Ulrich H. Waßner. Berlin: De Gruyter, pp. 222–238.
- Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) (2010). *Die interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS-Konkordat) vom 14. Juni 2007*. URL: <https://edudoc.ch/record/24711?ln=en>.

- Shardlow, Matthew, Richard Evans, and Marcos Zampieri (2022). “Predicting lexical complexity in English texts: The Complex 2.0 dataset”. In: *Language Resources and Evaluation* 56, pp. 1153–1194.
- Shatz, Itamar, Theodora Alexopoulou, and Akira Murakami (2024). “The potential influence of crosslinguistic lexical similarity on lexical diversity in L2 English writing”. In: *Corpora* 19.2, pp. 131–156.
- Sheehan, Kathleen, Irene Kostin, and Hilary Persky (2006). *Predicting item difficulty as a function of inferential processing requirements: An examination of the reading skills underlying performance on the NAEP grade 8 reading assessment*.
- Sheehan, Kathleen M. et al. (2014). “The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment”. In: *The Elementary School Journal* 115.2, pp. 184–209.
- Shen, Ye and Stephanie N. Del Tufo (2022). “The influence of orthographic depth on multilinguals’ neural networks”. In: *Neuropsychologia* 164, p. 108095.
- Shin, Jihye, Vedran Dronjic, and Boonjoo Park (2019). “The interplay between working memory and background knowledge in L2 reading comprehension”. In: *TESOL Quarterly* 53.2, pp. 320–347.
- Shiotsu, Toshihiko and Cyril J. Weir (2007). “The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance”. In: *Language Testing* 24.1, pp. 99–128.
- Sireci, Stephen G. (2005). “Using bilinguals to evaluate the comparability of different language versions of a test”. In: *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Ed. by Ronald K. Hambleton, Peter F. Merenda, and Charles D. Spielberger. Mahwah: Lawrence Erlbaum, pp. 117–138.
- (2009). “Packing and unpacking sources of validity evidence”. In: *The Concept of Validity: Revisions, New Directions and Applications*. Ed. by Robert W. Lissitz. Charlotte: Information Age Publishing, pp. 19–37.
- Sireci, Stephen G. and Avi Allalouf (2003). “Appraising item equivalence across multiple languages and cultures”. In: *Language Testing* 20.2, pp. 148–166.
- Sireci, Stephen G., James Harter, et al. (2003). “Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats”. In: *International Journal of Testing* 3.2, pp. 129–150.
- Sireci, Stephen G. and Shameem Nyla Khaliq (2003). *Comparing the Psychometric Properties of Monolingual and Dual Language Test Forms*. Center for Educational Assessment MCAS Validity Report 3. Amherst: School of Education, University of Massachusetts Amherst.
- Sireci, Stephen G., Yongwei Yang, et al. (2006). “Evaluating guidelines for test adaptations: A methodological analysis of translation quality”. In: *Journal of Cross-Cultural Psychology* 37.5, pp. 557–567.
- Slater, Wayne H., Michael F. Graves, and Gene L. Piché (1985). “Effects of structural organizers on ninth-grade students’ comprehension and recall of four patterns of expository text”. In: *Reading Research Quarterly* 20.2, pp. 189–202.
- Smith, Dean R. et al. (1989). *The Lexile Scale in Theory and Practice*. Final Report for NIH Grant. Washington, DC: MetaMetrics Inc.
- Smith, Reid et al. (2021). “The role of background knowledge in reading comprehension: A critical review”. In: *Reading Psychology* 42.3, pp. 214–240.
- Smolka, Eva, Katrin H. Preller, and Carsten Eulitz (2014). “‘Verstehen’ (‘understand’) primes ‘stehen’ (‘stand’): Morphological structure overrides semantic compositionality in the lexical representation of German complex verbs”. In: *Journal of Memory and Language* 72, pp. 16–36.

- Solano-Flores, Guillermo (2021). “The semiotics of test design: Conceptual framework on optimal item features in educational assessment across cultural groups, countries, and languages”. In: *Frontiers in Education* 6.
- Solano-Flores, Guillermo, Eduardo Backhoff, and Luis Ángel Contreras-Niño (2009). “Theory of Test Translation Error”. In: *International Journal of Testing* 9.2, pp. 78–91.
- Solano-Flores, Guillermo, Luis Ángel Contreras-Niño, and Eduardo Backhoff (2013). “The measurement of translation error in PISA-2006 items: An application of the Theory of Test Translation Error”. In: *Research on PISA: Research Outcomes of the PISA Research Conference 2009*. Ed. by Manfred Prenzel et al. Dordrecht: Springer Netherlands, pp. 71–85.
- Sonnleitner, Philipp (2008). “Using the LLTM to evaluate an item-generating system for reading comprehension”. In: *Psychology Science Quarterly* 50.3, pp. 345–362.
- Sparks, Richard L. (2021). “Identification and characteristics of strong, average, and weak foreign language readers: The Simple View of Reading model”. In: *The Modern Language Journal* 105.2, pp. 507–525.
- Spelberg, Henk C. Lutje, Paulien de Boer, and Kees P. van den Bos (2000). “Item type comparisons of language comprehension tests”. In: *Language Testing* 17.3, pp. 311–322.
- SPLASH-db.eu (2014). *Policy: "Intercantonal Agreement on Harmonisation of Compulsory Education (HarmoS Agreement) of 14 June 2007" (Information provided by Regula Julia Leemann & Sandra Hafner)*. URL: <https://splash-db.eu/policydocument/intercantonal-agreement-on-harmonisation-of-compulsory-education-harmos-agreement-of-14-june-2007/> (last accessed on 06/15/2022).
- Stanke, Luke and Okan Bulut (2019). “Explanatory item response models for polytomous item responses”. In: *International Journal of Assessment Tools in Education* 6.2, pp. 259–278.
- Stanovsky, Gabriel et al. (2014). “Intermediary semantic representation through proposition structures”. In: *Proceedings of the ACL 2014 Workshop on Semantic Parsing*. Baltimore: Association for Computational Linguistics, pp. 66–70.
- Stegmann, Tilbert Dídac (2005). “EuroComRom - Les sept tamis pour le plurilinguisme réceptif”. In: *Synergies Italie* 2, pp. 35–42.
- Stiller, Jurik et al. (July 2016). “Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty”. In: *Assessment & Evaluation in Higher Education* 41.5, pp. 721–732.
- Strohner, Hans (2006). “Textverstehen aus psycholinguistischer Sicht”. In: *Text - Verstehen. Grammatik und darüber hinaus*. Ed. by Hardarik Blühdorn, Eva Breindl, and Ulrich H. Wafner. Berlin: De Gruyter, pp. 187–204.
- Stubbe, Tobias C. (2011). “How do different versions of a test instrument function in a single language? A DIF analysis of the PIRLS 2006 German assessments”. In: *Educational Research and Evaluation* 17.6, pp. 465–481.
- Svetina, Dubravka, Joanna S. Gorin, and Kikumi K. Tatsuoka (2011). “Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach”. In: *International Journal of Testing* 11.1, pp. 1–23.
- Swami, Viren and David Barron (2019). “Translation and validation of body image instruments: Challenges, good practice guidelines, and reporting recommendations for test adaptation”. In: *Body Image* 31, pp. 204–220.
- Swinton, Spencer S. and Donald E. Powers (1980). *Factor analysis of the Test of English as a Foreign Language for several language groups*. TOEFL Research Report RR-80-32. Princeton: Educational Testing Service.
- Tardieu, Claire (2010). “Votre B1 est-il mon B1 ? L’interculturel dans les tests d’évaluation en Europe”. In: *Recherches en didactique des langues et des cultures* 7.2, pp. 225–239.
- Taşçı, Samet and Ümit Deniz Turan (2021). “The contribution of lexical breadth, lexical depth, and syntactic knowledge to L2 reading comprehension across different L2 reading proficiency groups”. In: *English Teaching & Learning* 45.2, pp. 145–165.

- Tatsuoka, Kikumi K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. New York: Routledge.
- Taylor, Jessica Nelson and Charles A. Perfetti (2016). “Eye movements reveal readers’ lexical quality and reading experience”. In: *Reading and Writing* 29.6, pp. 1069–1103.
- Thompson-Schill, Sharon L. et al. (1997). “Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation”. In: *Proceedings of the National Academy of Sciences* 94.26, pp. 14792–14797.
- Thürer, Daniel and Thomas Burri (2006). “Zum Sprachenrecht der Schweiz”. In: *Zur Entstehung des modernen Minderheitenschutzes in Europa*. Ed. by Christoph Pan and Beate Sibylle Pfeil. Vienna: Springer, pp. 242–266.
- To, Vinh, Si Fan, and Damon Thomas (2013). “Lexical density and readability: A case study of English textbooks”. In: *Internet Journal of Language, Culture and Society* 37, pp. 61–71.
- Todirascu, Amalia et al. (2013). “Coherence and cohesion for the assessment of text readability”. In: *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science*. Marseille, pp. 11–19.
- Torres Irribarra, David and Rebecca Freund (2014). *Wright Map: IRT item-person map with ConQuest integration*. URL: <https://github.com/david-ti/wrightmap> (last accessed on 07/19/2024).
- Toyama, Yukie (2018). “Item difficulty modeling of reading comprehension items using the Rasch latent regression Linear Logistic Test Models”. In: *Rasch Measurement Transactions* 32.1, pp. 1672–1676.
- (2019). “What Makes Reading Difficult? An Investigation of the Contribution of Passage, Task, and Reader Characteristics on Item Difficulty, Using Explanatory Item Response Models”. Doctoral thesis. University of California, Berkeley.
- Trace, Jonathan et al. (2017). “Determining cloze item difficulty from item and passage characteristics across different learner backgrounds”. In: *Language Testing* 34.2, pp. 151–174.
- Tunmer, William E., Michael L. Herriman, and Andrew R. Nesdale (1988). “Metalinguistic abilities and beginning reading”. In: *Reading Research Quarterly* 23.2, pp. 134–158.
- Urhahne, Detlef and Lisette Wijnia (2021). “A review on the accuracy of teacher judgments”. In: *Educational Research Review* 32, p. 100374.
- Urquhart, Sandy and Cyril J. Weir (1998). *Reading in a Second Language: Process, Product and Practice*. London, New York: Routledge.
- Ushey, Kevin (2023). *renv: Project Environments*. URL: <https://CRAN.R-project.org/package=renv> (last accessed on 04/16/2023).
- Vallerand, Robert J. (1989). “Vers une méthodologie de validation trans-culturelle de questionnaires psychologiques: Implications pour la recherche en langue française”. In: *Canadian Psychology / Psychologie canadienne* 30, pp. 662–680.
- Van de Vijver, Fons and Ronald K. Hambleton (1996). “Translating tests: Some practical guidelines”. In: *European Psychologist* 1.2, pp. 89–99.
- Van de Vijver, Fons and Norbert K. Tanzer (2004). “Bias and equivalence in cross-cultural assessment: an overview”. In: *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology* 54.2, pp. 119–135.
- Van Gorp, Koen and Bart Deygers (2014). “Task-based language assessment”. In: *The Companion to Language Assessment*. Ed. by Anthony John Kunnan. Vol. 2. Chichester: Wiley Blackwell, pp. 578–593.
- Vanhove, Jan (2014). “Receptive Multilingualism Across the Lifespan”. Doctoral thesis. University of Fribourg/Switzerland.
- Vanhove, Jan and Raphael Berthele (2017a). “Interactions between formal distance and participant-related variables in receptive multilingualism”. In: *International Review of Applied Linguistics in Language Teaching* 55.1, pp. 23–40.

- (2017b). “Item-related determinants of cognate guessing in multilinguals”. In: *Crosslinguistic Influence and Crosslinguistic Interaction in Multilingual Language Learning*. Ed. by Gessica De Angelis, Ulrike Jessner, and Marijana Kresić. London, New York: Bloomsbury, pp. 95–118.
- Vanhove, Jan, Audrey Bonvin, et al. (2019). “Predicting perceptions of the lexical richness of short French, German, and Portuguese texts using text-based indices”. In: *Journal of Writing Research* 10.3, pp. 499–525.
- Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and Computing* 27.5, pp. 1413–1432.
- Verhoeven, Ludo, Marinus Voeten, and Anne Vermeer (2019). “Beyond the simple view of early first and second language reading: The impact of lexical quality”. In: *Journal of Neurolinguistics* 50, pp. 28–36.
- Verner, Martin and Laura Helbling (2019). *Sampling ÜGK 2017. Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017*. URL: https://www.uegk-schweiz.ch/wp-content/uploads/2019/05/%C3%9CGK2017_Verner_Helbling_2019_Sampling_%C3%9CGK_2017.pdf (last accessed on 07/12/2022).
- Vidal, Karina (2011). “A comparison of the effects of reading and listening on incidental vocabulary acquisition”. In: *Language Learning* 61.1, pp. 219–258.
- Waddington, Christopher (2001). “Different methods of evaluating student translations: The question of validity”. In: *Meta : journal des traducteurs / Meta: Translators’ Journal* 46.2, pp. 311–325.
- Walczyk, Jeffrey J. et al. (2001). “Overcoming inefficient reading skills”. In: *Journal of Educational Psychology* 93.4, pp. 750–757.
- Watering, Gerard van de and Janine van der Rijt (2006). “Teachers’ and students’ perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items”. In: *Educational Research Review* 1.2, pp. 133–147.
- Wauters, Kelly, Piet Desmet, and Wim Van Den Noortgate (2012). “Item difficulty estimation: An auspicious collaboration between data and judgment”. In: *Computers & Education* 58.4, pp. 1183–1193.
- Weinert, Franz E. (1999). *Concepts of Competence: Definition and Selection of Competencies*. Ed. by SFSO, NCES & OECD.
- Weir, Cyril J. (2005). *Language Testing and Validation. An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.
- Weiss, Zarah, Xiaobin Chen, and Detmar Meurers (2021). “Using broad linguistic complexity modeling for cross-lingual readability assessment”. In: *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*. Online: LiU Electronic Press, pp. 38–54.
- Weiss, Zarah and Detmar Meurers (2018). “Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe: Association for Computational Linguistics, pp. 303–317.
- (2019a). “Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school”. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence: Association for Computational Linguistics, pp. 380–393.
- (2019b). “Broad linguistic modeling is beneficial for German L2 proficiency assessment”. In: *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*. Ed. by Andrea Abel et al. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 419–435.

- White, Sheida (2012). “Mining the text: 34 text features that can ease or obstruct text comprehension and use”. In: *Literacy Research and Instruction* 51.2, pp. 143–164.
- Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.
- Wickham, Hadley et al. (2019). “Welcome to the Tidyverse”. In: *Journal of Open Source Software* 4.43, p. 1686.
- Wikipédia, ed. (2025). *Langues en Suisse*. URL: https://fr.wikipedia.org/w/index.php?title=Langues_en_Suisse&oldid=226263000 (last accessed on 08/05/2025).
- Wiktionary (2022). *Appendix: German cognates with English*. URL: https://en.wiktionary.org/w/index.php?title=Appendix:German_cognates_with_English&oldid=68797583 (last accessed on 06/26/2023).
- Williams, Joanna P. (2007). “Literacy in the curriculum: Integrating text structure and content area instruction”. In: *Reading Comprehension Strategies: Theories, Interventions, and Technologies*. Ed. by Danielle S. McNamara. New York, London: Lawrence Erlbaum, pp. 199–219.
- Wilson, Mark, Paul De Boeck, and Claus H. Carstensen (2008). “Explanatory item response models: A brief introduction”. In: *Assessment of Competencies in Educational Contexts*. Ed. by Eckhard Klieme, Detlev Leutner, and Johannes Hartig. Göttingen: Hogrefe & Huber, pp. 83–110.
- Wisniewski, Katrin (2012). “Lexikalische Kompetenzen in der Fremdsprache testen: ein Modellierungsansatz”. In: *KOLIPSI II. Gli studenti altoatesini e la seconda lingua: indagine linguistica e psicosociale. / Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung*. Ed. by Chiara Vettori and Andrea Abel. Bolzano: EURAC research, pp. 23–48.
- Wisniewski, Katrin et al. (2013). “MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data”. In: *ICT for Language Learning. Conference Proceedings 2013*.
- Wu, Margaret (2022). *A Course on Test and Item Analyses*. URL: <https://www.edmeasurement-surveys.com/> (last accessed on 12/20/2022).
- Wu, Winston and David Yarowsky (2018). “Creating large-scale multilingual cognate tables”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), pp. 3411–3418.
- Xie, Xuping (2005). “The influence of schema theory on foreign language reading comprehension”. In: *The English Teacher* 34, pp. 67–75.
- Yarkoni, Tal, David Balota, and Melvin Yap (2008). “Moving beyond Coltheart’s N: A new measure of orthographic similarity”. In: *Psychonomic Bulletin & Review* 15.5, pp. 971–979.
- Yen, Wendy M. (1984). “Effects of local item dependence on the fit and equating performance of the three-parameter logistic model”. In: *Applied Psychological Measurement* 8.2, pp. 125–145.
- Zaccoletti, Sonia, Gianmarco Altoè, and Lucia Mason (2020). “The interplay of reading-related emotions and updating in reading comprehension performance”. In: *British Journal of Educational Psychology* 90.3, pp. 663–682.
- Zenisky, April L., Ronald K. Hambleton, and Frederic Robin (2004). “DIF detection and interpretation in large-scale science assessments: informing item writing practices”. In: *Educational Assessment* 9.1-2, pp. 61–78.
- Zhang, Zhiyong and Lijuan Wang (2017). *Advanced Statistics Using R*. Granger: ISDSA Press.
- Zhao, Xueyu and Guillermo Solano-Flores (2021). “Testing across languages in international comparisons: cultural adaptation of consensus-based test translation review procedures”. In: *Journal of Multilingual and Multicultural Development* 42.7, pp. 677–691.
- Zhao, Xueyu, Guillermo Solano-Flores, and Ming Qian (2018). “International test comparisons: Reviewing translation error in different source language-target language combinations”. In: *International Multilingual Research Journal* 12.1, pp. 17–27.

- Zhu, Xuelian and Vahid Aryadoust (2019). “Examining test fairness across gender in a computerised reading test: A comparison between the Rasch-based DIF technique and MIMIC”. In: *Papers in Language Testing and Assessment* 8.2, pp. 65–90.
- Zumbo, Bruno D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF)*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- (2003). “Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests”. In: *Language Testing* 20.2, pp. 136–147.

Glossary

- ÜGK** Überprüfung der Erreichung der Grundkompetenzen 18, 19, 20, 22, 23, 26, 27, 28, 29, 30, 33, 34, 39, 46, 50, 51, 53, 59, 70, 71, 74, 75, 77, 89, 90, 93, 131, 133, 134, 135, 136, 137, 138, 139, 142, 143, 144, 145, 150, 153, 157, 158, 160, 161, 162, 163, 164, 165, 167, 168, 184, 186, 191, 192, 193, 205, 208, 221, 222, 223, 224, 225, 228, 230, 231, 232, 236, 292, 293, 300, 301
- ACD** Automatic Cognate Detection 75
- ACTFL** American Council on the Teaching of Foreign Languages 66, 67
- ADB** Aufgabendatenbank der EDK 145, 164, 166, 300
- AFA** Autoconcepto Forma A 98
- AI** Artificial Intelligence 235
- AIC** Akaike information criterion 170
- ALTE** Association of Language Testers in Europe 30
- ANFIS** Adaptive Neuro-Fuzzy Inference System 110, 125
- ANN** Artificial Neural Network 108, 109, 110, 114, 125
- ANOVA** Analysis of variance 170
- aQ3** Adjusted Q3 283
- BIC** Bayesian information criterion 170
- BLUP** Best Linear Unbiased Predictor 213
- BNC** British National Corpus 151
- CART** Classification and Regression Tree 108, 109, 110, 125, 214
- CAT** Computer-Adaptive Test 130
- CD** Contextual Diversity 72
- CDA** Cognitive Diagnostic Assessment 103, 104, 107, 108, 109, 110, 120
- C-I** Construction-Integration 33, 34, 35, 37, 40, 42, 45, 51, 52, 110, 114, 119
- CEFR** Common European Framework of Reference for Languages 27, 28, 67, 72, 73, 91, 93, 137, 152, 155, 156, 223, 224, 229
- COFO** Vérification de l'atteinte des compétences fondamentales 19

- CTAP** Common Text Analysis Platform 57, 58
- CTT** Classical Test Theory 22, 23, 78, 79, 80, 81, 82, 84, 86, 89, 91, 95, 166, 168, 173, 184, 187, 192, 194, 225, 294
- CWI** Complex Word Identification 73
- DESI** Deutsch Englisch Schülerleistungen International 56
- DIF** Differential Item Functioning 20, 21, 63, 79, 84, 85, 86, 87, 89, 91, 92, 93, 95, 96, 97, 98, 99, 100, 101, 102, 114, 121, 125, 129, 130, 131, 145, 166, 167, 168, 190, 194, 221, 230, 234, 238, 297
- DLT** Dependency Locality Theory 58
- EAP** Expected A Posteriori 82, 188
- EBAFLS** European Bank of Anchor Items for Foreign Language Skills 91, 92
- EDK** Schweizerische Konferenz der kantonalen Erziehungsdirektoren 135
- EFA** Exploratory Factor Analysis 203
- EFL** English as a Foreign Language 66, 108
- ERP** Event-Related Potential 33
- ESL** English as a Second language 130
- ETS** Educational Testing Service 80, 103, 104, 111
- FLE** French as a Second language 57
- fMRI** Functional Magnetic Resonance Imaging 33
- GDP** gross domestic product 99
- GLMM** Generalized Linear Mixed Model 88, 163, 216
- HarmoS** Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule 20, 134, 136, 137
- HLM** Hierarchical Linear Modelling 107, 124
- ICC** Item Characteristic Curve 82, 83, 84, 85, 86
- IELTS** International English Language Testing System 125
- IoM** Institute of Multilingualism 139, 143, 144
- IRT** Item Response Theory 22, 23, 78, 79, 82, 83, 84, 85, 86, 88, 89, 95, 107, 108, 123, 127, 130, 163, 166, 168, 173, 188, 189, 192, 193, 194, 215, 225, 294
- ITC** International Test Commission 93
- KMO** Kaiser-Meyer-Olkin 169, 203
- LLTM** Linear Logistic Trait Model 89, 107, 235, 295, 296

- LMG** Lindeman-Merenda-Gold 170, 208, 209
- LOO** leave-one-out 171, 215, 217, 295, 296
- LQ** Lexical Quality 65
- LQH** Lexical Quality Hypothesis 65
- LR-LLTM** Latent Regression-Linear Logistic Trait Model 89, 107, 124
- LSA** Latent Semantic Analysis 55, 58, 59, 69, 116, 119
- MC** Multiple-Choice 116, 117, 141, 149, 158, 169, 184, 186, 187, 197
- MCQ** Multiple-Choice Question 30, 53, 60, 61, 62, 66, 100, 116, 117, 123, 126, 138, 139, 141, 143, 149, 158, 174, 175, 186, 193, 197, 283, 301
- MDS** Weighted Multidimensional Scaling Analysis 98
- MELAB** Michigan English Language Assessment Battery 109, 111
- MERLIN** Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context 67
- MET** Michigan English Test 109
- MFR** Many-facet Rasch 163, 166, 167, 170, 171, 173, 184, 188, 189, 191, 192, 193, 194, 211, 213, 215, 216, 282, 283, 295, 296, 299
- MH** Mantel Haenszel 86
- MLP** Multilayer Perceptron 109, 125
- MML** Marginal Maximum Likelihood 166
- MSE** mean squared error 171
- MTLD** Measure of Textual Lexical Diversity 68, 69, 147, 153, 154, 180, 209, 226, 227, 228, 281
- MUC** Memory, Unification, Control 37, 38
- NI** Necessary Information 62, 112, 113, 115, 116, 117, 119, 120, 121, 122, 123, 124, 125, 127, 128, 131, 149, 150, 151, 154, 155, 156, 157, 158, 159, 160, 161, 162, 165, 169, 170, 174, 176, 178, 180, 181, 183, 184, 186, 194, 195, 197, 198, 199, 201, 202, 208, 209, 210, 211, 212, 214, 215, 216, 217, 218, 219, 221, 222, 223, 224, 225, 227, 230, 231, 232, 233, 234, 238, 294, 297, 301
- NLP** Natural Language Processing 45, 47, 57, 73, 103, 110, 116, 151, 226
- OM** Ontogenesis Model of the L2 Lexical Representation 65
- PCA** Principal Component Analysis 58, 101, 283
- PCM** Partial Credit Model 296
- PET** Positron Emission Tomography 33
- PET** Psychometric Entrance Test 100

PIRLS Progress in International Reading Literacy Study 99

PISA Programme for International Student Assessment 18, 19, 20, 21, 30, 53, 56, 63, 85, 93, 99, 101, 113, 162

POS Part of Speech 59, 73, 116, 151, 152, 153, 154, 161, 162, 165, 169, 170, 173, 176, 177, 179, 197, 198, 199, 200, 201, 202, 203, 204, 209, 210, 214, 218, 219, 222, 223, 224, 226, 228, 233, 235, 236, 298, 299

PTE Pearson Test of English 120

PV Plausible Value 29, 144, 145, 150, 163, 167, 193, 300

QOL Quality of Life 97

RLD Reference Level Description 72, 73

RMSE Root Mean Squared Error 171, 211, 214, 219, 288

RMSEA Root Mean Squared Error of Approximation 203, 204

RPT Reading Proficiency Test 67

RQ Research Question 22, 221, 222, 223, 224, 225

RSF Reading Systems Framework 33, 34, 37, 38, 40, 42, 51

RUM Reparameterized Unified Model 108

SAIP School Achievement Indicators Program 101

SAQ Short Answer Question 116, 117, 143

SAT Stanford Achievement Test 61

SD Standard Deviation 163, 167, 215

SEM Structural Equation Modeling 98

SES socioeconomic status 62, 88, 125

SLA Second Language Acquisition 75, 236

STTS Stuttgart Tübingen Tag Set 151

SVR Simple View of Reading 33, 34, 35, 36, 37, 41, 49, 51, 66

TBLT Task-Based Language Teaching 30

TIMMS Trends in International Mathematics and Science Study 99

TLI Tucker-Lewis Index 203, 204

TLU Target Language Use 229

TOEFL Test of English as a Foreign Language 21, 61, 63, 74, 91, 103, 104, 105, 108, 111, 114, 117, 119, 121, 122, 129

TOEIC Test of English for International Communication 124

TTR Type-Token Ratio 68, 69, 115, 116, 147, 152, 227

TTTE Theory of Test Translation Error 96, 97

US United States 61, 66, 113

VeCoF Verifica del raggiungimento delle competenze fondamentali 19

VIF Variance Inflation Factor 171, 208

VL Vocabulary Levels Test 130

WLE Weighted Likelihood Estimate 82, 166, 167, 188, 192, 193, 294, 296

Appendix A

Use of Artificial Intelligence

A.1 Erklärung zur Nutzung von künstlicher Intelligenz

Im Verlauf der Arbeit wurde das KI-System ChatGPT (OpenAI, GPT 3 und 4) punktuell eingesetzt. Die KI unterstützte insbesondere bei der Entwicklung des Codes in R, z. B. für die Umstrukturierung von Datensätzen oder die Erstellung von Grafiken mit ggplot2. Sämtliche Codevorschläge der KI basierten auf konkreten Vorgaben durch die Autorin, wurden von dieser geprüft und selbsttätig in den R-Skripten umgesetzt. Im Laufe der statistischen Analyse wurde die KI ausserdem gelegentlich zu Interpretationen einzelner Ergebnisse befragt, um das Verständnis zu sichern. Bei Widersprüchen wurden externe Quellen hinzugezogen. Die in der Arbeit dargestellten Schlüsse basieren vollständig auf den Überlegungen und Recherchen der Autorin.

Darüber hinaus wurde die KI auch eingesetzt, um den Schreibprozess zu unterstützen. Dabei umfasste der Einsatz der KI vor allem Formulierungsvorschläge, das Prüfen der Idiomatik im Englischen sowie gelegentliche Ideenfindung bei der Strukturierung von Textabschnitten. Sämtliche inhaltlichen Entscheidungen, Analysen und Interpretationen basieren dabei weiterhin vollständig auf der eigenen wissenschaftlichen Arbeit der Autorin. Die generierten Vorschläge wurden kritisch geprüft, in den allermeisten Fällen angepasst und in den Gesamtkontext der Arbeit integriert. Der theoretische Teil wurde vollständig ohne Einsatz der KI geschrieben und formuliert, sämtliche Literaturrecherchen wurden ebenfalls ohne KI-Unterstützung durchgeführt.

Zu keinem Zeitpunkt wurden die Datensätze, die R-Skripte oder der Text als Ganzes für die KI zugänglich gemacht. Wenn Auszüge zur Verfügung gestellt wurden, wurden identifizierende Informationen zuvor entfernt. Die Einstellungen des Accounts sind ausserdem so gewählt, dass die Eingaben nicht für das weitere Training der KI genutzt werden können.

A.2 Declaration on the Use of Artificial Intelligence

During the preparation of this work, the AI system ChatGPT (OpenAI, GPT-3 and GPT-4) was used on a limited basis. Its primary role was to assist with R programming tasks, such as restructuring datasets and creating visualisations using ggplot2. All code suggestions generated by the AI were based on explicit instructions from the author, carefully reviewed, and manually implemented in the R scripts. During the course of the statistical analysis, the AI was occasionally consulted to clarify the interpretation of individual results, with the aim of ensuring a sound understanding. In cases of inconsistency, external sources were consulted. The conclusions presented in this work are based entirely on the author's own reasoning and research.

The AI was also employed to support the writing process, particularly for phrasing suggestions, checking the idiomatic use of English, and, on occasion, generating ideas for structuring text sections. All substantive decisions, analyses, and interpretations remain wholly grounded in the author's independent scholarly work. The suggestions provided by the AI were critically assessed, in most cases adapted, and integrated into the overall context of the thesis. The theoret-

ical section was written entirely without AI assistance, and all literature searches were conducted independently by the author.

At no point were complete datasets, R scripts, or the full text of the work made accessible to the AI. Where excerpts were provided, all identifying information was removed beforehand. Furthermore, the account settings are configured to ensure that inputs are not used for further training of the AI.

(This translation was written by Chat-GPT 5, though.)

Appendix B

Additional results

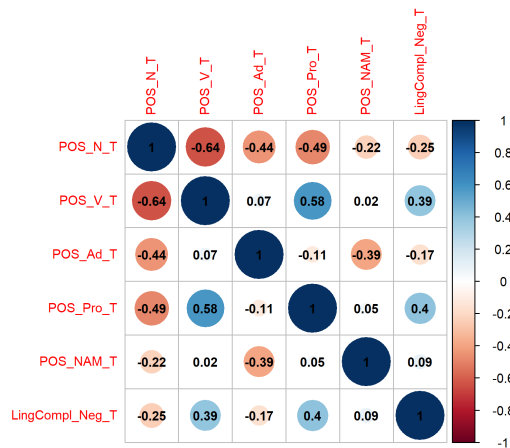
B.1 Descriptive analyses

B.1.1 Correlations of incidences of POS across language versions

The three graphs below are a graphic depiction of a correlation matrix: for each combination of variables, a correlation coefficient was calculated (Pearson's r). The correlation coefficient represents the strength and direction of the linear relationship between the two variables. In the plot, large, dark circles represent a strong correlation between the two variables, small light, or no circles represent a weak or no relationship. When the circles are blue, a positive relationship is indicated, which means that if the value of one variable increases, so will the value of the other. Red circles represent the inverse: if the value of one variable increases, the value of the other will decrease.

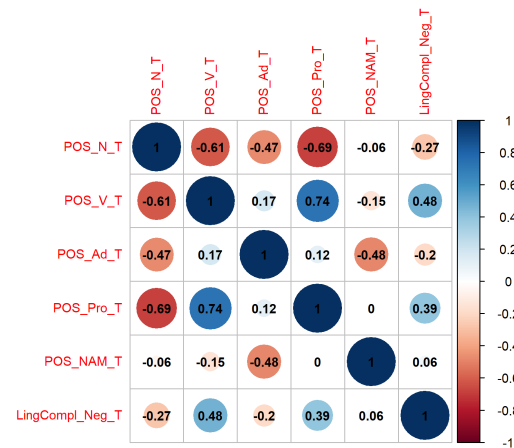
Note that this graph is drawn irrespective of whether the relationship between two variables is statistically significant, or whether the two values can be considered to have a linear relationship at all. All weaknesses of the Pearson's correlation coefficient apply.

POS variables of text length - Language Version: DEfr



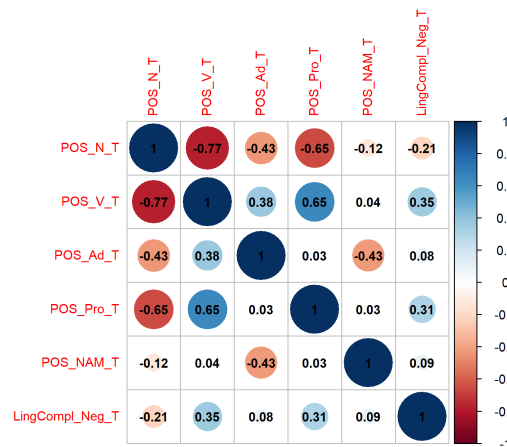
(a) German input texts

POS variables of text length - Language Version: ENde



(b) English input texts

POS variables of text length - Language Version: FRde



(c) French input texts

Figure B.1: Correlations of incidences of parts of speech in input texts

B.1.2 Descriptive statistics of lexical variables in input texts

Note: In the table below, the variable names are displayed using a dash (-) instead of an underscore (_) for technical reasons. This does not change their meaning. The lexical difficulty variables represent the proportion of A1 or A2 words among all content words of each text.

Variable	Lg	N	Min	Max	Median	Mean	SD
LexDiv-Var	German	38	43.56	286.72	90.16	105.519	53.192
LexDiv-Eve	German	39	0.32	3.323	1.279	1.363	0.724
LexDispersion	German	39	0	0.307	0.107	0.11	0.062

Variable	Lg	N	Min	Max	Median	Mean	SD
LexDisparity-form	German	39	0.098	0.149	0.131	0.129	0.013
LexDens-contentW	German	39	11	121	52	55.692	32.955
LexDens	German	39	37.143	73.171	54.545	54.761	6.694
LexSoph-tokens	German	39	4.992	6.862	6.175	6.125	0.486
LexSoph-content	German	39	233.294	3373.251	1510.207	1593.543	800.774
LexSoph-low-freq	German	39	-1.025	4.881	2.169	1.982	1.258
LexDiff-Deca-A1	German	39	0.294	0.773	0.609	0.589	0.122
LexDiff-CEFRlex-A1	German	39	0.667	0.965	0.816	0.819	0.084
LexDiff-Deca-A2	German	39	0	0.6	0.132	0.14	0.098
LexDiff-CEFRlex-A2	German	39	0	0.238	0.057	0.069	0.054
LexDiv-Var	English	38	38.725	117.03	70.359	70.407	23.306
LexDiv-Eve	English	39	0.209	3.643	1.325	1.418	0.839
LexDispersion	English	39	0	0.273	0.111	0.134	0.074
LexDisparity-form	English	39	0.078	0.123	0.107	0.104	0.011
LexDens-contentW	English	39	13	121	55	59.205	34.749
LexDens	English	39	35.333	73.469	55.172	54.456	7.311
LexSoph-tokens	English	39	5.412	7.929	6.966	6.871	0.54
LexSoph-content	English	39	110.647	1523.206	740.749	785.924	352.952
LexSoph-low-freq	English	39	1.328	5.665	3.205	3.279	0.887
LexDiff-Deca-A1	English	39	0.263	0.701	0.5	0.503	0.091
LexDiff-CEFRlex-A1	English	39	0.462	0.8	0.667	0.655	0.076
LexDiff-Deca-A2	English	39	0	0.526	0.136	0.145	0.085
LexDiff-CEFRlex-A2	English	39	0	0.154	0.073	0.072	0.04
LexDiv-Var	French	39	42.063	246.96	80.92	90.038	41.113
LexDiv-Eve	French	39	0.465	3.555	1.553	1.552	0.805

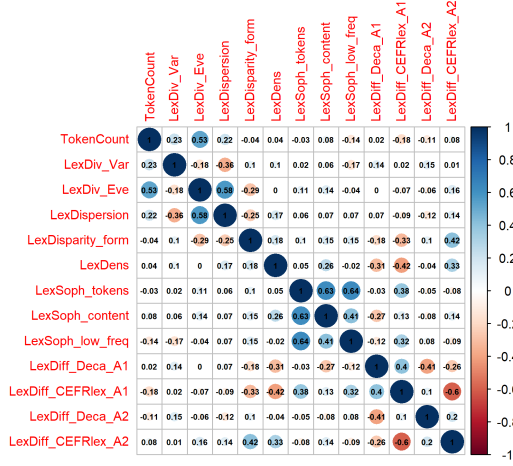
Variable	Lg	N	Min	Max	Median	Mean	SD
LexDispersion	French	39	0.019	0.317	0.091	0.101	0.06
LexDisparity-form	French	39	0.101	0.145	0.125	0.125	0.009
LexDens-contentW	French	39	15	133	55	61.718	36.267
LexDens	French	39	34.058	73.81	53.659	53.379	7.869
LexSoph-tokens	French	39	5.563	7.405	6.401	6.47	0.496
LexSoph-content	French	39	36.9	4228.952	1220.694	1212.186	779.925
LexSoph-low-freq	French	39	-0.777	5.108	2.594	2.548	1.24
LexDiff-Deca-A1	French	39	0.333	0.705	0.529	0.516	0.091
LexDiff-CEFRlex-A1	French	39	0.647	0.938	0.81	0.817	0.061
LexDiff-CEFRlex-Beacco-A1	French	39	0.412	0.867	0.722	0.712	0.096
LexDiff-Deca-A2	French	39	0.029	0.429	0.127	0.131	0.076
LexDiff-CEFRlex-A2	French	39	0	0.143	0.023	0.034	0.039
LexDiff-CEFRlex-Beacco-A2	French	39	0	0.143	0.038	0.047	0.038
LexDiff-CEFRlex-Beacco-A1	French	39	7	107	37	44.846	28.305
LexDiff-CEFRlex-Beacco-A2	French	39	0	10	2	2.718	2.533

Table B.1: Descriptive statistics of lexical variables in input texts per language version

B.1.3 Correlations of lexical variables and token counts across languages

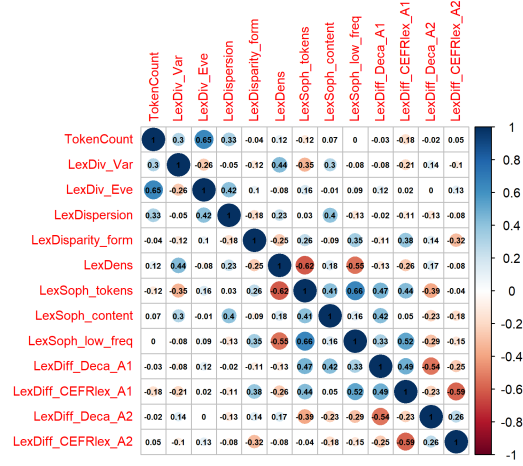
See for information on correlations and how to read the graphs section B.1.1.

Necessary information: Lexical variables and text length
Language Version: DEfr



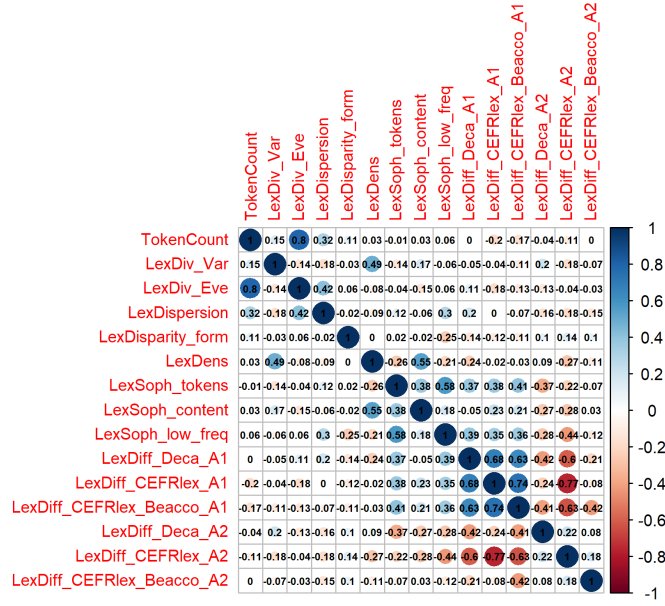
(a) German input texts

Necessary information: Lexical variables and text length
Language Version: ENde



(b) English input texts

Necessary information: Lexical variables and text length
Language Version: FRde



(c) French input texts

Figure B.2: Correlations of token count and lexical variables in NIs

B.1.4 Lexical difficulty and other linguistic characteristics of input texts

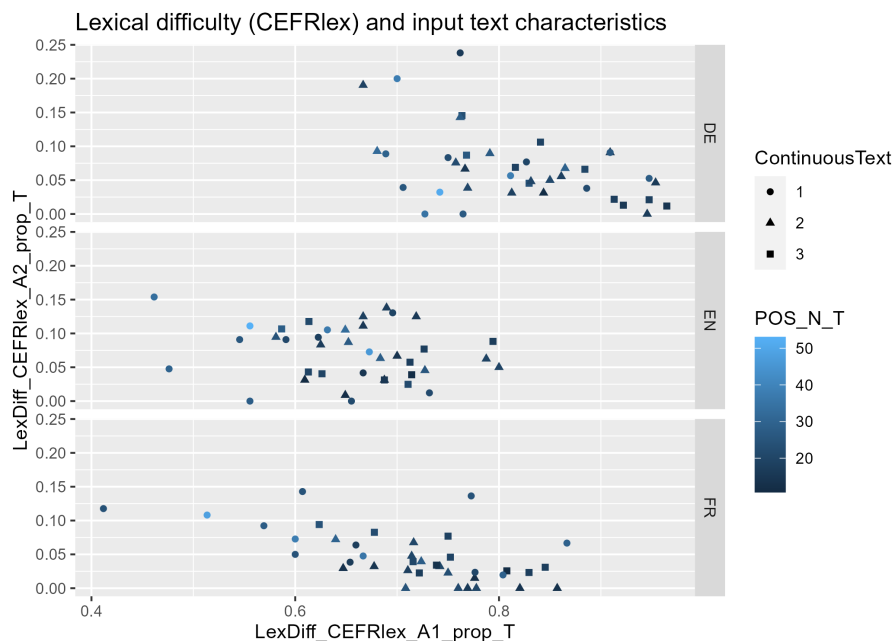


Figure B.3: Scatterplots of CEFRlex lexical difficulty variables, `POS_N` and `ContinuousText` across languages

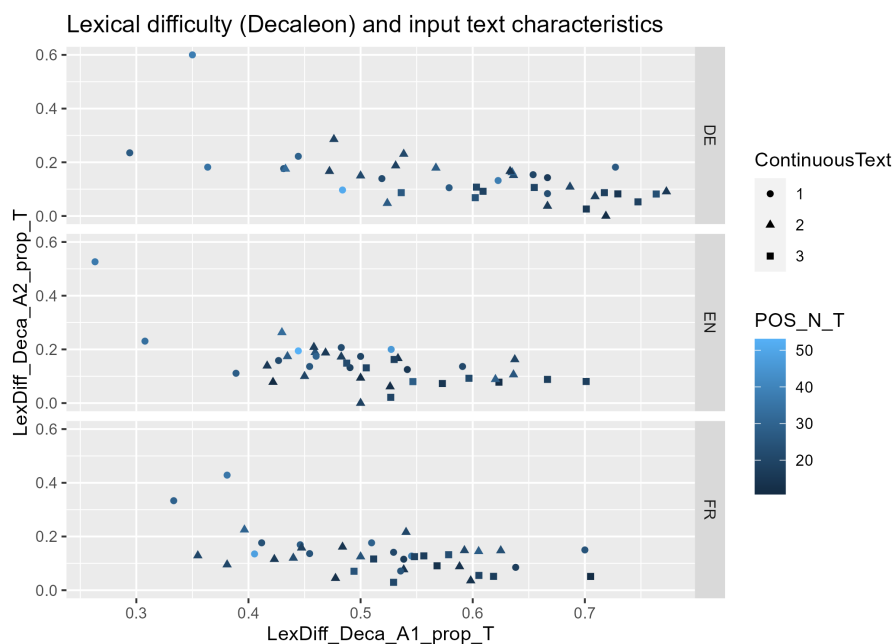


Figure B.4: Scatterplots of Decaleon lexical difficulty variables, `POS_N` and `ContinuousText` across languages

B.1.5 Lexical diversity: MTLD across input texts and language versions

Two graphs are shown: the first plots the MTLDs of the input texts in the three foreign language versions, their distributions and bivariate scatterplots. The second plot shows the MTLDs of each input text in the three foreign languages. This shows in which texts the values differ from each other and where they do not.

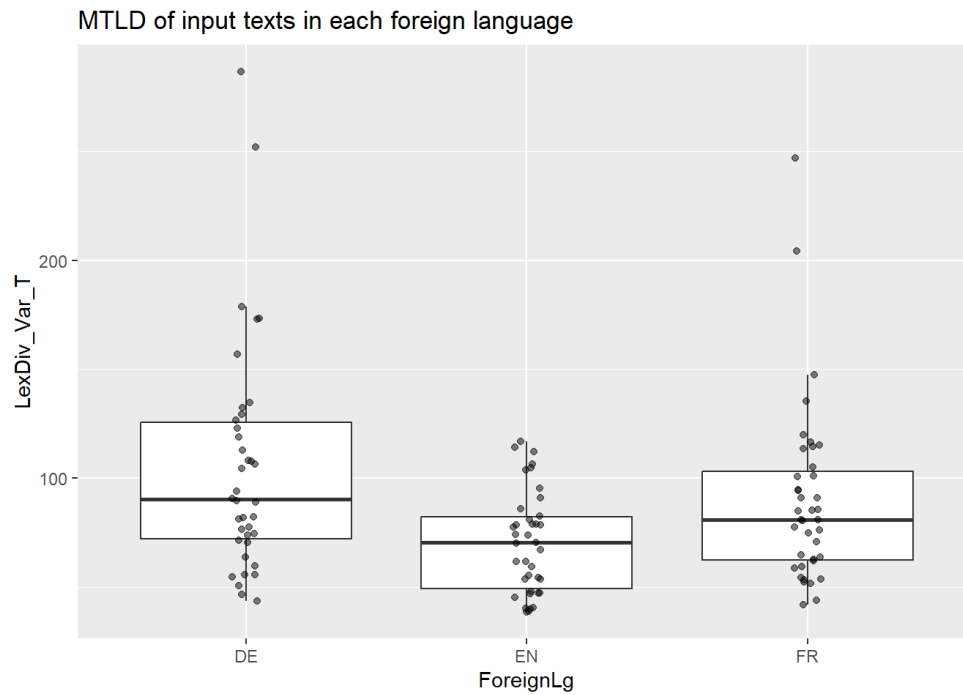


Figure B.5: Boxplots of MTLDs of input texts in German (DE), English (EN) and French (FR)



Figure B.6: MTLT by input text and language version

B.1.6 MFR: model fit

Both item fit and person fit of the MFR model's parameter were checked using TAM's fit functions (Robitzsch, Kiefer, and M. Wu 2022, see the R script `2_b_VariableExtraction_rsd` for more details and comments). They were deemed acceptable: the same items that had been identified to misfit in the Rasch models were also flagged in the MFR model. Since their presence in the data set did not change anything in terms of item difficulty, they were left in the model. Person fit was also checked, but only a small number of test takers showed significant misfit (only 23 individuals had t-values above 3 or below -3, which translated into infit values of 1.40 and above), so no changes were made to the original data set.

Additionally, model fit statistics were extracted using the package's `tam.modelfit()` function.

It outputs a variety of fit statistics, including *Yen's Q3* statistic (1984) and effect sizes of model fit (e.g. MADaQ3, SRMS). Most of the extracted fit statistics suggest a moderate overall fit ($100 * MADCOV = 1.652$, $SRMR = 0.083$, and $SRMSR = 0.107$). The MADaQ3 value was 0.07, with a maximum Adjusted Q3 (aQ3) of 0.56 ($p < .001$)²⁵⁹. Since aQ3 values should be close to 0, item pairs which showed high aQ3 values were extracted. All of them were items from the FRit language version. Since this language version was solved by the lowest number of test takers, it is plausible that its aQ3 values do not fit the overall MFR model well. Nevertheless, a PCA on the residuals of the FRit language version was added to the analyses to detect a possible unexpected factor structure. The analysis revealed two dominant components explaining $\approx 15\%$ of variance each, suggesting potential structured residuals. A more detailed analysis of the items which loaded most strongly on the two factors did not yield any particular reason for their clustering. If anything, the items on the first cluster were all Matching items or MCQ with more than three answer options. In the second cluster, many items were particularly easy in Italian compared to the other language versions. A plot of item difficulties and the loadings on this second cluster confirm this tentative finding. Following Fan and Bond (2019, pp. 96 sq.) however, since no “reasonable substantive interpretation” (ibid., p. 97) could be formulated, these findings are attributed to the lower number of test takers in the FRit version and the resulting interdependence of the item difficulties rather than to any detectable violation of the unidimensionality requirement of Rasch modelling.

B.1.7 Wright Maps: Item difficulties per language version

Wright Maps depict item difficulties and test taker abilities of an entire test in the same graph. On the left side, the barplot shows the distribution of the respondents' θ values. Bars which are higher up indicate higher ability levels. On the right side, the items are listed by numbers at their respective difficulty level. Higher positions of numbers indicate more difficult items. If the majority of test takers and items are approximately on the same level, the test is considered well-suited for the population of test takers.

These Wright Maps show the item difficulties estimated by an MFR model, as explained in section 10.1.2. For each item, its own item difficulty, the aggregated item difficulty of the language version, and a separate item difficulty of each item in each language version are summed up to provide one item difficulty value per item and language version. These correlate highly with item difficulties from separate Rasch models.

²⁵⁹aQ3s residuals are correlations between item residuals after correcting for person ability and item difficulty. If these correlations are very different from 0, certain item pairs diverge from their estimated item difficulty systematically in the same way. This is a violation of the principle of local independence which underlies Rasch modelling (for a more detailed explanation see e.g. J. Fan and Bond 2019).

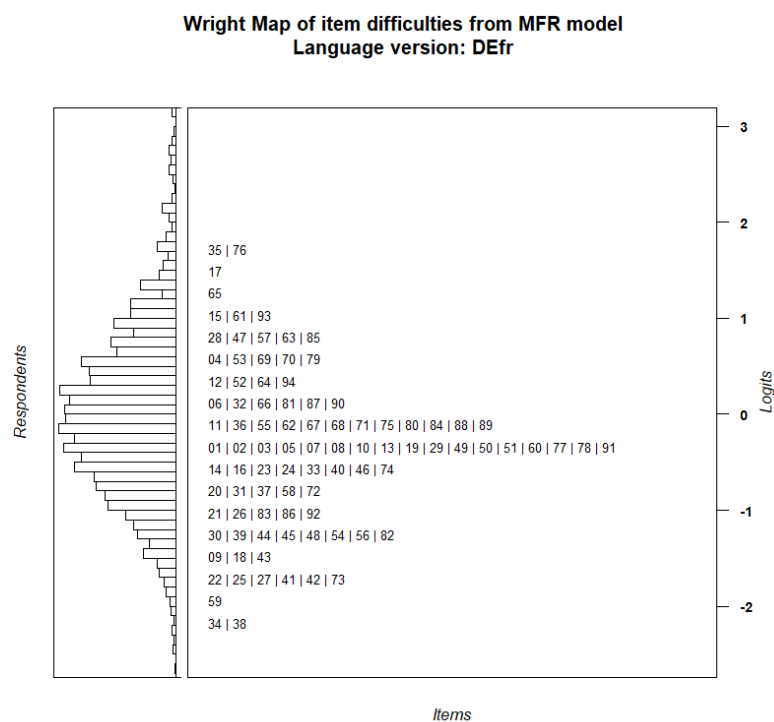


Figure B.7: Wright Map of item difficulties in language version DEfr. *Item difficulties from MFR model. Foreign language: German. Language of schooling: French.*

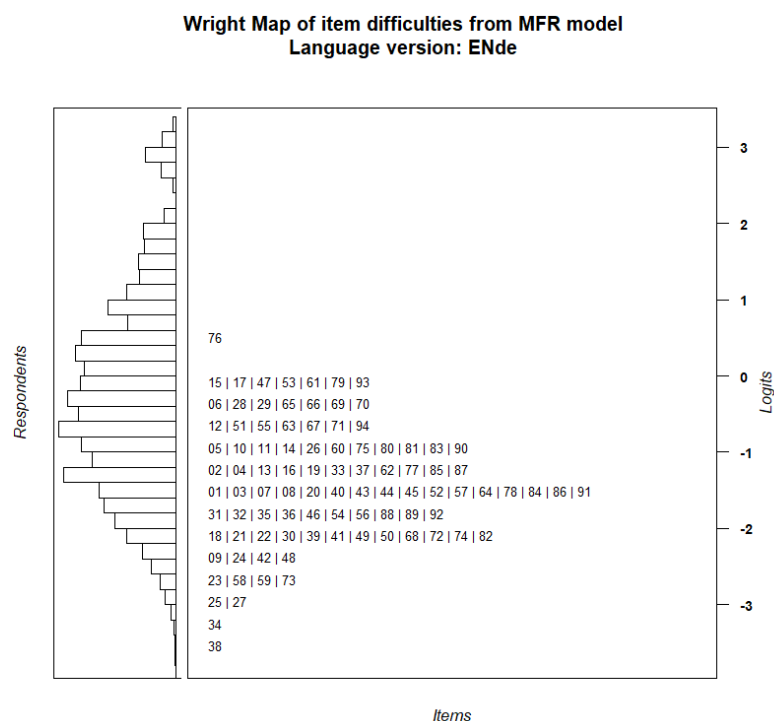


Figure B.8: Wright Map of item difficulties in language version ENde. *Item difficulties from MFR model. Foreign language: English. Language of schooling: German.*

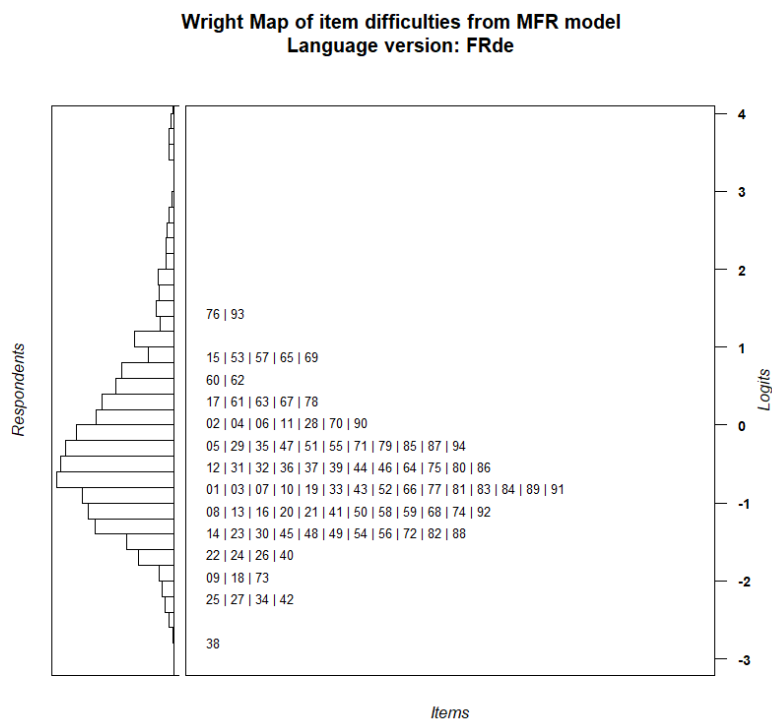


Figure B.9: Wright Map of item difficulties in language version FRde. *Item difficulties from MFR model. Foreign language: French. Language of schooling: German.*

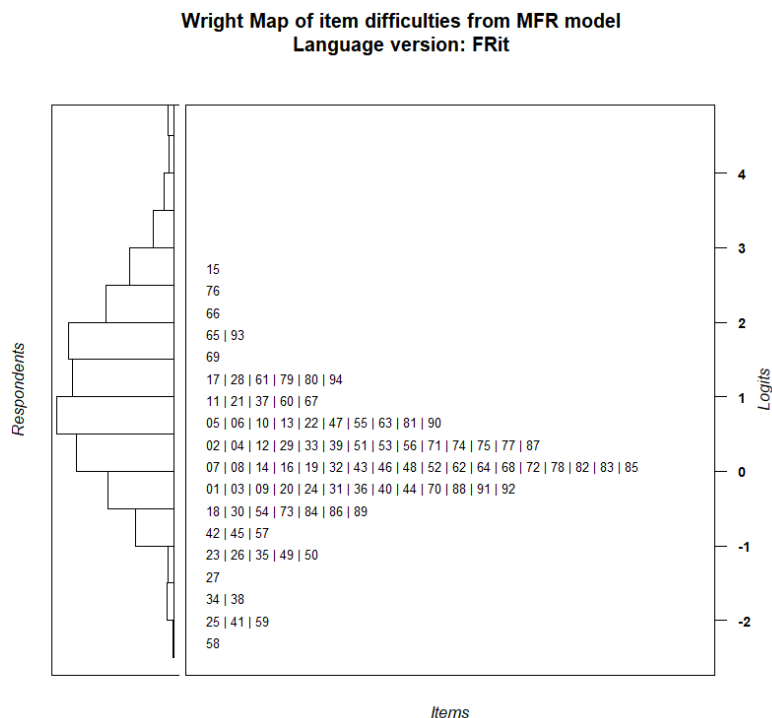


Figure B.10: Wright Map of item difficulties in language version FRit. *Item difficulties from MFR model. Foreign language: French. Language of schooling: Italian.*

B.1.8 Factor analyses of the linguistic characteristics of the input texts

This section contains the factor loadings and important statistical information on the final factor models for the FRit, FRde and DEfr language versions. Cf. section 12.2.1 for more information.

B.1.8.1 FRit language version

Variable	TextLength	POS	LxSophTok	LexDiff	LxSophCon
TokenCount	0.937				
POS_N		-0.658	-0.330		
POS_V		0.991			
POS_Pro		0.509			
LingCompl_Neg	0.312				0.690
LexDiv_Eve	0.923				
LexSoph_tokens			0.966		
LexSoph_content					0.822
LexSoph_low_freq			0.618		
LexDf_CEFRlx_A1				0.738	
LexDf_CEFRlx_A2				-0.824	

Table B.2: Factor loadings from exploratory factor analysis on linguistic characteristics of FRit input texts

B.1.8.2 FRde language version

Variable	LexSoph	CL	POS
POS_N			0.950
POS_V			-0.708
POS_Pro			-0.441
LingCompl_Neg	0.465	-0.314	
LexSoph_tokens	0.893		
LexSoph_content	0.656		
LexSoph_low_freq	0.746		
CL_Cogn		0.987	
CL_Lev		0.822	

Table B.3: Factor loadings from exploratory factor analysis on linguistic characteristics of FRde input texts

B.1.8.3 DEfr language version

Variable	POSLxSoph	CL	TextLength	LexDiv
TokenCount			0.860	
POS_N	-0.535	0.383		
POS_V	0.461	-0.314		
POS_Pro	0.646			0.475
LingCompl_Neg	0.309		0.481	
LexDiv_Var	-0.427			0.605
LexDiv_Eve			0.952	
LexDispersion			0.331	-0.711
LexSoph_tokens	0.765			
LexSoph_content	0.869			
LexSoph_low_freq	0.655		0.326	
CL_Cogn		0.896		
CL_Lev		0.961		

Table B.4: Factor loadings from exploratory factor analysis on linguistic characteristics of DEfr input texts

B.2 Inferential statistics

B.2.1 Regression models of French test data

B.2.1.1 Final model based on FRit data

The final fixed-effects linear model based on the FRit data explains a variance of 34% ($R^2 = .399$, adjusted $R^2 = .336$) and is significant ($F(5, 47) = 6.26$, $p < 0.001$). Its residual standard error is .675. Visual checks of residuals are acceptable. A mixed-effects linear model with `TextID` as a random factor has similar characteristics and explanation of variance.

Predictor	Estimate	2.5% CI	97.5% CI	p-value	rel. imp.
Intercept	-4.824	-7.734	-1.915	.002**	
LexSoph_tokens_T	0.884	0.415	1.353	.0004***	0.41
CL_Cogn_T	-2.114	-4.278	0.050	.055.	0.06
POS_V_NI	-0.026	-0.047	-0.006	.013*	0.16
POS_N_Q	-0.017	-0.033	-0.002	.029*	0.28
POS_Pro_correct	-0.019	-0.044	0.005	.114	0.08

Table B.5: Linear regression model predicting FRit item difficulty. *Note:* . indicates $p < 0.1$, * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, CI = confidence interval, rel. imp. = relative importance according to LMG metric, summed up to 100%.

B.2.1.2 Final model based on FRde data

The final fixed-effects linear model based on the FRde data explains a variance of 24% ($R^2 = .277$, adjusted $R^2 = .243$) and is significant ($F(4, 85) = 8.158$, $p < 0.001$). Its residual standard error is 0.723. Visual checks of residuals are acceptable. This model is based on 90 data points, which is considerably more than the final models in the other languages. It is also the only model which does not contain any item-related variables among its predictors.

Predictor	Estimate	2.5% CI	97.5% CI	p-value	rel. imp.
Intercept	-1.44	-2.478	-.405	.006**	
POS_N_T	0.033	-0.008	.074	.112	0.28
TokenCount_T	0.013	0.006	.02	.0004***	0.29
CL_Cogn_NI	-.614	-1.336	.108	.095.	0.15
POS_N_T:TokenCount_T	-.0005	-0.0008	-0.002	.003**	0.29

Table B.6: Linear regression model predicting FRde item difficulty. The model contains an interaction between two predictors, denoted by a colon (:). *Note:* . indicates $p < 0.1$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, CI = confidence interval, rel. imp. = relative importance according to LMG metric, summed up to 100%.

B.2.2 Decision tree based on random forests analysis

A single regression tree was fit to the dataset using the variables identified in the random forest analysis (see section 12.3.5). It has 22 splits and uses 13 predictors. This tree shows overfitting according to the tree complexity measures obtained from `printcp()`. The optimal tree based on 10-fold cross validation should only have 3 splits ($CP = 0.025$) and would achieve a minimum cross-validated relative error of 0.91 ($SE = 0.07$) and an RMSE of .799. Here, the overfitting tree is shown for illustrative purposes. Its cross-validated relative error is .96 ($SE = 0.07$) and its RMSE is .64.

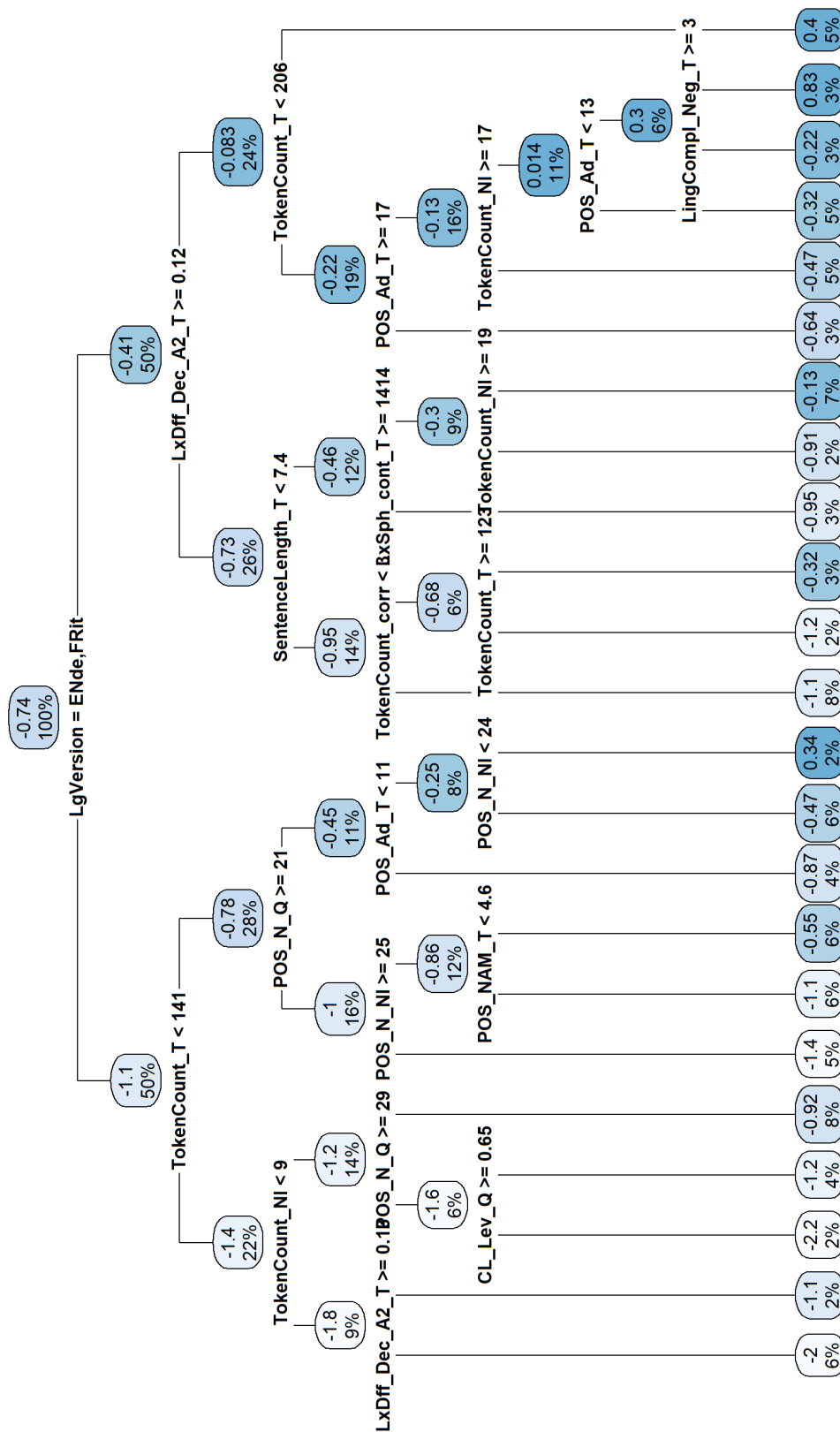


Figure B.11: Single tree based on variables from random forest modelling. *Note:* this tree only represents one possible solution of how item difficulty can be modelled using the variables identified during random forest modelling.

B.2.3 Comparison of item difficulties from Bayesian and parametric IRT models

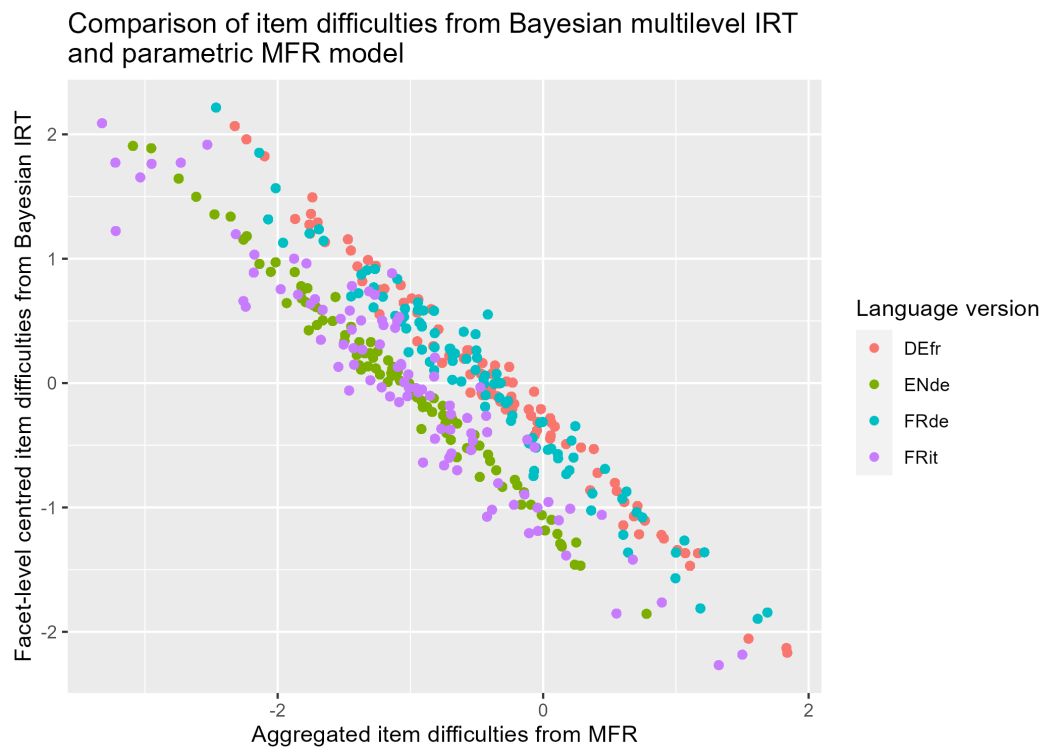


Figure B.12: Comparison of item difficulties from Bayesian and parametric IRT modelling

Appendix C

Electronic appendix: R project

The complete appendix for this thesis contains the original item texts and original item and text IDs. Due to legal requirements, it was only available to the supervising committee of this thesis. The public appendix contains a subsample of the files mentioned here.

C.1 Access to electronic appendices

The public appendix of this thesis can be accessed via https://osf.io/s25xf/?view_only=f6783d0c09a44a0280166b3411bfa68a or the QR code below.



Figure C.1: Access to public appendix

C.2 File index

The following table contains the filenames of all files in the electronic appendices, ordered by their general type and appearance in the analytic process. Dependencies between files are noted. Files which are only available in the non-public part of the electronic appendix are marked as such. All `.Rmd` files have associated `.html` files which can be opened by any browser and contain the code, remarks and outputs of the `.Rmd` files in a format more convenient to human readers. These are listed as files without a file ending. Files which were downloaded from the internet were renamed to fit the data structure. The original file names are usually noted in the corresponding R script. In some cases, there are multiple files which have similar filenames except for some elements. These are marked by brackets.

File	Description	Important dependencies	Restr. access
Diss_Data-Rproj	R Project file which contains all data and scripts. Open this to use project		X
Data sources			
0_Decaleon.xml	Original Decaleon database	Loaded by 0_Decaleon	X
0_Decaleon.rds	Contains relevant data extracted from Decaleon data base	Created by 0_Decaleon Loaded by 2_a_VariableExtraction_itdat	X
0_EFLlex.txt 0_FLElex.txt	Original reference lists from CEFRlex databases	Loaded by 2_a_VariableExtraction_itdat	X
0_refCorpus_de.xlsx 0_refCorpus_fr.tsv 0_refCorpus_it.csv 0_refCorpus_uk.xlsx	Original reference lists from SUBTLEX databases	Loaded by 2_a_VariableExtraction_itdat	X
0_LimeSurvey_head.txt	Head of import files for LimeSurvey questionnaires	Used by 1_DataPreparation	X
1_a_ItemData.xlsx	Original Excel file which contains all input and item texts and some structural information on items	Loaded by 1_DataPreparation	X
1_b_ResponseData.csv	Original response data provided by ÜGK project	Loaded by 1_DataPreparation	X
1_c_TestTakerData.csv	Original questionnaire data provided by SWISSbase	Loaded by 1_DataPreparation	X
1_c_ttdat_UEGK_2017_ItemScaleDescription.csv	Information on columns in original test taker file		X
2_a_correct-tags.csv	Helper file for semi-automatic correction of POS tagging	Loaded by 2_a_VariableExtraction_itdat	X
2_a_taggedTexts_dataCons.xlsx	POS tagged input texts with reference list information after manual correction	Based on 2_a_taggedTexts_[lgV].csv Loaded by 2_a_VariableExtraction_itdat	X
2_a_taggedTexts_humanR.xlsx	Results of manual coding of negations and cognates for input texts	Based on 2_a_taggedTexts_[lgV].csv Loaded by 2_a_VariableExtraction_itdat	X
2_a_itemTexts_dataCons.xlsx	POS tagged item texts with reference list information after manual correction	Based on 2_a_itemTexts_[lgV].csv Loaded by 2_a_VariableExtraction_itdat	X
2_a_itemTexts_humanR.xlsx	Results of manual coding of negations and cognates for item texts	Based on 2_a_itemTexts_raw_[lgV].csv Loaded by 2_a_VariableExtraction_itdat	X

File	Description	Important dependencies	Restr. access
2_a_taggedNIs.xlsx	Results of Limesurvey questionnaire about NIs, background knowledge and plausability rating	Transferred manually to 1_a_ItemData.xlsx and 2_a_itemTexts_dataCons.xlsx	X
2_newIDs-Match.csv	Table which matches original ÜGK item and text IDs with anonymised IDs	Created by 2_a_VariableExtraction_itdat	X
R scripts and functions			
0_Preliminaries	Installs koRpus packages and initialises renv package Needs to be run manually		
0_Decaleon	Prepares Decaleon data base for use in data analyses	Loads Decaleon.xml Creates 0_Decaleon.rds	X
0_LexDisparity.R 0_LexDispersion.R 0_LexSophistication.R	Contain R functions which are used to calculate lexical diversity measures	Used by 2_a_VariableExtraction_itdat	X
1_DataPreparation	Prepares item, response and test taker data for use in data analyses	Loads 1_a_ItemData.xlsx, 1_b_ResponseData.csv, 1_c_TestTakerData.csv Creates 1_a_itdat.rds, 1_b_rsdats.rds, 1_c_ttdat.rds	X
2_a_VariableExtraction_itdat	Extracts variables based on input texts and items	Loads 1_a_itdat.rds, reference lists, 2_a_correct-tags.csv, 2_a_taggedTexts_humanR.xlsx, 2_a_taggedTexts_dataCons.xlsx, 2_a_itemTexts_humanR.xlsx, 2_a_itemTexts_dataCons.xlsx Creates 2_a_taggedTexts_[lgV].csv, 2_a_itemTexts_[lgV].csv, 2_a_itemTexts_raw_[lgV].csv, 2_a_itdat.rds, 2_a_itdat.csv Outputs 2_a_taggedTexts_[lgV]_level-cognate-compare.csv, 2_a_taggedTexts_[lgV]_cognate-compare.png, 2_newIDs-Match.csv	X
2_b_VariableExtraction_rsdats	Extracts item difficulties and other item parameters from test response data	Loads 1_b_rsdats.rds Creates 2_b_ridat.rds, 2_b_ridat.csv Outputs files in folder 2_rsdats/	X

File	Description	Important dependencies	Restr. access
2_c_VariableExtraction_ttdat	Extracts test taker variables from questionnaire data	Loads 1_c_ttdat.rds Creates 2_c_ttdat.rds, 2_c_ttdat.csv	X
3_a_DescriptiveStatistics_itdat	Outputs various univariate descriptive statistics about assessment characteristics, also across languages	Loads 2_a_itdat.rds Creates 3_a_itdat.rds, 3_a_itdat.csv Outputs files in 3_a_plots and 3_a_tables	
3_b_DescriptiveStatistics_rsdar	Outputs descriptive statistics, mostly plots about response patterns and results from CTT and IRT analyses	Loads 2_b_ridar.rds, 2_a_itdat.rds Outputs files in 3_b_plots	
3_c_DescriptiveStatistics_ttdat	Outputs descriptive statistics about test takers, including performance on assessments (WLEs)	Loads 2_c_ttdat.rds, 2_rsdar/MFR_WLEs.csv	
4_AnalyticalStatistics_Multicollinearity	Analyses variables in data set for multicollinearity and possible data reduction	Loads 3_a_itdat.rds, 2_b_ridar.rds Creates 4_itdat.rds, 4_itdat.csv Outputs files in 4_plots, 4_tables	
4_AnalyticalStatistics_Regression	Contains correlation and regression analyses, predicting item difficulty	Loads 4_itdat.rds Outputs files in 4_plots, 4_tables	
4_AnalyticalStatistics_RandomForests	Contains random forests analyses for triangulation	Loads 4_itdat.rds, 4_tables/Correlations-xsi.csv	
4_AnalyticalStatistics_Bayesian	Contains Bayesian statistical analysis triangulation	Loads 4_itdat.rds, 1_b_rsdar, 2_c_ttdat.rds Outputs several .rds files containing model and validation data	X
Output files			
0_LimeSurvey_exp_[lgV].txt	Import files for LimeSurvey on identification of NIs	Created by 1_DataPreparation	X
1_a_itdat.rds	Item data	Created by 1_DataPreparation Loaded by 2_a_VariableExtraction_itdat	X
1_b_rsdar.rds	Test response data	Created by 1_DataPreparation Loaded by 2_b_VariableExtraction_rsdar	X
1_c_ttdat.rds	Test taker data	Created by 1_DataPreparation Loaded by 2_c_VariableExtraction_ttdat	X
2_a_taggedTexts_[lgV].csv	POS tagged input texts with information on reference lists	Created by 2_a_VariableExtraction_itdat Source of 2_a_taggedTexts_dataCons.xlsx and 2_a_taggedTexts_humanR.xlsx	X

File	Description	Important dependencies	Restr. access
2_a_itemTexts_[lgV].csv	POS tagged item texts with information on reference lists	Created by 2_a_VariableExtraction_itdat Source of 2_a_itemTexts_dataCons.xlsx	X
2_a_itemTexts_raw_[lgV].csv	POS tagged item texts	Created by 2_a_VariableExtraction_itdat Source of 2_a_itemTexts_humanR.xlsx	X
2_a_taggedTexts_[lgV] _level-cognate-compare.csv 2_a_taggedTexts_[lgV] _cognate-compare.png	Comparison of lexical difficulty and cognate variable extraction	Created by 2_a_VariableExtraction_itdat	X
2_a_itdat.csv 2_a_itdat.rds	Item data and assessment variables (using anonymised item and text IDs)	Created by 2_a_VariableExtraction_itdat	
2_b_ridat.csv 2_b_ridat.rds	Item response information and parameters (using anonymised item IDs)	Created by 2_b_VariableExtraction_rsdats	
2_c_ttdat.csv 2_c_ttdat.rds	Test taker data, limited for use in thesis	Created by 2_c_VariableExtraction_ttdats	
3_a_itdat.csv 3_a_itdat.rds	Item data and assessment variables, including recorded and eliminated variables, condensed incorrect answer option variables	Created by 3_a_VariableExtraction_itdat	
4_itdat.csv 4_itdat.rds	Data set with all variables on items, including item difficulty estimations	Created by 4_AnalyticalStatistics_Multicollinearity	
4_brms_MFR.rds	Data file containing model data of Bayesian MFR-like model (for quicker loading)	Created by 4_AnalyticalStatistics_Bayesian Loaded by 4_AnalyticalStatistics_Bayesian	X
4_brms_MFR_loo.rds	Data file containing results of LOO validation of Bayesian MFR-like model (for quicker loading)	Created by 4_AnalyticalStatistics_Bayesian Loaded by 4_AnalyticalStatistics_Bayesian	X
4_brms_LLTM.rds	Data file containing model data of LLTM-like Bayesian logistical regression (for quicker loading)	Created by 4_AnalyticalStatistics_Bayesian Loaded by 4_AnalyticalStatistics_Bayesian	X
4_brms_LLTM_loo.rds	Data file containing results of LOO validation of LLTM-like Bayesian logistical regression (for quicker loading)	Created by 4_AnalyticalStatistics_Bayesian Loaded by 4_AnalyticalStatistics_Bayesian	X
4_brms_LLTM_wInt.rds	Data file containing model data of LLTM-like Bayesian logistical regression with interactions (for quicker loading)	Created by 4_AnalyticalStatistics_Bayesian Loaded by 4_AnalyticalStatistics_Bayesian	X

File	Description	Important dependencies	Restr. access
4_brms_LLTM_wInt_loo.rds	Data file containing results of LOO validation of LLTM-like Bayesian logistical regression with interactions (for quicker loading)	Created by 4_AnalyticalStatistics_Bayesian Loaded by 4_AnalyticalStatistics_Bayesian	X
<i>Files in 2_rsdat</i>			X
Rasch_mod-summary_[lgV].Rout	Model summaries of language-specific Rasch models	Created by 2_b_VariableExtraction_rsdat	X
MFR_mod-summary.Rout	Model summary of multilingual MFR model	"	X
MFR_mod-xsi-facets.csv	Item difficulties estimated by MFR model	"	X
MFR_itemFits.csv	Item fit estimates extracted from MFR model	"	X
MFR_WLEs.csv	Person ability estimates (WLEs) extracted from MFR model	"	X
MFR_WrightMap_[lgV].png	Wright maps of item difficulties and person ability estimates per language version based on MFR model	"	X
MFR-COV-comparison.csv	WLEs extracted from MFR models with and without test taker covariates	"	X
PCM_mod-summary_ENde.Rout	Model summary of ENde Partial Credit Model (PCM) model	"	X
PCM_ENde_mod.rds	ENde PCM model as an R-readable data object	"	X
PCM-ENde_WrightMap.png	Wright maps of item difficulties and person ability estimates based on ENde PCM model	"	X
<i>Files in 3_a_plots</i>			Created by 3_a_DescriptiveStatistics_itdat
[000]_DistGraph_[Variable].png	Distribution plot of a variable (identified by number and name), including recoded variables		
[000]_BoxPlot_[Variable].png	Box plot of a variable (identified by number and name)	"	
LexVars_corrs_[lgV].png	Correlation matrix of all lexical variables of input texts in a language version (plot)	"	
LexDiv-pairs_[1 2]_[lgV].png	Pairs plot of lexical diversity variables of input texts in a language version (two plots)	"	

File	Description	Important dependencies	Restr. access
LexVars_corrs_NI_[lgV].csv	Correlation matrix of all lexical variables of NIs in a language version (plot)	"	
LexDiv-pairs_NI_[1 2]_[lgV].png	Pairs plot of lexical diversity variables in NIs in a language version (two plots)	"	
LexVars_T-NI_[lgV].png	Plot of relationship between lexical variables in input texts and NIs	"	
[Variable]-per-LgV.png	Plots of all lexical variables by input text and language version	"	
Overl_vars-comp_Q_[lgV].png	Compares overlap measures of stem/question and NI	"	
Overl_vars-comp_Corr_[lgV].png	Compares overlap measures of correct answer option and NI	"	
Overl_vars-comp_Incorr_[lgV].png	Compares overlap measures of incorrect answer option and input text	"	
Overl_corr-incorr_Q-T_[lgV].png	Compares overlap measures of correct and incorrect answer option and NI and input text	"	
<i>Files in 3_a_tables</i>			
itdat-describe_[lgV].csv	Output of <code>describe()</code> function on all variables in a language version	Created by <code>3_a_DescriptiveStatistics_itdat</code>	
LexVars_corrs_[lgV].csv	Correlation matrix of all lexical variables of input texts in a language version (table)	"	
LexVars_corrs_NI_[lgV].csv	Correlation matrix of all lexical variables of NIs in a language version (table)	"	
LexVars_DescriptiveStats.csv	Selected statistical values of lexical variables	"	
<i>Files in 3_b_plots</i>			
[00]_Distractors_[lgV].png	Response pattern of items in each language version	Created by <code>3_b_DescriptiveStatistics_rsdat</code>	
MFR-itemDiff_facetByLgV.png	Item differences per language version, with DIF information	"	

File	Description	Important dependencies	Restr. access
MFR_ItemDiff-comp.png	Plots of item difficulties in each combination of language versions	"	
<i>Files in 4_tables</i>			
CorrM_vars_[lgV].csv	Correlation matrices across all assessment variables per language version	Created by 4_AnalyticalStatistics_Multicollinearity	
POS-corrs_[fgnLg].csv	Correlation matrices of POS variables in input texts per language version	"	
POS_T-NI_corrs_[lgV].csv	Correlation matrices of POS variables in input texts and NIs per language version	"	
Correlations-xsi.csv	List of correlations between item difficulty and all assessment variables	Created by 4_AnalyticalStatistics_Regression	
Correlations-xsi-public.csv	List of significant correlations between item difficulty and all assessment variables (formatted for L ^A T _E Ximport)	"	
Correlations-xsi_texts.csv	List of significant correlations between mean item difficulty per input text and all input text variables	"	
Correlations-xsi-public_Texts.csv	List of significant correlations between mean item difficulty per input text and input text variables (formatted for L ^A T _E Ximport)	"	
<i>Files in 4_plots</i>			
Heatmaply_[lgV].csv	Heatmap across all assessment variables per language version (based on correlation matrices)	Saved manually from 4_AnalyticalStatistics_Multicollinearity	
Heatmaply_sig_[lgV].csv	Heatmap across all assessment variables which correlate significantly with item difficulty (based on correlation matrices, per language version)	Saved manually from 4_AnalyticalStatistics_Multicollinearity	
IndR_[Variables].png	Various plots of individual variables interacting with each other	Created by 4_AnalyticalStatistics_Multicollinearity	
POS-corrs_[fgnLg].png	Correlation plots of POS variables in input texts per language version	"	

File	Description	Important dependencies	Restr. access
POS-pairs_ <i>[fgnLg]</i> .png	Pairs plots of POS variables in input texts per language version	"	
POS_T-NI_corrs_ <i>[lgV]</i> .png	Correlation plots of POS variables in input texts and NIs per language version	"	
POS_T-NI_pairs_ <i>[lgV]</i> .png	Pairs plots of POS variables in input texts and NIs per language version	"	
Corrs-xsi_ <i>[Variable]</i> .png	Scatterplots of relationship between item difficulty and variables (significant correlation in at least one language version)	Created by 4_AnalyticalStatistics_Regression	
Corrs-xsi-text_ <i>[Variable]</i> .png	Scatterplots of relationship between mean item difficulty per input text and variables (significant correlation in at least one language version)	"	
Regr_ <i>[lgV]</i> _effects.png	Prediction of item difficulties according to major predictors in final linear regression model	"	
Regr_Ml_Mixed_randomeff_TextID.png	Plots effect of the input texts on item difficulty prediction in multilingual regression	"	
Regr_Ml_Mixed_randomeff_LgVers.png	Plots effect of language version on item difficulty prediction in multilingual regression	"	
Bayesian-MFR_comp.png	Plot of item difficulties estimated by MFR and Bayesian MFR-like model	Created by 4_AnalyticalStatistics_Bayesian	

Table C.1: File list of electronic appendix

C.3 Overview of tables

Three tables form the data source for this thesis: `1_a_ItemData.xlsx`, `1_b_ResponseData.csv` and `1_c_TestTakerData.csv`. The first contains all texts of the assessment and additional information on the texts and items. It was created by the author of this thesis by copying the texts from the original assessment. The second table contains the test takers' responses to the assessment. It was provided as-is by the ADB, the institution responsible for the analysis of the ÜGK 2017 assessment. Finally, the third table contains information on all test takers who were involved in the ÜGK assessment. Both test results (in the form of PVs) from the L1 and L2 assessments and answers from the background questionnaire are included in this file. All other tables in the electronic appendix are created by the provided R scripts and will not be described in more detail here.

C.3.1 Original item data

The table below describes all columns in the table `1_a_ItemData.xlsx`, which contains all raw texts and rated information for each item in all language versions. It is read by the R script `1_DataPreparation` and output without change to the file `1_a_itdat`. This is then the data source for the extraction of all linguistic variables in the script `2_a_VariableExtraction_itdat`.

Column identifier	Description
Item_ID	The technical item identifiers used in the original CBA ItemBuilder files.
Item_ID_UeGK	The item identifiers used by the ÜGK project during test administration.
Session	The test administration session. All items belong to the main study, this column is only present for completeness' sake.
LgVersion	The language version of the item.
ForeignLg	The foreign language of the item. FR : French, DE : German, EN : English.
LgSchool	The language of schooling of the item, used in the instructions and the item. fr : French, de : German, it : Italian
Page	The page of the computer-based task on which the item is located.
Text_Input	The entire input text located in the yellow part of the page.
TextID	A text ID created for the purposes of this study.
Answer_Format	An identifier of the item format used.
InputEqualsItems	In some cases, the input text contains the answer options of the item. In these cases, this column contains a y(es).
Text_Situation	The description of the input text's contextual situation.
Text_Instruction	An additional instruction to explain how the item works.
Text_Question	If the item contains a question, it is found in this column.
Text_Stem	If the question is a stem (the beginning of a sentence), it is found in this column.
Correct_Answer	The identifier of the correct answer.
Correct_Answer_Position	The position of the correct answer.
ID_AnswerOptionX	For each answer option, the technical ID.
Text_AnswerOptionX	For each answer option, the text it contains.
Pic_AnswerOptionX	For each answer option, a short description of the picture, if there is one.
ContinuousText	This text variable indicates to what extent the input text can be considered continuous.
TextFunction	This text variable indicates the primary text function of the input text.
ItemType	This item variable distinguishes various MCQ and Matching items.
StemType	This item variable indicates whether the item stem is a question or not.
RespProcess	This text-by-item variable contains the item's passage/question relation.
BackgroundKn	This text-by-item variable indicates whether raters considered background knowledge to be necessary for solving the item.
NI_Dist	This text-by-item variable indicates how the NI is distributed in the input text.
Fals	This text-by-item variable indicates how many (incorrect) answer options can be explicitly falsified using the input text.

Table C.2: Description of the columns in the data file `1_a_ItemData.xlsx`. *Note*: the file itself is not part of the open source data package of this thesis.

C.3.2 Original response data

The table below describes the columns in the table `1_b_ResponseData.csv`. It contains all test taker responses from the ÜGK 2017 foreign language reading assessment. First, the table indicates for each item which response the test taker chose or whether and why they did not choose any.

Second, for each test response, the table indicates whether it was correct or not. The table is read by the R script `1_DataPreparation` and output without change to the file `1_b_rsd.dat`. This is then the data source for the item analyses and estimation of test taker abilities in the script `2_b_VariableExtraction_rsd.dat`.

Column identifier	Description
<code>id_student</code>	The identifier of the test taker. This is the same ID used in the file <code>1_c_TestTakerData.csv</code> .
<code>(ItemID)_resp</code>	The actual response for the item identified by its task and item ID. There 94 columns of this type, one for each item.
<code>(ItemID)_correct</code>	The response for the item identified by its task and item ID, identifying whether it was correct or false. There 94 columns of this type, one for each item.

Table C.3: Description of the columns in the data file `1_b_ResponseData.csv`. *Note:* the file itself is not part of the open source data package of this thesis.

C.3.3 Original test taker data

The table `1_b_ResponseData.csv` contains background information on the test takers as well as additional test results. The table was provided by SWISSUbase. Documentation on its exact contents can be found in the file `1_c_ttdat_UEGK_2017_ItemScaleDescription.csv`. The table is read by the R script `1_DataPreparation` and output in a reduced form to the file `1_c_ttdat.rds`. This is then the data source for some descriptions of the test takers in the script `3_c_DescriptiveStatistics_ttdat` and, more importantly, the analyses in step 4.