

Department of Psychology

PhD Program "Psychology, Linguistics and Cognitive Neuroscience", 34th cycle

Computerized system for bilingual language and reading screenings

Maren Rebecca Eikerling
Registration number: 849495

Tutor: Prof. Dr. Maria Teresa Guasti
Department of Psychology
University Milan-Bicocca

Co-tutor: Dr. Maria Luisa Lorusso
Unit of Neuropsychology of Developmental Disorders - Department of Child Psychopathology
Scientific Institute IRCCS E. Medea - Associazione La Nostra Famiglia

Coordinator: Marco Perugini
Department of Psychology
University Milan-Bicocca

Academic year 2020/21

Abstract

In line with relevance and complexity of language and reading assessment in bilingual children, this work incorporates recommended language assessment of bilingual children's first as well as their second language and making use of innovative technologies. Previous research relates to the suitability of tasks assessing language with language-specific clinical markers, and reading skills in the bilingual population across languages. Some of these tasks have been implemented as computerized tasks allowing for automatic administration and evaluation. However, these computerized tasks are restricted to the characteristics defined by developers and cannot be adapted for use in school or clinical settings.

This manuscript includes a review of the scientific literature concerning language and reading screening approaches and six empirical studies on the construction, administration, suitability and usefulness of computerized bilingual language and reading screenings.

The first study concerns how Speech and Language Therapists assess bilingual children. In a survey with 300 Speech and Language Therapists it was investigated how far Speech and Language Therapists consider assessment of both languages spoken relevant and realizable with computerized screening tasks. Results suggest an imbalance between the knowledge of the specific requirements for assessment of multilingual children and their application at work. Speech and Language Therapists also indicated openness towards computerized solutions. The second study investigates the potential of reading assessment in both languages through automatic task administration and evaluation. The performance in standardized Italian reading tests of 33 Mandarin-speaking children living in Italy, attending grades 3 and 4 in primary school, is compared to task performance in computerized screening tasks, implemented on E-Prime 3.0, administered in presence of the examiner. Results suggest that computerized reading and language tests reliably assess reading performance in L2 speakers of Italian.

The third, fourth, fifth and sixth study investigate the potential of the modifiable screening platform MuLiMi, created for this dissertation, in the risk identification of Developmental Language Disorders and Developmental Dyslexia risk in bilingual children. These studies covering five language groups, two age ranges and two different screening purposes, testing 109 child participants indicate that it is possible to automatically assess language and reading skills in both languages with the screenings implemented on the MuLiMi web app. Further interesting findings suggest an advantage of paradigms such as "Who says it right?" over judgement tasks for screening tasks on morphosyntactic processing and phonological awareness as well as to the potential difference in contribution to risk identification between different

item subcategories, e.g. verbs vs. noun comprehension and language-specific vs. non-language-specific nonwords. In addition to that, study six assesses the usability of this screening platform as perceived by the examiners administering the screenings and by the children. The results generally suggest user-friendliness and ease in use even when administering these screenings remotely. Some of the information obtained indicate the need of improvement concerning the system speed.

In summary, findings concern the possibility of using fully automatically administered and scored tasks for the identification of bilingual children at risk for language or reading disorders, meeting the requirements concerning reliability and validity of the tests and allowing an assessment perceived as easy, useful and pleasant. The application of the screening platform MuLiMi appears to meet the requirements concerning bilingual language and reading assessment, allowing for the comparison of both languages as well as conscious decision-making for diagnoses and intervention need.

Un sistema computerizzato per lo screening delle competenze di linguaggio e lettura in bambini bilingui

Sommario

Considerando la rilevanza e complessità della valutazione del linguaggio e della lettura nei bambini bilingui, il presente lavoro raccomanda di valutare il linguaggio nella prima e nella seconda lingua e l'utilizzo di tecnologie innovative. Ricerche precedenti hanno indagato, in varie lingue, l'adeguatezza di prove che valutano il linguaggio utilizzando marcatori clinici lingua-specifici, e la lettura nella popolazione bilingue. Alcune di queste prove sono state implementate in forma computerizzata per la somministrazione e la valutazione automatizzate. Spesso queste prove computerizzati di screening sono limitate alle caratteristiche definite dagli sviluppatori e non possono essere adattate per un utilizzo in scuola o clinica.

Il presente elaborato include una revisione della letteratura scientifica relativa agli strumenti di screening del linguaggio e della lettura e sei studi empirici riguardanti la costruzione, la somministrazione, la validità e l'utilità di screening computerizzati di linguaggio e lettura nei bambini bilingui.

Il primo studio riguarda i metodi con cui i logopedisti valutano i bambini bilingui. Attraverso un questionario compilato da 300 logopedisti è stato indagato se loro considerano la valutazione di tutte le lingue parlate rilevante e realizzabile attraverso dei compiti computerizzati. I risultati suggeriscono uno squilibrio tra conoscenze delle prassi raccomandate e la loro effettiva applicazione. I logopedisti si dichiarano aperti alla possibilità di utilizzare delle procedure computerizzate. Il secondo studio indaga il potenziale della valutazione della lettura in entrambe le lingue attraverso procedure computerizzate. Confrontando le prestazioni di 33 bambini, parlanti mandarino-italiano e frequentando la terza e quarta elementare in Italia, in test standardizzati di lettura in italiano con quelle conseguite in compiti di screening computerizzati, implementati con E-Prime 3.0, somministrati in presenza, emerge che i test computerizzati valutano in modo affidabile le competenze di lettura nei parlanti di italiano L2.

Il terzo, quarto, quinto e sesto studio indagano il potenziale della piattaforma di screening modificabile MuLiMi, creata per questo lavoro di tesi, nell'identificare il rischio di Disturbo Primario del Linguaggio e di Dislessia Evolutiva in bambini bilingui. Questi studi includono cinque gruppi linguistici, due fasce di età, due target di studio e 109 bambini e confermano che è possibile valutare le competenze di linguaggio e lettura in entrambe le lingue attraverso la procedura di screening implementata sulla web app MuLiMi. Altri risultati suggeriscono un vantaggio di paradigmi come "Chi lo dice giusto?" rispetto ai giudizi di correttezza nelle prove

morfosintattiche e di consapevolezza fonologica, e la potenziale differenza nel contribuire all'identificazione del rischio da parte di diverse sottocategorie di item, per esempio comprensione di verbi vs. nomi e nonparole lingua-specifiche vs. non lingua-specifiche. Il sesto studio indaga l'usabilità di MuLiMi analizzando le impressioni degli esaminatori e dei bambini valutati. I risultati suggeriscono che entrambi i gruppi trovano l'uso della piattaforma facile, anche utilizzata a distanza. Alcune delle risposte indicano la necessità di migliorare la velocità del sistema.

In conclusione, i risultati riguardano la possibilità di utilizzare compiti somministrati e corretti in modo automatico per l'individuazione dei soggetti bilingui a rischio per disturbi di linguaggio e di lettura, rispettando i requisiti di affidabilità e validità delle prove e consentendo una valutazione valutata come facile, utile e piacevole. L'utilizzo di MuLiMi sembra soddisfare i requisiti relativi alla valutazione del linguaggio e della lettura nei bilingui, consentendo il confronto tra le competenze in entrambe le lingue, e favorendo negli esaminatori un processo decisionale consapevole per la diagnosi e l'intervento.

Go straight to pine trees
To learn pine
And to bamboo stalks
To know bamboo

– Basho (1644–1694)

Acknowledgments

My first and very special thanks go to Maria Luisa Lorusso, the scientific supervisor of this project, for sharing her outstanding expertise and advising me throughout the project. I would also like to thank Maria Teresa Guasti for all her support during the PhD phase.

I would like to thank all university students who have contributed to this project for their thesis or course participation. In particular, I would like to thank Matteo Secco for his enormous contribution to the creation of the screening platform. I would also like to thank Shari Gandelli, Giulia Cha, Gloria Marchesi, Bianca Luculli and Marco Andreoletti for their support in data collection. This goes together with thanks to my colleagues who have supported these very projects. Many thanks to Franca Garzotto and Francesco Vona from the i3Lab at Politecnico University, Milan; Sofia Limarzi and Annalaura Filippo from Corso di Laurea in Logopedia - Università degli Studi di Milano, Bosisio Parini; and Stefania Gazzola and Daniela Sarti from the IRCCS Istituto Neurologico Carlo Besta. I would particularly like to thank all the cooperating schools, kindergartens, SLT clinics and individuals who made recruitment and data collection possible despite pandemic-related restrictions.

Without the support of many different native speakers as well as colleagues and friends with a talent for language, who were essential in the selection, creation and validation of the screening items, the implementation of computerized, bilingual screenings would not have been possible. I would therefore like to thank Andrea Bigagli, Shenai Hu, Giulia Cha, Maria Luisa Lorusso, Sofia González Castro, Theo Marinis, Sheila Keeshan, Giuliana Genovese and all the people who have participated in the nonword rating studies. I would also like to thank the colleagues from the BiSLI COST Action for kindly sharing material and expertise, especially Myriam Cantú Sánchez and Tanja Rinker.

I would also like to thank my colleagues from Bilinguismo Conta, the BiL group and the Eichstätt lab meetings for the inspiration. Besides that, I am grateful for the peer-support along the way through the MultiMind ITN (and especially Theresa Bloder and Jasmijn Bosch), movement (Bielefeld University), the peer groups of the dbs and GISKID, as well as fellow PhD students from University Milano-Bicocca. Also, I found mentors' support from former colleagues from Bielefeld University as well new colleagues from the Martin-Luther-University in Halle.

I would like to thank my friends and family for any form of personal support. Special thanks go to my Italian SLT friends who helped me to navigate around work and free time, my flatmates, who tolerated my long nights of work and my long-time friends Angelie Kraft as

proofreader and Miguel Waltereit as polyglot. I thank all my friends and relatives for interest and support concerning my work. I want to especially thank my parents Heinz-Josef Eikerling and Ute Glunz-Eikerling as well as my brother Hendrik Eikerling and my partner Philipp Förster for their unconditional, steady support.

I feel honoured to be part of the MultiMind project as an Early Stage Researcher and Marie Skłodowska Curie Fellow funded by the European Union's Horizon2020 research and innovation programme and express my sincere gratitude for this opportunity.

Table of Content

1	Introduction	1
2	Theoretical background.....	1
2.1	Developmental Language Disorders.....	1
2.1.1	Prevalence (Epidemiology).....	2
2.1.2	Causes (Etiology) and risk factors	2
2.1.3	Symptoms	2
2.2	Developmental Dyslexia	3
2.2.1	Prevalence (Epidemiology).....	3
2.2.2	Causes (Etiology) and risk factors	4
2.2.3	Symptoms	4
2.3	Child bi-/multilingualism.....	5
2.3.1	Terminology and subtypes of bilingualism	5
2.3.2	Child bi-/multilingualism in Europe and its implications for the educational & health sector.....	7
2.4	Language and reading assessment in bilingual children.....	7
2.4.1	Language assessment in bilingual children	10
2.4.1.1	Language-specific clinical markers for DLD diagnosis in bilingual children.	10
2.4.1.2	Language universal clinical markers for DLD.....	11
2.4.2	Reading assessment in bilingual children.....	13
2.4.2.1	Language-specific risk indicators for DD diagnosis in bilingual children..	14
2.4.2.2	Dynamic Assessment.....	15
2.4.3	Caregiver questionnaires.....	16
2.4.4	Computerized screening tools	17
2.5	Structure of the thesis	17
3	Attitudes and practices concerning multilingualism in the context of SLTs' diagnoses and interventions.....	18
3.1	Participants	18
3.2	Questionnaire design	18
3.3	Data analysis.....	19
3.4	Procedure.....	19
3.5	Results	20
3.6	Discussion.....	23
4	Piloting computerized Chinese and Italian DD screening tasks	23
4.1	Methods & material	24
4.1.1	Participants	24

4.1.2	Material	24
4.1.2.1	Screening tasks.....	24
4.1.3	Procedure & data analysis.....	28
4.2	Results	28
4.2.1	Comparison of screening results and the teachers' risk indication	28
4.2.2	Comparison of screening results and the DDE-2.....	29
4.3	Discussion.....	30
5	MuLiMi screening studies	31
5.1	Goals.....	32
5.2	Experimental design.....	33
5.2.1	Development of screening tasks.....	34
5.2.1.1	Development of computerized DLD screening tasks	34
5.2.1.2	Development of computerized DD screening tasks	35
5.2.2	Description of the web app	35
5.2.3	Recruitment of participants.....	41
5.2.4	Validation	41
5.2.4.1	Teacher & SLT questionnaire.....	42
5.2.4.2	Caregiver questionnaires.....	42
5.2.5	General testing procedure	43
5.2.6	Data analysis.....	44
5.3	Bilingual, computerized DLD screening for Spanish-speaking children living in Italy	44
5.3.1	Hypotheses	44
5.3.2	Material & methods	45
5.3.2.1	Participants	45
5.3.2.2	Screening tasks.....	46
5.3.2.3	Standardized tests.....	53
5.3.2.4	Procedure	54
5.3.2.5	Risk score creation.....	55
5.3.3	Results & discussion	56
5.3.3.1	Comparison of performance in screening tasks and risk level	56
5.3.3.2	Comparison of performance in the screening's and standardized tasks	57
5.3.3.3	Comparison of performance in screening tasks and SLT & teacher questionnaires	60
5.3.3.4	Comparison of performance in the screening tasks and caregiver questionnaires	64

5.3.3.5	Comparison of performance in the screening tasks	64
5.3.3.6	Comparison of performance in the screening tasks at t1 and at t2	66
5.3.3.7	Interim discussion	68
5.4	Bilingual, computerized DLD screening for Italian-speaking children living in Germany	71
5.4.1	Hypotheses	72
5.4.2	Material & methods	72
5.4.2.1	Participants	73
5.4.2.2	Screening tasks.....	73
5.4.2.3	Standardized tests.....	77
5.4.2.4	Procedure	79
5.4.2.5	Risk score creation.....	79
5.4.3	Results & Discussion.....	80
5.4.3.1	Comparison of performance in screening task and risk level	80
5.4.3.2	Comparison of performance in the screening tasks and language background.....	82
5.4.3.3	Comparison of performance in the screening tasks and standardized tests	83
5.4.3.4	Comparison of performance in the screening's and teacher questionnaires	85
5.4.3.5	Comparison of performance in the screenings and caregiver questionnaires	87
5.4.3.6	Comparison of performance in the screening tasks	88
5.4.3.7	Comparison of performance in the screening tasks at t1 and at t2	90
5.4.3.8	Interim discussion	93
5.5	Bilingual, computerized DD Screening for Italian-speaking children living in Germany	96
5.5.1	Material & methods	96
5.5.1.1	Participants	96
5.5.1.2	Screening tasks.....	97
5.5.1.3	Standardized tests.....	104
5.5.1.4	Procedure	104
5.5.1.5	Risk score creation.....	105
5.5.2	Results & discussion	105

5.5.2.1	Comparison of screening tasks and standardized test results within languages	106
5.5.2.2	Comparison of screening tasks and standardized test results across languages and task types	108
5.5.2.3	Comparison of screening tasks and risk scores.....	111
5.5.2.4	Comparison of screening tasks and caregiver questionnaires	113
5.5.2.5	Comparison of screening tasks and teacher questionnaires.....	115
5.5.2.6	Interim discussion	117
5.6	Computerized, remote DD screening for bilingual children living in Italy	120
5.6.1	Hypotheses	121
5.6.2	Material & methods	121
5.6.2.1	Participants	121
5.6.2.2	Screening tasks.....	121
5.6.2.3	Standardized/traditional reading tests.....	126
5.6.2.4	Usability questionnaire	127
5.6.2.5	Procedure	128
5.6.2.6	Risk score creation.....	129
5.6.3	Results & discussion	130
5.6.3.1	Comparison of Italian screening results and the DDE-2	130
5.6.3.2	Comparison of L1 screening task and the standardized/traditional test results	133
5.6.3.3	Comparison of screening results and risk level.....	133
5.6.3.4	Comparison of screening results and caregiver and teacher questionnaires	134
5.6.3.5	Comparison of screening results within and across languages.....	136
5.6.3.6	Screening usability	139
5.6.3.7	Interim discussion	144
6	General discussion.....	149
6.1	Interpretation of results.....	150
6.1.1	Bilingual DLD screenings	150
6.1.1.1	Validity	150
6.1.1.2	Predictivity.....	154
6.1.2	Validity of bilingual DD screenings	155
6.1.3	Usability	158
6.2	Methodological discussion and future work	160
6.2.1	Limitations.....	160

6.2.2	Implications for future work.....	162
7	Conclusion	165
8	Appendix	I
8.1	Appendix A – Questionnaires	I
8.2	Appendix B – Screening items.....	VI
8.2.1	Morphosyntactic processing tasks.....	VII
8.2.1.1	WSIR – subject-verb agreement.....	VII
8.2.1.2	WSIR – finiteness.....	VII
8.2.1.3	Case matching	VIII
8.2.1.4	Subject-verb agreement	IX
8.2.1.5	Clitic pronoun judgement.....	X
8.2.1.6	Tense judgement	XI
8.2.2	Dynamic Novel Word Learning (testing phase).....	XI
8.2.3	Reading tasks	XIII
9	References.....	a

List of tables

Table 1: Overview of results for comparisons of screening performance (chapter 4).	30
Table 2: Clinical status of the participants of the DLD screening studies.	33
Table 3: Clinical status of the participants of the DD screening studies.	34
Table 4: Overview of NWs selected for the Spanish-Italian NWRT.	47
Table 5: Standardized test performance according to the children's risk level assignment.	56
Table 6: Overview of significant correlations (chapter 5.3).....	68
Table 7: Overview of NWs selected for the Italian-German NWRT.	73
Table 8: Descriptive statistics for age and standardized test results in the three groups.	80
Table 9: Associations between screening task performance at t1 and t2.	90
Table 10: Correlations between children's NWRT repetition performance at t1 and t2.	91
Table 11: Correlations between children's CLT subtest performance at t1 and t2.	91
Table 12: Overview of significant associations (chapter 5.4).....	94
Table 13: Overview of significant associations (chapter 5.5).....	117
Table 14: Overview of significant associations (chapter 5.6).....	145

List of figures

Figure 1: Distribution of SLTs' responses (experience & approaches).....	20
Figure 2: Distribution of SLTs' responses (language comparison).....	21
Figure 3: Distribution of SLTs' responses (diagnostic material).....	22
Figure 4: Distribution of SLTs' responses (computerized tasks).....	22
Figure 5: Line drawing for clitic pronoun judgement task as visual stimulation.....	27
Figure 6: Item presented in the radical position judgement task	27
Figure 7: Item presented in the left-right inversion judgement task	28
Figure 8: Groups of target users, user flow & functions of the MuLiMi screening platform.	36
Figure 9: Admin interface for content upload, item & screening compilation.	37
Figure 10: Video-based instruction.	38
Figure 11: Interface for test setup to initiate a remote screening session.....	39
Figure 12: Interface for screening result examination in the examiner section.	40
Figure 13: Examiner interface to enable screen sharing during a remote testing session ...	40
Figure 14: Examples from coloured line drawings presented during the NWRT.....	47
Figure 15: Examiner interface during remote administration CLT subtest.	48
Figure 16: Examiner interface during remote administration of the WSIR subtest.....	49
Figure 17: Examiner interface during remote administration of the DNWL subtest.	52
Figure 18: Mean task performance for screening tasks according to risk levels.....	57
Figure 19: Nonword repetition performance in the standardized test & in the screening	58
Figure 20: Task performance in the standardized sentence repetition test & WSIR task	59
Figure 21: Plot of kindergarten teachers' evaluation & WSIR task performance	63
Figure 22: Comparison of performance at t1 & t2 in the NWRT screening task.	66
Figure 23: Comparison of performance at t1 & t2 in the Italian & Spanish WSIR tasks.....	67
Figure 24: Comparison of performance at t1 & t2 in the Italian & Spanish CLT subtests. ...	68
Figure 25: Example from the German case matching screening task.	75
Figure 26: Example from the German subject-verb agreement screening task.	76
Figure 27: Examinee interface during the Italian clitic pronoun judgement task.	76
Figure 28: Mean task performance for screening tasks according to risk levels.	81
Figure 29: Mean repetition performance for NWs according to the risk levels.....	82
Figure 30: Scatterplot of performance in the CLT at t1 compared to t2.....	92
Figure 31: Scatterplot of NW repetition performance at t1 compared to t2.....	93
Figure 32: Examinee interface during the German self-paced syllable reading task.	97
Figure 33: Examinee interface during the German self-paced sentence reading task.....	98
Figure 34: Examinee interface during the German word identification task.....	99

Figure 35: Examinee interface during the German NW identification task.....	100
Figure 36: Examinee interface during the judgement screening tasks.	100
Figure 37: Examinee interface during German case matching screening (DD screening).	102
Figure 38: Examinee interface during the Italian RAN (digits) screening task	103
Figure 39: Examinee interface during the Italian clitic pronoun judgement screening task	103
Figure 40: Comparison of self-paced reading time & standardized word reading time	107
Figure 41: Comparison of self-paced reading time standardized word & NW reading.....	110
Figure 42: Self-paced reading time of German syllables according to overall risk.....	112
Figure 43: Accuracy in Italian clitic pronoun judgement according to overall risk.	113
Figure 44: Examinee interface during the Italian word stress identification task.	122
Figure 45: Examinee interface during the Mandarin character judgement tasks.	125
Figure 46: Examinee interface during the Mandarin phonological awareness tasks.....	125
Figure 47: Comparison of screening & standardized Italian word reading time	131
Figure 48: Comparison of screening & standardized word reading accuracy	132
Figure 49: Comparison of caregiver questionnaire & word identification screening task ...	135
Figure 50: Distribution of examiners' responses in the usability questionnaire (E2-E9).....	140
Figure 51: Distribution of examiners' responses in the usability questionnaire (E2-E14)...	141
Figure 52: Distribution of examiners' responses in the usability questionnaire (E15-E26).	142
Figure 53: Distribution of examiners' responses in the usability questionnaire (E27-E46).	143
Figure 54: Distribution of examinees' responses in the usability questionnaire	144

List of abbreviations

AoO	Age of Onset
ASR	Automatic speech recognition
BVL	Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni (BVL)
CDI	Communicative Development Inventories
CLT(s)	Cross-linguistic Lexical Task(s)
CPM	CPM-Coloured Progressive Matrices Italian version: Belacchi et al. 2008; German version: Bulheller & Häcker, 2001
CUP	Common Underlying Proficiency
DA	Dynamic Assessment
DDE-2	Batteria per la Valutazione della Dislessia e della Disortografia Evolutiva 2
DLD	Developmental Language Disorder
DD	Developmental Dyslexia
DNWL	Dynamic Novel Word Learning
DSA	Disturbi Specifici di Apprendimento, in English: Specific Learning Disorders
DSM-5	Diagnostic and Statistical Manual of Mental Disorders 5
FIGS	Family language input global score (QUIR-DC)
FIRS	Family language input risk score (QUIR-DC)
GS	General score (QUIR-DC)
ICD-11	International Classification of Diseases 11 th Revision
IPA	International Phonetic Alphabet
L1	First language, also referred to as family or home language
L2	Second language, also referred to as societal language
LITMUS	Language Impairment Testing in Multilingual Settings
LiSeDaZ	Linguistische Sprachstandserhebung Deutsch als Zweitsprache
LoE	Length of Exposure
ms	Milliseconds
MVC	Model-view-controller
n.a.	Not applicable
n.s.	not significant
NW(s)	Nonword(s)
NWRT(s)	Nonword repetition task(s)
PPVT-4	Peabody Picture Vocabulary Test
PS	phonological score (QUIR-DC)

QUIR-DC	Questionario per l'Identificazione del Rischio di Disturbo della Comunicazione, Questionnaire for the Identification of Risk for Communication Disorders
QUIS	Questionnaire for User Interface Satisfaction
RAN	Rapid Automatized Naming
RESTful	Representational state transfer
RS	Risk score (QUIR-DC)
RQ	Research question
SD	Standard deviation
SES	Socio-economic status
SLI	Specific Language Impairment
SLT	Speech and Language Therapy
SLTs	Speech and Language Therapists
SUS	System Usability Scale
t1	First time of testing
t2	Second time of testing
TD	Typically developing
TOWRE-2	Test of Word Reading Efficiency-Second Edition
TD	typically developing
UN	United Nations
WHO	World Health Organization
WSIR	Who says it right? (task paradigm)

1 Introduction

All over the world, bi-/multilingualism and multiculturalism are widespread and firmly rooted in science, pedagogical and speech therapy practice. This also holds for the context of Speech and Language Therapy (SLT). Assessing language competences of bilingual children however has been proven to be challenging, as shown both in policy reports (Garraffa et al. 2019) and survey studies with Speech and Language Therapists (SLTs, e.g. Bloder et al. 2021) especially pointing to the risk of misdiagnoses (Grimm & Schulz, 2014). This thesis faces the aforementioned topics in different ways: Previous research assessed the satisfaction and preparedness of SLTs with the service they provide for bilingual children through survey studies. In addition to that, a current study on this topic was conducted and is described in chapter 3. Strengthened by the agreement of SLTs concerning the application of bilingual and computerized screenings for automatic language and reading assessment in both languages spoken by bilingual children, the suitability of computerized screening tasks is empirically investigated in chapter 4. Furthermore, the features of a newly developed screening platform MuLiMi are presented. The core part of this thesis focuses on the preliminary validation of the screenings. In particular, chapter 5 evaluates and discusses their capability to detect the risk of Developmental Language Disorder (DLD) and Developmental Dyslexia (DD) in different language and age groups of multilingual Italian-speaking children. The selected languages Italian, Mandarin, German, English and Spanish represent different language families and degrees of orthographic depth allowing to assess the compatibility of the newly designed screening platform with different language characteristics and requirements. Moreover, in chapter 5.6.3.6 the issue of screening usability is addressed. All findings described in those chapters will be consolidated in chapter 6.

2 Theoretical background

To allow for adequate interpretation of the results of the survey and screening studies, context on the phenomena of DLD, DD and child bi-/multilingualism are given below (chapters 2.1, 2.2 and 2.3). Furthermore, an overview on the current literature available concerning the appropriate assessment of language and reading skills in bilingual children is provided (chapter 2.4).

2.1 Developmental Language Disorders

The term Developmental Language Disorders (DLD) refers to difficulties in understanding and/or producing speech or language and in communication that occur during childhood (World Health Organization [WHO], 2022a). Those difficulties are considered and classified as DLD when they cannot be traced back to regional, social, cultural or linguistic variations

and do not (exclusively) result from anatomical or neurological conditions. Note however, that the aforementioned internal and external factors do not rule out a diagnosis of DLD (Bishop et al., 2017). Furthermore, DLD diagnoses underlie the principle of persistence over time as well as the condition that the children's language performance significantly differs from the performance of the majority of their peers (WHO, 2022a). These language problems can refer to all of the linguistic areas affected (Bishop et al., 2017).

2.1.1 Prevalence (Epidemiology)

The prevalence of DLD is estimated at between 7% (Tomblin et al., 1997) and 10% (Norbury et al., 2016). This range can also be expected in the bilingual population (Grimm & Schulz, 2014).

2.1.2 Causes (Etiology) and risk factors

Despite early evidence for heritability of DLD (Gopnik & Crago, 1991; Rapin, 1996), the identification of causes of DLD is considered "complex and multifactorial" (Bishop et al., 2017). However, Newbury and Monaco (2010) did identify genes and gene clusters that are related to certain types of child language impairment, DLD (formerly Specific Language Impairment, SLI) included. These findings are extended by the observation that for one out of four children with developmental language delay a causative genetic diagnosis was found (Plug et al., 2021). Circumstances like family history of language impairment, male gender, and a low level of parental education and/or socioeconomic status (SES) are considered risk factors frequently associated with DLD (Arrhenius et al., 2018; Boivin et al., 2015; Sansavini et al., 2021).

2.1.3 Symptoms

According to Bishop et al. (2017), DLD may affect all linguistic areas. Thus, children may have problems related to the distinction and production of phonemes (phonetics and phonology) or the processing and realization of morphemes and syntactic structures (morphology and syntax). Furthermore, difficulties in associating words with meanings as well as retrieving them can be impaired (semantics and lexicon). Also, non-verbal communication and metaphorical use of language can be problematic in children with DLD (pragmatics). The linguistic deficits may also have an impact on the child's general communication abilities and, thus, might also affect the social development. DLD co-occurs with social-emotional, attentional, motor as well as with reading and spelling problems, but their causal relationship is unclear (Bishop et al., 2017). Overall, linguistic, communicative and social difficulties are associated with later scholar achievements and professional career. In particular, children with DLD are more likely to also show deficits in reading and writing acquisition and, DLD is a risk factor for

DD. This relates to potential long-term effects of DLD related to poorer academic performance and lower occupational status (Johnson et al., 2010).

Beyond the concrete symptoms, also long-term effects of DLD are discussed in the literature. For children whose language impairment persisted above the age of 5;5, the occurrence frequency of social and attentional difficulties was shown to be increased (Snowling et al., 2006). These observations relate to the United Nations' Sustainability Goals concerning "Goal 3: Good health and well-being", "Goal 4: Quality education" and, in the long run, also "Goal 8: Decent work and economic growth" (United Nations [UN], 2022).

2.2 Developmental Dyslexia

While formal reading and writing instruction starts at five to six years of age in most countries, reading-related skills are generally developed before (Luk & Bialystok, 2008). Despite these rough indications, reading acquisition largely varies individually due to language and orthography characteristics as well as schooling and home environment. However, typically developing (TD) children can be distinguished from children with a clinical condition regarding an impairment of reading acquisition referred to as Developmental Dyslexia (DD) or as "Developmental learning disorder with impairment in reading" according to the International Classification of Diseases (ICD-11, WHO, 2022b). Since this classification was more recently published, in this work the ICD-11's classification is applied (see chapter 2.2.3).

2.2.1 Prevalence (Epidemiology)

A recent study, Di Folco and colleagues (2020), compared the classification according to the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5, American Psychiatric Association, 2013) to the ICD-11 classification (WHO, 2022b). According to Di Folco and colleagues (2020), the prevalence of DD is ranging between 2 and 20%. A notable difference is found concerning variability in orthographic depth (Seymour et al., 2003), namely regarding the continuum from shallow or transparent orthographies (conversion of oral language into script comes with high correspondence of phonemes and graphemes, e.g. in Italian, written "ramo" (branch) is pronounced ['ra.mo] over deep (less or less consistent correspondence between phonemes and graphemes, e.g. in English, written "which" and "witch" both are pronounced [wɪtʃ]) to logographic (symbols or characters standing for words or word units, e.g. in Mandarin, written "也" (also) is pronounced [je], Frost et al., 1987; C.-F. Hu & Catts, 1998; Seymour et al., 2003). Accordingly, depending on the characteristics of languages and orthographies, different tasks can be more or less suitable for the detection of DD in a certain language (Landerl et al., 2013).

According to a recent study by Barbiero and colleagues (2019) prevalence of DD in Italy is estimated at 3.5%, while the number of children who are actually holding diagnoses is lower: “[in] two out of three children with dyslexia the disorder had not been previously diagnosed”. Even though German authors provide different prevalence estimates (2 to 12%), they describe DD as a very common, if not the most common, developmental disorder (Schumacher et al., 2007; Siegmüller & Heide, 2011).

2.2.2 Causes (Etiology) and risk factors

There is scientific evidence for a genetic predisposition of DD (Bishop, 2015). This has been shown in studies considering the cumulative occurrence of DD in families (Thompson et al., 2015; Wolff & Melngailis, 1994), in studies analysing genes with results on “inherited factors [...] estimated to account for up to 80%” (Schumacher et al., 2007). These studies indicate high family risk and, accordingly, the genetic component is considered the main cause of DD. In subjects with DD, differences were also found on a brain-structural and -functional level (Skeide et al., 2015). More specifically, neurobiological evidence suggests underlying phonological and sensorimotor deficit (Ramus, 2004). Such differences were found even before the beginning of formal reading instruction and are thus considered as reflective of the etiology, too (Norton et al., 2015). At the same time, there are many linguistic and behavioural influencing factors like interaction of reading skills with literacy interest and literacy-promoting practices (Hume et al., 2015) and parental reading skills (Torppa et al., 2011). In general, there is no evidence that multilingualism per se increases the risk of reading difficulties (Everatt et al., 2000).

2.2.3 Symptoms

According to the aforementioned classification by the WHO (2022b), this impairment refers to difficulties in reading fluency, accuracy and comprehension, that persist over time, while intellectual functioning is not impaired. These skills are measured and compared to TD children of the same age (WHO, 2022b). Reading errors and slow reading pace are generally considered core symptoms of DD (Adlof & Hogan, 2018). More specifically, often the transfer from auditory into written information (or vice versa) is impaired (Costenaro & Pesce, 2012).

Beyond the concrete symptoms, the ICD-11 (WHO, 2022b) mentions long-term effects of DD on academic and occupational achievements that also have to do with the UN Sustainable Development Goals (UN, 2022). It mentions, in particular, long-term effects concerning the inclusion in schools and academic achievements (Goal 4 “Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all”), as well as professional

career (Goal 8 “Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all”; UN, 2022). A formal diagnosis entails that the symptoms described above cannot (exclusively) be explained by another internal factor like impaired intellectual development, neurological and/or sensory impairment (vision or hearing), limited access to education and/or the language of instruction or psychosocial adversity (WHO, 2022b).

2.3 Child bi-/multilingualism

Grosjean and Li (2013) define bi- and multilingualism¹ as the circumstance in which a person is confronted with more than one language in everyday life, irrespective of the proficiency level in any of the languages (Stow & Dodd, 2003). Linguistic diversity within geographically close regions, globalization as well as migratory processes result in multilingualism having become normal in everyday life all over the world (Romaine, 2017). On a societal level, it is relevant to mention that multilingualism is also a wide-spread phenomenon in particular in educational and SLT contexts (Marini, 2019).

2.3.1 Terminology and subtypes of bilingualism

The L1 refers to the first or heritage language (Polinsky & Scontras, 2020) that a person is confronted with from birth in the family context (thus also home language, Giguere & Hoff, 2020) or family language, often mentioned in the context of family language policy (see Surrain, 2021). The L2, in contrast, refers to the second or societal language (Giguere & Hoff, 2020) that bilinguals encounter mostly outside the family context for example at kindergarten, school or at work. Beyond terminology, the distinctions of L1 and L2 are less clear-cut in real life. It has been known for many years that this distinction does not exactly reflect the contexts of use of the languages spoken: mixing languages in families is a frequently occurring phenomenon and does not imply “interlinguistic confusion” (Goodz, 1989).

Further, the chronology of bi- or multilingual language acquisition is represented in terminology. Critical periods (Werker & Hensch, 2015) may have an impact on how well children manage (bi-/multilingual) language acquisition at a certain point in time. Generally, there is agreement that children’s second language acquisition differs from first language acquisition for children above age seven (Meisel, 2011). In child bi-/multilingualism research, also for children below the age of 7;0, further distinctions are made: Simultaneous bilingualism refers to exposure to both L1 and L2 from or shortly after birth (also referred to as 2L1, see

¹ Note that in the theoretical background, bi- and multilingual(ism) are used interchangeably. In the studies described in chapter 5 only children speaking two languages were included and they are referred to as bilingual.

Schulz and Grimm (2018)). In the context of child (second) language acquisition, sequential bilingualism refers to exclusive exposure to L1 in the first years of life. In the case of first exposure to L2 in early childhood this is also referred to as eL2 (“e” standing for “early” Schulz & Grimm, 2018). Besides this distinction of the so-called age of onset (AoO), also quality (Tsimpli, 2014) as well as quantity of input as measured e.g. through length of exposure (LoE, Roesch & Chondrogianni, 2016) are taken into consideration. The concept of amount and quality of language exposure is related to the construct of language dominance (referring to whether or not performance in one of the languages spoken is higher than in another one, see Treffers-Daller, 2019, for an overview).

Even though in general, the milestones of language acquisition in bi-/multilingual children coincide with the ones during monolingual language acquisition (Kohnert, 2010; Meisel, 2006), the overall process is not comparable (Oller et al., 2007) and can be accelerated or decelerated by the diversity and variability of language input (Unsworth, 2016). In general, the acquisition of languages under diverse conditions regarding individual language contact patterns leads to highly variable expressions of the child’s language competences (Carroll, 2017). Depending on the language combination and whether or not the languages are similar, positive or negative transfer effects – also referred to as cross-linguistic influence (CLI) – can be observed in both child (Leśniewska & Witalisz, 2014) and adult language (Kupisch, 2019), as well as in reading acquisition (Daniels & Share, 2018; Koda, 2007). In fact, there is evidence that language and literacy acquisition do not necessarily happen in parallel, but inter-dependently – positive transfer effects can be observed (Cummins, 2012). For adult L1 and L2 reading-related performance, this was also shown at the neural level (Tan et al., 2003). A second, similar study suggested that these effects may depend on the characteristics of the languages and their specific combinations (Jasińska et al., 2017). Taken all these factors together, clinicians, teachers, multilingual families, as well as children themselves observe that bilingual children’s language performance patterns are highly individual and hardly comparable (Kohnert & Danahy, 2007). Besides the heterogeneous input- and exposure patterns, also language attrition (Kim & Kim, 2022), especially in the context of DLD (Blom et al., 2019) and language change over time – in particular in the light of the establishment of community as well as prestige languages in (migrant) communities – play a role when studying bilingualism. In fact, knowledge on intra- (language erosion) and interindividual (language change) language differences is required for contextualization and adequate evaluation of the child’s language performance (Gagarina, 2014). Furthermore, aforementioned variables referring to proficiency and usage were also shown to interact with each other (Luk & Bialystok, 2013).

2.3.2 Child bi-/multilingualism in Europe and its implications for the educational & health sector

Recent survey studies discuss a high proportion of multilingual children frequenting SLT services (Scharff Rethfeldt, 2019). At the same time, SLTs are not confident with the current situation and SLT service provided to multilingual children (Stankova et al., 2021).

Ever since the guest worker program from the 1950s to 1970s (Heckmann, 1981), Germany has experienced several waves of immigration (Baumert & Maaz, 2012). In Italy, instead, immigration is a rather recent (early 1990s) phenomenon, but more than 24 million people have emigrated from the country up to the late 1970s (Fondazione ISMU, 2022).

In 2020, the foreign-born population in Germany was at 18.1% (15.0 million persons) and at 10.3% (6.2 million persons) in Italy (Eurostat, 2022). These numbers, however, do not contain all people of foreign descent/with migration background – accordingly, the amount of speakers of a heritage language are expected to be higher. In Germany, the five most prevalent countries of origin are Turkey, Poland, Syria, Romania and Italy. In Italy, the five most prevalent countries of origin for Italian immigrants are Romania, Albania, Morocco, China and Ukraine (Eurostat, 2022). More specifically, among the foreigners living in the Italian region of Lombardy, 11.8% come from Central and Southern America – countries where Spanish is (one of) the official language(s). Most Spanish-speaking immigrants to Lombardy come from Peru (ca. 4% of the total foreign population in Lombardy). Furthermore, in Lombardy, ca. 6% of the foreign population comes from China (Istituto Nazionale di Statistica [ISTAT], 2021). In Tuscany, 17.3% of the foreign population is Chinese (ISTAT, 2022).

2.4 Language and reading assessment in bilingual children

Cummin's conceptualization of the "Common Underlying Proficiency" (CUP, Cummins, 1980, 1981) highlights that in bilingual children, the two languages do not develop independently from one another, but they are interrelated. In empirical studies, this model was recently confirmed for bilingual oral language (Cummins, 2016) and code-related (i.e. pre-reading) skills (Goodrich & Lonigan, 2017). However, the model of CUP is not consistently resembled and incorporated when assessing and evaluating language and reading skills in bilingual children.

It has been known for several decades that there is disproportionate occurrence of language and literacy problems in children from families with low SES or migrant background (Laing & Kamhi, 2003). Despite evidence-based statements that multilingualism does not cause language impairment (Paradis et al., 2003; Paradis, 2016), associations between language disorders and a migrant background are continuously found (Lehti et al., 2018). They

further appear to be related to the concept of over- and underdiagnosis (Grimm & Schulz, 2014). Misdiagnoses refer to two different concepts: In the case of overdiagnosis, Paradis (2005) speaks of “mistaken identity” (bilingual children who are mistakenly considered as having DLD) and about “missed identities” in the case of underdiagnoses (bilingual children who are mistakenly considered as TD, while in fact they have DLD). Diagnostic procedures that are inappropriate for bilingual children may lead to such misdiagnoses (de Lamo White & Jin, 2011). In particular, Paradis (2005) showed how the application of English standardized language tests with children who are English language learners can lead to mistakenly diagnosing children who have no language impairment, but weak or weaker English skills due to specific language exposure patterns that differ from monolingual children. Misdiagnoses carry the risk that no, insufficiently early or inappropriate language support and therapy measures are initiated (Werker & Hensch, 2015). This is a relevant notion considering the ongoing research concerning the “optimum starting point for intervention” (Law et al., 2003).

Several solutions to solve this issue are discussed in the scientific literature. In their review, Ebert and Kohnert (2016) explain how language assessment ideally considers all languages spoken, but acknowledge that this poses a particular challenge to clinicians that can be overcome through the application of suitable methods and resources. De Lamo White and Jin (2011) claim the reduction of biases in contexts or assessment material that puts multilingual children at disadvantage and suggest to take environmental factors influencing the child’s language use into account. More specifically, language test results should be interpreted in the context of the LoE and the AoO of the language of assessment (Cline, 2000; Letts, 2013). Bedore and Peña (2008) suggest to not only take duration and frequency of language contact, but also cultural background into account when choosing, administering and evaluating language tests and in particular when referring to normative data. With the aim of reduction of biases resulting from inappropriately applying monolingual normative data also Laing and Kamhi (2003) make suggestions on alternative solutions, e.g. dynamic assessment procedures, see MultiMind ITN (2022) for an overview.

Ever since the scientific community has been heavily lamenting the inappropriate application of standardized tests designed for monolingual children with multilingual children, the availability of standardized testing material in the L2 providing norms for multilingual children has increased. Standardized testing material in the L2 refer both to reading assessment (see e.g. for Italian: ALCE. Assessment di Lettura e Comprensione per l’Età Evolutiva, Bonifacci et al., 2014) and to language assessment (see e.g. for German: Linguistische Sprachstandserhebung Deutsch als Zweitsprache (LiSe-DaZ), Schulz & Tracy, 2011).

Testing all languages spoken by a child, as recommended in recent policy reports (Garraffa et al., 2019; MultiMind ITN, 2022), is more and more considered a solution (see for Italian the “Prove per la valutazione delle competenze verbali e non verbali in bambini bilingui”, BaBiL, Contento et al., 2013; for German the “Evozierte Diagnostik grammatischer Fähigkeiten für mehrsprachige Kinder”, ESGRAF-MK, Motsch, 2011). Furthermore, there are research attempts for the adaptation of the Italian standardized test battery “Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni” (BVL, Marini et al., 2015) in other languages. The German version is to be applied with bilingual children living in the North-Eastern part of Italy around Bolzano, where both Italian and German are spoken in family and educational contexts (Marini et al., 2019). Being able to assess the child’s L1, however, does not imply that the assessment of the L2 can be neglected, but should be considered because of its importance in the children’s every day and school life in- and outside the family (Jordaan, 2008). Similar attempts are also made within the development of the LITMUS-tasks (Language Impairment Testing in Multilingual Settings, BISLI Cost Action, 2022). There, tasks are constructed in various languages in a comparable way. In the case of Cross-linguistic Lexical Tasks (CLTs, Haman et al., 2017), for example, items to test children’s receptive and productive lexical skills were comparatively selected across languages. In their computerized version (Zinn, unpublished) CLTs can be automatically administered for bilingual children.

Besides the efforts concerning theoretical concepts, policy papers as well as the construction, standardization and publication of suitable assessment methods, it is relevant to consider the particular perspectives of practitioners, i.e. the clinicians in charge of the diagnosis. Despite the (expected) high proportion of multilingual children frequenting SLT services, a current survey study shows that the particular needs described above do currently not seem to be consistently met (Scharff Rethfeldt, 2019). It appears to be common to assess and treat the children exclusively in their L2 (Williams & McLeod, 2012), not taking the children’s L1 into account (Jordaan, 2008; Stankova et al., 2021) even though SLTs do not agree with this practice (Marinova-Todd et al., 2016). In general, the majority of SLTs does not feel confident concerning the treatment of multilingual children from various cultural backgrounds (Stankova et al., 2021) and acknowledges their lack of knowledge concerning appropriate normative data, reliable diagnostic procedures, competence in providing guidance to the children’s caregivers (Grandpierre et al., 2018) and children’s heritage languages (Roseberry-McKibbin et al., 2005). More precisely, SLTs believe that it would be beneficial for them to speak another language (Williams & McLeod, 2012) and when it is possible to speak to them in another language, this is more often done (D’Souza et al., 2012). The support of interpret-

ers however is not easily accessible (D'Souza et al., 2012). Trying to overcome the aforementioned shortcomings concerning adequate resources and (self-)satisfactory skills, often informal procedures are applied (Williams & McLeod, 2012). Overall, the results of the survey studies generally imply the multilingualism-friendly attitude of SLTs. Questionnaire studies showed that they, for example, tend to not consider multilingualism a risk factor of DLD. SLTs, thus, usually do not advise families to sacrifice the L1, but instead recommend active use of the language they know best (Williams & McLeod, 2012). This attitude and resource-oriented recommendation concerning family language use represented in their questionnaire responses might be explained by an increase of input on this topic in academic SLT training (Roseberry-McKibbin et al., 2005). In our recent study (see chapter 3), a series of factors like country of place of work (migration history and health care system) and experience in working with bilingual children were found to be relevant factors influencing the SLTs practices and approaches when providing SLT to bilingual children (Bloder et al., 2021).

2.4.1 Language assessment in bilingual children

The identification of DLD in bilingual children has been described as a diagnostic challenge (Armon-Lotem, 2012). Also, in the current classification of DLD (WHO, 2022a), particular attention is drawn to the specific diagnostic requirements relating to the reduction of biases in the assessment of bilingual children. This is facilitated through the application of assessment tasks that are less likely to be influenced by language exposure (phonological rather than lexical assessment) as well as the application of language background questionnaires (see chapter 2.4.3).

Since DLD affects all languages spoken by a child (Bishop et al., 2017), accurate diagnoses would ideally require assessing the child's language skills in both languages to distinguish between differences in bilingual language profiles vs. clinical conditions (Kohnert & Medina, 2009). Since clinicians and teachers typically do not speak all languages spoken by the child, in their recent policy report, Garraffa and colleagues (2019) explain how language-specific clinical markers and grammatical features, in particular, can be used to differentiate between the linguistic profiles of monolingual children with DLD and those of multilingual TD children.

2.4.1.1 *Language-specific clinical markers for DLD diagnosis in bilingual children*

In this section, the most relevant clinical markers that are used to detect DLD for the languages considered in the screening studies (see chapter 4 and 5) are described.

Clinical markers for DLD in Italian. In Italian, the omission of articles and clitics and the replacement of the third person singular of the verb with the third person plural are considered clinical markers (Bortolini et al., 1997; Bortolini et al., 2002; Bortolini et al., 2006; Dispaldro et al., 2013) and have been (partly) confirmed for the bilingual population (Guasti et al., 2021; Vender et al., 2016).

Clinical markers for DLD in German. Also, Clahsen et al. (1997) found subject-verb agreement in German to be a clinical marker of DLD, which was recently confirmed by Rothweiler and colleagues (2017), Ruberg and colleagues (2020) and Scherger (2022) for bilingual children. Furthermore, in German, case marking can be considered a clinical marker of DLD (Hasselaar et al., 2019), and, thus is also taken into consideration as potential clinical marker of DD. Despite large variance of multilingual children concerning their performance in case marking (Ulrich et al., 2021), the suitability of case marking as a clinical marker has been confirmed for bilingual children by Schönenberger and colleagues (2012) and Scherger (2015). Moreover, Scherger (2022) further highlights the impact of age- and time-related input variables of multilingual language acquisition on performance in these tasks.

Clinical markers for DLD in Spanish. In Spanish, the third person plural is considered a clinical marker (Bedore & Leonard, 2001). Furthermore, Grinstead and colleagues (2014) also found the use of infinitive verbs instead of inflected verbs to be an indicator for DLD.

2.4.1.2 *Language universal clinical markers for DLD*

Within the BiSLI COST Action IS0804 (2022), language-universal tasks assessing linguistic performance independently of the language(s) a child speaks were developed. In their study, Boerma and Blom (2017) found performance on such tasks (nonword repetition and narrative skills) to be associated with (later) language proficiency.

Nonword repetition. Nonword repetition tasks (NVRTs) have been shown to be reliable diagnostic tools for both mono- and bilingual children (Schwob et al., 2021). In NVRTs, children first listen to a so-called nonword (NW). NWs are strings of phonemes that in the language of assessment do not have a meaning, but are constructed in accordance with the language's linguistic features such as they could be real words in that language. Upon the presentation of a NW, the child repeats it and an examiner evaluates the accuracy of the child's repetition performance. According to Ebert (2014), discrimination, encoding and production of phonemic sequences are the processes involved in NW repetition. Children's performance in NW repetition is associated with other language measures concerning lexicon (Farabolini et al., 2021; Hoover & Storkel, 2006) and grammar (Rispen & Been, 2007). Thus, NVRTs were shown to have the potential to identify language difficulties. The repetition of

NWs appears to be particularly useful for their application with bilingual children since the repetition performance of items that are free of meaning in nature is less influenced by language exposure patterns and language-specific trajectories of phoneme inventories (Chiat, 2015). However, studies have shown that experience or the lack of experience with a certain language does play a role when the NWs administered are based on characteristics of this language (Boerma et al., 2015). This is why a series of studies investigates the particular benefit of repetition tasks that include NWs that are not language-specific (Chiat, 2015). The role of language-specific vs. non-specific NWs however is not clear: On the one hand, several studies showed that language experience does not have an influence on repetition performance (de Almeida et al., 2017; Tuller et al., 2018). On the other hand, in their study Abed Ibrahim and Hamann (2017) found that repetition performance of children with DLD was worse compared to TD children for all NW categories, but that this was more evident for language-specific than for non-language-specific NWs. Besides the idea of specifically designing language-specific and non-language-specific NWs to be applied in repetition tasks with multilingual children, other research projects addressed the suitability of well-established standardized NW repetition tests for their application with multilingual children. Wild and Fleck (2013), for example, provide norm data including bilingual and monolingual children for the German Mottier-Test, a nonword repetition test including 30 nonwords with simple consonant-vowel combinations consisting of two to six syllables. However, they provide separate norm data for mono- and bilingual children respectively for the age group of 5-year old children only. No norm data for children under the age of 5;0 is provided. Also, in this study, norm data from bilingual children with various language backgrounds is provided, not taking into consideration specific language characteristics and/or exposure patterns.

Dynamic Assessment and Fast Mapping. Hunt and colleagues' (2022) recent review indicates – besides further research needs – Dynamic Assessment (DA) to be a “suitable and time-efficient complementary method of diagnosis of language disorder in multilingual children” (Hunt et al., 2022). DA refers to the assessment of learning potential by systematically giving and recording feedback (Camilleri et al., 2014). Typically, the concept of “test-teach-re-test” is applied (Gutiérrez-Ciellen & Peña, 2001). A recent meta-analysis including bilingual children has shown that overall, DA has good diagnostic accuracy (Orellana et al., 2019). One subtype of DA is called “Fast Mapping” referring to the familiarization and consolidation of associating an unknown word or NW with a meaning or object (Girbau, 2016). In their study, Jackson and colleagues (2016) showed that children with impaired phonological short term memory and lexical deficits – core symptoms of DLD – also have problems in this kind of task. Concerning the application of Fast Mapping tasks with bilingual children, some studies

point to the particular suitability of this task (Roseberry & Connell, 1991), while other studies indicated that also bilingual TD children show difficulties in this task (Kohnert & Danahy, 2007). Kan and Kohnert (2008), however, point to bilingual children's learning potential for such tasks in their L1 being higher than in their L2.

2.4.2 Reading assessment in bilingual children

Reading accuracy and speed are also commonly assessed in standardized reading tests. For this assessment, in alphabetic languages usually both words and NWS (strings of graphemes following the phonological and orthographic constraints of a language system) are applied (see for Italian: Sartori et al., 2007; for English: Torgesen et al., 2012). In addition to reading accuracy and speed, the German Zürcher Lesetest (ZLT-II, Petermann & Daseking, 2019) also assesses reading-related skills like rapid naming and NW repetition, since children with DD often show impaired language skills alongside their reading and writing difficulties. This oftentimes is of relevance especially in the context of early diagnoses of multilingual children, see below). Those difficulties refer to phonological development (Durkin, 2000), the repetition of NWS (Vender et al., 2020) as well as morphosyntactic structures, in the case of Italian clitics (Arosio et al., 2016). In their paper, Arosio et al. (2016) confront the issue of a potential underidentification of DLD in poor readers. In general, there is agreement on the co-occurrence and association of DLD and DD (Del Tufo & Earle, 2020; Lautenschläger et al., 2020). In clinical practice, the observation of co-occurrence of reading and language impairment also comes with the application of linguistic clinical markers as early predictors of potential future reading impairment. These relate to basic abilities at the phoneme and syllable level (Bastien-Tonizzo et al., 2010; Durkin, 2000), word level (Quinn et al., 2015), as well as more complex skills at sentence or text level (Hulme et al., 2015); see Eikerling and Wendt (2016) for an overview. These observations were generally confirmed, but informative value and appropriateness may vary in bilingual children (Riva et al., 2020).

Reading is based on a series of skills related to language knowledge (Cummins, 1980, 1981) and accordingly, learning to read in the L2 requires (meta-)linguistic knowledge in both languages (Koda, 2007). Due to the DD underdiagnoses as indicated by Barbiero and colleagues (2012), diagnosing DD seems not only to be challenging for multilingual, but for all children. However, the variables resulting from child bi-/multilingual language and reading acquisition pose particular challenges to making reliable diagnoses (Everatt et al., 2000; MultiMind ITN, 2022). The complexity relates back to the specific circumstances of bilingual language and reading acquisition (chapter 2.3). In particular, variability of languages and orthographic systems as well as their influences on each other, see Kroll and Bialystok (2013),

along with issues related to appropriateness of testing material as described above for the case of language assessment (chapter 2.4.1) are influencing factors. In addition to that, language skills are normally acquired within the child's family (L1) and consolidated in the context of schooling, while formal reading and writing instruction takes place predominantly in schools and thus usually in the children's L2. Accordingly, depending on the domain under assessment, the children's performance can be evaluated in both languages (Gersten & Geva, 2003; Geva, 2000). Yet, the assessment of reading and writing as such depends on the literacy acquisition that usually takes place in the children's language of instruction, i.e. the societal language (L2). Task types that appear to be applicable both as well in the L1 as in the L2 across languages are Rapid Automated Naming (RAN) and phonological awareness (Landerl et al., 2013).

Rapid Automated Naming (RAN) refers to a group of tasks in which children name a limited set of repeatedly visually displayed familiar objects, colours or digits, as fast as they can. It was found to be predictive of later reading abilities (Norton & Wolf, 2012). Rapid automated digit naming was found to be a good predictor also in bilingual readers (Li et al., 2012), but there are mixed findings regarding RAN's relation with linguistic and orthographic influences. Differences for RAN (objects) in groups acquiring orthographies of varying orthographic depth were found by Georgiou and colleagues (2012). But based on their meta-analytic review, Araújo and Faísca (2019), however, highlight that neither orthographic complexity nor the writing system seem to have an influence on RAN performance.

Phonological awareness refers to the ability to process sound-related properties of spoken words irrespective of their meaning. For example, recognizing and manipulating phonemes, syllables, rhymes or word onset. It has been found to be impaired in DD and moreover to be an early predictor of later reading skills (Kenner et al., 2017; Melby-Lervåg et al., 2012). Phonological awareness appears to be independent of the language (Da Silva et al., 2020) and orthographic system acquired (Ho & Bryant, 1997; C.-F. Hu & Catts, 1998). Besides the processing of sound-related properties of spoken words irrespective of their meaning, the repetition of NWs (see chapter 2.4.1.2 for their application in DLD risk identification) is considered a valid indicator for DD in monolingual children (see Mottier-Test included in the German standardized reading test ZLT-II, Petermann & Daseking, 2019) and in bilingual children (Vender et al., 2020).

2.4.2.1 Language-specific risk indicators for DD diagnosis in bilingual children

DD is believed to generate from the same brain functional basis cross-culturally, but vary in its characteristics and extent due to cultural differences (Paulesu et al., 2001). This highlights

– similar to what was described for DLD in chapter 2.4.1.2 – the need for language-specific risk indicators. While RAN and phoneme deletion tasks seem to be suited for the detection of DD across languages (Landerl et al., 2013) along with word and NW reading accuracy and speed for alphabetic orthographies, there are specific, well-tested risk indicators for each language. Besides phonological development (Durkin, 2000), those extend to morphological (Siegel, 2008), syntactic and lexical skills (Ben-Dror et al., 1995). Since these differ from language to language, they are described individually for the various languages under consideration in this work.

In Italian, difficulties in subject-verb agreement (Cantiani et al., 2013) as well as comprehension and production of clitic pronouns (Arosio et al., 2016; Vender et al., 2018) are considered risk indicators for DD (see chapter 2.4.1 for overlap with clinical markers of DLD; see Arosio et al. (2016); see Ramus et al. (2013) for discussion on co-occurrence vs. underidentification of DLD in poor readers). Vender and colleagues (2018) extend the findings regarding the suitability of clitics in DD risk identification also to bilingual children.

In English, past tense is considered both a clinical marker for DLD and a risk indicator for DD (Robertson et al., 2013). The study by Jacobsen and Schwartz (2005) also suggest the suitability of past tense as clinical marker of DLD in bilingual children and, thus, is also likely to be a suitable clinical marker of DD in bilingual children. In German, case marking can be considered a clinical marker of DLD (see chapter 2.4.1.1).

In Mandarin, poor readers were shown to perform less well in the comprehension of negative sentences (S. Hu et al., 2018). It is further noteworthy that despite Mandarin not being an alphabetic language, phonological processing skills have been shown to relate to reading skills (C.-F. Hu & Catts, 1998). Accordingly, as well as in the aforementioned languages, also for Mandarin, language and phonological processing tasks are commonly used in DD (risk) detection. In Mandarin, a logographic script, tasks on the detection of incorrectly placed radicals (Chung et al., 2010) as well as on the detection of inverted characters (Chung & Ho, 2010) can be applied for DD risk identification.

2.4.2.2 Dynamic Assessment

Dynamic Assessment (DA) refers to the potential of learning. Accordingly, DA assesses the ability of readers to recognize and recall new information and is thus to be considered more independent from concrete language knowledge (see chapter 2.4.1.2). Hence, due to relative independence of specific language exposure patterns, it is suited for the application in the multilingual population, as shown both for multilingual children (Gellert & Elbro, 2018) and adults (Elbro et al., 2012). Applying DA in the context of reading assessment, examinees are

trained to learn a new orthographic code, i.e. to recognize, synthesize and read novel letters: teaching and testing the new skills is not happening sequentially, but alternately (Haywood & Lidz, 2006). During DA, assistance such as corrective feedback and repetitions are systematically provided by the examiner and recorded (Gellert & Elbro, 2018). These principles can also be implemented in computerized DA tasks and applied with children of young age (Horbach et al., 2015) to predict successfully later reading outcomes (Horbach et al., 2018). Suitability of DA for young bilingual children was confirmed in a study with Spanish-speaking children (Petersen & Gillam, 2015).

2.4.3 Caregiver questionnaires

Caregiver questionnaires fall under the category of indirect assessment opposed to direct assessment in which language or reading skills as such are tested. They are commonly used for the identification of the need of an SLT intervention, see for example the MacArthur Bates Communicative Development Inventories questionnaires (Heilmann et al., 2005) that have been adapted in many languages (MacArthur-Bates Communicative Development Inventories [MB-CDIs], 2022; Mayor & Mani, 2019).

The application of caregiver questionnaires is particularly recommended in language and reading assessment of bilingual children since (as mentioned in chapter 2.4 and claimed by de Lamo White and Jin, 2011) children's language skills need to be contextualized according to a series of factors related to quantity and quality of language exposure shaping bilingual child language acquisition (see chapter 2.3). Furthermore, caregiver questionnaires allow for systematic investigation of caregivers' impressions of the child's L1 language skills which is of particular benefit when L1 language performance cannot be assessed by the examiner because of a lack of appropriate resources and/or language knowledge. Pua and colleagues (2017) extend this observation to reports on language performance by the children's teachers.

Several questionnaires specifically designed for their application in the assessment of bilingual children are developed and validated across languages (Bonifacci et al., 2016). Evidence was found that caregivers' questionnaire responses on early language experience (Paradis et al., 2010), current language exposure (Paradis, 2011) and on both early experience and current language exposure (Tuller, 2015) are associated with children's performance in standardized language tests. It is worth mentioning that the combination of caregiver questionnaires such as the Alberta Language and Development Questionnaire (ALDeQ, Bonifacci et al., 2016) with direct language measures increases the accuracy in identifying DLD (Bonifacci et al., 2020).

2.4.4 Computerized screening tools

Besides the aforementioned possibilities to apply indirect methods in language and reading assessment protocols (see chapter 2.4.3) and to use non-language-specific and dynamic assessment tasks (see chapter 2.4.1.2 and 2.4.2.2), direct assessment of language-specific tasks in all languages spoken can be realized through automatic scoring and administration of such tasks. Besides the aforementioned computerized solutions for dynamic assessment for children attending kindergarten (Horbach et al., 2015), solutions are also provided for computerized reading assessment (Haridas et al., 2017), optionally with game-based elements (Hautala et al., 2020). Advantages are seen in the circumstance that administering computerized screenings circumvents the need for pen-and-paper documentation and evaluation of the children's reactions and, additionally, allows for remote testing (Hodge et al., 2019). Previous studies have shown that reading results of such screenings are (at least to a certain extent) significantly associated with the performance levels in standardized tests (Brookes et al., 2011).

Compared to primary school age children, younger children are less frequently considered in literature on the diagnostic and interventional use of innovative technologies. Nevertheless, there is evidence for their suitability in SLT intervention (Zwitzerlood et al., 2022) and in the screening of speech production (Speakaboo, 2022) as well as reading-related skills (Rauschenberger et al., 2019). Different from the computerized reading screenings mentioned above, Speakaboo (2022) specifically targets articulatory and phonology assessment of multilingual children. The examiner needs to familiarize with the target sounds in the children's language(s) and then manually evaluate the child's production performance. Rauschenberger and colleagues (2019) also provide insight in the construction of such an application and give examples for the conduction of usability studies. Usability studies can be carried out with the goal of standardized assessment of usability as perceived by users. For this purpose, the "System Usability Scale" (SUS, Sauro, 2022) and different versions of the Questionnaire for User Interface Satisfaction (QUIS, Wallace et al., 1988; Chin et al., 1988) are among the most common tools.

2.5 Structure of the thesis

The previous sections highlight the relevance of multilingualism all over the world. Furthermore, it is pointed out that the identification of DLD and DD in bilingual children comes with certain obstacles and the requirement to assess all languages spoken to prevent misdiagnoses. Besides that, previous studies suggest the potential of computerized screenings also in young children. The work presented here, thus, starts with a study on whether SLTs already

sufficiently do test both languages spoken by a bilingual child and whether they would use computerized screenings to do so (chapter 3). It will be furthermore addressed, whether computerized assessment of language and reading skills is applicable with Italian-speaking bilingual children as indicated by associations between standardized and computerized screening tasks (chapter 4). The core part of the thesis systematically assesses the potential of the novel screening platform MuLiMi in DLD and DD risk identification (chapter 5). All studies were approved by the IRCCS Medea's ethics committee and all caregivers signed informed consent forms according to the Declaration of Helsinki.

3 Attitudes and practices concerning multilingualism in the context of SLTs' diagnoses and interventions

A modified version of this chapter has been published in the peer-reviewed open access article "Speech and Language Therapy Service for Multilingual Children: Attitudes and Approaches across Four European Countries" (Bloder et al., 2021) in the special issue "Multidisciplinary Approaches to Multilingual Sustainability" of the open access journal "sustainability". While the paper cited contains a series of comparisons across variables as well as as well as investigations related to the country, for this chapter, only the parts relevant are mentioned.

The aim of this survey study was to investigate the current situation of attitudes and practices concerning multilingualism in the context of SLTs' diagnoses and interventions following up on observations by previous research in other countries (see chapter 2.4). For this study, Italy and Germany were chosen since they are the countries where children were recruited and tested for the screening studies (see chapter 5). The German questionnaire was also distributed in Austria and the German-speaking part of Switzerland due to the linguistic compatibility (German as official language) and differences in the health care and educational political system concerning the funding and responsible institution of SLT service provision

3.1 Participants

A total of $N = 300$ SLTs were included in this study. Of those, $n = 103$ lived in Italy and $n = 197$ were from a German-speaking country. More specifically, $n = 45$ SLTs lived in Austria, $n = 85$ lived in Germany and $n = 67$ lived in the German-speaking part of Switzerland.

3.2 Questionnaire design

Data was collected using a questionnaire consisting of 24 questions developed in Italian and piloted in Italy. This questionnaire was then translated into German.

The 24 questions address SLTs beliefs and practices concerning multilingualism in the context of diagnosis and SLT interventions they provide, using multiple-choice, open questions and closed (yes-no) questions. More specifically, the questionnaire targets experience in working with multilingual children, beliefs towards DLD in multilingual children, approaches applied when diagnosing and treating multilingual children and perceived barriers of effective assessment and treatment.

3.3 Data analysis

Data was analysed using IBM SPSS Statistics v.20. Chi-square goodness-of-fit analyses were applied to assess the prevalence of responses within the overall response distribution. Whenever responses could be categorized as ordinal variables (i.e. when the response options were “never, rarely, sometimes, yes” or similar), Somer’s D and gamma statistics were applied. In a first step, the general attitudes and practices were assessed for the whole group. In a second step, the association between responses given was evaluated.

3.4 Procedure

In Italy, an online version implemented on “Google Forms” (<https://docs.google.com/forms/>) as well as a pen-and-paper version of the same questionnaire were created and distributed. For the German-speaking countries, an online version was implemented on “qualtrics” (<https://www.qualtrics.com/>). Participant recruitment took place through the sharing of the respective link on social media, mailing lists as well as newsletters and websites of SLT associations. Data was collected completely anonymously and was initiated only upon the participants’ consent to participate and their anonymous and aggregated responses to be used for scientific purposes and publications.

For the purpose of this dissertation, only the questionnaire responses that are most related to the topics addressed are reported in this chapter. For an overview of the complete analyses, see Bloder and colleagues (2021). More specifically, the following questionnaire items will be discussed:

1. Have you ever provided therapy to multilingual children with DLD?
2. Do you think that different approaches are needed in the diagnosis and treatment of multilingual children compared to monolingual children?
3. In the context of SLT for multilingual children with DLD, do you think that it is useful to compare a child’s language performance in their first and second language?

4. If yes/sometimes/rarely (3) ... (A) were your comparisons based on information provided by the parents or (B) were you able to directly observe child behaviour in both languages (possibly in the presence of parents)?
5. Are you aware of any testing or other diagnostic material that have been developed specifically for multilingual children with DLD?
6. Do you use special diagnostic material/tools for multilingual children with DLD?
7. Do you think that computerized tasks to assess the level of proficiency in the child's other language would be useful?
8. If such tasks (7) existed, would you use them when working with multilingual children with DLD?

3.5 Results

Only 7.0% of all respondents have no experience in working with multilingual children, 12.7% indicated to have some experience, while 80.3% declared to regularly provide SLT services for multilingual children ($X^2(3, N = 300) = 34.85, p < .001$, see figure 1). The majority of SLTs also think that different approaches for diagnosis and treatment for multilingual children are always (70.0%) or sometimes (28.7%) needed in the context of SLT service provision, while only 1.3% of the respondents are of the opinion that this is never useful ($X^2(3, N = 300) = 215.12, p < .001$, see figure 1).

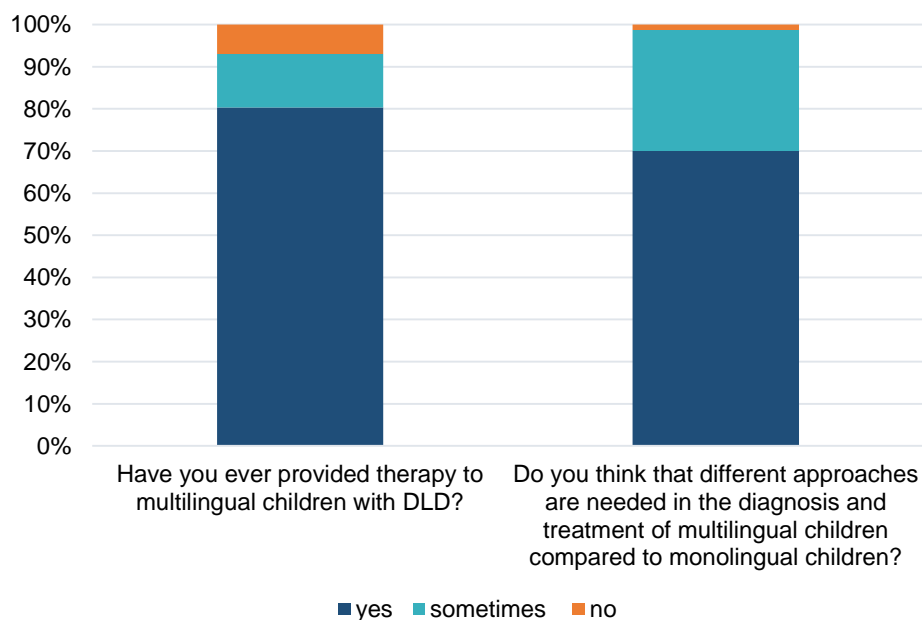


Figure 1: Distribution of SLTs' responses to the questionnaire items on experience and approaches in working with multilingual children.

More specifically relating to the context of language assessment of multilingual children, the majority ($X^2(3, N = 300) = 294.26, p < .001$) of the SLTs are of the opinion that it is in general (63.4%) or sometimes (29.1%) useful to compare the child's language performance in the different languages spoken. Only few SLTs consider it never (2.0%) or rarely (4.7%) useful (see figure 2). Of those who consider it useful to compare the performance in all languages spoken, the majority of respondents base (32.2%) or tend to base (35.5%) their language comparisons on parental reports, while others observe language behaviour usually (17.6%) or more often (14.5%) directly ($X^2(3, N = 290) = 38.41, p < .001$, see figure 2).

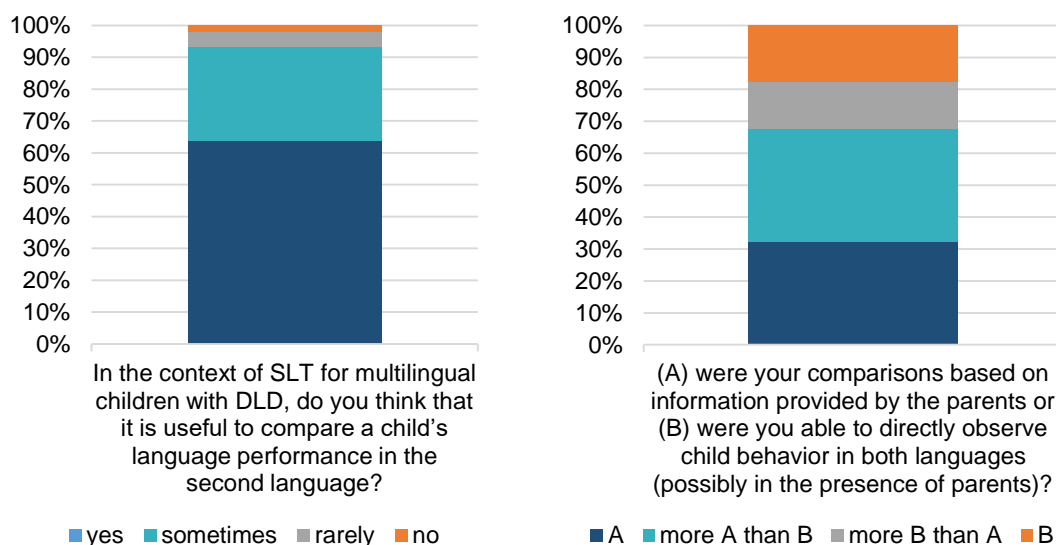


Figure 2: Distribution of SLTs' responses to the questionnaire items on the usefulness and modality of language comparison when working with multilingual children.

The majority (64.3%) of the SLTs also declared to be aware of diagnostic material that is specifically developed for multilingual children, but 35.7% are not aware of such material ($X^2(3, N = 300) = 24.65, p < .001$, see figure 3). Despite the awareness, only 18.1% of the SLTs regularly use such material when assessing multilingual children. While 22.4% of the respondents declare to sometimes use such material and 13.0% to use it rarely, the majority of SLTs (46.5%) does not use them at all ($X^2(3, N = 300) = 78.89, p < .001$, see figure 3).

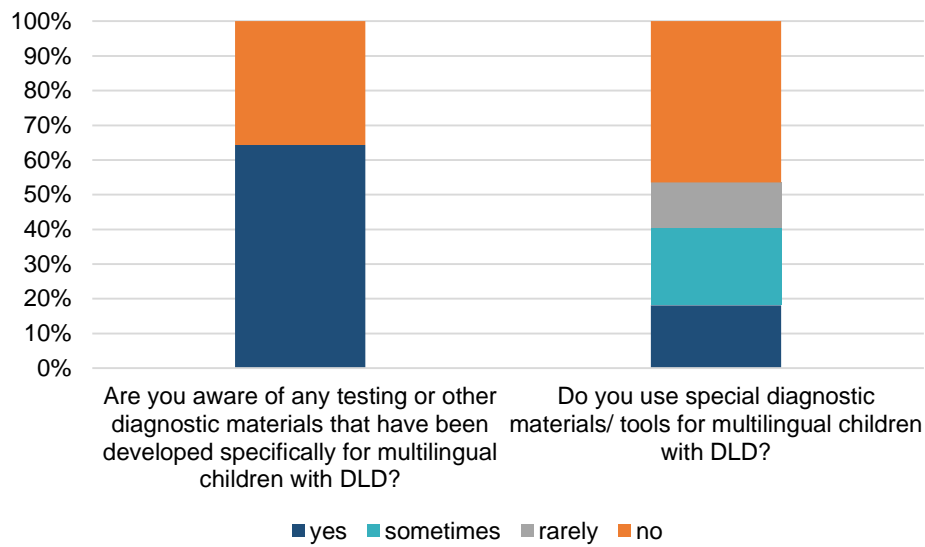


Figure 3: Distribution of SLTs' responses to the questionnaire items on diagnostic material for working with multilingual children.

The majority of SLTs also is of the opinion that it would be generally (56.7%) or sometimes (32.7%) useful to apply computerized tasks for the assessment of the child's heritage language ($X^2(3, N = 300) = 220.22, p < .001$). Only 5.3% each are of the opinion that this is rarely or never useful (see figure 4). The distribution of responses was very similar when asked whether SLTs would also administer these computerized tasks with multilingual children (see figure 4, 54.3% "yes", 36.0% "sometimes", 4.7% "rarely", 5.0% "no", $X^2(3, N = 300) = 215.39, p < .001$).

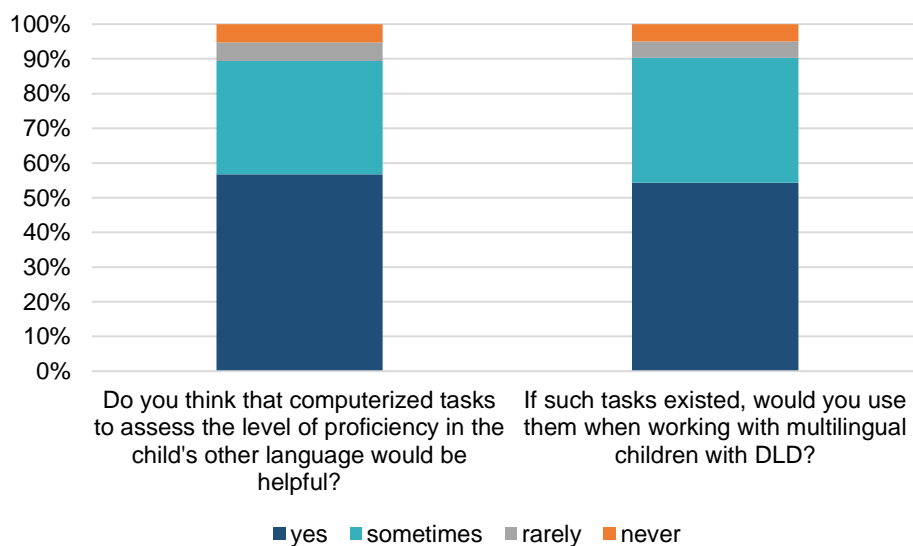


Figure 4: Distribution of SLTs' responses to the questionnaire items on computerized tasks for the language assessment of multilingual children.

3.6 Discussion

The responses by the SLTs support previous findings indicating that multilingualism is a very common phenomenon in SLT service provision (Marini, 2019; Scharff Rethfeldt, 2019). It furthermore underlines previous notions concerning the necessity of approaches different from the one applied for monolingual children and also the assessment of all languages spoken by the child (Armon-Lotem, 2012; Garraffa et al., 2019). Direct comparison of the children's language performance in both languages are hardly drawn even though the majority thinks that language comparisons are useful. However, despite the awareness of the need of material that is specifically designed for these children, less than 20% of the SLTs apply those. These findings suggest a gap between the SLTs' knowledge and the common practice applied. Further investigation is needed to identify the reasons behind this mismatch.

Generally, SLTs seem to be open to the application of innovative technology and think that computerized tasks are useful for the assessment of the child's heritage language. Only 10% of the respondents are of the opinion that computerized screening tasks are not useful in the context of language assessment of bilingual children and one tenth of the respondents would not apply them. Their openness towards computerized screening tasks for the automatic assessment of the heritage language reinforces the idea of computerized assessment for bilingual children as investigated in chapters 4 and 5.

4 Piloting computerized Chinese and Italian DD screening tasks

As mentioned in chapter 2.3.2, many people from China live in Tuscany. It is considered relevant to assess both language and orthographic systems acquired by the children (see chapter 2.4.). Similar to previous pilot studies in their final theses by Bigagli and Lorusso (2014) as well as Draghi (2015, unpublished), in this study, Mandarin and Italian screening tasks were implemented as experimental paradigms on "E-Prime 3.0" (<https://pstnet.com/products/e-prime/>). This allows for automatic administration of the tasks as well as evaluation concerning the accuracy and measurement of response time of the children's responses. This project is based on the following research questions: (1) Are standardized and computerized screening task results assumed to assess the same skills in the same language significantly associated with each other? (2) Are standardized and computerized screening task results assumed to assess related skills in different languages significantly associated with each other?

4.1 Methods & material

To answer these research questions, the following methods and material were applied in this study.

4.1.1 Participants

A total of $N = 33$ successive bilingual children which speak Mandarin at home, attend Italian public primary schools (grades 3, 4 and 5) and live in Prato (Tuscany, Italy) were tested. In addition to that, some of the children were orally exposed to (but not fluent in) Wenzhounese. For $n = 9$ of these children, teachers had indicated that they have or are at risk of DD. All children received formal Mandarin reading and writing instruction in their free time.

4.1.2 Material

Children were administered the standardized Italian reading test Batteria per la Valutazione della Dislessia e della Disortografia Evolutiva 2 (DDE-2, Sartori et al., 2007). Besides that, computerized screening tasks in Italian and Mandarin were administered assessing rapid automatized digit naming, phonological awareness, morphosyntactic processing and reading subtests. For all tasks, accuracy and response time were measured automatically. All audio clips were pre-recorded by a native speaker with natural voice and accent.

4.1.2.1 Screening tasks

Since all study participants attend a public Italian primary school and thus, their school reading and writing acquisition occurs predominantly in Italian, this study focused on Italian screening tasks.

Self-paced syllable reading. Participants are asked to read aloud and as fast as possible a series of syllables consisting each of a consonant followed by a vowel. Syllables are presented one by one on the screen. The next syllable is presented upon the child pressing the space bar. This self-paced reading time (time elapsed between the presentations of two subsequent syllables) is automatically recorded. The syllables presented are specific to the Italian orthographic system and each consist of one consonant followed by a vowel, e.g. “ve”. The task consists of 30 training and a total of 3 screening items. Self-paced reading time is automatically measured and stored, due to ceiling effects in pilot studies, accuracy is not tracked. Find more examples in appendix B.

Self-paced sentence reading. Participants are asked to read aloud and as fast as possible a list of five Italian sentences (consisting of high-frequent words, e. g. “La mamma porta un regalo al bambino” [The mum brings a present to the child.]), increasing in syntactic complexity and sentence length and presented one by one on the PC screen. Like the self-paced

syllable reading task, the next sentence is presented upon the child pressing the space bar. Again, this self-paced reading time is automatically recorded. The task consists of 1 training and a total of 5 screening items. Self-paced reading time is automatically measured and stored, due to ceiling effects in pilot studies, accuracy is not tracked. Find more examples in appendix B.

Word identification. A pre-recorded Italian word is played to the child participant. The orthographic form of this word is displayed on the screen along with two more words acting as distractors that differ in spelling and pronunciation. Across items, one of the distractors consists in a visual distractor. For each distractor, one grapheme was substituted with respect to the target in either case, for example the Italian word “colto” [educated] was presented auditorily and displayed on the screen along with the distractors “corto” [short] (phonological distractor) and “cotto” [cooked] (visual distractor). The task consists of 2 training and a total of 8 screening items. Response time and accuracy is automatically measured and stored. Find more examples in appendix B.

Nonword identification. The NW identification task underlies the same principle as the word identification task. Again, the distractor construction underlies the principle applied in the Italian word identification task, for example, the children listened to the NW “penko”, and they had to select the correct orthographic form among the following: “penco” (target), “benco” (phonological distractor) and “pencio” (which according to the Italian orthographic rules would be pronounced as “pentfo”, orthographic distractor). The task consists of 2 training and a total of 8 screening items. Response time and accuracy is automatically measured and stored. Find more examples in appendix B.

Phonological blending judgement. Two pre-recorded audio clips are played to the child participant. The first one consists of a series of phonemes that when blended would make an Italian word, presented at a one-second rate (e.g. “a-p-r-e” [(s)he opens] while the second audio presented is either the same or a slightly different string of phonemes, but an Italian word spoken normally e.g. “arpe” [harps]. The child is asked to judge whether the blending of the phonemes that were presented auditorily in the first audio correspond exactly to the word that was presented afterwards (in this example, no correspondence) by clicking on the corresponding buttons ✓ for correct and × for incorrect phonological blending. In 50% of the items, the audio clips did not correspond due to phoneme inversion. Items were presented in random order. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Syllabic inversion judgement. Again, two pre-recorded audio clips are played to the child participant, the first one being an Italian word (e.g. “dado” [dice, cube]) while the second audio presented contained the correctly (i.e. “do-da”) or incorrectly (i.e. “don-da”) inverted syllables of the same word, all of them being syllables in agreement with Italian-specific phonotactic rules. The child is asked to judge whether the inversion of the syllables that were presented auditorily in the first audio correspond exactly to the word that was presented afterwards by clicking on the corresponding buttons ✓ for correct and × for incorrect syllabic inversion. In 50% of the items, the audio clips did not correspond due to phoneme inversion. Items were presented in random order. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Subject-verb agreement. A pre-recorded sentence is presented auditorily that either do or do not contain violations in subject-verb agreement in number, for example “Le galline grasse mangia* sul prato” [the fat hens eats* on the lawn] is incorrect because there is no agreement between the subject “Le galline grasse” (plural) and the verb inflection “mangia” (3rd person singular), the latter ought to be “mangiano” (3rd person plural). The sentences are presented in random order with 50% correct and 50% incorrect subject-verb agreement. The child is asked to indicate whether the presented sentence is correct or not by selecting the corresponding buttons ✓ for correct and × for incorrect sentences. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Clitic pronoun judgement screening. A pre-recorded question in Italian is presented auditorily consisting each of a verb, a subject, and an object attached to the inflected version of the preposition “a”. Those are followed by answers in which the subject pronoun is dropped and the object nouns are replaced with clitic object pronouns. The latter may contain violations in agreement with the gender and case of the corresponding object mentioned in the preceding question. E.g. in “Che cosa fa il bambino alla bambina? La* dà i fiori” [What is the boy doing to the girl? He is giving her* flowers], the clitic pronoun “la” is incorrect, because it should be the dative-feminine clitic “le” instead of the accusative-feminine “la”. The stimuli are presented in random order with 50% correct and 50% incorrect clitic pronoun use, but they are each accompanied by visual support consisting in a line-drawing depicting the scene described in the auditorily presented sentence (see figure 5). Also here, the child is asked to indicate whether the presented sentence is correct or not by selecting the corresponding buttons ✓ for correct and × for incorrect sentences. The task consists of 1 training and a total of

12 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

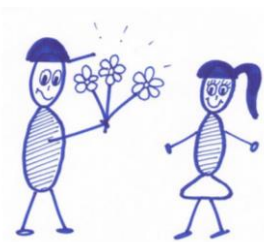


Figure 5: Line drawing used in the clitic pronoun judgement task as visual stimulation on the screen.

Radical position judgement. Mandarin characters are displayed on the screen one by one. Some of these characters have been manipulated regarding their position of the radicals (visual elements that are placed on the left part of the character, see figure 6 in which that radical mistakenly is placed on the right side). For each one of the characters presented, the child is asked to judge the correctness of the position of the radical by clicking on the corresponding buttons ✓ for correct and × for incorrect radical position of the character (see Chung et al., 2010). The task consists of 2 training and a total of 18 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.



Figure 6: Item presented in the radical position judgement task to be judged for correctness of radical placement.

Left-right inversion judgement. Similar to the radical position judgement, Mandarin characters are displayed on the screen one by one. Here, some of the characters have been manipulated regarding their orientation, more precisely have been (partly) mirrored (see figure 7 and Chung & Ho, 2010). For each one of the characters presented, the child is asked to judge the correctness of the orientation of the character by clicking on the corresponding buttons ✓ for correct and × for incorrect orientation of the character. The task consists of 2 training and a total of 18 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.



Figure 7: Item presented in the left-right inversion judgement task to be judged for correctness of radical orientation.

4.1.3 Procedure & data analysis

Data was collected in the schools in a separate and quiet room across two testing sessions with at least two hours of break in between. In the first session, the standardized reading test DDE-2 was administered. In the second session, screening tasks implemented on E-Prime 3 were administered using a Lenovo YOGA 720-15IKB laptop PC under the Windows 10 Pro operating system. Due to the Covid-19 pandemic, data collection was interrupted resulting in missing data from the teacher and caregiver questionnaires (thus not included in this study), an unbalanced sample of children with and without DD risk (see chapter 4.1.1), small sample size and missing data for the DDE-2 for $n = 5$ children who were excluded from the analyses when comparing DDE-2's and screening task results.

Correlational analyses (Spearman's rho) were performed on the scores obtained in the experimental tasks and in the standardized tests. Furthermore, Mann-Whitney U tests were run investigating group differences (DD risk vs. no risk) in the scores obtained in the screening tasks.

4.2 Results

Screening task results were compared to the teachers' risk indication and to the performance in the DDE-2.

4.2.1 Comparison of screening results and the teachers' risk indication

Mann-Whitney U tests were run to compare screening task performance of children without ($n = 24$) and with risk of DD ($n = 9$) as indicated by their teachers. Self-paced reading time for sentences was significantly different in the group of children with vs. without DD (risk) as indicated by the teacher ($U = 52.00, p = .024$). A difference in the DD (risk) vs. no DD (risk) group was also found for accuracy in the word identification ($U = 55.50, p = .003$) and accuracy in the clitic pronoun judgement ($U = 55.00, p = .028$) screening tasks. A marginally significant effect was found for accuracy in the phoneme blending judgement screening task ($U = 62.00, p = .056$).

4.2.2 Comparison of screening results and the DDE-2

None of the Mandarin screening tasks was shown to be significantly associated with raw scores obtained in the DDE-2 ($p_s > .05$). Italian screening tasks instead were shown to be correlated with standardized test performance. In particular, self-paced reading time of sentences, but not syllables were significantly associated with word reading time measured in the DDE-2 ($n = 28$, $\rho = .664$, $p < .001$). Self-paced sentence reading time was significantly associated with the amount of word ($n = 28$, $\rho = .473$, $p = .011$) and NW reading errors ($n = 28$, $\rho = .378$, $p = .047$). Self-paced sentence reading time was also significantly correlated with the NW reading time in the DDE-2 ($n = 28$, $\rho = .416$, $p = .028$). The percentage of correctly identified words in the Italian word identification task was significantly associated with word reading errors in the DDE-2 word reading subtest ($n = 28$, $\rho = -.460$, $p = .014$). Word identification accuracy in the screening was significantly associated with DDE-2 word reading time ($n = 28$, $\rho = -.388$, $p = .041$) and NW reading errors ($n = 28$, $\rho = -.462$, $p = .013$). Response time in the word identification was not significantly associated with these measures ($p_s > .05$). No significant correlations emerged comparing performance in the NW identification screening task and screening tasks on phonological awareness with the performance in the DDE-2 ($p_s > .05$). Significant associations instead emerged for the accuracy and response time in both tasks assessing the children's performance in morphosyntactic processing. In particular, response time in the subject-verb agreement judgement task was significantly associated with word reading errors ($n = 28$, $\rho = -.466$, $p = .012$), word reading time ($n = 28$, $\rho = -.482$, $p = .009$) and NW reading errors ($n = 28$, $\rho = -.423$, $p = .025$) in the DDE-2. The percentage of sentences correctly judged in the same task was significantly correlated with word reading errors in the DDE-2 ($n = 28$, $\rho = -.418$, $p = .027$). Both response time and accuracy in the clitic pronoun judgement task were significantly associated with word reading errors in the DDE-2 (response time: $n = 28$, $\rho = -.411$, $p = .030$; accuracy; $n = 28$, $\rho = -.410$, $p = .030$). For an overview of these results, see table 1.

Table 1: Overview of results for comparisons according to screening performance, teachers' risk evaluation (9 with, 24 without risk) and correlations of task performance in both languages with the DDE-2 raw scores.

Screening task performance	teachers' risk evaluation	DDE-2 (raw scores)
syllables (IT) self-paced reading time	n.s.	n.s.
sentences (IT) self-paced reading time	$U = 52.00, p = .024$	<i>word reading time:</i> $n = 28, \rho = .664, p < .001$
word identification (IT) response time	n.s.	n.s.
word identification (IT) accuracy (%)	$U = 55.50, p = .003$	<i>word reading errors:</i> $n = 28, \rho = -.460, p = .014$
nonword identification (IT) response time	n.s.	n.s.
nonword identification (IT) accuracy (%)	n.s.	n.s.
phoneme blending (IT) response time	n.s.	n.s.
phoneme blending (IT) accuracy (%)	$U = 62.00, p = .056$	n.s.
syllabic inversion (IT) response time	n.s.	n.s.
syllabic inversion (IT) accuracy (%)	n.s.	n.s.
subject-verb agreement (IT) response time	n.s.	<i>word reading errors:</i> $n = 28, \rho = -.466, p = .012$
subject-verb agreement (IT) accuracy (%)	n.s.	<i>word reading errors:</i> $n = 28, \rho = -.418, p = .027$
clitic pronouns (IT) response time	n.s.	<i>word reading errors:</i> $n = 28, \rho = -.411, p = .030$
clitic pronouns (IT) accuracy (%)	$U = 55.00, p = .028$	<i>word reading errors:</i> $n = 28, \rho = -.410, p = .030$
left-right inversion (MAND) response time	n.s.	n.s.
left-right inversion (MAND) accuracy (%)	n.s.	n.s.
radical position (MAND) response time	n.s.	n.s.
radical position (MAND) accuracy (%)	n.s.	n.s.

4.3 Discussion

This study investigated whether standardized and computerized screening task results assumed to assess the same skills within and across languages are significantly associated with each other. For Italian, a series of significant associations emerged between screening task performance and standardized test results as well as with the teachers' risk evaluation.

So, the research question on whether standardized and computerized screening task results assumed to assess the same skills in the same language are significantly associated with each other can be answered positively. This finding underlines the general appropriateness of computerized tasks in the assessment of children (Haridas et al., 2017; Rauschenberger et al., 2019).

The research question “Are standardized and computerized screening task results assumed to assess the related skills in different languages significantly associated with each other?” instead was not supported. Potential explanations are discussed here: On the one hand, this indicates that information concerning only one language is not informative with respect to the whole performance profile. Furthermore, the two reading and writing systems are very different and thus reflect different processes and mechanisms and also the language-specific screening tasks reflect different processes. They might thus not be of exactly similar difficulty as perceived by the participants. However, in children who do have DD, reading performance is expected to be similar in all language and orthographic systems acquired. It may also be that in general, the children’s performance in Mandarin was too low to show that they were at risk for DD.

5 MuLiMi screening studies

In the survey study, SLTs have reported that they find the application of computerized screenings useful (see chapter 3). Furthermore, a lack of suitable material for the assessment of bilingual children was identified. The idea to develop a computerized screening platform was strengthened by the SLTs’ indication that computerized screenings are considered useful, and the associations found between computerized screening task performance and risk evaluation as well as standardized test performance, a screening platform was developed for automatic administration and evaluation of tasks in various languages. The aim of the project is to provide a screening tool which allows clinicians and teachers to automatically assess language and reading performances of multilingual children in the languages they speak to appropriately identify their risk of DLD and DD. This screening tool consists of computerized tasks to automatically assess their language and/or reading along with related skills and provides automatic scoring of the children’s responses for teachers and clinicians.

Screening batteries for two different age groups with two different clinical foci were developed: DLD screenings were developed for children attending kindergarten aged between 4 and 6 years of age. DD screenings were developed for children attending 2nd to 4th grade of primary school, aged between 7 and 10 years. For the development and preliminary validation of the screening tool, specific language combinations were chosen representing

different linguistic features. DLD screenings were developed for a) Spanish-Italian-speaking children (see chapter 5.3), representing a combination of languages that are both Romance and, thus, similar regarding morphosyntax and vocabulary, and b) Italian-German-speaking children (see chapter 5.4), representing two alphabetic languages that are more distinct from each other and derive from different families of languages (German as Germanic language). While the Spanish-Italian DLD screening is designed for children living in Italy and speaking Italian as their majority language (L2), the Italian-German screening is designed for Italian-speaking children living and schooled in Germany. The same is true for the Italian-German DD screening (see chapter 5.5). Besides, English- and Mandarin-Italian DD screenings are constructed for children living and schooled in Italy (see chapter 5.6). Covering Italian, German, English and Mandarin, a range of orthographic depths from transparent/shallow (Italian) over deep (English) to logographic (Mandarin) is represented in the screening studies.

5.1 Goals

To assess the validity of the newly developed screening platform and the informative value of the screening task results, the following general research question (RQ) was posed: Do computerized screenings that automatically assess children's language and/or reading performance in all their spoken languages contribute to the identification of DLD and DD risk?

In accordance with the general research question of this project, for each study specific hypotheses emerge and will be described individually in the respective chapters.

For this, the general research question is further subdivided into more detailed research questions. These are then used to derive specific hypotheses for the individual studies in the following subchapters and which will be veri- or falsified based on the data collected.

RQ1: The results of standardized/traditional tests and screening tasks declared to measure the same skills are associated with each other (concurrent validity).

RQ2: Children with a higher risk score (based on information from SLT, caregiver and teacher questionnaires and standardized/traditional tests) can be distinguished from children with no/lower risk scores through their screening tasks performance, considering all languages spoken (discriminant validity).

These research questions lead to the following hypotheses.

1. Performance in screening tasks is associated with the risk levels determined by performance on the standardized tests (above or below clinical cut-offs) and with clinical status (Hypothesis 1).

2. Performance in standardized tests is associated with the screening tests assessing the same ability or closely related abilities (Hypothesis 2).
3. Performance on different screening tasks assessing similar skills is associated with each other (Hypothesis 3).
4. Performance on various screening tasks assessing the same linguistic area in the two different languages is associated. (Hypothesis 4).
5. The children who have been identified with language or reading difficulties show an impairment in both L1 and L2 on screening tasks assessing the same linguistic area (Hypothesis 5).

To address general research questions in the context of this project, several studies were carried out: chapter 5.3 presents a study on Spanish-Italian-speaking children attending kindergartens in Italy, chapter 5.4 presents a study on Italian-German-speaking children attending kindergartens in Germany, chapter 5.5. presents a study on Italian-German-speaking children attending primary schools in German and chapter 5.6 presents a study with Italian-Mandarin, Italian-English and monolingual Italian-speaking children attending primary schools in Italy.

5.2 Experimental design

All the MuLiMi screening studies carried out aimed at the preliminary validation of the novel web-based screening platform designed for automatic administration and evaluation of language and/or reading skills of bilingual children in all their languages spoken. Depending on the language exposure patterns of the single target groups, language and/or reading assessment using conventional methods like caregiver and teacher questionnaires and standardized tests were compared to accuracy and response time in the computerized screening tasks.

DLD screenings were constructed, implemented on the newly developed screening platform MuLiMi and preliminary validated with $n = 36$ Spanish-speaking children living in Italy (remotely) and $n = 37$ Italian-speaking children living in Germany (in presence). See table 2 for the distribution of participants per language-pair across clinical status and study characteristics.

Table 2: Clinical status of the participants of the DLD screening studies.

languages	DLD diagnosis	DLD risk	TD	testing mode	follow-ups
Spanish-Italian	$n = 16$	$n = 11$	$n = 9$	remotely	$n = 5$
Italian-German	$n = 7$	$n = 17$	$n = 13$	in presence	$n = 14$

DD screenings were constructed, implemented on the newly developed screening platform MuLiMi and preliminarily tested with $n = 11$ monolingual Italian children, $n = 12$ English-speaking and $n = 7$ Mandarin-speaking children living in Italy and tested remotely. Another $n = 26$ Italian-speaking children living in Germany were tested in presence. See table 3 for the distribution of participants per language-pair across clinical status and study characteristics.

Table 3: Clinical status of the participants of the DD screening studies.

languages	DD suspect	DD risk	TD	testing mode
Italian-German	$n = 3$	$n = 2$	$n = 21$	in presence
Mandarin-Italian	n.a. ²	n.a	$n = 7$	remotely
English-Italian	$n = 1$	$n = 1$	$n = 10$	
monolingual Italian	n.a.	n.a	$n = 11$	

5.2.1 Development of screening tasks

The most reliable clinical markers for the identification of DLD/DD-risk in the target languages were identified through in-depth literature research. For those markers, computerized tasks were designed and implemented.

5.2.1.1 Development of computerized DLD screening tasks

To identify the risk of DLD in bilingual children three linguistic areas to be assessed were selected:

- Phonology, i.e. the ability to process single sounds (phonemes)
- Morphosyntax, i.e. the ability to process grammatical features
- Lexicon, i.e. the vocabulary (words (both languages) recognized by the child).

The latter was tested using comparable tasks for both language pairs focusing on (verb) comprehension skills using picture matching tasks from the CLTs (Haman et al., 2017). To assess phonological skills in Spanish- and German-Italian-speaking children, NWRT were used (cf. Boerma et al., 2015); i.e. oral repetition of an auditory stimulus (NW). For morpho-syntactic processing, for each language, language-specific clinical markers were selected based on previous literature.

In order to assess the ability to process grammatical features, language-specific clinical markers have been selected. One focus is on verb morphology since subject-verb agreement is considered a clinical marker in Italian (Bortolini et al., 2006, 3rd. person plural), German (Rothweiler et al., 2017) and Spanish (Bedore & Leonard, 2001). In addition to that, for

² n.a. stands for not applicable

Spanish the use of finite vs. non-finite (i.e. inflected) verbs (Grinstead et al., 2014), for Italian the correct use of direct clitic object pronouns (Bortolini et al., 2006; Dispaldro et al., 2013) and for German case marking (Scherger, 2015) are taken into consideration as useful clinical markers.

According to the selected clinical markers, tasks assessing the comprehension skills and/or the abilities to process these grammatical phenomena (requirement for automatic administration and evaluation) were adapted from previous research projects where possible and newly developed when needed. These include matching as well as judgement tasks, and beyond that, a task combining grammaticality judgement and sentence comprehension (amount of correct and incorrect sentences counterbalanced). Furthermore, “Who says it right?” grammaticality judgement tasks were implemented. In the latter, a grammatically incorrect sentence and the same sentence in its grammatically correct form are presented auditorily. The child responds to the question “Who says it right?” (in the respective language) by pressing a corresponding key (see 5.3.2.2 for a detailed description of this task and examples).

5.2.1.2 Development of computerized DD screening tasks

To identify the risk of DD in bilingual children reading skills, for the alphabetic target languages, tasks assessing self-paced syllable and sentence speed as well as word and NW identification speed and accuracy were constructed following the example of Bigagli and Lorusso (2014). For Mandarin as logographic script instead, tasks assessing the ability to identify Chinese characters were implemented (S. Hu et al., 2018). Furthermore, for all languages, phonological awareness was assessed through the judgement of linguistic units manipulated on a phoneme level for example the judgement of phoneme blending and syllabic inversion or onset detection. Besides that, RAN of digits in the children’s L1 was chosen (Meyer et al., 1998, for detailed descriptions of the tasks, see 5.5.1.2 and 5.6.2.2).

In addition to that, also here, grammaticality judgement and matching tasks were used to assess the children’s abilities in morphosyntactic processing. When possible, those tasks were designed according to the specific features of each language in comparable ways.

5.2.2 Description of the web app

The construction of the web app was based on the following rationale: Based on the specific goal concerning the creation of language-specific, bilingual screenings, the screening platform was expected to be easily extendable for that assessment of various language combinations. Furthermore, this being an ongoing research project, it was considered important that language-combinations, tasks and items could be modified and added to the screenings.

It was thus decided to create a web app that could be easily accessed by potential users. Furthermore, it was designed to be a modifiable screening platform, allowing one group of target users (administrators) to add language combinations, items, tasks and screenings. Due to the specific requirement of testing remotely during the Covid19-pandemic's restrictions concerning access to schools, kindergarten and clinics, the web app was further developed to enable remote assessment (see screening studies in chapters 5.3 and 5.6). Remote screening administration was enabled through the connection between the two devices (examiner and the examinee), established exploiting WebRTC to instantiate a peer-to-peer communication, allowing for real-time screen sharing.

The functions of MuLiMi were implemented on a three-tier system in which the various functions of presenting, processing and managing/storing data are operated individually. The software architecture style Representational state transfer (RESTful) is characterized by its potential to sequentially add data and functionalities component by component to the software without recreating or modifying the architecture. The software architecture follows the Model-view-controller (MVC) pattern, in which data management, visualization and user interactions are handled individually.

Figure 8 visualizes the three different target users of the MuLiMi web app and how it allows for the creation and the administration of screenings as well as for the visualization and storage of the screening results and examinee characteristics.

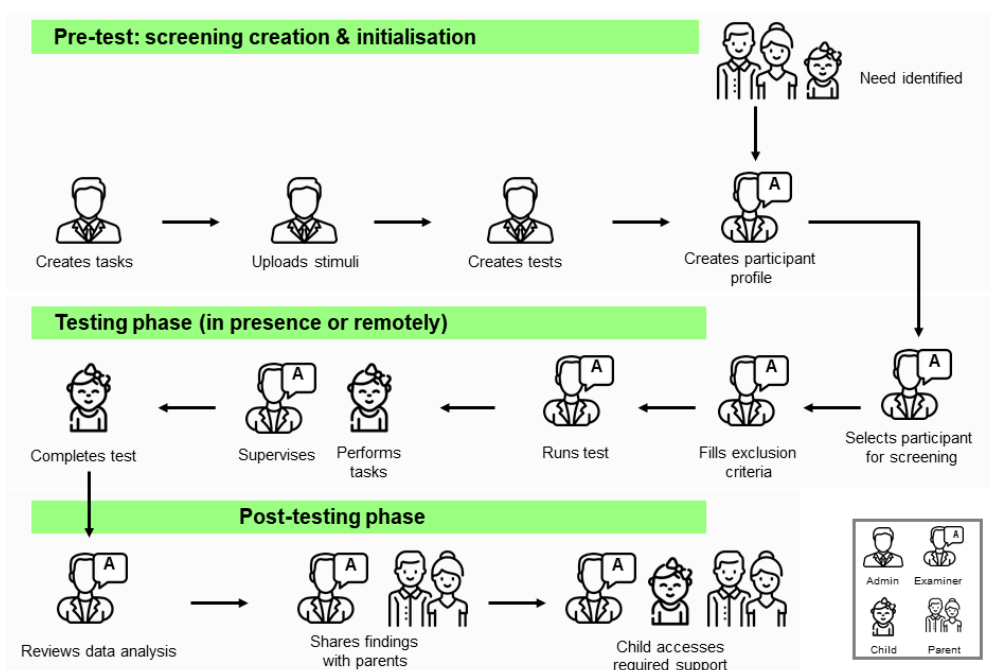


Figure 8: Groups of target users, user flow and functions of the MuLiMi screening platform.

To facilitate easy access to the screening platform for future users, a web-app was developed. This allows users to access the screening platform via web-browser from any device that is connected to the internet. A native app, in contrast, needs to be downloaded and installed on the users' device first.

The web app is used in different ways by three groups of users: *administrators* (*admins*, responsible for the screening creation and configuration), *examiners* (carrying out the screenings implemented by the administrator) and the *examinees* (who are administered the screenings by the respective examiner, see figure 9).

Admins can upload single audio, video, picture GIF-files and Boolean values to the MuLiMi screening platform in the “*content*” section (see figure 9). Additionally, text-based content can be loaded directly to the platform.

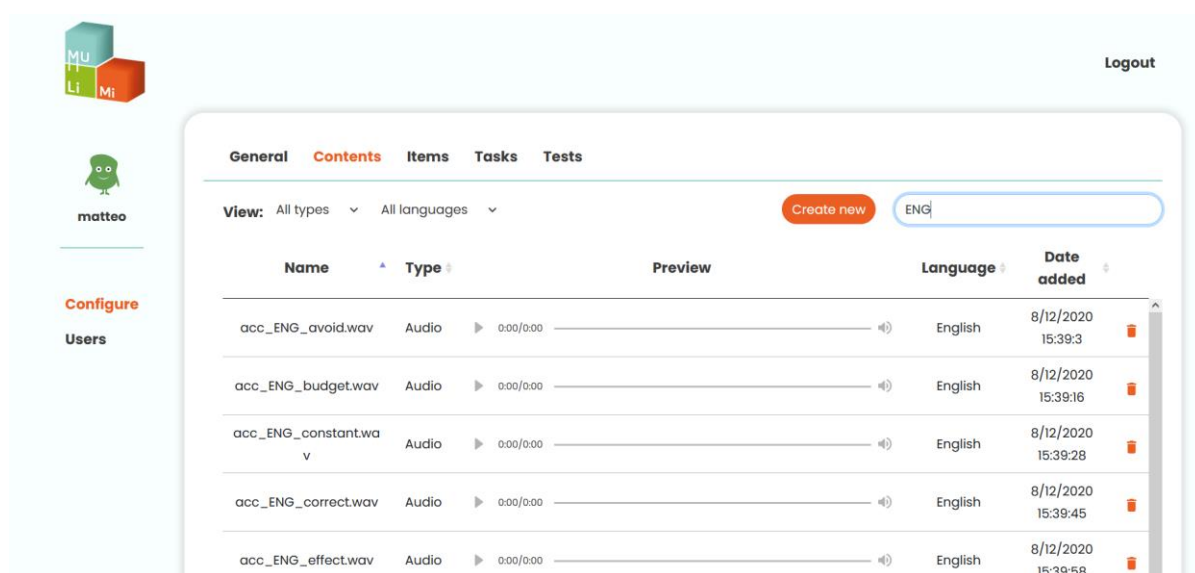


Figure 9: Admin interface for content upload, item and screening compilation.

From those contents, *screening tasks instructions* as well as *items* can be created by merging the contents required for a certain task or item. Already uploaded content as well as defined items can be sorted by date, type, language and retrieved through the search function. To facilitate and unify the instruction and item creation process, the patterns of the most common testing paradigms were selected and implemented. The patterns implemented on MuLiMi allow for the creation of NWRs, Dynamic Reading Assessment, Dynamic Novel Word Learning, self-paced reading and naming (RAN) tasks, matching tasks and judgement tasks. This paradigm-specific stencil for items is filled by the admin with the content previously uploaded. Items are specific to one task type and can be used to compose tasks of that type only, as the task type changes the semantic meaning conveyed by the contents inside the

item (and consequently the logic used to render them). The admin defines the order of presentation of the audio-visual contents that the item is made up of as well as the target response, which the admin defines as the expected response to be given by the examinee under assessment. In this way, a set of items that have been created in a comparable manner and that are testing the same linguistic phenomenon can be merged into a *task*, preceded by an instruction page (text, audio or video instruction, see figure 10).

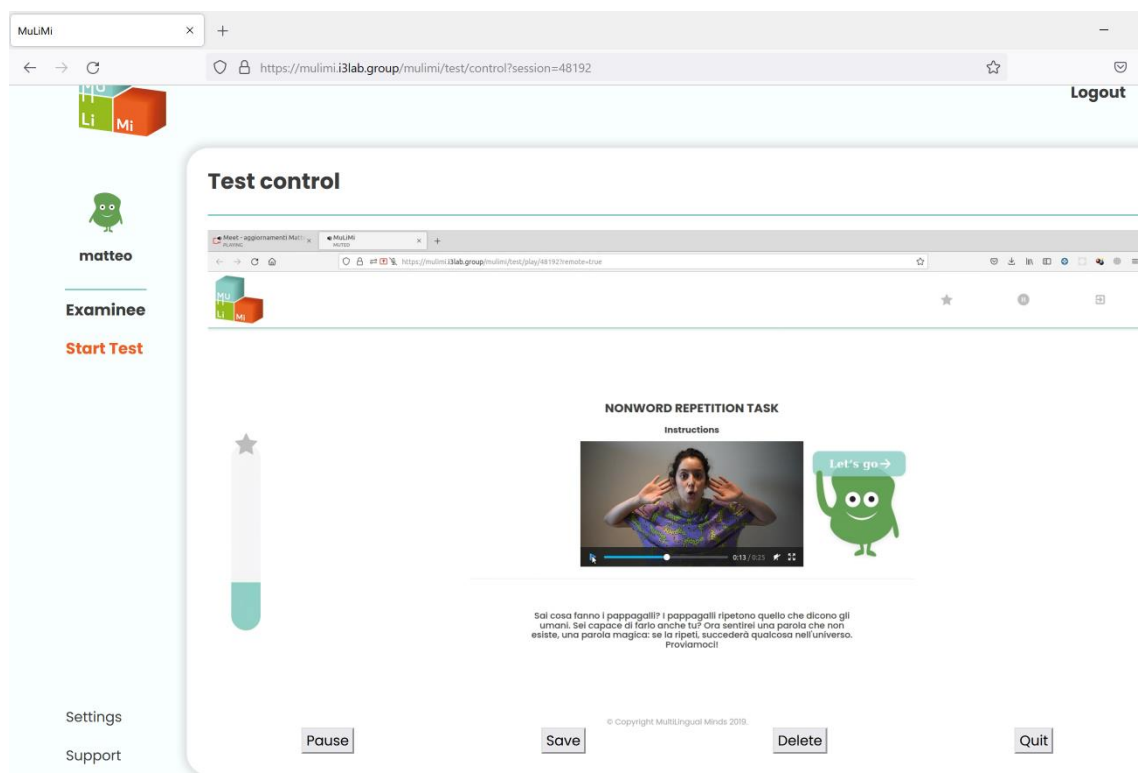


Figure 10: Video-based instruction in the child’s L1 with text-translation for the examiner in the child’s L2.

Those tasks in turn can then be merged into a so-called *test* which is ready to be used by examiners with their examinees for screening purposes (see figure 9 for the various clickable content, item, task and test buttons enabling access to the respective interface for upload, managing and modifying contents, items, tasks and tests). Admins configure the test to a specific language (pair of L1 and L2) and age group (an interval of ages), that are connected to the inclusion criteria the examiners insert for each examinee (see figure 11).

Examiners – as opposed to admins – are not entitled to create or make changes to the screenings. They can create, modify and manage examinee profiles, where both information on the examinee and their screening results can be safely stored. When starting the screening, based on the characteristics “languages” and “age” of the examinee, a pre-selection of available screenings is presented in a drop-down list that can be chosen from. When the inclusion criteria of the screening match the examinee’s characteristics, the screening can

be initiated. Optionally, the examiner can choose to initiate a remote testing session by checking the “Is this a remote test?” check box (see figure 11)

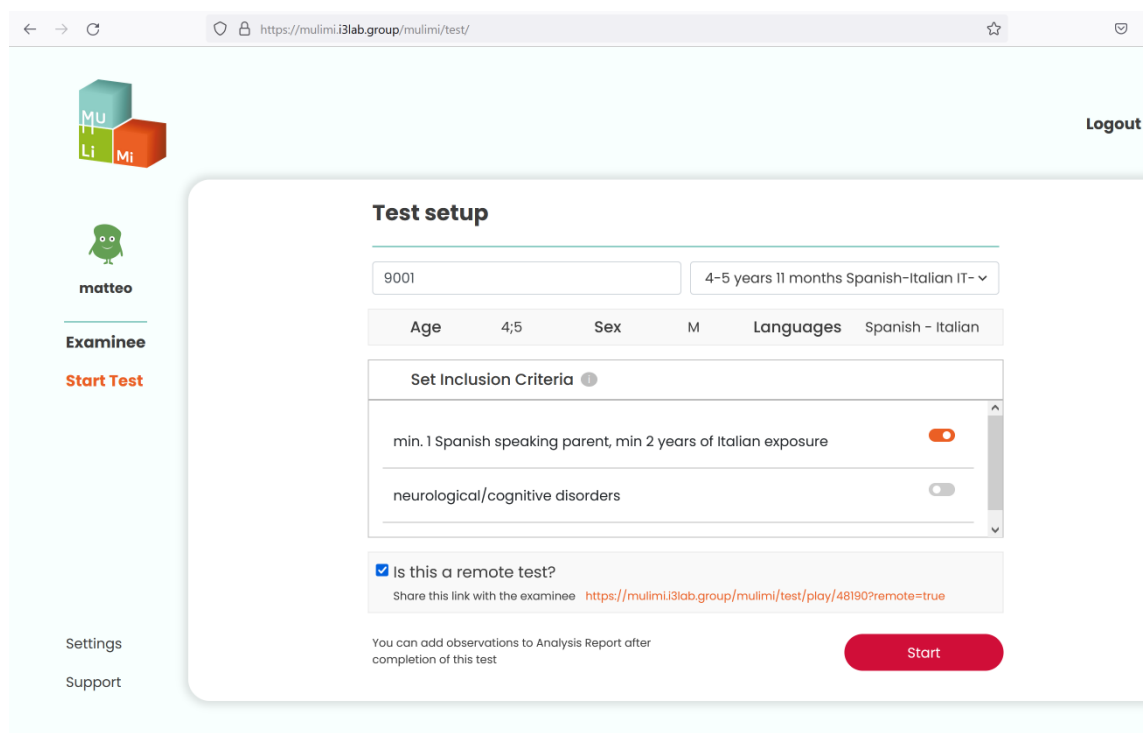


Figure 11: Interface for test setup to initiate a remote screening session.

During the screening session the examiner guides the examinee in screening execution, i.e. provides support when technical issues arise, makes sure that the child has understood the task instructions and how to respond to the stimuli presented automatically in the screening. Furthermore, during a session, the examiner can click on a “save”-button to ensure the examinee’s responses are safely stored. They can, moreover, pause the screening and re-start it in case of interruptions (see “pause”, “save”, “delete” and “quit” buttons on the bottom of figure 10).

After a screening has been administered, the single responses and response times per item of an examinee can be viewed and downloaded from a data file that can be exported as .ods, .csv or .xlsx files (see figure 12). Furthermore, the examinee’s overall performance regarding accuracy (number of correct and incorrect answers) and response time (total and mean response time in ms) of the responses within one task across all presented items can be viewed in the examinee profile. In the examinee profile, performance of all the tests that have been carried out using this screening platform are stored. To view the results of one particular test, the respective test needs to be selected from the list on the left. Next to it, on

the right, a second list will appear representing all the tasks that were included in the respective test (see figure 12).

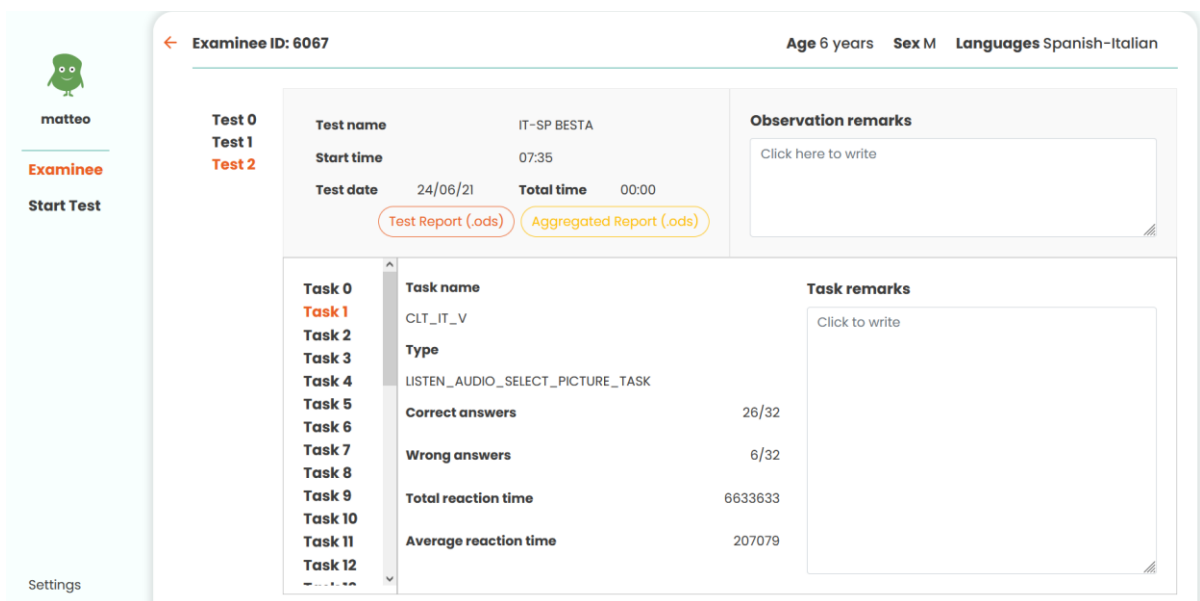


Figure 12: Interface for screening result examination in the examiner section.

Other than administrators and examiners, **examinees** do not sign up on the website. When tested in presence of the examiner, the child looks at and interacts with the examiner’s device and the examiner initiates the session. When the child is tested remotely, he/she receives the link and is expected to be seated in front of a desktop or laptop PC, reacting to the presented stimuli directly from their device. When using the MuLiMi screenings remotely, clicking on the landing page of the screening triggers a pop-up notification asking the examinee for permission to share their screen (see figure 13). Upon permission, the examinee’s screening interface is displayed on the examiner’s screening interface who is also able to initiate and pause the screening and able to save the results.

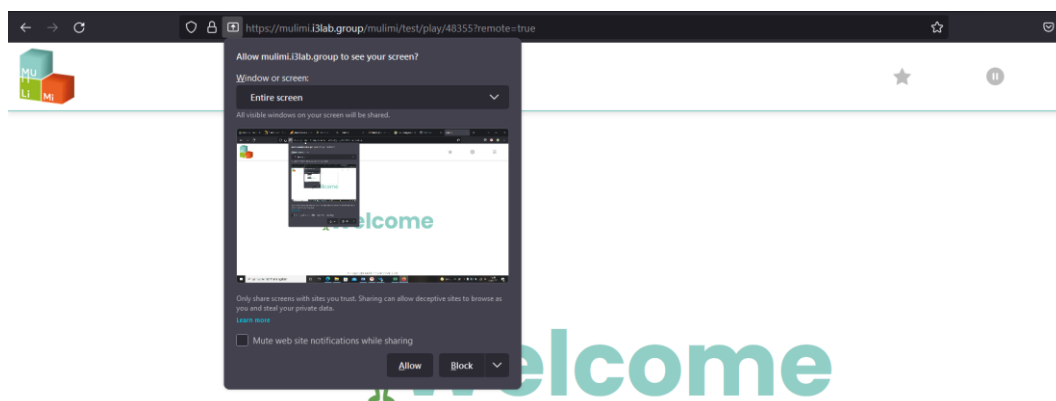


Figure 13: Examiner interface to enable screen sharing during a remote testing session

Overall, the interface was designed considering the examinees' young age, ranging from 4 to 10 years. Colourful animations in between the screening tasks (see figure 10 for the progress bar on the left side and the green character on the left of the screen reappearing in between screening tasks) were incorporated to improve enjoyability and to avoid boredom as the screening proceeds.

5.2.3 Recruitment of participants

All participants were recruited in SLT clinics, schools or kindergartens. Children of two different age groups (kindergarten children between the ages of 3;11 and 6;1 and primary school children in grades two through four between the ages of 7 and 10 years), with different clinical risk statuses ("risk of" or "diagnosed with" DLD or DD or TD) and different language backgrounds (see above) participated in the study. All children were exposed to their L1 through at least one of the caregivers regularly in the family context and to the L2 for at least two years through kindergarten and/or school. While children recruited in Germany and monolingual children recruited in Italy spoke Italian as their family language (L1), the group of Spanish-, English-, and Mandarin-speaking children spoke Italian as societal language (L2). All the German- and English-speaking children who participated in the DD-screening studies attended bilingual primary schools. Most of the German children who participated in the DLD-screening study attended bilingual kindergartens. Children with developmental disorders beyond language and/or reading impairment were not included in this study.

Before inclusion of the child participants in the study, potential participants and their caregivers were provided with information on the study and its purposes. After having given exhaustive information, oral consent to participation by the child and consent forms signed by the participant's caregivers were collected and safely stored.

5.2.4 Validation

The following constructs were investigated in the screening studies: validity, predictivity and usability. Discriminant validity is assessed comparing diagnosed, at risk and TD children. Concurrent validity is assessed through tests declared to be measuring the same construct. Content validity is assessed through the comparison of L1 and L2 screening task performance. Follow-up studies give insights about the persistence of the difficulties and predictive value as assessed by comparing screening performance with later achievements in standardized tests. Whether or not the screening platform is suitable for its application in clinical practice is investigated in usability studies.

The preliminary validation of the screening platform and the single screenings was carried out through comparison of the children's performance in the screening tasks between

risk level, as well as correlations with related direct (i.e. standardized tests) and indirect measures (i.e. teacher and caregiver questionnaires). A detailed description of the screening tasks and standardized tests used in the screening studies can be found in the respective subchapters 5.3.2 and onwards. Teacher and caregiver questionnaires, that were part of all the studies, are described in the following chapters 5.2.4.1 and 5.2.4.2.

5.2.4.1 *Teacher & SLT questionnaire*

For each language group, a questionnaire for teachers or SLTs on the participant's reading and/or language skills was collected (see appendix A). This questionnaire requests judgment of receptive and productive skills in relevant linguistic domains (phonology, vocabulary, morphosyntax and pragmatics). The respondents evaluate the child's receptive and productive skills separately for the various linguistic domains on a 5-point Likert scale ranging from "not problematic" to "very problematic". These responses were converted into numerical values between 0 and 4 for data processing. From the responses to the single questionnaire items, the compound scores *total* (sum of all responses), *total receptive* (sum of all responses regarding the receptive skills in the linguistic domains) and *total productive* (sum of all responses regarding the productive skills in the linguistic domains) are derived. For the teacher questionnaire, the compound score *literacy* consists of the teachers' responses regarding the children's abilities in reading aloud, reading comprehension and writing.

5.2.4.2 *Caregiver questionnaires*

All the participants' caregivers were asked to fill in either online or pen-and-paper questionnaires including questions on demographic data, the socio-anamnestic situation, caregivers' occupation and the child's language background. Further, they were asked to answer questions on the specific disorder targeted in the applied screening and both routine screening procedures on the IRCCS Medea Institute online platform. Caregivers could choose between the Italian version of these questionnaires and the version in their respective L1.

QUIR-DC (Questionario per l'Identificazione del Rischio di Disturbo della Comunicazione – Questionnaire for the Identification of Risk for Communication Disorders, Lorusso & Dolzadelli, 2016). Based on the Italian online version of the QUIR-DC (designed for clinical use on the IRCCS Medea Institute online platform), a German and a Spanish translation and adaptations as a pen-and-paper questionnaires were prepared to be filled in by children's caregivers. The questionnaire consists of 96 questions on anamnestic data and information on the child's language background. Caregivers' responses are merged into scores that are used for data analyses. Precisely, the general score (GS) expresses the level of development, whereas the risk score (RS) expresses probable risk of a developmental delay or disorder

(including a phonological score (PS) for the pronunciation difficulties). Besides, the family language input global score (FIGS) and family language input risk score (FIRS) express the quality of language input in the child's L2. Caregivers could choose between the Italian and German/Spanish versions of the questionnaire. All collected data was entered and scored automatically to obtain the compound scores introduced above using the "Formfacade" web application (<https://formfacade.com/>).

DSA Questionnaire (Lorusso & Milani, unpublished, DSA stands for the Italian term "Disturbi Specifici di Apprendimento" – Specific Learning Disorders). Caregivers of children who participated in the DD screening studies were asked to fill in the DSA questionnaire containing questions on the child's general development and school achievements. The questionnaire is structured into four categories with questions on school discomfort (1), general learning difficulties (2), in-depth analyses of learning difficulties for reading and writing acquisition (3) and maths (4). The responses of the caregivers are scored in the following manner: Whenever a caregiver responds positively ("yes") to the negatively connoted question (indicating the risk for DD), 1 point is assigned. From these answers, compound scores for the four different categories are created by summing up all the points assigned in each category. All data was either collected with – or, in pen-and-paper questionnaire case, transferred to – "Google Forms" (<https://docs.google.com/forms/>) for further analyses.

5.2.5 General testing procedure

Each child was tested within two to three testing sessions (45 to 60 minutes each), with at least one session for the standardized and at least one session for the screening tasks. The time foreseen for the final version of the screenings is about 40 minutes, but this longer time was needed to collect data on validity (standardized tests) and on a large number of items from which the best items will then be selected through a process of item analysis. For the administration of the screening tasks, children were first of all familiarized with the tasks. They received exhaustive standardized explanations on how the tasks work and were briefly trained on the use of the computerized interface. After that, the novel computerized screening tool was conducted. Each of the tests included a small number of practice trials. Short breaks were provided after each task. Caregiver and teacher questionnaires were collected outside these screening sessions and handed back to the researcher. For each of the studies, the screening tasks and the standardized tests that were selected will be described in the chapters below.

5.2.6 Data analysis

Data from standardized tests was manually scored by trained native speakers with professional background in the field of SLT or psychology. Most of the data collected with the help of computerized tasks was automatically coded and stored. Databases were constructed using codes corresponding to each participant to ensure anonymity in the data storage. Correlational analyses were performed on the scores obtained on the new screening tasks and standardized tests using non-parametric (Spearman's correlation) and parametric (Pearson's correlation) tests, depending on the sample's and variable's characteristics. For the comparison of screening task performance and dichotomous variables in small samples, point-biserial correlations were run and compared to Mann-Whitney U test results. Further correlations were computed between the experimental tasks and variables obtained from the SLT, caregiver and teacher questionnaires.

5.3 Bilingual, computerized DLD screening for Spanish-speaking children living in Italy

As mentioned in chapter 2.3.2, in Lombardy, Spanish is among the most common languages of the population with a migration background. Accordingly, an Italian-Spanish DLD screening was constructed and administered to children with, at risk of and without DLD remotely.

5.3.1 Hypotheses

For every study, hypotheses depend on the respective language-specific markers chosen. In addition to that, in this study, also a task based on the principles of Fast Mapping and Dynamic Assessment was included. For this task, an additional hypothesis is formulated below:

Do children with diagnosis of or at risk of DLD a) show more difficulty in correctly recognizing and associating the novel words (NWs) introduced in the dynamic novel word learning task (DNWL, Hypothesis 1a) and b) have more difficulties in producing them? (Hypothesis 1b)

Moreover, based on the assumption that a child at risk of or with DLD shows deficits in various tasks assessing the same linguistic area, Hypothesis 3 (see chapter 5.1) was further subdivided for the purpose of this study, given the specific tasks chosen. It was hypothesized that:

- a) the tasks of subject-verb agreement and finiteness are associated with each other in the respective languages (Hypothesis 3a).

Furthermore, based on the assumption that children with DLD show difficulties in both languages spoken, also Hypothesis 4 was subdivided:

- a) the Spanish test of NW repetition is associated with the performance in the Italian test of NW repetition (Hypothesis 4a).
- b) performance in Spanish verb comprehension is associated with Italian verb comprehension (Hypothesis 4b).
- c) performance in the Italian task of subject-verb agreement is associated with performance in the Spanish task of subject-verb agreement (Hypothesis 4c).
- d) performance in the Italian test of finiteness is associated with performance in the Spanish test of finiteness (Hypothesis 4d).
- e) performance in the Italian DNWL subtest is associated with performance in the Spanish DNWL subtests (Hypothesis 4e).

In addition to the hypotheses described in chapter 5.1, it was hypothesized that performance on the screening tests at first time of testing (t1) would show a general trend of associations with performance at second time of testing (t2) (follow-up, Hypothesis 6) and more precisely that performance levels at t1 are comparable to performance at t2. Due to the small sample size, this comparison was only assessed upon visual inspection of the data and more specifically, comparison of performance levels displayed in the bar graphs (Hypothesis 6a).

5.3.2 Material & methods

To answer these research questions, the following methods and material were applied in this study.

5.3.2.1 *Participants*

Thirty-six early-sequential or simultaneous bilingual Spanish-Italian-speaking children aged 4 to 6 (mean age in months: $M = 64.50$, $SD = 7.87$) participated in this study. Sixteen of these children had already been diagnosed with DLD by an SLT and received treatment in specialized rehabilitation centres (public services) in Italy. While $n = 9$ of these children did neither hold a DLD diagnoses nor scored below cut-off in the standardized tests, $n = 11$ children had not been diagnosed with DLD, but scored below cut-off according to the standardized test manuals. All children lived in Italy, attended kindergarten and had been exposed to the Italian language there for at least two years. At least one of the child's caregivers is a native speaker of Spanish. Native speakers were not further classified as being speakers of a certain variety of Spanish. Recruitment took place in kindergartens and clinical centres providing SLT services. Ten to eleven months after the first time of testing (t1), $n = 5$ children with an existing

DLD diagnoses and under SLT treatment were tested again using the same standardized and screening tasks. In addition to that, the DNWL and the “Who says it right?”-judgement task on use of infinitive versus inflected verbs were administered.

5.3.2.2 Screening tasks

Since the Spanish-Italian online screening is designed with the aim of assessing multilingual children's language skills in both their family and societal language, it offers the same screening tasks in both Spanish and Italian.

Nonword Repetition Task (NWRT). An instruction video by a Spanish native speaker is presented. Training items are not provided. In the instruction, the context of parrots' abilities to imitate and repeat human language is given and the child is told that upon repetition of “magic words” (i.e. NWs) something in the outer space scenery displayed on the screen will change. NWs are then presented auditorily one by one. The child is asked to repeat after each word. All NWs were previously recorded by native speakers of the respective language. They were constructed such that they either comply with the phonotactic constraints of a target language (language-specific, LS NWs) or include only phonemes that are present in both languages (non-language-specific, NLS NWs). LS items could, thus, be considered as more complex than the NLS items (see also Dos Santos & Ferré, 2016). Language-specific trajectories of phoneme acquisition were taken into account when constructing the NWs. LS items were recorded with language-specific prosodic features of the respective language while NLS items were recorded with flat, neutral prosody (without word stress on a certain specific syllable, see Chiat, 2015; Mottier, 1951). The selection of LS and NLS NWs was based on a two-step rating procedure (Bloder, Eikerling & Lorusso, submitted): In a first step, monolingual adult native speakers of the respective language repeated NWs and rated them for L1-alikeness and pronounceability on a scale from 1 to 5 (1: not L1-alike/not easy to pronounce, 5: very L1-alike/very easy to pronounce). Based on these scores, a subset of NWs was selected and rated again for pronounceability and language-specificity by a new group of adult native speakers using an online questionnaire implemented on “qualtrics” (<https://www.qualtrics.com/>). Following up this selection, further NW properties were assessed, namely the inter-rater-reliability between a native Spanish and a native Italian speaker as well as intra-scale reliability. For inter-rater-reliability, Cronbach's Alpha was $\alpha \geq .602$ for all NWs. Internal consistency as expressed by Cronbach's Alpha was $\alpha = 0.747$. The list of NWs used in the Spanish-Italian screening consists in a total of 10 NWs. Precisely, it consists in $n = 4$ Italian LS non-words (LS IT, e.g. ['blandjeza]), $n = 2$ NLS NWs spoken by a

native Italian speaker (NLS IT, e.g. [fulsami]) and $n = 4$ Spanish LS NWs (LS SP, e.g. [ajukom'jon]), see table 4. Syllable length varies from two to four syllables and is fairly distributed among the four categories.

Table 4: Overview of NWs selected for the Spanish-Italian NWRT.

category	NWs selected
LS IT	[ˈmudjo], [foˈda:na], [blanˈdjeza], [maŋkeˈtale]
NLS IT	[fulsami], [melinak]
LS SP	[ˈxano], [ˈnwelo], [ˈlaxo], [ajukomˈjon]

All $n = 10$ LS IT, NLS IT and LS SP NWs were automatically presented using the MuLiMi screening platform. They were presented in random order to prevent the children from accustoming themselves to the phonological and prosodic features of a certain language before switching to repetition of NWs of a different category. In order to maintain an adequate level of motivation and attention of the child, a coloured line drawing is presented on the screen depicting a space scenery during the NWRT (see figure 14).

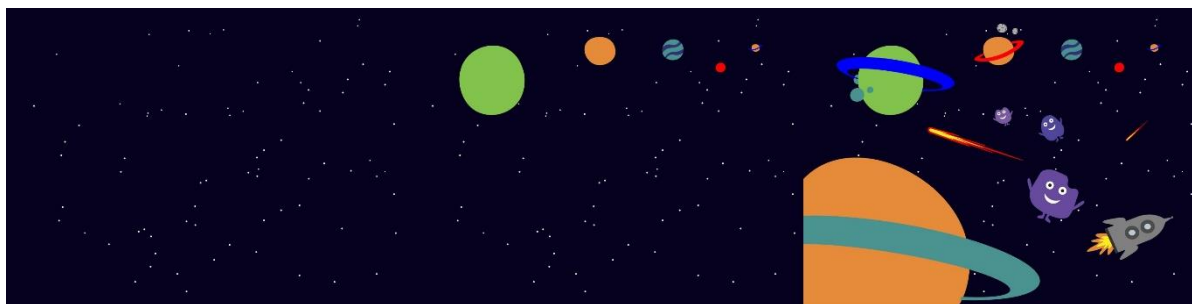


Figure 14: Examples from coloured line drawings presented on the screen during the NWRT.

Each repetition is followed by a minimal change of the image displayed, e.g. the appearance of a planet, disappearance of litter and apparition of aliens in space. Based on an audio recording of the child's NW repetition, the examiner manually assigns 1 point for correct repetition of the target and 0 points, when incorrectly repeated. In order for the scores to resemble the children's specific language acquisition conditions acquiring both Italian and Spanish, the scores used for further analyses were based on the mean score of the Spanish- and the Italian-speaker.

Cross-linguistic Lexical Tasks (CLTs). Verb comprehension subtests in Italian and Spanish from the CLTs (Haman et al., 2017) were used. Prior to each verb comprehension subtest, an audio explaining the task execution is presented in Italian or Spanish by a respective native speaker, depending on the language version. Like in the original CLTs, training

items are not provided. Following-up the auditory presentation of a pre-recorded question which embeds the target verb in gerund for both Italian (e.g. “Chi sta giocando a golf?” [Who is playing golf?], see figure 15) as well as for Spanish (e.g. “¿Quién esta pescando?” [Who is fishing?]), children are asked to identify the target picture among four options and select it via mouse click or touch screen. The developers of the CLT subtests agreed to implement the subtests on the MuLiMi screening platform that allows for automatic administration and scoring (matching accuracy) of the children’s responses. In the case of Italian, the audio files for instructions and items are the same ones that were used in the CLT app (Zinn, unpublished); for Spanish, instead, the instruction and items were recorded by a native speaker of Spanish as indicated in the Spanish version of the CLTs (Cantù Sanchez et al., unpublished). Each of the subtests in both languages contain 32 items (64 in total). Accuracy of the responses are automatically measured and stored.

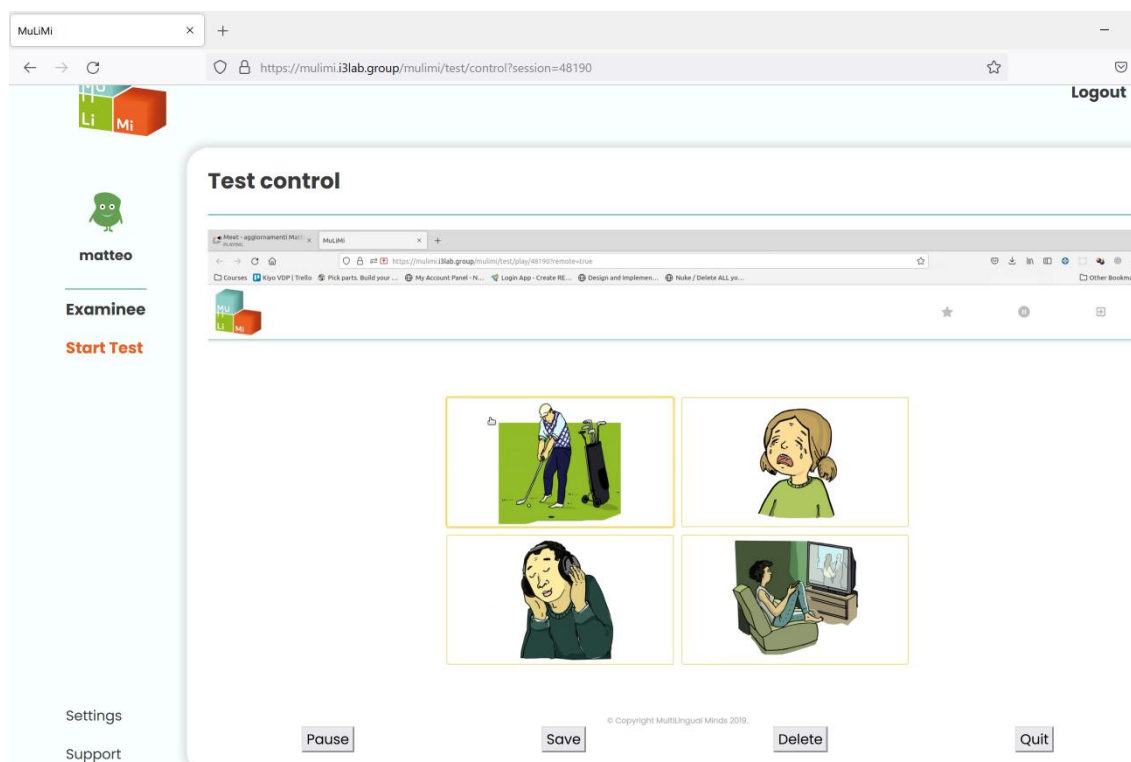


Figure 15: Examiner interface during remote administration of the Italian CLT verb comprehension subtest.

Who says it right? (WSIR). The pre-recorded question for grammaticality judgement “Who says it right?” is presented auditorily in the target language. Then, one correct and one incorrect sentence are subsequently presented auditorily in random order accompanied by two kinds of visual support: a) two different figures (GIF-files) that seem to be saying one of the sentences each and b) a coloured line drawing depicting the scene described (see figure 16). The child is asked to indicate which of the two sentences presented is correct by selecting

its corresponding “speaker”. Two types of WSIR tasks are used in both languages resulting in a total of four WSIR tasks in the Spanish-Italian screening. For both languages, one of the tasks targets incorrect sentences manipulated for subject-verb agreement and the second one targets sentences manipulated for non-inflected infinitive verbs. Find more examples in appendix B.

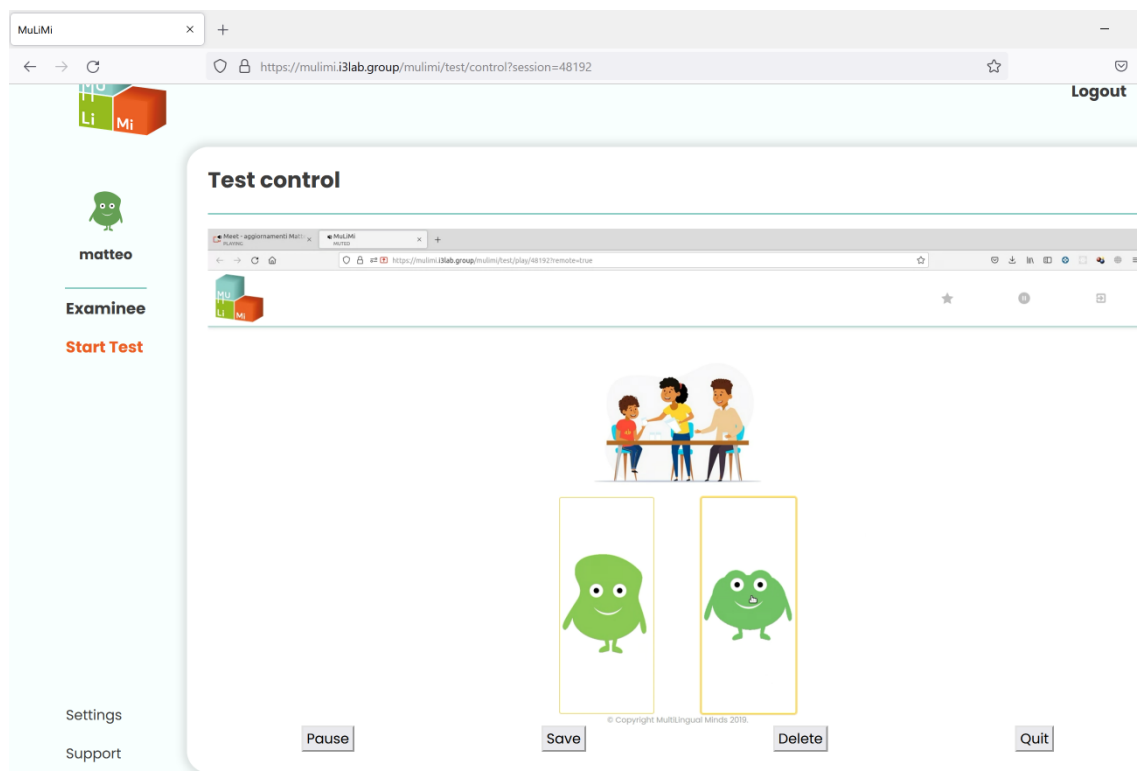


Figure 16: Examiner interface during remote administration of the WSIR subtest.

WSIR – Subject-verb agreement. For each of the language versions of this task, an instruction video with a native speaker explaining the task is presented prior to task completion and followed by two example items. Upon presentation of the pre-recorded question in Italian “Chi lo dice giusto?” one correct (e.g. “lo metto la sciarpa.”, [I put on the scarf.]) and one incorrect sentence (e.g. “lo mette* la sciarpa.”, [I puts* on the scarf.]) are presented. The latter is incorrect due to the verb that is inflected for use with a third person singular subject while the subject of this sentence is in first person singular. In the Spanish version of this task, the question “¿Quién lo dice bien?” [Who says it right?] is followed by one correct (e.g. “Él duerme mucho.”, [He sleeps a lot.]) and one incorrect sentence (e.g. “Él duermen mucho.”, [He sleep_{PL}* a lot.]). Both versions of this task consist of 2 training and a total of 16 screening items for each of the two languages. For 50% of the items, there is correspondence between

the amount of syllable length between target and distractor while for 50% there is no correspondence. Accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

WSIR – Finiteness. Since the children have been familiarized with the principle of the task in the WSIR-subject-verb agreement tasks, no instruction video is presented and neither are training items provided. Here, upon presentation of the pre-recorded WSIR-question, one correct (e.g. “Lei piange sempre.”, [She always cries]) and one incorrect sentence (e.g. “Lei piangere* sempre.”) are presented. The latter is incorrect due to the verb that is not inflected (infinitive) for use with a third person singular subject according to the subject of this sentence. Also in the Spanish version of this task, the question “¿Quién lo dice bien?” [Who says it right?] is followed by one correct (e.g. “Él come pollo.”, [He eats chicken.]) and one incorrect sentence (e.g. “Él comer* pollo.”, [He eat chicken.]). Both versions of this task consist of 2 training and a total of 8 screening items for each of the two languages. Accuracy of the given responses are automatically measured and stored. Due to time constraints, this task was not administered to all children. Find more examples in appendix B.

Dynamic Novel Word Learning (DNWL). This task investigates the ability to learn new words in a dynamic way, providing repetitions and corrective feedback, according to the child’s performance. To avoid biases depending on language exposure and (cultural) familiarity with certain words or semantic concepts, not real words but NWs were used. For both language versions, three NWs were associated with one character each (see figure 17). Those were constructed such that they did not resemble creatures or objects that children of that age are familiar with. For the Italian version of the task, the three NWs were “galpo” ([galpo]), “domio” ([domjo]) and “felio” ([fe:ljo]) while “mokal” ([mokal]), “flado” ([flado]) and “leñon” ([lejon]) were used for the Spanish version of this subtest. The NWs have been rated as specific to either Italian or Spanish in online ratings by adult native speakers of the respective language (see rating procedure described in the section on NWRT). The test consists of four phases: presentation, consolidation, test and naming. The task instructions and stimuli (linguistic context in which the NWs are presented) have been recorded by native speakers.

DNWL – Presentation. First of all, the child is presented with the three figures and their “names” (the NWs) one by one (e.g. in Italian “Ehi ciao, vorrei presentarti dei miei amici. Lui è Galpo.– Ciao, io sono Galpo. – Clicca su Galpo per dirgli ciao.”, in Spanish “Oye hola, me gustaría que conocieras algunos amigos míos. El es Mokal. – Hola, soy Mokal. – Haz clic en Mokal para decir hola.”, in English “Hey there, I would like to introduce you to my friends. This is Galpo. – Hello, I am Galpo. – Click on Galpo to say hello.”). When the child selects

the wrong character, the software first gives him/her feedback about the incorrect choice, points out whom the child actually clicked on and then proposes the task again (e.g. in Italian “No, questo non è Galpo. Questo è Domio. Clicca su Galpo per dirgli ciao”, in Spanish “No, este no es Mokal. Este es Flado. Haz clic en Mokal para decir hola”, in English “No, this is not Galpo. This is Domio. Click on Galpo to say hello”). In this initial phase of presentation of the NWs to be learned, children are asked to react to the stimuli by clicking on the corresponding character not to test their comprehension and recognition performance, but to make them actively interact. In addition to that, most children selected the target in nearly 100% (ceiling effect), so that children’s responses are neither collected nor analysed.

DNWL – Consolidation. In the second phase, the examinee is asked to select each character one by one (e.g. in Italian: “Clicca su Galpo”, in Spanish: “Haz clic en Mokal”, in English: “Click on Galpo”). When the child selects the corresponding picture at their first attempt, corroborative feedback is provided and the next item is presented. Also here, when the child selects the wrong character instead, the software first gives him/her feedback (e.g. in Italian “No, questo non è Galpo. Questo è Domio”, in Spanish: “No, este no es Mokal. Este es Flado”, in English “No, this is not Galpo. This is Domio”), and then the child is asked the same question again until the corresponding figure has been correctly identified. This phase continues until the child recognises all three characters consecutively at a first attempt. The frequency of correct character identification along with the total number of attempts is automatically recorded. In a second step, task administrators manually calculate the ratio from the latter.

DNWL – Test. In the test phase, the NWs previously associated with the characters are used within semantic contexts. Coloured line drawings depict the three characters in different scenes where they need and use various objects (high-frequent nouns, choice was based on the CDI database (MB-CDIs, 2022) confirming high-frequent use of these words in the target languages in early childhood). This subtest consists of a minimum of two and maximum of three scenes, depending on the child’s performance. In each scene, all three characters are addressed once in random order, so this phase consists of $n = 9$ items. A pre-recorded audio is played to the child with a sentence describing the scene (“Now they go out.”), followed by a description of a condition of a certain character (“Felio notices the rain.”). Finally, a question concerning the mapping of the NW with one of the characters (“Whom do you give the umbrella to?”) is presented. The child is asked to indicate the character whose condition requires the object mentioned in the question. Other than for the presentation and consolidation phases, no feedback is provided. A subtle audio signal indicates that the answer

has been recorded, irrespective of correctness of the child’s response. However, the system acts differently depending on the response the child provides. Whenever in the first or second scene the examinee answers incorrectly for at least one of the questions proposed, he/she is sent back to the consolidation phase until the child recognises all three characters consecutively at a first attempt. If the examinee provides exclusively correct answers in the first two scenes, he/she is never sent back to the consolidation phase and the third environment is not presented. Again, the frequency of correct character identification as well as the amount of attempts is automatically recorded. In a second step, task administrators manually calculate the ratio depending on whether all three scenes (minimum 0 and maximum 9 correct mappings) or only the first two scenes (100% accuracy) were administered. Find more examples in appendix B.

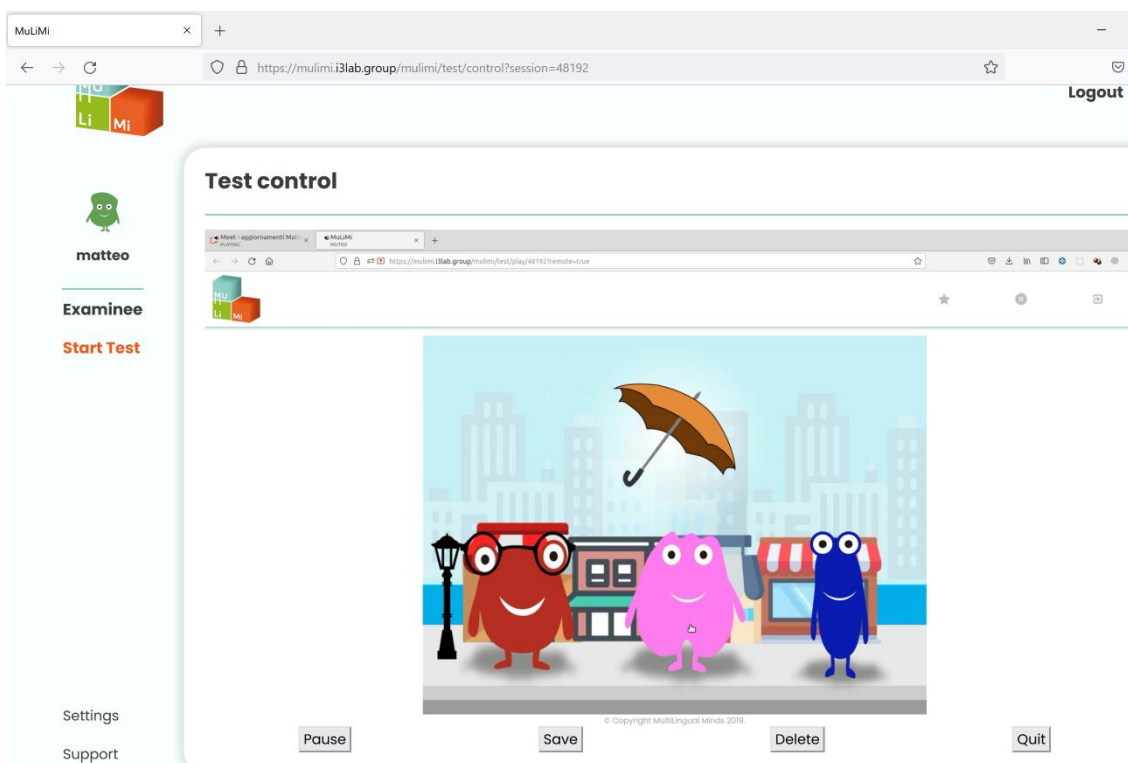


Figure 17: Examiner interface during remote administration of the DNWL testing phase subtest.

DNWL – Naming. Finally, the child is presented each character one by one on the screen and asked to name the three characters. This is embedded in the context of the characters needing to go home after the trips illustrated in the scenes in the test phase (in Italian: “Adesso devono tornare a casa, aiutami a chiamarli. Chiama... lui!”, in Spanish: "Ahora es hora de irse. Llama a mis amigos para que vengan con nosotros. ¡Llamalo!", in English "They have to go home now, help me call them. Call... him!"). When the child does not answer after 5 seconds or says that he/she does not know the answer, he/she was primed with the onset

phoneme (phonological onset clue), since screening administrators are fluent in Italian but not in Spanish, this clue was always given in Italian, i.e. “Ascolta, lui si chiama fla-...”, in English "Listen, his name is fla-..."). The children’s answers are recorded and then transcribed using the International Phonetic Alphabet (IPA). The transcripts are evaluated manually by the examiner: 1 point is assigned when the child’s response is precisely the target stimulus or when a minimum of three out of five phonemes – including all vowels – overlap with the target stimulus. 0.5 points are assigned whenever the child only upon the phonological onset clue provides an answer that is correct or contains a minimum of three correct phonemes and all syllables are correctly produced. 0 points are assigned whenever the child does not respond, or the response given varies from the target stimulus in more than two out of five phonemes.

5.3.2.3 Standardized tests

In order to assess the children’s language skills in their societal language, with the ultimate goal of validating the screening tasks, subtests from common standardized language tests for Italian-speaking preschool children were used.

Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni (BVL). Children were administered the following subtests of the BVL (Marini et al., 2015): NW repetition to assess phonological as well as sentence repetition, grammaticality judgement and sentence completion subtests to assess morphosyntactic skills. All tasks were administered and evaluated according to test manual and are described in more detail below.

Nonword repetition (BVL). The subtest consists of 18 NWs (3 NWs each consisting of 1, 3 and 4 syllables, 6 NWs are bisyllabic) that are presented auditorily to the child by the examiner. The child is asked to repeat them and 1 point is manually assigned by the examiner for each NW that was correctly repeated by the child.

Sentence repetition (BVL). The subtest consists of 20 sentences (of which 3 are complex sentences) that are presented auditorily one by one to child by the examiner. After each sentence the examiner presents, the child is asked to repeat it. 1 point is manually assigned by the examiner for each sentence that was correctly repeated by the child. The test is interrupted when the child repeats 5 sentences in a row incorrectly.

Grammaticality judgement (BVL). The subtest consists of 18 sentences of which 9 are grammatically correct and another 9 are not grammatically correct. The examiner auditorily presents the sentences and asks the child to judge after each sentence, whether it is correctly

formed or not. 1 point is manually assigned by the examiner for each sentence for which grammaticality was correctly judged.

Sentence completion (BVL). Upon a sentence as produced by the examiner, the child is asked to complete a second sentence. All the semantic information needed is provided in the first sentence. The examiner also defines the subject (with information on number) and tense of the second sentence that the child is asked to complete with an inflected verb. The subtest consists of 14 items and 1 point is manually assigned for each item for which the child's response matches the expected answer.

Test Fono-Lessicale: Valutazione delle abilità lessicali in età prescolare (TFL). To further assess the children's lexical skills in their societal language Italian, the "Comprensione lessicale" (vocabulary comprehension) of the (Vicari et al., 2007) was used. The test consists of 47 words (of which 2 are training items; 40 nouns, 7 verbs). The target word is presented auditorily in the context of a question (in Italian: "Qual è [target]?", "What is [target]?") by the examiner. The child is asked to indicate the corresponding picture (coloured line drawings) among the four pictures presented. For each of the items, one of the three distractors is semantically unrelated, one of them is semantically related and one of them is a phonological distractor that together with the target word makes a minimal pair (two words that differ in a single phoneme (Vicari et al., 2007)). 1 point is assigned each time a word is correctly identified.

CPM-Coloured Progressive Matrices (Belacchi et al., 2008). Nonverbal intelligence was tested by means of the *CPM-Coloured Progressive Matrices*, under the supervision of a trained psychologist if the scores of a standardized nonverbal intelligence tests were not retrievable from clinical records and not older than one year. It consists of 3 blocks of 12 pictures (coloured line drawings with geometrical forms, so-called "matrices", 36 items), each with a missing piece. Out of six options, the child chooses which piece needs to be added to complete the picture displayed. The performance is then compared to peers through t-scores.

These standardised test results were evaluated according to the criteria in the respective manuals and the norm data provided in the latter. Furthermore, the raw scores were converted into percentages to facilitate comparison with the results of the experimental tests.

5.3.2.4 Procedure

All children were tested individually in the SLT clinic or kindergarten where they were recruited, in the presence of their SLT, kindergarten teacher and/or a student researcher in a quiet room. In the first of two testing sessions, the trained student researchers or an SLT (all

native Italian speakers) administered the subtests of the Italian standardized tests BVL (Marini et al., 2015), TFL (Vicari et al., 2007) and the Raven's CPM (Belacchi et al., 2008) as described above according to the test manuals which lasted around 45 to 60 minutes. In the second session, they took on the role of a supporter and the examiner administering the screening tasks was connected via conference call from remote. The MuLiMi screening was administered through link and screen share (see chapter 5.2.2). For the simulation of a realistic testing scenario, the child was connected to the examiner using the laptops or desktop PCs available to the student researcher or in the clinic/kindergarten where they were tested. Due to technological constraints and comparability of screen size for the screening and standardized test visualisation, participation to the study from tablet or smartphone was not possible. The supporter assisted in establishing the connection to the screening's examiner, and recorded the child's responses to the NWRT and naming tasks in the DNWL Italian and Spanish version on a portable recording device. Additionally, they ensured that the child was feeling comfortable, gave neutral feedback to keep the child focused and motivated and also assisted when the task requirements were not clear to the child after the instruction and examples. When the device available did not allow for reacting to the stimuli presented via touchscreen and the child did not manage to use the mouse, he/she was instructed to indicate the button to click with the index finger and the supporter clicked on the button correspondingly. The heterogeneity in responding to the stimuli was considered acceptable, since not response time but accuracy is considered in this step of the Spanish-Italian screening validation. Also this session lasted around between 45 and 60 minutes. A break of minimum 60 minutes separated the first and the second testing session. Whenever necessary, the child took a short break. For $n = 5$ children with diagnoses of and under treatment for DLD, this procedure was repeated in summer 2021 (t2) which was around ten to twelve months after the first time of testing (t1).

Kindergarten teachers and SLTs working in the institution where the children were recruited filled in the pen-and-paper version of the teacher questionnaire in Italian. Caregivers chose between the online and pen-and-paper version of the QUIR-DC and a short language background questionnaire. Their choice depended on their own preference or that of the institution where the children were recruited and that managed correspondence with the caregivers. Caregiver and teacher questionnaires were completed exclusively at t1.

5.3.2.5 *Risk score creation*

Children who had already been diagnosed with and treated for DLD were assigned to the DLD group ($n = 16$). For the rest of the participants, children's scores in standardized tests

were analysed regarding whether or not they had scored below the cut-off indicated in the respective test manual, i.e. 2 standard deviations (SD) below norm. Children without diagnosis who scored below cut-off in at least one subtest were part of the at-risk group ($n = 11$). Children without diagnosis and with no results in the standardized subtests below the cut-off as indicated in the test manual were considered TD ($n = 9$). This three-level variable is referred to as “risk level”.

5.3.3 Results & discussion

None of the results of the screening tasks were significantly associated with the children’s test scores in the nonverbal intelligence tests and was thus not taken into account as a control variable in the analyses ($p_s > .05$). See table 5 for an overview of mean standardized test performance per group.

Table 5: Standardized test performance (M , SD) according to the children’s risk level assignment.

	TD ($n = 9$)	at risk ($n = 11$)	DLD ($n = 16$)
age (in months)	$M = 65.44$, $SD = 9.83$	$M = 60.64$, $SD = 7.45$	$M = 66.78$, $SD = 5.14$
TFL (%)	$M = 80.00$, $SD = 10.12$	$M = 72.12$, $SD = 12.18$	$M = 80.00$, $SD = 11.33$
CPM (z-scores)	$M = .826$, $SD = .848$	$M = .225$, $SD = .687$	$M = .130$, $SD = .824$
BVL grammaticality judgement (%)	$M = 66.67$, $SD = 19.25$	$M = 51.51$, $SD = 21.24$	$M = 66.67$, $SD = 16.20$
BVL sentence comprehension (%)	$M = 61.90$, $SD = 17.50$	$M = 56.96$, $SD = 21.57$	$M = 42.06$, $SD = 22.43$
BVL sentence repetition (%)	$M = 82.22$, $SD = 18.89$	$M = 35.91$, $SD = 17.58$	$M = 47.22$, $SD = 28.08$
BVL nonword repetition (%)	$M = 80.00$, $SD = 13.33$	$M = 51.51$, $SD = 21.24$	$M = 48.15$, $SD = 21.55$

5.3.3.1 Comparison of performance in screening tasks and risk level

The children’s risk level was significantly associated with their performance in the NWRT incorporating LS Spanish, LS Italian and NLS items ($n = 36$, $\rho = -.545$, $p > .001$). The risk level was also significantly correlated with both the Italian ($n = 36$, $\rho = -.524$, $p = .004$) and the Spanish version of the finiteness task ($n = 36$, $\rho = -.535$, $p = .003$). Furthermore, the children’s risk level was significantly associated with the Italian version of the WSIR subject-verb agreement task ($n = 36$, $\rho = -.432$, $p = .009$) and the Spanish version of the WSIR subject-verb agreement task (though not significantly, $n = 36$, $\rho = -.322$, $p = .056$). No significant correlations emerged comparing the CLT subtests and the subtests of the DNWL in both languages to the risk level. Figure 18 illustrates that for all tasks in both languages, the TD children – who did neither score below cut-off in one of the standardized tests nor hold an

official diagnosis of DLD – on average scored highest. Interestingly, visual inspection of figure 18 suggests that the children with an official diagnosis of DLD did not consistently score worse compared to the children without official diagnosis who scored below cut-off in at least one of the standardized tests. Note that the children in the DLD group all received treatment.

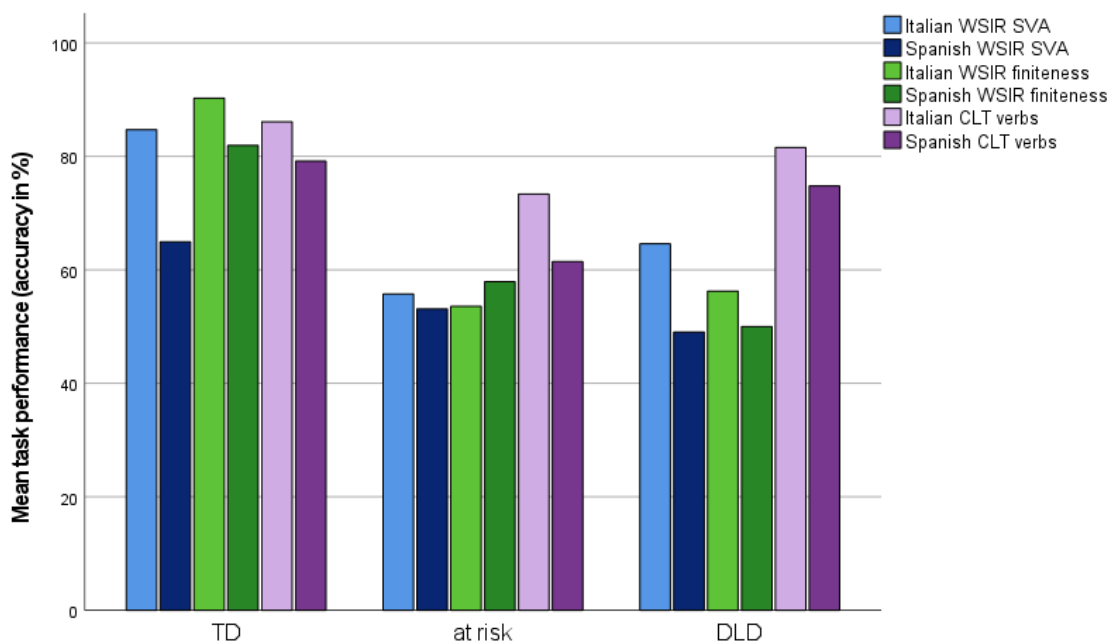


Figure 18: Mean task performance for screening tasks in both languages according to the participants' DLD risk levels.

5.3.3.2 Comparison of performance in the screening's and standardized tasks

The percentage of correctly repeated NWs in the Italian-Spanish NWRT screening was significantly associated with the percentage of correctly repeated NWs in the standardized BVL subtest ($n = 36, r = .780, p < .001$, see also figure 19) and the z-scores obtained for the latter ($n = 36, r = .780, p < .001$).

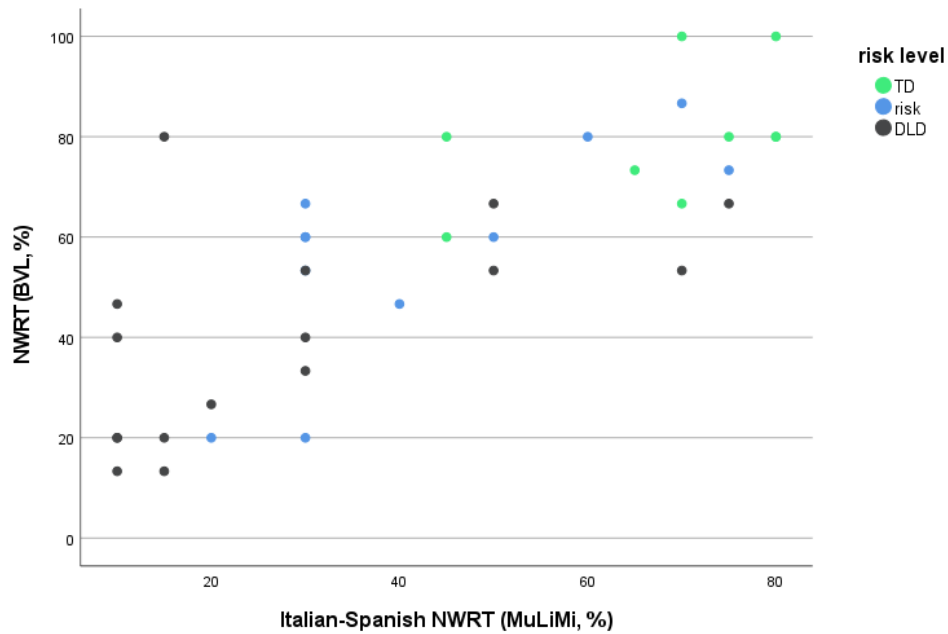


Figure 19: Repetition performance (accuracy in %) in the standardized test (BVL, Marini et al., 2015) and in the MuLiMi NWRT screening task according to the risk levels.

Both accuracy (in %) in the Italian WSIR subject-verb agreement and finiteness tasks were significantly associated with accuracy (in %) and the z-scores deriving from the latter in the sentence repetition (see figure 20) and the sentence completion subtests of the BVL ($n = 36$, r_s ranging from .435 to .634, $p_s < .008$). Accuracy in the Spanish finiteness task was significantly correlated with accuracy (in %) and the corresponding z-scores with the same BVL subtests ($n = 36$, r_s ranging from .420 to .558, $p_s < .008$), but there were no significant associations between the performances in the Spanish WSIR subject-verb agreement task and the BVL subtests ($p_s > .05$). Interestingly, none of the Italian screening tasks but accuracy (in %) in the Spanish WSIR finiteness task was significantly associated with the z-scores children obtained in the BVL grammaticality judgement subtest ($n = 36$, $r = .457$, $p = .013$).

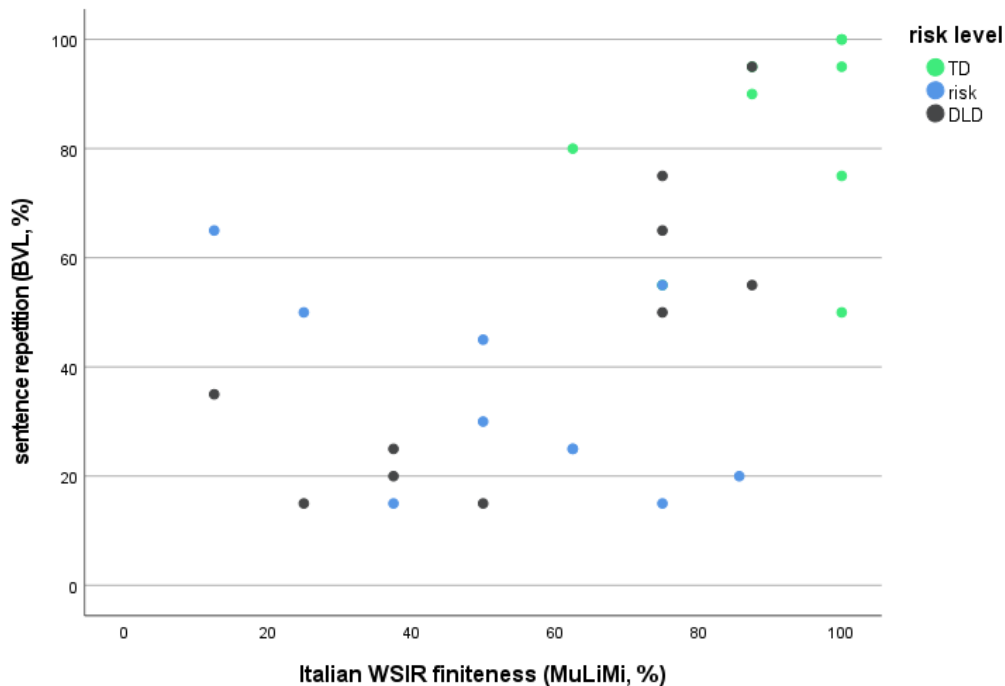


Figure 20: Task performance (accuracy in %) in the standardized sentence repetition test (BVL, Marini et al., 2015) and in the Italian WSIR finiteness screening task according to the risk levels.

The percentages of correctly identified items in both the Italian and the Spanish version of the CLT verb comprehension tasks were significantly associated with accuracy (in %) in the Italian standardized test TFL as well the z-scores deriving from the latter ($n = 36$, r_s ranging from .495 to .743, $p_s < .002$). Accordingly, performance (in %) in the Italian and the Spanish version of the CLTs were also significantly correlated ($n = 36$, $r = .701$, $p < .001$). Interestingly, neither the standardized nor the screening tasks assessing word comprehension were associated with the information on the presence of a diagnosis or the risk level ($p > .05$).

Even though they are not assessing identical linguistic skills, the children's performance in the Italian and Spanish versions of the DNWL were compared to the results obtained in the standardized tests to investigate the diagnostic potential of the newly developed DNWL screening tasks. The children's performance in the Italian version of the consolidation phase was significantly associated with all raw scores obtained in the standardized subtests with ρ_{os} ranging from .527 to .606 and $p_s < .036$ ($n = 16$). Performance in the Italian version of the naming subtest was significantly correlated with the BVL's sentence repetition ($n = 16$, $\rho = .549$, $p = .028$). Comparing the z-scores obtained in the standardized tasks, only one significant association emerged between the children's performance in the BVL NW repetition subtest and the performance in the naming phase in the Italian DNWL ($n = 16$, $\rho = .520$, $p = .039$). Children's performance in the consolidation and naming phases in the Spanish version

of the DNWL was significantly associated with the raw scores obtained in all subtests of the BVL and the TFL (except for the BVL's sentence comprehension and DNWL naming phase) with ρ_s ranging from .522 to .718 and $p_s < .046$ ($n = 15$). When running these comparisons using z-scores instead of raw scores for the standardized tests, the children's performance in the TFL was significantly correlated with the performance in the Spanish testing ($n = 16$, $\rho = .544$, $p = .036$) and naming phases ($n = 16$, $\rho = .716$, $p = .003$).

5.3.3.3 Comparison of performance in screening tasks and SLT & teacher questionnaires

In a first step, the questionnaires for children who were recruited in SLT clinics and accordingly filled in by the participants' SLTs were compared to performance in the screening tasks. The children's repetition NW performance (accuracy in %) was significantly associated with the SLTs' evaluation of the children's receptive ($n = 17$, $\rho = -.592$, $p = .012$) and productive ($n = 10$, $\rho = -.502$, $p = .040$) morphosyntactic skills. Their evaluation of the children's productive lexical skills was marginally significantly (10% alpha-level) associated with the children's NWRT performance ($n = 16$, $\rho = -.488$, $p = .055$). Significant correlations also emerged comparing the children's NWRT performance to the productive ($n = 17$, $\rho = -.573$, $p = .016$) and the total compound score ($n = 17$, $\rho = -.547$, $p = .023$) deriving from the SLTs' responses to the questionnaire.

Accuracy in the Italian WSIR subject-verb agreement task however, was significantly associated only with the children's receptive lexical skills as judged by the SLT ($n = 16$, $\rho = -.584$, $p = .017$). Accuracy (in %) in the Italian WSIR finiteness task was significantly correlated with the judgement on the child's productive lexicon skills ($n = 10$, $\rho = -.713$, $p = .021$). Despite the fact that the number of children who were administered all DNWL subtests and who were recruited through SLT clinics is very low ($n = 8$), significant associations with the SLTs' judgements on their linguistic skills and performance in the Italian DNWL test and naming subtests emerged. In particular, accuracy in the test phase of the Italian DNWL was significantly associated with the SLTs total ($n = 8$, $\rho = -.794$, $p = .019$) and total productive score ($n = 8$, $\rho = -.859$, $p = .006$). Accuracy in the Italian DNWL's test phase was furthermore associated with SLTs' judgements of receptive ($n = 8$, $\rho = -.856$, $p = .007$) and productive morphosyntax skills ($n = 8$, $\rho = -.898$, $p = .002$) as well as with SLTs' evaluation of pragmatic skills ($n = 7$, $\rho = -.805$, $p = .029$).

All three, the total, reception and production compound scores correlated significantly with the accuracy (in %) in the Italian CLT verb comprehension subtest ($n = 17$, ρ_s ranging from $-.492$ to $-.643$, $p_s < .045$). Interestingly, their judgement of the children's receptive lexical

skills – the purpose of the CLT Italian verb comprehension subtest – was not significantly correlated ($n = 16$, $\rho = -.473$, $p = .064$).

Despite the fact that the SLTs were able to exclusively judge the children's Italian language skills, significant associations also emerged between their evaluations and the children's performance in the Spanish screening tasks. Similar to the Italian version of the WSIR subject-verb agreement task, the only significant association emerged for accuracy (in %) in the Spanish version of this task with lexical receptive skills ($n = 16$, $\rho = -.556$, $p = .025$). Also for the Spanish versions of the test and naming DNWL subtests significant associations with the SLTs judgements on the children's skills in various linguistic domains emerged: Accuracy (in %) in the testing phase of the Spanish DNWL was significantly associated with the receptive total score ($n = 8$, $\rho = -.708$, $p = .049$), the morphosyntax receptive score ($n = 8$, $\rho = -.723$, $p = .043$) and the lexical receptive ($n = 8$, $\rho = -.914$, $p = .001$) and productive score ($n = 8$, $\rho = -.901$, $p = .002$). Significant correlations emerged also when comparing accuracy (in %) in the Spanish DNWL naming subtest to the receptive ($n = 8$, $\rho = -.797$, $p = .018$) and the productive score ($n = 8$, $\rho = -.838$, $p = .009$). Furthermore, accuracy in the Spanish version of the CLT verb comprehension task was significantly associated with the total ($n = 17$, $\rho = -.586$, $p = .013$), the total production scores ($n = 17$, $\rho = -.556$, $p = .021$), which is resembled in the significant associations emerging with the judgement of morpho-syntactic productive ($n = 17$, $\rho = -.536$, $p = .027$) and lexical productive skills ($n = 17$, $\rho = -.671$, $p = .004$).

Before comparing the children's performances in the screening to the judgement of their language skills by their kindergarten teachers, the kindergarten teachers' judgements were compared to the children's risk level since these are based on subjective, standardized testing procedures carried out within this study. Associations between the children's risk level and all compound scores deriving from the kindergarten teachers' questionnaire responses and accordingly with all single scores emerged ($n = 19$, ρ_s ranging from $-.456$ to $-.710$, $p_s < .05$), except for teachers' judgement of children's phonological, morphosyntactic and lexical receptive skills ($p_s > .05$). The lack of associations between the kindergarten teachers' judgement of receptive skills and the risk level can be explained by the lack of tests (and time to administer them) available to pedagogic staff in kindergartens. However, the associations show that it is reliable to compare the children's screening task performances to the kindergarten teachers' judgements. All single and compound scores deriving from the kindergarten teachers' questionnaire responses were significantly associated with the children's repetition accuracy (%) in the MuLiMi NWRT screening task with ρ_s ranging from $-.476$ to $-.818$, $p_s <$

.039 ($n = 19$). The accuracy (in %) of children's responses in the Italian WSIR subject-verb agreement task was significantly correlated with all single scores ($n = 19$, ρ s ranging from -.491 to -.552, p s < .033) but the receptive morphosyntax and lexical as well as pragmatics skills (p s > .05). Accordingly, WSIR subject-verb agreement performance was significantly associated with the total ($n = 19$, $\rho = -.555$, $p = .014$) as well as the total productive score ($n = 19$, $\rho = -.552$, $p = .014$) emerging from the single kindergarten teachers' judgements (see figure 21). Despite the associations not consistently reaching significance, a similar pattern was observed comparing children's performance (accuracy in %) in the Italian WSIR finiteness task to the responses of the kindergarten teacher questionnaire: Again, the total score ($n = 19$, $\rho = -.469$, $p = .049$), the total production score ($n = 19$, $\rho = -.454$, $p = .058$) and the morphosyntax production score ($n = 19$, $\rho = -.454$, $p = .058$) were associated with accuracy in the Italian WSIR finiteness task. In addition to that, the kindergarten teachers' judgement of receptive phonological skills was associated ($n = 19$, $\rho = -.467$, $p = .051$). Accuracy (in %) in the Italian CLT verb comprehension test was significantly correlated with the receptive compound score ($n = 19$, $\rho = -.484$, $p = .036$) and the kindergarten teachers' judgement of receptive morphosyntax skills ($n = 19$, $\rho = -.535$, $p = .018$). While none of the kindergarten teachers' questionnaire response categories was significantly correlated with accuracy in the Italian DNWL's consolidation and test phase (p s > .05), several associations emerged between accuracy in the Italian DNWL's naming phase and the responses by the kindergarten teachers. Namely, the total score ($n = 19$, $\rho = -.690$, $p = .058$) as well as kindergarten teachers' judgement on receptive ($n = 19$, $\rho = -.784$, $p = .021$) and productive phonological skills ($n = 19$, $\rho = -.876$, $p = .004$) were associated with the children's accuracy in the naming phase of the Italian DNWL.

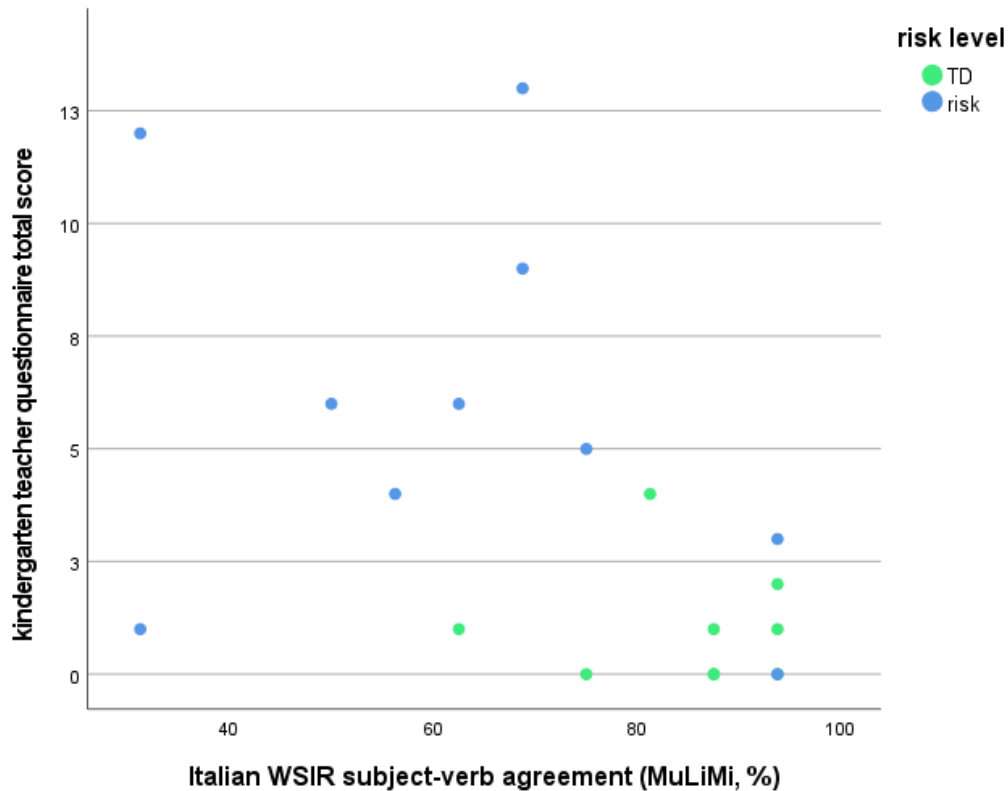


Figure 21: Plot of kindergarten teachers' evaluation of children's language performance against task performance (accuracy in %) in the Italian WSIR subject-verb agreement screening tasks according to the risk levels.

Moreover, the children's performance in the Spanish screening tasks was significantly associated with the kindergarten teachers' judgement of their linguistic performance. Similar to the SLTs' evaluation of the children's language skills compared to Italian screening task performance, fewer associations emerged between the kindergarten teachers' evaluation of the children's language skills and the performance in the Italian screening tasks. While none of the kindergarten teachers' judgements of the children's linguistic skills was found to be associated with the children's performances in the Spanish version of the WSIR finiteness task, some associations emerged for the WSIR subject-verb agreement task, namely with the total score ($n = 19$, $\rho = -.444$, $p = .057$), the total production score ($n = 19$, $\rho = -.471$, $p = .042$) and the productive phonological score ($n = 19$, $\rho = -.483$, $p = .036$). For the Spanish version of the CLT verb comprehension subtest, only the kindergarten teachers' evaluation of productive morphosyntax skills was significantly associated ($n = 19$, $\rho = -.484$, $p = .036$). Other than for the Italian version of the DNWL subtests, associations with subtests of the Spanish DNWL task emerged for the testing phase only with the total score ($n = 7$, $\rho = -.782$, $p = .038$) and the receptive phonological score ($n = 19$, $\rho = -.718$, $p = .069$).

5.3.3.4 Comparison of performance in the screening tasks and caregiver questionnaires

First of all, the children's risk level deriving from the performance in standardized Italian tests and the information about the presence/absence of an official diagnoses of DLD was compared to the responses to the caregiver questionnaire QUIR-DC. All the scores of the response categories that were significantly associated with the children's risk levels were analysed regarding their association with screening task performance. Depending on the type of the response category, either Spearman rho's or Pearson's correlations were calculated.

The children's performance in the NWRT screening task was significantly associated with the QUIR-DC GS ($n = 34$, $r = .553$, $p > .001$) and RS ($n = 34$, $r = -.517$, $p = .002$). Neither the QUIR-DC FIGS nor the FIRS correlated significantly with NW repetition performance in the MuLiMi NWRT ($p_s > .05$). Accuracy (in %) in the Italian WSIR subject-verb agreement screening task was significantly associated with the QUIR-DC GS ($n = 34$, $r = .416$, $p = .015$) and the QUIR-DC RS ($n = 34$, $r = -.398$, $p = .020$). No significant associations were found comparing the performance in the WSIR finiteness task to the responses in the caregiver questionnaire. A tendency, however, was observed for the QUIR-DC GS ($n = 27$, $r = .362$, $p = .063$). The children's performance in the Italian CLT verb comprehension subtest (accuracy in %) was significantly correlated with the QUIR-DC RS ($n = 34$, $r = -.372$, $p = .030$). While for the consolidation and test phase of the Italian DNWL no significant associations with the caregiver questionnaire emerged ($p_s > .05$), children's performance in the naming phase (accuracy in %) was significantly associated with the QUIR-DC RS ($n = 16$, $rho = -.517$, $p = .040$).

Significant correlations emerged when comparing performance in certain Spanish tasks to responses to the caregiver questionnaire. None of the response categories from the QUIR-DC were significantly associated with children's performance (accuracy in %) in the Spanish WSIR screening task on subject-verb agreement ($p_s > .05$). For the Spanish WSIR finiteness task however, a significant association emerged with the QUIR-DC GS ($n = 27$, $r = .422$, $p = .028$) and a non-significant one with the QUIR-DC RS ($n = 27$, $r = -.373$, $p = .055$). Accuracy in the Spanish CLT verb comprehension task was significantly correlated with the QUIR-DC RS ($n = 34$, $r = -.373$, $p = .030$). Other than for the Italian version of the DNWL naming subtest, no significant associations between the Spanish DNWL subtests and the responses to the caregiver questionnaire emerged ($p_s > .05$).

5.3.3.5 Comparison of performance in the screening tasks

The children's performance in the Italian-Spanish NWRT MuLiMi screening task was significantly associated with performance in other tasks from different linguistic domains. More spe-

cifically, significant associations were observed comparing the children's percentages of correctly repeated NWs to the performance in the Italian CLT verb comprehension subtest ($n = 36, r = .469, p = .004$) and their performance in the WSIR subject-verb agreement screening task ($n = 36, r = .334, p = .046$). Also, the children's performance in the naming phase of the Italian version of the DNWL was significantly correlated with children's NW repetition performance ($n = 16, rho = .757, p < .001$). NW repetition performance in the screening was also significantly associated with some of the Spanish screening task performances, namely with the performance (accuracy in %) in the Spanish version of the CLT verb comprehension subtest ($n = 36, r = .542, p < .001$) and in the Spanish version of the WSIR finiteness screening ($n = 29, rho = .372, p = .047$).

Correlations were also found for children's performance (accuracy in %) in the Italian WSIR subject-verb agreement screening task and their performance in the Italian WSIR finiteness screening tasks ($n = 29, r = .791, p < .001$). The same pattern was found comparing the children's performance in the two Spanish WSIR screening tasks to each other ($n = 29, r = .506, p = .005$). Furthermore, crosslinguistic associations were found comparing the WSIR subject-verb agreement screening task in its Italian and Spanish version ($n = 29, r = .413, p = .012$). And similarly, the performance in the Italian version of the WSIR finiteness screening task was significantly associated with performance in the Spanish version ($n = 29, r = .544, p = .002$).

Performance (accuracy in %) in the Italian verb comprehension CLT subtest was significantly associated with the performance in the Spanish version of the same task ($n = 36, r = .701, p < .001$). In contrast, no associations between the performance in the DNWL screening task subtests were neither found within nor across languages ($p_s > .05$).

As mentioned in chapter 2.4.1, an insufficient number of diagnostic tools that assess both languages spoken by the child yield the potential for misdiagnoses. Consequently, the appropriateness of the diagnoses of the children with DLD diagnosis was assessed by comparing their performances in the same screening task in both language versions. Even though, due to the small amount of NWs selected and used in the MuLiMi NWRT, for the purpose of the investigation of crosslinguistic associations between the repetition performance of LS Italian and LS Spanish NWs, the DLD children's repetition performance of the $n = 4$ LS Italian NWs were compared to the repetition performance of the $n = 4$ LS Spanish NWs and a significant association emerged ($n = 16, rho = .498, p = .049$). No significant association instead was found comparing the children's performance in the Italian WSIR subject-verb agreement ($n = 16$) and finiteness ($n = 10$) screening tasks to the performance in

the Spanish version of these tasks ($p_s > .05$). A significant association instead emerged comparing the children’s performance (accuracy in %) in the Italian version of the CLT verb comprehension to the Spanish version ($n = 16$, $\rho = .610$, $p = .012$). Also, for none of the DNWL subtests did significant associations with the same or other subtests in the other language emerge ($n = 8$, $p_s > .05$), with the exception of performance in the Italian naming and the Spanish testing phase ($n = 16$, $\rho = .818$, $p = .013$).

5.3.3.6 Comparison of performance in the screening tasks at t1 and at t2

Since the small sample size of children included in the follow-up study does not allow for statistical analyses, the performance levels at t1 is compared to t2 upon visual inspection of graphs (figures 22, 23). These illustrate the performances in the same tasks for both languages at t1 and t2 respectively for the single participants. All of these children had been diagnosed with DLD at t1 and received SLT treatment in Italian between t1 and t2. At this point, it should be added that these observations have only limited informative value and have the aim to grasp tendencies.

While for the majority of children repetition accuracy increased from t1 to t2 (see Participants 1 to 3), performance level was maintained for Participant 4 and decreased for Participant 5, see figure 22.

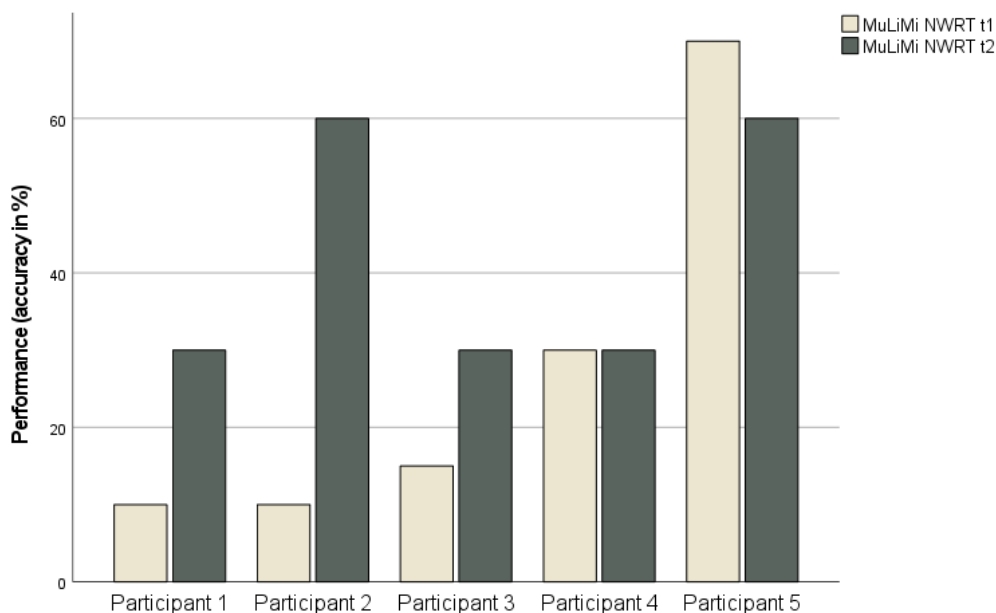


Figure 22: Comparison of performance (accuracy %) at t1 and t2 per participant in the NWRT screening task.

Figure 23 displays the children’s performance in the WSIR subject-verb agreement (SVA) task at t1 (light blue for Italian, light green for Spanish) and t2 (dark blue for Italian, dark green for Spanish). All children exception of participant 4 did show improvement in the

Italian version of this task. While participant 4 and 5 showed improvement for the same task also in its Spanish version, the other children did show worsening in this task.

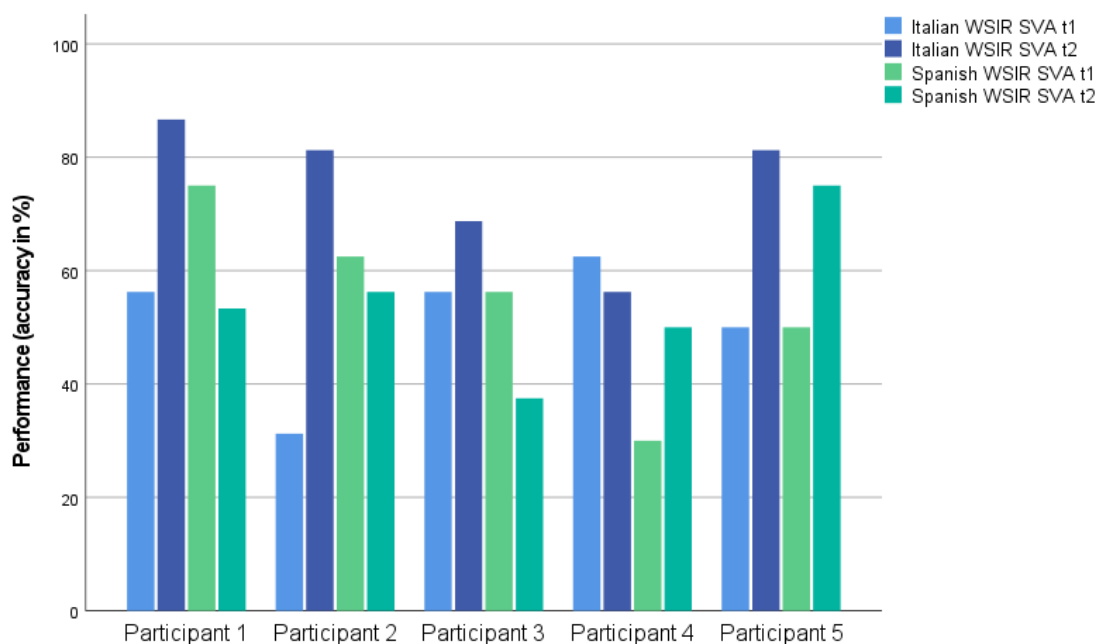


Figure 23: Comparison of performance (accuracy %) at t1 and t2 per participant in the Italian and Spanish versions of the WSIR screening tasks.

The pattern observed in figure 23 is very similar to the one observed in figure 24 in which the children’s performances in the CLT verb comprehension tests in Italian and Spanish at t1 and t2 are displayed. While again for improvement from t1 to t2 in the Italian version of this task all children except for participant show improvement, all children but not participant 1 show improvement in the Spanish version of this task from t1 to t2.

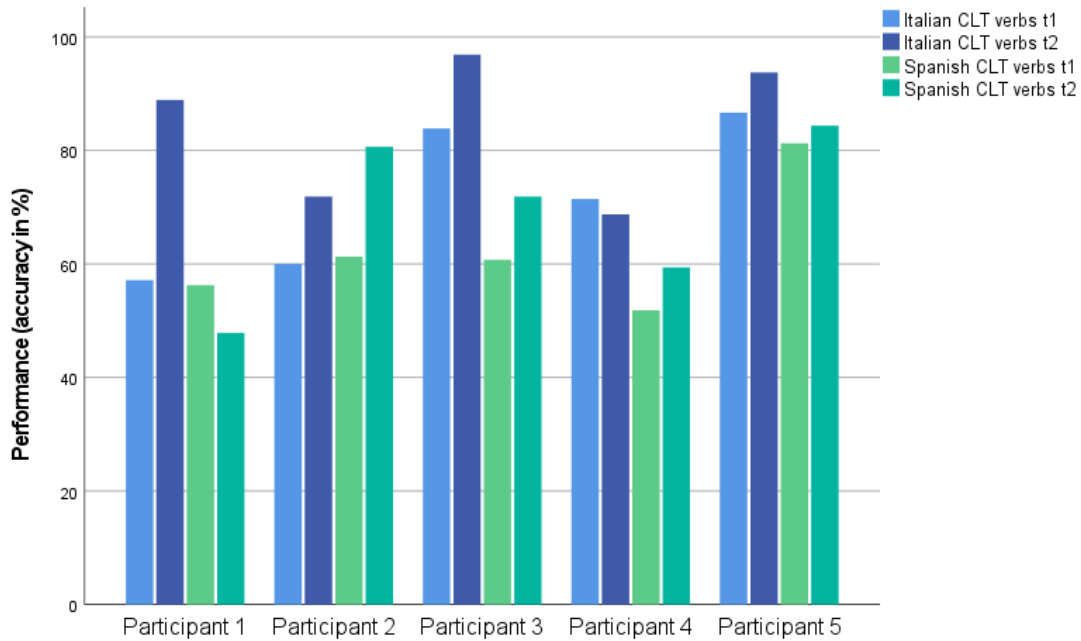


Figure 24: Comparison of performance (accuracy %) at t1 and t2 per participant in the Italian and Spanish versions of the CLT verb comprehension subtests.

5.3.3.7 Interim discussion

The results presented allow to answer the research questions on concurrent and discriminant validity. Across linguistic areas, significant associations between screening task performance and the variables concerning the risk level, standardized tests and caregiver, SLT and kindergarten teacher questionnaire emerged (for an overview, see table 6) indicating concurrent and discriminant validity.

Table 6: Overview of significant correlations between screening task performance (accuracy in %) risk level, standardized tests and SLT, kindergarten teacher and caregiver questionnaires.

Screening task (accuracy %)	risk level (TD, risk, DLD)	standardized tests (raw scores)	questionnaire		QUIR-DC
			SLTs' responses	teachers' responses	
NWRT	$n = 36,$ $\rho = -.545,$ $p > .001$	BVL nonword repetition $n = 36, r = .780,$ $p < .001$	compound score $n = 17, \rho = -.547,$ $p = .023$	compound score $n = 19, \rho = -.804,$ $p < .001$	general score $n = 34, r = .553,$ $p > .001$
WSIR finiteness (IT)	$n = 36,$ $\rho = -.524,$ $p = .004$	BVL sentence compr. $n = 36, r = .515,$ $p < .001$	productive lexicon $n = 10, \rho = -.713,$ $p = .021$	compound score $n = 19, \rho = -.469,$ $p = .049$	n.s.
WSIR finiteness (SP)	$n = 36,$ $\rho = -.535,$ $p = .003$	BVL sentence compr. $n = 29, r = .420,$ $p = .023$	n.s.	n.s.	general score $n = 27, r = .422,$ $p = .028$
WSIR	$n = 36,$ $\rho = -.432,$	BVL sentence compr.	receptive lexicon $n = 16, \rho = -.584,$	compound score $n = 19, \rho = -.555,$	general score $n = 34, r = .416,$

subject-verb agreement (IT) $p = .009$ $n = 29, r = .556, \rho = .002$ $p = .017$ $p = .014$ $p = .015$

Screening task (accuracy %)	risk level (TD, risk, DLD)	standardized tests (raw scores)	questionnaire		QUIR-DC
			SLTs' responses	teachers' responses	
WSIR subject-verb agreement (SP)	$n = 36,$ $\rho = -.322,$ $p = .056$	n.s.	receptive lexicon $n = 16, \rho = -.556,$ $p = .025$	compound production $n = 19, \rho = -.471,$ $p = .042$	n.s.
CLT verb comprehension (IT)	n.s.	TFL $n = 36, r = .743,$ $p < .001$	compound score $n = 17, \rho = -.643,$ $p = .005$	compound score $n = 19, \rho = -.484,$ $p = .036$	risk score $n = 34, r = -.372,$ $p = .030$
CLT verb comprehension (SP)	n.s.	TFL $n = 36, r = .495,$ $p = .002$	compound score $n = 17, \rho = -.586,$ $p = .013$	productive morphosyntax $n = 19, \rho = -.484,$ $p = .036$	risk score $n = 34, r = -.373,$ $p = .030$
DNWL consolidation (IT)	n.s.	BVL sentence repetition $n = 16, \rho = .606,$ $p = .013$	n.s.	n.s.	n.s.
DNWL testing (IT)	n.s.	n.s.	compound score $n = 8, \rho = -.856,$ $p = .007$	n.s.	n.s.
DNWL naming (IT)	n.s.	BVL sentence repetition $n = 16, \rho = .549$ $p = .028$	n.s.	productive phonology $n = 19, \rho = -.876,$ $p = .004$	risk score $n = 16,$ $\rho = .517,$ $p = .040$
DNWL consolidation (SP)	n.s.	n.s.	n.s.	n.s.	n.s.
DNWL testing (SP)	n.s.	BVL sentence repetition $n = 15, \rho = .718,$ $p = .003$	receptive lexicon $n = 8, \rho = -.914, p = .001$	compound score $n = 7, \rho = -.782,$ $p = .038$	n.s.
DNWL naming (SP)	n.s.	BVL sentence repetition $n = 15, \rho = .578,$ $p = .024$	compound productive $n = 8, \rho = -.838,$ $p = .009$	n.s.	n.s.

Data and results indicate that the screening is a contribution to DLD risk identification. Generally, according to the data collected and analysed upon administering the Spanish-Italian DLD screening to children with, at risk and without DLD attending kindergartens in Italy, the computerized Spanish-Italian MuLiMi DLD screening that automatically assesses children's language performance in both languages spoken has the potential to contribute to the identification of DLD. This has been shown by the observation that children with a diagnosis or at risk of DLD (based on information by the SLT, from caregiver and teacher questionnaires and standardized/traditional tests) underperformed children with no or a lower risk

score in screening tasks (Hypothesis 1). Children with diagnosis of or at risk of DLD as indicated by the SLTs or kindergarten teachers did show more difficulty in correctly recognizing and associating the novel words (NWs) introduced in the dynamic novel word learning task (DNWL). They also had more difficulties in recalling them (Hypothesis 1a). Interestingly, not risk level or standardized test scores were significantly associated with the children's performance in the DNWL subtests, but the caregivers', SLTs' and teachers' responses to the respective questionnaires only. This somehow reflects the novelty of this task that other than standardized tests per se, the risk level deriving from the latter or an official diagnosis does not assess one specific linguistic skill, but a series of abilities that interact with each other. This specific situation of acquiring and using new words in the DNWL screening task incorporating a series of linguistic skills is somewhat more comparable to the situations in which children naturally use language. Those situations are the basis of the caregivers', SLTs' and kindergarten teachers' responses to the questionnaire and might explain why not standardized test results, but caregiver and teacher questionnaire responses are significantly associated with the children's performance in the DNWL. Furthermore, significant associations between the children's results obtained in standardized tests with screening tasks declared to measure the same skills indicate the suitability of the task paradigms and items applied (Hypothesis 2). Also, performance in the different screening tasks on morphosyntactic processing are associated with each other (Hypothesis 3). In particular, performance in the subject-verb agreement and finiteness screening tasks were associated with each other in the respective languages (Hypothesis 3a). Similarly, performance on various screening tasks across linguistic areas assessing the same linguistic areas in the two different languages are correlated with each other (Hypothesis 4). More specifically, performance in the repetition of Spanish specific NWs was associated with repetition accuracy of Italian NW (Hypothesis 4a). Also, performance in the Spanish verb comprehension CLT subtest was significantly associated with the Italian version of this task (Hypothesis 4b). Significant associations were also found comparing performance in the Italian WSIR subject-verb agreement screening tasks with performance in the Spanish version of the same task (Hypothesis 4c). Similar effects were found comparing children's task performance in the Spanish and Italian version of the WSIR finiteness screening tasks (Hypothesis 4d). However, not for all subtests in the Italian DNWL subtest significant associations with performance in the Spanish DNWL subtests were found (Hypothesis 4e). A possible explanation might be the circumstance that due to the sample size, the presentation of screening tasks was not randomized or inverted for a subgroup of study participants. It is possible that during the Spanish DNWL, task that has always been administered towards the end of the screening session, children were less focused compared to

when doing the Italian version of the DNWL in the beginning of the screening session. Hypothesis 5 instead was only partly supported. When analysing the screening task performances across languages of children with a DLD diagnoses only, not for all subjects and screening tasks across linguistic areas performance in the Italian and the Spanish version of this task were considered similar to each other upon visual inspection (no statistical analyses possible due to small sample size). It is not clear whether this mismatch of screening task performance between the Italian and Spanish screening tasks can be explained by a) the effects of language dominance, b) misdiagnoses or c) the inappropriateness to assess Spanish-Italian children attending kindergartens in Italy with the MuLiMi DLD-screening. Predictive value of the task performance at t1 for task performance at t2 could not be assessed due to the small sample size. According to visual inspection of the data, generally children's performance did improve from t1 to t2 in the Italian versions of the tasks. While for the Spanish CLTs the majority of children did show improvement from t1 to t2, the pattern was not as clear for the Spanish WSIR subject-verb agreement task (Hypothesis 6). These observations have only limited value and serve to observe tendencies in the data.

Besides the assessment of the suitability of the Spanish-Italian screening for DLD risk identification, the high amount of children who participated in this study and managed to carry out all of the screening tasks indicates that generally the child's interaction with the online screening tool is a) feasible and b) motivating for children of this age (see Hautala et al., 2020). This is of particular importance in the light of the remote modality in which this screening study was carried out in collaboration with SLT clinics and kindergartens. The high and consistent participation rate, thus, also indicates the appropriateness and feasibility of remote assessment for DLD risk identification in bilingual children attending kindergartens.

5.4 Bilingual, computerized DLD screening for Italian-speaking children living in Germany

As mentioned in chapter 2.3.2, Italy is among the most frequent countries of origin for foreigners residing in Germany. Due to the guest worker programs established in the 1950s, Italian-speakers have been living in Germany for a long time. Accordingly, an Italian-German DLD screening was constructed and administered to children with, at risk of and without DLD in the presence of the examiner. Modified versions of parts of this chapter have been published in *Frontiers in Psychology*: "A Nonword Repetition Task Discriminates Typically Developing Italian-German Bilingual Children from Bilingual Children with Developmental Language Disorder: The Role of Language-Specific and Non-Language-Specific Nonwords" (Eik-

erling, et al., 2022a) and been accepted for publishing in *Lingue e Linguaggio*: “A web-platform for DLD screening in Italian-German-speaking children: preliminary data on concurrent and predictive validity” (Eikerling & Lorusso, in press). Modified versions of other parts of this chapter have been submitted to *Clinical Linguistics and Phonetics* in November 2021.

5.4.1 Hypotheses

Given the specific tasks used in this study, Hypothesis 3 (see chapter 5.1) was further subdivided for the purpose of this study and it was hypothesized that:

- a) performance in noun comprehension is associated with performance in verb comprehension (Hypothesis 3a).
- b) performance in the German case-marking picture matching task is associated with performance in the German subject-verb agreement judgement task (Hypothesis 3b).
- c) performance in the Italian subject-verb agreement judgement task is associated with the judgement of clitic object constructions (tested at t2 only, Hypothesis 3c).

Also Hypothesis 4 was subdivided:

- a) performance on the German test of NW repetition is associated with the performance in the Italian test of NW repetition (Hypothesis 4a).
- b) performance in German word comprehension is associated with Italian word comprehension (in particular, German noun comprehension with Italian noun comprehension and German verb comprehension with Italian verb comprehension) (Hypothesis 4b).
- c) performance in the Italian test of subject-verb agreement is associated with performance in the German test of subject-verb agreement (Hypothesis 4c).

In addition to the hypotheses described in chapter 5.1 it was hypothesized that performance on the screening tests at t1 would show a general trend of associations with performance at t2 (follow-up, Hypothesis 6) and more precisely that:

- a) performance at t1 would be associated with performance at t2 (Hypothesis 6a)
- b) performance at t1 would be predictive of improvement from t1 to t2 (Hypothesis 6b)
- c) other variables such as age, language dominance and/or IQ could show an influence on improvement from t1 to t2 (Hypothesis 6c)

5.4.2 Material & methods

To answer these research questions, the following methods and material were applied in this study.

5.4.2.1 Participants

Thirty-nine early-sequential or simultaneous bilingual Italian-German-speaking children aged from 3;10 to 6;2 years participated in this study (mean age in months: $M = 59.22$, $SD = 8.69$). Two children had to be excluded due to conflicting information regarding whether they receive(d) SLT or not between the caregiver questionnaire and kindergarten teachers' statements. Of those 37 study participants, $n = 7$ of these children had already been diagnosed with DLD by an SLT and received treatment in Germany. Another $n = 17$ children had not been diagnosed with DLD, but scored below cut-off according to the standardized test manuals, while $n = 13$ had neither scored below cut-off nor been diagnosed with DLD and are thus considered TD. All children lived in Germany, attended kindergarten and had been exposed to the German language for at least two years. At least one of the caregivers is a native speaker of Italian. Recruitment took place through either kindergartens or the SLT clinics.

5.4.2.2 Screening tasks

All audio clips were recorded by female native speakers with natural voice and accent. Depending on the language characteristics, see chapter 2.4.1.1, screening tasks were chosen and implemented accordingly.

Nonword repetition task (NWRT). The Italian-German version of the NWRT underlies the same principle as the Spanish-Italian described in chapter 5.3. Also these NWs were chosen based on the direct and online ratings by adult native speakers regarding the language-specificity of the NWs and then selected depending on their reliability scores. For inter-rater-reliability (based on $n = 3$ raters, 1 German-Italian bilingual, 1 German and 1 Italian native speaker), Cronbach's Alpha was $\alpha > .70$ for all NWs. Internal consistency as expressed by Cronbach Alpha was $\alpha = .903$. The list of NWs used in the Italian-German DLD screening consisted in $n = 6$ LS Italian, $n = 6$ LS German and $n = 9$ NLS NWs, see table 7.

Table 7: Overview of NWs selected for the Italian-German NWRT.

amount of syllables	LS German NWs ($n = 6$)	LS Italian NWs ($n = 6$)	NLS NWs ($n = 9$)	
			NLS German NWs ($n = 4$)	NLS Italian NWs ($n = 5$)
2	[fe'larŋk]	['dalmo]	[maful]	[lefum]
	['nɛ:ɣlax]			
	[p'a'molt]			
	[resol'ant]	['spulfarɔ]	[nisala]	[famelep]

	[kleŋ'ketɕ]	[stal'mo:no]	[lifena]	[fulsamit]
3				[melinak]
4	[tulmefo'kans]	[bjɛla'nare]	[minalefe]	[nufalemik]
		[maŋke'tale]		
		[rako'denso]		

Also these pre-recorded NWs were automatically administered through the MuLiMi screening platform providing visual feedback on the screen (see chapter 5.3.2.2). They were manually evaluated for correct (1 point) vs. incorrect (0 points) repetition by the examiner. To represent the specific language acquisition conditions of the Italian-German-speaking child participants, the NWRT score used for further analyses was based on the German-Italian bilingual speaker's scores.

Cross-linguistic Lexical Tasks (CLTs). The German version of the CLTs underlies the same principle as the Italian and Spanish CLT subtests described in chapter 5.3.2.2. Again audio instructions by native speakers of Italian and German respectively are provided. Here instead, both noun and verb comprehension CLT subtests in Italian and German (Haman et al., 2017) were administered as described above. While again the Italian verbs are embedded in a question using the gerund (see above), in the German verb comprehension the request “Zeige mir...” [Show me...] is followed by an infinitive (e.g. “öffnen” [to open]). Also here, each of the subtests contains 32 items (64 items per language, 128 in total). Other than described in chapter 5.3.2.2, the Italian-German-speaking children were not administered the CLT items using the MuLiMi web app, but tested using the CLT app (Zinn, unpublished) that allows for automatic administration, but not for automatic evaluation of the children's responses. Matching accuracy was assessed manually.

Case matching. A pre-recorded sentence containing an object marked for case – either accusative or dative – is presented auditorily, for example „Der Esel rennt in das Zimmer.“ (accusative object, [The donkey runs into the room.]). Simultaneously, two pictures are displayed on the screen. One of them corresponds to the sentence that is presented (target), while the distractor depicts the same protagonists and objects as the first one, but they are arranged such that the auditorily presented sentence does not match because of variation in case-marking of the object (i.e. „Der Esel rennt in dem Zimmer.“ (dative object, [The donkey runs inside the room.]), see figure 25). For this task, only prepositional phrases containing feminine or neuter objects – avoiding masculine objects due to the phonologically challenging distinction of “dem” (dative) vs. “den” (accusative) – are used. The child is asked to indicate

which of the two pictures presented matches the sentence he/she has heard by clicking on the corresponding picture. The task consists of 2 training and a total of 10 screening items with 50% being accusative and 50% being dative objects. Accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.



Figure 25: Example from the German case matching screening task: „Der Esel rennt in das Zimmer.“ [The donkey runs into the room.] with target picture displayed on the left and distractor on the right.

Subject-verb agreement. The screening task on subject-verb agreement is different from the subject-verb agreement task used in the Spanish-Italian screening since it incorporated both a matching and a judgement component. It was used for both the Italian and German part of the screening. Simple subject-verb-phrases were presented auditorily accompanied with a picture depicting the scene. In 50% of these sentences, subject-verb agreement was not met, e.g. “Die Hasen läuft*.” [The rabbits runs*:], see figure 26. It is the child’s task to indicate incongruence between subject and verb by selecting the sad face displayed on the screen. Conversely, for congruent sentences (e.g. “Der Hase läuft.” [The rabbit runs.]), it is the child’s task to identify the picture depicting the scene corresponding to the auditorily presented sentence. Along with the latter and the sad-faced button, another picture is displayed depicting the same scene as in the target sentence, but diverging in number of the subject (in our example this would be more than one rabbit running). Screening tasks in both language versions consisted of 2 items for training. The Italian version consisted of 10, the German version of 8 screening items. Accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

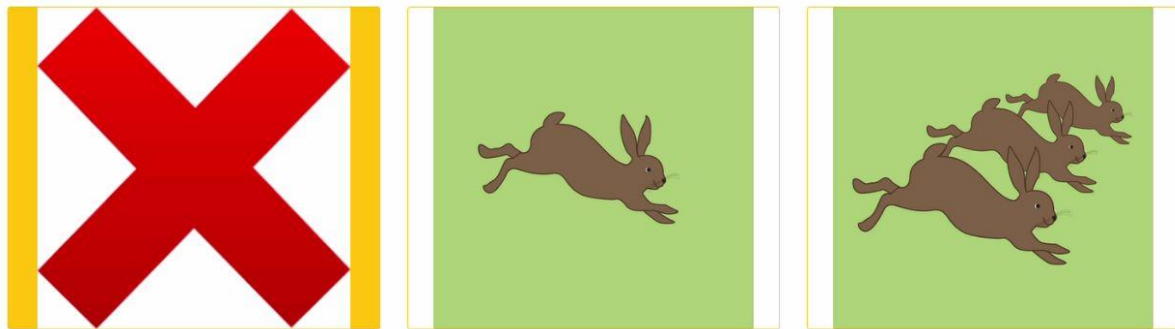


Figure 26: Example from the German subject-verb agreement screening task.

Clitic pronoun judgement. A simpler version of the clitic pronoun judgement task described in chapter 4.1.2.1 is used. For example, “Che cosa fa la mamma alla bambina? Lo* bacia.” [What is the mother doing to the girl? She is kissing him*.] see figure 27, the clitic pronoun “lo” is incorrect, because it should be the accusative-feminine clitic “la” instead of the accusative-masculine “lo”. The stimuli are presented in random order with 50% correct and 50% incorrect clitic pronoun use. They are each accompanied by visual support, comic-like colour paintings depicting the scene described in the auditorily presented sentence. The child is asked to indicate whether the presented sentence is correct or not by selecting the corresponding buttons ☺ for correct and ☹ for incorrect sentences. The task consists of 2 training and a total of 8 screening items. Accuracy of the given responses are automatically measured and stored. Due to time constraints, this task was not administered to all children. Find more examples in appendix B.

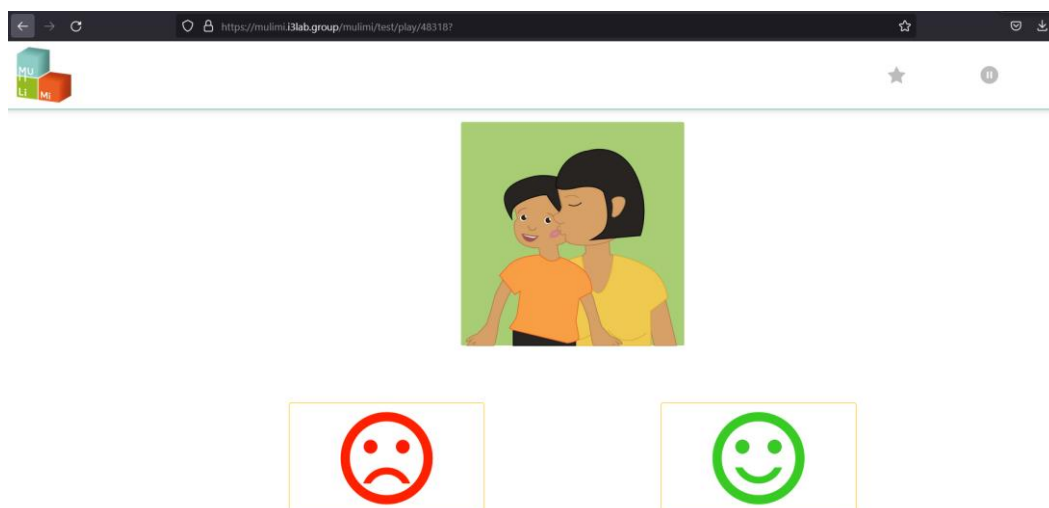


Figure 27: Examinee interface during the Italian clitic pronoun judgement task: “Che cosa fa la mamma alla bambina? Lo* bacia.” [What is the mother doing to the girl? She is kissing him*.] The child is expected to click the sad face on the left standing for incorrect clitic pronoun use.

DNWL (consolidation and testing phase). The German version of the DNWL was structured in a comparable way as the Spanish one. The Italian version was the same in both screening studies. In German, the NWs to be associated with the figures were ['tu:maχ], ['harɔk], and [pʰa'mɔlt]. In the Italian-German version of this task, the results of the naming phase had to be excluded from statistical analyses since phonological onset clue was not consistently provided. Due to time constraints, this task was not administered to all children.

5.4.2.3 Standardized tests

Peabody Picture Vocabulary Test (PPVT-4). The German standardized PPVT-4 (Lenhard et al., 2015) was used to assess receptive vocabulary skills. It consists of 19 blocks each containing twelve items – single nouns, verbs or adjectives – of increasing complexity and decreasing frequency. The task was administered presenting four pictures on a computer screen instead of a picture book. The target words were pre-recorded by a native speaker of German presented one by one via a computer. Children are asked to indicate the picture that corresponds to the word presented. Matching accuracy is assessed manually. Abort criteria and the scoring of responses followed the test manual.

Mottier-Test. The Mottier-Test (original by Mottier, 1951; norm data published by Kiese-Himmel and Risse, 2009) is a NWRT consisting of 30 NWs increasing in length containing between two and six syllables. The complexity of the single syllables instead remains stable (only CV structures). Children are asked to immediately repeat each NW after presentation. Repetition accuracy is assessed manually assigning 1 point for each correctly repeated NW, with a maximum score of 30. The NWs were pre-recorded by a native speaker of German at a one-syllable per second rate (according to Wild & Fleck, 2013) and presented one by one via a computer.

Linguistische Sprachstandserhebung Deutsch als Zweitsprache (LiSe-DaZ). Morpho-syntactical and semantic skills are assessed in the elicited production and comprehension tasks of the LiSeDaZ (Schulz & Tracy, 2011) that provides norms for mono- and multilingual German-speaking children. The task administration was carried out presenting the pictures on a computer screen instead of in a picture book. From the sample of the child's spontaneous speech elicited through a picture story and semi-structured storytelling interview, the following domains are analysed manually according to the test manual: verb placement, subject-verb agreement, case-marking, word classes and comprehension of wh-questions.

Verb placement (LiSeDaZ). All utterances produced by the child that contain a verb are scored on a scale ranging from 1, indicating an utterance of a single word only, to 4, indicating the use of subordinate clause sentences with the verb in sentence-final position. At

least three utterances corresponding to a certain level need to be recorded in order to consider this level achieved. For example, one child in this sample produced the sentence "... wenn wir die Tonne umkippen." [if we turn over the container], correctly producing the verb "umkippen" [turn over] in the final position of the subordinate clause.

Subject-verb agreement (LiSeDaZ). All utterances containing a subject and a verb are counted and then assessed for correct subject-verb agreement. The ratio between the number of occurrences of correct subject-verb agreement and all the utterances consisting in a subject and a verb is calculated. The maximum score is 1.0. The ratio is then classified into a four-point scale. For example, one child in this sample produced the sentence "Was hast du in deinem Rucksack?" [What do you have in your backpack?], correctly inflecting the verb "hast" [have, 2nd person singular].

Word classes (LiSeDaZ). The amount of utterances per word class (prepositions, focus particles (i.e. "auch" [too], "nicht" [not], "nur" [only]), verbs, auxiliaries and conjunctions) is counted.

Case marking (LiSeDaZ). Some of the questions to be asked by the examiner are constructed such as they evoke case marked sentences, for example upon "Und wen kannst du hier noch sehen?" [And who else can you see here?], children are expected to respond with the accusative object "den Hund" [the_{ACC} dog]. Raw scores are converted into t-scores for further data processing.

Wh-questions (LiSeDaZ). This task is not directly part of the spontaneous speech produced by the child in the context of a picture story, but also here, a picture in the context of the story is presented to the child. The examiner first describes the picture, gives background information on the sentence not illustrated in the picture and then asks a question containing a wh-question word, for example "Der Hase frisst die Karotte. Lise hat sie von zuhause für den Hasen mitgenommen. Was frisst der Hase?" [The rabbit eats the carrot. Lise took it from home for the rabbit. What does the rabbit eat?]. All responses (even if semantically incorrect) that are in accordance with the wh-question word "was" [what] are assigned on point. Raw scores are converted into t-scores for further data processing.

CPM Raven's Progressive Matrices and Vocabulary Scales - Coloured Progressive Matrices von John Carlyle Raven (Bulheller & Häcker, 2001). Nonverbal intelligence was tested by the means of the *Raven's CPM* under the supervision of a trained psychologist if the scores of a standardized nonverbal intelligence test were not retrievable from clinical records and not older than one year. For the description of the test, see chapter 5.3.2.3.

All these standardised test results were evaluated according to the criteria in the respective manuals and the norm data provided in the latter. Furthermore, the raw scores were converted into percentages to facilitate comparison with the results of the experimental tests.

5.4.2.4 Procedure

Children were tested individually in the kindergartens or clinics where they were recruited. Whenever entrance to kindergartens or clinics was restricted during the pandemic, children were tested in a quiet room at their home. Each child was tested in two separate testing sessions. In the first session, children were administered the Mottier-Test (Kiese-Himmel & Risse, 2009), the PPVT-4 (Lenhard et al., 2015) the LiSeDaZ (Schulz & Tracy, 2011) and the Raven's CPM (Bulheller & Häcker, 2001), which lasted between 45 to 60 minutes. After a break of minimum one hour, in the second session, the German and Italian screening tasks were administered using the MuLiMi web-based screening platform accessed through the browser Firefox on a Lenovo laptop, model YOGA 720-15IKB under the Windows 10 Pro operating system. This lasted another 40 to 50 minutes. Caregivers filled in the pen-and-paper versions of the caregiver questionnaires QUIR-DC (Lorusso & Dolzadelli, 2016) and the exhaustive language background measuring the amount of in- and output in both languages spoken (Bloder et al., unpublished) in Italian or German. For children who were recruited in SLT clinics and kindergartens, their teachers/SLT filled in the pen-and-paper version of the teacher questionnaire. For $n = 14$ children the procedure was repeated in follow-up sessions (t2) that took place 8 to 14 months after the first time of testing (t1).

5.4.2.5 Risk score creation

Group assignment based on existing diagnoses or the absence/presence of subtest results below cut-off is referred to as risk level. Children who had already been diagnosed with and treated for DLD were assigned to the DLD group ($n = 7$). For the rest of the participants, children's scores in standardized tests were analysed regarding whether or not they had scored below the cut-off indicated in the respective test manual, i.e. a t-score of 40 for the LiSeDaZ (Schulz & Tracy, 2011) and the PPVT-4 (Lenhard et al., 2015) and of 43.3 for the Mottier-Test (norm data provided by Kiese-Himmel & Risse, 2009). Children without diagnoses but at least one subtest below cut-off were part of the at-risk group ($n = 17$). Children without diagnoses and with no results in the subtests below the cut-off indicated in the test manual were considered TD ($n = 13$). This three-level variable based on existing diagnoses or the absence/presence of subtest results below cut-off is referred to as risk level.

5.4.3 Results & Discussion

Children’s screening task performance was analysed with respect to a series of variables related to their DLD risk and risk level allocation and compared to standardized test performance. Table 8 gives an overview of standardized test performance in each of the three groups.

Table 8: Descriptive statistics for age and standardized test results in the three groups.

	TD ($n = 13$)	at risk ($n = 17$)	DLD ($n = 7$)
age (in months)	$M = 60.83, SD = 10.69$	$M = 58.47, SD = 8.46$	$M = 57.86, SD = 6.54$
Mottier (raw scores, max. 30)	$M = 15.08, SD = 3.62$	$M = 9.47, SD = 4.65$	$M = 3.14, SD = 3.89$
CPM (t-scores)	$M = 56.10, SD = 15.45$	$M = 44.90, SD = 9.23$	$M = 38.80, SD = 6.96$
LiSeDaZ verb placement (1-4)	$M = 4.00, SD = .00$	$M = 3.0, SD = 1.55$	$M = 1.75, SD = 1.50$
LiSeDaZ subject-verb agreement (1-4)	$M = 4.00, SD = .00$	$M = 2.73, SD = 1.68$	$M = .50, SD = .58$
PPVT-4 (raw scores)	$M = 101.71, SD = 20.43$	$M = 85.45, SD = 36.36$	$M = 42.00, SD = 33.73$

5.4.3.1 Comparison of performance in screening task and risk level

Performance (repetition accuracy in %) in all the single NWRT subcategories and compound scores was significantly associated with the risk level ($n = 37$, ρ s ranging from $-.559$ to $-.739$, p s $< .001$). Neither for the Italian clitic pronoun judgement screening task ($n = 17$) nor for the German case matching screening task ($n = 37$) there was a significant association with their risk level (p s $> .05$). However, for the case matching tasks notable non-significant associations emerged ($\rho = -.294$, $p = .078$). Screening task performance related to subject-verb agreement judgement and matching instead in both language versions was significantly correlated with the child’s risk level ($n = 37$, Italian: $\rho = -.364$, $p = .027$; German: $\rho = -.385$, $p = .019$, for the German and Italian version of this task, this effect was no longer significant when inserting the children’s performance in the Raven as control variable: $p > .05$). Performance (accuracy in %) in the German and Italian DNWL subtests ($n = 17$) was not significantly associated with the risk level (p s $> .05$). Children’s performance (accuracy in %) in the German CLT verb and noun comprehension subtests as well as the compound score deriving from the two single subtests were significantly correlated with the risk level ($n = 37$, ρ s ranging from $-.570$ to $-.600$, p s $< .001$). None of the scores in the Italian version of the CLT word comprehension tasks were significantly associated with the risk level. Figure 28 illustrates that for the morphosyntactic processing screening tasks (the Italian clitic pronoun

judgement task was not included since due to time constraints, it was administered only to a small subset of children who participated in this study) in both languages as well as in the German, but not the Italian CLT subtests, the children with a diagnosis of DLD on average scored lowest. Furthermore, visual inspection of figure 28 suggests that the children with a DLD diagnosis (with the exception of the Italian CLT subtests) consistently scored worse compared to the children without DLD diagnoses who scored below cut-off in at least one of the standardized tests.

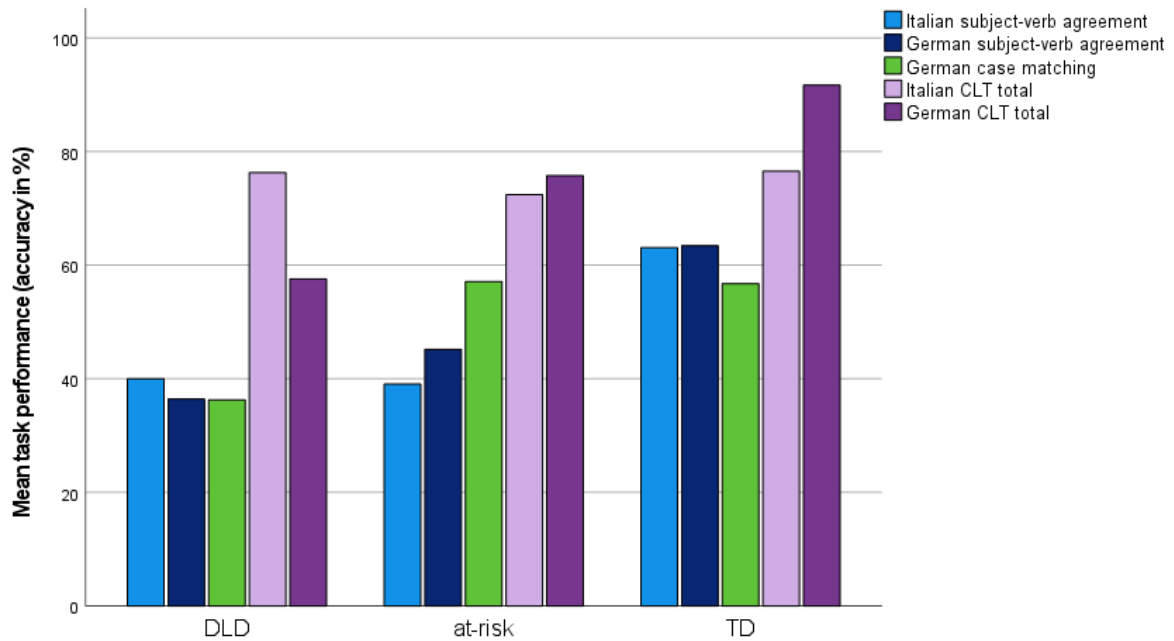


Figure 28: Mean task performance for screening tasks in both languages according to the child participants' DLD risk levels.

Similar to the pattern described for the morphosyntactic processing screening tasks and the CLT subtests in both languages, children with a diagnosis of DLD on average consistently obtained the lowest scores in the NWRT as well. Upon visual inspection of the graphs representing repetition performance in the various subcategories, for at risk and TD children, the repetition of LS NWs seems to be less challenging than the repetition of NLS NWs. Children with a diagnosis of DLD instead on average appear to perform better in the repetition of NLS opposed to LS NWs (see figure 29).

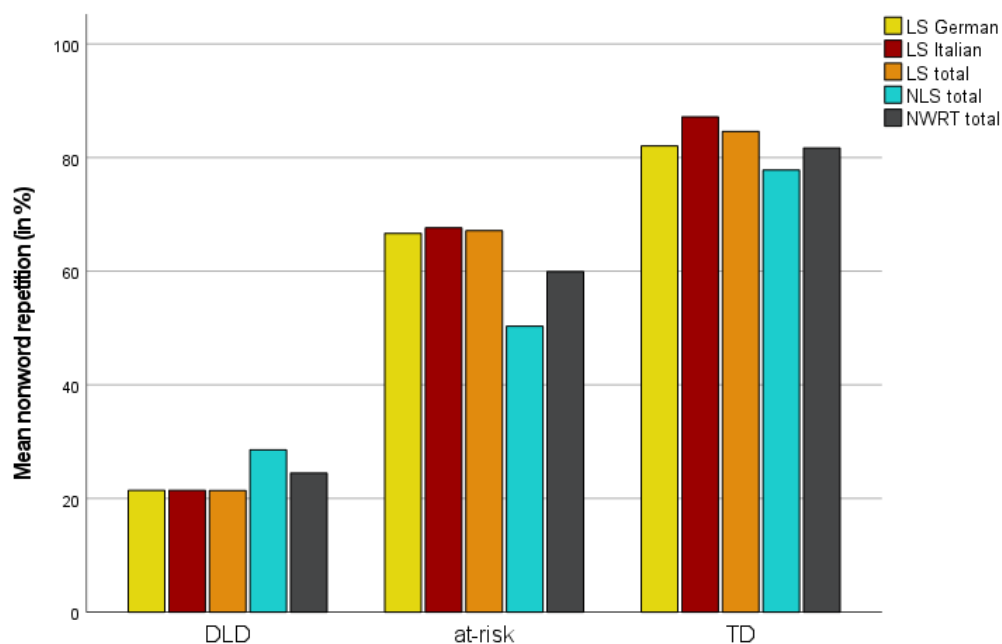


Figure 29: Mean repetition performance for NWs of different subcategories according to the child participants' DLD risk levels.

5.4.3.2 Comparison of performance in the screening tasks and language background

Since one of the child participants' caregivers did not return the questionnaire, this participant had to be excluded from these analyses. Furthermore, since for one of the child participants there was conflicting information regarding the amount of Italian output, also this participant was excluded from the analyses in which Italian output was concerned.

Since language in- and output scores are represented in a continuous variable with no gross violations of normality, Pearson's correlations were run. All the CLT subtest and total scores per language were significantly associated with the children's language in- and output in German and Italian respectively as measured by the language background questionnaire (Bloder et al., unpublished), with r_s ranging from .441 to .609, $p_s < .001$ (n_s either 35 or 36). This effect was no longer found to be significant for Italian nouns when performance in the Raven's CPM was inserted as control variable ($p > .05$). However, with the exception of accuracy (in %) in the German case matching screening tasks which was significantly and negatively associated with the amount of children's Italian output ($n = 35$, $r = -.369$, $p = .029$), the screening tasks on morphosyntactic processing ($n = 16$ for Italian clitic pronoun judgement, $n = 35/36$ for subject-verb agreement judgement and matching in both languages separately) and none of the repetition performance (accuracy in %) in the single and compound categories of the NWRT were significantly associated with the children's language in- and output for Italian and German ($p_s > .05$). Regarding the children's performance (accuracy in %) in the DNWL subtests in both languages, only the children's performance in the test phase of the

German DNWL version was significantly correlated with the amount of German input ($n = 15$, $r = .595$, $p = 0.19$).

5.4.3.3 *Comparison of performance in the screening tasks and standardized tests*

Significant associations between the children's NWRT performance in the single NW categories and compound scores emerged for both the Mottier-test's raw ($n = 37$, r_s ranging from .505 to .788, $p_s < .001$) and t-scores ($n = 37$, r_s ranging from .621 to .810, $p_s < .001$).

Since NW repetition skills are known to be associated with vocabulary knowledge and complex syntax (see chapter 2.4.1.2), the NWRT performances in single and compound categories were also compared to the performances in the respective standardized tests. For the assessment of associations between NWRT performance and vocabulary knowledge, first the PPVT-4's results were used. Significant associations between the children's NWRT performance in the single NW categories and compound scores emerged for both the PPVT-4's raw ($n = 37$, r_s ranging from .397 to .600, $p_s < .015$) and t-scores ($n = 37$, r_s ranging from .328 to .483, $p_s < .047$). When inserting the children's performance in the Raven's CPM, this effect remained stable for the German LS NWs and the LS compound score only. Since the NWs were constructed such, that they are very specific or non-specific to one of the languages spoken by the children as rated by adult native speakers, also the non-standardized CLT verb and noun comprehension subtest results were also compared to the children's repetition performances of LS NWs. Repetition performance of LS Italian NWs was significantly associated with both German ($n = 37$, $r = .348$, $p = .035$) and Italian noun comprehension ($n = 37$, $r = .371$, $p = .024$), but not verb comprehension in any of the languages ($p_s > .05$). Repetition performance of LS German NWs instead was significantly correlated with the results of the German CLT subtests only, namely German verb ($n = 37$, $r = .469$, $p = .003$) and noun comprehension ($n = 37$, $r = .377$, $p = .021$). Interestingly, also the repetition performance of NWs that were rated as NLS was significantly associated with all CLT subtests in both languages ($n = 37$, r_s ranging from .325 to .438, $p_s < .05$ except for German noun comprehension ($p > .05$)). Accordingly, significant associations emerged comparing the children's NW repetition performance total score incorporating all LS and NLS NWs from both languages to all of the single subtest results of both verb and noun comprehension for Italian and German ($n = 37$, r_s ranging from .341 to .372, $p_s > .039$). However, when inserting the children's performance in the Raven's CPM as control variable, no significant associations between word comprehension in any of the subtests and NW repetition performance in any of the NW categories remained significant ($p_s > .05$).

For the assessment of associations between NWRT performance, the indices representing children's abilities in the field of complex syntax namely the LiSeDaZ subtests subject-verb agreement, verb placement, conjunctions and wh-questions were used as well. For the LiSeDaZ subtests for which the results are interpreted on a 4-point scale and whenever raw scores for a certain subtest – especially for the word class subtests – do not represent a wide range of values, Spearman correlations comparing the latter to screening task results were calculated. While the amount of conjunctions produced by the children was not significantly associated with any of the NWRT single and compound scores ($p_s > .05$), several significant associations emerged with the remaining subtests selected. Children's repetition performance of LS German NWs (accuracy in %) was significantly correlated with the children's performance levels regarding verb placement, subject-verb agreement (4-point scale), case marking and comprehension of wh-questions (raw scores) as measured in the LiSeDaZ with rho_s ranging from .328 to .587, $p_s < .048$ ($n = 37$). Similarly, significant associations between the children's repetition performance of LS Italian NWs emerged for verb placement, subject-verb agreement and case marking as measured by the LiSeDaZ with rho_s ranging from .348 to .428, $p_s < .035$ ($n = 37$). Also, the children's performance in the comprehension of wh-questions was associated, but not significantly ($n = 37$, $rho = .313$, $p = .059$). Accordingly, the LS total compound score was significantly associated with all of the LiSeDaZ subtest results mentioned above with rho_s ranging from .371 to .512, $p_s < .024$ ($n = 37$). Performance levels in the LiSeDaZ subject-verb agreement ($n = 37$, $rho = .472$, $p = .003$), but not verb placement, case marking and wh-question comprehension, were significantly associated with repetition accuracy of NLS NWs ($p_s > .05$). Subsequently, significant correlations were found for the children's NWRT total score and the LiSeDaZ verb placement ($n = 37$, $rho = .360$, $p = .029$), subject-verb agreement ($n = 37$, $rho = .494$, $p = .002$) and wh-question ($n = 37$, $rho = .477$, $p = .003$) subtest results. When inserting the children's performance in the Raven's CPM, the associations between performance in the NWRTs and the scores obtained in the LiSeDaZ subtests (with the exception of scores obtained in the LiSeDaZ wh-questions subtest) were no longer significant ($p_s > .05$).

Both German screening tasks assessing children's morphosyntactic processing skills were significantly associated with the LiSeDaZ verb placement ($n = 37$, subject-verb agreement: $rho = .373$, $p = .023$; case matching: $rho = .386$, $p = .018$) and the LiSeDaZ subject-verb agreement score ($n = 37$, subject-verb agreement: $rho = .540$, $p < .001$; case matching: $rho = .369$, $p = .025$). Both tasks were also significantly correlated with the raw scores obtained in the wh-question comprehension subtest ($n = 37$, subject-verb agreement: $rho = .518$, $p < .001$; this association was no longer significant when inserting the children's performance

in the Raven CPM as control variable: $p > .05$; case matching: $\rho = .443$, $p = .006$). Furthermore, a non-significant association between accuracy (in %) in the case matching screening task and the amount of conjunctions produced emerged ($n = 37$, $\rho = .303$, $p = .069$). Since the range of raw scores for the word class conjunction subtest ranged between 0 and 9, additionally Pearson's correlations were calculated, resulting in a significant association of $r = .364$, $p = .027$ ($n = 37$). None of the Italian screening tasks assessing children's morpho-syntactic processing skills were significantly correlated with the LiSeDaZ subtest results. Also, performance levels in both versions of the DNWL subtests were not significantly associated with the performance in any of the standardized tests ($p_s > .05$).

As highlighted in chapter 5.4.3.2, significant correlations were found between the children's performance in the single subtests of the CLTs and children's language in- and output. The latter is also significantly associated with the children's raw and t-scores obtained in the German standardized PPVT-4 vocabulary test ($n = 36$, r_s ranging from .466 to .569, $p_s < .004$). Accordingly, none of the Italian CLT single and compound scores was significantly correlated with the raw and t-scores of the PPVT-4 ($p_s > .05$). Instead, all the German CLT single subtest and the total scores were significantly associated with the raw and the t-scores obtained in the PPVT-4 with r_s ranging from .689 to .841, $p_s < .001$ ($n = 37$).

5.4.3.4 Comparison of performance in the screening's and teacher questionnaires

Since the responses to the questionnaire are hypothesized to be different for SLT vs. kindergarten teacher respondents, responses from SLTs ($n = 3$) and kindergarten teachers ($n = 20$) were not merged for the analyses. For one of the children with an existing DLD diagnosis, the kindergarten teacher filled in the questionnaire. In a first step, the kindergarten teachers' responses were compared to the risk level incorporating the information on an existing diagnosis and performance above or below cut-off in any of the standardized tests. While kindergarten teachers' responses to all of the single productive phonology, morphosyntax and lexicon, the productive compound as well as the total compound score were significantly correlated with the children's risk level ($n = 19$, with ρ_s ranging from .470 to .593, $p_s < .042$), neither the receptive single nor the receptive total scores were significantly associated with the child's risk level ($p_s > .05$). Based on these findings, these scores were further compared to the children's performance in the screening tasks.

Children's repetition accuracy (in %) of the LS German NWs was somehow associated with the kindergarten teacher questionnaire's total compound score, but not significantly ($n = 20$, $r = -.423$, $p = .063$). This association was significant when the children's performance in

the Raven's CPM was inserted as a control variable ($n = 20$, $r = -.526$, $p = .012$). Also repetition accuracy of the LS Italian NWs was almost significantly associated with the total scores of kindergarten teacher questionnaires ($n = 19$, $r = -.429$, $p = .059$). Again, this association became significant when the children's performance in the Raven's CPM was inserted as control variable ($n = 19$, $r = -.515$, $p = .014$). For repetition accuracy of these NWs, significant associations were found with teachers' judgement of children's productive phonology skills ($n = 20$, $r = -.477$, $p = .033$) and the productive compound score ($n = 19$, $r = -.459$, $p = .042$). Subsequently, also the accuracy score for both German and Italian LS NWs combined was significantly correlated with productive ($n = 20$, $r = -.456$, $p = .043$) and total compound score ($n = 19$, $r = -.457$, $p = .043$) of the kindergarten teacher questionnaires. Since also for repetition accuracy of the NLS NWs and kindergarten teachers' judgements of the children's phonological productive skills ($n = 19$, $r = -.521$, $p = .018$) significant associations emerged, accordingly also repetition accuracy total score incorporating NWs of all categories was significantly associated with the kindergarten teachers' judgements of children's phonological productive skills ($n = 20$, $r = -.500$, $p_s = .025$) and the production compound score ($n = 20$, $r = -.449$, $p = .047$).

Comparing the children's performance in the subtests of the DNWL screening tasks, only one significant and positive association emerged between the kindergarten teachers' judgements of productive phonological skills and performance in the consolidation phase of the Italian version of this task ($n = 10$, $r = .667$, $p = .035$) indicating that the worse the child's productive phonology skills were judged, the better they scored in this subtest.

None of the Italian or German screening tasks assessing children's morphosyntactic processing skills correlated significantly with the teachers' judgements of children's language skills, neither for the single nor the compound scores ($p_s > .05$).

Similarly, none of the subtest and compound scores (accuracy in %) of the Italian version of the CLTs was significantly associated with the responses by the kindergarten teachers ($p_s > .05$). For the German versions instead, the kindergarten teachers' judgements of children's competences regarding morphosyntax and lexicon productive skills were associated with German noun and verb comprehension and subsequently also with the German CLT total score ($n = 19$, with r_s ranging from $-.521$ to $-.692$, $p_s < .015$). The production compound score was significantly correlated with German noun comprehension ($n = 19$, $r = -.521$, $p = .018$). Accuracy (in %) in the German noun comprehension task was also significantly associated with the compound score incorporating all the responses by kindergarten teachers to the questionnaire ($n = 19$, $r = -.464$, $p = .039$).

5.4.3.5 Comparison of performance in the screenings and caregiver questionnaires

Since one of the child participants' caregivers did not return the questionnaire, he/she had to be excluded from the analyses. Before comparing the children's performance in the screening tasks to the responses given by the caregivers in the QUIR-DC to the screening tasks, the QUIR-DC's appropriateness for this comparison was assessed comparing it to the children's risk level. The children's risk level was associated with all compound scores, but not with the PS ($n = 36$, $\rho = -.177$, $p = .303$). Still, all QUIR-DC compound scores were included in the further analyses comparing the children's screening task performances to the compound scores deriving from the caregivers' responses to the QUIR-DC.

All the single and compound scores representing repetition accuracy in the NWRT correlated significantly with the GS ($n = 36$, with r_s ranging from .442 to .522, $p_s < .007$) and the RS ($n = 36$, with r_s ranging from -.403 to -.453, $p_s < .016$) deriving from the caregivers' responses to the QUIR-DC. Due to the nature of the phonological as well as the FIGS and FIRS, comparing these scores to the screening tasks, Spearman-rho's were calculated. All scores except the one for repetition performance of NLS NWs ($n = 36$, $\rho = .305$, $p = .070$) were significantly associated with the PS deriving from the caregivers' responses to the QUIR-DC with ρ_s ranging from .331 to .395, $p_s < .049$ ($n = 36$). While the FIGS was significantly correlated with the repetition of LS German NWs ($n = 36$, $\rho = .364$, $p = .029$), it was non-significantly associated with the repetition score of LS Italian NWs ($n = 36$, $\rho = .313$, $p = .063$) and subsequently also not with the LS compound score ($n = 36$, $\rho = .388$, $p = .070$). When inserting the children's performance in the Raven's CPM as control variable, the association between the FIGS and the LS compound score was significant ($n = 36$, $r = .357$, $p = .035$). The FIRS instead was significantly and negatively associated with the NLS ($n = 36$, $\rho = -.387$, $p = .020$) and total compound scores ($n = 36$, $\rho = -.382$, $p = .021$), but these effects were no longer present when inserting the children's performance in the Raven's CPM as control variable ($p_s > .05$).

Comparing the children's performance in the screening tasks on morphosyntactic processing in both languages separately to the caregivers' responses represented in the QUIR-DC GS and RS, only one significant association emerged comparing the QUIR-DC GS to the accuracy (in %) in the German case matching screening task ($n = 36$, $r = .340$, $p = .042$). Furthermore, associations between the children's performance in both the German subject-verb agreement as well as the case matching screening tasks and both input variables reached significance: significant associations were found for the FIGS and accuracy (in %) in the subject-verb agreement ($n = 36$, $\rho = .440$, $p = .007$) as well as the case matching ($n =$

36, $\rho = .436$, $p = .008$) screening tasks. Also the FIRS was significantly associated with accuracy (in %) in the subject-verb agreement ($n = 36$, $\rho = -.449$, $p = .006$, this association was no longer significant when inserting children's performance in the Raven's CPM as control variable: $p > .05$) and the case matching ($n = 36$, $\rho = -.355$, $p = .034$) screening tasks. None of the Italian screening task results were significantly correlated with these scores ($p_s > .05$).

Due to the small sample size, the results in the subtests of the DNWL screening task in both languages were compared to the QUIR-DC compound scores calculating Spearman's correlations. Significant associations emerged comparing the QUIR-DC RS to the children's performances in the Italian ($n = 14$, $\rho = .567$, $p = .034$) and the German ($n = 15$, $\rho = -.572$, $p = .026$) version of the DNWL's consolidation phase. For accuracy (in %) in the German consolidation phase another significant association with the QUIR-DC's PS was found ($n = 15$, $\rho = .645$, $p = .009$).

While children's performance in none of the Italian CLT subtests was significantly associated with any of the GS and RS compound scores deriving from the caregivers' responses to the QUIR-DC questionnaire, significant associations were found for the German subtests and the compound score. Namely, the QUIR-DC GS correlated significantly with German noun ($n = 36$, $r = .524$, $p = .001$), verb ($n = 36$, $r = .502$, $p = .002$) comprehension and subsequently the compound score incorporating both latter scores ($n = 36$, $r = .561$, $p = .002$). The PS was significantly associated with the performance of Italian noun ($n = 36$, $\rho = .388$, $p = .019$) and verb ($n = 36$, $\rho = .337$, $p = .044$) comprehension and subsequently also the compound score deriving from the latter ($n = 36$, $\rho = .337$, $p = .045$). The PS was also – though not significantly – associated with German noun comprehension ($n = 36$, $\rho = .320$, $p = .057$). Concerning the FIGS and FIRS, again only significant associations with the German CLT scores emerged, namely family global input score with German noun ($n = 36$, $\rho = .541$, $p < .001$) and verb ($n = 36$, $\rho = .630$, $p < .001$) comprehension and subsequently also the compound score incorporating the latter ($n = 36$, $\rho = .649$, $p < .001$) as well as the FIRS with the German word comprehension total score ($n = 36$, $\rho = -.373$, $p = .025$).

5.4.3.6 Comparison of performance in the screening tasks

Furthermore, screening task performance levels were compared intra- and crosslinguistically. Both repetition accuracy (in %) of German and Italian LS NWs was significantly associated with the LS, NLS and total NWRT compound scores with r_s ranging from .705 to .947, $p_s < .001$ ($n = 37$). Neither significant associations emerged comparing the Italian clitic pronoun judgement tasks to the Italian subject-verb agreement task ($n = 17$, $p > .05$), nor comparing

the German case matching task to the German subject-verb agreement tasks ($n = 37, p > .05$). Regarding the CLT subtests instead, for both Italian and German performance in the CLT subtests, noun comprehension was significantly correlated with performance in the CLT verb comprehension subtest in the same language ($n = 37$, Italian: $r = .834, p < .001$; German: $r = .675, p < .001$). Performance levels in the single phases of the DNWL screening tasks within languages were not found to be significantly associated ($n = 14, p > .05$).

In a next step, task performance in tasks assessing similar abilities were compared crosslinguistically. Repetition accuracy (in %) of LS German NWs correlated significantly with repetition accuracy of LS Italian NWs ($n = 37, r = .791, p = .001$). While the children's performances (accuracy in %) in the Italian and German version of the subject-verb agreement tasks were significantly associated with each other ($n = 37, r = .604, p < .001$), none of them was significantly correlated with the children's performance in the Italian clitic pronoun judgement or the German case matching task's performance ($p_s > .05$). The children's performance in the Italian CLT noun and verb comprehension subtests were not significantly associated with the performance in the German versions of these tasks ($p_s > .05$). Performance (accuracy in %) in the testing phase of the German version of the DNWL screening task was found to be significantly correlated with performance in the test phase of the Italian version of the DNWL screening task ($n = 15, rho = .565, p = .035$).

Finally, to assess the appropriateness of the diagnoses of the children who had already been diagnosed with DLD in the light of a potential misdiagnoses, their performances in comparable screening tasks in both language versions was compared. Repetition accuracy (in %) for both LS German and LS Italian NWs was associated with the LS, NLS and total compound scores. Repetition accuracy of Italian LS NWs was significantly ($n = 7, rho_s$ between $.791$ and $.898, p_s < .034$) and repetition accuracy of German LS NWs was almost significantly associated ($n = 7, rho_s$ between $.704$ and $.824, p_s < .078$) with the LS, NLS and total compound scores. However, the repetition accuracy of LS Italian and LS German NWs were not significantly correlated with each other ($n = 7, rho = .574, p = .178$). While the cross-linguistic associations had emerged comparing the whole group's performance levels in the German and Italian version of the subject-verb agreement task, analysing the NW repetition performance (accuracy in %) of the children with a DLD diagnosis only, the association was no longer significant ($n = 7, rho = .991, p < .001$). For the CLTs, the Italian noun ($n = 7, rho = .680, p = .093$) and verb ($n = 7, rho = .739, p = .058$) comprehension subtest as well as the LS total compound scores ($n = 7, rho = .937, p = .002$) were (significantly) associated with the score of the same subtest and compound score in German. Since only two DLD children

were administered the DNWL screening tasks at t1, no language comparisons for the DNWL screening tasks in the DLD group only were calculated.

5.4.3.7 Comparison of performance in the screening tasks at t1 and at t2

Since the DNWL and clitic pronoun judgements tasks were not administered at t1 and t2, only the results from the CLT, NWRT and subject-verb agreement tasks were analysed. Of the $n = 14$ children participating in the follow-up study, $n = 5$ had been diagnosed with DLD, $n = 5$ are considered at risk due to their performance below cut-off in the standardized test and $n = 3$ were TD children.

The follow-up study showed that the child participants' performance level at t1 across tasks was significantly associated with the performance in the same task at t2 (see table 9).

Table 9: Associations between screening task performance at t1 and screening task performance at t2 in the same task ($n = 14$).

Language	Screening task type	Correlation coefficient
German	Subject-verb agreement	$\rho = .641, p = .013$
	Case matching	$\rho = .508, p = .064$
	Word comprehension (CLTs)	$\rho = .937, p < .001$
Italian	Subject-verb agreement	$\rho = .649, p = .012$
	Word comprehension (CLTs)	$\rho = .937, p < .001$
	NWRT total	$\rho = .559, p = .038$

Based on these results, for the tasks that consist of different items, analyses investigating the correlations in each different item category were calculated. Significant correlations were found between performance level at t1 in a certain NWRT category and performance in the same NW category at t2 (see table 10). Associations across NWRT categories were found for most, but not all tasks. Fewest associations were found for NLS NW repetition performance.

Table 10: Correlations between children's MuLiMi NW repetition performance at t1 and t2 (n = 14). Correlations between performance levels in the same NW category's scores in bold.

		t2				
		LS-Italian	LS-German	LS-total	NLS-total	NWRT total
t1	LS-Italian	rho = .648, p = .012	rho = .548, p = .042	rho = .637 p = .014	rho = .590 p = .026	rho = .656 p = .011
	LS-German	rho = .591, p = .026	rho = .536, p = .048	rho = .615 p = .019	rho = .426 p > .05	rho = .502 p > .05
	LS-total	rho = .634, p = .015	rho = .551 p = .041	rho = .644 p = .013	rho = .469 p > .05	rho = .566 p = .035
	NLS-total	rho = .605, p = .022	rho = .376 p > .05	rho = .525 p > .05	rho = .574 p = .032	rho = .545 p = .044
	NWRT total	rho = .602, p = .023	rho = .443 p > .05	rho = .556 p = .039	rho = .562 p = .037	rho = .559 p = .038

Furthermore, performance level at t1 in a certain CLT subtest was significantly associated with the performance in the same subtest at t2 (see table 11). While both Italian noun and verb comprehension at t1 correlated significantly with both Italian noun and verb comprehension at t2, none of them was significantly associated with German verb or noun comprehension at t2. The same pattern can be observed comparing the German verb and noun comprehension at t1 to t2 Italian noun and verb comprehension ($p_s > .05$).

Table 11: Correlations between children's CLT subtest performance at t1 and t2 (n = 14). Correlations between performance levels in the same subtest's scores in bold.

	Italian nouns t2	Italian verbs t2	German nouns t2	German verbs t2
Italian nouns t1	rho = .852, p < .001	rho = .747, p = .003	rho = .020, p > .05	rho = .301, p > .05
Italian verbs t1	rho = .900, p < .001	rho = .896, p < .001	rho = -.186, p > .05	rho = .064, p > .05
German nouns t1	rho = .223, p > .05	rho = -.099, p > .05	rho = .781, p < .001	rho = .713, p = .004
German verbs t1	rho = .039, p > .05	rho = -.286, p > .05	rho = .886, p < .001	rho = .923, p < .001

Besides directly comparing screening task performance at t1 to screening task performance at t2 also the degree of difference between the results at t2 compared to t1 (t1 performance minus t2 performance = difference score) were compared to t1 performance.

Here, significant associations were found between t1 screening task performance and the difference score for the German case-matching task (n = 14, rho = -.581, p = .029), the NWRT total (n = 14, rho = -.543, p = .045) as well as the NLS total score (n = 14, rho = -.615, p = .019) and the German CLT total score (n = 14, rho = -.570, p = .033). In these screening

tasks, the lower the performance at t1, the more improvement children showed from t1 to t2. This pattern is also illustrated in figure 30 showing that children with high performance levels at t1 show less improvement from t1 to t2 and that the highest improvement is yielded by a child holding a diagnosis with the group's poorest performance at t1.

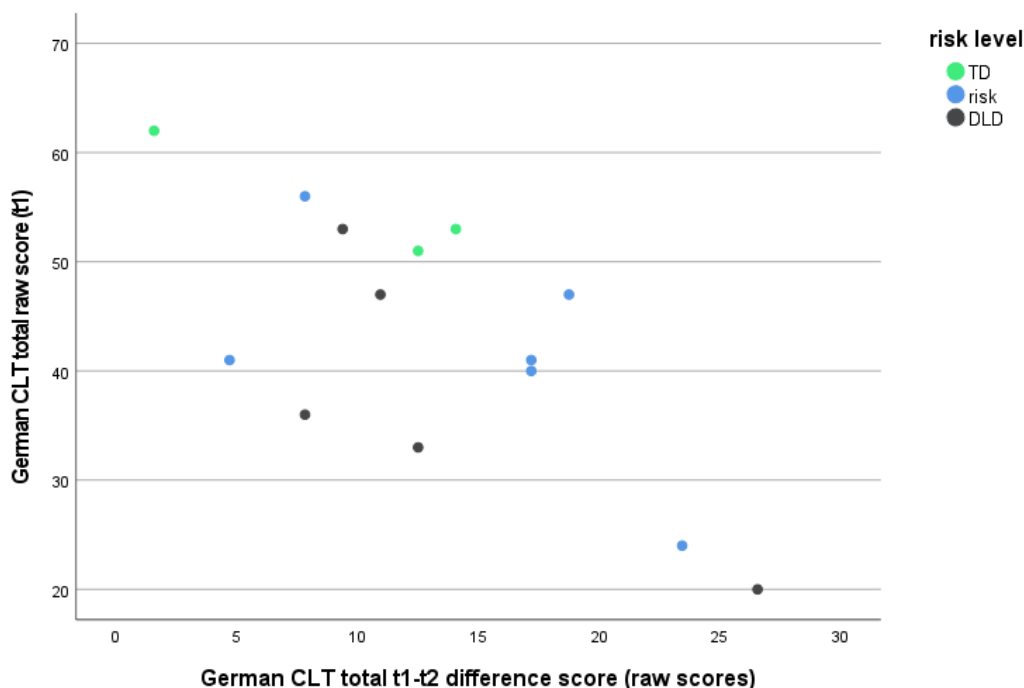


Figure 30: Scatterplot of performance in the German word comprehension CLT compound score at t1 (y-axis) against the difference regarding performance in the same task between t1 and t2 (x-axis) according to the child participants' risk levels.

Post-hoc analyses investigating the variables that determine improvement from t1 to t2 revealed that overall, difference scores were not associated with age, IQ or language dominance across tasks ($p_s > .05$) with some exceptions: The amount of Italian output at t1 was significantly and negatively associated with improvement in the German LS NW repetition performance ($n = 13$, $\rho = -.553$, $p = .050$) suggesting that the lower the input in Italian the more the child improves in the German LS NW repetition. Furthermore, age was significantly associated with the improvement in the repetition performance belonging to the categories LS German ($n = 14$, $\rho = -.904$, $p < .001$), LS total ($n = 14$, $\rho = -.626$, $p = .017$) and the total compound score ($n = 14$, $\rho = -.562$, $p = .037$) as well as with the Italian CLT total score ($n = 14$, $\rho = -.557$, $p = .048$) suggesting that in these tasks, the younger the children are the more they improve from t1 to t2, illustrated in figure 31. Across risk levels, the youngest children improved most from t1 to t2. This effect is most pronounced in the DLD group (red dots).

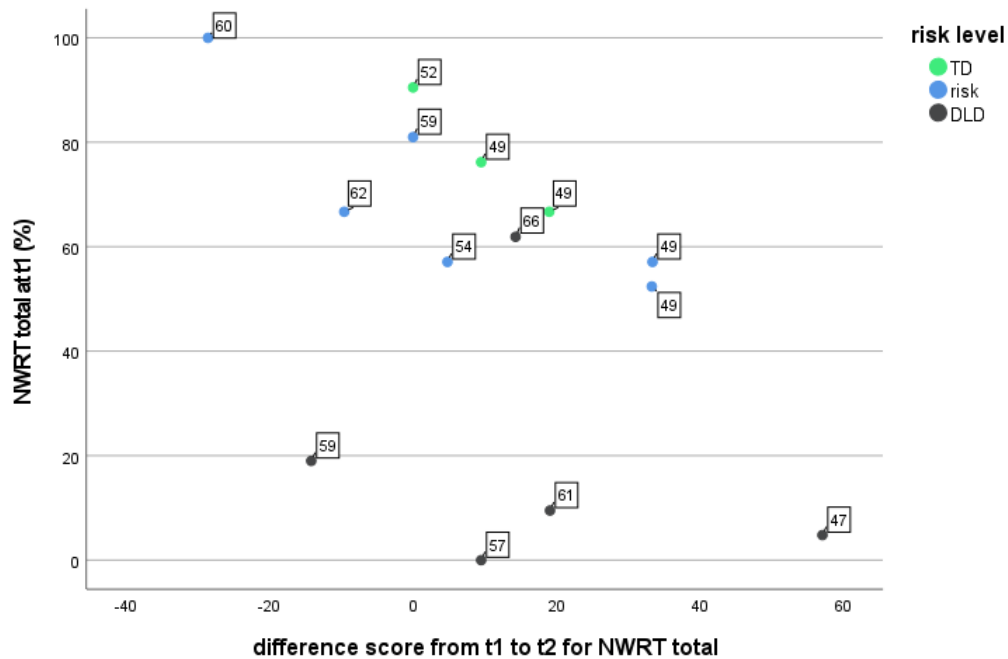


Figure 31: Scatterplot displaying repetition performance of German NWs at t1 (y-axis) against difference in performance in the same task between t1 and t2 (x-axis) according to the child participants' risk level. Labels indicate the child's age in months at t1.

5.4.3.8 Interim discussion

The research questions on concurrent and discriminant validity can be positively answered based on the significant associations between performance in the screening tasks and the variables on DLD risk, standardized test performance as well as caregiver and kindergarten teacher questionnaires that emerged (see table 12 for an overview).

Table 12: Overview of significant associations between screening task performance (accuracy in %) risk level, standardized tests as well as kindergarten teachers and caregiver questionnaires.

Screening task (accuracy in %)	risk level (TD, risk, DLD)	standardized tests (raw scores)	teacher questionnaire	QUIR-DC
NWRT total	$n = 36,$ $\rho = -.739,$ $p > .001$	Mottier-Test $n = 37, r = .752,$ $p < .001$	productive compound score $n = 19, r = -.459,$ $p = .042$	global score $n = 36, r = .453,$ $p = .006$
Subject-verb agreement (IT)	$n = 37,$ $\rho = -.364,$ $p = .027$	n.a.	n.s.	n.s.
Subject-verb agreement (GER)	$n = 37,$ $\rho = -.385,$ $p = .019$	LiSeDaZ (subject-verb agreement) $n = 37, \rho = .540,$ $p < .001$	n.s.	family input risk score $n = 36,$ $\rho = -.440,$ $p = .007$
Clitic pronoun (IT)	n.s.	n.a.	n.s.	n.s.
Case marking (GER)	n.s.	LiSeDaZ (subject-verb agreement) $n = 37, \rho = .386,$ $p = .018$	n.s.	global score $n = 36, r = .340,$ $p = .042$
CLT (IT) noun comprehension	n.s.	n.a.	n.s.	phonological score $n = 36, \rho = .388,$ $p = .019$
CLT (GER) noun comprehension	$n = 37,$ $\rho = -.570,$ $p < .001$	PPVT-4 $n = 37, r = .709,$ $p < .001$	total score $n = 19, r = -.464,$ $p = .039$	global score $n = 36, r = .524,$ $p = .001$
CLT (IT) verb comprehension	n.s.	n.a.	n.s.	phonological score $n = 36, \rho = .337$ $p = .044$
CLT (GER) verb comprehension	$n = 37,$ $\rho = -.574,$ $p < .001$	PPVT-4 $n = 37, r = .821,$ $p < .001$	productive lexicon $n = 19, r = -.521,$ $p = .022$	risk score $n = 36, r = .502,$ $p = .002$
DNWL (IT) consolidation	n.s.	n.s.	productive phonology $n = 10, \rho = .667,$ $p = .035$	risk score $n = 14, \rho = .567$ $p = .034.$
DNWL (IT) testing	n.s.	n.s.	n.s.	n.s.
DNWL (GER) consolidation	n.s.	n.s.	n.s.	risk score $n = 15, \rho = -.572$ $p = .026$
DNWL (GER) testing	n.s.	n.s.	n.s.	n.s.

Similar to the interim discussion of the study on the Spanish-Italian DLD screening (see chapter 5.3.3.7), this study assessing the potential of the MuLiMi DLD screening for Italian-German-speaking children living in Germany also indicates that in general, computerized

screenings automatically assessing children's language performance in both languages spoken do contribute to the identification of DLD. This is sustained by the significant associations that emerged when comparing children's performance in screening tasks to their risk levels determined by performance in the standardized tests (above or below clinical cut-offs) and the presence vs. absence of a DLD diagnoses (Hypothesis 1). Though, due to limitations in sample size, discriminant validity could not be directly assessed through statistical tests, the general trend as observed by visual inspections of graphs representing the children's screening task performances grouped by risk level indicates that children with a higher risk score (presence of a diagnosis based on information by SLTs' and performance in standardized tests below cut-off) underperform in comparison to the children with no and lower risk scores in screening tasks in all languages spoken. Furthermore, the presence of a series of significant associations between results of standardized tests and performance in the screening tasks declared to assess the same linguistic skills indicate concurrent validity of the screening (Hypothesis 2). Moreover, it was shown for several screening tasks that performance on different screening tasks assessing similar skills in the same language are associated with each other (Hypothesis 3). More specifically, performance in noun comprehension was significantly associated with performance in verb comprehension in both languages (Hypothesis 3a). However, performance in the German case matching screening task instead did not correlate significantly with performance in the German subject-verb agreement task (Hypothesis 3b). Similarly, performance in the Italian subject-verb agreement task was not significantly associated with the of clitic pronoun judgement task (Hypothesis 3c). For a series of screening tasks, it was found that performances in screening tasks assessing the same linguistic area in the two different languages are associated with each other (Hypothesis 4). More specifically, performance in the repetition of LS German NWs was significantly correlated with repetition performance of LS Italian NWs (Hypothesis 4a). Performance in German word comprehension instead was not significantly associated with Italian word comprehension for neither the single noun and verb comprehension nor the LS compound scores (Hypothesis 4b). The children's performance in the Italian version of the subject-verb agreement screening task was significantly correlated with performance in the German version of this screening task (Hypothesis 4c). Associations within the same language however were only observed in the NWRT and CLT screening tasks when analysing the data separately in the small sample of seven children who already have been diagnosed with DLD (Hypothesis 5). Besides that, several comparisons indicate predictive value of the screening (Hypothesis 6). More precisely, performance in a series of screening tasks at t1 correlated significantly with perfor-

mance in the same task at t2 (Hypothesis 6a). Also, performance at t1 is significantly associated with improvement from t1 to t2 (Hypothesis 6b). Furthermore, not IQ but in certain screening tasks, age and language dominance or more specifically language exposure and production patterns were shown to influence the improvement from t1 to t2 (Hypothesis 6c).

5.5 Bilingual, computerized DD Screening for Italian-speaking children living in Germany

As mentioned in chapter 2.3.2 and 5.4, Italy is among the top-five countries of origin of foreigners living in Germany. Many years after the beginning of frequent migration from Italy to Germany ever since guest worker programs were established in the 1950s, Italian-German bilingual public as well as state (primary) schools were established in bigger cities. Accordingly, an Italian-German DD screening was constructed and administered to Italian-speaking children living in Germany with, at risk of and without DD in the presence of the examiner. A modified version of this chapter has been published in the German article (conference proceedings) “Computergestütztes, bilinguales Lese-Screening mit der Screening-Plattform MuLiMi” (Computerized, bilingual reading screening using the screening platform MuLiMi) published in the German open access e-journal “Sprachtherapie aktuell: Forschung – Wissen – Transfer”, ed. 2: Schwerpunktthema: Perspektiven auf Beeinträchtigungen der Schriftsprache: e2021-40; doi: 10.14620/stadbs210740 (Eikerling & Lorusso, 2021).

5.5.1 Material & methods

The methods applied as well the research questions, hypotheses and how they were investigated followed the principle introduced in chapters 5.3 and 5.4. Several screening tasks follow the same procedure as described in chapter 4.1.2.1.

5.5.1.1 *Participants*

Twenty-six children participated in this study. They were all attending grade two or grade three at an Italian-German bilingual primary school in Germany and aged 7 to 9 (mean age in months: $M = 103.92$, $SD = 8.20$). At least one of their caregivers is a native Italian-speaker. The children had been living and schooled in Germany for at least two years. For $n = 3$ of these children, teachers have indicated difficulties in oral and written language; one of them at the time of the recruitment being in the process of DD diagnoses, one of them having a special educational need in the area of learning disabilities and the third one having been diagnosed with DLD receiving an SLT intervention that also touches reading-related areas like phonological awareness. Based on the performance below cut-off according to the manuals of standardized reading tests, another $n = 2$ children were considered at risk, (see 5.5.1.5). For one of the TD children, technical issues impeded safe storage of the responses

and response time in the screening tasks so that for some analyses, this participant was excluded from further analyses.

5.5.1.2 Screening tasks

Overall, the four relevant areas of reading assessment RAN, reading speed and accuracy, phonological awareness and grammatical measures were covered in the Italian-German reading screening. To ensure language skills in all languages spoken to be represented in the screening, language tasks, especially regarding phonological awareness were constructed and administered in a comparable manner in both languages. However, since the target group of this study are children living and schooled in Germany, an emphasis was put on German reading screening tasks. All audio clips were pre-recorded by a native speaker with natural voice and accent.

Self-paced syllable reading. The German version of the self-paced syllable reading task underlies the same principle as the Italian self-paced syllable reading task described in chapter 4.1.2.1. Differently from the syllables specific to the Italian orthographic system, all syllables used in the German version of this task adhere to the German phonological system and increased from simple ($n = 10$, e.g. “ka”, see figure 32) over intermediate ($n = 10$, e.g. “spo”) to complex syllables ($n = 10$, e.g. “knau”). The task consists of 3 training and a total of 30 screening items. Self-paced reading time is automatically measured and stored, but due to ceiling effects in Italian pilot studies, accuracy is not tracked. Find more examples in appendix B.



ka

Figure 32: Examinee interface during the German self-paced syllable reading task.

Self-paced sentence reading. The German version of the self-paced sentence reading task underlies the same principle as the German self-paced sentence reading task described in chapter 4.1.2.1. Participants are asked to read aloud and as fast as possible a list of five German sentences (consisting of high-frequent words, e.g. “Die Mama liest in dem großen Buch.” [The mum reads in the big book.], see figure 33), increasing in syntactic complexity and sentence length, presented one by one on the PC screen. The task consists of 1 training and a total of 5 screening items. Self-paced reading time is automatically measured and stored, but due to ceiling effects in Italian pilot studies, accuracy is not tracked. Find more examples in appendix B.



Die Mama liest in dem großen Buch.

Figure 33: Examinee interface during the German self-paced sentence reading task.

Word identification. The German version of the word identification screening task underlies the same principle as the Italian word identification screening task described in chapter 4.1.2.1. Here, a pre-recorded German word is played to the child participant, e.g. “scheinen” [to shine] or [to seem], see figure 34. The orthographic form of this word is also displayed on the screen along with two more words acting as distractors that differ in spelling and pronunciation. Across items, one of the distractors consists in a visual distractor (e.g. “schienen” [shone] or [seemed], substitution of “ei” with “ie”), while the third word displayed is a phonological distractor with inversion or elision of one of the graphemes (“schneien”, [to snow]). The child is asked to indicate which one of the three orthographic forms displayed on the screen corresponds to the word that was presented auditorily by selecting the corresponding button. The task consists of 2 training and a total of 8 screening items. Response time and accuracy

of the given responses are automatically measured and stored. Find more examples in appendix B.

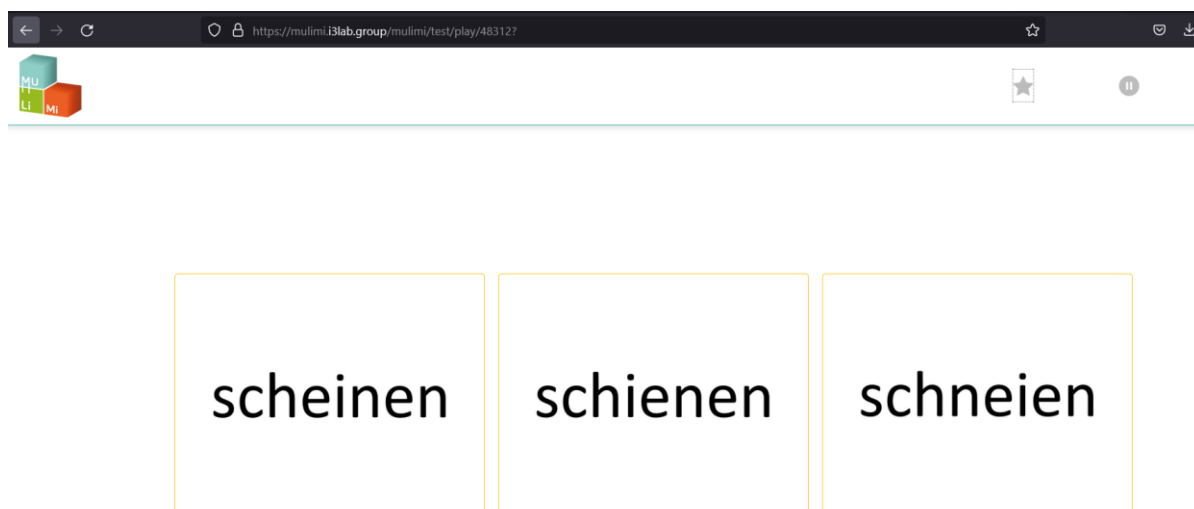


Figure 34: Examinee interface during the German word identification task.

Nonword identification. The German version of the NW identification screening task underlies the same principle as the Italian NW identification screening task described in chapter 4.1.2.1. Similar to the word identification task, a pre-recorded NW that follows German-specific phonotactic rules, is played to the child participant, e.g. “quaftē” [ˈkwɛftə], see figure 35. The orthographic form of this NW is displayed on the screen along with two NWs acting as distractors that differ in spelling. Across items, one of the distractors consists in a visual/orthographic distractor (e.g. “qualte”, substitution of “f” with “l”), while the third word displayed consists in a grapheme elision (e.g. “quate”, elision of “f”). Again, the child is asked to indicate which one of the three orthographic forms displayed on the screen corresponds to the NW that was presented auditorily by selecting the corresponding button. The task consists of 2 training and a total of 8 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

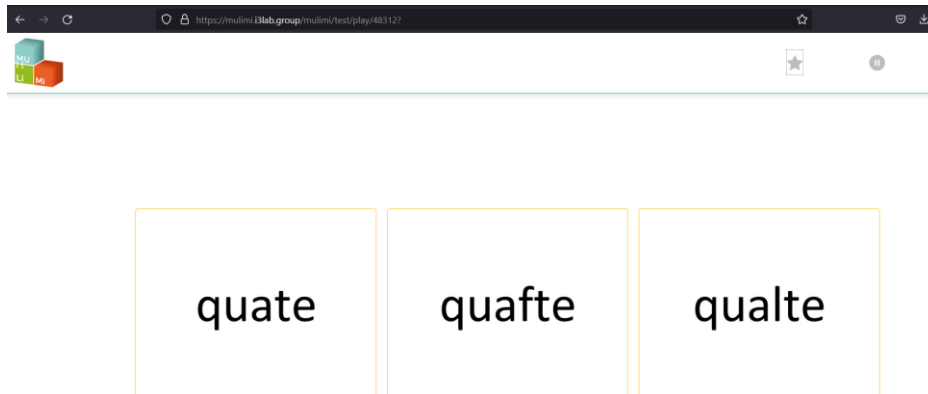


Figure 35: Examinee interface during the German NW identification task.

Phonological blending judgement. The description of the Italian versions of this task (see chapter 4.1.2.1) applies also to its German version. Two pre-recorded audio clips are played to the child participant, the first one consisting of a series of phonemes that when blended would make a German word, presented at a one-second rate (e.g. “L-a-d-e-n” [store] while the second audio presented is a German word spoken normally e.g. “Nadel” [needle]. The child is asked to judge whether or not the blending of the phonemes that were presented auditorily in the first audio correspond exactly to the word that was presented afterwards (in this example, no correspondence) by clicking on the corresponding buttons ✓ for correct and × for incorrect phonological blending, see figure 36. In 50% of the items, the audio clips did not correspond due to phoneme inversion. Items were presented in random order. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

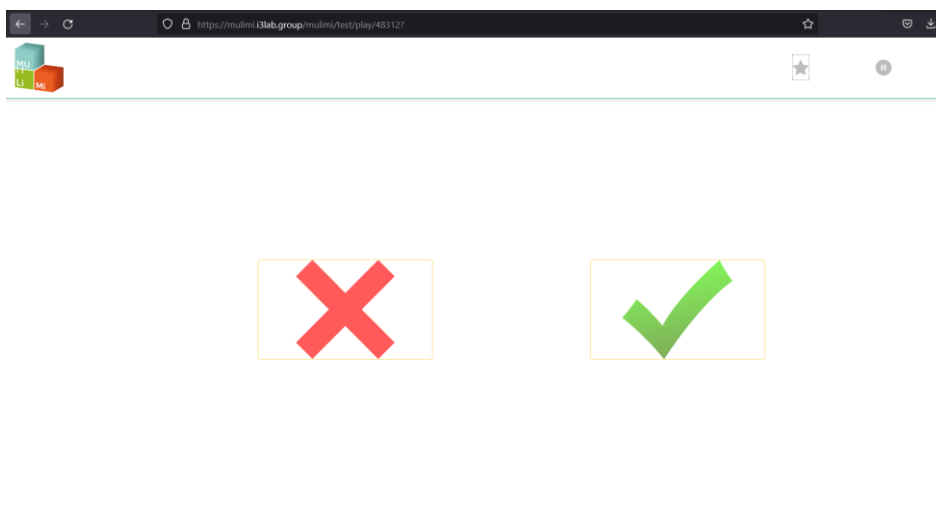


Figure 36: Examinee interface during the judgement screening tasks.

Syllabic inversion judgement. The description of the Italian versions of this task (see chapter 4.1.2.1) applies also to its German version. Again, two pre-recorded audio clips are played to the child participant, the first one being a German word (e.g. “schlafen” [to sleep]) while the second audio presented contained the correctly (i.e. “fen-schla”) or incorrectly (i.e. “fen-schal”) inverted syllables of the same word, all of them being syllables in agreement with German-specific phonotactic rules. The child is asked to judge whether or not the inversion of the syllables that were presented auditorily in the first audio correspond exactly to the word that was presented afterwards by clicking on the corresponding buttons ✓ for correct and × for incorrect syllabic inversion, see figure 36. In 50% of the items, the audio clips did not correspond due to phoneme inversion. Items were presented in random order. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Case matching. The version of the case matching screening task for this study underlies the same principle as the case matching screening task described in chapter 5.4.2.2 with the difference that for this task, only prepositional phrases containing masculine objects were used, meaning that the case distinction depends merely on the distinction of “dem” (dative) vs. “den” (accusative), representing processing of small linguistic units that can be equally classified as phonemes as well as as morphemes. So again, a pre-recorded sentence containing an object marked for case – either accusative or dative – is presented auditorily, for example „Der Esel rennt in den Stall.“ (accusative object, [The donkey runs into the barn.]) while the distractor image depicts a scene that varies regarding the case-marking of the object (i.e. „Der Esel rennt in dem Stall.“ (dative object, [The donkey runs inside the barn.]), see figure 37. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

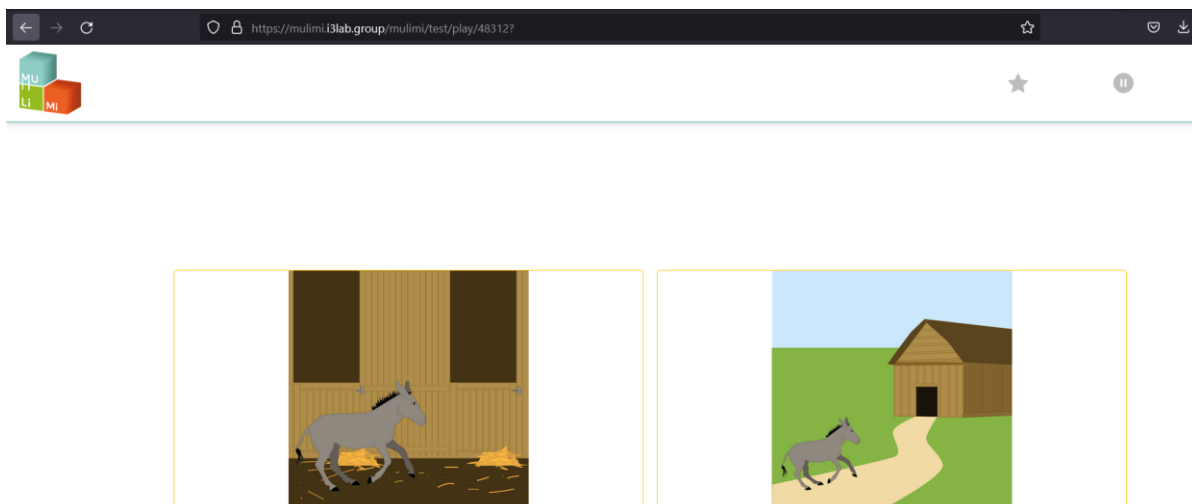
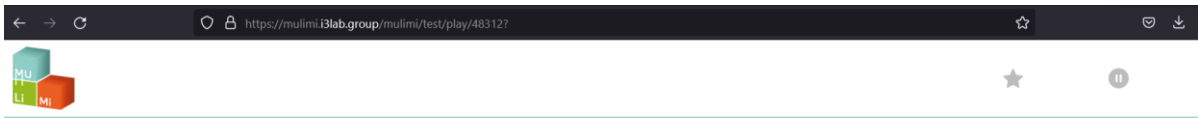


Figure 37: Examinee interface during the German case matching screening task in the DD screening: „Der Esel rennt in den Stall.“ [The donkey runs into the barn.] with target picture displayed on the right and distractor on the left.

Due to variance in the acquisition of Italian orthographic rules in this sample due to different circumstance in formal Italian reading acquisition, not decoding accuracy, but speed in reading of sentences including high-frequent words and naming as well as speed and accuracy in language tasks concerning phonologic awareness and morphosyntactic processing are represented in the Italian reading screening. All audio clips were pre-recorded by a native speaker with natural voice and accent.

RAN. Participants are asked to name a series of digits displayed one by one (see figure 38) on the screen as fast as possible. The next digit is presented upon the child pressing the space bar. The time elapsed between the presentations of two subsequent digits is automatically recorded. The task consists of 5 training and a total of 30 screening items. Due to ceiling effects in pilot studies, accuracy is not tracked.



2

Figure 38: Examinee interface during the Italian RAN (digits) screening task in the DD screening.

For the description of the Italian self-paced sentence reading task, the phonological blending judgement, the syllabic inversion judgement and the clitic pronoun judgement see chapter 4.1.2.1. These tasks were implemented on the MuLiMi screening platform instead of E-Prime for this study. Figure 39 illustrates the implementation of the clitic pronoun judgement screening task when administered through the MuLiMi screening web app. Find examples in appendix B.

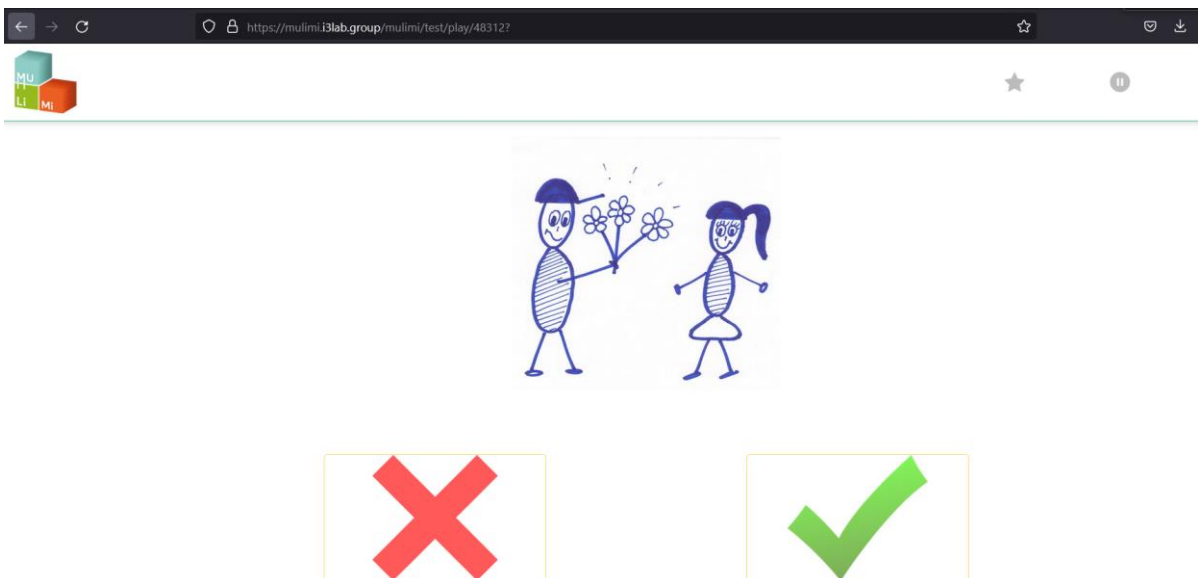


Figure 39: Examinee interface during the Italian clitic pronoun judgement screening task in the DD screening.

5.5.1.3 Standardized tests

The word and NW reading subtests of the German Zürcher Lesetest – II (ZLT-II, Petermann & Daseking, 2019) were used, providing norms on monolingual German-speaking children only. Depending on the grade the children attend, they are asked to read two sets of words all displayed in two to four columns on a single word reading card (“Wortlesekarte”) that they are presented with. 18 to 32 words are presented on two different word cards with one of them being the same (“Wortlesekarte 2”) and one of them differing for the children in this sample from grade 2 (“Wortlesekarte 1”) and 3 (“Wortlesekarte 3”). In addition, both children read the 15 NWs displayed on a separate word card. Children are asked to read as fast and as accurate as possible.

All children recruited for this study attended Italian-German bilingual schools in Germany where they did acquire some literacy skills in Italian, and were thus also administered the word- and NW-reading subtests of the *Batteria per la Valutazione della Dislessia e della Disortografia Evolutiva-2* (DDE-2, Sartori et al., 2007) providing norms on monolingual Italian-speaking children only. In this test, all children read the same four vertically displayed lists of 28 words each and a NW reading subtest including three lists of 16 NWs (non-existing words). Each list is presented to the child one column at a time. Children are asked to read the words or NWs displayed as accurately and fast as possible.

In both tests, the test administrator measures the reading time for each word list or word card with a stopwatch. After the test administration, the child’s reading performance is scored by a trained native speaker based on audio recordings as indicated in the test manual. These results of standardised tests were evaluated according to the criteria in the respective manuals and the norm data provided in the latter. In order to uniquely process the data obtained, the t-scores obtained for the ZLT-II were transformed into z-scores. Furthermore, the raw scores were converted into percentages to facilitate comparison with the results of the experimental tests.

Furthermore, nonverbal intelligence was tested by the means of the *CPM Raven’s Coloured Progressive Matrices* (Bulheller & Häcker, 2001). For the description of the test, see chapter 5.3.2.3.

5.5.1.4 Procedure

The data collection took place in February and March 2020 in the bilingual schools in three German cities where the children were recruited. Subtests for reading words and NWs of the German standardised ZLT-II (Petermann & Daseking, 2019) as well as the Italian standardised DDE-2 (Sartori et al., 2007) were carried out, which lasted roughly 30 minutes in total.

They were conducted with each child individually and evaluated according to the criteria described in the respective manuals by native speakers later on based on audio recordings. The children's individual results were then compared to the norm data provided in the test manuals. Furthermore, teachers and caregivers filled in the pen-and-paper versions of the respective questionnaires choosing between the Italian or German version. In addition, the Italian-German MuLiMi reading screening was administered to each child in an individual setting in the presence of the examiner, which took approximately another 30 minutes and was administered in a separate session after a break following up the administration of the standardized tests. In order to use the screening platform in a realistic setting, the PCs available in the schools where recruitment and testing took place were used whenever possible.

5.5.1.5 Risk score creation

In order to process the data irrespective of the presence of an official diagnosis that might be of particular complexity in bilingual children due to the diagnostic dilemma and the potential of misdiagnoses described in chapter 2.4., risk scores were created. In a first step, for Italian and German separately, the information on whether or not the children had scored two standard deviations below norm in accuracy and speed for the word and NW reading standardized subtests was scored with 1 point each, summed up and divided by four (due to 4 values reading time and errors for words and NWs separately that were considered). In a further step, the latter score for Italian and German separately was summed up and considered the total balanced risk. In that way, despite differences in test construction, both tests equally contributed to this score. Here, children obtained scores of minimum 0 and maximum 2. Based on the total balanced risk, the overall risk was derived. Children who scored above 1 in the total balanced risk – indicating that they either had difficulties in both languages or severe difficulties in one language for both reading accuracy and time – were assigned an overall risk of 1, while the other children had a risk of 0. A further score was created indicating the characteristics of reading performance, i.e. whether the children had scored two standard deviations below the norm in the German reading subtests only, the Italian subtests only or in both standardized tests. This qualitative information was then further processed quantitatively in terms of the degree of the impairment with scores ranging from 0 to 2 and the language(s) in which a deficit is present (ZLT-II only, DDE-2 only or both). In addition to the $n = 3$ children who were classified as at risk of DD according to the teacher, another $n = 2$ children were identified as at risk of DD based on these risk scores.

5.5.2 Results & discussion

Data collection in this study was severely constrained since, due to the restrictions of the Covid-19-pandemic, data collection was interrupted. For this reason, not all children were

administered the Raven's CPM. In fact, this test was carried out with $n = 10$ children. Despite the small sample size, the children's performance in the Raven's CPM were compared to screening task performances. Only one significant association emerged with accuracy (%) in the Italian clitic pronoun judgement task ($n = 10$, $\rho = .760$, $p = .011$). Due to the inappropriateness of parametric tests with a sample size of $n = 10$ children, whenever significant associations with the children's accuracy in Italian clitic pronoun judgement emerged, the children's performance in the Raven's CPM could not be inserted as control variable.

Since for these analyses, children from two different grades and raw scores for both the standardized and screening tasks were used, partial correlations controlling for age in months were calculated when comparing them. Because of the fact that in the ZLT-II, depending on the grade, the participants read different word lists, raw score comparisons for the whole group are only carried out for the word card 2 that all of the participants read. Due to the nature of z-scores, when they were used for comparisons, the effect of age instead is not factored out. A group of $n = 5$ children had been assigned the overall risk. All $n = 3$ children who were known to have language, learning or reading difficulties as indicated in chapter 5.5.1.5 were part of this risk group.

5.5.2.1 Comparison of screening tasks and standardized test results within languages

Reading time. Reading times in the German screening tasks were significantly associated with reading time raw scores in the ZLT-II. In particular average reading time in the self-paced syllable reading task correlated significantly with word (word card 2, $n = 25$, $r = .759$, $p < .001$, with a $R^2 = .578$, see figure 40) and NW reading time in the ZLT-II ($n = 25$, $r = .774$, $p < .001$). Similarly, self-paced sentence reading time was significantly associated with word (word card 2, $n = 25$, $r = .774$, $p < .001$) and NW reading time in the ZLT-II ($n = 25$, $r = .679$, $p < .001$). Also average response time in the word identification screening task was associated with word (word card 2, $n = 25$, $r = .703$, $p < .001$) and NW reading time in the ZLT-II ($n = 25$, $r = .386$, $p = .057$). Associations also reached significance for both ZLT-II subtests when being compared to the average response time in the NW identification task (words: $n = 25$, $r = .590$, $p = .002$; NWs: $n = 25$, $r = .624$, $p < .001$).

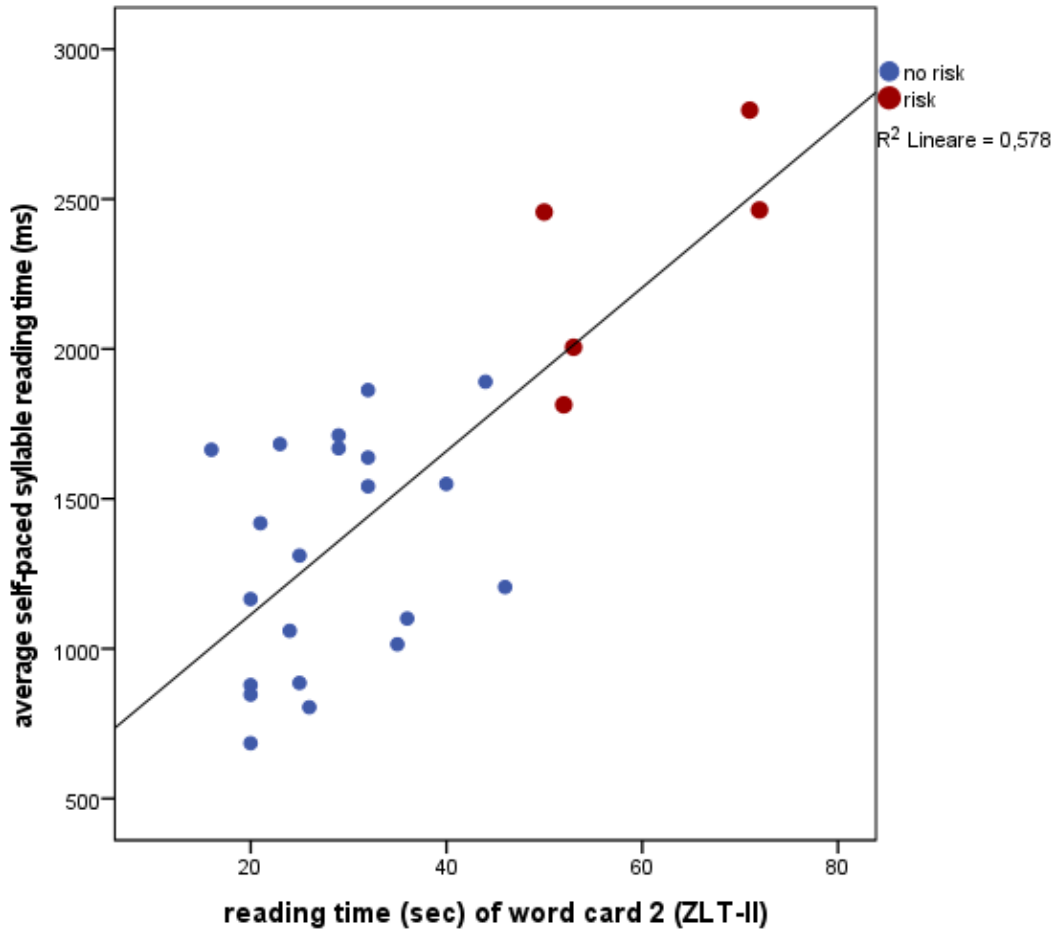


Figure 40: Comparison of average self-paced reading time of German syllables (ms, screening) and word reading time (sec, ZLT-II) according to the child participants' risk level.

Response time in the phoneme blending and syllabic inversion judgement screening tasks were not significantly associated with reading time raw scores in the ZLT-II ($p_s > .05$). Response time in the case-marked sentence picture matching task instead was significantly, but negatively associated with reading time of words ($n = 24$, $r = -.433$, $p = .035$) in the ZLT-II, suggesting that fast readers take more time in the case matching task or slow readers are faster in the case matching task.

Also the Italian self-paced sentence reading time correlated significantly with reading time raw scores in the word ($n = 24$, $r = .971$, $p < .001$) and NW ($n = 24$, $r = .690$, $p < .001$) reading subtests of the DDE-2. Response time in the phonological awareness and clitic pronoun judgement screening tasks were not significantly associated with reading times in the DDE-2 ($p_s > .05$).

Reading accuracy. Accuracy in the word identification screening task was significantly and negatively associated with the raw scores of errors obtained in the word (word card 2, n

= 25, $r = -.765$, $p < .001$) and NW ($n = 25$, $r = -.643$, $p < .001$) reading subtests of the ZLT-II. A similar pattern was found for accuracy in the NW identification screening tasks and the errors obtained in the word (word card 2, $n = 25$, $r = -.567$, $p = .003$) and NW ($n = 25$, $r = -.628$, $p < .001$) subtests of the ZLT-II. Also accuracy in the German blending judgement screening task was significantly and negatively associated with errors in the word ($n = 25$, $r = -.544$, $p = .005$) and NW ($n = 25$, $r = -.501$, $p = .011$) reading subtests of the ZLT-II. No significant associations instead emerged for accuracy in the inversion judgement screening task and the standardized test results. Accuracy in the picture matching task of case-marked sentences was significantly and negatively associated with errors in the word ($n = 24$, $r = -.642$, $p < .001$) and NW ($n = 24$, $r = -.540$, $p = .006$) reading subtests of the ZLT-II, too.

For Italian, accuracy in phoneme blending judgement was significantly and positively associated with errors (raw scores) in word ($n = 24$, $r = .451$, $p = .027$) and NW ($n = 24$, $r = .505$, $p = .012$) reading subtests of the DDE-2. Syllabic inversion instead was significantly and negatively associated with errors in the word ($n = 24$, $r = -.418$, $p = .042$) but not the NW reading task in the DDE-2 ($p > .05$). Also accuracy in the Italian clitic pronoun judgement screening task was significantly and negatively associated with errors in the word ($n = 24$, $r = -.588$, $p = .003$) but not the NW ($p > .05$) reading task in the DDE-2.

5.5.2.2 Comparison of screening tasks and standardized test results across languages and task types

Associations between tasks of the same language within as well as across linguistic areas were found for both Italian and German.

In particular, reading and response times obtained in the German self-paced sentence and syllable reading as well as obtained in the word and NW identification tasks were significantly associated with each other ranging from $r = .500$ to $r = .656$, $p_s < .004$. Also German word and NW identification accuracy were significantly correlated with each other ($n = 26$, $r = .660$, $p < .001$). Furthermore, accuracy in the German blending judgement task was associated with accuracy in the German word ($n = 26$, $r = .609$, $p < .001$) and NW identification task ($n = 25$, $r = .683$, $p < .001$). Also response time in the German syllabic inversion judgement task was significantly associated with accuracy in the German NW identification task ($n = 25$, $r = .468$, $p = .018$). Accuracy in the German NW identification and the case matching screening tasks were significantly correlated, too ($n = 25$, $r = .448$, $p = .025$). Moreover, average response time in the German phoneme blending judgement task was significantly associated with accuracy in the German case matching screening task ($n = 25$, $r = .423$, $p = .035$). Significant correlations were also found for response time in the German case matching

task and response time in the judgement task on syllabic inversion in German ($n = 25$, $r = .639$, $p < .001$).

Similarly, accuracy in the Italian syllabic inversion judgement was significantly associated with self-paced sentence reading time of Italian sentences ($n = 25$, $r = -.529$, $p = .006$). Moreover, response time in the Italian phoneme blending judgement task was significantly correlated with response time in the Italian syllabic inversion judgement task ($n = 25$, $r = .441$, $p = .027$). Accuracy in the Italian syllabic inversion judgement task was significantly associated with accuracy in the Italian clitic pronoun judgement task ($n = 25$, $r = .522$, $p = .007$).

Besides associations within languages – in the case of average self-paced reading time for Italian sentences with the z-scores on reading from the DDE-2 – associations emerged also comparing average self-paced reading time for Italian sentences to z-scores on reading time in the German ZLT-II ranging from $r_s = -.951$ to $-.677$ ($p_s < .001$). Furthermore, Italian self-paced sentence reading time correlated significantly with z-scores for Italian word reading performance ($n = 25$, $r = -.679$, $p < .001$), but not NW reading subtest ($p > .05$) of the DDE-2 while reading speed in this task was significantly associated with z-scores obtained in German NW ($n = 25$, $r = -.576$, $p = .003$), but not word reading subtest ($p > .05$). Linear regression analyses showed that the z-scores in the Italian word ($R^2 = .904$) and German NW ($R^2 = .701$) reading tasks explained more variance than the Italian NW ($R^2 = .458$) and German word reading standardized subtests ($R^2 = .575$), see figure 41.

Crosslinguistic associations were also found comparing performance in the screening tasks with each other. Reading time in the German and Italian self-paced sentence reading task correlated significantly with each other ($n = 25$, $r = .841$, $p < .001$). For most results obtained in the phonological awareness judgement tasks, significant associations between performance in the same task across languages regarding both response time (phoneme blending: $n = 25$, $r = .565$, $p = .003$; syllabic inversion: $n = 25$, $r = .720$, $p < .001$) and accuracy (syllabic inversion: $n = 25$, $r = .550$, $p = .004$) were found. A similar pattern was not found for the grammatical tasks on Italian clitic pronoun judgement and German case matching tasks ($p > .05$).

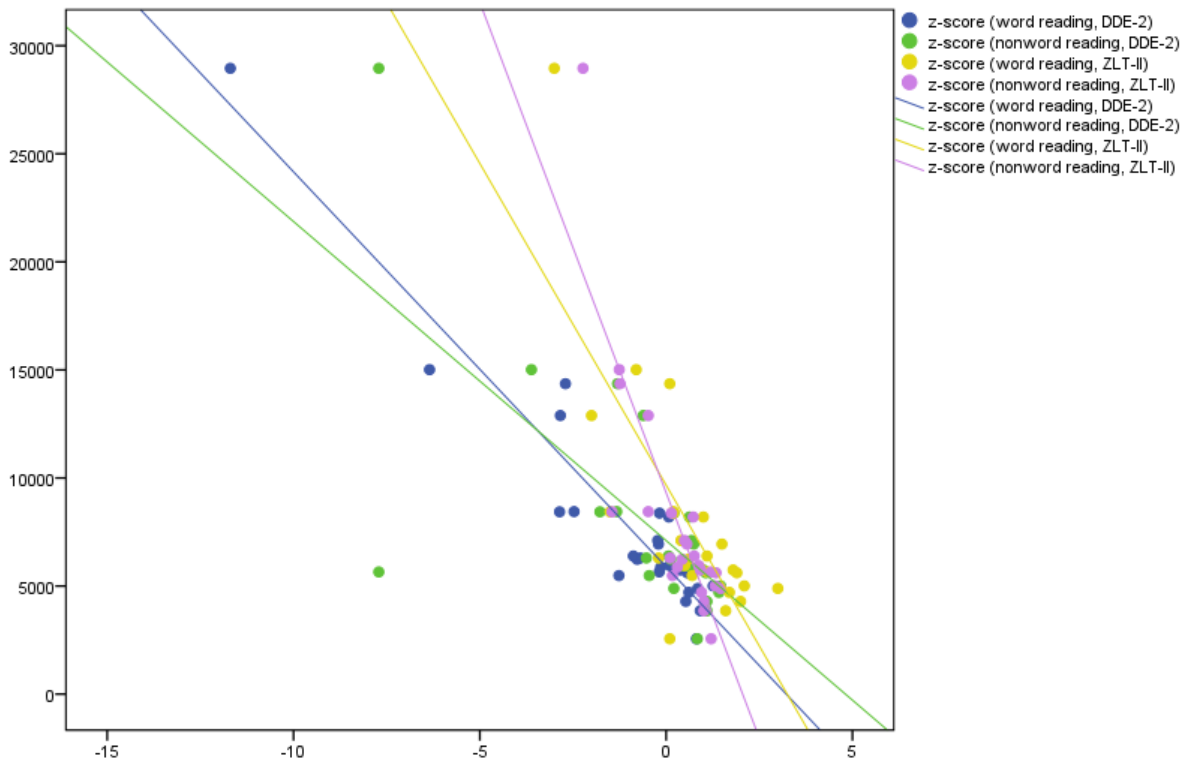


Figure 41: Comparison of average self-paced reading time (ms) for Italian sentences (y-axis) and z-scores (reading time) in Italian and German standardized word and NW reading subtests (x-axis).

Interesting effects were also found comparing different task types across languages, for example comparing self-paced reading time of German syllables to response time in the RAN task targeting Italian digits ($n = 25$, $r = .406$, $p = .044$) as well as to reading time Italian self-paced sentence reading task ($n = 25$, $r = .683$, $p < .001$). Interestingly, neither accuracy nor response time in the German word identification ($p > .05$), but accuracy ($n = 25$, $r = -.414$, $p = .040$) and average response time ($n = 25$, $r = .557$, $p = .004$) in the German NW identification task were significantly associated with Italian self-paced sentence reading time (see patterns described in figure 41). Response time in the German syllabic inversion judgement task was not significantly, but somewhat associated with response time in the Italian clitic pronoun judgement task ($n = 25$, $r = .388$, $p = .055$). Also response time in the German case matching task was significantly correlated with response time in the Italian judgement tasks on phoneme blending ($n = 25$, $r = .648$, $p < .001$) as well as syllabic inversion ($n = 25$, $r = .534$, $p = .006$). Accuracy in the Italian blending judgement was significantly and negatively associated with German word identification accuracy ($n = 25$, $r = -.545$, $p = .005$). Accuracy in the Italian syllabic inversion judgement screening task was significantly associated with German average self-paced syllable ($n = 25$, $r = -.428$, $p = .033$) and sentence reading time ($n = 25$, $r = -.454$, $p = .023$) as well as NW ($n = 25$, $r = -.406$, $p = .044$), but not word identification response time ($p > .05$).

5.5.2.3 Comparison of screening tasks and risk scores

In a first step, non-parametric correlations comparing the children's age in months with the risk scores obtained revealed no significant association between them ($p_s > .05$).

Point-biserial correlations were run to determine the relationship between the overall risk score (risk: $n = 5$, vs. no risk: $n = 21$) and screening task performance. Significant correlations emerged between the overall risk score and response time in the German self-paced syllable reading ($n = 26$, $r_{pb} = .730$, $p < .001$), the self-paced sentence reading ($n = 26$, $r_{pb} = .608$, $p < .001$, see figure 42), word identification task ($n = 26$, $r_{pb} = .474$, $p = .014$) and in the nonword identification task ($n = 26$, $r_{pb} = .621$, $p < .001$). Furthermore, significant correlations emerged between the overall risk score and Italian self-paced sentence reading time ($n = 25$, $r_{pb} = .768$, $p < .001$) as well as with accuracy in the Italian syllabic inversion judgement task ($n = 25$, $r_{pb} = -.522$, $p = .007$) and accuracy in the Italian clitic pronoun judgement task ($n = 25$, $r_{pb} = -.586$, $p = .002$, see figure 43). These significant correlations were confirmed by Mann-Whitney U tests. From this data, it can be concluded that children belonging to the risk group, compared to children not belonging to the risk group scored significantly lower in the German self-paced syllable reading task ($U = 2.00$ $p = .001$), the German self-paced sentence reading task ($U = 7.00$ $p = .003$), the German word identification task (response time: $U = 20.00$ $p = .034$), the German nonword identification task (response time: $U = 12.00$ $p = .008$), the Italian self-paced sentence reading task ($U = 1.00$ $p < .001$), the Italian syllabic inversion judgement task (accuracy: $U = 13.50$ $p = .012$) and the Italian clitic pronoun judgement task (accuracy $U = 9.50$ $p = .004$).

The difference in reading performance in screening performance depending on the risk status can not only be observed by visual inspection of figure 40 in chapter 5.5.2.1. All but one of the German reading screening tasks' (self-paced syllable and sentence reading, word and NW identification task) reaction and reading times correlated significantly with all of the risk scores created (see chapter 5.5.1.5), ranging from $\rho_{os} = .423$ to $.657$, $p_s < .031$. Response time in the German word identification task however was non-significantly associated with the balanced risk score ($n = 26$, $\rho = .387$, $p = .051$) and the Italian risk score ($p > .05$). The difference in the self-paced syllable reading time according to the child's overall risk is displayed in figure 42.

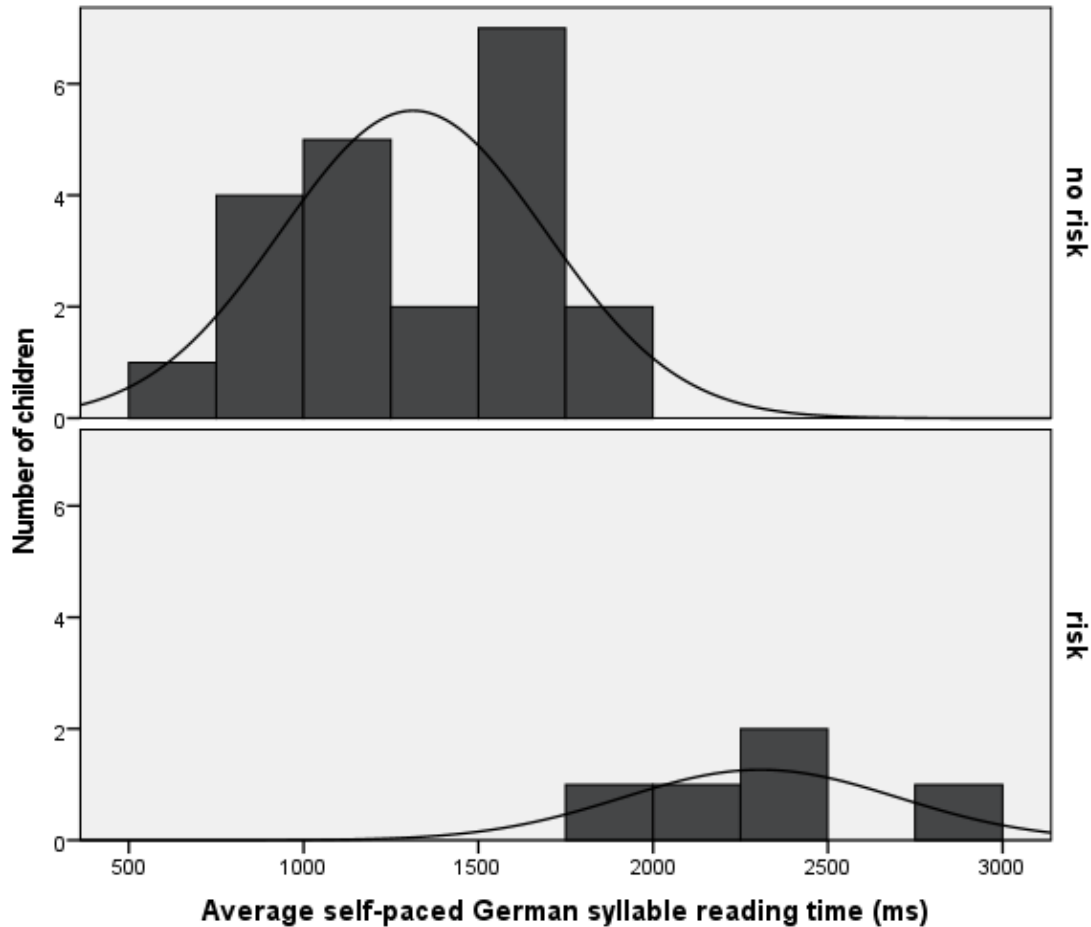


Figure 42: Number of children grouped by average self-paced reading time of German syllables (ms, screening) according to the child's overall risk.

In addition to that, accuracy in the German case matching screening task was significantly associated with all risk scores with ρ_s ranging from $-.423$ to $-.607$, $p_s < .035$.

Similarly, the Italian self-paced sentence reading time was significantly correlated with all of the risk scores (see chapter 5.5.1.5), with ρ_s ranging from $.513$ to $.630$, $p_s < .009$. In addition to that, also accuracy in both syllabic inversion and phoneme blending judgement was significantly associated with all of the above mentioned risk scores with ρ_s ranging from $-.469$ to $-.569$, $p_s < .018$ (except for the German risk score, $p > .05$). Furthermore, accuracy in the clitic pronoun judgement task correlated significantly with the Italian risk score ($n = 25$, $\rho = -.482$, $p = .015$). This is exemplified in figure 43 displaying the accuracy in the clitic pronoun judgement screening task according to the child's overall risk. Note however that the bimodal distribution in the graph shows that not all children show difficulties in this task.

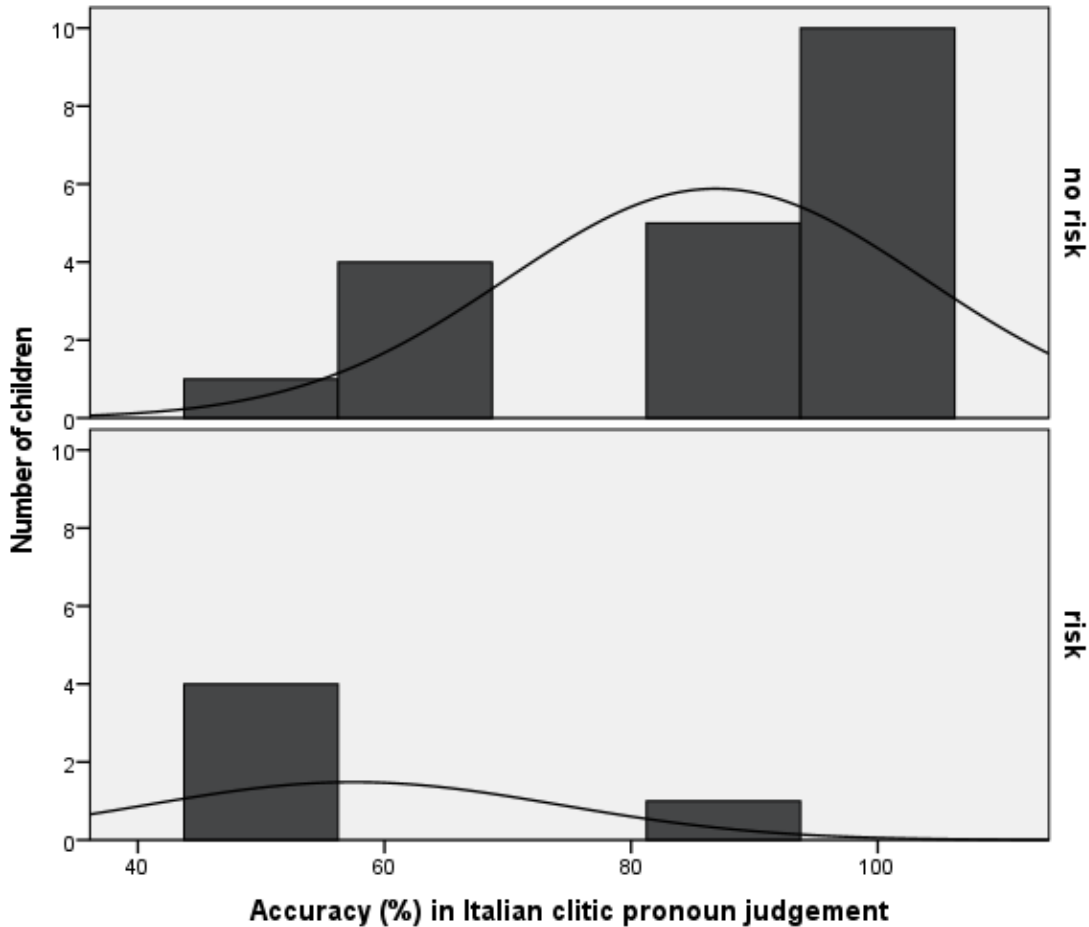


Figure 43: Number of children grouped by accuracy in Italian clitic pronoun judgement (% , screening) according to the child's overall risk.

5.5.2.4 Comparison of screening tasks and caregiver questionnaires

Since caregivers' questionnaire responses were not normally distributed, as assessed by a Shapiro-Wilk-Test, $p < .05$, non-parametric tests were applied. Before using the caregiver questionnaire for the preliminary validation of the screening tasks, correlations between the questionnaire responses and z-scores on word and NW reading time were calculated for the German standardized reading tests with rho_s ranging from $-.374$ to $-.445$, $p_s < .072$, but not with z-scores obtained in the DDE-2 ($p_s > .05$). Accordingly, the total and the individual categories' scores of the DSA questionnaires were associated with risk scores derived from German standardized reading subtests ($rho_s = .466$ to $.400$ with $p_s < .059$), but not with the risk score deriving from the Italian standardized reading subtests ($p_s > .05$). However, significant associations between the score in the "School discomfort" category and the balanced risk ($n = 24$, $rho = .419$, $p = .042$) and the overall risk score ($n = 24$, $rho = .453$, $p = .026$) – both incorporating the results from German and Italian standardized tests equally – were found.

Furthermore, non-parametric correlations comparing the children's age in months with the risk scores obtained revealed no significant association between them ($p_s > .05$).

The DSA questionnaire total score (sum of all the times caregivers respond "yes" to one of the questions indicating risk) was shown to be significantly associated with average self-paced reading time of German syllables ($n = 24$, $\rho = .412$, $p = .045$) as well as response time in the German word ($n = 24$, $\rho = .446$, $p = .029$) and NW ($n = 24$, $\rho = .553$, $p = .005$) identification. Investigating the responses for the subcategories individually, further associations with reading or response time in various screening tasks emerged. More specifically, average self-paced reading time of German sentences was associated with caregivers' answers in the categories of "Scholar achievements" ($n = 24$, $\rho = .420$, $p = .041$) as well as with "Reading and writing development" ($n = 24$, $\rho = .383$, $p = .071$), though not significantly. Average response time in the German word identification task was associated with answers in the categories "School discomfort" ($n = 24$, $\rho = .407$, $p = .048$) and "Scholar achievements" ($n = 24$, $\rho = .396$, $p = .055$), though not significantly. Average response time in the German NW identification task instead was shown to be associated with all of categories mentioned above, namely the categories "School discomfort" ($n = 24$, $\rho = .597$, $p = .002$) and "Scholar achievements" ($n = 24$, $\rho = .404$, $p = .050$), and with "Reading and writing development" ($n = 24$, $\rho = .406$, $p = .055$), though not significantly. Significant associations between the performance in the German phoneme blending judgement and case matching screening tasks and the DSA questionnaire total and individual subcategories' scores emerged. In particular, average response time in phoneme blending judgement was significantly associated with the total score ($n = 24$, $\rho = .446$, $p = .029$) and caregivers' responses in the "Reading and writing development" category ($n = 23$, $\rho = .491$, $p = .017$). Accuracy in the German case matching task was shown to be associated with the total score ($n = 23$, $\rho = -.444$, $p = .034$) as well as with caregivers' responses in the "Scholar achievements" ($n = 23$, $\rho = -.549$, $p = .007$) and "Reading and writing development" category ($n = 23$, $\rho = -.399$, $p = .066$), though not significantly.

The amount of (significant) associations between performance in Italian screening tasks and responses in the DSA questionnaire was lower. In particular, self-paced reading time of Italian sentences was associated with the DSA questionnaire's total score ($n = 23$, $\rho = .390$, $p = .066$) and the score of the "Reading and writing development" category ($n = 22$, $\rho = .456$, $p = .033$). Response time in the syllabic inversion judgement task was also significantly associated with the "Reading and writing development" category ($n = 22$, $\rho = .501$, $p = .017$).

5.5.2.5 Comparison of screening tasks and teacher questionnaires

Since all the single categories of the teacher questionnaire – except for the responses related to the child’s productive oral lexical skills ($p > .05$) – and all the compound scores deriving from the latter were significantly correlated with the children’s overall risk score deriving from their performance in Italian as well as German standardized reading tests ($n = 25$, ρ_s ranging from .400 to .673, $p_s < .043$), the responses given by the teachers were considered reliable and also compared to the children’s performance in the screening tasks. When comparing the children’s performance in the screening tasks to the teacher questionnaire’s responses, due to the nature of the data (5-point Likert-scale), Spearman-rho’s were calculated.

Average reading time in the German self-paced syllable reading screening task was significantly associated with the teachers’ ratings of phonological and receptive phonological as well as morphosyntactic skills and also writing and reading comprehension ($n = 25$, ρ_s ranging from .394 to .496, $p_s > .047$). Accordingly, the average self-paced reading time for German syllables correlated significantly with the teacher total ($n = 26$, $\rho = .452$, $p = .021$), production ($n = 26$, $\rho = .530$, $p = .005$) and literacy compound scores ($n = 26$, $\rho = .438$, $p = .025$). Average reading time as measured in the German self-paced sentence reading screening task was significantly associated with all single response categories ($n = 25$, ρ_s ranging from .414 to .588, $p_s < .036$) except for the teachers’ judgements of the child’s pragmatic competences ($n = 19$, $\rho = .435$, $p = .063$). A significant correlation was also found for the compound score incorporating all the single scores and the average reading time in the German self-paced sentence reading task ($n = 26$, $\rho = .541$, $p = .004$). Significant associations were also found comparing the children’s response times in the German word and NW identification tasks to the teachers’ judgements of the children’s language and literacy skills, but not for accuracy ($p_s > .05$). In particular, average response time in the German word identification task was significantly associated with the teacher’s total compound score based on the sum of responses to all single categories ($n = 26$, $\rho = .416$, $p = .034$). More significant associations emerged comparing the children’s average response times in the German NW identification task to the responses to the teacher questionnaire: Teachers’ judgements regarding the child’s receptive phonological skills ($n = 25$, $\rho = .484$, $p = .014$), their competence in reading out aloud ($n = 25$, $\rho = .442$, $p = .027$) as well as the total ($n = 26$, $\rho = .416$, $p = .035$) and the production ($n = 26$, $\rho = .411$, $p = .037$) compound scores were significantly correlated with the children’s average response times in the German NW identification screening task. While for neither of the German phonological awareness screening tasks significant associations with the responses to the teacher questionnaire emerged, significant associations were found for both accuracy and response time in the German case

matching task. For accuracy in the German case matching task, significant associations emerged with the teachers' judgements of the children's phonological ($n = 25$, $\rho = -.409$, $p = .042$), morphosyntactic, lexical ($n = 25$, $\rho = -.465$, $p = .019$) productive skills ($n = 25$, $\rho = -.434$, $p = .030$) as well as with the total compound score ($n = 26$, $\rho = -.415$, $p = .039$). Average response time in the same screening task was found to be significantly associated with teachers' judgements of phonological productive ($n = 25$, $\rho = -.396$, $p = .050$), morphosyntactic productive ($n = 25$, $\rho = -.442$, $p = .027$), writing ($n = 25$, $\rho = -.488$, $p = .013$) and reading comprehension skills ($n = 25$, $\rho = -.422$, $p = .040$). Notably, the faster the children responded, the worse were the teachers' judgements of the children's phonology, morphosyntax, writing and reading comprehension skills.

Interestingly, for some of the judgements by German-speaking teachers who evaluated the children's language and literacy abilities in German significant associations with Italian screening tasks emerged. Similar as for the associations found between the German self-paced syllable reading task, all but two (phonology and lexicon production skills) teacher questionnaire's single categories and all compound scores were significantly associated with average self-paced reading time for Italian sentences with ρ_s ranging from .446 to .673, $p_s < .034$ ($n_s > 18$). While none of the German phonological awareness task results were significantly correlated with any of the single or compound scores of the teacher questionnaires, accuracy (in %) the Italian phonological blending judgement task was positively associated with the total receptive ($n = 25$, $\rho = .428$, $p = .033$), total productive ($n = 25$, $\rho = .428$, $p = .033$) and total literacy ($n = 25$, $\rho = -.437$, $p = .029$) compound scores deriving from the responses to the single categories in the teacher questionnaires, indicating that the better the children scored in the phonological blending judgement task, the worse the teachers' judgements of their competences. Accuracy in the syllabic inversion judgement task was not associated with any of the compound scores, but negatively associated with teachers' judgements regarding the children's lexical receptive skills ($n = 25$, $\rho = -.411$, $p = .046$) as well as their abilities in reading out aloud ($n = 25$, $\rho = -.459$, $p = .024$) indicating that the worse teachers evaluated their language and reading skills, the worse the children scored in the phonological blending judgement task. Children's performance (accuracy in %) in the clitic pronoun judgement task was significantly associated with all but three (phonological, morphosyntactic and lexical production skills) of the single scores in the teacher questionnaire with ρ_s ranging from -.424 to -.670, $p_s < .035$ ($n_s > 18$). Also the total compound score incorporating all the single scores from the teacher questionnaire correlated significantly with accuracy (in %) in the clitic pronoun judgement task ($n = 25$, $\rho = -.508$, $p = .009$).

5.5.2.6 Interim discussion

Across domains, significant associations between accuracy and reaction/reading time in the screening tasks with variables related to DD risk, standardized tests as well as caregiver and teacher questionnaires emerged (see table 13 for an overview), allowing for a confirmation of the research questions raised on concurrent and discriminant accuracy of the Italian-German computerized reading screening administered in face-to-face settings.

Table 13: Overview of significant associations between screening task performance (accuracy in %) risk level, standardized tests as well as teacher and caregiver questionnaires.

Screening task	Risk score	standardized tests	teacher questionnaire	DSA questionnaire
		(raw scores)		
		ZLT-II		
syllables (GER) self-paced reading time	overall risk: $n = 26, r_{pb} = .730, p < .001$	word card 2: reading time $n = 25, r = .759, p < .001$	<i>total score:</i> $n = 25, rho = .452, p = .021$	<i>total score:</i> $n = 24, rho = .412, p = .045$
sentences (GER) self-paced reading time	overall risk: $n = 26, r_{pb} = .608, p < .001$	word card 2: reading time $n = 25, r = .774, p < .001$	<i>total score:</i> $n = 25, rho = .541, p = .004$	<i>scholar achievements:</i> $n = 24, rho = .420, p = .041$
word identification (GER) response time	overall risk: $n = 26, r_{pb} = .474, p = .014$	word card 2: reading time $n = 25, r = .703, p < .001$	<i>total score:</i> $n = 26, rho = .416, p = .034$	<i>total score:</i> $n = 24, rho = .446, p = .029$
word identification (GER) accuracy (%)	n.s.	word card 2: errors $n = 25, r = -.765, p < .001$	n.s.	n.s.
nonword identification (GER) response time	overall risk: $n = 26, r_{pb} = .621, p < .001$	nonwords: reading time $n = 25, r = .624, p < .001$	<i>total score:</i> $n = 26, rho = .416, p = .035$	<i>total score:</i> $n = 24, rho = .553, p = .005$
nonword identification (GER) accuracy (%)	n.s.	nonwords: errors $n = 25, r = -.628, p < .001$	n.s.	n.s.
phoneme blending (GER) response time	n.s.	n.s.	n.s.	<i>total score:</i> $n = 24, rho = .446, p = .029$
phoneme blending (GER) accuracy (%)	n.s.	word card 2: errors $n = 25, r = -.544, p = .005$	n.s.	n.s.
syllabic inversion (GER) response time	n.s.	n.s.	n.s.	n.s.
syllabic inversion (GER) accuracy (%)	n.s.	n.s.	n.s.	n.s.
case marking (GER) response time	n.s.	word card 2: reading time $n = 24, r = -.433, p = .035$	n.s.	n.s.
case marking (GER) accuracy (%)	German risk: $n = 25, rho = -.538, p = .006$	word card 2: errors $n = 24, r = -.642, p < .001$	<i>total score:</i> $n = 26, rho = -.415, p = .039$	<i>total score:</i> $n = 23, rho = -.444, p = .034$

DDE-2				
RAN (IT) response time	Italian risk: $n = 25$, $\rho = .038$	n.s.	n.s.	<i>reading & writing:</i> $n = 22$, $\rho = .033$
sentences (IT) self-paced reading time	overall risk: $n = 25$, $r_{pb} = .768$, $p < .001$	<i>word reading time:</i> $n = 24$, $r = .971$, $p < .001$	<i>total score:</i> $n = 25$, $\rho = .514$, $p = .009$	n.s.
phoneme blending (IT) response time	n.s.	n.s.	n.s.	n.s.
phoneme blending (IT) accuracy (%)	n.s.	<i>word reading errors:</i> $n = 24$, $r = .451$, $p = .027$	n.s.	n.s.
syllabic inversion (IT) response time	n.s.	n.s.	n.s.	<i>reading & writing:</i> $n = 22$, $\rho = .501$, $p = .017$
syllabic inversion (IT) accuracy (%)	overall risk: $n = 25$, $r_{pb} = -.522$, $p = .007$	<i>word reading errors:</i> $n = 24$, $r = -.418$, $p = .042$	<i>reading aloud:</i> $n = 24$, $\rho = -.459$, $p = .024$	n.s.
clitic pronouns (IT) response time		n.s.	<i>total score:</i> $n = 25$, $\rho = -.508$, $p = .009$	n.s.
clitic pronouns (IT) response time	overall risk: $n = 25$, $r_{pb} = -.586$, $p = .002$	<i>word reading errors:</i> $n = 24$, $r = -.588$, $p = .003$	n.s.	n.s.

Overall, the various associations of screening results with both raw and z-scores (Hypothesis 2) as well as risk scores (Hypothesis 1) derived from standardized tests both within and across languages point to the potential of the computerized screening task to assess a child's reading performance in both languages and thus to identify the risk of DD in Italian-speaking children living in Germany. Based on the amount of significant associations that emerged between the standardized reading measures, the risk scores derived therefrom and the caregiver/teacher questionnaires with the screening tasks, some tasks seem to be more useful in this context than others. For German, self-paced reading of syllables seems to be more revealing than self-paced sentence reading. While this difference between the German word and NW identification screening task is less pronounced, it is evident that the German phoneme blending judgement screening task is more often significantly associated with other reading measures, screening tasks and caregivers' responses to the DSA questionnaire. Notably, for the Italian versions of these tasks, this effect was not found. Both German case matching and Italian clitic pronoun judgement were shown to be associated with other reading measures, screening tasks and caregivers' responses to the DSA questionnaire, but not with each other. This highlights the potential differences in language acquisition across languages. The performances in judgement tasks on phonological awareness instead were shown to be

associated crosslinguistically which points to them being more independent from language knowledge and exposure (Hypothesis 4). Also self-paced sentence reading in both languages was shown to be correlated across languages and Italian RAN (digits) correlated significantly with self-paced syllable and sentence reading in German pointing to the potential of processing speed irrespective of language exposure and accuracy evaluation on the side of the examiner in DD risk identification. The observation that significant associations between grammatical and phonological awareness tasks both within (Hypothesis 3) and across languages could be interpreted in manifold ways, i.e. both are equally impaired when reading and writing problems occur and should thus be equally considered in the diagnostic process or that the general underlying functions of processing small linguistic units (phonemes in the phonological awareness and morphemes in the morphosyntactic processing tasks) are similar. However, not all children who were considered at risk according to their performance in the Italian and German standardized tests did perform poorly in the clitic pronoun judgement task. Since the subtests administered from the standardized tests did neither directly test phonological awareness nor morphosyntactic processing performance, this finding cannot be exhaustively discussed and interpreted. Here, the specific tasks types that were implemented allowing for automatic administration and evaluation of the screening tasks in both languages spoken might also be at play – both phonological awareness and clitic pronoun judgement tasks as well as case matching tasks similarly require attentional and cognitive skills. Future validation studies should thus incorporate more traditional phoneme and morpheme manipulation tasks to understand whether or not the specific format of a judgement task sufficiently represents the child's phonological awareness and morphosyntactic processing skills. This might also be relevant to consider when further investigating the negative association that emerged between the German word identification and Italian phoneme blending judgement tasks and the positive association between number of word reading errors in the DDE-2 and accuracy (%) in the Italian phoneme blending judgement task through item analyses and the thorough assessment of ceiling and floor effects. This adds to the observation that fast readers take more time in the case matching task and/or slow readers respond faster in the case matching task. The above mentioned observations point to the necessity of examining the data for particular patterns on a thorough item-based analysis and in particular – where possible – analyses of response time only for the items in which the target stimulus was correctly identified. This could give interesting insights regarding the nature of DD and how slow reading time is an indicator for an impairment, but poor readers might also read fast and incorrectly. Along these lines, it is important to take both reading time and accuracy into account

in the diagnostic processes of DD. Regarding the use of caregiver questionnaires in diagnostic processes, the finding that the questionnaire responses are associated mostly with the German and not Italian standardized and screening tasks might reflect that either the questions from the questionnaire or the caregivers' impressions are mostly focusing on the societal language (language of schooling). Especially in this particular group of children attending bilingual Italian-German schools in Germany, it would be very useful to further analyse these associations taking language exposure patterns into account. The notion that teachers' judgements of the children's language and reading skills – based on their observations in every day school life and a series of tests and exams at school – are associated with the children's performance in the screening tasks shows that these tasks are a useful and timesaving resource, grasping aspects of children's language and literacy competences. Due to the small amount of children ($n = 3$) who had been identified with reading difficulties, their performance in L1 and L2 screening tasks assessing the same linguistic area were not statistically assessed separately (Hypothesis 5). However, the general associations between risk score and screening task performance (Hypothesis 1) does suggest the screening's potential in DD risk identification.

5.6 Computerized, remote DD screening for bilingual children living in Italy

As mentioned in chapter 2.3.2, China is among the most frequent countries of origin for foreigners residing in Lombardy and Tuscany. Accordingly, an Italian-Mandarin DD screening was constructed and applied with children remotely. Furthermore, due to the orthographic constraints and the importance of English as lingua franca, also an English DD screening was constructed and administered remotely. With the focus on the assessment of the applicability and usability of remote computerized screenings, the Italian part of the screenings was also administered to monolingual children. A modified version of this chapter has been published in the peer-reviewed journal article "Remote Dyslexia Screening for Bilingual Children" published in the open-access journal "Multimodal Technologies and Interaction" (6, 7) (Eikerling et al., 2022b).

In this study, screenings for the risk identification of DD were carried out exclusively remotely with both mono- and bilingual children living in Italy. In addition to the caregiver and teacher questionnaires, the standardized tests and the web-based screening, a further online questionnaire on the usability of the screening tool was administered to both the monolingual Italia-speaking children who were tested as well as to a group of examiners who have administered MuLiMi screenings.

5.6.1 Hypotheses

In addition to the hypotheses and research questions described in chapter 5.1 and answered individually for each of the screening studies, in this study, a research question on the screening platform's usability was added:

RQ 3: In usability studies, both target user groups, the examinees (child participants) and the examiners (screening administrators) perceive the screening platform to sufficiently respond to the requirements, to be enjoyable in use and to meet their expectations.

5.6.2 Material & methods

To answer these research questions, the methods and material comparable to the ones described in chapters 4.1.2.1 and 5.5.1.2 were applied in this study.

5.6.2.1 *Participants*

Thirty children participated in this study who were all attending either the last months of grade two or grade three to five at a primary school in Italy. The children thus covered an age span of around three years (7 to 10 years of age, mean age in months: $M = 101.57$, $SD = 15.86$). While $n = 11$ children were monolingual Italian, a total of $n = 19$ children were multilingual. Among these children, $n = 12$ children spoke English both in their families (L1) and attending bilingual English-Italian schools while $n = 7$ children attending mainstream Italian public schools spoke Mandarin in their families (L1). None of these children was known to have DD, but based on the children's performance in the standardized reading tests, risk scores were calculated (see chapter 5.6.2.6).

The participants of the usability studies consisted of a) the $n = 11$ monolingual children who participated in the aforementioned screening study and b) $n = 10$ examiners who had administered screenings using the MuLiMi web app remotely at least once.

5.6.2.2 *Screening tasks*

Again, for all three languages relevant to this screening study, the four relevant areas of reading assessment RAN, reading speed and accuracy, phonological awareness and grammatical skills were examined.

In addition to the Italian reading screening tasks that were already described in chapter 5.5.1.2 for the Italian-German screening, some more tasks were administered to the group described in this screening studies considering that the child participants are attending primary schools in Italy and thus the emphasis of reading and writing acquisition is on the Italian language. All audio clips were pre-recorded by a native speaker with natural voice and accent. For the description of the Italian self-paced syllable reading task, the self-paced sentence

reading, the word identification, the NW identification, the phonological blending, the syllabic inversion, subject-verb agreement and clitic pronoun judgement tasks used in these screenings, see description in chapter 4.1.2.1. Find examples in appendix B.

Word stress identification. A pre-recorded three-syllabic word is presented auditorily. At the same time, the same word segmented into the three syllables is displayed on the screen. Children are asked to click on the stressed syllable, for example, they listened to the word “lampada”, [ˈlampada] (lamp), saw the three syllable “lam”, “pa” and “da” (see figure 44) and were expected to click on the “lam”. The task consists of 2 training and a total of 8 screening items. Response time and accuracy is automatically measured and stored. Find more examples in appendix B.

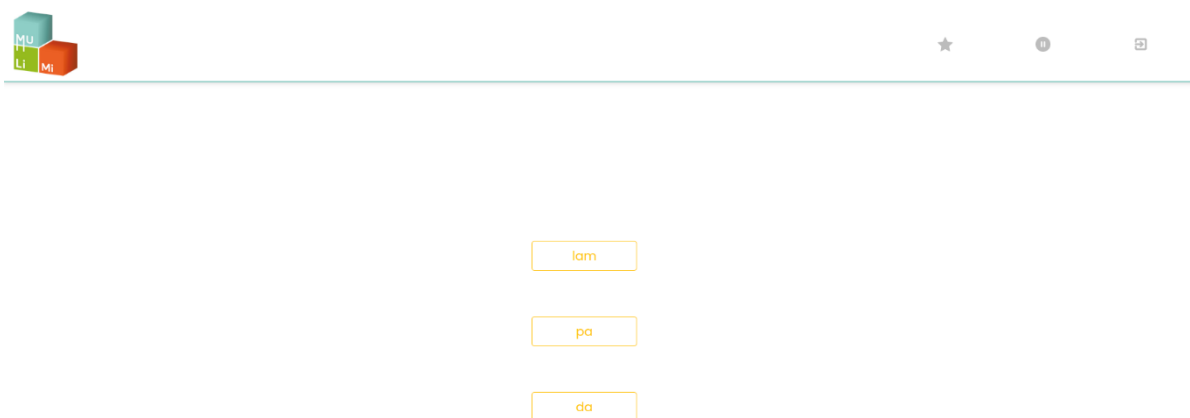


Figure 44: Examinee interface during the Italian word stress identification task.

The nature and structure of the English reading screening tasks both differ from the Italian ones due to the difference in orthographic depth, but screening structure and task construction principles are maintained. All audio clips were pre-recorded by a native-like speaker with natural voice and accent.

RAN. The description of the Italian version of this task (see chapter 5.5.1.2) applies also to its English version, in which English-speaking children are asked to name the digits that appear on the screen as fast as possible in English. The task consists 5 training and a total of 30 screening items. Self-paced naming speed is automatically measured and stored, due to ceiling effects in pilot studies, accuracy was not tracked.

Self-paced sentence reading. The description of the German version of this task (see chapter 5.5.1.2) applies also to its English version, in which English sentences increasing in complexity are used. The task consists of 1 training and a total of 5 screening items. Find more examples in appendix B.

Orthographic form identification. A pre-recorded sentence is played to the school-aged child participant, e.g. “Every girl will dress up as a witch.”. The orthographic form of the last word of this sentence is displayed on the screen (“witch” - [wɪtʃ]) along with two distractors varying in spelling. Across items, one of the distractors consists in an existing word pronounced in the same way as the target but spelled differently (“which” - [wɪtʃ]), while another distractor is a non-existing word that could be pronounced in the same way as the target (“whitch” - [wɪtʃ]). The child is asked to indicate which one of the three orthographic forms displayed on the screen is the correct one to use in the context of the sentence that was played by selecting the corresponding button. The task consists of 3 training and a total of 9 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Phonological form identification. A written sentence is visually presented to the child participant on the screen. One of the words displayed is highlighted, e.g. “In class, sometimes teachers project slideshows.”, highlighted word: “project” [prəˈdʒekt]. Simultaneously, three buttons are displayed below the sentence on the screen. The buttons flash one after another, while flashing one pre-recorded word per button is presented auditorily. One of those words represents the phonological form of the highlighted word of the written sentence displayed (target), the other two are distractors, e.g. [ˈproject] (variation in word stress: [ˈprɛdʒɛkt] and [proˈtɛkt] (substitution of a phoneme: [prɛˈtɛkt]). The child is asked to indicate which one of the three phonological forms presented is the correct one to use in the context of the written sentence displayed by selecting the corresponding button. The task consists of 3 training and a total of 8 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Word stress identification. The description of the Italian version of this task (see above) applies also to its English version, in which pre-recorded English bisyllabic words are used. The task consists of 3 training and a total of 8 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find examples in appendix B.

Sound deletion. A pre-recorded question on the process of sound deletion of a certain word is presented auditorily, e.g. “What word would be left if the “b”-sound was taken away

from “block”?”. Then the target answer (“lock” [lɔk]) and one distractor (“bock” [bɔk]) are subsequently presented auditorily in random order accompanied with visual support (two different figures that seem to be saying either the target or distractor word, see WSIR screening tasks in chapter 5.3.2.2). The child is asked to indicate which of the two words presented is the one resulting from the introductory question by selecting the corresponding figure. The task consists of 2 training and a total of 10 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Tense judgement. Also this task is similar to the WSIR subject-verb agreement and finiteness tasks in the Spanish-Italian screening described in chapter 5.3.2.2. Again, upon presentation of the pre-recorded question “Who says it right?” one correct (“Last summer it rained a lot.”) and one incorrect sentence (“Last summer it rains* a lot.”) are subsequently presented auditorily in random order accompanied with visual support (two different figures that seem to be saying one of the sentences each). The child is asked to indicate which of the two sentences presented is correct by selecting the corresponding figure. The task consists of 2 training and a total of 12 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

The nature and structure of the Mandarin reading screening tasks again differs from the Italian ones due to the logographic nature of Mandarin script (see chapter 2.4.2.1), but screening structure and task construction principles are maintained as much as possible. All audio clips were pre-recorded by a native-like speaker with natural voice and accent.

RAN. The description of the Italian version of this task (see above) applies also to its Mandarin version, in which Mandarin-speaking children are asked to name the digits that appear on the screen as fast as possible in Mandarin. The task consists of 5 training and a total of 30 screening items. Self-paced naming speed is automatically measured and stored, due to ceiling effects in pilot studies, accuracy was not tracked.

Left-right inversion and radical position judgement. See chapter 4.1.2.1. for the task descriptions and figure 45 for the screening task implementation on MuLiMi. Find examples in appendix B.

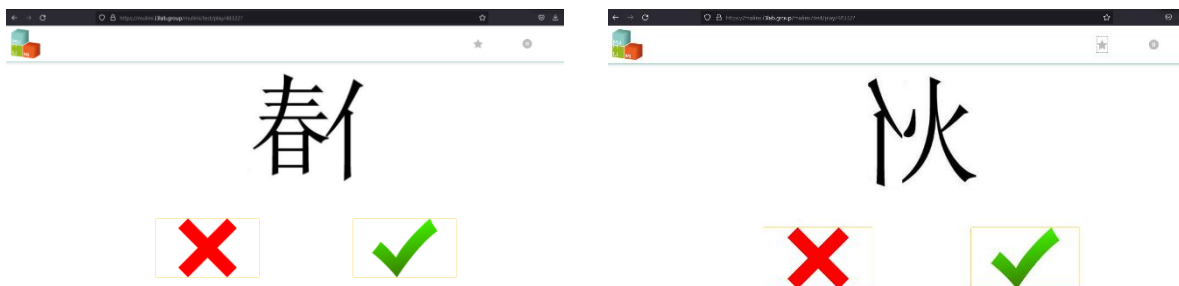


Figure 45: Examinee interface during the Mandarin radical position (left) and left-right inversion (right) tasks.

Tone detection. In each trial, four buttons are displayed on the screen (see figure 46). Those buttons flash one after another while for each button flashing, one of four pre-recorded monosyllabic Mandarin words are presented auditorily. In this task, three of them had the same tone (tán, luó, lán) while one of them deviated (huā). The child is asked to indicate which one of the four words presented auditorily deviates in tone by selecting the corresponding button. The task consists of 2 training and a total of 8 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

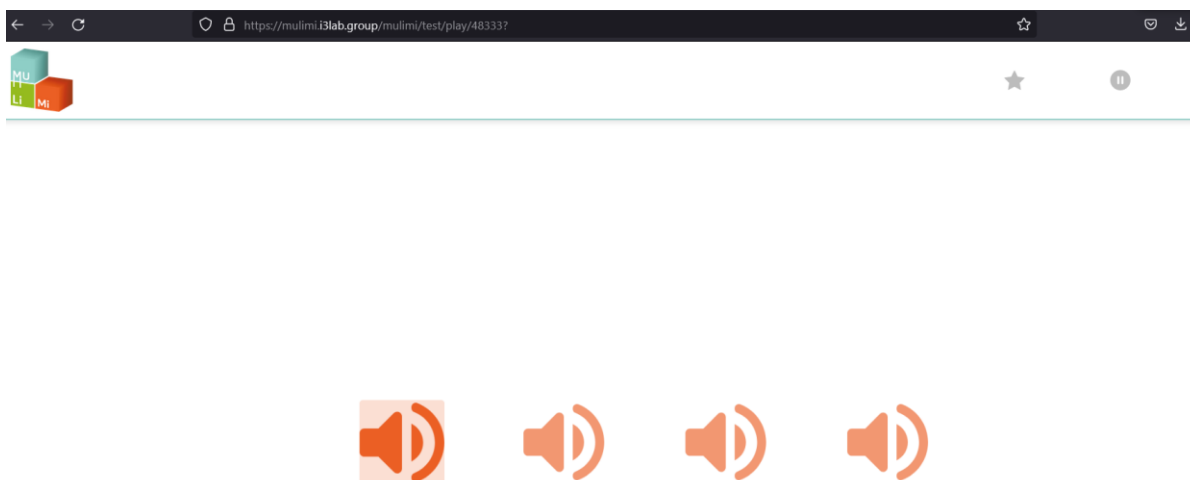


Figure 46: Examinee interface during the Mandarin phonological awareness tasks.

Onset detection. For the task structure see above in the description of the tone detection screening task. In this task however, three monosyllabic of the Mandarin words presented auditorily had the same onset (tán, tǐng, téng) while one of them deviated (luó). The child is asked to indicate which one of the four words presented auditorily deviates regarding its onset

by selecting the corresponding button. The task consists of 2 training and a total of 16 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

Rhyme detection. For the task structure see above in the description of the tone detection screening task. In this task however, three monosyllabic Mandarin words had the same rhyme (tǎng, dǎng, láng) while one of them deviated (qíng). The child is asked to indicate which one of the four words presented auditorily deviates in rhyme by selecting the corresponding button. The task consists of 2 training and a total of 12 screening items. Response time and accuracy of the given responses are automatically measured and stored. Find more examples in appendix B.

5.6.2.3 Standardized/traditional reading tests

Since the children were tested exclusively remotely, all standardized tests used had to be adapted for remote assessment to be displayed on the child's screen in a comparable manner across participants. The reading material was displayed to the child through screen share activated by the examiner in a video call after the examiner had verified that the size of the version of the word lists displayed was comparable to the original size of the characters in the standardized test.

Batteria per la Valutazione della Dislessia e della Disortografia Evolutiva-2 (DDE-2, Sartori et al., 2007). Also in this study, both the word- and NW-reading subtests of the DDE-2 (Sartori et al., 2007) and norm data provided for monolingual Italian-speaking children was used. Accordingly, also in this study the evaluation of the child's reading performance was based on the instructions given in the test manual and was carried out by native-speakers of Italian.

Test of Word Reading Efficiency-Second Edition (TOWRE-2, Torgesen et al., 2012). Also English standardized word- and NW-reading subtests were conducted. Again, children are presented both words and NWs separately in vertical lists on the child's screen. Here instead, they are asked to read as many of them as possible within 45 seconds for words and NWs each and interrupted after time is up by the examiner.

Furthermore, nonverbal intelligence was tested by the means of the *CPM Raven's Coloured Progressive Matrices* (Belacchi et al., 2008). or the description of the test, see chapter 5.3.2.3.

Standardised test results were evaluated according to the criteria in the respective manuals and the norm data provided in the latter. Furthermore, the raw scores were converted into percentages to facilitate comparison with the results of the experimental tests.

Chinese reading test (Hu, in preparation). This unpublished test is especially designed for Mandarin-speaking children living in Italy and acquiring oral Mandarin in their families and written Mandarin in weekend schools. It consists of 150 Mandarin words, that are each made up of two characters and were selected from textbooks used in Chinese schools in Italy. In the test, they are ordered with an increase in difficulty. Those characters are displayed in vertical lists on the child's screen. For this study, the evaluation of the child's reading performance in this test was carried out by a speaker of Mandarin (checked by a native Mandarin-speaker) and does not take reading speed but accuracy only into account. 1 point is assigned when the whole word (both characters) are correctly read. When only one of the two characters is read correctly instead, 0.5 points is assigned. Total reading performance as measured in this test is displayed in raw scores (percent accuracy) since norm data has not been published yet.

5.6.2.4 Usability questionnaire

In order to assess the satisfaction with the design, functionalities and application of the MuLiMi screening platform, usability questionnaires for both child participants tested remotely (examinees) as well as clinicians and graduate students in psychology and SLT who were involved in the application of MuLiMi screenings for scientific purposes in 2021 (examiners) were created. Questionnaire items that are relevant for the assessment of usability of the MuLiMi screening platform for examinees and examiners were selected from the most commonly applied usability questionnaires available, i.e. the "System Usability Scale" (SUS, Sauro, 2022) and two different versions of the Questionnaire for User Interface Satisfaction (QUIS, Chin et al., 1988; Wallace et al., 1988). In a second step, these were adapted for their application in the context of usability judgement of the MuLiMi screening platform, translated into Italian and implemented on "Google Forms" (<https://docs.google.com/forms/>) as online questionnaire. According to the specific needs of examiners and examinees, two different versions of the usability questionnaire were created.

Examiner usability questionnaire. The usability questionnaire for examiners consists of 46 questions and addresses questions on the examiners' perceptions and opinions regarding the application of MuLiMi remotely in clinical practice with pre- and primary school children (see complete list of English version of questionnaire items in appendix A with indexed with

an “E” for examiner). Questionnaire items are statements followed by two more remarks representing the degree of agreement on the two extremes of a five- to nine-point scale, that the examiner chooses from.

Examinee usability questionnaire. From the previously described usability questionnaire designed for examiners, 12 items were selected, adapted for language style and used to assess the examinees’ opinions regarding their experience with MuLiMi remote reading screenings (see complete list of English version of questionnaire items in appendix A with indexed with a “P” for child participant). In order to reduce the complexity of the questionnaire tool in its version for the examinees, the questionnaire items presented (statements) are followed by remarks representing the degree of agreement on the two extremes on a 5-point Likert scale, instead of a nine-point scale (see above).

5.6.2.5 Procedure

While all of the components of the direct assessment fit in one testing session lasting 90 minutes (10-minute break in between) for monolingual children who participated in this study, two testing sessions lasting 45 to 60 minutes each were needed for children speaking Italian as societal language in addition to either English or Mandarin due to the additional standardized/traditional reading tests and L1-screening tasks used. Both the standardized/traditional reading tests (through screen share, see chapter 5.6.2.3) and the MuLiMi reading screening (by sharing of the link directly through the MuLiMi platform, see chapter 5.2.2) were carried out entirely remotely, so examinees were connected to the examiner via video conferencing tools allowing for screen share for the standardized/traditional test administration and voice communication during the screening task administration. Depending on the preference of the institutions involved in the recruitment procedure, either “zoom” (<https://zoom.us/>) or “google meet” (<https://meet.google.com/>) were used as video conferencing tools. For the simulation of a realistic testing scenario, both examiners and examinees used their own work, personal or school laptops or desktop PCs depending on whether they were tested or ran the testing session from work, home or school (according to their preferences and feasibility). Due to technological constraints and comparability of screen size for the screening and standardized test visualisation, participation to the study from tablet or smartphone was not possible.

Also in this study, questionnaires to be filled in by caregivers were used. According to the preferences expressed by the schools regarding the collection of filled-in caregiver questionnaires, caregivers of Mandarin-speaking study participants filled in a pen-and-paper version of the DSA questionnaire, while caregivers of monolingual and English-speaking caregivers responded to the questions in its online version implemented on “Google Forms”

(<https://docs.google.com/forms/>). Links to the online questionnaire were shared with the examinees' caregivers outside the screening session. Caregivers of the multilingual study participants could choose between the Italian version of the DSA questionnaire and its translation into the other language spoken in the children's home. The teacher questionnaire (see chapter 5.2.4.1) was filled in exclusively by the teachers of the multilingual children participating in this screening study in its pen-and-paper version.

While the online usability questionnaire for monolingual examinees was short and could thus be filled in immediately after screening administration, the examiners were asked to fill in a more detailed version, after minimum one screening administration. Based on the results of the usability study with examiners, a follow-up questionnaire was spread among them for an in-depth investigation and facilitation of interpretation of the responses given.

5.6.2.6 *Risk score creation*

For each study participant, a risk score was generated based on the standardized/traditional reading test results in all languages spoken in order to be able to compare the screening results to this score irrespective of the presence of a diagnosis or misdiagnosis of DD.

In a first step, with the aim of obtaining a risk score representing the children's reading performance in Italian, the z-scores obtained in the DDE-2 on both accuracy and reading speed were converted according to whether a z-score in one of the two subtests is (+1) or is not (0) at or below minus two standard deviations. To generate the DDE-2 risk score, the number of times a child had scored at or below two standard deviations in the DDE-2 was summated.

For the multilingual children participating in this study a further risk score was created. Besides the DDE-2 risk score, a second reading risk score was created based on the standardized/traditional reading tests in the children's family languages English (TOWRE-2, Torgesen et al., 2012) or Mandarin (Chinese reading test, Hu, unpublished). Equivalently to the transfer of the DDE-2 z-scores into the DDE-2 risk score, the z-scores the English-speaking children obtained in the TOWRE were converted into the L1 risk score. Since the Chinese reading test does not provide norm data, a risk of 1 was assigned to Mandarin-speaking children who were not able to read more than a third (33.33%) of the Chinese words. In a third step, the reading compound risk score (ranging from 0 to 4) based on the sum of those two risk scores described above was calculated.

5.6.3 Results & discussion

Data analysis relied on the general constraints described in chapter 5.2.6. Whenever significant associations between the children's performance in the Raven's CPM and a certain screening task emerged, analyses were run again with the CPM score as control variable when the sample size allowed for running parametric tests. Each time the association did not remain stable, this is reported.

5.6.3.1 Comparison of Italian screening results and the DDE-2

Since all children who participated in this study learn to read and write in Italian schools, they were all assigned both the DDE-2 (Sartori et al., 2007) and the Italian screening tasks.

Reading time. DDE-2 total word reading time was significantly associated with mean reading time per item in the self-paced syllable ($n = 27$, $r = .461$, $p = .015$) and in the sentence screening task ($n = 27$, $r = .942$, $p < .001$). The same pattern was revealed comparing the DDE-2 total NW reading time with mean reading time per item in the self-paced syllable ($n = 27$, $r = .406$, $p = .036$) and sentence screening task ($n = 30$, $r = .942$, $p < .001$). Response time in the Italian word identification screening task was also significantly associated with total reading time of words ($n = 30$, $r = .531$, $p = .003$) and NWs ($n = 30$, $r = .582$, $p = .001$) in the DDE-2, while the response time in the Italian NW identification screening task was not ($p > .05$). This general pattern was confirmed when comparing DDE-2 and screening task results for each language group individually with smaller sample sizes and lower significance levels and is exemplified in the figure 47.

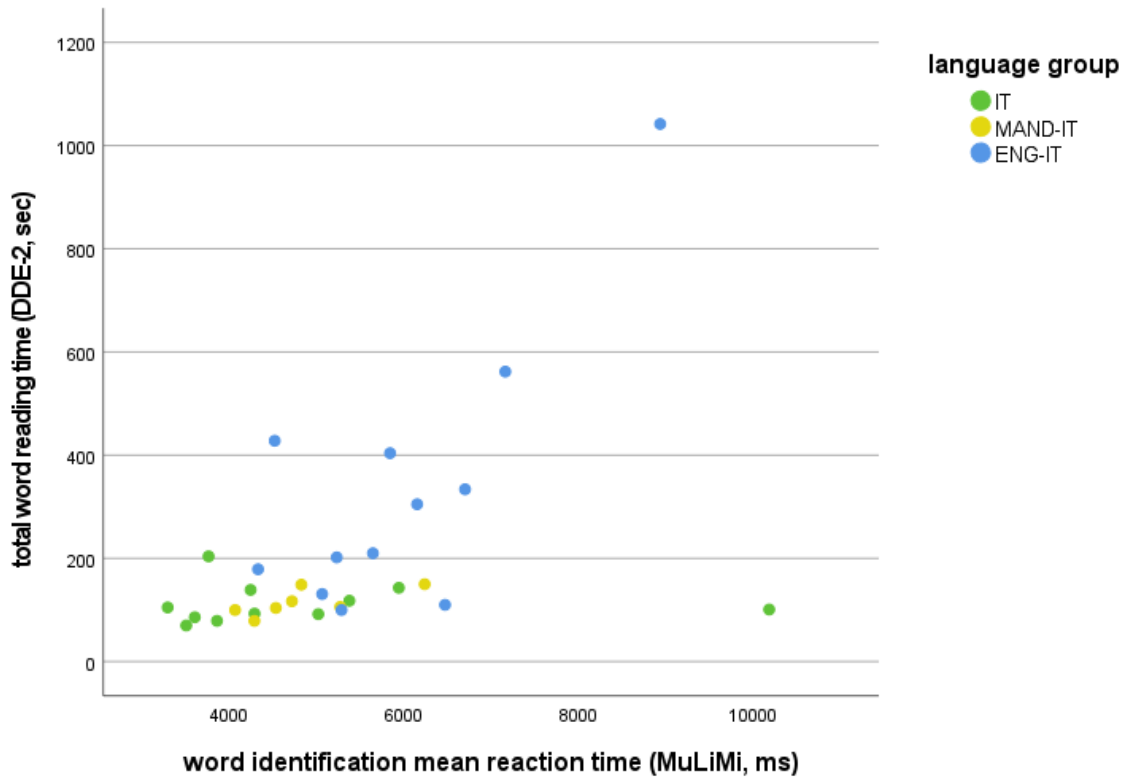


Figure 47: Comparison of total word reading time (sec) for Italian words (DDE-2, y-axis) and response time (ms) in Italian word identification screening task (x-axis) according to language groups.

Reading accuracy. Word identification accuracy (%) in the Italian word matching screening task correlated significantly with reading accuracy of words in the DDE-2 ($n = 30$, $r = .880$, $p = .001$, with a $R^2 = .774$). Also this pattern was confirmed when analysing the associations for each of the three language groups individually which is exemplified in figure 48. Similarly, NW identification accuracy (%) in the Italian NW identification screening task was significantly associated with reading accuracy of NWs in the DDE-2 ($n = 30$, $r = .569$, $p = .001$).

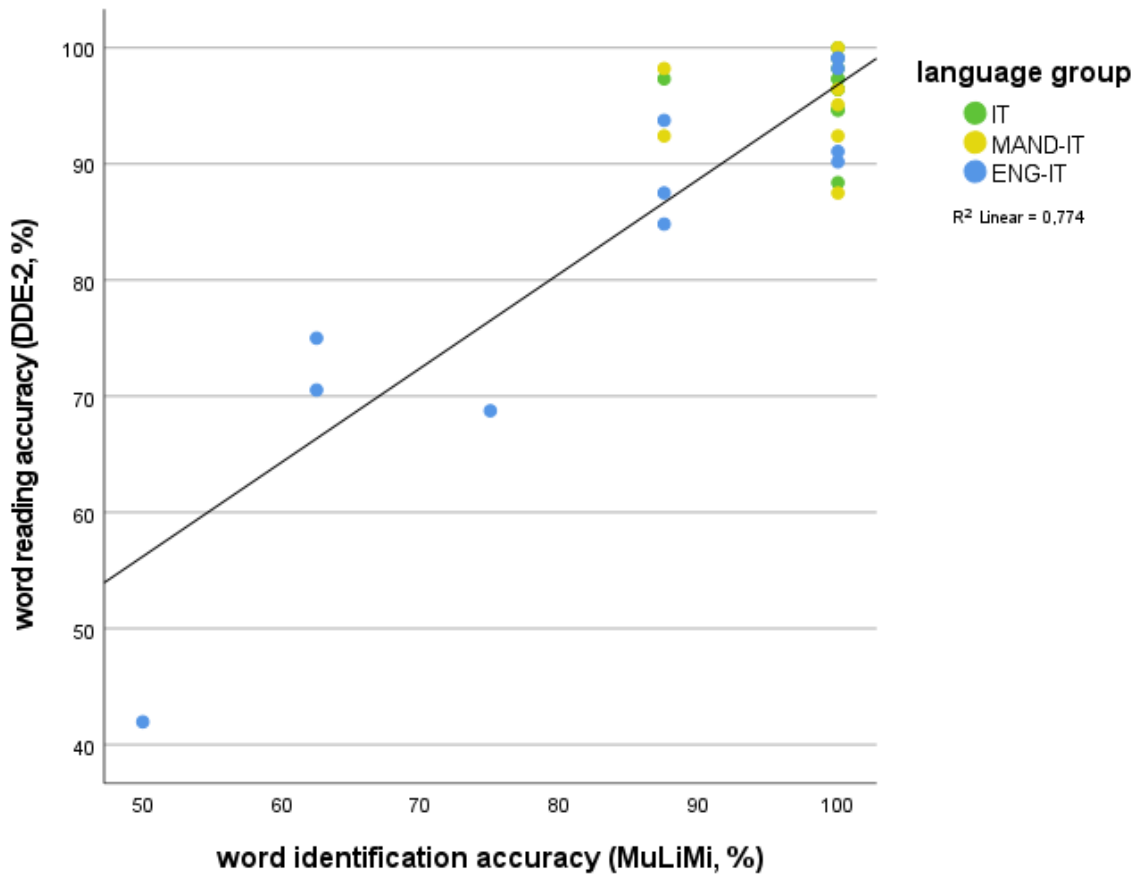


Figure 48: Comparison of word reading accuracy (%) for Italian words (DDE-2, y-axis) and accuracy (%) in the Italian word identification screening task (x-axis) according to language groups.

Phonological awareness. The phonological blending judgement accuracy (%) in the Italian screening task was significantly associated with both word ($n = 30$, $\rho = .492$, $p = .006$) and NW ($n = 30$, $\rho = .356$, $p = .053$) reading accuracy in the DDE-2. No significant associations instead were found between the DDE-2 subtests and the accent identification ($n = 27$, $\rho_s < .333$, $p_s > .05$) or syllabic inversion judgement screening tasks (carried out with monolingual Italian- and Mandarin-speaking children only, $n = 18$, $\rho_s < .052$, $p_s > .05$).

Clitic pronoun judgement. The clitic pronoun judgement accuracy (%) in the Italian screening task correlated significantly with word reading accuracy ($n = 30$, $\rho = .513$, $p = .004$), but not NW reading accuracy ($n = 30$, $p > .05$) in the DDE-2. Re-running these analyses for the three language groups individually, significant associations between the clitic pronoun judgement screening task and both word ($n = 12$, $\rho = .713$, $p = .009$) and NW ($n = 12$, $\rho = .676$, $p = .016$) reading accuracy in the DDE-2 emerged for the English-Italian subgroup.

5.6.3.2 Comparison of L1 screening task and the standardized/traditional test results

Associations between the scores obtained in L1 standardized/traditional reading tests and screening tasks were computed for each language group separately.

English screening. Reading time measured in the self-paced sentence reading screening task was significantly associated with mean reading time for words ($n = 12$, $\rho = .888$, $p < .001$) and NWS ($n = 12$, $\rho = .690$, $p = .013$) in the TOWRE-2. Performance in both aforementioned TOWRE-2 subtests were also significantly correlated with performance in the orthographic form identification screening task (words: $n = 12$, $\rho = .713$, $p = .009$; NWS: $n = 12$, $\rho = .701$, $p = .011$). Significant associations were also found comparing word reading accuracy as measured in the TOWRE-2 with accuracy (%) in the orthographic form identification screening task ($n = 12$, $\rho = .748$, $p = .005$). None of the English phonological awareness tasks (neither sound deletion nor word stress identification) was shown to be significantly associated with reading time or accuracy in the TOWRE-2 ($p > .05$). Word reading accuracy in the TOWRE-2 instead correlated significantly with accuracy (%) in the tense judgement screening task ($n = 12$, $\rho = .740$, $p = .006$).

Mandarin screening. Non-significant associations emerged from the comparison of the amount of correctly read words in the Chinese reading test (%) and accuracy (in %) in the Mandarin screening tasks left-right inversion ($n = 7$, $\rho = .574$, $p = 0.178$) and radical position judgement ($n = 7$, $\rho = .574$, $p = 0.178$). No significant associations or associations close to significance were found comparing the children's performance in the Chinese reading test to performance (accuracy in %) in the screening tasks on phonological awareness.

5.6.3.3 Comparison of screening results and risk level

Since the performance of multilingual Italian-speaking children in the screening tasks – other than for the monolingual Italian-speaking children – cannot be directly compared to the norm data from the Italian standardized reading test, a risk score representing standardized/traditional reading test results in both languages (see chapter 5.6.2.6) was compared to the screening results of multilingual children ($n = 19$).

Italian screening tasks. Similar to the associations found across language groups between the Italian screening tasks and DDE-2 results, the L1 risk score was significantly associated with reading time in the Italian self-paced sentence reading screening task ($n = 19$, $\rho = .473$, $p = .041$) as well as with accuracy (%) in the word identification screening task ($n = 19$, $\rho = -.500$, $p = .029$). Also the compound risk score, incorporating both the reading performance in the L1 and the Italian standardized/traditional tests were found to be significantly correlated with response time in the word identification screening task ($n = 19$, $\rho =$

.457, $p = .049$) and response time in the clitic pronoun judgement screening task ($n = 19$, $\rho = -.531$, $p = .019$). Further non-significant associations were found between the compound risk score and the self-paced sentence reading time ($n = 19$, $\rho = .428$, $p = .068$) and accuracy (%) in the NW identification screening task ($n = 19$, $\rho = -.406$, $p = .085$).

English screening tasks. Both the compound risk score and the L1 risk score deriving from the English standardized reading test TOWRE-2 only were found to be associated with the screening task performance. In particular, reading time in the English self-paced sentence reading task was associated with the L1 ($n = 12$, $\rho = .717$, $p = .009$) and the compound risk score, though not significantly ($n = 12$, $\rho = .566$, $p = .055$). Further associations emerged between the L1 risk score and the accuracy (%) in the orthographic form identification ($n = 12$, $\rho = -.570$, $p = .053$), the phonological form identification ($n = 12$, $\rho = -.493$, $p = 0.104$) and the tense judgement screening task ($n = 12$, $\rho = -.781$, $p = .003$). The L1 risk score was also shown to be associated, though not significantly, with response time in the phonological form identification task ($n = 12$, $\rho = -.512$, $p = .089$) and in the sound deletion judgement screening task ($n = 12$, $\rho = .512$, $p = .089$).

Mandarin screening tasks. Also in the group of Mandarin-speaking children, the compound risk score was found to be associated with the performance in the Mandarin screening tasks, in particular significantly associated with accuracy (%) in both the left-right inversion ($n = 7$, $\rho = -.882$, $p = .009$) and the radical position judgement screening task ($n = 7$, $\rho = -.882$, $p = .009$) as well as with response time in the onset detection task ($n = 7$, $\rho = -.886$, $p = .012$). Furthermore, the compound score was associated (though non-significantly) with response time in the judgement of radical position ($n = 7$, $\rho = -.577$, $p = 0.175$), rhyme ($n = 7$, $\rho = -.577$, $p = 0.175$) and tone detection ($n = 7$, $\rho = -.577$, $p = 0.175$) as well as RAN ($n = 7$, $\rho = .577$, $p = 0.175$). A comparable pattern was revealed comparing these screening task performances to the L1 risk score.

5.6.3.4 Comparison of screening results and caregiver and teacher questionnaires

Caregiver questionnaire. The number of caregiver's positive ("yes") responses to all questions on the child's problems with reading and writing acquisition were significantly associated with accuracy (%) and response time in several screening tasks in Italian. The aforementioned score correlated significantly with response time in the self-paced sentence reading ($n = 30$, $\rho = .568$, $p = .001$), the word ($n = 30$, $\rho = .399$, $p = .029$ with $R^2 = .318$, no longer significant when inserting the children's performance level in the CPM as control variable: $p > .05$) and NW identification ($n = 30$, $\rho = .426$, $p = .019$) as well as in the blending judgement ($n = 30$, $\rho = .580$, $p = .001$) screening task. The same score was also significantly associated with

accuracy (%) in the word ($n = 30$, $\rho = -.625$, $p < .001$) and NW identification ($n = 30$, $\rho = -.477$, $p = .008$), the blending judgement ($n = 30$, $\rho = -.427$, $p = .018$) and the screening tasks on grammaticality judgement for subject-verb agreement ($n = 19$, $\rho = -.728$, $p = .001$) and clitic pronouns ($n = 30$, $\rho = -.482$, $p = .007$). Italian word identification accuracy (%) was found to be associated with both the compound score of questions related to school discomfort ($n = 30$, $\rho = -.372$, $p = .043$) and mathematical problems ($n = 30$, $\rho = -.535$, $p = .002$), but when inserting the children's performance in the Raven CPM as control variable, these associations no longer were found to be significant ($p_s > .05$). Further associations with screening task performance were found with the compound scores of general learning difficulties in particular with accuracy (%) in the Italian NW identification ($n = 30$, $\rho = -0.371$, $p = .044$) and with response time in the clitic pronoun grammaticality judgement task ($n = 30$, $\rho = 0.384$, $p = .036$). A comparable pattern was found when analysing all three subgroups individually (see figure 49).

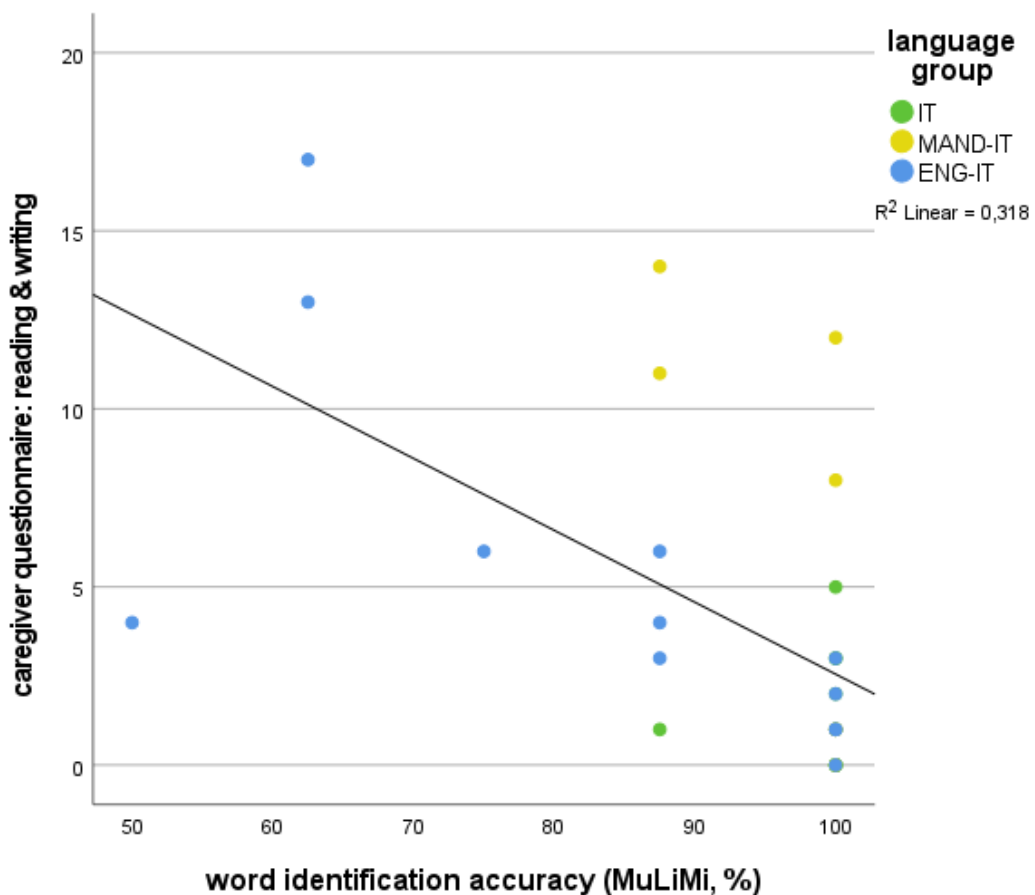


Figure 49: Comparison of caregivers' evaluation of children's reading and writing performance (DSA questionnaire) and accuracy (%) in the Italian word identification screening task (x-axis) according to language groups. Note that all $n = 30$ participants were included in these analyses, but due to overlapping data points are not visualized here.

Teacher questionnaire. Teachers' judgement on children's language and reading skills were obtained for multilingual children only and were exclusively significantly associated with children's accuracy (%) in the Italian word identification screening task. In particular, significant associations with word identification accuracy were found for the teachers' judgements on children's phonological receptive ($n = 17$, $\rho = -.482$, $p = .050$) and productive phonological ($\rho = -.482$, $p = .050$) as well as for receptive morphological ($\rho = -.494$, $p = .044$) and receptive vocabulary skills ($\rho = -.494$, $p = .044$), but these associations were no longer present when the children's performance in the CPM Raven was used as control variable ($p_s > .05$).

5.6.3.5 Comparison of screening results within and across languages

Several associations within the same language were found when comparing child participants' performances in different screening tasks to each other.

Namely for Italian, self-paced sentence and syllable reading average reading were significantly associated ($n = 27$, $r = .475$, $p = .012$). Also average response time in the Italian word and NW identification screening tasks correlated significantly with each other ($n = 30$, $r = .765$, $p < .001$). Children's mean response time in the word, but not NW identification screening tasks was significantly associated with average self-paced sentence ($n = 30$, $r = .471$, $p = .009$), but not syllable reading time ($p > .05$). Self-paced syllable reading time was also not significantly correlated with the mean response time in the NW identification screening task ($p > .05$).

Also accuracy (in %) in the Italian word identification screening task was significantly associated with accuracy in the NW identification task ($n = 30$, $\rho = .551$, $p = .002$). The accuracy (in %) in the two Italian screening judgement tasks on phonological awareness (phoneme blending and syllabic inversion) did not correlate significantly with each other ($p > .05$), but both were significantly associated with accuracy in the subject-verb agreement judgement task (phoneme blending: $n = 18$, $\rho = .477$, $p = .045$; syllabic inversion: $n = 18$, $\rho = .496$, $p = .036$). Accuracy in the blending task was significantly correlated also with the Italian clitic pronoun judgement tasks ($n = 30$, $\rho = .519$, $p = .003$, but when the children's performance in the CPM Raven was used as control variable, this association was no longer significant $p > .05$). Furthermore, accuracy (in %) in the clitic pronoun judgement task was significantly associated with accuracy in the subject-verb agreement judgement task ($n = 18$, $\rho = .769$, $p < .001$).

Associations within languages were also found for the English and Mandarin screening tasks. While response time in the Mandarin version of the RAN screening task was not

significantly associated with any of the other Mandarin screening tasks ($p_s > .05$), a significant association emerged for accuracy in the left-right inversion and the radical position judgement task associated ($n = 7$, $\rho = .963$, $p < .001$). With the exception of a significant association between the mean response time in the Mandarin rhyme and tone detection tasks ($n = 7$, $\rho = .857$, $p = .014$) no significant associations emerged for the Mandarin phonologic awareness screening tasks.

For the English screening tasks, mean self-paced sentence reading time was significantly associated with performance in the orthographic form identification tasks (accuracy: $n = 12$, $\rho = -.811$, $p = .001$; response time: $n = 12$, $\rho = .790$, $p = .002$), accuracy (in %) in the phoneme deletion task ($n = 12$, $\rho = -.733$, $p = .007$) as well as performance in the tense judgement screening task (accuracy: $n = 12$, $\rho = -.750$, $p = .005$; reaction: $n = 12$, $\rho = -.580$, $p = .048$). Accuracy in the English orthographic form identification screening task was significantly associated with mean response time in the same tasks ($n = 12$, $\rho = -.814$, $p = .001$) as well as accuracy in the phoneme deletion ($n = 12$, $\rho = .787$, $p = .002$) and tense judgement ($n = 12$, $\rho = .764$, $p = .004$) screening tasks. Mean response time in the English orthographic form identification screening task was also significantly, but negatively associated with mean response time in the English tense judgement task ($n = 12$, $\rho = -.650$, $p = .022$).

Furthermore, crosslinguistic associations emerged comparing the Italian screening task performance to performance in the English and Mandarin screening tasks. For the group of English-speaking children, several significant associations between screening task performance in the English and Italian screening tasks were found. Self-paced reading time of English sentences was significantly associated with Italian self-paced syllable ($n = 12$, $\rho = .601$, $p = .039$) and sentence ($n = 12$, $\rho = .909$, $p < .001$) reading time. Significant associations also emerged comparing the English self-paced sentence reading time to response time in the Italian NW identification ($n = 12$, $\rho = .776$, $p = .003$) and in the blending judgement tasks ($n = 12$, $\rho = .608$, $p = .036$). Furthermore, self-paced reading time of English sentences was significantly correlated with accuracy (on %) in the Italian word identification ($n = 12$, $\rho = -.706$, $p = .010$) and clitic pronoun judgement ($n = 12$, $\rho = -.788$, $p = .002$) screening tasks. Accuracy in the orthographic form identification task was significantly associated with accuracy in the Italian word identification task ($n = 12$, $\rho = .708$, $p = .010$) and in the clitic pronoun judgement task ($n = 12$, $\rho = .763$, $p = .004$). Accuracy in the orthographic form identification task also correlated significantly with reading time in the Italian self-paced syllable ($n = 12$, $\rho = -.612$, $p = .034$) and sentence reading task ($n = 12$, $\rho = -.680$, $p = .015$) as well as with response time in the Italian blending judgement task ($n = 12$, $\rho = -.666$, $p = .018$). Also

response time in the English orthographic from identification task was shown to be significantly associated with response time in Italian reading screening tasks (Italian self-paced syllable reading: $n = 12$, $\rho = .832$, $p < .001$; Italian self-paced sentence reading: $n = 12$, $\rho = .664$, $p = .018$; Italian NW identification task: $n = 12$, $\rho = -.734$, $p = .007$). Response time in the English orthographic from identification task was also shown to be significantly correlated with accuracy in the clitic pronoun judgement task ($n = 12$, $\rho = -.807$, $p = .002$). Children's performance (accuracy in %) in the English phonological form identification task was significantly associated with accuracy in the Italian blending judgement tasks ($n = 12$, $\rho = .749$, $p = .005$). Response time in the English phonological form identification task correlated significantly with accuracy in the Italian word identification ($n = 12$, $\rho = .670$, $p = .017$, this effect was no longer significant when the children's performance in the CPM was used as control variable: $p > .05$) and accent identification task ($n = 12$, $\rho = .864$, $p < .001$). Accuracy in the English phoneme deletion task is significantly associated with accuracy in the Italian blending ($n = 12$, $\rho = .596$, $p = .041$) and clitic object judgement tasks ($n = 12$, $\rho = .739$, $p = .006$). Furthermore, accuracy in the English phoneme deletion task is significantly correlated with response time in the Italian self-paced sentence reading ($n = 12$, $\rho = -.630$, $p = .028$), NW identification ($n = 12$, $\rho = -.605$, $p = .037$) and blending judgement screening task ($n = 12$, $\rho = -.726$, $p = .008$). Also response time in the English phoneme deletion task is significantly associated with self-paced reading time of Italian syllables ($n = 12$, $\rho = .615$, $p = .033$) and response time in the Italian word identification screening task ($n = 12$, $\rho = .636$, $p = .026$). Also performance (accuracy in %) in the English tense judgement screening task correlated significantly with response time and accuracy in Italian reading screening tasks. In particular, performance (accuracy in %) in the English tense judgement was significantly associated with accuracy in the Italian word identification ($n = 12$, $\rho = .874$, $p < .001$) and in the clitic pronoun judgement tasks ($n = 12$, $\rho = .591$, $p = .043$). Performance (accuracy in %) in the English tense judgement was furthermore significantly correlated with self-paced reading time of Italian syllables ($n = 12$, $\rho = -.682$, $p = .015$) and sentences ($n = 12$, $\rho = -.721$, $p = .008$) and with response time in the word ($n = 12$, $\rho = -.735$, $p = .006$) and NW identification task ($n = 12$, $\rho = -.679$, $p = .015$).

Also for the group of Mandarin children, notable significant associations emerged between response time in the Mandarin RAN and the Italian self-paced syllable speed ($n = 7$, $\rho = .707$, $p < .001$). Mandarin RAN was also found to be significantly associated with accuracy in the Italian subject-verb agreement judgement task ($n = 7$, $\rho = .767$, $p = .044$). Response time in the Mandarin onset detection task was found to correlate significantly with response time in Italian word ($n = 7$, $\rho = .786$, $p = .036$) and NW identification ($n = 7$, $\rho =$

.821, $p = .023$), blending judgement ($n = 7$, $\rho = .821$, $p = .023$) and with performance (accuracy in %) in the subject-verb agreement judgement screening task ($n = 7$, $\rho = -.767$, $p = .044$). Accuracy in the Mandarin rhyme detection task was significantly associated with response time in all three Italian phonological awareness tasks (accent: $n = 7$, $\rho = -.956$, $p = .003$; blending: $n = 7$, $\rho = -.756$, $p = .049$; inversion: $n = 7$, $\rho = -.756$, $p = .049$) and with accuracy in the clitic pronoun judgement task ($n = 7$, $\rho = .784$, $p = .037$). Also response time in the Mandarin rhyme detection task correlated significantly with response time in the Italian self-paced syllable reading task ($n = 7$, $\rho = .893$, $p = .007$).

5.6.3.6 Screening usability

For the facilitation of data processing, analysis and interpretation of scores for selected questions were inverted (see appendix A) so that a higher score is always representative of a positive user experience.

Examiner usability. Figure 50 visualizes the distribution of responses to the usability online questionnaires. Examiners' responses to questions E3 to E9 on examiners' impressions on ease of use of the system (taken from the System Usability Scale, Sauro, 2022) were generally positive, with medians of $\bar{x}_s = 4$, on a scale from 1 (negative) to 5 (positive impression). Distributions of the answers across participants however differed across questions. None of the respondents rated system complexity (E2) as well as time (E6) and knowledge acquisition (E9) for system familiarization below 3. Distributions of responses ranged from 5 to 2 for the questions on the need of the support of a technician (E4), the integration of system functions (E5), cumbersomeness of use (E7) and confidence in using the system (E8). Examiners' ratings regarding the ease of system use ranged between 2 and 4 (E3).

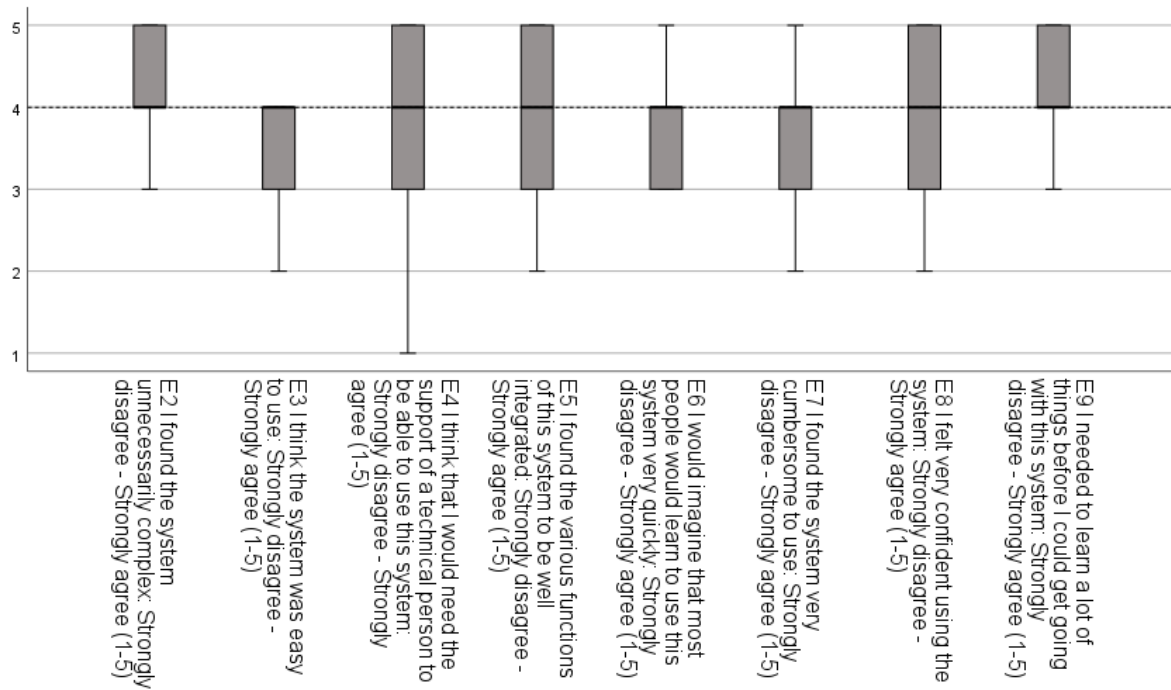


Figure 50: Distribution of examiners' responses to items E2-E9 in the usability online questionnaire.

Several questions targeted the overall feedback to the software (QUIS, Chin et al., 1988; Wallace et al., 1988). Again, the use of the software was not considered difficult (E11, $\bar{x} = 6.5$). A median of $\bar{x}_s = 7$ on a scale from 1 (negative) to 9 (positive impression) indicated a moderately positive user experience (E10: “wonderful”, E13: “stimulating”) with sizable distribution across the scale (see figure 51). Distributions were also rather high for the responses on rigidity (E14, $\bar{x} = 5$) and frustration (E12, $\bar{x} = 6$) perceived when using the software.

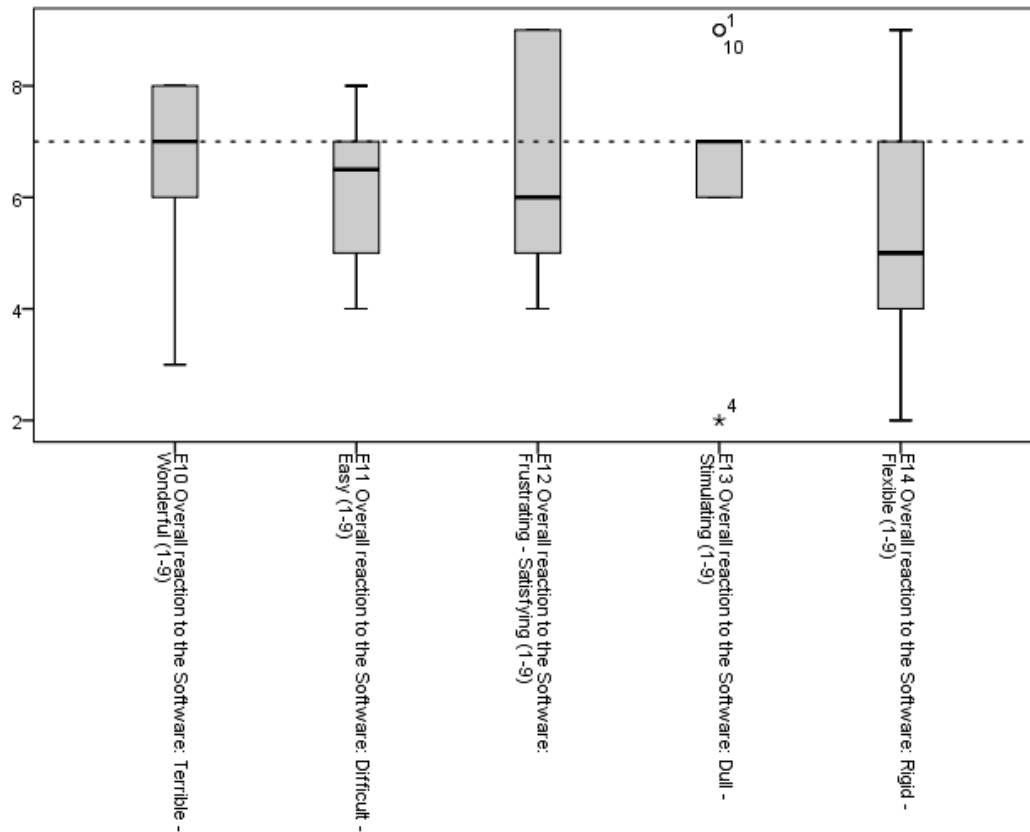


Figure 51: Distribution of examiners' responses to items E2-E14 in the usability online questionnaire.

Overall, examiners indicated again the perceived ease in operating the system (E17, $\bar{x} = 7.5$), task performance (E18, $\bar{x} = 7$) and that they felt both comfortable (E23, $\bar{x} = 8$) as well as satisfied with the system use (E26; $\bar{x} = 7$). The examiners expressed high to moderate satisfaction with the characters displayed (E15, $\bar{x} = 8.5$), the screen layout (E16, $\bar{x} = 7$) and the design of the interface (E24, $\bar{x} = 8$). Examiners' ratings on the speed of the system (E19, $\bar{x} = 6$), reliability (E20, $\bar{x} = 6$) and appropriateness for all levels of users (E21, $\bar{x} = 5$) were slightly lower and more distributed across the scale (see figure 52). However, the perceived benefit employed by the software for the work context of the examiners was rated as medium to high regarding work efficiency (E22, $\bar{x} = 7$) as well as functionalities (E25, $\bar{x} = 6.5$) and was again fairly distributed across the whole scale.

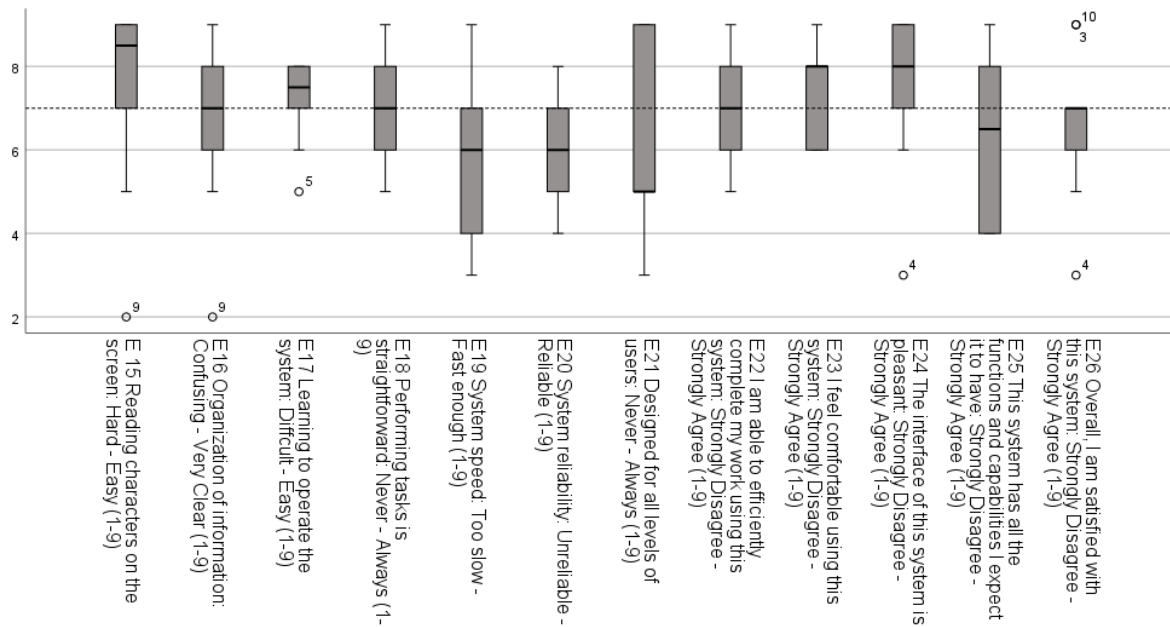


Figure 52: Distribution of examiners' responses to items E15-E26 in the usability online questionnaire.

The examiners expressed medium to high satisfaction with the screen design and layout in general (E44, $\bar{x} = 8$) and in particular the use of colours (E45, $\bar{x} = 9$) as well as with the resolution (E27, $\bar{x} = 9$), shape (E28, $\bar{x} = 9$) and contrast with background (E29, $\bar{x} = 8.5$) of the characters when displayed during the screenings (see figure 53). While also the amount (E33) and the arrangement of information (E34) on the screen was rated as rather satisfactory ($\bar{x}_s = 9$), the highlighting of screen items (E30), use of colours for the latter (E31) as well as the screen layouts (E32) in the tasks however were perceived as slightly less positive (all $\bar{x}_s = 7$). Despite having generally reached medium to high ratings, the distributions of the answers show that at least for some of the examiners, the ease in finding (E36; $\bar{x} = 9$) and selecting (E35, $\bar{x} = 7$) an item as well as the item selection area size (E37, $\bar{x} = 7.5$) appeared to be less convincing. Distributions are even more remarkable and ratings are generally lower when it comes to the system's responsiveness expressed in the system's feedback on cursor location (E38, $\bar{x} = 7$), item selection in general (E39, $\bar{x} = 6$) and when the finger is pulled away from the touch screen (E40, $\bar{x} = 6$). Similarly, the response time of the system is generally perceived as moderate to high (E43, $\bar{x} = 6.5$), but answers are distributed across the whole scale again. Less distribution was found regarding the questionnaire items on general satisfaction and in particular ease (E41, $\bar{x} = 7$) and time (E42, $\bar{x} = 7$) needed for the system familiarization. Most examiners described the use of the system as pleasant (E46, $\bar{x} = 7.5$).

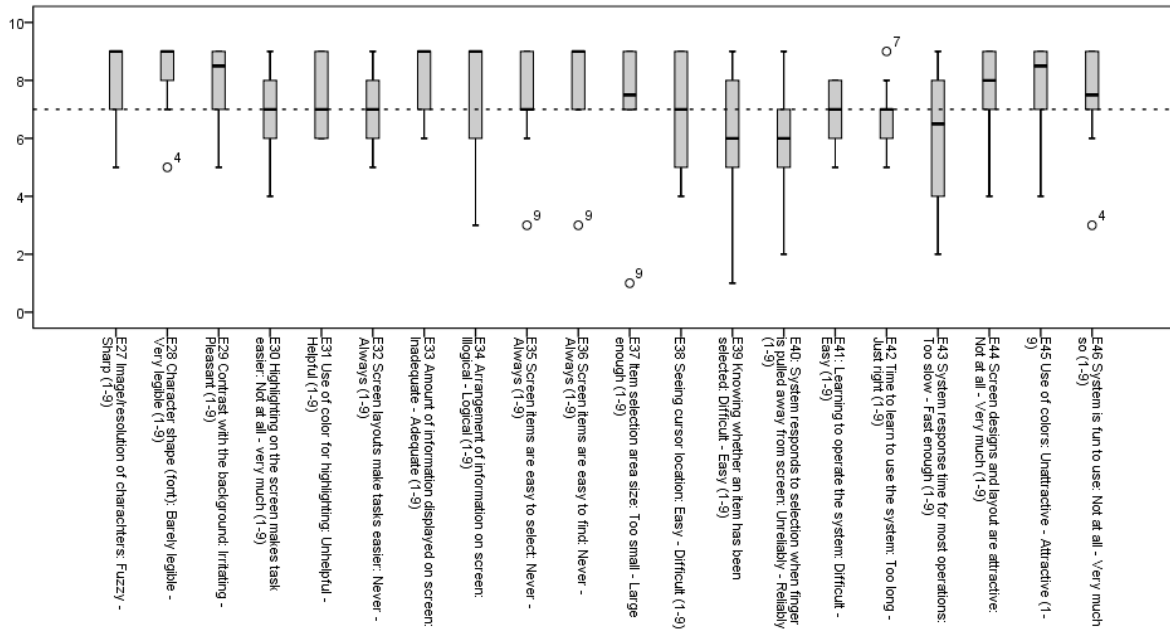


Figure 53: Distribution of examiners' responses to items E27-E46 in the usability online questionnaire.

Examinee usability. The age of the examinees participating in the usability study was not significantly associated with any of the questionnaire items ($p_s > .05$) except for ease of learning to use the system (P4; $n = 11$, $\rho = 0.807$, $p = .003$). Figure 54 visualizes the distribution of responses to the usability online questionnaire. Examinees' responses to questions P1, P4, P6 and P7 ($\bar{x}_s = 4$) suggest ease of overall familiarization, accessibility and use of the remote screenings they were administered. While most of the children enjoyed the use of the software (P2, $\bar{x} = 4$), few participants considered the screening administration boring (P3, $\bar{x} = 4$). In particular, they enjoyed screen graphics (P12, $\bar{x} = 5$) and were content about character readability (P8, $\bar{x} = 4$) as well as about cursor position visualization (P10, $\bar{x} = 5$). Expectations on screen element selection (P9, $\bar{x} = 4$), system feedback related to the latter (P11, $\bar{x} = 4$) as well as system speed (P5, $\bar{x} = 3$) however were not consistently met.

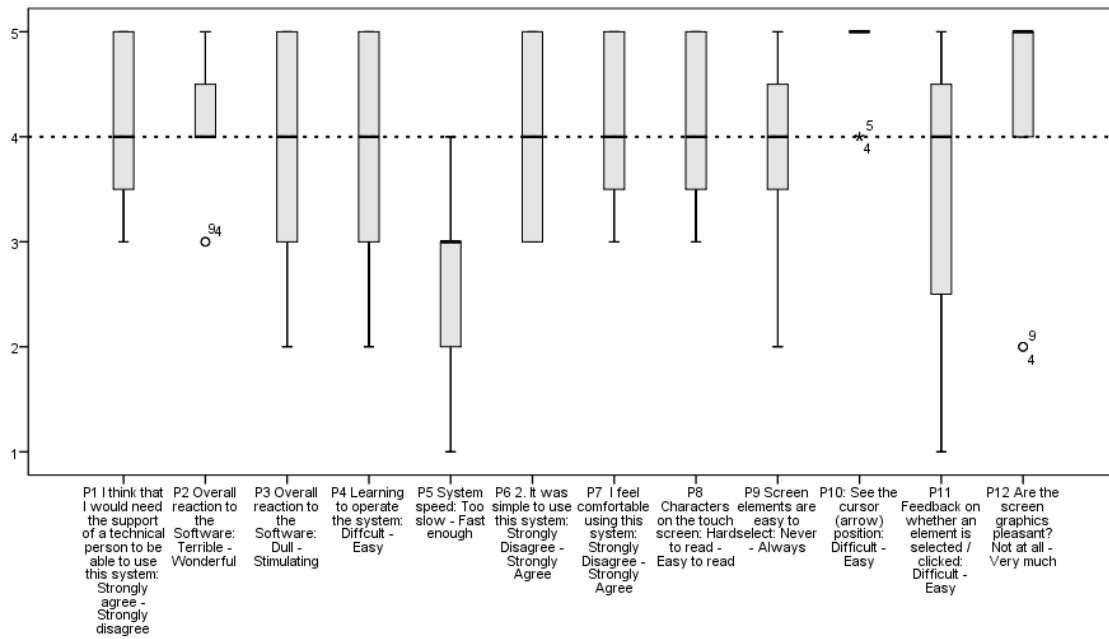


Figure 54: Distribution of examinees' responses to the usability online questionnaire.

Post-hoc-analyses revealed that the perceived ease of using the system (P6) was significantly correlated with overall reaction to the software (P2; $n = 11$; $\rho = 0.634$, $p = .036$), how boring they found the screening (P3; $n = 11$; $\rho = 0.926$, $p < .001$) and how comfortable they felt using it (P7; $n = 11$, $\rho = 0.946$, $p < .001$). The latter (P7) was found to be significantly associated with the degree of dullness perceived (P3; $n = 11$, $\rho = 0.898$, $p < .001$). Cursor position detection (P10) was associated (but not significantly) with the ease of learning how to use the systems (P4, $n = 11$, $\rho = 0.588$, $p = .057$).

5.6.3.7 Interim discussion

Overall, screening tasks in Italian and the children's family languages were shown to be associated with results obtained in standardized/traditional tests in the respective languages, the risk scores deriving from the latter and caregivers' and teachers' evaluation of the children's language and reading development and skills as well as academic achievements, see table 14 for an overview. Based on these findings, the research questions raised concerning concurrent and discriminative validity can be positively answered.

Table 14: Overview of significant associations between screening task performance (accuracy in %) risk level, standardized tests as well as teacher and caregiver questionnaires.

Screening task performance	Risk score	standardized tests (raw scores)	Teacher questionnaire	DSA questionnaire
DDE-2				
syllables (IT) self-paced reading time	n.s.	<i>word reading time:</i> $n = 27, r = .461,$ $p = .015$	n.s.	n.s.
sentences (IT) self-paced reading time	L1 risk: $n = 19, rho = .473,$ $p = .041$	<i>word reading time:</i> $n = 24, r = .942,$ $p < .001$	<i>total score:</i> $n = 25, rho = .514,$ $p = .009$	<i>reading & writing:</i> $N = 30, rho = .568,$ $p = .001$
word identification (IT) response time	compound risk: $n = 19, rho = .457,$ $p = .049$	<i>word reading time:</i> $N = 30, r = .531,$ $p = .003$	n.s.	<i>reading & writing:</i> $N = 30, rho = .399,$ $p = .029$
word identification (IT) accuracy (%)	L1 risk: $n = 19, rho = -.500,$ $p = .029$	<i>word reading accuracy:</i> $N = 30, r = .880,$ $p = .001$	<i>receptive phonology:</i> $n = 17, rho = -.482,$ $p = .050$	<i>reading & writing:</i> $N = 30, rho = -.625,$ $p < .001$
nonword identification (IT) response time	n.s.	n.s.	n.s.	<i>reading & writing:</i> $N = 30, rho = .426, p = .019$
nonword identification (IT) accuracy (%)	n.s.	<i>word reading accuracy:</i> $N = 30, r = .569,$ $p = .001$	n.s.	<i>reading & writing:</i> $N = 30, rho = -.477,$ $p = .008$
phoneme blending (IT) response time	n.s.	n.s.	n.s.	<i>reading & writing:</i> $N = 30, rho = .580, p = .001$
phoneme blending (IT) accuracy (%)	n.s.	<i>word reading accuracy:</i> $N = 30, rho = .492,$ $p = .006$	n.s.	<i>reading & writing:</i> $N = 30, rho = -.427,$ $p = .018$
stress identification (IT) response time	n.s.	n.s.	n.s.	n.s.
stress identification (IT) accuracy (%)	n.s.	n.s.	n.s.	n.s.
syllabic inversion (IT) response time	n.s.	n.s.	n.s.	n.s.
syllabic inversion (IT) accuracy (%)	n.s.	n.s.	n.s.	n.s.
subject-verb agreement (IT) response time	n.s.	n.s.	n.s.	n.s.
subject-verb agreement (IT) accuracy (%)	n.s.	n.s.	n.s.	<i>reading & writing:</i> $n = 19, rho = -.728,$ $p = .001$

clitic pronouns (IT) response time	compound risk: $n = 19, \rho = -.531,$ $p = .019$	<i>n.s.</i>	<i>n.s.</i>	<i>general learning difficulties</i> $N = 30, \rho = .384,$ $p = .036$
clitic pronouns (IT) accuracy (%)	<i>n.s.</i>	<i>word reading accuracy:</i> $N = 30, \rho = .513,$ $p = .004$	<i>n.s.</i>	<i>reading & writing:</i> $N = 30, \rho = -.482,$ $p = .007$
TOWRE-2				
sentences (IT) self-paced reading time	L1 risk: $n = 12, \rho = .717,$ $p = .009$	<i>word reading time:</i> $n = 12, \rho = .888,$ $p < .001$	<i>n.s.</i>	<i>n.s.</i>
orthographic form (ENG) response time	<i>n.s.</i>	<i>word reading time:</i> $n = 12, \rho = .713,$ $p = .009$	<i>n.s.</i>	<i>n.s.</i>
orthographic form (ENG) accuracy (%)	<i>n.s.</i>	<i>word reading accuracy:</i> $n = 12, \rho = .748,$ $p = .005$	<i>n.s.</i>	<i>n.s.</i>
phonological form (ENG) response time	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
phonological form (ENG) accuracy (%)	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
stress identification (ENG) response time	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
stress identification (ENG) accuracy (%)	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
sound deletion (ENG) response time	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
sound deletion (ENG) accuracy (%)	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
tense judgement (ENG) response time	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
tense judgement (ENG) accuracy (%)	L1 risk: $n = 12, \rho = -.781,$ $p = .003$	<i>word reading accuracy:</i> $n = 12, \rho = .740$ $p = .006$	<i>n.s.</i>	<i>n.s.</i>
Chinese reading test				
left-right inversion (MAND) response time	<i>compound risk:</i> $n = 7, \rho = -.882,$ $p = .009$	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
left-right inversion (MAND) accuracy (%)	<i>compound risk:</i> $n = 7, \rho = -.882,$ $p = .009$	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
radical position (MAND) response time	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>
radical position	<i>compound risk:</i> $n = 7, \rho = -.882,$	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>

(MAND) accuracy (%)	$p = .009$			
onset detection (MAND) response time	<i>compound risk:</i> $n = 7, \rho = -.886,$ $p = .012$	n.s.	n.s.	n.s.
onset detection (MAND) accuracy (%)	n.s.	n.s.	n.s.	n.s.
rhyme detection (MAND) response time	n.s.	n.s.	n.s.	n.s.
rhyme detection (MAND) accuracy (%)	n.s.	n.s.	n.s.	n.s.
tone detection (MAND) response time	n.s.	n.s.	n.s.	n.s.
tone detection (MAND) accuracy (%)	n.s.	n.s.	n.s.	n.s.

Since the results of standardized/traditional tests and screening tasks declared to measure the same skills are associated with each other – suggesting good concurrent validity – Hypothesis 1 is also supported in this sample. Furthermore, children with a higher risk score (based on information from standardized/traditional tests) did perform worse in the screening tasks than children with no or lower risk scores which suggests the screening’s potential towards discriminant validity and supports Hypothesis 2. Investigating the contribution of the screening results to DD risk identification further, the concepts of reading accuracy and speed as measured directly through the standardized/traditional tests were shown to be associated with the same concepts measured in the screening tasks (cf. results for self-paced syllable and sentence reading speed and word and NW identification speed and accuracy). In addition to that, screening tasks that did not involve exclusively reading-related decoding processes but grammatical (cf. clitic pronoun judgement) or metaphonological abilities (cf. phonological bending) were also shown to be associated with results in the standardized/traditional reading tests and thus their potential in DD risk identification was supported. These relationships were not only found for Italian, but also for the screening tasks in the children’s family languages, i.e. English and Mandarin. For both languages again not only reading-related tasks (English self-paced sentence reading and orthographic form identification task, Mandarin radical position and left-right inversion) are significantly associated with scores obtained in traditional/standardized tests, but also language- and metaphonological tasks, like English tense judgement as well as Mandarin onset and rhyme detection and RAN. This finding supports the potential of contribution of language tasks in the detection of DD risk (Arosio et al., 2016; Vender et al., 2018; Vender et al., 2020), which is of importance in the diagnostic processes

for multilingual children with a high degree of variance regarding the alphabetization in the family language. For alphabetization, variance presumably is even higher than variance in quality, frequency and duration of language exposure, which is also reflected in our sample with some children being schooled in the societal language attending weekend schools for literacy acquisition in the family language (Mandarin-speaking children) vs. children attending bilingual schools (English-speaking children). Knowing that a) it is important to consider all languages spoken by a child (see explicit indication by the WHO in the case of DLD by the WHO, 2022a) and b) that not only reading, but also language tasks in the children's family and heritage language contribute to the DD risk identification (cf. chapter 2.4.2, Gersten & Geva, 2003; Geva, 2000) is of high relevance for diagnostic procedures of multilingual children. Despite the fact that neither the screening tasks nor the caregiver questionnaire have been validated, the associations between them show that the performance in the screening tasks is somewhat representative for caregivers' and teachers' perception of the situation. The fact that the behaviour shown in a screening in a one-on-one remote setting corresponds to the teachers' observation of the child in the classroom and the caregivers' experiences and perceptions indicates that computerized screenings can be a useful first step in order to identify DD risk in children. This however does not mean that caregiver and teacher questionnaires are no longer needed. Diagnostic processes and the planning of an intervention should always also include the caregivers' and teachers' perspectives to base diagnosis and intervention on the children's concrete situation which cannot be fully grasped through direct assessment like the MuLiMi screenings only. Performance on different screening tasks assessing similar skills in the same language were shown to be associated with each other (Hypothesis 3). Furthermore, performance on various screening tasks assessing the same linguistic area in the two different languages were shown to be associated. (Hypothesis 4). Due to the lack of children with an existing DD diagnosis in this sample, Hypothesis 5 inquiring about whether children who have been identified with language or reading difficulties do show an impairment in both L1 and L2 on screening tasks assessing the same linguistic area could not be finally answered.

Further studies on larger samples including subjects with a diagnosis are needed to understand the role of each single screening task, for example to assess why none of the English metaphonological tasks was found to be associated with reading performance in the TOWRE-2. These data could provide insight in whether this finding results from differences in orthographic depth, the linguistic phenomena tested (phoneme blending and word stress identification), the paradigm (judgement and matching tasks) or their adaptation to a fully

automatized version of the tasks. A further limitation is that the Mandarin screening does not contain tasks testing grammatical knowledge and morphosyntactic processing skills.

In the usability studies, both target user groups, the examinees and the examiners indicated that for the most part, the screening platform sufficiently responds to their requirements, meets their expectations and is enjoyable in use (RQ 3). This underlines the potential of computerized screenings not only from a risk identification point of view, but also regarding the actual application of such tools in clinical practice. Like in other studies where different electronic and digital means in other populations were tested and validated (Haridas et al., 2017; Hautala et al., 2020; Horbach et al., 2018; Rauschenberger et al., 2019), the potential of transfer or adaptation of well-tried and reliable tasks into computerized screening tasks was shown. This finding is not only supported by the associations found in the work presented between the screening tasks and standardized/traditional reading measures as well as with questionnaires, but also with the fairly good ratings by both examiners and examinees in the usability online survey. Despite the need of improvement for system speed and feedback, both groups of users showed general satisfaction and were able to use the system autonomously and satisfyingly.

6 General discussion

From the studies described above, a series of observations, interpretations and indications for clinical practice as well as future research emerge. Those do relate to the potential of the screenings in the assessment of bilingual children's language and reading skills. Overall the hypotheses related to concurrent (as assessed through standardized test performance) and discriminant validity (as assessed through the risk level determined by the standardized test performance and information on an existing diagnoses) were supported for a series of screening tasks across languages, age groups and screening purposes (see the interim discussions in chapters 5.3 to 5.6). The studies carried out indicate that it is possible to automatically assess language, reading as well as reading-related skills in the children's L1 without any requirement for the examiner to speak this language. Even if screening task performance is not sufficient to issue formal diagnoses, scores obtained in screening tasks proved to be useful indicators for timely initiation of diagnostic and interventional processes (Law et al., 2003).

It is important to mention that due to the diagnostic dilemma described in chapter 2.4 concerning the difficulty of appropriately assessing bilingual children's language performance, it was considered important to conduct studies avoiding circularity between classification and verification as much as possible. The validation was thus grounded in comparison

of standardized test performance, caregiver, teacher and SLT questionnaires, pre-existing diagnoses or risk indicated by the teacher as well as associations within and across languages among the screening tasks. This combination of several approaches might still not be sufficient to exclude the risk of misdiagnoses for the children already holding a diagnosis in our sample, but the risk is minimized. From the significant associations of screening task performance with caregiver and teacher questionnaires, it is inferred that the screening tasks have the potential to grasp some of the children's language and reading skills in ecological family and school settings. In this general discussion, certain aspects concerning the appropriateness of screening tasks in DLD and DD risk detection will be highlighted and interpreted in the context of the study design, screening construction, the participants' characteristics and language-specific features.

6.1 Interpretation of results

The survey study with SLTs confirmed that many SLTs regularly diagnose and treat bilingual children and that SLTs are aware of good practice and requirements when diagnosing and treating multilingual children. They also revealed openness towards new methods like computerized screenings. The identified gap between awareness and use of specific material for bilingual children points to the fact that the situation still needs to be improved. In the following subchapters, it is thoroughly discussed in how far bilingual computerized screenings implemented on the screening platform MuLiMi can contribute to closing this gap.

6.1.1 Bilingual DLD screenings

For screening tasks across linguistic areas, significant associations emerged between task performance and variables indicating DLD risk in both versions of the screening, suggesting potential in DLD risk identification.

6.1.1.1 *Validity*

Overall, in both the Spanish-Italian as well as in the Italian-German screening tasks, NWRT performance was consistently associated with variables on DLD risk, standardized test performance as well as with SLT, kindergarten teacher and caregiver questionnaire responses. Also, as suggested by previous studies, the children's performance in the NWRT was consistently significantly correlated with other screening tasks across linguistic areas in both languages (Farabolini et al., 2021; Hoover & Storkel, 2006; Rispens & Been, 2007), which holds for both the Spanish-Italian as well as the Italian-German DLD screening study. This observation supports previous research that highlights the potential of NWRT in DLD risk identification also in bilingual children (Schwob et al., 2021). Also, the effort of a two-step rating and selection procedure for the NWs constructed, implemented and administered considering LS

and NLS NW characteristics seems to be relevant (Chiat, 2015). However, small sample size in the case of the Italian-German screening and low amount of selected NWs in the different categories in the Spanish-Italian screening do not allow for in-depth analyses of the advantages of the single NW subcategories. While in the case of the Spanish-Italian sample, NWRT performance was not significantly associated with nonverbal intelligence as measured by the children's Raven's CPM performance, significant associations between CPM and NWRT scores emerged in the Italian-German screening. Since most of the significant associations remained stable when inserting the children's CPM t-scores as control variable, it can nonetheless be concluded that the NWRT is a suitable screening task for bilingual children with varying levels of nonverbal intelligence. One further advantage of the NWRT implemented on the MuLiMi screening platform relates to the possibility to automatically administer NWs in the child's L2 as well as in the child's L1. Despite the current lack of automatic evaluation of the children's responses through automatic speech recognition (ASR), high levels of inter-rater reliability suggest that the evaluation of NW repetition performance is feasible even when evaluating child repetition performance of NWs that are specific to a language that is not spoken by the rater in charge of the scoring.

Also performance levels in the CLTs were consistently associated with all variables related to DLD risk in the Spanish-Italian screening study, while these associations were found for the German version of the CLT subtests only in the Italian-German screening study. This might be explained by the circumstance that significant associations between CLT scores and the information from language background questionnaires on language input and output in both languages were revealed in the Italian-German sample. As expected due to the nature of the CLT screening tasks, performance in the German versions of these tasks correlated significantly with German standardized test performance and the risk level deriving from the latter. Since the language background questionnaire was exclusively administered to the caregivers of children participating in the study on the Italian-German DLD screening, the interaction between CLT performance and language input and output could not be investigated in the Spanish-Italian sample. It can thus not be clarified whether also here the language in- and output are associated with CLT performance levels. The difference between the Spanish-Italian screening study, where the variables relevant for DLD risk identification were found to correlate with both CLT versions opposed to the Italian-German screening study, where only significant associations with the German(-based) variables emerged, might be explained by the higher amount of language similarity and cognates across languages for Spanish vs. Italian compared to Italian vs. German. The successful implementation of the

CLT comprehension subtests as screening task to be automatically administered and evaluated as indicated by the significant associations between standardized and CLT scores in both screenings indicates that the transfer of well-tried testing paradigms into computerized screenings is possible and useful for functional diagnostic purposes as indicated by the amount of significant associations emerging for the CLT subtest performance in the language of the standardized tests administered and the performance in the latter.

In the Italian-German screening, while some of the screening tasks were found to be associated with all or many of the variables assessed, data analyses revealed that the children's performance in screening tasks on morphosyntactic processing were less consistently associated with those variables (see table 12). For the Italian screening tasks, this is easily justified since the standardized tests and the risk level deriving from the latter as well as the teacher questionnaire (all in German) are not designed to grasp the Italian language performance levels of the children. The lack of significant associations between the German morphosyntactic processing screening tasks and the other variables might have to do with the fact that for both case marking and subject-verb agreement in German, time-related variables (age, LoE, AoO) seems to play an important role (Scherger, 2022). Despite the ambiguity in previous research findings and unclear indications concerning the suitability as clinical marker, children's performance in the German case marking screening task was found to be associated with improvement in the same task, suggesting that in fact, case marking does have diagnostic and predictive potential. Interestingly, in the Spanish-Italian DLD screening study instead, the Italian finiteness and subject-verb agreement tasks were all significantly associated with all the other variables (with the exception of Spanish finiteness and the caregiver questionnaire). From the comparison of appropriate DLD risk detection properties of the Italian and Spanish finiteness screening tasks (WSIR) opposed to the German subject-verb agreement (judgement and matching) and case-marking (picture matching) screening tasks, it may be hypothesized that the WSIR paradigm implemented in the MuLiMi screening platform is more appropriate for the DLD risk detection in bilingual children than judgement and matching tasks. Here, it is also worth mentioning, that the tasks on morphosyntactic processing in the screening task are different in nature from the LiSeDaZ (elicited spontaneous speech sample in a semi-structured interview) and some of the BVL subtests (oral production in sentence repetition and completion). Despite these differences, associations between the screening and standardized tasks were observed indicating good concurrent validity.

Another interesting observation relates to the inconsistency of associations between screening task performance and language in- and output scores deriving from the language

background questionnaire, which was filled in by caregivers in the Italian-German screening study only. Two different interpretations are reasonable: (1) the caregivers' perception of the children's language abilities based on their daily language interaction, i.e. spontaneous speech, do not directly compare with the screening (processing) tasks that might have the potential for DLD detection despite artificiality (NWs, grammaticality judgement, phonological awareness): (2) It is possible that the tasks chosen and identified as suitable clinical markers also for bilingual children actually are less sensitive to language in- and output and thus are not associated with screening task performance. However, for some tasks, the language in- and output scores were significantly correlated with screening task performance indicating that it is relevant to assess the exposure patterns (Parra et al., 2011), especially when lexical skills cannot be directly assessed (Engel de Abreu, 2011). It is also possible that the children despite different exposure patterns as revealed by the language background questionnaire, the eL2 and 2L1 speakers included in this study all have already had sufficient exposure and thus the effects of language experience were minimized in this sample (Thordardottir & Brandeker, 2013). This might be explained by the sampling method used in the Italian-German screening study, for which most children were recruited in Italian-German kindergartens as well as the inclusion criteria of minimum two years of German language exposure.

Overall, language acquisition in eL2 and 2L1 language learners is believed to evolve in a comparable way. Furthermore, DLD is believed to manifest itself in all languages spoken. It was thus investigated whether the children's screening task performance in their L1 is associated with screening task performance in their L2. In the Spanish-Italian screening, significant crosslinguistic associations emerged for all screening tasks except the DNWL when including the whole group in the analyses. When conducting the same analyses for the subgroup of DLD-children only, this effect was only significant for the NWRT and the CLT verb comprehension subtest. Interestingly, this pattern was also found for the same analyses on the Italian-German sample: significant associations across tasks emerged for the whole sample, but analysing the data collected with children with DLD only, the effects were present exclusively in the NWRT and CLT screening tasks. Due to the small sample size, this effect can be interpreted in different ways: (1) The DLD children's disorder could be more evident for phonological and lexical skills and morphosyntax was thus not similarly impaired; (2) When analysing data from children of all risk levels, those were significantly associated with both finiteness task as well as with the Italian subject-verb agreement screening tasks from the Spanish-Italian screening and in the German-speaking sample with both versions of the subject-verb agreement task (but neither case marking nor clitic object pronoun task performance). However, non-significant associations across languages in these tasks emerged

when analysing the children with DLD only. This might as well be related to insufficient informative value of these screening tasks in this population. Item analyses will allow for a selection of the most discriminative items; (3) The absence of crosslinguistic associations in the screening tasks on morphosyntactic processing are related to the fact that children had been mistakenly identified with DLD, but in fact perform poorly in one of the languages only. Due to the associations of caregivers' responses to the questionnaire with children's screening performance as well as with DLD children's low performance levels in the standardized tests (see tables 6 and 12), this seems unlikely but cannot be excluded. (4) In both the Spanish-Italian as well as the Italian-German screening, the morphosyntactic processing tasks appear to be more demanding compared to the CLT and the NWRT in terms of memory. The role of task complexity is believed to be of high relevance when running these comparisons. However, despite the fact that when factoring out children's performance in the Raven's CPM, screening results were generally maintained, tables 5 and 8 indicate that in both the Spanish-Italian and the Italian-German samples, children in the DLD group on average scored lowest in the Raven's CPM. Again, larger samples are needed in order to a) understand the relationships between nonverbal intelligence and language task performance and in how far these potential interactions need to be considered in the construction of automatically administered and evaluated screening tasks as well as b) to be able to run statistical analyses for group comparisons and have sufficiently big sample sizes to run reliable partial correlational analyses in the clinical group without violating the requirements for parametric tests.

6.1.1.2 Predictivity

Overall, children's performance level at t1 was found to be significantly associated with their performance at t2 as well as with the degree of improvement in the Italian-German DLD screening follow-up study. This finding suggests that the screening tasks do capture the current as well as predict future language outcomes.

Both in the follow-up study of children that had been administered the Spanish-Italian screening and in the follow-up study on the Italian-German screening, associations between screening tasks performance at t1 and improvement were found in the respective L2 (societal language) more consistently than for the screening tasks in the children's L1. For the Italian-German screening, this interpretation is grounded in result of non-parametric tests while due to the small sample of $n = 5$ children, for the Spanish-Italian screening this interpretation is based on visual inspection of figures 22, 23 and 24 comparing t1 and t2 screening results (with the exception of one participant). This observation might be either traced back to poor predictive value of the L1-screening tasks or the circumstances that schooling and also SLT

intervention in the case of children with DLD diagnosis takes place in the children's L2. For the Italian-German sample it was found that these language-specific effects do not relate to the children's performance in the NWRT since performance at t1 was not significantly correlated with improvement in the single and compound LS scores. It is hypothesized that this effect is due to the greater complexity concerning the repetition of NLS NWs and thus offered more room for improvement. This related to the expected but still remarkable effect that children with lower performance at t1, showed highest levels of improvement (see figure 31).

More specifically, in order to better understand the complex interactions of variables, post-hoc analyses were run. Overall, few significant associations with the variables representing nonverbal intelligence as measured by the Raven's CPM, age (in months) and language dominance were found. However, certain significant associations give indications that are relevant for both research and clinical practice: The children's age at t1 was significantly associated with improvement in the NWRT and CLTs suggesting that the younger a child, the more potential of improvement there is. Figure 31 shows that this effect seems to be of particular importance in the DLD and risk groups. This is relevant considering potential effects of early vs. later onset of SLT interventions (Law et al., 2003). Furthermore, the amount of Italian output was significantly (negatively) associated with the improvement in German LS NW repetition performance, which points to the necessity of contextualizing the children's language(-specific) performance depending on the linguistic background and supports the requirement of the assessment of both languages spoken. This interpretation is undergirded by the observation that in the follow-up study on Italian-German-speaking children, no crosslinguistic associations for the t1 performance and improvement were found. This shows that the assessment of one language only is insufficient to determine the language performance in the other language. The Italian-German DLD screening follow-up study indicates that the assessment of both languages can be realized with the help of computerized screening tasks that can be automatically administered and give useful information concerning later reading outcomes.

6.1.2 Validity of bilingual DD screenings

In all DD screening studies, the standardized reading test results were significantly associated with performance on several screening tasks (Bigagli & Lorusso, 2014; Brookes et al., 2011). Especially in the diagnosis of reading difficulties, it should be accentuated that for self-paced reading time in particular, robust findings concerning the association with reading time in standardized screening tasks emerged. This shows that time-consuming and cumbersome manual assessment of reading times in standardized assessment like the ZLT-II (Petermann

& Daseking, 2019) or DDE-2 (Sartori et al., 2007) might be replaced with automatic assessment of reaction and reading time at least for screening purposes. Similarly, screening tasks in different languages indicate the potential in automatic assessment of reading accuracy. In particular German and Italian word and NW identification screening tasks, English orthographic form identification tasks as well as Mandarin radical position and left-right inversion tasks were correlated with standardized test performance in the respective language. A very practical implication concerning the screening tasks assessing reading accuracy irrespective of the assessment of bilingual children might be that in school contexts, children could be administered with the screening tasks all at once (provided that each child has a suitable device and headphones) and evaluated automatically, which on the long run (after familiarization with the screening tool and tasks) might have a positive impact concerning teachers' resources (this needs to be quantified in future studies).

Besides that, associations between the screening task results and caregiver as well as teacher questionnaire responses for the evaluation of the children's language and reading performance emerged: this finding, considering potential circularity of verification and identification when relying on the standardized test results and risk scores deriving from the latter only, further strengthens data screening validity. Furthermore, screening task performance in the same or comparable tasks across languages was compared to each other and association were found to be significant. This is line with literature suggesting parallel development of linguistic skills in eL2 and 2L1 learners (de Lamo White & Jin, 2011; Letts, 2013; Riva et al., 2020), but it can also be expected in the samples involved in the screening studies, since they all acquired both languages at school or in the case of the Mandarin-speaking children in (online) weekend schools. This fact however makes it difficult to generalize the present findings to the whole population of bilingual children, since bilingual and weekend schools do not exist for all language combinations and access to them is associated with costs and effort. It is thus necessary to understand the specific role of reading-related screening tasks in the children's L1 not directly assessing the reading skills, but known to be associated with reading performance (see chapter 2.4.2, discussion below).

The evidence that reading-related skills like RAN, phonological awareness as well as morphosyntactic processing tasks are useful in DD risk detection (see chapter 2.4.2), was supported across screening versions, since significant associations of screening scores with reading-related skills and standardized reading test performance were found (Araújo & Faisca, 2019; Ben-Dror et al., 1995; Da Silva et al., 2020; Melby-Lervåg et al., 2012; Norton

& Wolf, 2012). Nevertheless, compared to reading tasks, fewer and in some cases no associations were found between performance in some of the screening tasks assessing reading-related skills and standardized reading tests. This might be explained by the non-fully appropriate direct comparison between screening tasks assessing reading-related skills and standardized test performance since the skills tested in screening and standardized tests are not measuring the same skill. This finding indicates that, especially when the focus is on functional diagnosis for the planning of intervention, the assessment of reading-related abilities is not sufficient to infer the child's training needs. Nonetheless, in both screening studies, associations between standardized reading tests and phonological awareness tasks were found in all languages for certain tasks and variables. The results suggest that phonological awareness tasks may contribute to risk detection and might be especially useful when the child has not or hardly formally instructed to read in the language of the examiner and of assessment.

Across language groups, most significant associations between the standardized test, risk and questionnaire variables were found for the Italian phonological awareness tasks. This might be explained by the transparent nature of Italian orthography suggesting that the decoding process of graphemes and converting them into phonemes (which is required in the standardized screening tasks) is most similar to the phoneme and syllable manipulations in the Italian phonological awareness tasks. Since good evidence for the appropriateness of phonological awareness tasks in DD identification was found across languages in previous studies (Ho & Bryant, 1997; Landerl et al., 2013), another reason might be the construction of the tasks since in the case of the Mandarin and English screening, the paradigms applied differed from the ones used in the Italian screening (see figures 36 and 46).

When interpreting the results of the screening studies, it is important to consider that response time in the phonological awareness and morphosyntactic processing tasks might need an interpretation that is different from interpreting (self-paced) reading times. In the course of the preliminary validation of the DD screening studies, the analyses of such patterns was not statistically investigated. It was observed that long reading times were often associated with poor reading performance as measured in the standardized tests, which is expected. Also short response times for phonological awareness and morphosyntactic processing tasks were found to be associated with poor reading performance in standardized tests. This shows that poor readers respond quicker to phonological awareness and morphosyntactic processing tasks. In this case, short response times does not necessarily mean better task performance, but could also indicate fast and inaccurate responses. Thus, ideally, the computerized screening system would not only display the amount of incorrect and correct

responses along with mean and total response time (see figure 12), but also the response time for correct items only (if the amount of correct responses is sufficient to be considered representative of the child's performance).

As mentioned above, the caregivers' and teachers' responses to the questionnaires in addition to the standardized tests and the risk levels deriving from the latter, are significantly associated with screening task performance (Pua et al., 2017). Language-specific patterns for these associations were observed: while for teachers' responses to the questionnaire, it can be expected that more significant associations occur with the screening tasks in the (main) language of schooling, the same pattern was observed across screening studies for caregivers' responses. This indicates that a) the caregivers' impressions of their children's reading skills seem to heavily depend on the children's achievement at school and b) that the DSA questionnaire seems to mostly grasp information related to the (main) language of schooling. Point (a) reinforces the idea that the direct assessment of reading skills in the children's L1 cannot be fully replaced by indirect assessment since in this study, questionnaire responses were not often associated with children's screening task performance in their L1. Point (b) suggests that the DSA questionnaire that was originally designed for caregivers of monolingual Italian children is representative of the caregivers' evaluation of reading and writing performance in the child's language of schooling. If the DSA questionnaire should also give indications on the child's reading and writing skills in the L1, it might need some adaptation to more ideally suit the needs in the reading assessment of bilingual children.

6.1.3 Usability

The results previously described and discussed across age groups, language combinations and screening purposes indicate that the MuLiMi screening platform offers a flexible solution that can be easily adapted to various requirements. It should be remarked that the studies indicate the potential of computerized assessment in children in general (Brookes et al., 2011) and even for preschool children (Horbach et al., 2018; Rauschenberger et al., 2019). In our case however, preschool children's (chapters 5.3. and 5.4) satisfaction with the screening tool was only indirectly assessed through the responses by their examiners described in chapter 5.6.3.6. For Italian monolingual primary school children instead, usability was directly assessed.

Overall, the usability study yielded positive results: the screening platform's graphics were described as enjoyable and the fact that in the majority of cases, children who were tested remotely autonomously responded to the screening tasks without support from their caregivers indicates that overall, the system is easy to use. This means that the tasks seem

to be sufficiently intuitive to be handled by the children and at the same time to be representative of their reading skills, supporting their potential for DD screening and risk identification. It is remarkable that also data collected exclusively remotely (studies described in chapter 5.3. and 5.6) appear to be valid and discriminant, suggesting the suitability of remote DLD and DD screenings (Hodge et al., 2019). This is of high relevance considering the accessibility of health care providing structures that can pose a challenge for families especially considering the resources required related to time and travel expenses, as well as to organizational issues (e.g. incompatibility with working hours).

Besides the fact that user experience for both examiners and examinees was generally described positively, limitations concerning usability relate to the system's speed. Some of the responses to the usability questionnaire suggest that the software from time to time appeared to be unresponsive. This may lead to uncertainties concerning safe data storage on the side of the examiners and boredom, frustration or disruption of the flow on the side of the examinee. When piloting the system within the team of researchers and developers, latency effects hardly occurred; however, researchers and developers were all working from well-equipped labs where a stable internet connection is granted through cabled networks, which might not be the case in all children's homes. It is thus hypothesized that the latency effects observed both from examiners' and examinees' perspectives relate to poor internet connections. Future studies should systematically assess whether properties like speed of the internet connection (through speed tests) and the system's working memory may cause systematic errors in the data collection. Despite the finding that examiners and examinees did mention the system speed to sometimes be too slow, for the data collected, significant associations between self-paced reading time in the screening and reading time in the standardized reading tasks emerged. This indicator of concurrent validity is interpreted such as that response time despite latency effects was reliably assessed.

Furthermore, alternative solutions concerning loading the data no longer with eager (all data loaded at once) but lazy loading methods might play a role in improving the system. Another solution would be to construct a progressive web app that allows for online screening construction and testing offline, being less affected by variation in internet connection. However, when defining the system's requirements, it was considered more appropriate to display all the contents on the screen at once. It must thus be considered carefully whether the lazy loading approach might result in scattered and non-parallel display of contents belonging to the same item which might result in systematic errors or bias in the children's responses since the children would be confronted with one of the item's components earlier than with another

one that requires longer loading times. This is of particular relevance when a single item's components build up on each other, like in the tasks following the WSIR-approach with GIF-files demonstrating figures that open and close their mouths while an audio file is played or the word as well as the orthographic form identification tasks.

Besides the system speed, some examiners bemoaned the lack of a feature to resume a screening session. When developing the system, this was somehow expected and again can be explained with certain considerations during the construction phase: It was considered relevant for the prevention of misuse of the screenings to not enable jumping back and forth within the screening which might result in the examiner – be it mistakenly or on purpose – to present the same item more than once to the same child which in turn might lead in systematic errors. It was therefore decided that the examiners' freedom in screening use would be restricted and that a screening can only be administered following the sequence of items and tasks defined by the admin on the respective screening platform interface. It is possible to restart a screening session any time. Post-hoc analyses revealed significant associations between the missing expected feature of resuming a session and perceived rigidity of the system ($n = 10$, $r = 0.837$, $p = 0.003$).

Finally, some of the examiners reported that they do not consider the handling of the screening platform appropriate for all skill levels of users. In a follow-up questionnaire, respondents explained that this concern is mostly due to the level of difficulty they observed when children responded to the screening items. This hypothesis will be further investigated and appropriateness of the level of difficulty of the tasks will be improved upon envisaged item selection.

6.2 Methodological discussion and future work

From the study results and interpretations described above, limitations of the study as well as concrete implications for future research and clinical practice derive.

6.2.1 Limitations

The most remarkable limitation across studies refers to the sample sizes which restrict the generalizability of results. In order to assess the suitability of the screenings implemented on a modifiable screening platform, the screenings were piloted on many small samples differing in country of residence, bilingual vs. monolingual schooling, language background, age and screening purpose. Having confirmed the potential of computerized DLD and DD screening in bilingual populations, in future studies applying the MuLiMi screening it is recommended to work with larger samples and select language combinations that also allow for the recruitment

of a sufficiently high number of children with DLD or DD diagnoses enabling statistical analyses to determine cut-offs for each screening task and levels of sensitivity and specificity.

Another limitation relates to the inappropriateness of the application of standardized tests normed for the monolingual population speaking the target language when assessing bilingual children (Garraffa et al., 2019; MultiMind ITN, 2022). In all studies, norm data based on monolingual populations was applied which a) resembles the average clinical practice (Jordaan, 2008; Stankova et al., 2021; Williams & McLeod, 2012) and b) can be explained by the lack of appropriate and time-efficient resources (Bloder et al., 2021). Despite the commonness of this approach as well as lack of alternatives, this procedure may appear circular and insufficient for the determination of DLD or DD risk. This limitation was encountered by also considering other variables like caregiver, SLT and teacher questionnaires (Pua et al., 2017) as well as pre-existing diagnoses (see chapter 5.3 and 5.4). The latter are expected to be based on standardized tests in the L2 for an estimation of language performance for functional assessment and, in addition, on thorough investigation of the developmental history, family risk, language background, language acquisition and linguistic behaviour in the L1. Although it cannot be excluded that SLTs and other figures in charge of the diagnostic processes have based their evaluation exclusively on L2 standardized tests (Williams & McLeod, 2012), consideration of all the mentioned information is considered to be the appropriate procedure for clinical diagnosis.

Levels of exposure to the two languages spoken were very diverse across and within the different language groups, which becomes most evident in the DD screening studies. The levels of automation of the reading processes in both languages might be very different and thus the interaction with Italian reading acquisition accordingly might vary as well across language groups. The most limiting factor in this respect is the fact that in the DD screening studies (chapters 5.5 and 5.6) and in the Spanish screening-study (chapter 5.3), language background questionnaires were not available. A thorough assessment of the child's language background however would have been relevant to assess the impact of language exposure on performance in the different language tests in the single languages tested. This missing element for the in-depth investigation of suitability of the screening platform use and screening administration is of particular importance considering the preliminary results on language dominance and screening task performance from the study conducted with German-speaking preschool children. For the German children, language exposure was systematically assessed with Bloder et al.'s (unpublished) language background questionnaire. The results indicate that language dominance (defined as ratio of in- and output in both languages

spoken) is not significantly correlated with NWRT performance and the majority of morpho-syntactic processing tasks and DNWL subtests. In the course of this study, it could not be thoroughly investigated whether it is in the nature of the tasks that they appeared to be unaffected by various levels of language exposure. In future studies on the properties of the screenings, the potential effect of exposure to the languages under assessment must be thoroughly investigated. It is furthermore desirable to increase the sample size allowing for more elaborate statistical analyses in which the quantity and quality of language input can be controlled for. Both the administration and evaluation of such detailed language background questionnaire can be effortful. In line with the digital nature of the project described, it is thus desirable to implement computerized versions of such tools that facilitate administration and analyses of responses in order to be able to directly relate them to screening results. Along the same lines, it might have been relevant to control for the language variety that the children tested were exposed to. This might be of particular relevance for the Spanish-speaking sample due to differences in phoneme inventories and thus in particular for our NWRT.

DA procedures for both DLD (see chapter 2.4.1.2) and DD risk identification (see chapter 2.4.2.2), due to time constraints, were only administered to sub-samples of Spanish-Italian- and Italian- German-speaking preschool children. Moreover, a DA screening task was implemented on the screening platform and piloted with a small number of mono- and bilingual Italian-speaking school-aged children. These pilot studies had the goal to test and improve the procedures for the administration of the task (feedback, repetition, looping) and for automatic scoring of the children's performance and do not allow for systematic description, analyses and interpretation of the results. Future studies should incorporate the DA task in the DD screening studies and thoroughly assess its potential in risk identification.

6.2.2 Implications for future work

The screening studies supports the appropriateness of computerized screening in DLD and DD risk identification, but replication of these preliminary validation studies on larger samples is needed to confirm the results presented and provide further indications for large scale implementation of this approach in SLT practice.

Future studies will also contribute to the identification of the best combination of methods relevant for the diagnosis of bilingual children: while standardized tests are necessary for the functional diagnostic component in order to establish and define therapy goals when SLT intervention is only provided in the L2, they seem to be of particular use when combined with information collected from teachers and/or caregivers concerning language/reading development, children's (linguistic) behaviour in class and groups as well as family risk and language

background (Bonifacci et al., 2020). The screening studies presented do show that direct assessment of both languages spoken in a comparable manner that does not require unpaid additional time for the familiarization with the child's L1 because of the application of computerized is appropriate for DLD and DD risk identification. The diagnostic value of such screenings can be further improved through item selection which will a) shorten the screenings and thus enhance applicability and b) improve the screenings potential in discriminating TD children from children with DLD and/or DD. This step is being carried out for the tasks for which sufficiently large samples and a sufficient number of children with a clinical diagnosis are ensured. For this purpose, in the case of NWRT in the Italian-German sample, in depth analyses of item categories give interesting indications related to the best combination of screening task items across languages: the combination of L1 and L2 LS items was more consistently associated with variables indicating DLD risk (risk level, standardized tests, questionnaires, Eikerling et al., 2022a) suggesting that in future studies, compound scores across languages might be applied also in other linguistic areas. While this is common practice in the field of lexicon (i.e. conceptual and total lexicon, Core et al., 2013), for morphosyntactic processing skills, specific, meaningful combinations still need to be identified.

Future studies should also take into consideration the potential interaction of task types and clinical markers chosen. While WSIR tasks in the Spanish- and English-Italian screening seemed to be appropriate, the grammaticality judgement tasks in the DD screening studies as well as combined grammaticality judgement and picture matching for subject-verb agreement tasks and picture matching for case marking tasks in the Italian-German screening seem to be less appropriate. In future studies, the same clinical markers for which good evidence was found in the previous literature should be implemented in the WSIR paradigms, that seem to be feasible for their implementation with young children (see chapter 5.3) as well as with older children (see chapter 5.6). Such a procedure would allow for the evaluation the appropriateness of screening task paradigms and clinical markers separately. Potentially, administering the screening tasks selected for the purpose of bilingual screenings to monolingual peers with and without DLD or DD could contribute to a better understanding of the functioning and appropriateness of the selected screening tasks. In-depth analyses of screening results and more specifically error patterns across aforementioned groups may give indications on similarities and differences between functioning patterns of mono- and bilingual children with and without DLD or DD and thus also contribute to more theory-driven research on typical and atypical (second) language acquisition.

The survey study with SLTs proved widespread awareness of the necessity of adaptation of SLTs' diagnostic and therapeutic practices and at the same time openness towards new methods, e.g., computerized screenings. Further studies are needed to investigate whether this discrepancy between knowledge and practice can be explained by limitations in resources that either relate to a) the costs for purchase of the material that potentially due to limited caseload of specific language-combinations might be applied only with a very small proportion of children or b) the additional effort and time assessing and evaluating the children's L1 which frequently is not spoken by the SLTs (Roseberry-McKibbin et al., 2005). In order to answer this question, it would have been beneficial to learn more about the SLTs' own language background and further demographics. More specific information is being collected from SLTs through additional and follow-up questionnaires. Despite these open questions, the SLTs' general openness to the assessment of both languages and application of computerized tools point to the potential applicability of bilingual computerized screening tools.

In order for computerized screenings to be actually used in the future, the problems revealed in the usability study related to system speed, internet connection and perceived rigidity of the system and screening administration should be solved. Usability studies with larger samples of examiners/examinees as well as direct assessment of usability as perceived by our pre-schoolers are needed. Usability could be assessed more thoroughly with observations of users, semi-standardized interviews and heuristic evaluations in addition to usability questionnaires. Despite these limitations, both in presence and remote administration of computerized screening as well as reliable recording of the children's reactions (as indicated by the associations between computerized screening and standardized test scores) seem feasible. The following two advantages could not be systematically assessed in the usability studies, but were observed by the student researchers administering the screenings. Other than in traditional reading and language tests, the examiner was not required to record the reading time and take notes on the child's responses during screening administration. Instead, the examiner could observe the child when reacting to the stimuli which in turn can provide some useful insights when planning further diagnostic procedures or the intervention. The automation of processes might be of particular relevance considering the background of dynamic assessment or adaptive testing, which is complex when carried out manually, but well controllable and feasible when automated (see potential of DNWL described in chapter 5.3. Further studies need to investigate the potential of DA in reading screenings (see chapter 6.2.1).

Other technical features could further improve the applicability of computerized screenings. These relate to the implementation of ASR mechanisms to allow for automatic evaluation of NW repetition performance as well as of other clinical markers that, due to their incompatibility with requirements concerning the automatic administration and evaluation, could not be integrated in the preliminary validation studies described here. Concerning the NWRT and its potential for ASR mechanisms, it is considered relevant to then conduct in depth-analyses of repetition accuracy on syllable- and phoneme-level which will allow for more systematic and useful insights concerning the children's articulatory skills.

7 Conclusion

The results of the screening studies underline that the identified need for the assessment of both languages of bilingual children for DLD and DD risk detection as indicated by previous literature, policy reports and as perceived by SLTs can be met by the bilingual screening tasks implemented on the MuLiMi platform. They offer solutions for various languages and orthographies, age groups and screening purposes. Furthermore, screenings can be expanded for other languages, age groups and screening purposes due to the modifiability of the screening platform. Most tasks implemented across linguistic areas were found to have good preliminary validity and seem to have the potential in discriminating different levels of DLD or DD risk across languages and linguistic areas. Preliminary data on predictivity furthermore suggests that performance in the screening tasks are generally representative of later language performance. On the whole, the associations between screening task performance and variables indicating DLD or DD risk were robust across age and language exposure patterns, also when controlling for nonverbal intelligence. Despite limitations concerning system speed, ease of familiarization and rigidity of the screening as identified in the usability studies, the screening platform was well received for both examiners and examinees. Overall, they described it as easy-to-use and pleasant which indicates the potential of automatic, efficient and non-complicated identification of DLD and DD risk in bilingual children increasing the probability of opportunities for (early) intervention.

8 Appendix

Appendix A contains the unpublished teachers, SLT and usability questionnaires used in this study. Appendix B contains example of the unpublished screening tasks used in this study.

8.1 Appendix A – Questionnaires

Appendix A1: Teacher & SLT questionnaire (Italian kindergarten version).I
 Appendix A2: Teacher questionnaire (Italian primary school version).....II
 Appendix A3: Items contained in the online survey on usability of MuLiMi for examinees.....III
 Appendix A4: Items contained in the online survey on usability of MuLiMi for examiners V

Valutazione del linguaggio del bambino		senza problemi	leggermente problematico	problem- atico	molto prob- lematico
sistema dei suoni/ fonologia	recettiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	produttiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
grammatica/ morfosintassi	recettiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	produttiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
lessico/ vocabolario/semantica	recettiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	produttiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
pragmatica		<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix A1: Teacher & SLT questionnaire (Italian kindergarten version).

Valutazione del linguaggio del bambino		senza problemi	leggermente problematico	problem- atico	molto prob- lematico
sistema dei suoni/ fonologia	recettiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	produttiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
grammatica/ morfosintassi	recettiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	produttiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
lessico/ vocabolario/semantica	recettiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	produttiva	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
pragmatica		<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Valutazione delle capacità di lettura e scrittura					
leggere ad alta voce		<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
scrittura		<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
comprensione della lettura		<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix A2: Teacher questionnaire (Italian primary school version).

Code	Question	Response scale (5-point-scale)
P1 ³	I think that I would need the support of a technical person to be able to use this system	Strongly agree - Strongly disagree
P2	Overall reaction to the Software	Terrible - Wonderful
P3	Overall reaction to the Software	Dull - Stimulating
P4	Learning to operate the system	Difficult - Easy
P5	System speed	Too slow - Fast enough
P6	It was simple to use this system	Strongly Disagree - Strongly Agree
P7	I feel comfortable using this system	Strongly Disagree - Strongly Agree
P8	Characters on the touch screen	Hard to read - Easy to read
P9	Screen elements are easy to select	Never - Always
P10	See the cursor (arrow) position	Difficult - Easy
P11	Feedback on whether an element is selected/clicked	Difficult - Easy
P12	Are the screen graphics pleasant?	Not at all - Very much

Appendix A3: Items contained in the online survey on usability of MuLiMi for examinees, adapted from (Chin et al., 1988; Sauro, 2022; Wallace et al., 1988).

³ In order to facilitate the interpretation of the responses, the response options of this question were inverted so that a higher score is always representative of a positive user experience.

Code	Question	Response Scale	Scale
E1	How many times have you administered a screening using the MuLiMi screening platform?	Once once to 5 times more than 5 times	N/A
E2 ³	I found the system unnecessarily complex	Strongly Disagree - Strongly Agree	1-5
E3	I think the system was easy to use	Strongly Disagree - Strongly Agree	1-5
E4 ³	I think that I would need the support of a technical person to be able to use this system	Strongly Disagree - Strongly Agree	1-5
E5	I found the various functions of this system to be well integrated	Strongly Disagree - Strongly Agree	1-5
E6	I would imagine that most people would learn to use this system very quickly	Strongly Disagree - Strongly Agree	1-5
E7 ³	I found the system very cumbersome to use	Strongly Disagree - Strongly Agree	1-5
E8	I felt very confident using the system	Strongly Disagree - Strongly Agree	1-5
E9 ³	I needed to learn a lot of things before I could get going with this system	Strongly Disagree - Strongly Agree	1-5
E10	Overall reaction to the Software	Terrible - Wonderful	1-9
E11	Overall reaction to the Software	Difficult - Easy	1-9
E12	Overall reaction to the Software	Frustrating - Satisfying	1-9
E13	Overall reaction to the Software	Dull - Stimulating	1-9
E14	Overall reaction to the Software	Rigid - Flexible	1-9
E15	Reading characters on the screen	Hard - Easy	1-9
E16	Organization of information	Confusing - Very Clear	1-9
E17	Learning to operate the system	Difficult - Easy	1-9
E18	Performing tasks is straightforward	Never - Always	1-9
E19	System speed	Too slow - Fast enough	1-9
E20	System reliability	Unreliable - Reliable	1-9
E21	Designed for all levels of users	Never - Always	1-9
E22	I am able to efficiently complete my work using this system	Strongly Disagree - Strongly Agree	1-9
E23	I feel comfortable using this system	Strongly Disagree - Strongly Agree	1-9
E24	The interface of this system is pleasant	Strongly Disagree - Strongly Agree	1-9
E25	This system has all the functions and capabilities I expect it to have	Strongly Disagree - Strongly Agree	1-9
E26	Overall, I am satisfied with this system	Strongly Disagree - Strongly Agree	1-9
E27	Image/resolution of characters	Fuzzy - Sharp	1-9
E28	Character shape (font)	Barely legible - Very legible	1-9
E29	Contrast with the background	Irritating - Pleasant	1-9
E30	Highlighting on the screen makes task easier	Not at all - Very much	1-9

E31	Use of color for highlighting	Unhelpful - Helpful	1-9
E32	Screen layouts make tasks easier	Never - Always	1-9
E33	Amount of information displayed on screen	Inadequate - Adequate	1-9
E 34	Arrangement of information on screen	Illogical - Logical	1-9
E35	Screen items are easy to select	Never - Always	1-9
E36	Screen items are easy to find	Never - Always	1-9
E 37	Item selection area size	Too small - Large enough	1-9
E 38 ²	Seeing cursor location	Easy - Difficult	1-9
E39	Knowing whether an item has been selected	Difficult - Easy	1-9
E40	System responds to selection when finger is pulled away from screen	Unreliably - Reliably	1-9
E41	Learning to operate the system	Difficult - Easy	1-9
E42	Time to learn to use the system	Too long - Just right	1-9
E 43	System response time for most operations	Too slow - Fast enough	1-9
E44	Screen designs and layout are attractive	Not at all - Very much	1-9
E 45	Use of colours	Unattractive - Attractive	1-9
E 46	System is fun to use	Not at all - Very much so	1-9

Appendix A4: Items contained in the online survey on usability of MuLiMi for examiners, adapted from (Chin et al., 1988; Sauro, 2022; Wallace et al., 1988).

8.2 Appendix B – Screening items

Appendix B1: Examples from the Spanish & Italian WSIR subject-verb agreement tasks..	VII
Appendix B2: Examples from the Spanish & Italian WSIR finiteness tasks.....	VII
Appendix B3: Examples from the German case matching task.....	VIII
Appendix B4: Examples from the subject-verb agreement tasks (DLD screening).....	IX
Appendix B5: Examples from the subject-verb agreement tasks (DD screening).....	IX
Appendix B6: Examples from the Italian pronoun judgement task (DLD screening).....	X
Appendix B7: Examples from the Italian pronoun judgement task (DD screening).....	X
Appendix B8: Examples from the English tense judgement task (WSIR paradigm).	XI
Appendix B9: Examples from the Italian, Spanish & German DNWL testing phase.	XI
Appendix B10: Object pictures used in the DNWL testing phase.....	XII
Appendix B11: Scene pictures used in the above DNWL testing phase.....	XII
Appendix B12: Examples from the Italian & German self-paced syllable reading tasks. ...	XIII
Appendix B 13: Examples from the self-paced sentence reading tasks.	XIII
Appendix B14: Examples from the Italian & German word identification tasks.....	XIII
Appendix B15: Examples from the Italian & German NW identification tasks.....	XIII
Appendix B16: Examples from the Mandarin character judgement tasks.....	XIV
Appendix B17: Examples from the English orthographic form identification tasks.....	XIV
Appendix B18: Examples from the English phonological form identification tasks.	XIV
Appendix B19: Examples from the Italian & German phonological awareness tasks.	XV
Appendix B20: Examples from the Italian & English word stress identification tasks.....	XV
Appendix B21: Examples from English sound deletion task.....	XV
Appendix B22: Examples from Mandarin phonological awareness tasks.....	XV

8.2.1 Morphosyntactic processing tasks

Find here examples from the morphosyntactic processing tasks for the different clinical markers, task paradigms, languages and age groups.

8.2.1.1 WSIR – subject-verb agreement

Language	Target response	Distractor
Italian	lo gioco a palla.	lo giochi* a palla.
	lo scendo subito.	Lei scendo* subito.
	Loro vivono in Italia.	Loro vive* in Italia.
Spanish	El duerme mucho.	El duermen* mucho.
	Ella lee el libro.	Ella leen* el libro.
	Tu hablas inglés.	Tu hablo* inglés.



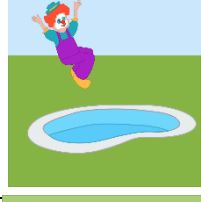
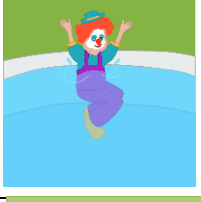


Appendix B1: Examples from the Spanish and Italian WSIR subject-verb agreement tasks.

8.2.1.2 WSIR – finiteness

Language	Target response	Distractor
Italian	Lei chiude la porta.	Lei chiudere* la porta.
	Lui corre veloce.	Lui correre* veloce.
	Tu porti la borsa.	Tu portare* la borsa.
Spanish	El come pollo.	El comer* pollo.
	Ella cae siempre,	Ella caer* siempre.
	Tu llevas la mochila .	Tu llevar* la mochila.

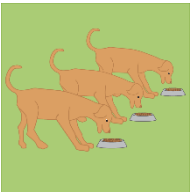
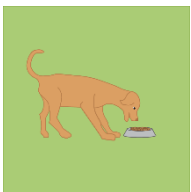



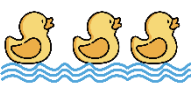
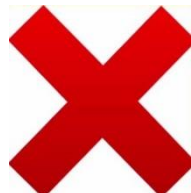
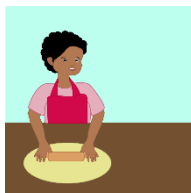


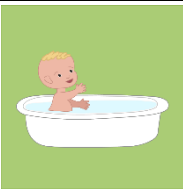

Appendix B2: Examples from the Spanish and Italian WSIR finiteness tasks.

8.2.1.3 Case matching

DLD screening version	DD screening version	Target	Distractor
Das Baby krabbelt in der Pfütze.	Das Baby krabbelt in dem Teich.		
Der Clown springt in das Becken	Der Clown springt in den Pool		
Die Biene fliegt in die Tasse.	Die Biene fliegt in den Becher.		

Appendix B3: Examples from the German case matching task.

8.2.1.4 Subject-verb agreement




Stimulus	Target	Distractor 1	Distractor 2
I cani bevono.			
La papera nuotano*.			
Die Mama backen*.			
Die Babys baden.			

Appendix B4: Examples from the German and Italian subject-verb agreement tasks (DLD screening).

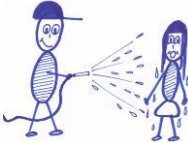
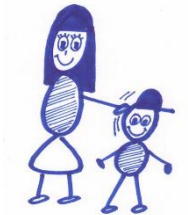

Stimulus	Expected response
Il lupo grigio corre nel bosco.	✓
I pesci colorati nuota nel mare.	×
La maestra buona scrivono alla lavagna.	×

Appendix B5: Examples from the Italian subject-verb agreement tasks (DD screening).

8.2.1.5 Clitic pronoun judgement

Stimulus	Visual support	Expected response
Che cosa fa il bambino la bambina? Lo* colora.		x
Che cosa fa la donna all'uomo? Le legge un libro.		✓
Che cosa fa il bambino al pesce? Gli dà da mangiare.		✓

Appendix B6: Examples from the Italian pronoun judgement task (DLD screening).

Stimulus	Visual support	Expected Response
Che cosa fa il bambino alla bambina? Le* bagna.		x
Che cosa fa la mamma al bambino? Lo accarezza.		✓
Che cosa dà il bambino alla bambina? La* dà i fiori.		x

Appendix B7: Examples from the Italian pronoun judgement task (DD screening).

8.2.1.6 Tense judgement

Target	Distractor
Yesterday he raked the leafs.	Yesterday he rakes* the leafs.
When the sun went down we started to freeze.	When the sun went down we start* to freeze.
As long as she lived there she never answered my emails.	As long as she lived there she never answers* my emails*.

Appendix B8: Examples from the English tense judgement task (WSIR paradigm).

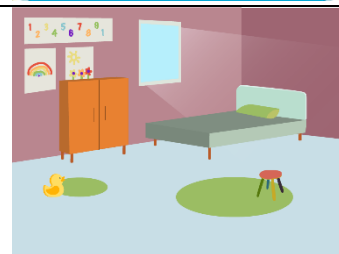
8.2.2 Dynamic Novel Word Learning (testing phase)

Language	Scene	Prompt
Italian	Corridor	Galpo vuole uscire di casa. A chi dai le scarpe?
	Outside	Felio vede che piove. A chi dai l'ombrello?
	Pool	Domio vuole nuotare. A chi dai il costume da bagno?
Spanish	Playground	Mokal tiene sed. ¿A quién le das el vaso?
	Living room	Nuelo quiere sentarse. ¿A quién le das la silla?
	Bedroom	Flado quiere dormir. ¿A quién le das el pijama?
German	Playground	Pamolt hat Durst. Wem gibst du das Glas?
	Living room	Harock möchte Papier schneiden. Wem gibst du die Schere?
	Bedroom	Tumach friert. Wem gibst du die Decke?

Appendix B9: Examples from the Italian, Spanish and German DNWL testing phase. See corresponding pictures in appendix B10.



Appendix B10: Object pictures used in the above mentioned examples from the Italian, Spanish and German DNWL testing phase. See corresponding scenes in appendix B11.



Appendix B11: Scene pictures used in the above mentioned examples from the Italian, Spanish and German DNWL testing phase.

8.2.3 Reading tasks

Language	Syllables					
Italian	si	me	ca	lo	gu	si
German	ju	we	bä	wö	dau	pfa

Appendix B12: Examples from the Italian and German self-paced syllable reading tasks.

Language	Easy	Medium	Complex
Italian	La farfalla vola sui fiori colorati.	Le bambine giocano felici con le loro amiche.	Alcuni ragazzi corrono veloci nel giardino della scuola
German	Die Katze trinkt von der kalten Milch.	Die Hunde laufen auf dem grünen Gras.	Kein Auto fährt schnell auf den Parkplatz.
English	A horse eats green grass.	The dog is barking loudly at the neighbour.	Some children are climbing quickly on top of a tree.

Appendix B 13: Examples from the Italian, German and English self-paced sentence reading tasks.

	Stimulus	Target	Distractor 1	Distractor 1
Italian	[ˈpeʃe]	pesce	pece	pesche
	[ˈfɔʎa]	foglia	voglia	fogna
German	[li:t]	Leid	Lied	Leib
	[ˈʃaɪnən]	scheinen	schienen	schneien

Appendix B14: Examples from the Italian and German word identification tasks.

	Stimulus	Target	Distractor 1	Distractor 1
Italian	[ˈke:ɾdo]	cherdo	cerdo	gherdo
	[pu:ˈla dʒe]	pulage	bulage	pulaghe
German	[ˈʃvyke]	schwücker	schwucker	schwücher
	[ˈʃri:sal]	schrießal	schießal	schriebal

Appendix B15: Examples from the Italian and German NW identification tasks.

Task type	Stimulus	Expected response
Radical Position	南	x
	河	✓
	你	✓
Left-right inversion	妈	✓
	鸡	x
	玩	✓

Appendix B16: Examples from the Mandarin character judgement tasks.

Stimulus	Target	Distractor 1	Distractor 2
The cat is on the road.	road	rode	wroad
The parents decided what the children had to wear.	wear	where	whare
In the park the children saw few squirrels running by.	by	bye	bhy

Appendix B17: Examples from the English orthographic form identification tasks.

Stimulus	Target	Distractor 1	Distractor 2
In class, sometimes teacher project slide shows.	[prə'dʒɛkt]	['prɒdʒɛkt]	[prə'tɛkt]
Before going to the wedding, he wants to polish his shoes.	['pɒlɪʃ]	['pəʊlɪʃ]	['dɒlɪʃ]
During workout, usually muscles contract.	[kən'trækt]	['kɒntrækt]	['kɒntækt]

Appendix B18: Examples from the English phonological form identification tasks.

Screening Task	Language	Stimulus	Expected answer
Phoneme blending	Italian	t i g r e	tigre ✓
		p e r s e	prese ✗
	German	b r a t e n	traben ✗
		m a n t e l	Mantel ✓
Syllabic inversion	Italian	lido	doldi ✗
		pila	lapi ✓
	German	Del-fin	fin-del ✓
		schmut-zig	tzig-schum ✗

Appendix B19: Examples from the Italian and German phonological awareness tasks.

Language	Stimulus	Option 1	Option 2	Option 3
Italian	sa'pone	sa*	po	ne*
	'favola	fa	vo*	la*
English	a'void	a*	void	n.a.
	'silent	si	lent*	n.a.

Appendix B20: Examples from the Italian and English word stress identification tasks.

Stimulus	sound to be deleted	Target	Distractor
black	/b/	lack	back
clamp	/l/	camp	lamp
ground	/g/	round	gound

Appendix B21: Examples from English sound deletion task.

Task type	Target	Option 1	Option 2	Option 3
Onset detection	wō	ruì	wèi	wén
	luò	nán	nuò	níng
Rhyme detection	qín	xīng	píng	mìng
	chuáng	juǎn	suàn	kuān
Tone detection	mǎng	qiàn	sì	tòng
	guō	háng	luó	bó

Appendix B22: Examples from Mandarin phonological awareness tasks.

9 References

- Abed Ibrahim, L., & Hamann, C. (2017). Bilingual Arabic-German and Turkish-German with and without Specific Language Impairment: Comparing Performance in Sentence Repetition and Nonword Repetition Tasks. In *Proceedings of the 41st Annual Boston University Conference on Language Development* (pp. 1–17). Somerville, MA: Cascadia Press.
- Adlof, S. M., & Hogan, T. P. (2018). Understanding Dyslexia in the Context of Developmental Language Disorders. *Language, Speech, and Hearing Services in Schools, 49*(4), 762–773. https://doi.org/10.1044/2018_LSHSS-DYSLC-18-0049
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5.: 5th edition*. American Psychiatric Association.
- Araújo, S., & Faisca, L. (2019). A Meta-Analytic Review of Naming-Speed Deficits in Developmental Dyslexia. *Scientific Studies of Reading, 23*(5), 349–368. <https://doi.org/10.1080/10888438.2019.1572758>
- Armon-Lotem, S. (2012). Introduction: Bilingual children with SLI – the nature of the problem. *Bilingualism: Language and Cognition, 15*(1), 1–4. <https://doi.org/10.1017/S1366728911000599>
- Arosio, F., Pagliarini, E., Perugini, M., Barbieri, L., & Guasti, M. T. (2016). Morphosyntax and logical abilities in Italian poor readers: The problem of SLI under-identification. *First Language, 36*(3), 295–315. <https://doi.org/10.1177/0142723716639501>
- Arrhenius, B., Gyllenberg, D., Chudal, R., Lehti, V., Sucksdorff, M., Sourander, O., Virtanen, J.-P., Torsti, J., & Sourander, A. (2018). Social risk factors for speech, scholastic and coordination disorders: A nationwide register-based study. *BMC Public Health, 18*(1), 739. <https://doi.org/10.1186/s12889-018-5650-z>
- Barbiero, C., Lonciari, I., Montico, M., Monasta, L., Penge, R., Vio, C., Tressoldi, P. E., Ferluga, V., Bigoni, A., Tullio, A., Carrozzi, M., & Ronfani, L. (2012). The submerged dyslexia iceberg: How many school children are not diagnosed? Results from an Italian study. *PloS One, 7*(10), e48082. <https://doi.org/10.1371/journal.pone.0048082>
- Barbiero, C., Montico, M., Lonciari, I., Monasta, L., Penge, R., Vio, C., Tressoldi, P. E., Carrozzi, M., de Petris, A., de Cagno, A. G., Crescenzi, F., Tinarelli, G., Leccese, A., Pinton, A., Belacchi, C., Tucci, R., Musinu, M., Tossali, M. L., Antonucci, A. M., . . . Ronfani, L. (2019). The lost children: The underdiagnosis of dyslexia in Italy. A cross-sectional national study. *PloS One, 14*(1), e0210448. <https://doi.org/10.1371/journal.pone.0210448>

- Bastien-Toniazzo, M., Stroumza, A., & Cavé, C. (2010). Audio-Visual Perception and Integration in Developmental Dyslexia: An Exploratory Study Using the McGurk Effect. *Current Psychology Letters*. Advance online publication. <https://doi.org/10.4000/cpl.4928>
- Baumert, J., & Maaz, K. (2012). Migration und Bildung in Deutschland. *DDS – Die Deutsche Schule*, 104(3), 279–302.
- Bedore, L. M., & Leonard, L. B. (2001). Grammatical Morphology Deficits in Spanish-Speaking Children With Specific Language Impairment. *Journal of Speech, Language, and Hearing Research*, 44(4), 905–924. [https://doi.org/10.1044/1092-4388\(2001/072\)](https://doi.org/10.1044/1092-4388(2001/072))
- Bedore, L. M., & Peña, E. D. (2008). Assessment of Bilingual Children for Identification of Language Impairment: Current Findings and Implications for Practice. *International Journal of Bilingual Education and Bilingualism*, 11(1), 1–29. <https://doi.org/10.2167/beb392.0>
- Belacchi, C., Scalisi, T. G., Cannoni, E., & Cornoldi, C. (2008). *CPM-Coloured Progressive Matrices*. Standardizzazione italiana. Giunti O.S.
- Ben-Dror, I., Bentin, S., & Frost, R. (1995). Semantic, Phonologic, and Morphologic Skills in Reading Disabled and Normal Children: Evidence from Perception and Production of Spoken Hebrew. *Reading Research Quarterly*, 30(4), 876–893. <https://doi.org/10.2307/748202>
- Bigagli, A., & Lorusso, M. L. (2014). *Predittori Della Lettura in Italiano L2 in Bambini di Madrelingua Cinese*. XXIII Congresso Nazionale AIRIPA, Lucca.
- Bishop, D. V. M. (2015). The interface between genetics and psychology: Lessons from developmental dyslexia. *Proceedings. Biological Sciences*, 282(1806), 20143139. <https://doi.org/10.1098/rspb.2014.3139>
- Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 58(10), 1068–1080. <https://doi.org/10.1111/jcpp.12721>
- BISLI Cost Action. (2022). *LITMUS Network*. <https://www.bi-sli.org/>
- Bloder, T., Eikerling, M., Rinker, T., & Lorusso, M. L. (2021). Speech and Language Therapy Service for Multilingual Children: Attitudes and Approaches across Four European Countries. *Sustainability*, 13(21), 12143. <https://doi.org/10.3390/su132112143>
- Blom, E., Boerma, T., & de Jong, J. (2019). First language attrition and Developmental Language Disorder. *Oxford Handbooks in Linguistics*. <https://www.narcis.nl/publication/RecordID/oai:dare.uva.nl:publications%2F47422d74-4d04-4b01-ac75-340ea8f3734a>

- Boerma, T., & Blom, E. (2017). Assessment of bilingual children: What if testing both languages is not possible? *Journal of Communication Disorders*, 66, 65–76.
<https://doi.org/10.1016/j.jcomdis.2017.04.001>
- Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children Learning Dutch as a Second Language. *Journal of Speech, Language, and Hearing Research*, 58(6), 1747–1760. https://doi.org/10.1044/2015_JSLHR-L-15-0058
- Boivin, M. J., Kakooza, A. M., Warf, B. C., Davidson, L. L., & Grigorenko, E. L. (2015). Reducing neurodevelopmental disorders and disability through research and interventions. *Nature*, 527(7578), S155-60. <https://doi.org/10.1038/nature16029>
- Bonifacci, P., Atti, E., Casamenti, M., Piani, B., Porrelli, M., & Mari, R. (2020). Which Measures Better Discriminate Language Minority Bilingual Children With and Without Developmental Language Disorder? A Study Testing a Combined Protocol of First and Second Language Assessment. *Journal of Speech, Language, and Hearing Research*, 63(6), 1898–1915. https://doi.org/10.1044/2020_JSLHR-19-00100
- Bonifacci, P., Mari, R., Gabbianelli, L., Ferraguti, E., & Porrelli, M. (2016). Sequential bilingualism and Specific Language Impairment: The Italian version of ALDeQ Parental Questionnaire. *Applied Psychology Bulletin*, 64(275), 50–63.
- Bonifacci, P., Tobia, V., Lami, L., & Snowling, M. (2014). *ALCE. Assessment di Lettura e Comprensione per l'Età Evolutiva*. Hogrefe. <https://iris.univr.it/handle/20.500.11768/88299>
- Bortolini, U., Arfé, B., Caselli, C. M., Degasperi, L., Deevy, P., & Leonard, L. B. (2006). Clinical markers for specific language impairment in Italian: The contribution of clitics and non-word repetition. *International Journal of Language & Communication Disorders*, 41(6), 695–712. <https://doi.org/10.1080/13682820600570831>
- Bortolini, U., Caselli, M. C., Deevy, P., & Leonard, L. B. (2002). Specific language impairment in Italian: The first steps in the search for a clinical marker. *International Journal of Language & Communication Disorders*, 37(2), 77–93.
<https://doi.org/10.1080/13682820110116758>
- Bortolini, U., Caselli, M. C., & Leonard, L. B. (1997). Grammatical deficits in Italian-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 40(4), 809–820. <https://doi.org/10.1044/jslhr.4004.809>
- Brookes, G., Ng, V., Lim, B., Hong Tan, W. P., & Lukito, N. (2011). The computerised-based Lucid Rapid Dyslexia Screening for the identification of children at risk of dyslexia: A

- Singapore study. *Educational and Child Psychology*, 28(2), 33–51.
<https://psycnet.apa.org/record/2011-16938-004>
- Bulheller, S., & Häcker, H. (2001). *CPM Raven's Progressive Matrices and Vocabulary Scales - Coloured Progressive Matrices von John Carlyle Raven*. Pearson.
- Camilleri, B., Hasson, N., & Dodd, B. (2014). Dynamic Assessment of bilingual children's language at the point of referral. *Educational & Child Psychology*, 31(2), 57–72.
<https://openaccess.city.ac.uk/id/eprint/4224/>
- Cantiani, C., Lorusso, M. L., Perego, P., Molteni, M., & Guasti, M. T. (2013). Event-related potentials reveal anomalous morphosyntactic processing in developmental dyslexia. *Applied Psycholinguistics*, 34(6), 1135–1162.
<https://doi.org/10.1017/S0142716412000185>
- Carroll, S. E. (2017). Exposure and input in bilingual development. *Bilingualism: Language and Cognition*, 20(1), 3–16. <https://doi.org/10.1017/S1366728915000863>
- Chiat, S. (2015). Nonword Repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Multilingual Matters.
- Chin, J. P., Diehl, V. A., & Norman, L. K. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88*. ACM Press.
<https://doi.org/10.1145/57167.57203>
- Chung, K. K. H., & Ho, C. S.-H. (2010). Second language learning difficulties in Chinese children with dyslexia: What are the reading-related cognitive skills that contribute to English and Chinese word reading? *Journal of Learning Disabilities*, 43(3), 195–211.
<https://doi.org/10.1177/0022219409345018>
- Chung, K. K. H., Ho, C. S.-H., Chan, D. W., Tsang, S.-M., & Lee, S.-H. (2010). Cognitive profiles of Chinese adolescents with dyslexia. *Dyslexia*, 16(1), 2–23.
<https://doi.org/10.1002/dys.392>
- Clahsen, H., Bartke, S., & Göllner, S. (1997). Formal features in impaired grammars: A comparison of English and German SLI children. *Journal of Neurolinguistics*, 10(2-3), 151–171. [https://doi.org/10.1016/S0911-6044\(97\)00006-7](https://doi.org/10.1016/S0911-6044(97)00006-7)
- Cline, T. (2000). Multilingualism and dyslexia: Challenges for research and practice. *Dyslexia*, 6(1), 3–12.
[https://doi.org/10.1002/\(SICI\)1099-0909\(200001/03\)6:1<3::AID-DYS156>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1099-0909(200001/03)6:1<3::AID-DYS156>3.0.CO;2-E)
- Contento, S., Bellocchi, S., & Bonifacci, P. (2013). *BaBIL. Prove per la valutazione delle competenze verbali e non verbali in bambini bilingui*. Guinti Psychometrics.

- Core, C., Hoff, E., Rumiche, R., & Señor, M. (2013). Total and Conceptual Vocabulary in Spanish–English Bilinguals From 22 to 30 Months: Implications for Assessment. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1637–1649. [https://doi.org/10.1044/1092-4388\(2013/11-0044\)](https://doi.org/10.1044/1092-4388(2013/11-0044))
- Costenaro, V., & Pesce, A. (2012). Dyslexia and the Phonological Deficit Hypothesis Developing Phonological Awareness in Young English Language Learners. *Undefined*. <https://www.semanticscholar.org/paper/Dyslexia-and-the-Phonological-Deficit-Hypothesis-in-Costenaro-Pesce/5a8fd9c687ee27203ec43b823604efb0aaa85f1c>
- Cummins, J. (1980). The Cross-Lingual Dimensions of Language Proficiency: Implications for Bilingual Education and the Optimal Age Issue. *TESOL Quarterly*, *14*(2), 175. <https://doi.org/10.2307/3586312>
- Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. *Schooling and Language Minority Students. A Theoretical Framework*. <https://ci.nii.ac.jp/naid/10029777893/>
- Cummins, J. (2012). The intersection of cognitive and sociocultural factors in the development of reading comprehension among immigrant students. *Reading and Writing*, *25*(8), 1973–1990. <https://doi.org/10.1007/s11145-010-9290-7>
- Cummins, J. (2016). Reflections on Cummins (1980), "The Cross-Lingual Dimensions of Language Proficiency: Implications for Bilingual Education and the Optimal Age Issue". *TESOL Quarterly*, *50*(4), 940–944. <http://www.jstor.org/stable/44984725>
- Da Silva, P. B., Engel de Abreu, Pascale M. J., Laurence, P. G., Nico, M. Â. N., Simi, L. G. V., Tomás, R. C., & Macedo, E. C. (2020). Rapid Automatized Naming and Explicit Phonological Processing in Children With Developmental Dyslexia: A Study With Portuguese-Speaking Children in Brazil. *Frontiers in Psychology*, *11*, 928. <https://doi.org/10.3389/fpsyg.2020.00928>
- Daniels, P. T., & Share, D. L. (2018). Writing System Variation and Its Consequences for Reading and Dyslexia. *Scientific Studies of Reading*, *22*(1), 101–116. <https://doi.org/10.1080/10888438.2017.1379082>
- De Almeida, L., Ferré, S., Morin, E., Prévost, P., Santos, C. d., Tuller, L., Zebib, R., & Barthez, M.-A. (2017). Identification of bilingual children with Specific Language Impairment in France. *Linguistic Approaches to Bilingualism*, 331–358. <https://doi.org/10.1075/lab.15019.alm>
- De Lamo White, C., & Jin, L. (2011). Evaluation of speech and language assessment approaches with bilingual children. *International Journal of Language & Communication Disorders*, *46*(6), 613–627. <https://doi.org/10.1111/j.1460-6984.2011.00049.x>

- Del Tufo, S. N., & Earle, F. S. (2020). Skill Profiles of College Students With a History of Developmental Language Disorder and Developmental Dyslexia. *Journal of Learning Disabilities, 53*(3), 228–240. <https://doi.org/10.1177/0022219420904348>
- Di Folco, C., Guez, A., Peyre, H., & Ramus, F. (2020). *Epidemiology of developmental dyslexia: A comparison of DSM-5 and ICD-11 criteria*. <https://doi.org/10.1101/2020.12.18.20248189>
- Dispaldro, M., Leonard, L. B., & Deevy, P. (2013). Clinical markers in Italian-speaking children with and without specific language impairment: A study of non-word and real word repetition as predictors of grammatical ability. *International Journal of Language & Communication Disorders, 48*(5), 554–564. <https://doi.org/10.1111/1460-6984.12032>
- D'Souza, C., Kay-Raining Bird, E., & Deacon, H. (2012). Survey of Canadian speech-language pathology service delivery to linguistically diverse clients. *Canadian Journal of Speech-Language Pathology and Audiology, 36*(1), 18–39.
- Durkin, C. (2000). Dyslexia and bilingual children? does recent research assist identification? *Dyslexia, 6*(4), 248–267. [https://doi.org/10.1002/1099-0909\(200010/12\)6:4<248::AID-DYS173>3.0.CO;2-O](https://doi.org/10.1002/1099-0909(200010/12)6:4<248::AID-DYS173>3.0.CO;2-O)
- Ebert, K. D. (2014). Role of auditory non-verbal working memory in sentence repetition for bilingual children with primary language impairment. *International Journal of Language & Communication Disorders, 49*(5), 631–636. <https://doi.org/10.1111/1460-6984.12090>
- Ebert, K. D., & Kohnert, K. (2016). Language learning impairment in sequential bilingual children. *Language Teaching, 49*(3), 301–338. <https://doi.org/10.1017/S0261444816000070>
- Eikerling, M., Bloder, T., & Lorusso, M. L. (2022a). A Nonword Repetition Task Discriminates Typically Developing Italian-German Bilingual Children from Bilingual Children with Developmental Language Disorder: The Role of Language-Specific and Non-Language-Specific Nonwords. *Frontiers in Psychology, 13*:826540. <https://doi.org/10.3389/fpsyg.2022.826540>
- Eikerling, M., & Lorusso, M. L. (in press). A web-platform for DLD screening in Italian-German-speaking children: preliminary data on concurrent and predictive validity, *Lingue e Linguaggio, 11*(1), 193.
- Eikerling, M., & Lorusso, M. L. (2021). Computergestütztes, bilinguals Lese-Screening mit der Screening – Plattform MuLiMi. *Sprachtherapie aktuell, 2021*(2), 1–10. <https://doi.org/10.14620/stadbs210740>

- Eikerling, M., Secco, M., Marchesi, G., Guasti, M. T., Vona, F., Garzotto, F., & Lorusso, M. L. (2022b). Remote Dyslexia Screening for Bilingual Children. *Multimodal Technologies and Interaction*, 6(1), 7. <https://doi.org/10.3390/mti6010007>
- Eikerling, M., & Wendt, C. (2016). Einflussfaktoren auf Lese- und Rechtschreibschwierigkeiten. *Forum Logopädie*, 30(4), 34–39. https://www.dbl-ev.de/fileadmin/inhalte/fl_archiv/2016/4/fl_2016_04_eikerling_neu.pdf
- Elbro, C., Daugaard, H. T., & Gellert, A. S. (2012). Dyslexia in a second language?—a dynamic test of reading acquisition may provide a fair answer. *Annals of Dyslexia*, 62(3), 172–185. <https://doi.org/10.1007/s11881-012-0071-7>
- Engel de Abreu, P. (2011). Working memory in multilingual children: Is there a bilingual effect? *Memory*, 19(5), 529–537. <https://doi.org/10.1080/09658211.2011.590504>
- Eurostat. (2022). *Migration and migrant population statistics*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Migration_and_migrant_population_statistics#Migrant_population:_23_million_non-EU_citizens_living_in_the_EU_on_1_January_2020
- Everatt, J., Smythe, I., Adams, E., & Ocampo, D. (2000). Dyslexia screening measures and bilingualism. *Dyslexia*, 6(1), 42–56. [https://doi.org/10.1002/\(SICI\)1099-0909\(200001/03\)6:1<42::AID-DYS157>3.0.CO;2-0](https://doi.org/10.1002/(SICI)1099-0909(200001/03)6:1<42::AID-DYS157>3.0.CO;2-0)
- Farabolini, G., Rinaldi, P., Caselli, M. C., & Cristia, A. (2021). Non-word repetition in bilingual children: the role of language exposure, vocabulary scores and environmental factors. *Speech, Language and Hearing*, 1–16. <https://doi.org/10.1080/2050571X.2021.1879609>
- Fondazione ISMU. (2022, March 11). *Dati sulle migrazioni: immigrati in Italia ed in Europa - Fond. ISMU*. <https://www.ismu.org/dati-sulle-migrazioni/>
- Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1), 104–115. <https://doi.org/10.1037/0096-1523.13.1.104>
- Gagarina, N. (2014). Diagnostik von Erstsprachkompetenzen im Migrationskontext. In S. Chilla (Ed.), *Handbuch Spracherwerb und Sprachentwicklungsstörungen: Vol. 3. Mehrsprachigkeit*. Elsevier Urban & Fischer.
- Garraffa, M., Vender, M., Sorace, A., & Guasti, M. T. (2019). *Is it possible to differentiate multilingual children and children with Developmental Language Disorder?* Apollo - University of Cambridge Repository. <https://doi.org/10.17863/CAM.37928>

- Gellert, A. S., & Elbro, C. (2018). Predicting reading disabilities using dynamic assessment of decoding before and after the onset of reading instruction: A longitudinal study from kindergarten through grade 2. *Annals of Dyslexia*, *68*(2), 126–144. <https://doi.org/10.1007/s11881-018-0159-9>
- Georgiou, G. K., Papadopoulos, T. C., Fella, A., & Parrila, R. (2012). Rapid naming speed components and reading development in a consistent orthography. *Journal of Experimental Child Psychology*, *112*(1), 1–17. <https://doi.org/10.1016/j.jecp.2011.11.006>
- Gersten, R., & Geva, E. (2003). Teaching reading to early language learners. *Educational Leadership: Journal of the Department of Supervision and Curriculum Development, N.E.A.*, *60*(7), 44–49.
- Geva, E. (2000). Issues in the assessment of reading disabilities in L2 children? Beliefs and research evidence. *Dyslexia*, *6*(1), 13–28. [https://doi.org/10.1002/\(SICI\)1099-0909\(200001/03\)6:1<13::AID-DYS155>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1099-0909(200001/03)6:1<13::AID-DYS155>3.0.CO;2-6)
- Giguere, D., & Hoff, E. (2020). Home language and societal language skills in second-generation bilingual adults. *International Journal of Bilingualism*, *24*(5-6), 1071–1087. <https://doi.org/10.1177/1367006920932221>
- Girbau, D. (2016). The Non-word Repetition Task as a clinical marker of Specific Language Impairment in Spanish-speaking children. *First Language*, *36*(1), 30–49. <https://doi.org/10.1177/0142723715626069>
- Goodrich, J. M., & Lonigan, C. J. (2017). Language-Independent and Language-Specific Aspects of Early Literacy: An Evaluation of the Common Underlying Proficiency Model. *Journal of Educational Psychology*, *109*(6), 782–793. <https://doi.org/10.1037/edu0000179>
- Goodz, N. S. (1989). Parental language mixing in bilingual families. *Infant Mental Health Journal*, *10*(1), 25–44. [https://doi.org/10.1002/1097-0355\(198921\)10:1<25::AID-IMHJ2280100104>3.0.CO;2-R](https://doi.org/10.1002/1097-0355(198921)10:1<25::AID-IMHJ2280100104>3.0.CO;2-R)
- Gopnik, M., & Crago, M. B. (1991). Familial aggregation of a developmental language disorder. *Cognition*, *39*(1), 1–50. [https://doi.org/10.1016/0010-0277\(91\)90058-C](https://doi.org/10.1016/0010-0277(91)90058-C)
- Grandpierre, V., Milloy, V., Sikora, L., Fitzpatrick, E., Thomas, R., & Potter, B. (2018). Barriers and facilitators to cultural competence in rehabilitation services: A scoping review. *BMC Health Services Research*, *18*(1), 23. <https://doi.org/10.1186/s12913-017-2811-1>
- Grimm, A., & Schulz, P. (2014). Specific Language Impairment and Early Second Language Acquisition: The Risk of Over- and Underdiagnosis. *Child Indicators Research*, *7*(4), 821–841. <https://doi.org/10.1007/s12187-013-9230-6>

- Grinstead, J., Lintz, P., Vega-Mendoza, M., de La Mora, J., Cantú-Sánchez, M., & Flores-Avalos, B. (2014). Evidence of optional infinitive verbs in the spontaneous speech of Spanish-speaking children with SLI. *Lingua*, *140*, 52–66. <https://doi.org/10.1016/j.lingua.2013.11.004>
- Grosjean, F., & Li, P. (2013). *The psycholinguistics of bilingualism* (First published.). *Ebrary online*. Wiley-Blackwell. <http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10658003>
- Guasti, M. T., White, M. J., Bianco, G., Arosio, F., Camilleri, B., & Hasson, N. (2021). Two clinical markers for DLD in monolingual Italian speakers: What can they tell us about second language learners with DLD? *Clinical Linguistics & Phonetics*, *35*(9), 829–846. <https://doi.org/10.1080/02699206.2020.1830303>
- Gutiérrez-Ciellen, V. F., & Peña, E. (2001). Dynamic Assessment of Diverse Children. *Language, Speech, and Hearing Services in Schools*, *32*(4), 212–224. [https://doi.org/10.1044/0161-1461\(2001/019\)](https://doi.org/10.1044/0161-1461(2001/019))
- Haman, E., Łuniewska, M., Hansen, P., Simonsen, H. G., Chiat, S., Bjekić, J., Blažienė, A., Chyl, K., Dabašinskienė, I., Engel de Abreu, P., Gagarina, N., Gavarró, A., Håkansson, G., Harel, E., Holm, E., Kapalková, S., Kunnari, S., Levorato, C., Lindgren, J., . . . Armon-Lotem, S. (2017). Noun and verb knowledge in monolingual preschool children across 17 languages: Data from Cross-linguistic Lexical Tasks (LITMUS-CLT). *Clinical Linguistics & Phonetics*, *31*(11-12), 818–843. <https://doi.org/10.1080/02699206.2017.1308553>
- Haridas, M., Vasudevan, N., Iyer, A., Menon, R., & Nedungadi, P. (2017). Analyzing the Responses of Primary School Children in Dyslexia Screening Tests. In *2017 5th IEEE International Conference on MOOCs, Innovation and Technology in Education (MITE)*. IEEE. <https://doi.org/10.1109/mite.2017.00022>
- Hasselaar, J., Letts, C., & McKean, C. (2019). Case marking in German-speaking children with specific language impairment and with phonological impairment. *Clinical Linguistics & Phonetics*, *33*(1-2), 117–134. <https://doi.org/10.1080/02699206.2018.1505955>
- Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V., Latvala, J.-M., & Richardson, U. (2020). Identification of Reading Difficulties by a Digital Game-Based Assessment Technology. *Journal of Educational Computing Research*, *58*(5), 1003–1028. <https://doi.org/10.1177/0735633120905309>
- Haywood, H. C., & Lidz, C. S. (2006). *Dynamic Assessment in Practice: Clinical and Educational Applications*. Cambridge University Press.

- Heckmann, F. (1981). *Die Bundesrepublik: Ein Einwanderungsland? : zur Soziologie der Gastarbeiterbevölkerung als Einwandererminorität*. Klett-Cotta. <https://fis.uni-bamberg.de/opus/38958>
- Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in Identifying Language Abilities of Late-Talking and Typically Developing Toddlers. *American Journal of Speech-Language Pathology*, *14*(1), 40–51. [https://doi.org/10.1044/1058-0360\(2005/006\)](https://doi.org/10.1044/1058-0360(2005/006))
- Ho, C. S.-H., & Bryant, P. (1997). Phonological skills are important in learning to read Chinese. *Developmental Psychology*, *33*(6), 946–951. <https://doi.org/10.1037/0012-1649.33.6.946>
- Hodge, M. A., Sutherland, R., Jeng, K., Bale, G., Batta, P., Cambridge, A., Detheridge, J., Drevensek, S., Edwards, L., Everett, M., Ganesalingam, K., Geier, P., Kass, C., Mathieson, S., McCabe, M., Micallef, K., Molomby, K., Ong, N., Pfeiffer, S., . . . Silove, N. (2019). Agreement between telehealth and face-to-face assessment of intellectual ability in children with specific learning disorder. *Journal of Telemedicine and Telecare*, *25*(7), 431–437. <https://doi.org/10.1177/1357633X18776095>
- Hoover, J. R., & Storkel, H. L. (2006). Using Nonword Repetition in Vocabulary Assessment. <https://kuscholarworks.ku.edu/handle/1808/19904>
- Horbach, J., Scharke, W., Cröll, J., Heim, S., & Günther, T. (2015). Kindergarteners' performance in a sound-symbol paradigm predicts early reading. *Journal of Experimental Child Psychology*, *139*, 256–264. <https://doi.org/10.1016/j.jecp.2015.06.007>
- Horbach, J., Weber, K., Opolony, F., Scharke, W., Radach, R., Heim, S., & Günther, T. (2018). Performance in Sound-Symbol Learning Predicts Reading Performance 3 Years Later. *Frontiers in Psychology*, *9*, 1716. <https://doi.org/10.3389/fpsyg.2018.01716>
- Hu, C.-F., & Catts, H. W. (1998). The Role of Phonological Processing in Early Reading Ability: What We Can Learn From Chinese. *Scientific Studies of Reading*, *2*(1), 55–79. https://doi.org/10.1207/s1532799xssr0201_3
- Hu, S., Vender, M., Fiorin, G., & Delfitto, D. (2018). Difficulties in Comprehending Affirmative and Negative Sentences: Evidence From Chinese Children With Reading Difficulties. *Journal of Learning Disabilities*, *51*(2), 181–193. <https://doi.org/10.1177/0022219417714775>
- Hulme, C., Nash, H. M., Gooch, D., Lervåg, A., & Snowling, M. J. (2015). The Foundations of Literacy Development in Children at Familial Risk of Dyslexia. *Psychological Science*, *26*(12), 1877–1886. <https://doi.org/10.1177/0956797615603702>

- Hume, L. E., Lonigan, C. J., & McQueen, J. D. (2015). Children's literacy interest and its relation to parents' literacy-promoting practices. *Journal of Research in Reading, 38*(2), 172–193. <https://doi.org/10.1111/j.1467-9817.2012.01548.x>
- Hunt, E., Nang, C., Meldrum, S., & Armstrong, E. (2022). Can Dynamic Assessment Identify Language Disorder in Multilingual Children? Clinical Applications From a Systematic Review. *Language, Speech, and Hearing Services in Schools, 1*–28. https://doi.org/10.1044/2021_LSHSS-21-00094
- Istituto Nazionale di Statistica. (2021). *Il Censimento permanente della popolazione in Lombardia*. <https://www.istat.it/it/archivio/253434>
- Istituto Nazionale di Statistica. (2022). *Il Censimento permanente della popolazione in Toscana - Anno 2020*. <https://www.istat.it/it/archivio/267740>
- Jackson, E., Leitao, S., & Claessen, M. (2016). The relationship between phonological short-term memory, receptive vocabulary, and fast mapping in children with specific language impairment. *International Journal of Language & Communication Disorders, 51*(1), 61–73. <https://doi.org/10.1111/1460-6984.12185>
- Jacobson, P. F., & Schwartz, R. G. (2005). English Past Tense Use in Bilingual Children With Language Impairment. *American Journal of Speech-Language Pathology, 14*(4), 313–323. [https://doi.org/10.1044/1058-0360\(2005/030\)](https://doi.org/10.1044/1058-0360(2005/030))
- Jasińska, K. K., Berens, M. S., Kovelman, I., & Petitto, L. A. (2017). Bilingualism yields language-specific plasticity in left hemisphere's circuitry for learning to read in young children. *Neuropsychologia, 98*, 34–45. <https://doi.org/10.1016/j.neuropsychologia.2016.11.018>
- Johnson, C. J., Beitchman, J. H., & Brownlie, E. B. (2010). Twenty-Year Follow-Up of Children With and Without Speech-Language Impairments: Family, Educational, Occupational, and Quality of Life Outcomes. *American Journal of Speech-Language Pathology, 19*(1), 51–65. [https://doi.org/10.1044/1058-0360\(2009/08-0083\)](https://doi.org/10.1044/1058-0360(2009/08-0083))
- Jordaan, H. (2008). Clinical intervention for bilingual children: An international survey. *Folia Phoniatica Et Logopaedica : Official Organ of the International Association of Logopedics and Phoniatics (IALP), 60*(2), 97–105. <https://doi.org/10.1159/000114652>
- Kan, P. F., & Kohnert, K. (2008). Fast mapping by bilingual preschool children. *Journal of Child Language, 35*(3), 495–514. <https://doi.org/10.1017/s0305000907008604>
- Kenner, B. B., Terry, N. P., Friehling, A. H., & Namy, L. L. (2017). Phonemic awareness development in 2.5- and 3.5-year-old children: An examination of emergent, receptive, knowledge and skills. *Reading and Writing, 30*(7), 1575–1594. <https://doi.org/10.1007/s11145-017-9738-0>

- Kiese-Himmel, C., & Risse, T. (2009). Normen für den Mottier-Test bei 4- bis 6-jährigen Kindern. *HNO*, 57(9), 943–948. <https://doi.org/10.1007/s00106-009-1958-4>
- Kim, K., & Kim, H. (2022). Sequential bilingual heritage children's L1 attrition in lexical retrieval: Age of acquisition versus language experience. *Bilingualism: Language and Cognition*, 1–11. <https://doi.org/10.1017/S1366728921001139>
- Koda, K. (2007). Reading and Language Learning: Crosslinguistic Constraints on Second Language Reading Development. *Language Learning*, 57, 1–44. <https://doi.org/10.1111/0023-8333.101997010-i1>
- Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders*, 43(6), 456–473. <https://doi.org/10.1016/j.jcomdis.2010.02.002>
- Kohnert, K., & Danahy, K. (2007). Young L2 learners' performance on a novel morpheme task. *Clinical Linguistics & Phonetics*, 21(7), 557–569. <https://doi.org/10.1080/02699200701374231>
- Kohnert, K., & Medina, A. (2009). Bilingual children and communication disorders: A 30-year research retrospective. *Seminars in Speech and Language*, 30(4), 219–233. <https://doi.org/10.1055/s-0029-1241721>
- Kroll, J. F., & Bialystok, E. (2013). Understanding the Consequences of Bilingualism for Language Processing and Cognition. *Journal of Cognitive Psychology*, 25(5), 497–514 <https://doi.org/10.1080/20445911.2013.799170>
- Kupisch, T. (2019). Italian as a heritage language in Germany : Acquisition outcomes and the role of cross-linguistic influence. In E. Bidese, J. Casalicchio, & M. C. Moroni (Eds.), *Studia Romanica et Linguistica Ser: v.57. La Linguistica Vista Dalle Alpi Linguistic Views from the Alps: Teoria, Lessicografia e Multilinguismo Language Theory, Lexicography and Multilingualism*. Peter Lang GmbH Internationaler Verlag der Wissenschaften.
- Laing, S. P., & Kamhi, A. (2003). Alternative Assessment of Language and Literacy in Culturally and Linguistically Diverse Populations. *Language, Speech, and Hearing Services in Schools*, 34(1), 44–55. [https://doi.org/10.1044/0161-1461\(2003/005\)](https://doi.org/10.1044/0161-1461(2003/005))
- Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., O'Donovan, M., Williams, J., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., Tóth, D., Honbolygó, F., Csépe, V., Bogliotti, C., Iannuzzi, S., Chaix, Y., Démonet, J.-F., . . . Schulte-Körne, G. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 54(6), 686–694. <https://doi.org/10.1111/jcpp.12029>

- Lautenschläger, T., Sachse, S., Buschmann, A., & Bockmann, A.-K. (2020). Folgeprobleme und begleitende Auffälligkeiten bei Sprachentwicklungsstörungen. In *Sprachentwicklung* (pp. 253–280). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-60498-4_12
- Law, J., Garrett, Z., & Nye, C. (2003). Speech and language therapy interventions for children with primary speech and language delay or disorder. *The Cochrane Database of Systematic Reviews*(3), CD004110. <https://doi.org/10.1002/14651858.CD004110>
- Lehti, V., Gyllenberg, D., Suominen, A., & Sourander, A. (2018). Finnish-born children of immigrants are more likely to be diagnosed with developmental disorders related to speech and language, academic skills and coordination. *Acta Paediatrica*, *107*(8), 1409–1417. <https://doi.org/10.1111/apa.14308>
- Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *Peabody picture vocabulary test - 4. Ausgabe: Deutsche Fassung von A. Lenhard, W. Lenhard, R. Segerer & S. Suggate.* (Deutsche Fassung). Pearson.
- Leśniewska, J., & Witalisz, E. (2014). Crosslinguistic Influence and Bilingual Children's Weaker Language. In M. Pawlak & L. Aronin (Eds.), *Second Language Learning and Teaching. Essential Topics in Applied Linguistics and Multilingualism* (pp. 225–233). Springer International Publishing. https://doi.org/10.1007/978-3-319-01414-2_13
- Letts, C. (2013). 3. What Are the Building Blocks for Language Acquisition? Underlying Principles of Assessment for Language Impairment in the Bilingual Context. In V. C. M. Gathercole (Ed.), *Solutions for the Assessment of Bilinguals* (pp. 36–56). Multilingual Matters. <https://doi.org/10.21832/9781783090150-005>
- Levorato, C., & Marini, A. (Eds.). (2019). *Il bilinguismo in età evolutiva. Aspetti cognitivi, linguistici, neuropsicologici, educativi.* Edizioni Centro Studi Erickson.
- Li, T., McBride-Chang, C., Wong, A., & Shu, H. (2012). Longitudinal predictors of spelling and reading comprehension in Chinese as an L1 and English as an L2 in Hong Kong Chinese children. *Journal of Educational Psychology*, *104*(2), 286–301. <https://doi.org/10.1037/a0026445>
- Lorusso, M. L., & Dolzadelli, C. (2016). *Uno strumento per la rilevazione dello stato di rischio per disturbi del linguaggio dai 20 ai 60 mesi.* IRCCS Fondazione Stella Maris. GIORNATE CLASTA VII, Calambrone.
- Luk, G., & Bialystok, E. (2008). Common and distinct cognitive bases for reading in English–Cantonese bilinguals. *Applied Psycholinguistics*, *29*(2), 269–289. <https://doi.org/10.1017/S0142716407080125>

- Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology, 25*(5), 605–621. <https://doi.org/10.1080/20445911.2013.795574>
- MacArthur-Bates Communicative Development Inventories. (2022). *Adaptations in Other Languages -- MacArthur-Bates Communicative Development Inventories*. <http://mb-cdi.stanford.edu/adaptations.html>
- Marini, A. (2019). Disturbo Primario del Linguaggio in bambini bilingui. In C. Levorato & A. Marini (Eds.), *Il bilinguismo in età evolutiva. Aspetti cognitivi, linguistici, neuropsicologici, educativi* (pp. 177–190). Edizioni Centro Studi Erickson.
- Marini, A., Marotta, L., Bulgheroni, S., & Fabbro, F. (2015). *Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni*. Giunti OS.
- Marini, A., Sperindè, P., Ruta, I., Savegnago, C., & Avanzini, F. (2019). Linguistic Skills in Bilingual Children With Developmental Language Disorders: A Pilot Study. *Frontiers in Psychology, 10*, 493. <https://doi.org/10.3389/fpsyg.2019.00493>
- Marinova-Todd, S. H., Colozzo, P., Mirenda, P., Stahl, H., Kay-Raining Bird, E., Parkington, K., Cain, K., Scherba de Valenzuela, J., Segers, E., MacLeod, A. A. N., & Genesee, F. (2016). Professional practices and opinions about services available to bilingual children with developmental disabilities: An international study. *Journal of Communication Disorders, 63*, 47–62. <https://doi.org/10.1016/j.jcomdis.2016.05.004>
- Mayor, J., & Mani, N. (2019). A short version of the MacArthur-Bates Communicative Development Inventories with high validity. *Behavior Research Methods, 51*(5), 2248–2255. <https://doi.org/10.3758/s13428-018-1146-0>
- Meisel, J. M. (2006). The Bilingual Child. In T. K. Bhatia & W. C. Ritchie (Eds.), *The Handbook of Bilingualism* (pp. 90–113). Blackwell Publishing Ltd. <https://doi.org/10.1002/9780470756997.ch4>
- Meisel, J. M. (2011). *First and Second Language Acquisition*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511862694>
- Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138*(2), 322–352. <https://doi.org/10.1037/a0026744>
- Meyer, M. S., Wood, F. B., Hart, L. A., & Felton, R. H. (1998). Selective predictive value of rapid automatized naming in poor readers. *Journal of Learning Disabilities, 31*(2), 106–117. <https://doi.org/10.1177/002221949803100201>
- Motsch, H.-J. (2011). *ESGRAF-MK: Mit 16 Abbildungen und 17 Tabellen : mit Diagnostik-Software auf CD-ROM*. Ernst Reinhardt Verlag.

- Mottier, G. (1951). Über Untersuchungen der Sprache lesegestörter Kinder. *Folia Phoniatrica Et Logopaedica*, 3(3), 170–177. <https://doi.org/10.1159/000262507>
- MultiMind ITN. (2022). *Policy Report: How to improve assesment and treatment of multilingual children with language and reading disorders*. https://www.multilingualmind.eu/_files/ugd/850b63_1143ca829d0f413ebb25a2dca26ae56b.pdf
- Newbury, D. F., & Monaco, A. P. (2010). Genetic advances in the study of speech and language disorders. *Neuron*, 68(2), 309–320. <https://doi.org/10.1016/j.neuron.2010.10.001>
- Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 57(11), 1247–1257. <https://doi.org/10.1111/jcpp.12573>
- Norton, E. S., Beach, S. D., & Gabrieli, J. D. E. (2015). Neurobiology of dyslexia. *Current Opinion in Neurobiology*, 30, 73–78. <https://doi.org/10.1016/j.conb.2014.09.007>
- Norton, E. S., & Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology*, 63, 427–452. <https://doi.org/10.1146/annurev-psych-120710-100431>
- Oller, D. K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics*, 28(2), 191–230. <https://doi.org/10.1017/S0142716407070117>
- Orellana, C. I., Wada, R., & Gillam, R. B. (2019). The Use of Dynamic Assessment for the Diagnosis of Language Disorders in Bilingual Children: A Meta-Analysis. *American Journal of Speech-Language Pathology*, 28(3), 1298–1317. https://doi.org/10.1044/2019_AJSLP-18-0202
- Paradis, J. (2005). Grammatical Morphology in Children Learning English as a Second Language. *Language, Speech, and Hearing Services in Schools*, 36(3), 172–187. [https://doi.org/10.1044/0161-1461\(2005/019\)](https://doi.org/10.1044/0161-1461(2005/019))
- Paradis, J. (2011). Internal and External Factors in Child Second Language Acquisition. *Linguistic Approaches to Bilingualism*, 1(3), 213–237. <https://doi.org/10.1075/lab.1.3.01par>
- Paradis, J. (2016). The Development of English as a Second Language With and Without Specific Language Impairment: Clinical Implications. *Journal of Speech, Language, and Hearing Research*, 59(1), 171–182. https://doi.org/10.1044/2015_JSLHR-L-15-0008

- Paradis, J., Crago, M., Genesee, F., & Rice, M. (2003). French-English Bilingual Children With SLI. *Journal of Speech, Language, and Hearing Research, 46*(1), 113–127. [https://doi.org/10.1044/1092-4388\(2003/009\)](https://doi.org/10.1044/1092-4388(2003/009))
- Paradis, J., Emmerzael, K., & Duncan, T. S. (2010). Assessment of English language learners: Using parent report on first language development. *Journal of Communication Disorders, 43*(6), 474–497. <https://doi.org/10.1016/j.jcomdis.2010.01.002>
- Parra, M., Hoff, E., & Core, C. (2011). Relations among language exposure, phonological memory, and language development in Spanish-English bilingually developing 2-year-olds. *Journal of Experimental Child Psychology, 108*(1), 113–125. <https://doi.org/10.1016/j.jecp.2010.07.011>
- Paulesu, E., Démonet, J. F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S. F., Cossu, G., Habib, M., Frith, C. D., & Frith, U. (2001). Dyslexia: Cultural diversity and biological unity. *Science (New York, N.Y.), 291*(5511), 2165–2167. <https://doi.org/10.1126/science.1057179>
- Petermann, F., & Daseking, M. (2019). *ZLT-II Zürcher Lesetest - II: Weiterentwicklung des Zürcher Lesetests (ZLT) von Maria Linder und Hans Grissemann: Manual* (4th ed.). Hogrefe.
- Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities, 48*(1), 3–21. <https://doi.org/10.1177/0022219413486930>
- Plug, M. B., van Wijngaarden, V., de Wilde, H., van Binsbergen, E., Stegeman, I., van den Boogaard, M.-J. H., & Smit, A. L. (2021). Clinical Characteristics and Genetic Etiology of Children With Developmental Language Disorder. *Frontiers in Pediatrics, 9*, 651995. <https://doi.org/10.3389/fped.2021.651995>
- Polinsky, M., & Scontras, G. (2020). A roadmap for heritage language research. *Bilingualism: Language and Cognition, 23*(1), 50–55. <https://doi.org/10.1017/S1366728919000555>
- Pua, E. P. K., Lee, M. L. C., & Rickard Liow, S. J. (2017). Screening Bilingual Preschoolers for Language Difficulties: Utility of Teacher and Parent Reports. *Journal of Speech, Language, and Hearing Research, 60*(4), 950–968. https://doi.org/10.1044/2016_JSLHR-L-16-0122
- Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study. *Child Development, 86*(1), 159–175. <https://doi.org/10.1111/cdev.12292>

- Ramus, F. (2004). Neurobiology of dyslexia: A reinterpretation of the data. *Trends in Neurosciences*, 27(12), 720–726. <https://doi.org/10.1016/j.tins.2004.10.004>
- Ramus, F., Marshall, C. R., Rosen, S., & van der Lely, H. K. J. (2013). Phonological deficits in specific language impairment and developmental dyslexia: Towards a multidimensional model. *Brain: A Journal of Neurology*, 136(Pt 2), 630–645. <https://doi.org/10.1093/brain/aws356>
- Rapin, I. (1996). Practitioner review: Developmental language disorders: A clinical update. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 37(6), 643–655. <https://doi.org/10.1111/j.1469-7610.1996.tb01456.x>
- Rauschenberger, M., Lins, C., Rousselle, N., Hein, A., & Fudickar, S. (2019). Designing a New Puzzle App to Target Dyslexia Screening in Pre-Readers. In A. Bujari, P. Manzoni, A. Forster, E. Mota, & O. Gaggi (Eds.), *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good* (pp. 155–159). ACM. <https://doi.org/10.1145/3342428.3342679>
- Rispens, J., & Been, P. (2007). Subject-verb agreement and phonological processing in developmental dyslexia and specific language impairment (SLI): A closer look. *International Journal of Language & Communication Disorders*, 42(3), 293–305. <https://doi.org/10.1080/13682820600988777>
- Riva, A., Musetti, A., Bomba, M., Milani, L., Montrasi, V., & Nacinovich, R. (2020). Language-Related Skills in Bilingual Children With Specific Learning Disorders. *Frontiers in Psychology*, 11, 564047. <https://doi.org/10.3389/fpsyg.2020.564047>
- Robertson, E. K., Joanisse, M. F., Desroches, A. S., & Terry, A. (2013). Past-tense morphology and phonological deficits in children with dyslexia and children with language impairment. *Journal of Learning Disabilities*, 46(3), 230–240. <https://doi.org/10.1177/0022219412449430>
- Roesch, A. D., & Chondrogianni, V. (2016). "Which mouse kissed the frog?" Effects of age of onset, length of exposure, and knowledge of case marking on the comprehension of wh-questions in German-speaking simultaneous and early sequential bilingual children. *Journal of Child Language*, 43(3), 635–661. <https://doi.org/10.1017/S0305000916000015>
- Romaine, S. (2017). Multilingualism. In M. Aronoff & J. Rees-Miller (Eds.), *The Handbook of Linguistics* (pp. 541–556). Wiley. <https://doi.org/10.1002/9781119072256.ch26>
- Roseberry, C. A., & Connell, P. J. (1991). The use of an invented language rule in the differentiation of normal and language-impaired Spanish-speaking children. *Journal of Speech and Hearing Research*, 34(3), 596–603. <https://doi.org/10.1044/jshr.3403.596>

- Roseberry-McKibbin, C., Brice, A., & O'Hanlon, L. (2005). Serving English Language Learners in Public School Settings. *Language, Speech, and Hearing Services in Schools*, 36(1), 48–61. [https://doi.org/10.1044/0161-1461\(2005/005\)](https://doi.org/10.1044/0161-1461(2005/005))
- Rothweiler, M., Schönenberger, M., & Sterner, F. (2017). Subject-verb agreement in German in bilingual children with and without SLI. *Zeitschrift für Sprachwissenschaft*(1), 79–106. <https://www.degruyter.com/document/doi/10.1515/zfs-2017-0005/html>
- Ruberg, T., Rothweiler, M., Veríssimo, J., & Clahsen, H. (2020). Childhood bilingualism and Specific Language Impairment: A study of the CP-domain in German SLI. *Bilingua-lism: Language and Cognition*, 23(3), 668–680. <https://doi.org/10.1017/S1366728919000580>
- Sansavini, A., Favilla, M. E., Guasti, M. T., Marini, A., Millepiedi, S., Di Martino, M. V., Vecchi, S., Battajon, N., Bertolo, L., Capirci, O., Carretti, B., Colatei, M. P., Frioni, C., Marotta, L., Massa, S., Michelazzo, L., Pecini, C., Piazzalunga, S., Pieretti, M., . . . Lorusso, M. L. (2021). Developmental Language Disorder: Early Predictors, Age for the Diagnosis, and Diagnostic Tools. A Scoping Review. *Brain Sciences*, 11(5) . <https://doi.org/10.3390/brainsci11050654>
- Sartori, G., Job, R., & Tressoldi, P. E. (2007). *DDE-2: Batteria per la valutazione della dislessia e della disortografia evolutiva-2*. Giunti O.S.
- Sauro, J. (2022, March 21). *Measuring Usability with the System Usability Scale (SUS) – MeasuringU*. <https://measuringu.com/sus/>
- Scharff Rethfeldt, W. (2019). Speech and Language Therapy Services for Multilingual Children with Migration Background: A Cross-Sectional Survey in Germany. *Folia Phoniatrica Et Logopaedica : Official Organ of the International Association of Logopedics and Phoniatrics (IALP)*, 71(2-3), 116–126. <https://doi.org/10.1159/000495565>
- Scherger, A.-L. (2015). Kasus als klinischer Marker im Deutschen. *Logos*, 23(3), 164–175.
- Scherger, A.-L. (2022). The role of age and timing in bilingual assessment: Non-word repetition, subject-verb agreement and case marking in L1 and eL2 children with and without SLI. *Clinical Linguistics & Phonetics*, 36(1), 54–74. <https://doi.org/10.1080/02699206.2021.1885497>
- Schönenberger, M., Rothweiler, M., & Sterner, F. (2012). Case marking in child L1 and early child L2 German. In K. Braunmüller & C. Gabriel (Eds.), *Multilingual Individuals and Multilingual Societies*. John Benjamins Publishing.
- Schulz, P., & Grimm, A. (2018). The Age Factor Revisited: Timing in Acquisition Interacts With Age of Onset in Bilingual Acquisition. *Frontiers in Psychology*, 9, 2732. <https://doi.org/10.3389/fpsyg.2018.02732>

- Schulz, P., & Tracy, R. (2011). *LiSe-DaZ Linguistische Sprachstandserhebung - Deutsch als Zweitsprache. Hogrefe-Vorschultests*. Hogrefe.
- Schumacher, J., Hoffmann, P., Schmäl, C., Schulte-Körne, G., & Nöthen, M. M. (2007). Genetics of dyslexia: The evolving landscape. *Journal of Medical Genetics*, *44*(5), 289–297. <https://doi.org/10.1136/jmg.2006.046516>
- Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using Nonword Repetition to Identify Developmental Language Disorder in Monolingual and Bilingual Children: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *64*(9), 3578–3593. https://doi.org/10.1044/2021_JSLHR-20-00552
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology (London, England : 1953)*, *94*(Pt 2), 143–174. <https://doi.org/10.1348/000712603321661859>
- Siegel, L. S. (2008). Morphological Awareness Skills of English Language Learners and Children With Dyslexia. *Topics in Language Disorders*, *28*(1), 15–27. <https://doi.org/10.1097/01.adt.0000311413.75804.60>
- Siegmüller, J., & Heide, B. von der. (2011). Störungen des Lesens und Schreibens bei Kindern. In J. Siegmüller & H. Bartels (Eds.), *Leitfaden Sprache Sprechen Stimme Schlucken*. Elsevier.
- Skeide, M. A., Kirsten, H., Kraft, I., Schaadt, G., Müller, B., Neef, N., Brauer, J., Wilcke, A., Emmrich, F., Boltze, J., & Friederici, A. D. (2015). Genetic dyslexia risk variant is related to neural connectivity patterns underlying phonological awareness in children. *NeuroImage*, *118*, 414–421. <https://doi.org/10.1016/j.neuroimage.2015.06.024>
- Snowling, M. J., Bishop, D. V. M., Stothard, S. E., Chipchase, B., & Kaplan, C. (2006). Psychosocial outcomes at 15 years of children with a preschool history of speech-language impairment. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *47*(8), 759–765. <https://doi.org/10.1111/j.1469-7610.2006.01631.x>
- Speakaboo. (2022). *Speakaboo*. <http://www.speakaboo.io/>
- Stankova, M., Rodríguez-Ortiz, I. R., Matic, A., Levickis, P., Lyons, R., Messarra, C., Kouba Hreich, E., Vulchanova, M., Vulchanov, V., Czaplewska, E., Ringblom, N., Hansson, K., Håkansson, G., Jalali-Moghadam, N., Dionissieva, K., Günhan Senol, N. E., & Law, J. (2021). Cultural and Linguistic Practice with Children with Developmental Language Disorder: Findings from an International Practitioner Survey. *Folia Phoniatrica Et Logopaedica : Official Organ of the International Association of Logopedics and Phoniatrics (IALP)*, *73*(6), 465–477. <https://doi.org/10.1159/000511903>

- Stow, C., & Dodd, B. (2003). Providing an equitable service to bilingual children in the UK: A review. *International Journal of Language & Communication Disorders*, 38(4), 351–377. <https://doi.org/10.1080/1368282031000156888>
- Surrain, S. (2021). ‘Spanish at home, English at school’: how perceptions of bilingualism shape family language policies among Spanish-speaking parents of preschoolers. *International Journal of Bilingual Education and Bilingualism*, 24(8), 1163–1177. <https://doi.org/10.1080/13670050.2018.1546666>
- Tan, L. H., Spinks, J. A., Feng, C.-M., Siok, W. T., Perfetti, C. A., Xiong, J., Fox, P. T., & Gao, J.-H. (2003). Neural systems of second language reading are shaped by native language. *Human Brain Mapping*, 18(3), 158–166. <https://doi.org/10.1002/hbm.10089>
- Thompson, P. A., Hulme, C., Nash, H. M., Gooch, D., Hayiou-Thomas, E., & Snowling, M. J. (2015). Developmental dyslexia: Predicting individual risk. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 56(9), 976–987. <https://doi.org/10.1111/jcpp.12412>
- Thordardottir, E., & Brandeker, M. (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders*, 46(1), 1–16. <https://doi.org/10.1016/j.jcomdis.2012.08.002>
- Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, 40(6), 1245–1260. <https://doi.org/10.1044/jslhr.4006.1245>
- Torgesen, J. K., Wagner, R., & Rashotte, C. A. (2012). *Test of word reading efficiency—second edition: (TOWRE-2)*. Pro-Ed.
- Torppa, M., Eklund, K., van Bergen, E., & Lyytinen, H. (2011). Parental literacy predicts children's literacy: A longitudinal family-risk study. *Dyslexia*, 17(4), 339–355. <https://doi.org/10.1002/dys.437>
- Treffers-Daller, J. (2019). What Defines Language Dominance in Bilinguals? *Annual Review of Linguistics*, 5(1), 375–393. <https://doi.org/10.1146/annurev-linguistics-011817-045554>
- Tsimpli, I. M. (2014). Epistemological issue with keynote article “Early, late or very late? Timing acquisition and bilingualism” by Ianthi Maria Tsimpli. *Linguistic Approaches to Bilingualism*, 4(3), 283–313. <https://doi.org/10.1075/lab.4.3.01tsi>

- Tuller, L. (2015). 11. Clinical Use of Parental Questionnaires in Multilingual Contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment* (pp. 301–330). Multilingual Matters. <https://doi.org/10.21832/9781783093137-013>
- Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., Dos Santos, C., Abed Ibrahim, L., & Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders*, 53(4), 888–904. <https://doi.org/10.1111/1460-6984.12397>
- Ulrich, T., Thater, S., & Mennicken, S. (2021). Kasusfähigkeiten mehrsprachiger Achtjähriger. Eine explorative Pilotstudie in Regelgrundschulen. *Logos : die Fachzeitschrift für Logopädie und Sprachtherapie*, 29(2), 84–95. <https://kups.ub.uni-koeln.de/54651/>
- United Nations. (2022). *THE 17 GOALS | Sustainable Development*. <https://sdgs.un.org/goals>
- Unsworth, S. (2016). Quantity and quality of language input in bilingual language development. In E. Nicoladis & S. Montanari (Eds.), *Bilingualism across the lifespan: Factors moderating language proficiency* (pp. 103–121). American Psychological Association. <https://doi.org/10.1037/14939-007>
- Vender, M., Delfitto, D., & Melloni, C. (2020). How do bilingual dyslexic and typically developing children perform in nonword repetition? Evidence from a study on Italian L2 children. *Bilingualism: Language and Cognition*, 23(4), 884–896. <https://doi.org/10.1017/S1366728919000828>
- Vender, M., Garraffa, M., Sorace, A., & Guasti, M. T. (2016). How early L2 children perform on Italian clinical markers of SLI: A study of clitic production and nonword repetition. *Clinical Linguistics & Phonetics*, 30(2), 150–169. <https://doi.org/10.3109/02699206.2015.1120346>
- Vender, M., Hu, S., Mantione, F., Delfitto, D., & Melloni, C. (2018). The Production of Clitic Pronouns: A Study on Bilingual and Monolingual Dyslexic Children. *Frontiers in Psychology*, 9, 2301. <https://doi.org/10.3389/fpsyg.2018.02301>
- Vicari, S., Marotta, L., & Luci, A. (2007). *TFL Test Fono-lessicale: Valutazione delle abilità lessicali in età prescolare*. Edizioni Erickson.
- Wallace, D. F., Norman, K. L., & Plaisant, C. (1988). *The American Voice and Robotics Guardian System: A Case Study in User Interface Usability Evaluation*. <http://www.cs.umd.edu/hcil/trs/88-10/88-10.pdf>
- Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, 66, 173–196. <https://doi.org/10.1146/annurev-psych-010814-015104>

- Wild, N., & Fleck, C. (2013). Neunormierung des Mottier-Tests für 5-bis 17-jährige Kinder mit Deutsch als Erst-oder als Zweitsprache. *Praxis Sprache*(3), 152–156. <http://www.schulpsychologie-sg.ch/pic-pdf-temp/neunormierung%20mottier-test.pdf>
- Williams, C. J., & McLeod, S. (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *International Journal of Speech-Language Pathology*, 14(3), 292–305. <https://doi.org/10.3109/17549507.2011.636071>
- Wolff, P. H., & Melngailis, I. (1994). Family patterns of developmental dyslexia: Clinical findings. *American Journal of Medical Genetics*, 54(2), 122–131. <https://doi.org/10.1002/ajmg.1320540207>
- World Health Organization. (2022a). *ICD-11 for Mortality and Morbidity Statistics: 6A01.2 Developmental Language Disorder*. <https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/33269655>
- World Health Organization. (2022b). *ICD-11 for Mortality and Morbidity Statistics: 6A03.0 Developmental learning disorder with impairment in reading*. <https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1008636089>
- Zwitsersloot, R., Harmsel, M. t., Schulting, J., Wiefferink, K., & Gerrits, E. (2022). To Game or Not to Game? Efficacy of Using Tablet Games in Vocabulary Intervention for Children with DLD. *Applied Sciences*, 12(3), 1643. <https://doi.org/10.3390/app12031643>