City Research Online

City St George's, University of London

Citation: Wallan, A. (2018). Evaluation of Arabic tests of sentence repetition and verbal short term memory for Saudi preschoolers. (Unpublished Doctoral thesis, City, Universtiy of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/19835/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Evaluation of Arabic Tests of Sentence Repetition and

Verbal Short Term Memory for Saudi Preschoolers

ASHWAG WALLAN

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

CITY UNIVERSITY LONDON

SCHOOL OF HEALTH SCIENCES

DIVISION OF LANGUAGE AND COMMUNICATION SCIENCE

April, 2018

Table of Contents

List of Tables vii
List of Figures and Illustrations xi
Dedication xiii
Abstract xiv
Declaration xv
List of Symbols, Abbreviations, and Nomenclature xvi
Epigraph xviii

CHAPTER ONE: SENTENCE REPETITION AS AN ASSESSMENT TOOL 1
1.1 Introduction 1
1.1.1 Outline of the thesis 2
1.2 Historical Background on Sentence Repetition 3
1.3 Sentence Repetition as an Assessment Tool 9
1.3.1 Examples of Sentence Repetition assessments and why they are language-specific ... 9
1.3.2 Targets 9
1.3.3 Scoring 14
1.3.4 Psychometric properties (valid measure of language) 15
1.3.5 Age and gender sensitivity (valid measure of development) 19
1.4 Conclusion 21

CHAPTER TWO: SENTENCE REPETITION AS A POTENTIAL CLINICAL MARKER OF SPECIFIC LANGUAGE IMPAIRMENT 23
2.1 Introduction 23
2.1.1 Key terminology 23
2.1.2 1. What constitutes a language impairment? 24
2.1.3 Nonverbal IQ 24
2.1.4 Comorbidity 24
2.1.5 Sentence Repetition as a marker of Specific Language Impairment in English 28
2.1.6 Conclusion 42
2.2 Diagnostic Accuracy Studies in other Languages 44
2.2.1 Heterogeneity 52
2.2.2 Scoring 52
2.3 Diagnostic Accuracy Studies and Underlying Processes 53
2.3.1 Correlational evidence and implications 54
2.3.1.1 Correlations between marker tasks 55
2.3.1.2 Sentence Repetition: Relations to Grammatical Morpheme probes versus Verbal Short Term Memory 55
2.3.1.3 Relations between Grammatical Morpheme probes and Verbal Short Term Memory measures 56
2.3.1.4 Relations between Verbal Short Term Memory measures: Same or different? 56
2.3.1.5 Correlations between broad language measures and marker tasks 56
2.3.2 Multiple regression and implications 62
2.3.3 Principal component analysis 63
2.3.4 Evidence from differential diagnosis studies and implications 63
2.3.5 Conclusion 63
2.4 A Closer Look at the Anomalous Findings of the Cantonese Study: Implications for Repetition Tasks 64
2.4.1 Differences in target content: The influence of language typology, syllable structure, and length of nonwords 64
2.4.2 Implications of nonword repetition findings: Test design and underlying processes. 67

2.4.2.1 Differences in target content: The influence of language typology, stimuli, and scoring method on Sentence Repetition 68
2.4.3 Implications for Sentence Repetition tests: Targets, scoring, and underlying processes 70

CHAPTER THREE: LINGUISTIC MANIPULATION OF IMMEDIATE REPETITION TARGETS: SERIAL RECALL AND SENTENCE REPETITION 72
3.1 Linguistic Manipulation of Serial Recall Tests 72
3.1.1 Lexicality effect 78
3.1.2 Frequency effect 79
3.1.3 Concreteness and imageablity effects 80
3.2 Syntactic Manipulation of Well-formed Sentences 82
3.2.1 Effects of syntax when length was controlled 83
3.2.2 Effects of syntax when length was manipulated 90
3.3 Linguistic Manipulation of Syntactically Simple Sentences 96
3.3.1 Manipulation of syntax, semantics, prosody and lexicality 98
3.4 Literature Review Summary 104

CHAPTER FOUR: METHODS 106
4.1 Participants 107
4.1.1 Demographics 109
4.2 Development and Piloting of Tests 110
4.3 Assessments 112
4.3.1 Verbal Short-Term Memory test 112
4.3.1.1 Working Memory Test Battery for Children 112
4.3.1.2 Arabic Verbal Short Term Memory test 113
4.3.1.3 Procedure of the Verbal Short Term Memory test 116
4.3.1.4 Scoring of the Verbal Short Term Memory test 117
4.3.2 Sentence Repetition test 119
4.3.2.1 Procedure of the Sentence Repetition test 121
4.3.2.2 Scoring of the Sentence Repetition test 122
4.3.3 Anomalous Sentence Repetition test 124
4.3.3.1 Procedures for the Anomalous Sentence Repetition test 127
4.3.3.2 Scoring for the Anomalous Sentence Repetition test 127
4.3.4 Nonverbal IQ test 128
4.4 General Procedure 128
4.4.1 Reliability 131

CHAPTER FIVE: RESULTS 132
5.1 Reliability 132
5.1.1 Inter-rater reliability 133
5.1.2 Test-retest reliability 134
5.1.3 Internal consistency 136
5.2 Validity 137
5.3 Typically Developing Participants 137
5.3.1 Background assessment: Nonverbal IQ scores 140
5.3.2 Gender and school type 141
5.3.3 Verbal Short Term Memory test 144
5.3.3.1 A comparison of Verbal Short Term Memory subtest span scores 144
5.3.3.2 A comparison of Total Span scores 149
5.3.4 Sentence Repetition test 151
5.3.4.1 A comparison of Lexical and Grammatical Morpheme scores 151
5.3.4.2 A comparison of Total Sentence Accuracy scores for age groups 155

5.3.5 Anomalous Sentence Repetition test 158
5.3.5.1 A comparison of morpheme and sentence type (descriptive statistics) 158
5.3.5.2 A comparison of morpheme and sentence type (inferential statistics) 162
5.4 Participants with Language Concerns 165
5.4.1 Verbal Short Term Memory test 166
5.4.1.1 A comparison of Verbal Short Term Memory subtests according to language status 166
5.4.1.2 A comparison of Total Span score according to language status 168
5.4.2 Sentence Repetition test 169
5.4.2.1 A comparison of Lexical and Grammatical Morpheme scores according to language status 169
5.4.2.2 A comparison of Total Sentence Accuracy scores according to language status 172
5.4.3 Anomalous Sentence Repetition test 173
5.4.3.1 A comparison of morpheme scores according to sentence type and language status 173
5.4.3.2 A comparison of morpheme, sentence type and language status (inferential statistics) 175
5.5 Z-scores 179
5.6 Error Analysis 188
5.6.1 Verbal Short Term Memory Errors 188

CHAPTER SIX: DISCUSSION 192
6.1 Main Findings 193
6.1.1 Reliability and replication 193
6.1.2 Validity 194
6.1.3 Levels and patterns of performance in the Typically Developing sample 194
6.1.4 Levels and patterns of performance in Language Concerns sample 200
6.2 Clinical Implications 203
6.3 Theoretical Implications: Immediate Repetition - Memory, Language or Both? 206
6.3.1 Implications of linguistic effects on repetition performance in Typically Developing children 206
6.3.2 Implications of linguistic effects across age groups 208
6.3.3 Implications of linguistic effects across language ability groups 210
6.4 Limitations and Future Directions 211
6.5 Conclusion 212

REFERENCES 214

APPENDIX A: INVITATION LETTERS, CONSENT, QUESTIONNAIRE 227
A.1. Invitation Letter to Heads of Nursery 227
A.2. Invitation Letter To Parents 228
A.3. Consent Form 229
A.4. Parental Questionnaire 229

APPENDIX B: PARENT DEMOGRAPHICS 230
B.1. Saudi Population (15 Years and Over) by Marital Status, Gender, and Educational Status (Riyadh Administrative region); adapted from General Authority for Statistics of Kingdom of Saudi Arabia (2010) 230

APPENDIX C: DEVELOPMENT STAGE AND PILOT RESULTS 231
C.1. Development Stage Results 231
C.2. Pilot Results 231

C.2.1. Correlation reseults between VSTM subtests: Pilot Stage 232
C.2.2. Correlation results betweenVSTM Total Span score and Total Sentence Accuracy Score (CELF): Pilot Stage.....233

APPENDIX D: DEVELOPMENT STAGE VERBAL SHORT TERM MEMORY WORD RECALL SUBTEST 234
D.1. Word List Recall-Noun 234
D.2. Word List Recall-Mixed 235

APPENDIX E: VERBAL SHORT TERM MEMORY SUBTESTS.....237
E.1. Digit Recall.....237
E.2. Word List Recall 237
E.3. Nonword List Recall 238

APPENDIX F: SENTENCE REPETITION TESTS 240

APPENDIX G: ANONMALOUS SENTENCE REPETITION TEST 242
G.1. Typical sentences that were used to derive the Anomalous sentences and their score sheet. 242
G.1.1. Summary Score Sheet for typical sentences 242
G.2. Semantically Anomalous sentences and their score sheets. 243
G.2.1. Summary Score Sheet for Semantically Anomalous sentences 243
G.3. Syntactically Anomalous sentences and their score sheets. 244
G.3.1. Summary Score Sheet for Syntactically Anomalous sentence.....245

APPENDIX H: TESTS OF NORMALITY AND SUPPLEMENTARY NONPARAMETRIC ANALYSIS 246
H.1. Verbal Short Term Memory Test.....246
H.1.1. Nonparametric Analysis (Subtest Span Score) 246
H.1.2. Nonparametric Analysis (Total Span Score).....246
H.2. Sentence Repetition Test.....247
H.2.1. Nonparametric Analysis (Morpheme Score) 247
H.2.2. Nonparametric Analysis (Total Sentence Accuracy Score).....247
H.3. Anomalous Sentence Repetition Test 248
H.3.1. Nonparametric Analysis 248

List of Tables

Table 1.1 *Summary of Sentence Repetition Assessments* 11

Table 1.2: *Psychometric Characteristics of the Sentence Repetition Tests* 16

Table 2.1 *Contingency Table for obtaining Diagnostic Accuracy Measures* 25

Table 2.2 *Summary of Conti-Ramsden et al. (2001)* 31

Table 2.3 *Summary of Poll et al. (2010)* 32

Table 2.4 *Summary of Baseline from Everitt et al. (2013)* 33

Table 2.5 *Summary of Follow-up from Everitt et al. (2013)* 34

Table 2.6 *Summary of Archibald and Joanisse (2009)* 37

Table 2.7 *Summary of Results from Archibald and Joanisse (2009)* 38

Table 2.8 *Summary of Botting and Conti-Ramsden (2003)* 39

Table 2.9 *Summary of Redmond et al. (2011)* 41

Table 2.10 *Summary of Sentence Repetition Tests used in Diagnostic Accuracy Studies in Languages Other than English* 44

Table 2.11 *Summary o*f *Stokes et al. (2006)* 47

Table 2.12 *Summary of Leclercq et al. (2014)* 48

Table 2.13 *Summary of Thordardottir et al. (2011)* 50

Table 2.14 *Correlations between Markers in Diagnostic Accuracy Studies of English Specific Language Impairment* 59

Table 2.15 *Spearman's Correlations between PLS-3 Scores and Markers Tasks in Everitt (2009)* 61

Table 2.16 *Strength and Weakness of the Four Scoring Methods of Stokes et al. (2006)* 70

Table 3.1 *Overview of Studies that Manipulated Linguistic Domains in Serial Recall Rests* 75

Table 3.2 *Summary of Diessel and Tomasello (2005)* 85

Table 3.3 *Summary of Frizelle and Fletcher (2014)* 87

Table 3.4 *Summary of Riches et al. (2010)* 89

Table 3.5 *Summary of Willis and Gathercole (2001) Experiment 1* 91

Table 3.6 *Summary of Willis and Gathercole (2001) Experiment 2* 92

Table 3.7 *Summary of Wilsenach (2006)* 94

Table 3.8 *Summary of Moll et al. (2015)* 96

Table 3.9 *Stimuli Examples from G. A. Miller and Isard (1963, p. 220)* 97

Table 3.10 *Summary of Studies that Linguistically Manipulated Syntactically Simple Sentence* .. 100

Table 4.1 *Background Information on Typically Developing Participants* 108

Table 4.2 *Background Information on Language Concerns Participants* 108

Table 4.3 *Education Levels of the Parent Samples and Saudi Population* 111

Table 4.4 *Summary of Participants in Development and Pilot Stages* 111

Table 4.5 *Transcription of Arabic Digits and Number of Syllables* 114

Table 4.6 *Distribution of Morphemes in the Sentence Repetition Test* 120

Table 4.7 *List of Sentences in the Sentence Repetition Test with English Gloss, in the Order Presented* 120

Table 4.8 *Distribution of Morphemes in Semantically Anomalous Sentences, Syntactically Anomalous Sentences, and Typical Sentences* 124

Table 4.9 *List of Semantically Anomalous Sentences in the Order Presented* 126

Table 5.1 *Interclass Correlation Coefficient between Rater 1 and Rater 2 Accuracy Scores* 133

Table 5.2 *Interclass Correlation Coefficients between Rater 1 and Rater 2 Error Classification for Verbal Short Term Memory Subtests* 134

Table 5.3 *Interclass Correlation Coefficients between Rater 1 and Rater 2 Error Classification for Sentence Repetition and Anomalous Sentence Repetition Tests* 134

Table 5.4 *Interclass Correlation Coefficients between Time 1 and Time 2 Accuracy Scores* 135

Table 5.5 *Interclass Correlation Coefficients for Time 1 and Time 2 on Error Classification for Verbal Short Term Memory Subtests.* 135

Table 5.6 *Interclass Correlation Coefficients for Time 1 and Time 2 on Error Classification for Sentence Repetition and Anomalous Sentence Repetition Tests.* 136

Table 5.7 *Internal Consistency for Sentence Repetition and Anomalous Sentence Repetition Tests* 136

Table 5.8 *Partial Correlation Controlling for Age in Months between the Verbal Short Term Memory and Sentence Repetition tests for Typical Sample (n = 140)* 137

Table 5.9 *Descriptive Statistics for Block Design and Object Assembly of Wechsler Preschool and Primary Scale of Intelligence-Third Edition for Typically Developing Participants, According to Age* 141

Table 5.10 *Distribution of Typically Developing Participants According to Age (6-month Age Bands) and Gender* 142

Table 5.11 *Distribution of Typically Developing Participants According to Age (6-month Age Bands) and School Type* 142

Table 5.12 *Descriptive Statistics for Typically Developing Participants on All Measures According to Gender* 143

Table 5.13 *Descriptive Statistics for Typically Developing Participants on All Measures According to School Type* 144

Table 5.14 *Descriptive Statistics for the Verbal Short Term Memory Subtests for Typically Developing Participants According to Age (6-month Age Bands)* 145

Table 5.15 *Comparison of Mean Difference in Span Scores between Age Groups* 148

Table 5.16 *Comparison of Mean Difference in Span Scores between Verbal Short Term Memory Subtests* 148

Table 5.17 *Descriptive Statistics for Verbal Short Term Memory Total Scores of Typically Developing Participants According to Age (in 1-year age Bands)* 150

Table 5.18 *Comparison of Mean Difference Total Span Scores between Age Groups* 151

Table 5.19 *Descriptive Statistics for the Sentence Repetition Test Morpheme Scores for Typically Developing Participants According to Age (in Percentages)* 152

Table 5.20 *Comparison of Mean Difference in Morpheme Scores between Age Groups* 154

Table 5.21 *Descriptive Statistics for the Sentence Repetition Test Total Sentence Accuracy Scores for Typically Developing Participants According to Age* 156

Table 5.23 *Descriptive Statistics for Lexical Morpheme Scores of Typically Developing Participants According to Age and Sentence Type (in Percentages)* 159

Table 5.24 *Descriptive Statistics for Grammatical Morpheme Scores of Typically Developing Participants According to Age and Sentence Type (in Percentages)* 161

Table 5.25 *Comparison of Mean Difference in Scores between Age Groups* 163

Table 5.26 *Comparison of Mean Difference in Scores between Sentence Types* 164

Table 5.26 *Descriptive Statistics for the Verbal Short Term Memory Subtests According to Language Status, Controls, and Language Concerns* 166

Table 5.27 *Comparison of Mean Difference in Span Scores between Verbal Short Term Memory Subtests* 168

Table 5.28 *Descriptive Statistics for Total Span Scores According to Language Status, Controls, and Language Concerns* 168

Table 5.29 *Descriptive Statistics for the Sentence Repetition Test Morpheme Scores (in Percentages) According to Language Status, Controls, and Language Concerns* 170

Table 5.30 *Descriptive Statistics for Total Sentence Accuracy Scores According to Language Status, Controls, and Language Concerns* 172

Table 5.31 *Descriptive Statistics for the Anomalous Sentence Repetition Test Morpheme Scores According to Language Status and Sentence Type (in Percentages), Controls, and Language Concerns* 174

Table 5.33 *Percentage of Children Scoring Above, Within, and Below Normal Range on the Verbal Short Term Memory and Sentence Repetition Tests* 182

Table 5.34 *Percentage of Children Scoring Above, Within, and Below Normal Range on the Anomalous Sentence Repetition Test* 182

Table 5.41 *Proportion of Errors in the Three Subtests of the Verbal Short Term Memory Test for Typically Developing Participants* 189

Table 5.42 *Proportion of Errors in the Three Subtests of the Verbal Short Memory Test for the Language Concerns Group* 189

Table 6.1 *Descriptive Statistics for the Sentence Repetition Test Scores in current study and AlKadhi (2012)* 193

Table 6.2 *Partial Correlation (Controlling for Age in Months) between Sentence Repetition Scores and the Arabic Picture Vocabulary Test (AlKadhi, 2012)* 194

Table 6.3 *Developmental Differences in Verbal Span Tasks* 196

List of Figures and Illustrations

Figure 2.1 Example of ROC curve.....26

Figure 4.1. Order of presentation of test battery. 130

Figure 5.1. Boxplot showing performance on a Sentence Repetition test according to gender of participants (Wallan, 2006).....138

Figure 5.2. Error bar chart of Lexical, Preposition, and Overall Gender Agreement scores according to age (Wallan, 2006). 139

Figure 5.3. Boxplots and error bars showing Typically Developing group's performance on Verbal Short Term Memory subtests according to age. 146

Figure 5.4. Age x subtest interaction graph. 149

Figure 5.5. Boxplots and error bars of Total Span scores of Typically Developing participants according to age groups.....150

Figure 5.6. Boxplots showing Typically Developing group's performance on the Sentence Repetition test according to age 152

Figure 5.7. Error bars for Sentence Repetition scores of Typically Developing participants according to age groups (in percentage) 153

Figure 5.8. The interaction between age and morpheme type. 155

Figure 5.9. Boxplots showing Typically Developing group's Lexical Morpheme scores according to age and sentence type. 159

Figure 5.10. Error bars showing Typically Developing group's Lexical Morpheme scores according to age and sentence type. 160

Figure 5.11. Boxplots showing Typically Developing group's Grammatical Morpheme scores according to age and sentence type. 161

Figure 5.12. Error bars showing Typically Developing group's Grammatical Morpheme scores according to age and sentence type. 162

Figure 5.13. The interaction between Morpheme and Sentence type.....164

Figure 5.14. The interaction between Morpheme and Sentence type and age.....165

Figure 5.15. Verbal Short Term Memory subtest span scores according to language status of participants.....167

Figure 5.16. Total Span score according to language status of participants.....169

Figure 5.17. Morpheme scores according to language status of participants. 170

Figure 5.18. The interaction between morpheme type and language status of participants. 172

Figure 5.19. Sentence Accuracy scores according to language status of participants.....173

Figure 5.20. Morpheme scores according to language status and sentence type. 175

Figure 5.21. The interaction between morpheme and sentence type..... 177

Figure 5.22. The interaction between morpheme type and sentence type according to language status. 178

Figure 5.23 Proportion of errors in the three subtests of the Verbal Short Term Memory Test in each language status group. 190

Figure 6.1 Digit comprehension levels of Najdi Arabic-Speaking children aged 24-60 months (*Note.* CP, Cardinal principal); adapted from "Grammatical morphology as a source of early number word meanings" by Almoammer et al. (2013) *Proceedings of the National Academy of Sciences, 110*(46), 18448-18453. 197

Dedication

To Mom and Dad for their unconditional love and support

To Shula and Penny for their guidance, generosity and encouragement

To the wonderful children who inspired me to undertake this project and to their parents for the trust they placed in me

Abstract

Background: Sentence Repetition (SR) is considered to be a good indicator of children's grammatical knowledge. Cross-linguistic evidence suggests that performance on SR improves with age, differentiates children with language difficulties, and shows relationships with other language assessments. However, there is debate about the underlying skills involved in SR with few studies directly investigating the impact of linguistic manipulation on SR performance. In the absence of standardized language assessments and lack of normative data, and building on evidence from typologically diverse languages, SR provides a potentially useful assessment tool in Arabic.
Aims: (1) To examine the clinical utility of a novel SR test and an adapted Verbal Short Term Memory (VSTM) test by investigating the psychometric properties of the tests and their sensitivity to age and language ability. (2) To evaluate the contribution of established linguistic knowledge to immediate repetition by comparing the patterns of performance across different linguistic factors 3) To determine whether patterns of performance are similar or dissimilar across different age groups of Typically Developing children and different language ability groups.
Methods: Three immediate repetition tests were developed or adapted: (1) a novel SR test targeting morphosyntactic structures of Arabic; (2) an adapted VSTM test based on the structure of the Working Memory Test Battery for Children (WMTB-C; Pickering & Gathercole, 2001) with three subtests of Digit Recall, Word List Recall, and Nonword List Recall; and (3) an Anomalous Sentence Repetition (ASR) test including sets of Semantically Anomalous and Syntactically Anomalous sentences created from and matched to a subset of sentences in the SR test in target Lexical and Grammatical Morphemes as well as length. The SR and ASR tests were scored for the number of Lexical and Grammatical Morphemes repeated correctly. VSTM tests were scored based on the highest number of items repeated in correct order. The SR and VSTM tests were administered to Typically Developing Arabic-speaking children aged 2;6 to 5;11 (n = 140) and a Language Concerns group in the same age range (n = 16), matched on age and nonverbal IQ. The ASR test was only administered to participants older than 4 years.
Results: The SR and VSTM tests were reliable, valid, and sensitive to age and language ability of participants. In the Typical sample a) Lexical Morphemes were easier to repeat than Grammatical Morphemes, (b) Digit span was higher than Word span and Word span was higher than Nonword span, and (c) Typical sentences were easier to repeat than Semantically Anomalous sentences followed by Syntactically Anomalous sentences. The gap between Digit and Word span, Grammatical and Lexical Morphemes in the SR test and Lexical Morphemes in Typical and Semantically Anomalous sentences showed a change with age. While performance was significantly reduced in the Language Concerns group, the profile of performance was largely similar. Like the younger children in the Typical sample, they showed a greater vulnerability in Grammatical Morphemes. Only four of 16 children in the clinical sample showed mismatches between their performance on the SR and VSTM tests.
Conclusions: The study's results are consistent with cross-linguistic evidence demonstrating that SR and VSTM tests are sensitive to developmental change and language difficulties and are informative about children's language processing abilities. These findings lay the foundations for creating standardized assessments for Arabic-speaking preschool children.

Declaration

The following statement is included in accordance with the Regulations governing the Research Studies Handbook, 2008/9 (regulation 5(e)):

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Ashwag Wallan

List of Symbols, Abbreviations, and Nomenclature

Symbol	Definition
ADHD	Attention Deficit Hyperactivity Disorder
ALI	Autism and Language Impairment
AM	Age-matched
ANOVA	analysis of variance
ASD	Autism Spectrum Disorders
ASHA	American Speech-Language-Hearing Association
AWMA	Automated Working Memory Assessment
BAS	British Ability Scales
CARS	Childhood Autism Rating Scales
CBCL DSM-ADHD	Child Behavior Checklist Diagnostic Statistics Manual ADHD subscale
CCC	Children's Communication Checklist
CELF-3	Clinical Evaluation of Language Fundamentals
CELF-4	Clinical Evaluation of Language Fundamentals 4th edition
CELF-P	Clinical Evaluation of Language Fundamentals-Preschool
CELF-R	Clinical Evaluation of Language Fundamentals-Revised
CELFST-4	Clinical Evaluation of Language Fundamentals Screening Test 4th edition
CMMS	Columbia Mental Maturity Scales
CNRep	Children's Test of Nonword repetition
CRDLS	Cantonese Reynell Developmental Language Scales
CRDLS–R	Cantonese Reynell Developmental Language Scales Receptive subtest
CRVT	Cantonese Receptive Vocabulary Test
CVC	Consonant Vowel Consonant
DLD	Developmental Language Disorder
DSM-5	Diagnostic and Statistical Manual of Mental Disorders 5th edition
DSS	Developmental Sentence Scoring
ELI	Elicited Language Imitation
ERB	Early Repetition Battery
EVT	Expressive Vocabulary Test
GAPS	Grammar and Phonology Screening
IQ	Intelligence Quotient
L2MA2	Battery for oral language, writing, memory and attention [Batterie langage oral, langage écrit, mémoire, attention]
LIPS-R	Leiter International Performance Scale-Revised
LR	Likelihood Ratio
MLU	Mean Length of Utterance
NNAT-I	Naglieri Nonverbal Achievement Test-Individual
NRT	Nonword Repetition Test
NWR	Nonword Recall
OOAQ	Quebec Association of Speech-Language Pathologists and Audiologists
OSV	Object Subject Verb
PLI	Pragmatic Language Impairment
PLS-3	Preschool Language Scale
PLS-3^{UK}	Preschool Language Scale 3rd edition
PLS-4^{UK}	Preschool Language Scale 4th edition
PPVT-R	Peabody Picture Vocabulary Test-Revised

PSRep	PreSchool Repetition
RAPT	Renfrew Action Picture Test
ROC	Receiver operating characteristics
SCQ	Social Communication Questioner
SELD	Specific Expressive Language Delay
SIT	Sentence Imitation Test
SLI	Specific Language Impairment
SLQ	Spoken Language Quotient
STM	Short Term Memory
TACL-R	Test of Auditory Comprehension of Language Revised
TDAM	Typically Developing Age Matched
TDY	Typically Developing Younger
TEGI	Test of Early Grammatical Impairment
TNL	Test of Narrative Language
TOAL-3	Test of Adolescent and Adult Language
TOLD	Test of Language Development
TROG	Test for the Reception of Grammar
TSA	Total Sentence Accuracy
VSO	Verb Subject Object
VSTM	Verbal Short Term Memory
WAIS-III	Wechsler Adult Intelligence Scale-III
WISC-III	Wechsler Intelligence Scale for Children 3rd edition
WISC-IV	Wechsler Intelligence Scale for children 4th edition
WMI	Working Memory Impairment
WMTB-C	Working Memory Test Battery for Children
YTD	Younger Typically Developing

Epigraph

كُلَّما أدَّبني الدَّهرُ

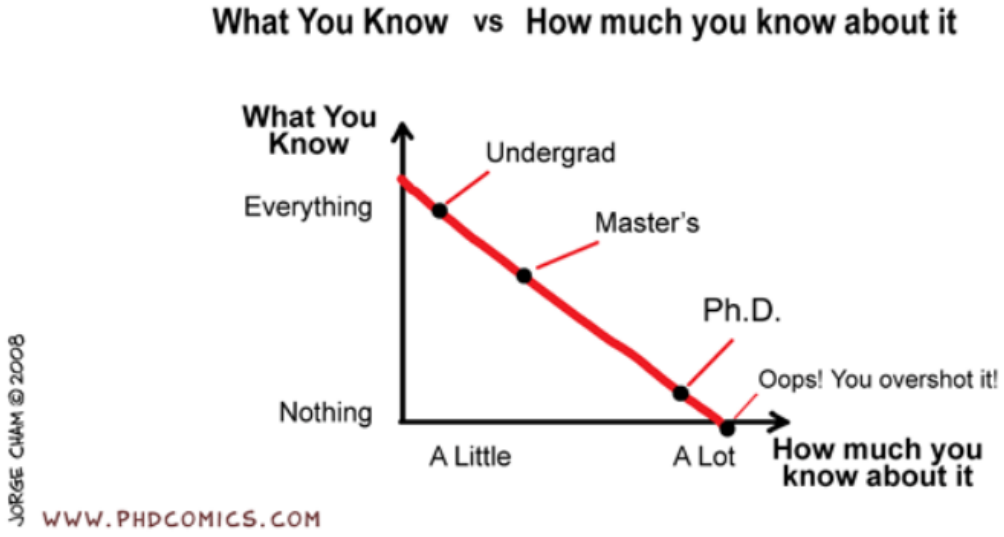
أراني نقصَ عقلِي

وإذا ما ازددتُ علماً

زادني علماً بجهلِي

الإمام الشَّافعي

"The more I learn, the more I realize how much I don't know." – Alberta Einstein



Chapter One: Sentence Repetition as an Assessment Tool

1.1 Introduction

In the Kingdom of Saudi Arabia speech-language therapy is a relatively new profession (AlAbdulkarim, 2015). In light of the newness of the profession and that Arabic is the country's official language, speech-language therapists face many challenges. These include shortages in the number of qualified professionals, large caseloads and waiting lists, limited service delivery models, lack of (or in some cases non-existent) research in language development and atypical language, and limited commercial tools for use in service delivery. In order to identify children with language impairment clinicians commonly resort to informal assessment, clinical judgment, and translated or adapted versions of standardized tests in other languages. In the absence of standardized language assessments in Arabic and a lack of normative data, a key aim of the study is to draw on evidence from English and other languages to develop an efficient and informative Arabic language assessment.

Researchers in the last 40 years generally agree that Sentence Repetition provides a window into the child's expressive language ability, more specifically morpho-syntax and lexical knowledge (Carrow, 1974; Chiat et al., 2013; Dockrell & Marshall, 2015; Newcomer & Hammill, 1997; Ratner, 2000). It is commonly used as part of expressive language assessments such as the Test of Language Development (TOLD; Newcomer & Hammill, 1997) and the Clinical Evaluation of Language Fundamentals (CELF-3; Semel, Wiig, & Secord, 1994), with good levels of reliability and validity. Sentence Repetition scores improve with age, differentiate between children with and without language impairment, and significantly correlate with broad assessments of expressive language (Chiat & Roy, 2008; Everitt, Hannaford, & Conti-Ramsden, 2013; Seeff-Gabriel, Chiat, & Roy, 2008).

Impaired performance on Sentence Repetition in individuals with Specific Language Impairment (SLI) is a robust finding across a number of typologically diverse languages (e.g., English: Conti-Ramsden, Botting, & Faragher, 2001; French: Leclercq, Quemart, Magis, & Maillart, 2014; Gulf Arabic: Shaalan, 2010; Cantonese: Stokes, Wong, Fletcher, & Leonard, 2006; Turkish: Topbaş & Güven, 2009); age groups (e.g., middle school: Hesketh & Conti-Ramsden, 2013; adults: Poll, Betz, & Miller, 2010; preschool; Seeff-Gabriel et al., 2008), and language modalities (British Sign Language: Marshall et al., 2015). As such, Sentence Repetition has been put forward as a potential clinical marker for SLI. Recently there has been increased interest in its use an assessment tool with children who are bilingual (Chiat et al., 2013; Komeili & Marshall, 2013; Thordardottir & Brandeker, 2013) and from different socio-economic backgrounds (Roy & Chiat, 2013).

While the strong relations found between Sentence Repetition and language assessments along with the poor performance of individuals with SLI can be argued as evidence in support of the contribution of established linguistic knowledge, the exact nature of the underlying skills involved in Sentence Repetition is poorly understood (Polisenska, Chiat, & Roy, 2015). Relatively few studies have directly manipulated linguistic factors such as semantics or intonation in order to examine their impact on children's repetition (e.g., Bonvillian, Raeburn, & Horan, 1979; Polisenska et al., 2015) and even fewer studies in children with language impairment (Frizelle & Fletcher, 2014). It is important to gain a better understanding of the factors that underpin Sentence Repetition to inform its use as an assessment.

Building on extensive research in English and other languages, Sentence Repetition may hold promise as an assessment tool for Arabic-speaking preschool children as well. This is not to suggest that it replaces the need for developing broad language assessments or other forms of assessments. With the shortage of normative data and absence of standardized assessments, Sentence Repetition has great merit as a clinical tool: (1) it is quick and easy to administer and score (Gardner, Froud, McClelland, & van der Lely, 2006; Redmond, 2005); (2) it is precise in that it can be designed to target specific grammatical structures (Fujiki & Willbrand, 1982); (3) it can examine a wide range of structures with relatively few instances (Fujiki & Brinton, 1983); and (4) it easily adaptable and can be customized to target problematic structures in different languages (Marinis & Armon-Lotem, 2015). In a small-scale study carried out as part of my Master's degree with 30 Typically Developing Najdi Arabic-Speaking children, Sentence Repetition was found to be sensitive to the age of participants and opened up questions addressed in my PhD research: Would the findings be replicated on a larger scale? Would Sentence Repetition identify children with Language Concerns? What is it testing? Could the linguistic manipulation of test items throw light on the underlying skills involved in Sentence Repetition? Would the pattern of performance across the different linguistic conditions be comparable between different age groups of Typically Developing children? Would it be comparable between Typically Developing children and children with Language Concerns? In addressing these questions, the aim of the current study was to examine the clinical utility of Sentence Repetition in Najdi Arabic-Speaking preschool children and gain a better understanding of the underlying processes involved in immediate repetition.

1.1.1 Outline of the thesis

This chapter commences with a historical overview of the use of Sentence Repetition in psycholinguistic research and language assessment. This is followed by an in-depth review of available Sentence Repetition tests, examining its use as a measure of language development with special consideration given to the design of tests (language, targets, and scoring), their psychometric properties, and their sensitivity to age and gender of participants.

Chapter 2 provides an overview of studies that evaluated the diagnostic accuracy of Sentence Repetition as a marker for SLI in English and other languages. It also explores the underlying processes involved in Sentence Repetition by looking at the relations between Sentence Repetition and other marker tasks that are viewed as either assessments of memory (e.g., digit span) or language (e.g., past tense elicitation task), relations between marker tasks and broad language assessments, and relations between different categories of qualitative scores on the same Sentence Repetition test. It concludes with an in-depth examination of the findings of a Cantonese study (Stokes et al., 2006), the only diagnostic accuracy study that explored how test construction on two immediate repetition tests (nonword and Sentence Repetition) influenced children's performance.

Chapter 3 highlights the influence of linguistic knowledge on repetition tests by examining studies that manipulated linguistic factors in serial recall and Sentence Repetition tests to determine how these factors affected children's performance.

Chapter 4 presents the research questions and the study design. It describes in detail how three measures of immediate repetition (Sentence Repetition, Verbal Short Term Memory span, and Anomalous Sentence Repetition tests) were developed/adapted, how targets were linguistically manipulated, and the findings of the pilot study. It goes on to list the recruitment criteria for the main study, participants' characteristics, test stimuli, and the procedures for the administration and scoring of the tests.

Results are presented in *Chapter 5* and examine the psychometric properties of the Sentence Repetition, Verbal Short Term Memory span, and Anomalous Sentence Repetition tests as well as the levels and patterns of performance across different age groups of Typically Developing children and between children with Language Concerns and their age and nonverbal Intelligence Quotient (IQ) matched controls.

A discussion of the findings is provided in *Chapter 6*, along with clinical implications and limitations of the study, and concludes with a roadmap for future research.

1.2 Historical Background on Sentence Repetition

One of the earliest mentions of repetition was by Jespersen (1922):

> *"One thing that plays a great role in children's acquisition of language, and especially in their early attempts to form sentences, is Echoism: the fact that children echo what is said to them"* (p. 135)

While the above quote highlights what ignited interest in repetition and the possible role it played in language acquisition, only when Chomsky's (1965) theory of Transformational Generative Grammar came to light in the 1960s that a flurry of articles were published on repetition. Chomsky hypothesized that children did not acquire language by repeating or memorizing the surface

structure of an adult sentence. The surface structure of an utterance carries the phonological relationships between sentence morphemes including stress and intonation and is specific to each language (Dale, 1976). Rather, children acquired language by using the deep structure of a sentence and a set of transformational rules to generate an infinite number of sentences modelled to them by adults and novel sentences never heard before. The deep structure of a sentence is innate and specifies the semantic relations between sentence morphemes. It is converted into surface structure by a set of transformational rules (Dale, 1976). According to this theory, the grammar used by children in their repeated utterances should be similar to the grammar used in spontaneously produced utterances. To investigate this hypothesis, a number of studies investigated the relationship between the grammar used in repeated and spontaneously produced sentences (Bloom, Hood, & Lightbown, 1974; Brown & Bellugi, 1964; Brown & Fraser, 1963; Ervin-Tripp, 1964; Menyuk, 1963; Moerk, 1977; Ramer, 1976).

Studies have been inconclusive with regard to the exact nature of the relationship between repeated and spontaneously produced utterances. Some studies found that the grammar used by children in both types of utterance were similar (Brown & Bellugi, 1964; Brown & Fraser, 1963; Ervin-Tripp, 1964). In contrast, Menyuk (1963) found that the grammar used in repeated utterances overestimated the grammar used in spontaneous utterances. Other studies argued that it was not a static relationship and that its direction was dependent on whether the target structure was already mastered or was in the process of being acquired (emerging) (Bloom et al., 1974; Moerk, 1977; Ramer, 1976). In the case of mastered structures, imitative utterances tended to underestimate their occurrence in spontaneous utterances while they tended to overestimate emerging structures (Bloom et al., 1974; Moerk, 1977; Ramer, 1976).

Rees (1975) pointed out that the discrepancy between findings of early studies that investigated repetition may be due to a lack of consensus with regard to terminology, definition, and methodology used. Among the terms used to refer to repetition were imitation, echoism, mimicry, copying, and matching (Rees, 1975). The commonality between the different terms used to refer to repetition was that each involved the presence of an adult model utterance that the child attempted to reproduce. Methodologically, the studies can be broadly categorized into two main types: experimental studies that investigated elicited imitation and descriptive studies that investigated spontaneous imitation.

In elicited imitation, the experimenter instructed the child to repeat a set of target sentences: for example, before the sentences were presented the experimenter used the phrase “say what I say” in the Brown and Fraser (1963) study. The target sentences were constructed carefully to include specific morphemes or transformational rules. In spontaneous imitation, the experimenter did not prompt the child to repeat specific utterances but rather focused on imitations

that occurred naturally in the communication cycle (Prutting & Connolly, 1976). Although this categorization was not always clear-cut, for example in the Moerk (1977) study, a communication partner other than the experimenter sometimes "**prodded**" (p. 189) the child to imitate a model sentence. If a child imitated the model sentence, it was not discounted, although the author stated that this occurred rarely.

Studies varied in how they defined repetition temporally. Some studies limited their investigation to utterances that were repeated immediately after a model utterance without allowing gaps of time or intervening utterances (Brown & Fraser, 1963; Ervin-Tripp, 1964). Other studies were not as strict with their definition of imitation temporally, allowing for both immediate and delayed repetitions of model sentences. For example, Bloom et al. (1974) allowed up to five utterances between the model and repeated utterance and Moerk (1977) made allowances for an intervening utterance and a short interval of silence between the model and repeated utterance. Both studies did not identify a specific time limit.

Studies also differed in how topographically similar the repeated utterance had to be to the model sentence in order for it to be included in the analysis. In the studies of spontaneous imitation discussed here, all made allowances for imitative utterances that included omission errors but varied when it came to substitution errors. Ervin-Tripp (1964) and Bloom et al. (1974) excluded imitative utterances that contained substitution errors. Moerk (1977) on the other hand allowed for substitutions that he described as:

> *"an assimilation to the child's system of grammatical rules, when, for example, the child imitated the modeled* ***I vacuum*** *as* ***Me vacuum****"* (p. 189)

Studies that found an equal relationship between the grammar used in repeated and spontaneous utterances will be addressed first. In the Brown and Fraser (1963) study, 13 simple sentences of various grammatical types were administered in random order to six children between the ages of 25 to 35 months. Three types of scores were calculated for each child: an average Mean Length of Utterance (MLU) score was calculated for (1) all imitated sentences, (2) the total number of correctly imitated morphemes for initial, medial, and final position, and (3) the total percent correct for two syntactic categories: content and function morphemes. Content morphemes belonged to an open class syntactic category with many possible members and were divided into nouns, adjectives, and verbs. Function morphemes belonged to a closed class syntactic category with few possible members and were divided into articles, pronouns, auxiliary verb, copula, and inflections. Results showed that morphemes in the final serial position showed the highest likelihood of retention in comparison to the initial or medial positions (recency effect). The authors used the term "Telegraphic" to described sentences imitated by younger children because of the similarities between their contents and telegrams, an expensive mode of communication used in the

day. Customers were charged according to the number of words in a telegram. To reduce cost, customers condensed the usual number of words in a sentence by omitting function words, which increased the cost of a message without influencing the gist of the message. While content words were retained because they were vital for the meaning of the telegram and were not predictable. In the same way, younger children veered towards systematically omitting more function words in comparison to content words and tended to omit inflections such as past tense "ed." As the age of the children increased, their imitative utterances were less telegraphic and closely approached the number of morphemes in the model sentence. Also, children tended to maintain the correct word order in their repetitions. The following is an example from the study page:

Model Sentence:	I showed you the book
Eve (25 ½ months) response:	I show book
June (35 ½ months) response:	Show you the book

The authors drew three similarities between the performance of the children on the elicited imitation task and the spontaneous language samples collected for each child: (1) they were similar in MLU; (2) the utterances of younger children tended to be telegraphic in nature with the omission of function morphemes in obligatory contexts in spontaneous speech as well, for example, "two ball" where the plural suffix s was omitted; and (3) they both maintained the adult word order of an utterance.

They argued that this systemic reduction of function morphemes in sentences could not be merely explained by the children's short-term memory capacity because content and function morphemes differed in five main linguistic features. Content morphemes tend to be located in the final position of a sentence; they are reference-making forms; they belong to a large and expandable syntactic category including verbs adjectives and nouns; they relatively speaking cannot be predicted from the context of sentences; and they usually receive heavier stress in English. Function morphemes tend to be located in the middle position of a sentence; they are not reference-making forms; they belong to a small and nonexpendable syntactic category that includes morphemes such as articles, pronouns and inflections; they are relatively unpredictable from the sentence context; and they usually receive weaker stress. Therefore, they hypothesized that the children's knowledge of grammar influenced their performance on the elicited imitation task. Where function morphemes had a better chance of being retained was when the child acquired the grammatical knowledge of the use of that particular morpheme.

Brown and Fraser (1963) were not alone in describing the nature of children's utterances as telegraphic. W. Miller and Ervin (1964) stated:

> ***"It is often striking that one can provide a translation of children's utterances into adult utterances by the addition of function words and inflectional affixes. It appears that the***

children select the stressed utterance segments, which usually carry the most information" (p. 13)

Their results were consistent with the findings of Brown and Bellugi (1964) and Ervin-Tripp (1964). Both studies were descriptive and focused on spontaneous rather than elicited imitation. Brown and Bellugi (1964) compared the imitative and non-imitative spontaneous utterances of Adam and Eve, from Brown's famous longitudinal study when both children were in the early stages of language development. Their MLU was equal to 1.75. Their findings corresponded to the three similarities identified above in the Brown and Fraser (1963) study with regards to MLU, the telegraphic nature of utterances, and maintained word order of adult utterances. Ervin-Tripp (1964) compared imitative and non-imitative spontaneous utterances of five children between the ages of 1;10 to 2;5. She also found that both utterance types were telegraphic in nature and maintained adult word order. Rather than using percent correct of different syntactic categories to compare the grammar of both utterance types, the researcher identified the transformational rules used in non-imitative utterances and their frequency was compared with imitative utterance. There was no difference found between the two utterance types.

In contrast to the above-mentioned studies, Menyuk (1963) reported that the grammar used by children in imitative utterances overestimated the grammar they used in their spontaneous utterances. A set of 27 sentences representing various transformational rules were administered to 14 nursery-school children (mean age: 3;8) and 25 kindergarten children (mean age: 3;3). There could be a number of reasons for the discrepancy between Menyuk's (1963) findings and Brown and Fraser (1963). Although both studies used an elicited imitation task, they differed in how they scored and analysed the data. Menyuk (1963) scored sentences as correct or incorrect repetitions (all or none) rather than comparing the total percent correct of content and function words as in the Brown and Fraser (1963) study. As for the analysis, it was limited to transformations that were spontaneously produced by less than 50% of the children in both age groups. However, when the analysis included all transformational rules, a significant correlation was found between the number of syntactic structures children produced in their spontaneous speech and elicited repetitions in both age groups. Moreover, age was an important factor, with children in the Kindergarten group performing better than children in the nursery group.

Most importantly and irrespective of whether some or all the transformational rules were included in the analysis, Menyuk (1963) argued that Sentence Repetition was not mere parroting and was also influenced by the type of structure used in the sentence. This was supported by the lack of correlation between the length of sentences and the accuracy of repetition in both age groups. In addition, the modification of transformational rules in repeated sentences were parallel to modifications generally observed in the spontaneous production of children in the same age group.

Based on these reported findings, it can be argued that Menyuk's (1963) study expanded the findings of Brown and Fraser (1963) to a wider age range and supported the finding that elicited imitation taps into linguistic knowledge.

In a follow study, Menyuk and Looney (1972) compared the performance of children with language impairment (mean age 6;2) and controls (mean age 4;6) on an elicited imitation task. Sentences were scored as ether correct or incorrect. Any deviation was considered an error. Overall, children with language impairment obtained poorer scores than controls. In addition, the profile of performance varied in the two groups. Children with language impairment were influenced by sentence type. The highest frequencies of errors were noted for negative subject and passive sentence types followed by questions and negative sentences respectively. In the control group, the influence of sentence type was less marked. The influence of length was also investigated. As in the Menyuk's (1963), no difference was found between the frequency of errors for three sentence lengths 3 vs. 4 vs. 5 words in both groups. Therefore, Menyuk and Looney (1972) extended the finding that linguistic knowledge played a role in the repetition of children with language impairment.

Finally, some studies found a dynamic relationship between imitative and spontaneous utterances. Bloom et al. (1974) and Moerk (1977) found that structures were not imitated in two instances: before they were acquired and after they were mastered in spontaneous productions. Structures that were beginning to emerge in spontaneous productions were imitated more than they were produced. Both studies were longitudinal and focused on spontaneous imitations. Bloom et al. (1974) included six children whose ages ranged between 16 to 21 months and were at stage 1 MLU (1.0-2.0) at the start of the study. Moerk (1977) included two older children aged 28 and 31 months.

Although the main findings of these two studies appear to conflict with Brown and Fraser (1963)'s study, they agree in several aspects. Children did not imitate structures that were completely absent from their spontaneous productions. Imitative utterances maintained the same word order as spontaneous productions. Imitative utterances were developmentally progressive. Finally, and most importantly, both studies concluded that imitative utterances reflected the established linguistic knowledge of the child and were not merely parroted versions of model utterances. This is because how and when children imitated the target structures was dependent on whether the structure was novel, emerging, or mastered.

To conclude, while there was no consensus on the exact relationship between the grammar used in imitation and spontaneous language, all of the studies agreed that children's imitation of model sentences taps into language and children were not merely parroting adult utterances. The studies also highlighted the need for a unified definition of Sentence Repetition. In the present

study, Sentence Repetition is defined as a form of elicited imitation where participants are instructed to verbally repeat a set of stimuli presented to them auditorily by the researcher without any accompanying pictures. Imitation is immediate, with no allowances for verbal utterances by researcher or participant and no prolonged period of silence between the presentation of the stimulus and the child's response.

Studies of elicited imitation provide evidence of the viability of the use of Sentence Repetition as an assessment tool. Both Brown and Fraser (1963) and Menyuk (1964) showed a developmental trend: scores improved as age increased. Menyuk and Looney (1972) showed that it differentiated between children with language impairment and controls. Finally, the systemic deviations found in the repetition of children with Typical and atypical development with regard to the superiority of content words over function words and the influence of sentence type on frequency of errors supports the notion that repetition is not merely parroting and receives support from established linguistic knowledge.

1.3 Sentence Repetition as an Assessment Tool

Sentence Repetition tests have been utilized by speech-language therapists and researchers interested in child language. They are most commonly available as a subtest of an expressive language assessment battery such as the Recalling Sentences subtest of the CELF-P (Wiig, Secord, & Semel, 2000) and the Sentence Imitation subtest of the TOLD (Newcomer & Hammill, 1997). Less commonly, Sentence Repetition tests are available as stand-alone tests such as the Elicited Language Imitation (ELI; Carrow, 1974) and the Sentence Imitation Test (SIT; Seeff-Gabriel et al., 2008)

The section commences with examples of Sentence Repetition tests that provided details of the performance of children across different age groups, followed by a close examination of the test targets and scoring. It goes on to assess the clinical utility of the tests by looking at the psychometric properties of these tests and their sensitivity to age.

1.3.1 Examples of Sentence Repetition assessments and why they are language-specific

A summary of the main Sentence Repetition tests discussed in this section is presented in Table 1.1 and shows the name, type of test, language of test and test sample. It also provides a brief description of the test stimuli, administration procedure, and scoring.

1.3.2 Targets

In focusing on the content of the Sentence Repetition tests featured in Table 1.1, Devescovi and Caselli (2007) stressed the importance of adapting assessments that reflect the topographic characteristics of a particular language rather than just merely translating a test from English. It is common practice in Italy, Qatar, and Saudi to assess children's language ability using translated

versions of an English test (Devescovi & Caselli, 2007; Shaalan, 2010). This is due to the lack of standardized tests available in these languages. The main drawback of translating from English to languages such as Italian or Arabic is that the test would not reflect the morphological richness of ether of the two languages. A key Grammatical Morpheme may be present in one language but not the other, For example in both Italian and Arabic adjectives are marked for agreement with nouns for number and gender (e.g., in Italian **piccol-a** small-feminine.singular and in Arabic **<u>s</u>iʁi:r-a** small-feminine.singular) while the same is not true for adjectives in English. Even in cases where the Grammatical Morpheme exists in both languages, they may follow different developmental trajectories with a Grammatical Morpheme developing early in one language and late in the other (Shaalan, 2009).

Table 1.1 *Summary of Sentence Repetition Assessments*

Name	Type	Language	Sample, Age	Stimuli	Administration	Scoring
Elicited Language Imitation (ELI) (Carrow, 1974)	Diagnostic stand-alone test	American English	Standardized on 475 children 3 to 7;11	51 sentences 2-10 words in length Wide range of Grammatical Morphemes targets. Sentences range from simple to complex syntax.	Children are instructed to repeat sentences and their responses are audio recorded for later transcription	Total Error score Two subscores: Grammatical category: noun verb adjective Type of errors substitution omission addition transposition
Grammar and Phonology Screening (GAPS) (Gardner et al., 2006)	Screening subtest designed to assess morpho-syntax	British English	Standardized on 668 children 3;6 to 6;6	11 sentences consisting of early acquired words with simple phonological structure. Targets morpho-syntactic structures mastered by Typically Developing children 3 to 4 years of age but difficult for children with language impairment.	Children are presented with a short picture storybook and are asked to repeat sentences to an alien character named "Bik," who can only understand children. 10 min to administer.	Sentence Score All/none

Sentence Imitation Test (SIT) (Seeff-Gabriel et al., 2008)	Diagnostic stand-alone test combined with the Preschool Repetition Test to form the Early Repetition Battery	British English	Standardized on 383 children 2;6 to 5;11	27 test items (6 to 9 words in length) Simple sentences constructed according to a graded developmental syntactic hierarchy Targets a wide range of morpho-syntactic structures Words are high frequency, semantically familiar, short, contain early acquired phonemes	Live administration and scoring Fixed Order Increase in length and complexity	Level 1: Sentence Score (All/none) Level 2: Content Word Function Word Inflection Level 3: Content: Nouns Verbs, Adjectives. Function: Copula, Determiners, Prepositions
Devescovi and Caselli (2007)	Novel stand-alone test developed for a research study to assess morpho-syntax	Italian	100 children 2 to 4 years	27 test items (3 to 6 words in length) Simple sentences containing familiar words and targets a range of early acquired morpho-syntactic structures.	Test presented live with illustrations for each sentence Responses recorded for later transcription	MLU-word Complete Sentences (all/none) Omissions and Errors in 5 grammatical categories: Articles Prepositions Verbs Nouns Modifiers
Shaalan (2010)	Novel diagnostic test developed for a research study	Gulf Arabic	112 children: 86 Controls 26 SLI 4;6 to 9;4	41 test items with a mix of simple and complex sentences targets selected based on spontaneous language samples collected from 35 Gulf Arabic speaking children between ages of 2;11 to 4;11 years old	Test presented and scored live Sentences presented in a fixed order with increasing length and grammatical complexity	CELF scoring: Each sentence 3 = no errors 2= one error 1= two or three errors 0= four or more errors.

Wallan (2006)	Novel stand-alone test developed for a research study	Najdi Arabic	30 children 3 to 5 years	12 test items (length: 5 to 6 words, 8-12 morphemes). Simple sentences Targets a range of Grammatical Morphemes	Test presented live and responses were audio recorded for later transcription Fixed order	Total Repetition score Grammatical scores: Lexical score Preposition score Gender agreement score Article score Verb tense score

It may also be necessary to adapt the Sentence Repetition tests for different dialects within the same language. This is especially true for Arabic, which is classified as a diglossic language. Ferguson (1959) defines diglossia as the co-existence of two varieties of the same language. Gulf Arabic and Najdi Arabic are spoken regional dialects and differ from Modern Standard Arabic in syntax, semantics, morphology, and phonology (Shaalan, 2010). Children are not exposed to Modern Standard Arabic until they enter school, with the exception of a few television programs. Parents communicate with children using the regional dialect. Asking a Gulf or Najdi Arabic speaking child to repeat a sentence in Modern Standard Arabic would be the equivalent of asking a child from Texas or East London to repeat "To thine own self be true" (Shakespeare, trans. 2000, *Ham.* 1.3.84-86), with the difference that Modern Standard Arabic is still spoken today in schools and formal settings and is the form of written language, while Early Modern English is rarely used. Moreover, Hemingway, Montague, and Bradley (1981) found that a screening version of the ELI (Carrow, 1974) identified three distinct subgroups groups of African American children based on error type and frequency: those who spoke Standard American English, those who spoke African American English, and those with language impairment. They argued that failure to take dialectal variation into account would run the risk of misdiagnosing children who speak African American English as language impaired. Rather than making allowances for dialectal variation when scoring a Sentence Repetition test, Shaalan (2010) and Wallan (2006) developed the tests in the respective regional dialect. Shaalan (2010) manipulated length and grammatical complexity simultaneously, making it difficult to pinpoint what particular grammatical structures were difficult for children with SLI. One way of overcoming this difficulty is by providing a fine-grained qualitative scoring method such as that featured in the English SIT, which provides a profile of performance according to morpho-syntactic category at its second level of scoring and allows for further breakdown within each category.

1.3.3 Scoring

The scoring systems featured in Table 1.1 fall on a continuum with regard to how discriminating they are and how much, if any qualitative information they provide. On the one end of the continuum, we have the all or none scoring method. This scoring method provides a purely quantitative score and is the least discriminating. If any deviation from the target sentence occurs, that item is scored as incorrect. As can be seen from Table 1.1 this scoring method is used in three tests: the Grammar and Phonology Screening (GAPS; Gardner et al., 2006), SIT (Seeff-Gabriel et al., 2008), and Devescovi and Caselli (2007). It is also a commonly used scoring system in Sentence Repetition subtests of standardised language assessments such as the TOLD (Newcomer & Hammill, 1997). The CELF (Wiig et al., 2000) scoring method falls further along the quantitative continuum, it is more graded/discriminating, and provides a leeway for up to three

errors per sentence. The ELI (Carrow, 1974) Total Error score and Wallan's (2006) Total Repetition score fall on the other end of the continuum. The Total Error score (Carrow, 1974) tallies the overall number of errors. Unlike the CELF scoring system, it does not cap the number of errors allowed per sentence. The Total Repetition score (Wallan, 2006) is a cumulative accuracy score of the five grammatical scores identified in Table 1.1.

The tests that utilise qualitative scoring methods differ in the scope of the morpho-syntactic categories they score. The ELI (Carrow, 1974), Devescovi and Caselli (2007) and Wallan (2006) scoring methods employ a mix of narrow morpho-syntactic categories, with the difference that the ELI (Carrow, 1974) and Devescovi and Caselli (2007) tally errors while Wallan (2006) tallies correct imitations. The SIT (Seeff-Gabriel et al., 2008) scoring method is unique in that it provides two levels of qualitative scoring: an intermediate scoring level which includes three broad morpho-syntactic categories, and a third level of scoring which breaks down the Content and Function word categories further into narrow morpho-syntactic categories such as nouns and prepositions. This method of scoring allows for the identification of patterns of performance across the three morpho-syntactic categories. Clinically, it requires less time to score and it is less tedious with the final level scored only when the child exhibits a deficit in the corresponding broad category. The SIT scoring method is the most comprehensive and strikes a balance between the quantitative and qualitative information it provides.

1.3.4 Psychometric properties (valid measure of language)

Table 1.2 presents the psychometric characteristics of the Sentence Repetition tests discussed in this section. It focuses on three measures of reliability (inter-rater reliability, test retest reliability, and internal consistency) and two measures of validity (concurrent and construct validity). Reliability of a test can be defined as its ability to produce consistent results under different conditions (Field, 2009). The different conditions include different points in time (test-retest), different assessors (inter-rater), and different items from the same test (internal consistency). According to Landis and Koch (1977) reliability coefficient values from .61 to .80 indicate substantial agreement, and from .81 to 1.00 indicate almost perfect agreement. Devescovi and Caselli (2007) reported inter-rater reliability as the proportion of agreement between two coders and it was found to be high: 97% for agreement for the complete sentence score and 92% for errors and omissions. To establish test-retest reliability, all participants were retested within 10 days. High correlation coefficients were reported for MLU-Word (r = .93) and complete sentence score (r = .94). Correlation coefficient scores for omissions were also high for the five grammatical categories: articles, prepositions, verbs, nouns, and modifiers and ranged from r = .73 to .95. The stability of omission scores is of particular interest because the pattern of omissions in grammatical categories according to age groups was the most informative score (discussed below). Wallan

(2006) reported the proportion of agreement between two coders on morpheme transcription was 90%. The reliability coefficient values are reported in Table 1.2 and show that the tests reported high reliability coefficients, indicating that the tests were stable across time, different scorers, and different items from the same tests.

Table 1.2: *Psychometric Characteristics of the Sentence Repetition Tests*

		Reliability			**Validity**	
		Inter-rater	**Test/ Retest**	**Internal Consistency**	**Concurrent**	**Construct**
Elicited Language Imitation (ELI)		.98	.98	n/a	✓	✓
Grammar and Phonology Screening (GAPS)		n/a	n/a	.86	✓	✓
Sentence Imitation Test (SIT)	Sentence	.98	.88	.92	✓	✓
	Content word	.99	.80	.95	✓	✓
	Function word	.99	.78	.95	✓	✓
	Inflection	.98	.92	.89	✓	✓
Devescovi & Caselli (2007)		✓	✓	n/a	✓	✓
Shalaan (2010)		n/a	✓	.89	✓	✓
Wallan (2006)		✓	n/a	n/a	n/a	n/a

Note. n/a = not available.

Validity is defined as the extent to which a test measures the construct it claims to measure (Mayers, 2013). It establishes the degree of confidence we can place on the assumptions made about individuals, based on their test scores (Streiner & Norman, 1995). There are a range of measures that examine validity of a test. This section will focus on two of those measures: concurrent and construct validity. Concurrent validity is part of criterion validity, and is established by determining the extent to which test scores agree with other valid tests that measure the same construct (Paul, 2007). All the tests with the exception of Wallan (2006) provided evidence to support their concurrent validity.

The total error score of the ELI test was compared with the results of the Developmental Sentence Scoring (DSS) test (L. L. Lee & Canter, 1971). The DSS is a standardized spontaneous

language analysis measure that focuses on grammatical complexity. According to a survey by Kemp and Klee (1997), the DSS was the most common standardized analysis method used by speech language therapists in the United States. High correlations were found between the ELI and DSS scores of Typically Developing children (r = .79; Carrow, 1974). Also, Werner and Kresheck (1981) investigated the relationship between the DSS and ELI for different age groups and found a significant and high correlation for 4-year-old participants (r = .66) and 5-year-old participants (r = .60). Furthermore, Dailey and Boxx (1979) found that children with Language Impairment obtained a similar score on Brown's (1973) 14 Grammatical Morphemes on the ELI and a spontaneous language sample: the overall percent correct score was 41% on the ELI and 44% for the spontaneous language sample.

During the pilot stage, which included 148 children of different age groups, the GAPS grammar subtest (Sentence Repetition) scores showed significant correlations (when age was partialed out) with two subtests of CELF-P (Wiig et al., 2000) Sentence Structure (r = .52) and Word Structure (r = .43; Gardner et al., 2006). The CELF-P is a commonly used test by speech language therapists and researchers to diagnose Language Impairment (Dockrell, 2001). The two subtests assess children's ability to understand and use grammatical markers respectively (Gardner et al., 2006).

In the standardization sample of the SIT, significant correlations were found when age was taken as a covariate between the SIT scores and the PreSchool Repetition (PSRep; Seeff-Gabriel et al., 2008) test scores. Correlations ranged from r =.48 to .53. The PSRep along with the SIT are part of the Early Repetition Battery (ERB; Seeff-Gabriel et al., 2008). Like the SIT, it assesses repetition but at a different level: single word/nonword level. In an earlier study (Chiat & Roy, 2008) the SIT was administered to 152 clinically referred children, significant correlations were found between the Function Word score of the SIT-16 (a shortened version of the SIT) and CELF-P UK Recalling Sentences Subtest (r = .62) and the Renfrew Action Picture Test (RAPT) grammar score (r = .58; Renfrew, 1997). The RAPT is a standardized test that assesses the use of grammatical markers through picture description. In both correlational analyses, age was a covariate.

To investigate the relationship between grammar used in spontaneous language and repeated in the Sentence Repetition test, Devescovi and Caselli (2007) selected 25 children from the original 100 (five from each age group) to undergo further testing. The spontaneous language sample and performance on the Sentence Repetition test were compared based on MLU-word, omission of articles, and the number of verbs. Significant partial correlations (controlling for age) were found for omission of articles (r = .33) and number of verbs (r = .37).

In order to establish the concurrent validity of the Sentence Repetition test, Shaalan (2010) compared the scores on the Sentence Repetition test to other novel tests developed for his project: the Sentence Comprehension test, the Expressive Language test, and the Arabic Picture Vocabulary test. Significant correlations were found ranging from $r = .34$ to .69. The highest correlation was found between the SR test and the Expressive Language test (.69), which assess the use of morpho-syntactic structures commonly used by Gulf Arabic speaking children.

The second type of validity presented here is construct validity. Construct validity can be defined as the test's ability to measure the construct it is intended to measure (Streiner & Norman, 1995). There are several ways to test construct validity: it can be established by conducting correlational analyses to examine whether the test scores relate to others tests measuring the same construct or other variables or constructs that are predicted to relate to the construct of interest (convergent validity) or not (divergent validity); and by examining whether the tests discriminates between groups of participants known to differ on the construct of interest (Streiner & Norman, 1995). The influence of age on test scores and the lack of influence of gender are presented below. The main focus here will be on how the Sentence Repetition tests featured in Table 1.2 relate to general language measures and their ability to discriminate between exceptional groups. With regards to correlation with broad language measures, the SIT scores of a subgroup from the standardization sample ($n = 321$) highly correlated when age was controlled for with the Preschool Language Scale 4th edition (PLS-4^{UK}; Zimmerman, Pond, & Steiner, 2009). Correlation coefficients for the SIT scores ranged from $r = .45$ to .59. The same relationship was reported for 152 clinically referred children in an earlier study (Chiat & Roy, 2008). High correlations were found between the Function Word score of the SIT-16 and the Preschool Language Scale 3rd edition (PLS-3^{UK}; Zimmerman, Steiner, Pond, Boucher, & Lewis, 1997) Auditory ($r = .48$) and Expressive subscales ($r = .66$).

Evidence to support the ability of Sentence Repetition tests to discriminate between children with language impairment from Typically Developing children was reported for all the tests featured in Table 1.2, with the exception of Devescovi and Caselli (2007) and Wallan (2006).

The ELI (Carrow, 1974) successfully discriminated between children according to their language ability. Children with Language Impairment obtained lower Total Error scores than Typical children (Cornelius, 1974). More specifically, they scored lower in the following morpho-syntactic categories: articles, adjectives, noun plurals, pronouns, verbs, negatives, prepositions, and conjunctions. To examine the ability of the GAPS (Gardner et al., 2006) to identify children with SLI, it was administered to 17 children who were previously diagnosed with SLI by speech and language therapist. At the cut-off percentile score of 10, the Grammar subtest identified 65% of children with SLI; the cut-off percentile score of 15, it identified 71% of children with SLI.

Finally, Shaalan (2010) used the four novel assessments developed for the study (mentioned above) to identify children with SLI. In order to receive a diagnosis of SLI, children had to obtain a Z-score of -2.0 or more on one test or -1.5 or more on at least two of the four language tests. Children with SLI and the age-matched controls were grouped according to their corresponding age band. As a group, children with SLI who ranged in age between 4;6 to 9;4 obtained a mean score lower than the youngest age group of controls aged 4;6 to 5;11. The SLI group obtained a mean score of 59.5 and the youngest Typical group obtained a mean score of 69.8 (out of a maximum score of 123). Overall children, with SLI obtained scores similar to age-matched controls who were 2 years younger than they were. Because of the nature of the test used, which manipulated length and complexity simultaneously, it was difficult to pinpoint specific areas of difficulty for children with SLI. The scoring method, which followed the CELF scoring system, was more discriminating at group level in comparison to an all or none scoring method. However, it did not have the advantage of the SIT or ELI scoring methods of providing information on which morpho-syntactic structures were most difficult for children with SLI.

1.3.5 Age and gender sensitivity (valid measure of development)

The findings related to age sensitivity will be presented first followed by gender sensitivity. All the tests without exception were able to discriminate between participants according to their age group and showed developmental trends with older age groups obtaining better scores in comparison to younger age groups. Due to the difference in scoring methods as discussed above, the tests differed in how much information they provided regarding which specific morpho-syntactic structures proved to be difficult for young children and how if at all the pattern of performance changed as children got older. The tests that incorporated a narrow or broad qualitative scoring system proved to be most informative.

Focusing on the quantitative group difference, the ELI (Carrow, 1974) Total Error score decreased as the age of participants increased. Age correlated most highly with the grammatical category of verbs and pronouns. The Complete Sentence score of the GAPS (Gardner et al., 2006) and Devescovi and Caselli (2007) study, and the CELF score of Shaalan (2010) and the Total Sentence score (Wallan, 2006) all showed improvement with age. All five grammatical scores of Wallan (2006) were sensitive to age. A common finding was that not all adjacent age groups differed significantly from each other. The Devescovi and Caselli (2007) results highlighted the fact that not all scoring methods were equally sensitive to age. Unlike the Complete Sentence score, the Mean Length of Utterance-word was only sensitive in the two youngest 6-month age groups. The SIT (Seeff-Gabriel et al., 2008) reported a strong correlation with age ($r = .60$). The raw scores were not normally distributed, showing floor effects in the youngest age groups and ceiling effects in the oldest age groups. The Sentence, Content Word, Function Word, and Inflection

scores showed an increase in mean scores as age increased. High standard deviation (SD) values were noted for the younger age groups indicating a large variability in scores. The SD values decreased as age increased due to less variability in scores.

The qualitative scoring systems revealed a common thread that ran through the results of three tests: SIT (Seeff-Gabriel et al., 2008), Devescovi and Caselli (2007), and Wallan (2006). Irrespective of language typology, the repetition of young children was telegraphic in nature: content/Lexical Morphemes were more likely to be retained while function words and inflections were more likely to be omitted. This pattern of performance/dominance was clearest in the youngest age groups and reduced in magnitude as children got older. Although the ELI (Carrow, 1974) test employed a qualitative scoring method, it did not provide a detailed breakdown of grammatical scores according to age.

The results of the SIT (Seeff-Gabriel et al., 2008) showed an advantage of the Content Word score over the Function Word score in all the six age groups. This advantage was greatest in the two youngest age groups and decreased in magnitude from the age group 3;6 to 3;11 onwards. For children in the youngest age group (2;6 to 2;11), then mean percentage score of Content Word was 54.53 in comparison to 44.21. For children in the oldest age group (5;0 to 5;11), the mean percentage score was 97.79 and 94.34 for Content and Function Word scores, respectively.

Devescovi and Caselli (2007) found that the omission of grammatical categories occurred most commonly in 2-year-olds. Their repetitions consisted mainly of nouns, modifiers, and a few verbs. The most frequently omitted category was articles. The omission of grammatical/function words almost completely disappeared by the age of 3 years. The omission of articles, however, continued even in the oldest age groups.

In the Wallan (2006) study, participants in the youngest age group obtained the highest mean percentage scores in the Lexical category (31.50), while the mean percentage scores of Preposition, Gender Agreement, Article, and Verb Tense ranged from 7.19 to 27.86 (the combined mean percentage score of the four grammatical categories was 16.90). By the age of 5 years, the mean percentage score of Verb Tense was at ceiling (98.33). The Preposition and Article category improved but remained difficult for 5-year-olds with the mean percentage scores equalling 61.67 and 79.69, respectively.

With regard to the effect of gender on repetition scores, Gardner et al. (2006) reported no effect of gender on the Complete Sentence score and no interaction between age and gender. This was also true for the Wallan (2006) study, which found no effect of gender on the Total Repetition score. In the SIT standardization sample (Seeff-Gabriel et al., 2008), boys and girls did not differ on the Content Word and Inflection score. However, a small but significant advantage for girls was found on the Sentence and Function Word scores. This advantage was limited to two age groups:

2;6 to 2;11 and 3;6 to 3;11. Since the advantage of girls' repetition scores was limited to these two age groups and two out of the four scores, the authors argued that this did not warrant deriving separate norms for boys and girls. Shaalan (2010), Devescovi and Caselli (2007), and Carrow (1974) did not report on gender effects.

To conclude, both scoring methods were able to discriminate at group level between different age groups. Qualitative scoring methods were more informative in comparison to purely quantitative scoring methods and showed that across different language typologies a simple repetition test can tap into the morpho-syntactic abilities of children. The advantage of content words over function words is consistent with the findings of Brown and Fraser (1963) study and extends its findings to older children, a different dialect (British English), as well as other languages (Najdi Arabic and Italian). This adds further support to the role that linguistic knowledge plays in repetition. Furthermore, the qualitative scoring methods allowed for a comparison between the repetition and production of Grammatical Morphemes and was presented above in the validity section for two studies: Devescovi and Caselli (2007) and Carrow (1974). Finally, gender does not seem to be a factor that influences repetition.

1.4 Conclusion

As assessment tools, the Sentence Repetition tests presented here were found to be sensitive to the age of participants. The SIT (Seeff-Gabriel et al., 2008) and Shaalan (2010) tests were sensitive to the language ability of participants. Quantitative scoring methods were able to pick up group differences with regard to the age and language status of participants but failed to identify specific areas of weakness. Qualitative scoring methods overcame this downside and showed that both young children and children with language impairment found Grammatical Morphemes the most difficult to repeat. They were also able to show similarities between the error patterns children produced in their spontaneous productions and their performance on Sentence Repetition tests. Sentence Repetition tests were psychometrically robust and showed adequate levels of reliability and validity.

The role that linguistic knowledge plays in Sentence Repetition was supported by the fact that content/lexical items tended to be retained more often than function/Grammatical Morphemes in young children and children with language impairment, and the significant correlations that were found between Sentence Repetition tests and productive tests of morpho-syntax as well as broad language measures.

Although two tests of Sentence Repetition were available in Arabic, there was still a need to develop a novel test for this study. This is due to the age range, structure, and scoring methods of both tests. The Shaalan (2010) test structure is in a different dialect of Arabic; it was designed to be administered to older children than the target age group for the present study and consisted of long

sentences and complex structures and only included a purely quantitative scoring method. The Wallan (2006) test was designed in the same dialect: it was tested on a small sample, and scoring methods and structure were limited. It included a quantitative scoring method that was too fine grained and needed to be replaced with one that is broader and easier to apply yet discriminating. Its qualitative scoring method needed to include an intermediate level like the SIT (Seeff-Gabriel et al., 2008), but adapted to reflect the morphological richness of Arabic. A further aim of this study was to test the underlying processes involved in Sentence Repetition through developing an Anomalous Sentence Repetition. Therefore, there was a need to develop a new set of sentences that yielded itself to systemic manipulation in order to create the Anomalous Sentence Repetition test.

Chapter Two: Sentence Repetition as a Potential Clinical Marker of Specific Language Impairment

2.1 Introduction

2.1.1 Key terminology

Specific Language Impairment (SLI) is a term first coined by Leonard in 1981 and has become commonly used by researchers since then (Reilly et al., 2014). Multiple definitions of SLI exist; the core concept in each is that language difficulty occurs in the absence of identifiable cognitive, neurological, neuro-developmental, sensory, behavioural, or psychological challenges. Internationally, a number of terms have been used to refer to SLI such as Primary Language Impairment in Canada (Thordardottir et al., 2011) and Developmental Dysphasia in the Czech Republic (Smolik & Vavru, 2014). It is thought to affect approximately 7 % of 5-year-olds based on a population study of over 6,000 school-aged children in the United States (Tomblin et al., 1997), with a higher prevalence for boys in comparison to girls (8% vs. 6%). Results from several longitudinal studies have shown that communicative difficulties experienced by most children with SLI are long lasting and show collateral effects on the children's academic abilities, social, and emotional wellbeing (Conti-Ramsden & Botting, 2008; Fujiki, Spackman, Brinton, & Hall, 2004; Johnson et al., 1999; Law, Rush, Schoon, & Parsons, 2009).

As the definition above implies, SLI is diagnosed through a set of exclusionary criteria, which are intended to confirm that the impairment is solely of language in the absence of an obvious cause (Hesketh & Conti-Ramsden, 2013; Marshall & Morgan, 2015). A benchmark for the diagnosis of SLI is the mismatch/discrepancy between language ability and non-verbal intelligence (Bishop, 2014). Standardized tests of language and nonverbal IQ need to show that the child has impaired language ability with age appropriate nonverbal abilities. Children with impairments in nonverbal abilities as well as are excluded. In addition, children are excluded if diagnosed with a sensory impairment such as hearing loss, neurological impairment such as traumatic brain injury or epilepsy, behavioural impairment such as Attention Deficit Hyperactivity Disorder (ADHD), neuro-developmental impairments such as autism, autism spectrum disorders, or dyslexia. Researchers interested in running group comparisons show a preference for the use of exclusionary criteria because it allows for the selection of a seemingly homogeneous group of participants (Reilly et al., 2014). Clinicians, on the other hand, shy away from diagnosing SLI because the information on nonverbal IQ is not always readily available in many if not most clinical settings (American Speech-Language-Hearing Association, 2012). There has been growing dissatisfaction and wide debate over the use of SLI as a diagnostic label and the exclusionary criteria used to diagnose it (Conti-Ramsden et al., 2001; Ebbels, 2014). This is exemplified in the removal of SLI

as a diagnostic category from the Diagnostic and Statistical Manual of Mental Disorders 5th edition (DSM-5) and replacement with the more general label Language Disorders. Why? Although it is beyond the scope of this thesis to discuss the reasons in detail (see Bishop, 2014; Reilly et al., 2014) below are some of the relevant issues.

2.1.2 1. What constitutes a language impairment?

A diagnosis of SLI relies primarily on obtaining a standardized score below a random and untested cut-off point that supposedly differentiates between impaired and unimpaired language ability (Reilly et al., 2014). For example, in the Tomblin et al. (1997) longitudinal study, participants had to score 1.25 SD below the standardized mean on two or more composite measures of expressive, receptive, vocabulary, grammar, and narrative abilities in order to be diagnosed as SLI. There is a lack of agreement on the profile of language impairment with regards to what core aspects of language are impaired and to what degree (Snowling, 2014). To receive a diagnosis of SLI, it is not clear whether a child with SLI needs to exhibit an impairment in language comprehension, language production, or both. It is also not clear which components of language need to be impaired (morphology, syntax, vocabulary, phonology, and/or pragmatics) and to what degree (Dockrell & Marshall, 2015; Marshall & Morgan, 2015). Due to the lack of guidelines, children with SLI form a heterogeneous group. As Snowling (2014) stated, "There is clearly no one SLI" (p. 438). Reilly et al. (2014) argued that this has consequences for the generalizability of SLI study findings. One additional problem with standardized tests is that the random cut-off points may put children who are bilingual or speakers of a non-standard dialect and children from low-socio-economic backgrounds at a disadvantage (American Speech-Language-Hearing Association, 2012; Roy & Chiat, 2013).

2.1.3 Nonverbal IQ

As with language ability, nonverbal IQ tests employ random and untested cut-off points to determine nonverbal ability (Reilly et al., 2014). Nonverbal IQ scores of children with a history of SLI showed a gradual decline from adolescence to early adulthood (Botting, 2005; Conti-Ramsden, St Clair, Pickles, & Durkin, 2012). This brings into question the criterion to exclude an impairment in nonverbal ability.

2.1.4 Comorbidity

Language impairment frequently occurs with other developmental disorders such as dyslexia (Moll, Hulme, Nag, & Snowling, 2015), ADHD (Redmond, 2005), Pragmatic Language Impairment (PLI; Botting & Conti-Ramsden, 2003), and autism (Riches, Loucas, Baird, Charman, & Simonoff, 2010). This further brings into question the use of the word "Specific" in SLI and the need to exclude these conditions when diagnosing SLI as it is a subject of wide debate (Bishop, 2014; Reilly et al., 2014). As Redmond, Thompson, and Goldstein (2011) point out, if the use of

exclusionary criteria requires differential diagnosis between developmental disorders and SLI, there needs to be an investigation into what are the optimal cut-off points for differential diagnosis.

Due to the dissatisfaction with the exclusionary criteria used to diagnose SLI, there has been an increased interest in investigating clinical markers that allow for the diagnosis of SLI through inclusionary criteria. A marker can be defined as a symptom that can accurately identify individuals who are affected with a particular condition or impairment (Roodenrys & Stokes, 2001).

There are two ways to establish suitability of a clinical marker (Poll et al., 2010). One way is to find a test that discriminates between the group performances of children diagnosed with SLI and age- or language-matched controls. This is done through the use statistical analysis measures such as t-tests, analysis of variance (ANOVA), or discriminant function analysis. The other is to quantify the degree of separation between children with SLI and controls by comparing the performance of the children on an established *gold standard* diagnostic assessment to the proposed clinical marker. A clinical marker needs to have high levels of sensitivity, specificity, and positive Likelihood Ratio (LR+) and low negative Likelihood Ratio (LR-). They are collectively known as Diagnostic Accuracy measures. Table 2.1 details how the diagnostic accuracy measures are calculated.

Table 2.1 *Contingency Table for obtaining Diagnostic Accuracy Measures*

	Diagnosis *Gold Standard*/Criterion	
	True Affected	**True Unaffected**
Positive test result	A True Positive	B False Positive
Negative test result	C False Negative	D True Negative
	A+C	B+D

Note. Sensitivity = A/[A+C]
Specificity = D/[B+D]
LR+ = $\frac{A/[A+C]}{B/[B+D]}$ or sensitivity/[1-specificity]
LR- = $\frac{C/[A+C]}{D/[B+D]}$ or [1-sensitivity]/specificity]

Sensitivity is defined as the ability of the test to consistently classify children with SLI as being affected with SLI. Specificity is defined as the ability of the test to consistently classify children without SLI as being unaffected with SLI (Redmond et al., 2011). Even though there are no generally accepted thresholds for sensitivity and specificity (Archibald & Joanisse, 2009), Plante and Vance (1994) suggested that the minimum value for both measures should be above 80%. The measures of sensitivity and specificity are highly dependent on the prevalence of the

disorder in the population and show a trade-off with higher levels of sensitivity leading to lower levels of specificity. LR+ and LR-, on the other hand, are independent from the prevalence of a disorder in the population (Archibald & Joanisse, 2009). LR+ reflects the ratio between the probability that a positive test score comes from a person with the disorder to the probability of a positive test score from a person without the disorder. LR+ values above 10 are desirable (Dollaghan, 2007). Redmond et al. (2011) further classified LR+ values into the following categories: LR+ = 1 neutral, LR+ = 3 moderately positive, LR+ ≥ 10 very positive. LR- reflects the ratio between the probability that a negative test score comes from a person with the disorder to the probability of a negative test score from a person without the disorder. LR- values < 0.10 are desirable (Dollaghan, 2007). According to Redmond et al. (2011) LR- = 1 neutral, LR- = 0.30 moderately negative LR- ≤ 0.10 extremely negative. A graphic way to examine diagnostic accuracy levels is through the use of Receiver Operating Characteristics (ROC) graphs. An example of a ROC graph is presented in Figure 2.1.

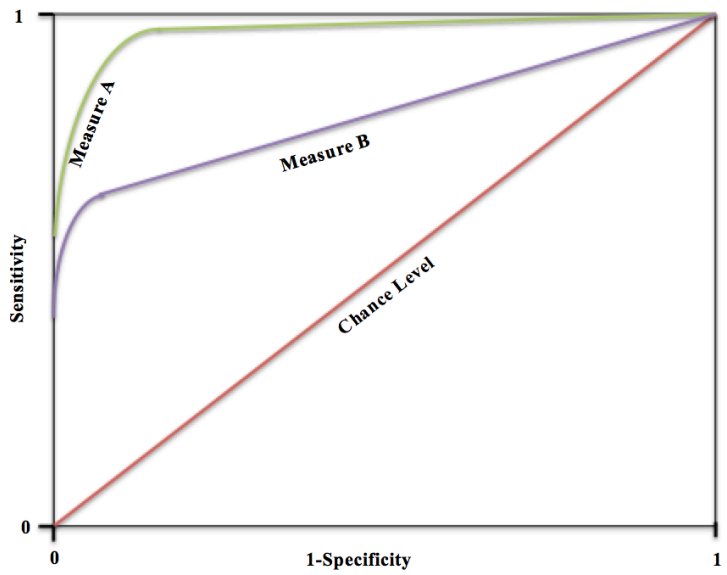


Figure 2.1 Example of ROC curve

A ROC curve is the graphic representation of the trade-off between sensitivity (y-axis) and specificity (x-axis). The ROC curve of a measure falling below the diagonal red line represents a measure that is performing at chance level and the rate of true positive is equal to the false positive rate. The upper left corner of Figure 2.1 (0,1) corresponds to a perfect classification accuracy. The higher the area under the ROC curve, the higher the diagnostic accuracy of a measure. In Figure 2.1, Measure A obtained higher diagnostic accuracy values compared to Measure B. It can also establish optimal cut-off points.

There are three widely considered clinical markers for English-speaking children with SLI. One is nonword repetition: children with SLI find it difficult to repeat fake/made-up words (Bishop, North, & Donlan, 1996; Dollaghan & Campbell, 1998). Verb tense is another area of weakness for English-speaking children with SLI (Rice & Wexler, 1996), children with SLI tend to omit the past tense suffix "–ed" and finally Sentence Repetition (Archibald & Joanisse, 2009; Conti-Ramsden et al., 2001). Studies have not been limited to English but also extended to other languages such as Cantonese (Stokes et al., 2006) and French (Leclercq et al., 2014; Thordardottir et al., 2011).

Relatively few published studies have reported diagnostic accuracy levels when investigating the potential use of Sentence Repetition as a marker for SLI. Most studies focused on English-speaking participants (Archibald & Joanisse, 2009; Botting & Conti-Ramsden, 2003; Conti-Ramsden et al., 2001; Everitt et al., 2013; Poll et al., 2010; Redmond et al., 2011). Even fewer studies focused on participants whose first language is not English (Leclercq et al., 2014; Stokes et al., 2006; Thordardottir et al., 2011).

With regards to English speaking participants, the studies cover a wide age range starting from 7-year-old children with SLI to 26-year-old adults with SLI. One study extends the lower limit to children as young as 3 years (Everitt et al., 2013). In this case, poor Sentence Repetition performance was labelled as a risk rather than clinical marker because children were at risk of or potentially had SLI but were too young to be diagnosed with the condition. This wide age range allows for an examination of the stability of Sentence Repetition as a marker of SLI in English.

Studies have also investigated whether the poor performance on Sentence Repetition tests was limited to children with SLI or extended to children with other developmental disorders such as autism, PLI, and ADHD (Botting & Conti-Ramsden, 2003; Redmond et al., 2011). Its utility as a screening tool has been examined as well (Archibald & Joanisse, 2009). No studies were population based, with most of the studies relying on participants who have been referred to clinical services or language units.

Concerning participants whose first language was not English, two languages have been examined: Cantonese (Stokes et al., 2006) and French (Leclercq et al., 2014; Thordardottir et al., 2011). In contrast to English, Cantonese is a tonal language with sparse morphology, while French is a morphologically rich language. In spite of few studies, these provide insight into the stability of Sentence Repetition as a marker of SLI across typologically different languages.

The use of Sentence Repetition as a marker of SLI in English will be presented first followed by its potential use as a marker in other languages. Within each section, the studies will be looked at from two angles: the diagnostic accuracy levels of Sentence Repetition tests in English

and other languages, as well as how these studies inform us about the underlying processes involved in Sentence Repetition.

2.1.5 Sentence Repetition as a marker of Specific Language Impairment in English

Two Sentence Repetition tests have been used in English marker studies: the Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals-Revised (CELF-R; Semel et al., 1994) and the Redmond (2005) test. The two tests differ with regards to test type, structure, procedure, and scoring. The Recalling Sentences subtest of the CELF-R, along with its subsequent versions presented below, is part of a standardized diagnostic assessment battery. It commences with grammatically simple short sentences and gradually increases in grammatical complexity and length simultaneously, covering a wide range of syntactic structures. It is presented live, in a fixed order and is discontinued after three consecutive no responses. Each sentence is scored based on the number of deviations from the target sentence: 3 for no deviations, 2 for a single deviation, 1 for two or three deviations, and 0 for four or more deviations. The Redmond (2005) test, on the other hand, is a non-standardized standalone research probe. It consists of 16 sentences with an equal mix of passive and active sentences. The sentences are equal in length consisting of 10 words and 10 to 14 syllables. All the target sentences are presented with no discontinue rule. The mode of presentation differed according to the study: it was presented via digital audio recording in the Archibald and Joanisse (2009) study, and live in the Redmond et al. (2011) study. Scoring, as with the CELF-R, is purely quantitative and error based but with a reduced range of scores for each sentence, commencing at a score of 2 for correct sentence, 1 for three errors or fewer, and 0 for four or more errors.

Conti-Ramsden et al. (2001) were the first to investigate diagnostic accuracy levels of a Sentence Repetition test. They compared the performance of 160 11-year-old participants with a documented history of SLI at age 7 years to 100 age-matched controls on four potential markers: third person singular, past tense marking, nonword repetition, and Sentence Repetition. Details of the study showing sample size, population, recruitment criteria, diagnostic accuracy levels, and the main findings can be found in Table 2.2.

The sensitivity and specificity of the individual marker tasks were calculated using different thresholds to predict group membership. Children scoring at or below a threshold were classified as impaired while children scoring above the threshold were classified as non-impaired. Three cut-off scores were examined that corresponded to the 2.5th, 10th, and 16th centiles. The Sentence Repetition test showed the best combination of sensitivity (90%) and specificity (85%) at the 16th centile compared to the other three markers investigated. Combining the marker tasks did not improve diagnostic accuracy levels. The ROC curve analysis further confirmed this finding by showing that the Sentence Repetition test covered the largest area (.92). Furthermore, the Sentence

Repetition test was also able to identify 60% of children who had a history of SLI at age 7 years but performed within 1 SD of the population mean on all language assessments at age 11 years. This is of particular relevance to genetic marker studies, where it is important to identify individuals who may have developed compensatory strategies and perform within the normal range on language assessments (Bishop, 2014).

In a follow-up study, Hesketh and Conti-Ramsden (2013) investigated more closely how receptive language ability and nonverbal IQ levels influenced the performance of the same group of children with a history of SLI on a Sentence Repetition test. The sample was subdivided into four groups: (1) children with SLI (n = 32) who obtained low scores on the receptive language test but whose nonverbal IQ levels were within norms; (2) children with non-specific language impairment with low scores on receptive language and nonverbal IQ (n = 56); (3) children with resolved receptive language and low nonverbal IQ scores (n = 34); and (4) children with resolved receptive language and nonverbal IQ levels within the norms (n = 75). The performance of the four groups on the CELF-R were compared to age-matched controls and revealed that all four groups performed significantly lower than controls. Irrespective of nonverbal IQ levels, children with language impairment obtained the lowest scores followed by children with resolved receptive language. Even children with resolved receptive language scored more than 1 SD below mean on the Sentence Repetition test, further confirming its utility as a marker.

Poll et al. (2010) extended the findings of Conti-Ramsden et al. (2001) to adults with SLI. The performance of 13 adult participants with SLI aged 18;0 to 25;11 was compared to 18 age-matched controls on three potential clinical markers: grammaticality judgment, nonword repetition, and the Recalling Sentences subtest of the CELF-3 (Semel et al., 1994; see Table 2.3 for details). The Test of Adolescent and Adult Language (TOAL-3) Spoken Language Quotient (SLQ; Hammill, Brown, Larsen, & Wiederholt, 1994) was used to classify participants. One of the four subtests that counted towards the SLQ was a Sentence Repetition subtest. To ensure that poor scores on the Recalling Sentences subtest of the CELF-3 was not a by-product of the classification process, participants were reclassified without the use of the TOAL-3 Sentence Repetition test. Reclassification resulted in two distinct groups: one group met the criteria for SLI while the other failed to without the TOAL-3 Sentence Repetition subtest. A comparison between the scores of the CELF-3 Recalling Sentences subtest of two groups did not show a significant difference; this may be due to the difference in scoring between both tests. The TOAL-3 Sentence Repetition subtest employs a purely quantitative scoring method but at a vastly reduced range with a single error resulting in a score of 0 while an exact repetition is awarded a score of 1. Rather than relying on arbitrary cut-off points, Poll et al. (2010) utilized ROC curves to identify optimal cut-off points that maximized classification accuracy for each marker task. Sentence Repetition was the best

individual marker with a sensitivity of 85% and specificity of 89%. A combination of the marker tasks improved diagnostic accuracy levels but did not differ greatly from the Sentence Repetition test with regards to overall classification accuracy 90% compared to 87% for the Sentence Repetition test alone.

Everitt et al. (2013) extended the findings of Conti-Ramsden et al. (2001) to younger children. The study's strength lies in its use of a longitudinal rather than cross sectional study design. At baseline, the performance of 47 children with Specific Expressive Language Delay (SELD) aged 3;0-4;0 was compared to 47 age and sex matched controls on 5 markers: Sentence Repetition, word repetition, nonword repetition, digit recall and third person singular. At follow-up, the performance of 37 children with SELD aged 4;0-5;0 and 54 controls on the same 5 markers at baseline along with an additional nonword repetition test and past tense measure. Details of the study at baseline and follow-up can be found in Table 2.4 and Table 2.5, respectively. Diagnostic accuracy values for the Sentence Repetition test and digit recall were obtained for raw and standard scores. This is due to slight modifications in test administration to increase child participation (modification to CELF-P previously discussed). Not all cut-off points investigated were included in the tables; only the most useful cut-off score for each marker was included. Both linguistic markers, third person singular and past tense, posed some challenges in administration and scoring. A pilot study revealed that children with SELD between ages 3;0-4;0 had difficulty engaging with the past tense measure; hence, it was not administered at baseline. Some children in SELD group at both baseline and follow-up were unable to consistently produce /s/ and/or /z/ in the case of third person singular or make consistent substitutions to mark tense in word final positions, which made their responses difficult to score. At baseline, the Sentence Repetition (standard score) at the 16th centile was the best group discriminator with a sensitivity of 89% and a specificity of 79%, with the largest area under the ROC curve (.92). At follow-up, again Sentence Repetition (standard) score was the best marker at the 16th centile a sensitivity of 95% and a specificity of 81%, with the largest area under the ROC curve (.94). In addition, Everitt et al. (2013) found that the Sentence Repetition (standard score) at baseline was the only marker task that significantly predicted persistent expressive language delay at follow-up, supporting the predictive validity of the test.

Table 2.2 *Summary of Conti-Ramsden et al. (2001)*

Sample, Age (mean)	**Participant Recruitment**	**Markers**	**Sensitivity (%)**			**Specificity (%)**			**Findings**
			2.5th	**10th**	**16th**	**2.5th**	**10th**	**16th**	
160 SLI	SLI:	Past Tense[a]	33	89	74	100	93	89	Sentence
100 AM	Recruited at age 7 and attended Year 2	Third Person Singular[b]	21	52	63	100	93	90	Repetition was
(10;9)	mainstream language units across England.	Sentence Repetition[c]	54	86	**90**	99	92	**85**	the most useful
	Exclusion criteria:	Nonword Repetition[d]	42	74	78	98	92	87	clinical marker
	1. Hearing loss.	Combination	**16th**			**16th**			at the 16th
	2. Major physical disability. 3. Diagnosis of autism or moderate learning		Both tests low	Either test low		Both tests low	Either test low		centile. Combination of
	difficulties. 4. Global delay at age 7: nonverbal IQ > 2 *SD* below population mean.	Third Person Singular Past Tense	54	82		97	82		markers did not improve diagnostic
		Third Person Singular Sentence Repetition	60	93		97	78		accuracy.
	Language ability: History of Language Impairment at age 7 with no specific criteria regarding level of performance on language measures at age 11.	Third Person Singular Nonword Repetition	55	86		96	81		
	AM: Recruited from 3 primary schools in rural & urban settings.	Past Tense Sentence Repetition	71	93		96	78		
	Exclusion criteria: 1. History of special needs education.	Past Tense Nonword Repetition	61	92		96	80		
	Broad language and nonverbal IQ measures were not administered.	Sentence Repetition Nonword Repetition	73	96		94	78		

Note. SLI = specific language impairment; AM = chronological age-matched controls; IQ = intelligence quotient.
[a] Past Tense Task (Marchman, Wulfeck, & Weismer, 1999) [b] Third Person Singular task (Simkin & Conti-Ramsden, 2001) [c]CELF-R Recalling Sentences subtest (Semel et al., 1994) [d] The Children's Test of Nonword repetition (Gathercole & Baddeley, 1990)
Only raw scores were used for all marker task.

Table 2.3 *Summary of Poll et al. (2010)*

Sample, Age (mean)	Participant Recruitment	Markers	Cut-off[d]	Sensitivity	Specificity	LR+	LR-	Findings
13 SLI 18 AM 18;00-25;11 (21)	All Participants: Current or recent students at a vocational post-secondary school in Pennsylvania. 1. Standard PIQ ≥ 80 on 3 subtests of WAIS-III: Picture completion Block design Digit-symbol coding. 2. Passed a hearing screening. SLI: If participant met one of following: 1. Expressive Language: 1 SD ≤ mean SLQ of TOAL-3 or 2 SD ≤ mean on a single subtest. 2. Receptive Language: 1 SD ≤ mean on the PPVT-R. AM: If participant met all of following criteria: 1. Expressive Language: 1 SD > mean SLQ of TOAL-3 & 2 SD > mean on all subtests 2. Receptive Language 1 SD > mean PPVT-R No self-reported history of language therapy, special education, or a known diagnosis of language or cognitive disorder.	Nonword Repetition[a]: (PPC)						Sentence Repetition was the best individual marker. A combination of the three marker tasks showed an improvement in diagnostic accuracy levels.
		Total	92	.615	.778	2.77	.495	
		3 syllable	98	.846	.556	1.91	.277	
		4 syllable	84	.692	.667	2.08	.462	
		Sentence Repetition[b] (raw score)	**62.5**	**.846**	**.889**	**7.62**	**.173**	
		Grammaticality Judgment[c]: (A`)[d]						
		Simple omitted finiteness	.90	.231	1.0	∞	.769	
		Complex omitted finiteness	.95	.538	.944	9.61	.489	
		Complex bad agreement	.95	.462	.833	2.68	.646	
		Complex missing progressive	.90	.077	1.0	∞	.923	
		Combination: Sentence Repetition, Complex omitted finiteness & 3 syllable PPC	**.44**	**.92**	**.89**	**8.3**	**.09**	

Note. SLI = specific language impairment; AM = chronological age-matched controls; PIQ = performance intelligence quotient; WAIS-III = Wechsler Adult Intelligence Scale-III (Wechsler, 1997); SLQ = spoken language quotient; TOAL-3 = Test of Adolescent and Adult Language; PPVT-R = Peabody Picture Vocabulary Test-Revised (Dunn & Dunn, 1981); PPC = percentage of phonemes correct; LR+ = positive likelihood ratio, LR- = negative likelihood ratio.
[a] Nonword Repetition task (Dollaghan & Campbell, 1998) [b]CELF-3 Recalling Sentences subtest [c] novel experimental task [d] Scored using A`(equivalent to percent correct for an unbiased forced choice task, ranges from .50 to 1.00, where .50 = random chance 1.00= perfect discrimination of well-formed vs. Anomalous sentences) [e] Cut-off that maximized classification accuracy using ROC.

Table 2.4 *Summary of Baseline from Everitt et al. (2013)*

Sample, Age (M)	Participant Recruitment	Markers	Sensitivity (%)		Specificity (%)		LR+		LR-		ROC Curve*	Findings
			16th	25th	16th	25th	16th	25th	16th	25th		
3;0-4;0	All children:	Sentence Repetition[a]										Sentence Repetition was the best individual marker at the 16th centile.
47 SELD (3;58)	All state nurseries (n = 58) and family centres (with children age 3 & above; n = 7) located in Aberdeen, Scotland were contacted. 18 nurseries and 1 family centre participated.	(raw score)	81	91	85	74	5.43	3.58	.23	.11	.92	
		(standard score)	**89**	96	**79**	64	**4.20**	2.65	**.14**	.07	**.92**	
47 AM (3;57)	1. Teacher identified.	Word Repetition[b] (% correct)	66	83	85	64	4.43	2.29	.40	.27	.81	
	2. No history of: hearing, oro-motor, behavioural or neurological difficulties.	Nonword Repetition[b] (% correct)	62	77	79	64	2.90	2.12	.49	.37	.77	
	3. Nonverbal IQ ≥ 80 on LIPS-R.	Digit Recall[c]										
	4. Monolingual, English 1st language and not from a multiple birth.	(raw score)	62	66	76	70	2.64	2.21	.50	.48	.76	
	5. EC on PLS-3: SELD > 1 *SD* below population mean; AM ≤ 1 *SD*	(percentile score)	55	62	85	74	3.71	2.42	.53	.51	.80	
		Third Person Singular[d] (% correct)	71	82	81	75	3.71	3.20	.36	.25	.82	
	SELD:											
	1. No more than 6 sessions of therapy.											
	2. SCQ < 15 Lifetime score.											

Note. SELD = specific expressive language delay; AM = chronological age and gender matched controls; IQ = intelligence quotient; LIPS-R = Leiter International Performance Scale-Revised (brief screener; (Roid & Miller, 1995, 1997); EC = expressive communication subscale; PLS-3 = preschool language scale-3 UK; SCQ = social communication questioner (Rutter, Bailey, & Lord, 2003).

[a] Recalling Sentences subtest of the CELF-Preschool UK (Wiig et al., 2000) [b] Preschool Repetition test (Seeff-Gabriel et al., 2008) [c] Recall of Digits Forward subtest of the British Ability Scales II (Elliott, 1997) [d] Test of Early Grammatical Impairment (Rice & Wexler, 2001)

* all significant at $p < .001$

Table 2.5 *Summary of Follow-up from Everitt et al. (2013)*

Sample, Age (M)	Classification	Markers	Sensitivity (%)		Specificity (%)		LR+		LR-		ROC Curve*	Findings
			16th	25th	16th	25th	16th	25th	16th	25th		
4;0 – 5;0	EC on PLS-3:	Sentence Repetition[a]										Sentence Repetition was the best individual marker at the 16th centile.
		(raw score)	81	92	85	74	5.47	3.54	.22	.11	.91	
SELD:	SELD at follow-up	(standard score)	**95**	100	**81**	72	**5.11**	n/a	**.07**	n/a	**.94**	
35/47 PELD	> 1 SD below	Word Repetition[b]										
11 Recovered	population mean	(% correct)	59	68	85	70	3.94	2.24	.48	.46	.73	
1 Dropout		Nonword Repetition[b]										
	PELD at baseline	(% correct)	51	73	79	68	2.47	2.28	.61	.40	.74	
AM:	and follow-up	Digit Recall[c]										
43/47	> 1 SD	(raw score)	41	49	83	76	2.43	2.02	.71	.68	.66	
2 SELD		(percentile score)	30	51	85	74	2.01	1.98	.82	.66	.71	
2 Dropout	Recovered	Third Person Singular[d]										
	baseline > 1 SD	(% correct)	47	66	85	76	3.11	2.68	.63	.46	.81	
Totals:	follow-up ≤ 1 SD	Past Tense[d] (% correct)										
37 SELD		Composite	61	73	76	74	2.47	2.75	.52	.37	.82	
(4;67)	AM ≤ 1 SD	Regular	73	76	85	74	4.82	2.87	.32	.33	.84	
		Irregular	64	64	62	62	1.69	1.69	.58	.58	.62	
54 AM		Irregular Past Finite	49	61	85	72	3.21	2.14	.61	.55	.76	
(4;57)		CNRep[e]										
		(standard score)	73	86	85	72	4.93	3.11	.32	.19	.87	

Note: n/a = not available; SELD = specific expressive language delay; AM = chronological age and gender matched controls; PELD = persistent expressive language delay; EC = expressive communication subscale; PLS-3 = preschool language scale-3 UK.

[a] Recalling Sentences subtest of the CELF-Preschool UK (Wiig et al., 2000) [b] Preschool Repetition test (Seeff-Gabriel et al., 2008) [c] Recall of Digits Forward subtest of the British Ability Scales II (Elliott, 1997) [d] Test of Early Grammatical Impairment (Rice & Wexler, 2001) [e] Children's Test of Nonword Repetition (Gathercole & Baddeley, 1990)

* all significant at $p < .05$ except irregular past tense

Like above studies, Archibald and Joanisse (2009) examined the utility of Sentence Repetition and nonword repetition as clinical markers of SLI. In addition, they were interested in whether the two marker tasks were able to identify children with Working Memory Impairment (WMI) and whether it occurred in the absence of Language Impairment in poor repeaters. Finally, whether sentence and nonword repetition were similar or better clinical markers for Language and/or Working Memory Impairments. In order to achieve the above aims, Redmond's (2005) Sentence Repetition test and a nonword repetition test were administered first to 400 children aged 5;0-9;0 followed by a detailed assessment of language and working memory skills in poor repeaters and a subset of average scores (see Table 2.6). The order of administration and the examination of Working Memory ability were unique to this study. The nonverbal IQ, Verbal Short Term Memory, and Visuo-spatial Short Term Memory abilities were assessed but did not influence inclusion or exclusion criteria. The Language Impaired group encompassed three profiles of language and memory skills: SLI, Language Impairment with WMI, and Language Impairment with unclassified WMI. All three subgroups were characterized by an equal deficit in language accompanied by a Verbal Short Term Memory Deficit. Diagnostic accuracy results (see Table 2.7) showed that the Sentence Repetition test was the best marker for Language Impairment at the 10th centile with a sensitivity of 84.6%, a specificity of 90.3%, and LR+ value of 8.7 (LR- was not reported). A combination of the two marker tasks reduced diagnostic accuracy levels; this could be because item rather than phoneme level scoring was used in the nonword repetition test.

So far, the focus of this review has been on the use of Sentence Repetition as a clinical marker solely for SLI across different age groups. Botting and Conti-Ramsden (2003) were the first to address the question of whether Sentence Repetition was a unique marker for SLI or a marker of general language impairment irrespective of underlying cause/developmental disorder. They also addressed whether Sentence Repetition as a marker was capable of differentiating between children with different communication disorders. They compared the performance of four groups of 11-year-old participants—(1) 29 children with a history of SLI at age 7, (2) 25 children with Pragmatic Language Impairment (PLI), (3) 13 children with Autism Spectrum Disorders (ASD), and (4) 100 age-matched controls—on three marker tasks: Sentence Repetition, past tense morphology, and nonword repetition (see Table 2.8 for details). Testing revealed that the PLI group encompassed two distinct groups: children with PLI Pure were characterized by a severe pragmatic language impairment coupled with linguistic difficulties (poor Sentence Repetition and Past Tense scores) in the absence of autistic traits. While the PLI Plus group were characterized by pragmatic language difficulties with some autistic traits and without any language difficulties. Therefore, diagnostic accuracy levels for the two PLI groups were reported separately.

Results revealed that the Sentence Repetition test showed the best combination of sensitivity and specificity at 10th centile for children with a history of SLI (90%, 92%) and at the 16th centile for children in the PLI Pure (93%, 85%) and ASD groups (85%,85%), respectively. Since the common denominator between the three groups was the presence of language impairment, this supports the use of Sentence Repetition as a marker for not only SLI, but other communication impairments as well. None of the markers investigated showed acceptable levels of sensitivity and specificity for the PLI Plus group. With regard to differential diagnosis, the PLI Pure group could not be distinguished from the SLI and ASD groups.

Redmond et al. (2011) examined the diagnostic accuracy of four marker tasks— Sentence Repetition, nonword repetition, tense morphology, and narratives—in differentiating between 20 children with SLI aged 7;0-8;0 and 20 age-matched controls and differentiating between children with SLI and 20 children with ADHD (see Table 2.9). Group comparisons of the four marker tasks showed the same pattern of results: children with SLI consistently obtained the lowest scores while children in the ADHD group performed similarly to controls. The distribution of scores on Sentence Repetition test showed no overlap between the SLI group and the other two groups with the exception of one SLI participant. Group comparison of Sentence Repetition scores also showed the largest effect size (Eta Squared $\eta^2 = .617$), 1.5-2 times larger than the other three markers.

The Sentence Repetition test was the only marker that maintained high diagnostic accuracy levels in both discriminating participants with SLI from controls and participants with SLI from those with ADHD. For diagnosis and differential diagnosis, sensitivity and specificity were above the recommended 80% (Plante & Vance, 1994) at the optimal cut-off points identified for each using ROC curves. ROC curve areas were also the largest for Sentence Repetition in SLI versus controls discrimination (.959) and SLI versus ADHD (.963). The LR+ value for SLI versus controls equalled 9, which is close to the desirable 10 recommended by Dollaghan (2007), and surpassed it in SLI versus ADHD equalling 18. The authors attributed the higher LR+ value in the SLI versus ADHD discrimination to the fact that fewer cases in the ADHD group fell below the cut-off point compared to controls. Finally, LR- equalled .111 for SLI versus controls and .105 for SLI versus ADHD, both which were extremely close to the recommended .1 (Dollaghan, 2007).

The findings of this study fell in line with the Archibald and Joanisse (2009) study with regard to the high diagnostic accuracy levels obtained for Sentence Repetition in discriminating between children with language impairment and controls. The differential diagnosis results for sensitivity were similar to the Botting and Conti-Ramsden (2007) study. Although the PLI Plus and ADHD groups differed in the underlying developmental disorder, both groups did not exhibit language difficulties according to the reference standard administered.

Table 2.6 *Summary of Archibald and Joanisse (2009)*

Participant Recruitment	Markers	Resulting Profiles	Sample	Classification Criteria	Classification Profiles
All children attending senior kindergarten to Grade 3 in 9 schools in Southwest region of Ontario Canada were invited to participate. 1,255 consent forms were distributed; 412 responses with 12 excluded (absent on day of testing or too young) 400 participants 5;03-9;42	Nonword Repetition[a] Sentence Repetition[b]	Low scorers: scored <15th percentile for their age on both screening tasks or <10th percentile on one of the screening tasks. *n* = 52. Average scorers: scored >35th percentile for their age on both tasks. *n* = 197 150 fell between their arbitrary cut-off scores presented above and were not investigated any further.	52 low scorers 36 average scorers 36 out of 197 average scorers were selected with the following constraints: 5 boys 5 girls from same grade level and school as low scorers. Classification assessments were administered to all low scorers and only a subset of average scorers due to economic and time constraints.	LI: composite language score > 1 *SD* below population mean on CELF-4. WMI: both verbal and visuo-spatial working memory composites > 1 *SD* below population mean on AWMA. Unclassified WMI: > 1 *SD* below population mean on either the verbal or visuo-spatial working memory composite.	26 children with LI: 7 LI only 12 LI and WMI 7 LI and unclassified WMI (26 low, 0 average) 7 with WMI only (4 low, 3 average) 18 unclassified WMI (7 low, 11 average) 36 children with no impairments in language or working memory (14 low, 22 average)

Note: LI = language impairment; WMI = working memory impairment; CELF-4 = Clinical Evaluation of Language Fundamentals 4th edition (Semel, Wiig, & Secord, 2003); AWMA = Automated Working Memory Assessment (Alloway, 2007); IQ = intelligence quotient;
[a]Nonword Repetition Test (Dollaghan & Campbell, 1998) scored at item- rather than phoneme-level correct [b]Sentence Repetition task (Redmond, 2005)

Table 2.7 *Summary of Results from Archibald and Joanisse (2009)*

Impairment	Subgroups	Markers	Sensitivity (%)			Specificity (%)			LR+			Findings
			10th	**15th**	**Both 10th or One 15th**	**10th**	**15th**	**Both 10th or One 15th**	**10th**	**15th**	**Both 10th or One 15th**	
Language	SLI	Nonword Repetition	19.2	46.1		66.1	62.9		0.6	1.2		The best clinical marker was sentence repetition at the 10th percentile with levels of sensitivity & specificity above 80% and highest positive likelihood ratio
	LI + WMI	Sentence Repetition	**84.6**	96.2		**90.3**	75.8		**8.7**	4.0		
	LI + unclass. WMI	Combination			100			58.1			2.4	
Working Memory	LI + WMI	Nonword Repetition	23.1	47.4		67.7	62.3		0.7	1.3		
	LI + unclass. WMI	Sentence Repetition	57.9	84.2		75.4	66.7		2.3	2.5		
	SWMI	Combination			84.2			47.8			1.6	
Combined Language & Working Memory	None	Nonword Repetition	16.7	41.7		68.4	60.5		0.5	1.1		
		Sentence Repetition	83.3	100		76.3	64.5		3.5	2.8		
		Combination			100			47.4			1.9	
Any Memory	LI + WMI	Nonword Repetition	22.7	40.1		63.6	61.4		0.6	1.1		
	LI + unclass. WMI	Sentence Repetition	45.5	59.1		77.8	70.5		2.0	2.0		
	SWMI	Combination			68.2			50			1.4	
	Unclass. WMI											

Note. Negative Likelihood Ratio was not reported; SLI = specific language impairment; LI = language impairment; WMI = working memory impairment; SWMI = specific working memory impairment; unclass = unclassified, LR+ = positive likelihood ratio.

Table 2.8 *Summary of Botting and Conti-Ramsden (2003)*

Sample, Age (M)	Participant Recruitment	Markers	Sensitivity[d] (%)			Specificity[d] (%)			Findings
			2.5th	10th	16th	2.5th	10th	16th	
29 SLI 10;2-11;9 (10;10) 25 PLI 10;2-12;5 (11;3) PLI split into two sub-groups: 11 PLI Plus 14 PLI Pure 13 ASD 10;2-12;6 (10;10) 100 AM 10;5-11;6 (11)	SLI, PLI ASD: Recruited at age 7 and attended Year 2 mainstream language units across England. 6/25 PLI and 5/13 ASD: recruited via local specialist schools for children with communication difficulties or autism. 1. All clinical groups: PIQ ≥ 70 WISC-III SLI: 1. History of language impairment at age 7. 2. EVT scores < the 10th centile. 3. TROG scores< the 50th centile. 4. No current status of pragmatic impairment: score > 132 on CCC. PLI Pure: 1. Pragmatic scale score < 132 on CCC. 2. Non-autistic: CARS score <30. PLI Plus: 1. Above criteria and restricted interest (24< Interest subtest CCC) or social impairment (28< Social Relationships subtest CCC). ASD: 1. CARS score ≥ 30 or clinical diagnosis of Autism. AM: Recruited from 3 primary schools in rural & urban settings.	**Past Tense[a]**				100	93	89	Sentence Repetition was the best marker for the SLI, PLI Pure and ASD groups. None of the markers were discriminating for the PLI Plus group. For differential diagnosis, Sentence Repetition showed the highest sensitivity but with reduced specificity values in distinguishing PLI Plus from SLI, PLI Pure and ASD and the SLI group from ASD.
		SLI	39	89	89				
		PLI Pure	57	86	86				
		ASD	39	73	73				
		PLI Plus	10	30	30				
		Sentence Repetition[b]				99	**92**	**85**	
		SLI	72	**90**	90				
		PLI Pure	43	79	**93**				
		ASD	39	69	**85**				
		PLI Plus	20	40	50				
		Nonword Repetition[c]				98	92	87	
		SLI	62	79	79				
		PLI Pure	29	43	64				
		ASD	15	54	62				
		PLI Plus	20	20	50				
		Past Tense[a]							
		PLI plus vs ASD		73			70		
		PLI Plus vs PLI Pure	N/A	86	N/A	N/A	70	N/A	
		PLI vs Plus SLI		89			70		
		SLI vs ASD		89			27		
		Sentence Repetition[b]							
		PLI plus vs ASD		**85**	N/A		50	N/A	
		PLI Plus vs PLI Pure	N/A	**93**		N/A	50		
		PLI Plus vs SLI		N/A	**90**		N/A	60	
		SLI vs ASD			**90**			30	
		Nonword Repetition[c]							
		SLI vs PLI Plus	N/A	79	N/A	N/A	80	N/A	

Note. SLI = specific language impairment; PLI = pragmatic language impairment; ASD = autism spectrum disorder; AM = chronologically age-matched controls; PIQ = performance intelligence quotient; WISC-III = Wechsler Intelligence Scale for Children 3rd ed, short form (Wechsler, 1992); EVT = Expressive Vocabulary Test

(Williams, 1997); TROG = Test for Reception of Grammar (Bishop, 1982); CCC = Children's Communication Checklist (Bishop, 1998); CARS = Childhood Autism Rating Scales (Schopler, Reichler, DeVellis, & Daly, 1980).
[a] Past Tense Task (Marchman et al., 1999) [b] CELF-R Recalling Sentences subtest (Semel et al., 1994) [c]The Children's Test of Nonword repetition (Gathercole & Baddeley, 1990) [d] For differential diagnosis between communication impairments, sensitivity & specificity values were only reported if the accuracy value was ≥70 and only for the best cut-off

Table 2.9 *Summary of Redmond et al. (2011)*

Sample, Age (M)	Participant Recruitment	Markers		Area Under Curve	Cut-off[b]	Sensi-tivity	Speci-ficity	LR+	LR-	Findings
7-8 20 SLI (7.85) 20 ADHD (7.86) 20 AM (7.83)	All: 1. Monolingual speakers of Standard American English 2. Passed hearing & phonological screener of TEGI 3. Nonverbal IQ ≥ 80 on NNAT–I 4. No diagnosis of Autism or PDD SLI: Speech and language therapist caseloads from two school districts and two university clinics in Utah. 1. Diagnosed as SLI and receiving therapy. 2. Score at or below -1 *SD* on CELFST-4. ADHD: Utah chapter of Children and Adults with ADHD and caseloads of clinical psychologists in same area. 1. Diagnosed with combined type ADHD and receiving treatment. 2. Score > 64 on CBCL DSM-ADHD. 3. No diagnosis of Language Impairment. AM: Same school districts as clinical group. Public notices in community bulletins 1. Score above -1 *SD* on CELFST-4. 2. Score ≤ 64 on CBCL DSM-ADHD.	TEGI[a] PC	SLI vs. AM	.954*	95.75	.84	.90	16.80	0.168	The best individual marker for both identification and discrimination was the Sentence Repetition test.
			SLI vs. ADHD	.900*	93.70	.79	.85	5.27	0.247	
		NWR PPC	SLI vs. AM	.924*	85.91	.95	.90	9.50	0.056	
			SLI vs. ADHD	.875*	84.90	.90	.70	3.00	0.143	
		SR raw score	SLI vs. AM	**.959***	**14.50**	**.90**	**.90**	**9.00**	**0.111**	
			SLI vs. ADHD	**.963***	**15.50**	**.90**	**.95**	**18.00**	**0.105**	
		TNL composite standard score	SLI vs. AM	.936*	95.50	.95	.80	4.75	0.063	
			SLI vs. ADHD	.882*	95.50	.95	.65	2.71	0.077	

Note. SLI = Specific Language Impairment; AM = chronologically age-matched controls; CELFST-4 = Clinical Evaluation of Language Fundamentals Screening Test—Fourth Edition (Semel, Wiig, & Secord, 2004); CBCL DSM-ADHD = Child Behavior Checklist Diagnostic Statistics Manual ADHD subscale (Achenbach & Rescorla, 2001); TEGI = Test of Early Grammatical Impairment (Rice & Wexler, 2001); NNAT-I = Naglieri Nonverbal Achievement Test-Individual (Naglieri, 2003); PDD = Pervasive Developmental Disorders; PC = Percent Correct; NWR = Nonword Recall (Dollaghan & Campbell, 1998); PPC = Phoneme Percent Correct; SR = Sentence Recall (Redmond, 2005); TNL = Test of Narrative Language (Gilliam & Pearson, 2004).
[a] The regular third-person and past-tense probes [b] Optimal cut-off as established by ROC curve * p <.001

2.1.6 Conclusion

The above studies have shown that Sentence Repetition is a stable marker of SLI across a wide age range. Sentence Repetition was found to be a potential risk marker of SELD in children as young as 3 years of age (Everitt et al., 2013) and a potential clinical marker of SLI for children as young as 7 years to adults up to 26 years. Without exception, all studies found that Sentence Repetition obtained the highest diagnostic accuracy measures in comparison to other marker tasks. The combination of marker tasks did not yield significant increases in diagnostic accuracy levels. Sentence Repetition was also found to be a "Universal Marker" (Botting & Conti-Ramsden, 2003, p. 523) of language impairment irrespective of underlying cause.

Sensitivity and specificity values for Sentence Repetition exceeded the 80% threshold identified by Plante and Vance (1994), ranging from 81%-95% for sensitivity and 81%-92% for specificity. This indicates a good degree of agreement between how the reference standard test classified children as having SLI or not and Sentence Repetition. The most common cut-off scores employed were the 10th and 16th centile, which corresponds to approximately 1.25 SD and 1 SD away from the mean. In two of six studies, ROC analysis had been used to determine optimal cut-off scores (Poll et al., 2010; Redmond et al., 2011). In three of six studies (Conti-Ramsden et al., 2001; Everitt et al., 2013; Redmond et al., 2011), the area under the ROC curve was measured and was the largest for Sentence Repetition in comparison to other markers; the area ranged from .922 to .959, all close the perfect classification accuracy of 1.

Not all studies calculated LR+ and LR- values for marker tasks. Four of six studies reported LR+ values (Archibald & Joanisse, 2009; Everitt et al., 2013; Poll et al., 2010; Redmond et al., 2011). All were above the moderately positive benchmark of 3 (Redmond et al, 2011) and ranged from 5.11 to 9; the upper range was near the extremely positive threshold of 10 (Redmond et al., 2011). This indicates that a score below the cut-off was 5 to 9 times more likely to come from a participant with SLI than from controls. Three of six studies reported LR- values (Everitt et al., 2013; Poll et al., 2010; Redmond et al., 2011). LR- values ranged from the moderately negative value of .23 to the extremely negative value of .07 (Redmond et al., 2011), indicating that it was unlikely that a score below the cut-off could have come from a participant with SLI. All the studies were based on clinical samples; none of them were population based. The diagnosis of SLI was dependent on an arbitrary cut-off score on a language measure. All but two of the studies were also dependent on arbitrary cut-off points for marker tasks.

If we take a closer look at the design of the studies included above we can see that they exemplified the notion of heterogeneity in SLI (presented in the introduction to this section) both across and within studies. Across studies it was mostly evident in the recruitment process: what if

any language or nonverbal IQ measure was used, whom it was administered to and what cut-off points were used.

With regard to language measures, Conti-Ramsden et al. (2001) relied solely on the history of SLI at a younger age without administering any language assessment at time of study. Botting and Conti-Ramsden (2003) relied on history of SLI for initial pool of participants but followed it up with an expressive vocabulary test and a receptive grammar test to differentiate participants with SLI from participants with other communication disorders. Redmond et al. (2011) required participants to have an existing diagnosis of SLI, be enrolled in an intervention program, and supplemented it with a language screening measure. Three studies employed a board diagnostic language measure focusing mainly or entirely on expressive language subtests (Archibald & Joanisse, 2009; Everitt et al., 2013; Poll et al., 2010). Although they all focused mainly on expressive language, the content of subtests and what they assessed varied between language measures. The cut-off score used was mainly 1 SD below population mean with only two exceptions, Botting and Conti-Ramsden (2003) and Poll et al. (2010). Not all the studies provided details regarding the profile of language impairment of participants. The two studies that did illustrated the heterogeneity of SLI with in their samples. For example, Everitt et al. (2013) reported that of the children identified as SELD at baseline, 11 had expressive language delay, 22 had expressive and receptive language delay, seven had expressive and articulation delay, and seven had expressive receptive and articulation delay. Botting and Conti-Ramsden (2003) reported that 16 of the 29 children with SLI had impaired expressive and receptive language. One potential limitation of these studies is that three studies ether did not administer language measures to controls (Botting & Conti-Ramsden, 2003; Conti-Ramsden et al., 2001) or excluded participants with borderline scores (Archibald & Joanisse, 2009), risking missing children with mild language impairment.

Central to the early definitions of SLI is the mismatch between nonverbal IQ and language ability. Two sources of heterogeneity relate to nonverbal IQ in these studies: (1) whether a nonverbal IQ test was administered at all, as is the case in the Archibald and Joanisse (2009) study, or (2) when it was administered, the cut-off score and who it was administered to. Of the five studies that administered a nonverbal IQ test, two studies used a nonverbal IQ score of 70 as the cut-off and only administered it to participants in SLI group and not the control group (Botting & Conti-Ramsden, 2003; Conti-Ramsden et al., 2001). A nonverbal IQ score of 80 was used in three studies; the test was administered to all participants (Everitt et al., 2013; Poll et al., 2010; Redmond et al., 2011).

Although the studies are relatively small in number, they provide strong evidence that irrespective of the nature of heterogeneity in SLI—whether it is a by-product of the recruitment

process, or difference in the profile of language impairment or nonverbal IQ ability—Sentence Repetition is a good marker for SLI. That, coupled with the evidence supporting its use as a universal marker for language impairment fits well with the viewpoint that argues against the use of specific in SLI.

2.2 Diagnostic Accuracy Studies in other Languages

Sentence Repetition has been used as a clinical marker in studies from two languages that differ typologically from English, French and Cantonese. In comparison to English, the two languages fall on the opposite ends of the spectrum with regard to inflectional morphology. Cantonese is morphologically sparse: verbs are not inflected for tense or verb-subject agreement. In addition, it is a tonal language with function and content words receiving equal stress. French, on the other hand, is morphologically rich. Verb forms are conjugated for tense, mood, and aspect and agree with the subject on person and number. In this section, we examine the following: (1) whether Sentence Repetition as a marker was able to establish adequate levels of diagnostic accuracy as it did in English; (2) how it fared in comparison to other markers within each study; (3) how the use of different scoring systems can inform the development of more discriminating/efficient Sentence Repetition tests; and (4) how uniform/different the selection process of participants with SLI can shed light on how SLI is defined across different language populations, and how that impacts the use of Sentence Repetition as a marker. A detailed description of the stimuli, administration, and scoring methods of each study is provided in Table 2.10.

Table 2.10 *Summary of Sentence Repetition Tests used in Diagnostic Accuracy Studies in Languages Other than English*

Study	Stokes et al. (2006)
Description	Novel stand-alone diagnostic test developed for a research study.
Stimuli	16 sentences, 9-10 syllables in length. 8 sentences with aspect marker and 8 sentences passive.
Administration	Children were instructed to repeat sentences presented via free field speaker. Responses were audio recorded for later transcription.
Scoring	Four scoring methods were used: 1. Complete Sentence Correct: All/none 2. Core Elements Correct: 1 point awarded for each sentence if core elements repeated correctly, regardless of errors in other elements. 3. Error Scoring (CELF): Each sentence awarded the following points: 3 = no errors, 2= one error, 1 = two or three errors, and 0 = four or more errors. 4. Percent of correct syllables: 1 point awarded for each syllable correct. Additions, transpositions ignored.
Study	Thordardottir et al. (2011)
Description	French adaptation of CELF-P (Wiig, Secord, & Semel, 1992) by Royle and Elin Thordardottir (2003). Preliminary normative data available for 3 age groups (4.5, 5, 5.6), *n* = 78.
Stimuli	n/a.
Administration	Sentences presented live with original picture book. Live scoring,

	Fixed order increases in length and grammatical complexity.
Scoring	Scoring modified, used percent of words correctly repeated.

Study	Leclercq et al. (2014)
Description	Standardized on 455 children between ages of 7-12 years with 90 children in each age subgroup.
Stimuli	13-15 sentences depending on age. Length: 6-17 words and 11-24 syllables.
Administration	Live administration and scoring. Fixed Order with increase in length and grammatical complexity
Scoring	Seven scoring methods used divided into three main categories: Global scoring: 1. Sentences Correct: All/none 2. Number of words Morpho-syntax: 3. Syntax:1 point awarded for each sentence repeated with 2 verbs and a connecting words (substitutions allowed) 4. Verb morphology: 1 point awarded for each verb inflection correctly repeated with regard to person number & tense (Lexical substitutions allowed). 5. Function Words: Number of function words correctly repeated. Lexico-semantic: 6. Lexical words: Number of content words (mainly verbs and nouns). 2 points for correct word stem, 1 point for synonym 7. Semantic: Point awarded if main idea repeated. If sentence contains more than one idea and only one repeated, no points awarded.

Further details of Stokes et al. (2006), Thordardottir et al. (2011), and Leclercq et al. (2014) are provided in Tables 2.11-13. Stokes et al. (2006) investigated the use of Sentence Repetition and nonword repetition as clinical markers for SLI in Cantonese. They compared the performance of 14 children with SLI aged 4;2-5;7 to 15 Typically Developing Age-Matched controls (TDAM) and 15 Typically Developing Younger (TDY) MLU-matched controls. Thordardottir et al. (2011) examined the diagnostic accuracy of several measures including the following: Sentence Repetition, nonword repetition, receptive vocabulary, receptive grammar, spontaneous language, narrative production, following directions, rapid automatized naming, and digit span. Leclercq et al. (2014) was the first and only study to focus solely on Sentence Repetition as a clinical marker and explore the diagnostic accuracy of seven different scoring methods. Performance of 34 children with SLI aged 7 to 12 years was compared to 34 aged-matched controls.

The findings of all three studies point towards Sentence Repetition as a useful clinical marker in spite the different language typologies, falling in line with results of the English studies presented above. Sensitivity and specificity values for Sentence Repetition were ether just below or well exceeded the 80% threshold identified by Plante and Vance (1994): sensitivity and specificity were 86% and 92% for French (Thordardottir et al., 2011) and 77% and 97% for Cantonese (Stokes et al., 2006), respectively. LR+ values were extremely positive at 10.46 for French and 22.66 for

Cantonese. LR- values fell between moderate to extreme negative at .16 for French and .24 for Cantonese. The optimal cut-off score was the 16th centile for French, which is in agreement with English findings; an optimal cut-off score was not established in Cantonese.

Leclercq et al. (2014) extended the findings of Thordardottir et al. (2011) to older participants. With regard to the discrimination ability of different scoring methods, Leclercq et al. (2014) found that overall sensitivity and specificity values exceeded 80% at the three cut-off points investigated for all seven Sentence Repetition scores. LR+ values were highest at the 3rd centile, falling well above the recommended 10 for five of seven scores. LR- values were close to or well below the extremely negative score of .10 for the three cut-off points. ROC curves were used to determine the optimal cut-off points for each scoring methods. The most stringent cut-off points were identified for scores under the morpho-syntactic category (-1.7 to -2 SD) followed by lexico-semantic scores (-1.43 to -1.57 SD) with the most lax cut-off scores for the two global scoring methods (-1.31 to -1.38 SD). As the authors point out, the stringent cut-off points for morpho-syntactic scores indicate an area of weakness. This falls in line with the findings of the SIT-61 in English (Seeff-Gabriel, Chiat, & Dodd, 2010), with lower function word score in comparison to content word scores in repetitions of children with SLI.

As with the diagnostic accuracy studies in English, all the studies were clinical rather than population based. Unlike English, none of the studies investigated the use of Sentence Repetition as a clinical marker for young children with SLI or Adults with SLI.

Table 2.11 *Summary of Stokes et al. (2006)*

Sample, Age (M)	Recruitment Criteria	Markers	Sensitivity (%)	Specificity (%)	LR+	LR-	Findings
14 SLI 4;2-5;7 (4;11) 15 TDAM 4;1-6;9 (5;0) 15 TDY 2;11-3;6 (3;3)	All children: 1. Nonverbal IQ no lower than 1 SD < mean on CMMS 2. Passed tests of oral-motor function; articulation, and hearing. 3. No reported neurological or psychosocial dysfunction. SLI: Local child assessment center and speech language therapist caseload. 1. Previously diagnosed by SLT 2. 1.2 SD < mean on Receptive subtest of CRDLS MLU: 3.70 TDAM: 1. No lower than .67 < SD on Receptive subtest of CRDLS MLU: 4.49 TDY: 1. No lower than .67 < SD Receptive subtest of CRDLS MLU: 3.97	Nonword (total percent correct) Sentence Repetition (Error Scoring/ CELF)	n/a 77	n/a 97	n/a 25.66	n/a .24	At group level: Nonword repetition: TDAM=SLI< TDY Sentence Repetition: TDAM < SLI= TDY Sensitivity fell just below the recommended 80% Cut-off points not established

Note. SLI = specific language impairment; TDAM = chronological age-matched controls; TDY = younger controls language matched on receptive grammar scores; CRDLS = Cantonese version of the Reynell Developmental Language Scales (Reynell & Huntley, 1987); and MLU = mean length of utterance; CMMS = Columbia Mental Maturity Scales (Burgemeister, Blum, & Lorge, 1972).

Table 2.12 *Summary of Leclercq et al. (2014)*

Sample, Age (M)	Recruitment Criteria	Markers	Sensitivity			Specificity			LR+			LR-			Findings
			3rd	10th	16th	3rd	10th	16th	3rd	10th	16th	3rd	10th	16th	
34 SLI 7-12 (9.11) 34 AM 7-12 (10.2)	All children: Schools in Liege, Belgium. 1. Native speakers of French 2. No history of neuro-developmental delay or disorder and sensory impairment. 3. Nonverbal IQ ≥ 80 on the WISC-IV. 4. Passed a hearing screening test. Assessments: 1. Nonword Repetition subtest of L2MA2. 2. French adaptation of the Peabody Picture Vocabulary Test[a]. 3. Oral Language Assessment[b], two subtests: a) Sentence Comprehension b) Sentence Production SLI: Recruited via language classes in special needs schools and diagnosed as SLI by SLT 1. Score > 1.25 SD below the mean on two assessments. AM: Recruited via schools 1. Within normal range on all assessments.	Sentence Repetition (Z scores):														Sensitivity and Specificity ranged from high to very high at all 3 cut-off points. ROC curves were used to identify ideal cut-off point for each score. This cut-off point varied between -1.31 SD and -2 SD (3rd percentile) depending on the score type. More stringent cut-off points were identified for the three scoring methods that fall under Morpho-syntax (Syntax, Function Words, Verb Morphology).
		Correct Sentences	.82	.97	1.0	1.0	.88	.85	____	8.08	6.67	.18	.03	.00	
		Words	.85	.94	.94	.97	.88	.82	28.33	7.83	5.22	.15	.07	.07	
		Syntax	.94	.97	.97	.88	.76	.71	7.83	4.04	3.34	.07	.04	.04	
		Verb Morphology	.74	.88	.97	.97	.85	.82	24.67	5.87	5.39	.27	.14	.04	
		Function Words	.97	.97	.97	.97	.82	.79	32.33	5.39	4.62	.03	.04	.04	
		Lexical Words	.82	.91	.97	.94	.85	.85	13.67	6.07	6.47	.19	.11	.04	
		Semantics	.82	.97	1.0	.94	.85	.79	13.67	6.47	4.76	.19	.04	.00	

Note: SLI = Specific Language Impairment; AM = chronologically age-matched controls; SLT = speech and language therapist; L2MA2 = Battery for oral language,

writing, memory and attention, Batterie langage oral, langage écrit, mémoire, attention (Chevrie-Muller, Maillart, Simon, & Fournier, 2010); WISC-IV = Wechsler Intelligence Scale for children fourth edition (Wechsler, 2005); LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio.
[a] Echelle de Vocabulaire en Images Peabody (Dunn, Thériault-Whalen, & Dunn, 1993) [b] Evaluation du langage oral (Khomsi, 2001)
Correct sentences: number of correctly repeated sentences; Words: number of correctly repeated words; Syntax: number of grammatically correct sentences that contain two verbs and a connecting word; Verb Morphology: number of correctly inflicted verbs for number person and tense; Function Words: number of function words correctly repeated; Lexical Words: number of correctly repeated lexical words; Semantics: the number of repeated sentences that accurately conveyed the main idea's/meanings; all raw scores were converted to standard scores

Table 2.13 *Summary of Thordardottir et al. (2011)*

Sample, Age (M)	Recruitment Criteria	Markers	Sensitivity		Specificity		LR+		LR-		Findings
			10th	16th	10th	16th	10th	16th	10th	16th	
14 PLI 4;5-5;9 (5;01) 78 AM 4;0-5;9 (4;11) Diagnostic accuracy: 14 PLI AM varies depending on test	All children: Located in Montreal. 1. Monolingual speakers of Quebec French, with no significant regular exposure to other languages 2. Nonverbal IQ ≥ 70 on LIPS-R 3. Passed hearing screening PLI: Recruited via speech and language therapy department at a rehabilitation center 1. Diagnosed by an SLP as having PLI according to OOAQ (2004) guidelines or Language disorder (a severe language delay that does not meet government criteria for special needs). 2. No diagnosis of primary developmental disorder such as autism or Down Syndrome. AM: Recruited via daycare centers 1. No parental concerns regarding developmental milestones.	Receptive Vocabulary[a]	.64	.88	.89	.85	6.27	5.11	.40	.25	Sentence Repetition best individual marker. 16th best cut-off for test.
		Receptive Language[b]	.64	.71	.90	.86	7.0	5.0	.40	.33	
		Spontaneous Language:									
		MLU_w	.20	.40	.94	.85	2.68	2.92	.91	.67	
		MLU_M	.21	.36	.96	.87	5.36	2.68	.82	.74	
		Assessment of Narrative[c]									
		Story Grammar	.46	.46	.87	.81	3.49	2.41	.62	.67	
		First Mentions	.15	.31	.93	.88	2.12	2.65	.91	.78	
		Nonword Repetition[d] (PPC)	.85	.85	.88	.86	6.77	5.92	.18	.18	
		Sentence Repetition[e] (PWC)	.72	**.86**	.93	**.92**	10.89	**10.46**	.31	**.16**	
		Following Directions[f]	.93	.93	.92	.86	11.61	6.63	.08	.08	
		Rapid Automatized[g] Naming									
		Errors	.71	.71	.91	.90	7.80	7.10	.32	.32	
		Time	.36	.43	.87	.86	2.82	3.04	.74	.67	
		Digit Span[h]	.39	.54	.95	.89	7.69	4.62	.65	.52	

Note. PLI = Primary Language Impairment; AM = chronologically age-matched controls; OOAQ = Quebec Association of Speech-Language Pathologists and Audiologists (Ordre des Orthophonistes et Audiologistes du Québec); LIPS-R = Leiter International Performance Scale-Revised (Roid & Miller, 1997); MLU_w = Mean Length of Utterance in words; MLU_m = Mean Length of Utterance in morphemes; PPC = Percentage of Phonemes Correct; PWC = Percent Words Correct
[a] Canadian French normed version of the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997; Dunn et al., 1993) [b] Unpublished adaptation of the Test for Auditory Comprehension of Language (Carrow-Woolfolk, 1985) [c] French adaptation of the Edmonton Narrative Norms Instrument (Elin Thordardottir & Gagné, 2006; Gagné & Elin Thordardottir, 2006; P. Schneider, Dubé, & Hayward, 2002) [d] unpublished Quebec French test of Nonword Repetition (Courcy, 2000) [e] French adaptation of the Recalling Sentences subtest of the Clinical Evaluation of Language Fundamentals-Preschool (Royle & Elin Thordardottir, 2003; Wiig et al., 1992) [f] French adaptation of the Following Directions subtest of the Clinical Evaluation of Language Fundementals-4th ed. (Boulianne & Labelle, 2006; Semel, Wiig, & Secord, 2006) [g] Rapid

Automatized Naming of Animals (Catts, 1993) [h] French adaptation of the Digit Span subtest of the Clinical Evaluation of Language Fundementals-4th ed. (Boulianne & Labelle, 2006; Semel et al., 2006)

2.2.1 Heterogeneity

In a similar fashion to English studies discussed above, heterogeneity was evident in the recruitment process both across and within studies. The cut-off score for nonverbal IQ varied between studies from a score of 70 (Thordardottir et al., 2011) to 84 (Stokes et al., 2006). While the same cut-off point was set for children with SLI and those in the two control groups (1 SD below mean) in the Stokes et al. (2006) study, similar nonverbal IQ ability in the three groups was not evident. Children with SLI scored significantly lower than younger children with similar language skills and age-matched controls. This further questions the requirement of a mismatch between nonverbal IQ and language ability that is central to the definition of SLI, and how that mismatch is operationally defined.

The language ability requirement also varied between the three studies with regard to the type of tests used, selected cut-off points, and which participants the tests were administered to. The Stokes et al. (2006) study relied solely on receptive grammar while the Leclercq et al. (2014) used a combination of five assessments covering expressive, receptive language, and Phonological Short Term Memory. Children in the SLI group were required to obtain low scores on any two of the five tests, leading to variability in language ability with in the SLI group as well. Cut-off scores in the two studies varied between 0.67 SD to 1.25 SD. Thordardottir et al. (2011) differed from the two studies in the terminology used. Primary Language Impairment was used rather than SLI and they did not rely on norm referenced tests, but rather the clinical experience of the diagnosing speech and language therapist as well as qualitative assessment of errors and communicative behaviour. The language ability of children in the age-matched control group was not assessed and depended on the absence of concern from parents.

2.2.2 Scoring

In determining the use of Sentence Repetition as a marker for SLI, it is not only important to look at its ability as a test, but also how the structure and scoring of the test can maximize diagnostic accuracy levels. Leclercq et al. (2014) was the first study to investigate the diagnostic accuracy of qualitative scores on a Sentence Repetition test in addition to quantitative scores. The three scores in the morpho-syntactic category showed the highest discriminative power followed by lexico-semantic scores and quantitative scores, respectively. This was evident from the lower cut-off point for morpho-syntactic scores, which was a result of the minimal overlap between scores of children SLI and control group in comparison to other scores. This finding falls in line with the disadvantage of function scores over content scores found in English (Seeff-Gabriel et al., 2008), Italian (Devescovi & Caselli, 2007), Arabic (Wallan, 2006), and extends it to older children with SLI (Leclercq et al., 2014). The Stokes et al. (2006) study did not investigate the diagnostic accuracy of the Core Element score, which most resembles qualitative scoring methods. However,

based on group differences, they found that it was not as discriminating as the quantitative CELF scoring methods. Specifically, this was due the lack of difference in scores between children with SLI and the age- and language-matched control groups, as well as higher percentage of ceiling scores from predominantly passive and not aspect sentences. The Core Elements score shares features from both qualitative and quantitative scoring methods. It focuses on the core elements of a sentence based on the structure (aspect or passive) and disregards errors in noncore elements; it is not clear whether substitution of core elements with other elements from the same category was allowed. Its limited range, a score of 1 or 2 awarded for each sentence most resembles an all/none scoring method.

2.3 Diagnostic Accuracy Studies and Underlying Processes

The studies presented above provide compelling evidence that supports the use of Sentence Repetition as a clinical marker for SLI in English, Cantonese, and French. In addition to establishing whether a Sentence Repetition test is discriminating according to age or language ability, it is important to identify the skills that underpin good/poor performance on a Sentence Repetition test. When a child is unable to perform at the same level as an aged counterpart on a Sentence Repetition test, what skills does the child lack? Is it a result of poor memory skills, poor language ability, or a mix of both? Are the underlying skills involved constant or do they change according to age? Are the same skills tapped in children with and without Language Impairment or are they different? If they are different are they similar to younger children or are they completely different? Are the findings the same across typologically different languages or are they language specific? Being able to address these questions would have both clinical and theoretical implications. From a clinical perspective, it would help optimize its use as an assessment by determining what it tells us about a child's language ability, what areas require further assessment, and what to target for treatment. From a theoretical standpoint, it would have implications for models of Verbal Short Term Memory (VSTM) and theories of SLI (Polisenska et al., 2015).

Although not the primary aim of diagnostic accuracy studies, they have attempted to address the question of underlying processes involved in Sentence Repetition. One approach was to use various statistical analyses to (1) investigate the relationship between performance on Sentence Repetition and other marker tasks, which are taken to assess different underlying skills; (2) investigate the relationship between Sentence Repetition performance and a broad language assessment; and (3) investigate the relationship between different categories of qualitative scores on the same Sentence Repetition test. Three analyses were used (correlational analysis, multiple regression, and principal component analysis) to establish whether a relationship existed, the direction of that relationship, the strength of that relationship, and how much of the variance in performance was explained by that relationship.

The interpretation of relations between performance on Sentence Repetition and other marker tasks hinges on a polarized view of the underlying skills involved in marker tasks: nonword repetition tests and serial recall tasks such as digit span are assumed to primarily assess VSTM, while Grammatical Morpheme probes are assumed to assess established linguistic knowledge. If performance on Sentence Repetition is found to relate to scores on nonword repetition or digit span tests, the view that VSTM is implicated in Sentence Repetition is compatible. On the other hand, if performance on Sentence Repetition relates to Grammatical Morpheme probes, the view that linguistic knowledge is associated with Sentence Repetition performance is supported. In the case that both VSTM and linguistic knowledge are implicated, is there a difference in the degree? Due to the design of diagnostic accuracy studies, almost all the studies, with the exception of Thordardottir et al. (2011), allow for a comparison of findings for children with SLI and their controls, which clarifies whether presence or absence of language impairment influences the underlying processes tapped by marker tasks.

An alternative approach to address the question of underlying processes is to turn to the findings of differential diagnosis studies. Sentence Repetition appears to be a universal marker of Language Impairment rather than a marker for SLI. It can be argued that this finding implicates language processing being involved in Sentence Repetition. By comparing the types of developmental delay profiles in which Sentence Repetition failed to achieve acceptable diagnostic accuracy levels to instances where it did achieve acceptable levels, it is possible to postulate what skills are not tapped by Sentence Repetition.

The following sections will present a summary of findings from the three statistical analyses followed by differential diagnosis study findings. Within each section of statistical analyses, the limitations of each method and possible implications will be examined.

2.3.1 Correlational evidence and implications

Of the three statistical analyses used to investigate the relations between tasks, correlational analysis was the most common method. The strength of the correlational findings was determined using Cohen's (1992) guidelines: a correlation coefficient $r = .1$ constitutes a small effect size, $r = .3$ a medium effect size and $r = .5$ a large effect size. Two categories of correlational analyses are presented: correlation between scores on marker tasks and correlations between broad language measures and marker tasks. The main aim of this section is to determine possible contributing factors to performance on Sentence Repetition. Further aims are (1) to look at the correlations between VSTM marker tasks and Grammatical Morpheme probes to examine whether they support the view that nonword repetition tests and serial recall tasks are a pure measure of VSTM, while Grammatical Morpheme probes are a pure measure of linguistic knowledge; and (2) to compare whether nonword repetition tests and serial recall tasks were interchangeable. Two

studies allow for comparison of findings according to participant age. Everitt (2009) compared correlation findings at two times (baseline, 12-month follow-up) while Stokes et al. (2006) allows for comparison between controls in two age groups (younger language matched TDY, older age-matched controls TDAM). In the case of English studies, a comparison across studies can be made between children as young as 3 years (Everitt, 2009), middle childhood (Conti-Ramsden et al., 2001; Hesketh & Conti-Ramsden, 2013), and adults with SLI (Poll et al., 2010). A cross-linguistic comparison will be possible for correlation findings between broad language assessments and marker tasks in three typologically different languages (Cantonese, English, and French). In interpreting the findings, it is important to recognize that correlational analysis has several limitations: it only looks at the relationship between two variables at a time; it does not control for shared variance; significant correlations may be a result of the two variables correlating with a third unknown variable; and it does not determine causation (Field, 2009; Mayers, 2013).

2.3.1.1 Correlations between marker tasks

Table 2.14 summarizes the results of the correlational analyses reported in the diagnostic accuracy studies, including relationships between Sentence Repetition with Grammatical Morpheme probes and with VSTM measures, between Grammatical Morpheme probes and VSTM measures, and between VSTM measures.

2.3.1.2 Sentence Repetition: Relations to Grammatical Morpheme probes versus Verbal Short Term Memory

Sentence Repetition scores and Grammatical Morpheme probes demonstrate mainly strong correlations in Language Impaired and Typically Developing participants. This was true for young children and children in middle school. The two exceptions were between Sentence Repetition and Third Person Singular in Conti-Ramsden et al. (2001) and Everitt et al. (2013); the authors postulated that could be due to ceiling effects and floor effects respectively. These findings are compatible with the view that linguistic knowledge contributes to performance on a Sentence Repetition test. Poll et al. (2010) was the only study that did not investigate this relationship, thus the findings could not be generalized to adults with SLI.

The relationship between Sentence Repetition and VSTM measures exhibited medium correlations. Three of the four non-significant correlations occurred in Language Impaired groups: SELD children at follow-up (Everitt, 2009) and adults with SLI (Poll et al., 2010). The findings suggest that VSTM is also a contributing factor to performance on Sentence Repetition. Due to the reduced strength of the relationship between Sentence Repetition and VSTM measures in comparison to Sentence Repetition and Grammatical Morpheme probes, VSTM may be implicated less when compared to linguistic knowledge. Unlike relations with Grammatical Morpheme

probes, there may be a discrepancy between correlation findings in participants with SLI and controls, possibly indicating that the contribution of VSTM is less in participants with SLI.

2.3.1.3 Relations between Grammatical Morpheme probes and Verbal Short Term Memory measures

Over half the correlations between Grammatical Morpheme probes and VSTM measures were significant with mainly moderate effect sizes. These findings suggest that VSTM measures are related to linguistic knowledge while Grammatical Morpheme probes are linked to VSTM, putting into question the polarized view of the underlying skills tapped by the two assessments. In comparison to the correlation findings between Sentence Repetition and Grammatical Morpheme probes, VSTM measures appear to be less linked to linguistic knowledge given the reduction in the strength of correlation coefficients.

2.3.1.4 Relations between Verbal Short Term Memory measures: Same or different?

Everitt (2009) and Poll et al. (2010) were the only studies that reported correlations between VSTM measures. Correlations between single item VSTM tests (word and nonword) and serial recall tasks (digit span) failed to reach significance for adults with SLI and controls (Poll et al., 2010) and children with SELD at baseline and follow-up (Everitt, 2009). For children in the SELD group, strong correlations were chiefly found between single-item tests at baseline and follow-up. This was true irrespective of whether the relationship was between the same category of linguistic items but from different tests (the nonword subtest of the PSRep and the CNRep) or whether the relationship was between different types of linguistic items (the nonword and word subtests of the PSRep, and the nonword subtest of the PSRep and the CNRep). The Typically Developing controls in Everitt's (2009) study were the only exception, showing significant medium to strong correlations irrespective of the type of VSTM test. These findings raise the possibility that single item tests and serial recall tasks may be related to different underlying skills or the same underlying skills but to varying degrees. This appears to be especially true for young children with SELD as well as adults irrespective of language ability.

2.3.1.5 Correlations between broad language measures and marker tasks

Three studies looked at the correlation between marker tasks and broad language measures (English: Everitt, 2009; Cantonese: Stokes et al., 2006; French: Thordardottir, Keheyia, Lessard, Sutton, & Trudeau, 2010). Two types of language assessments were utilized. The first was an omnibus assessment of expressive and/or receptive language—the Preschool Language Scale (PLS-3; Zimmerman et al., 1997) and the French adaptation of the Test of Auditory Comprehension of Language Revised (TACL-R; Carrow-Woolfolk, 1985). The second was a comprehensive assessment of a specific component of language such as receptive vocabulary—the Cantonese Receptive Vocabulary Test (CRVT; K. Y. S. Lee, Lee, & Cheung, 1996) and the French

adaptation of the Peabody Picture Vocabulary Test (Dunn et al., 1993)—and receptive grammar—the Receptive subtest of the Cantonese version of the Reynell Developmental Language Scales (CRDLS–R; Reynell & Huntley, 1987). All three studies found that Sentence Repetition was the strongest marker linked to broad language assessments, as reflected in the degree and number of significant correlations. This finding supports the relationship found between Sentence Repetition and Grammatical Morpheme probes in English and extends it to typologically different languages. The following sections will provide a summary of the findings of each of the three studies along with implications.

Everitt (2009) investigated the correlations between marker tasks and the three scores of the PLS-3 at baseline and follow-up: Auditory Comprehension, Expressive Communication, and Total Communication scores. Table 2.15 presents the correlations between the three PLS-3 scores and marker tasks and shows that Sentence Repetition scores were significantly correlated with the three PLS-3 scores for Language Impaired and control groups. At baseline, effect sizes were strong for controls and moderate for participants with Language Impairment. At follow-up, moderate to strong relationships were observed in both groups. The difference in the strength of the relationship at both timelines for children with Language Impairment suggests that the degree linguistic knowledge is implicated may vary according to the age of participants, with older children relying more on established linguistic knowledge than younger children with LI.

Stokes et al. (2006) also found that Sentence Repetition showed stronger correlations with language measures in comparison to nonword repetition. Sentence Repetition scores of the younger language control group TDY strongly correlated with receptive vocabulary and receptive grammar scores (r = .59 and .90, respectively), while Sentence Repetition scores of children with SLI strongly correlated with receptive grammar only (r = .58). For nonword repetition, the only significant correlation was a strong correlation with receptive vocabulary in the TDY group (r = .65, $p < .01$). The TDAM group did not show any significant correlations. Authors argued that the relationship between linguistic knowledge and Sentence Repetition in Typically Developing participants is dependent on age, with performance on Sentence Repetition reflecting linguistic knowledge in younger children and independent of it by 5 years.

Thordardottir et al. (2011) did not report any correlational analysis but did so in an earlier normative study. Thordardottir et al. (2010) showed that performance on Sentence Repetition for French-speaking children between the ages of 4;6 and 5;6 was more linked to linguistic knowledge than nonword repetition. This was evident from the strong correlations found between Sentence Repetition scores and scores on tests of receptive language and receptive vocabulary (r = .74 and .55, respectively), while nonword repetition scores only showed a moderate correlation with receptive language assessment (r = .46). Both Stokes et al. (2006) and Thordardottir et al. (2010)

supported findings in English that while Sentence Repetition and nonword repetition are linked to linguistic knowledge, Sentence Repetition appears to be more strongly linked.

The overall pattern of the correlational analyses results in relation to underlying processes highlights the advantage of Sentence Repetition as an assessment over the other marker tasks investigated. When comparing Sentence Repetition to nonword repetition, both tests appear to simultaneously reflect VSTM and established linguistic knowledge. However, established linguistic knowledge appears more linked to Sentence Repetition based on the stronger relationship found between Sentence Repetition and Grammatical Morpheme markers, and between Sentence Repetition and broad language measures cross linguistically. When comparing Sentence Repetition and Grammatical Morpheme probes, VSTM was implicated more in Sentence Repetition as is evident from the stronger relationship with VSTM markers. The findings suggest that it is important to investigate underlying processes in children with different language abilities and different age groups as well. With regard to VSTM tests, it highlights that not all tests are created equal with differences between serial recall and single item tests.

Table 2.14 *Correlations between Markers in Diagnostic Accuracy Studies of English Specific Language Impairment*

Study	Age	Language status	SR & Grammatical Morpheme probes		SR & VSTM measures		VSTM & Grammatical Morpheme probes		VSTM measures	
			Measure	***r*/*r*$_s$**	**Measure**	***r*/*r*$_s$**	**Measure**	***r*/*r*$_s$**	**Measure**	***r*/*r*$_s$**
Conti-Ramsden et al. (2001) (r_s)	11	AM	SR & 3prsg SR & Past Tense	.06 ns .62**	SR & CNRep	.34**	CNRep & 3prsg CNRep & Past Tense	.13 ns .39**	n/a	
		SLI	SR & 3prsg SR & Past Tense	.57* .62*	SR & CNRep	.55**	CNRep & 3prsg CNRep & Past Tense	.37** .39**	n/a	
Hesketh and Conti-Ramsden (2013) (r)	11	AM	SR & Past Tense	.53**	SR & CNRep	.43**	CNRep & Past Tense	.42**	n/a	
		SLI	SR & Past Tense	.64**	SR & CNRep	.54**	CNRep & Past Tense	.44**	n/a	
Poll et al. (2010) (r)	18;0-25;0	AM	n/a		SR & Digit span	.50*	n/a		NRT & Digit span	.30 ns
		SLI	n/a		SR & Digit span	.52 ns	n/a		NRT & Digit span	.19 ns
Everitt (2009) Baseline (r_s)	3;0-4;0	AM	SR & 3prsg	.47**	SR & PSRep: Word Nonword SR & Digit span	 .47** .41** .58**	PSRep & 3prsg: Word Nonword Digit span & 3prsg	 .37** .15 ns .33*	PSRep: Word & Nonword Word & Digit span Nonword & Digit span	 .37** .40** .44**
		SELD	SR & 3prsg	.16 ns	SR & PSRep: Word Nonword SR & Digit span	 .29* .34* .45*	PSRep & 3prsg: Word & 3prsg Nonword & 3prsg Digit span & 3prsg	 .23 ns .06 ns .10 ns	PSRep: Word & Nonword Word & Digit span Nonword & Digit span	 .68** .21 ns .25 ns

Everitt (2009) Follow-up (r_s)	4;0-5;0	AM	SR & 3prsg	.32*	SR & PSRep:		PSRep & 3prsg:		PSRep:	
			SR & Past Tense composite	.43*	Word	.30*	Word	.34**	Word & Nonword	.59**
					Nonword	.41**	Nonword	.17 ns	Word & Digit span	.38**
							Digit span & 3prsg	.09 ns	Nonword & Digit span	.27**
					SR & Digit span	.46**	CNRep & 3prsg	.24 ns	CNRep & Word	.60**
					SR & CNRep	.29 ns			CNRep & Nonword	.47**
							Word & Past Tense	.34*	CNRep & Digit span	.56**
							Nonword & Past Tense	.32*		
							Digit span & Past Tense	.13 ns		
							CNRep & Past Tense	.24 ns		
		SELD	SR & 3prsg	.37*	SR & PSRep:		PSRep & 3prsg:		PSRep:	
			SR & Past Tense composite	.44*	Word	.23 ns	Word	.38*	Word & Nonword	.54**
					Nonword	.30 ns	Nonword	.53**	Word & Digit span	-.04 ns
							Digit span & 3prsg	.27 ns	Nonword & Digit span	-.03 ns
					SR & Digit span	.40*	CNRep & 3prsg	.40*	CNRep & Word	.57**
					SR & CNRep	.41**			CNRep & Nonword	.42*
							PSRep & Past Tense		CNRep & Digit span	.17 ns
							Word	.38*		
							Nonword	.51**		
							Digit span & Past Tense	.27 ns		
							CNRep & Past Tense	.43*		

Note. For full details of tests with references, please refer to each study table. All studies used raw scores with the exception of (Everitt, 2009), who used standard score for SR and percentile score for Digit span. SLI = specific language impairment; SELD = specific expressive language delay; AM = chronological age-matched controls; SR = recalling sentences subtest of CELF; 3rdpsg = third person singular probe; CNRep = Children's Test of Nonword repetition (Gathercole & Baddeley, 1990); PSRep = Preschool Repetition test. (Seeff-Gabriel et al., 2008); n/a = not available; ns = not significant.
* $p < .05$
** $p < .001$

Table 2.15 *Spearman's Correlations between PLS-3 Scores and Markers Tasks in Everitt (2009)*

	Language group	PLS-3	SR	Grammatical morpheme probes		VSTM Markers			
						PSRep		CNRep	Digit span[b]
				3prsng[a]	Past Tense Composite[a]	Word	Nonword		
Baseline	AM	PLS-3 AC	.65**	.50**		.29*	.48**		.60**
		PLS-3 EC	.70**	.41**		.57**	.38**		.51**
		PLS-3 Total	.75**	.52**	n/a	.47**	.48**	n/a	.62**
	SELD	PLS-3 AC	.34*	-.18 ns		-.17 ns	.05 ns		.37*
		PLS-3 EC	.45*	-.13 ns		.16 ns	.33*		.28 ns
		PLS-3 Total	.39*	-.20 ns		-.09 ns	.16 ns		.39*
Follow-up	AM	PLS-3 AC	.51**	.26 ns	.23 ns	.34*	.26 ns	.26	.46**
		PLS-3 EC	.48**	.22 ns	.12 ns	.25 ns	.24 ns	.30*	.31*
		PLS-3 Total	.55**	.31*	.20 ns	.35**	.30*	.33*	.44**
	SELD	PLS-3 AC	.41*	.04 ns	-.05 ns	-.11 ns	-.02 ns	-.08 ns	.37*
		PLS-3 EC	.70**	.20 ns	.30 ns	.16 ns	.11 ns	.28 ns	.17 ns
		PLS-3 Total	.63**	.14 ns	.10 ns	-.01 ns	.04 ns	.07 ns	.36*

Note. PLS-3 = preschool language scale-3 UK (Zimmerman et al., 1997); SR = Recalling Sentences subtest of the CELF-Preschool UK (Wiig et al., 2000); 3rdpsg = third person singular Probe from the Rice/Wexler Test of Early Grammatical Impairment (Rice & Wexler, 2001); PSRep = Preschool Repetition test (Seeff-Gabriel et al., 2008); CNRep = Children's Test of Nonword repetition (Gathercole & Baddeley, 1990); AM = chronological age-matched controls; SELD = specific expressive language delay; PLS-3 AC: Auditory Comprehension subscale from the Preschool Language Scale-3 (UK); PLS-3; EC: Expressive Communication subscale from the Preschool Language Scale-3 (UK); PLS-3 Total = Total language from the Preschool Language Scale-3 (UK); n/a = not available; ns = not significant.
[a]Test of Early Grammatical Impairment (Rice & Wexler, 2001) [b]Recall of Digits Forward subtest of the British Ability Scales II (Elliott, 1997)
* $p < .05$
** $p < .001$

2.3.2 Multiple regression and implications

Hesketh and Conti-Ramsden (2013) did not rely solely on correlation to address the issue of underlying processes. They took it a step further and utilized multiple regression analysis. The advantage of multiple regression over correlational analysis is that rather than looking at two variables at a time, it looks at the relationship between multiple predictor variables and a single outcome variable. Multiple regression helps to clarify how much of the variance can be explained by the relationship between one of the predictor variables and the outcome variable by controlling for the other predictor variables, therefore teasing apart the influence of independent correlations from shared correlations (Field, 2009; Mayers, 2013). In this study, the predictor variables were scores on the CNRep and past tense morphology elicitation task, and the outcome variable was score on a Sentence Repetition test.

If we take a look back at Table 2.14, it would appear that middle school children with SLI and their age-matched controls showed a similar profile of correlations, scores on the CNRep, and past tense morphology significantly correlated with Sentence Repetition scores. However, the multiple regression model revealed a different story. For children with SLI, VSTM as measured by CNRep and linguistic knowledge as measured by past tense morphology were predictive of Sentence Repetition scores and explained 51% of the variance. For children in the Typically Developing age-matched group, only past tense morphology significantly contributed to the model, explaining 33% of the variance.

Hesketh and Conti-Ramsden (2013) proposed an explanation for why VSTM appeared to be implicated in the Sentence Repetition of children with SLI and not their age-matched counterparts. Children in the control group can call on their linguistic knowledge to repeat a sentence either by chunking the target sentence to smaller easier-to-process elements, or by mapping the target sentence on to "predictable structural representations" (p. 6), therefore reducing the load on VSTM. Children with SLI, on the other hand, might not have established linguistic representations that would allow for breaking apart the target sentence or a structural template to map the items onto, therefore processing each sentence element individually like a serial recall task, which in turn places a greater demand on VSTM abilities.

A limitation of the study is that it relied on a single test, CNRep, to assess the contribution of VSTM. The authors argued that nonword repetition is not a pure measure of VSTM and places less demand on VSTM in comparison to a serial recall task such as digit span. We know from correlational findings presented above that single item and serial recall tasks are not interchangeable in preschoolers and adults with SLI, and that there is a relationship between Grammatical Morpheme probes and VSTM measures. There are also differences in the discrimination ability of different nonword repetition tests based on task design and scoring

(CNRep versus NRT) and within a single test based on syllable structure of the stimuli as was evident in the Cantonese nonword repetition test (Stokes et al., 2006) (attested vs. unattested) nonwords. To conclude, although multiple regression provides more information than just correlation, it does not tell the full story and is only as good as the predictor variables in the model.

2.3.3 Principal component analysis

The methods presented thus far compared Sentence Repetition to other tests with different task demands. Leclercq et al. (2014) investigated the relationship between different qualitative scores on a Sentence Repetition test to see whether they corresponded to the two broad categories they were designed to asses: lexico-semantic and morpho-syntactic abilities. The aim of principal component analysis is to reduce the observed scores to underlying linear factors (Field, 2009). As predicted, the five scores were reduced to two components explaining 96.48% of the variance: a lexico-semantic component with two scores (lexical words and semantics) explaining 43.92% of the variance, and a morpho-syntax component with three scores (syntax, function words, and verb morphology) explaining 52.56% of the variance. In addition to providing construct validity for the test scores, this study provides support for the implication of linguistic knowledge in the repetition of sentences.

2.3.4 Evidence from differential diagnosis studies and implications

Differential diagnosis studies indicate that Sentence Repetition is less sensitive to impairments in pragmatic language (Botting & Conti-Ramsden, 2003) and attention (Redmond et al., 2011) in comparison to language difficulties. Botting and Conti-Ramsden (2003) found that of four developmental delay groups, Sentence Repetition failed to identify children with PLI Plus (pragmatic language impairment as measured by the Children's Communication Checklist (Bishop, 1998), some autistic traits, no language difficulties) from controls. Sentence Repetition was able to distinguish children with PLI Plus from participants with SLI, PLI Pure, and autism (all three groups showed language difficulties). Sentence Repetition showed good diagnostic accuracy levels in discriminating children with SLI from age-matched controls and children with ADHD (Redmond et al., 2011). Children in both the ADHD and control groups did not exhibit any difficulties with language as measured by a language screening test and a tense marking task.

2.3.5 Conclusion

If we go back to the questions raised at the start, diagnostic accuracy studies do not provide conclusive answers regarding the underlying processes involved in Sentence Repetition. When it comes to the relationship between Sentence Repetition and linguistic knowledge, the indirect evidence presented here through statistical methods and differential diagnosis studies implicate linguistic processing. This appears to be true for English and typologically different languages such as Cantonese and French. The picture is less clear when we look at the relationship between VSTM

and Sentence Repetition. The weak link here appears to be the underlying assumption that linguistic knowledge can be untangled from VSTM measures. Rather than addressing the underlying process question by determining whether language processing or VSTM is implicated in Sentence Repetition, an alternative way is to directly and systemically manipulate the linguistic features of a serial recall task and Sentence Repetition to see how that influences the performance of participants. For example, the linguistic items can be varied in a serial recall task: digit versus word versus nonwords and the span scores for the three tasks can be compared. In Sentence Repetition, grammatical complexity, semantic plausibility, and grammaticality, among other linguistic features, can be manipulated and performance on the different sentence types can be compared.

2.4 A Closer Look at the Anomalous Findings of the Cantonese Study: Implications for Repetition Tasks

The Cantonese study (Stokes et al., 2006) presented two anomalous findings from previous research. While the Nonword Repetition test was sensitive to the age of participants, older participants with SLI and their TDAM controls performed better than the TDY MLU-matched controls. This was the first study to report a lack of difference between scores of children with SLI and their age-matched controls. Results of the Sentence Repetition test indicated that the core elements correct score, the only qualitative scoring system employed, was less discriminating than the other three purely quantitative scores. This result is inconsistent with the findings reported in English (Seeff-Gabriel et al., 2008), Italian (Devescovi & Caselli, 2007), and French (Leclercq et al., 2014). The following sections will address these two findings and explore possible explanations. Emphasis will be placed on how the manipulation of different variables in the design of the Nonword and Sentence Repetition tests can shed light on the underlying processes tapped by these two tests and on the clinical implications of these findings.

2.4.1 Differences in target content: The influence of language typology, syllable structure, and length of nonwords

Stokes et al. (2006) postulated that the failure of the Nonword Repetition test to discriminate between children with SLI and TDAM children could have resulted from distinct features of Cantonese that were reflected in the stimuli: simple phonotactic structures consisting of only singleton consonants; simple prosodic features with all syllables receiving equal stress; and limited phonemic inventory with a limited number of consonants that can occur in syllable final position in comparison to syllable initial position. These three factors are far more complex in other languages such as English, Swedish, and Dutch and have been shown to influence performance of children with SLI on nonword repetition tests. For example, nonwords containing consonant clusters were found to be challenging for children with SLI to repeat (English: Archibald &

Gathercole, 2006; Bishop et al., 1996). The presence or absence of stress and its position in a nonword are also important factors: children with SLI are more prone to omit weak syllables that occur in pre-stressed positions in some languages (English: Chiat & Roy, 2007; Dutch: de Bree, 2007; Swedish: Sahlén, Reuterskiöld-Wagner, Nettelbladt, & Radeborg, 1999). Interestingly, tone was one aspect of Cantonese typology that was not manipulated. Items of the same syllable length were appointed the same tonal pattern as a real word of the same length in order not to confound the influence of syllable structure, item length, and language group.

Another explanation that was put forward by Stokes et al. (2006) was related to how the stimuli were constructed and how that, in turn, influenced the underlying skills tapped by the test. The test consisted of two categories of stimuli: nonwords created from consonant-vowel structures that appear in Cantonese words and nonwords created from consonant-vowel structures that do not appear in Cantonese words. Within both SLI and TDAM groups, the first type of nonword was easier to repeat than the latter. Across-group comparisons revealed that children with SLI and TDAM children were equally able to repeat the nonwords consisting of unattested consonant vowel structures (both scored 45%). However, TDAM children were more accurate than children with SLI in repeating nonwords containing attested consonant vowel syllables (74% vs. 67%). The authors cautioned that the difference was not statistically significant and stated that it could be due to the study's small sample size. They proposed that this pattern of performance was due to the two sets of stimuli tapping different underlying skills. Nonwords with unattested syllables relied more heavily on VSTM because of the lack of representation of these syllables in Cantonese while established linguistic knowledge could be drawn upon to recall nonwords with attested syllables. The higher score obtained by children with SLI and TDAM controls on attested versus unattested nonwords shows that both groups benefit from drawing upon their established linguistic knowledge. However, the higher scores attained by TDAM controls on attested nonwords shows that they benefit to a greater degree from the aid of established linguistic knowledge than children with SLI. This finding falls in line with the greater discrimination ability of the CNRep test versus the Nonword Repetition Test (NRT) test in English (Graf Estes, Evans, & Else-Quest, 2007). Because of the characteristics of the CNRep stimuli, it is thought to draw on established linguistic knowledge more so than the NRT test.

In addition to syllable structure, Stokes et al. (2006) investigated the influence of item length (as measured by the number of syllables) on the performance of children with SLI and TDAM children. Repetition scores of both groups showed a similar pattern: as the number of syllables increased, the accuracy of repetition decreased and the gap between the mean scores of both language groups increased in magnitude. For shorter nonwords, children with SLI and TDAM children performed at a similar level, for one-syllable (77% vs. 75%) and two-syllable (both 76%).

However, for longer nonwords, children with SLI and TDAM children obtained lower scores. Repetition scores for three-syllable nonwords and four-syllable nonwords were 62% versus 52%, respectively, for children with SLI compared with 69% versus 62% for TDAM children. Again, no significant difference was found between the repetition scores of children with SLI and TDAM children. Although the gap between the two groups increased as length increased, no significant length x group interaction was found ether. In an attempt to unpack the influence of syllable structure from length on nonword repetition scores, Stokes et al. (2006) reported that while the total score for attested and unattested nonwords differed between groups, language groups did not differ on attested nonwords for individual syllable lengths, indicating that difference in performance could be attributed to syllable structure and was independent from the effect of item length.

Length effects were also found in English nonword repetition tests, where the difference between children with SLI and age-matched controls increased in magnitude as the length of nonwords increased. Graf Estes et al. (2007) conducted an exploratory meta-analysis of 16 studies that reported effect sizes for each nonword length and found that the degree of effect sizes increased from medium effect sizes in one- and two-syllable nonwords to large effect sizes for three- and four-syllable nonwords. The greater difficulty faced by children with SLI and age controls in repeating longer nonwords in both the Stokes et al. (2006) study and Graf Estes et al. (2007) meta-analysis could be explained by longer nonwords placing a greater demand on VSTM in comparison to short nonwords. Children with SLI may have a reduced VSTM capacity in comparison to age-matched controls and therefore can be disproportionately influenced by nonword length. However, VSTM cannot explain the full story. Although the effect size is reduced, children with SLI still perform worse than age controls on short nonwords. Graf Estes et al. (2007) argue that the reduced effect size for short nonwords may result from insufficient statistical power due to small study size or a lack of one-syllable nonword test items. To investigate the influence of length further, Graf Estes et al. (2007) ran a multiple regression model using nonword length as a predictor of effect size magnitude and found that the model only explained 18% of variance, indicating that factors other than length influenced the difference in performance between children with SLI and age-matched controls. Leonard (2014) in a paper reviewing SLI across languages argues that although length effects are generally maintained across languages, the degree to which length influences nonword repetition scores varies according to language typology. Morphologically rich languages such as Spanish and Italian tend to have on average longer words than morphologically sparse languages such as English. Italian-speaking children with SLI are more adapt at producing three- and four-syllable words in their day-to-day language in comparison to English-speaking children with SLI. This, in turn, is reflected in the higher repetition scores for three- (80%) and four-syllable nonwords (70%) obtained by Italian-speaking children in

comparison to their English-speaking counterparts who obtained 55% and 35% for the same nonword lengths, respectively (Deevy, Weil, Leonard, & Goffman, 2010; Dispaldro, Leonard, & Deevy, 2013; as cited in Leonard, 2014, p. 3). A possible explanation is that three- and four syllable nonwords place less of a demand on VSTM capacity of Italian-speaking children with SLI. To conclude, length effects can vary in a single language based on the child's language experience and can vary between languages.

2.4.2 Implications of nonword repetition findings: Test design and underlying processes

Although nonword repetition is not the focus of this study, it shares task demands with Sentence Repetition in that they both rely on the immediate reproduction of a modelled utterance by the child. The nonword repetition findings of the Stokes et al. (2006) study is informative with regard to the design and implementation of Sentence Repetition tests as well as the underlying processes involved in immediate repetition. With respect to test development, it highlights the importance of constructing a test that reflects language typology and how that can impact results.

Findings emphasize the importance of investigating the assessment's utility across different languages, age, and language ability groups within a single language. Just because a test is discriminating in many languages does not guarantee the same finding in all languages. The usual pattern of results is that older children with typical language development obtain higher scores than younger children, and children with SLI usually perform at a similar magnitude to younger language-matched children. The Cantonese study is a reminder that this is not always the case, since children with SLI obtained similar scores to their age controls on nonword repetition.

When it comes to the underlying processes involved in nonword repetition, originally it was considered a pure measure of VSTM. However, as the Cantonese study shows along with English studies, it is not free from influence of linguistic knowledge. The Stokes et al. (2006) study shows that it is possible to develop a clinical tool with the intention of simultaneously identifying children with language difficulties and investigating what underlying skills the assessment taps into. The knowledge gained from the investigation of underlying skills can then be utilized in developing a more discriminating clinical tool, for example, a nonword repetition test consisting of long nonwords with unattested syllables. It also highlights the difficulty of developing an assessment that systematically manipulates all the possible factors that could influence performance. In this study, tone was one aspect of Cantonese typology that was controlled for rather than manipulated.

2.4.2.1 Differences in target content: The influence of language typology, stimuli, and scoring method on Sentence Repetition

The core elements correct score is considered a qualitative scoring method because it depends on the correct repetition of particular features of a sentence (see Table 2.16 for details of scoring methods). These features vary depending on sentence type (aspect or passive). The other three scoring methods—error scoring (CELF), complete sentence score, and percent syllable correct—are broad/overall measures of accuracy or errors, without accounting for specific features of a sentence. The lack of discrimination of the core elements scoring method stemmed from the failure to differentiate between scores of children with SLI and their age-matched peers TDAM in the passive sentences condition, in contrast to the other three scoring methods that were able to differentiate between the two language groups in passive and aspect sentences. No scoring method was able to differentiate between children with SLI and younger MLU-matched controls TDY.

One could argue that the lack of discriminative power of the core elements correct scoring method specifically for passive sentences was due to the limited range of possible scores per sentence (all/none). Yet this limited range was sufficient to differentiate between children with SLI and their age-matched peers using the same scoring method for aspect sentences. Furthermore, the quantitative complete sentence correct score was adequate to discriminate between language groups for passive sentences. Looking more closely at the data it is apparent that the main difference in results between core passives and complete passives was that four children with SLI, five children in the age-matched control group, and four in the language-matched control group were able to obtain maximum core scores while none were able to do so for complete passives. This indicates that core passive elements were not as vulnerable as non-core passive elements or core aspect elements. It is difficult to pinpoint with certainty which non-core elements in passive sentences were a particular source of difficulty for children with SLI. The grammatical category of classifiers is a possible candidate and was included once or twice in each aspect and passive sentence. This would be in line with the findings of Stokes and So (1997) who reported that 14 Cantonese-speaking children with SLI (mean age 53 months) produced shape classifiers with less frequency than their age-matched controls.

Chiat et al. (2013) point out that there is no obvious explanation for why the qualitative scoring method core elements correct was not effective in discriminating between children with SLI and children in the TDAM group. In order to consider possible reasons why, it is important to take a closer look at aspect markers and passives in Cantonese. According to Fletcher, Leonard, Stokes, and Wong (2005), aspect markers are monosyllabic stressed morphemes that are placed after verbs and do not undergo any morpho-phonemic processes. The aspectal system in Cantonese includes a total of six morphemes. The studies presented here focus on perfective and imperfective

forms. For example, “gan2” is an aspect marker that refers to an ongoing activity (the number represents the tone, which in this case is high rising). Deleting an aspect marker from a sentence would not affect the grammaticality of a sentence. According to Leonard, Wong, Deevy, Stokes, and Fletcher (2006), passives in Cantonese share some commonalities with passives in English: they undergo word order changes with the patient moving to subject position and the insertion of a by-phrase “bei2” following the verb. Unlike English, the verb remains uninflected and tenseless; the by-phrase is stressed and omitting the by-phrase would render the sentence ungrammatical. Aspect markers and the by-phrase are similar in that they are both phonologically salient with the main difference being that aspect markers are optional while the by-phrase in passive sentences is obligatory.

Interestingly, the pattern of results for the core elements scoring method mirrors findings of two studies by the same research group with the same sample of participants but using elicitation tasks rather than Sentence Repetition (Fletcher et al., 2005; Leonard et al., 2006). The Fletcher et al. (2005) investigation focused on aspect markers. Black and white line drawings were used to probe the imperfective aspect marker “gan2” in past and present contexts, and the perfective aspect marker “zo2”, the same aspect markers that were utilized in the Stokes et al. (2006) Sentence Repetition test. Results found that children with SLI produced perfective and imperfective aspect markers significantly less frequently than children in the TDAM or TDY groups, which did not differ significantly from each other. In addition, the children’s use of a past-time temporal adverb was investigated. Unlike aspect markers, children with SLI performed at the same level as children in the TDY group, indicating that they did not have a difficulty with temporal notions. In order to elicit passive sentences, Leonard et al. (2006) employed two tasks that involved item manipulation. In the first task, the child was asked a question about the patient without priming. In the second task, a passive construction was modelled first by the examiner and then the child was asked to describe a second action while the examiner was holding up the patient. Contrary to the findings of Fletcher et al. (2005), although children with SLI produced passive structures less frequently than TDAM children, the numerical difference was not statistically significant. Close examination of the SLI group’s attempt at passives did not show a tendency to omit the by-phrase, confuse patient and agent, or produce an active sentence. The researchers in both studies examined whether the findings were in keeping with a number of SLI theories and were unable to find one that fully explained the pattern of results. For example, a surface account of SLI (Leonard, 1989) predicts that passives and aspect structures remain intact due to their phonological salience. A Morphological Richness account (Lukacs, Leonard, Kas, & Pléh, 2009) on the other hand, predicts that due to morphological sparseness of Cantonese, children with SLI would find both structures difficult. The researchers proposed that due to the optional nature of aspect markers in Cantonese,

children with SLI were more likely to omit them while the obligatory nature of the by-phrase in passive sentences guaranteed their retention. Taking the findings of Stokes et al. (2006) and the production studies, into account the strength and weakness of each of the 4 scoring methods are summarized in Table 2.16.

Table 2.16 *Strength and Weakness of the Four Scoring Methods of Stokes et al. (2006)*

Scoring Method	Strength	Weakness
Core elements correct	Mirrors findings of production studies investigating aspect and passives structures in Cantonese	13 out of 44 children from the three language groups (TDY, TDAM, SLI) obtained maximum scores for passive sentences Limited range of scores Diagnostic accuracy measures for aspect sentence were not reported
Complete correct	None	9 children scored 0% on complete passive and 15 children scored 0% on complete aspect largely in TDY and SLI for both sentence types. Purely quantitative Limited range of scores
Error/CELF	No floor/ceiling Graded The most differentiating between language groups	Purely quantitative
Percent syllables correct	No floor/ceiling	Discriminant function analysis for both sentence types entered together showed a high rate of false negatives, just over half the children with SLI were misclassified as TDAM. High false negatives were also obtained when both sentence types were entered separately.

2.4.3 Implications for Sentence Repetition tests: Targets, scoring, and underlying processes

Cantonese-speaking children with SLI did not have difficulty repeating the core elements of passive sentences or producing them in elicitation tasks. This contradicts findings in English, where passives are a known area of weakness for English-speaking children with SLI and were included in Redmond's (2005) Sentence Repetition test to ensure grammatical complexity and avoid ceiling performance by children in the age-matched control group. There is also variability in cross-linguistic acquisition studies. In some languages such as Sesotho, a Southern Bantu language (Demuth, 1990), Kiswahili and Kigiriama, Eastern Bantu languages (Alcock, Rimba, & Newton, 2012), and Inuktitut, the language of the Inuit of Arctic Canada (Allen & Crago, 1996) passives can be fully acquired by Typically Developing children 3 years of age or younger. In other languages such as Hebrew, Catalan, and Lithuanian, passives are not fully mastered by the age of 5 years and are therefore not the ideal markers to identify children with SLI (Armon-Lotem et al., 2016). This

variability in whether passives structures pose a serious difficulty for children with SLI and the age of mastery by Typically Developing children highlights the importance of identifying the elements within each language that are vulnerable for children with SLI and customizing the stimuli of the Sentence Repetition test accordingly.

Stokes et al. (2006) is the first study to compare the diagnostic utility of different scoring methods for a Sentence Repetition test. Determining which of the four scoring methods was found to be the most informative depends on the reason behind administrating the test. If the purpose was only to discriminate between children with SLI and their age-matched controls, the error/CELF scoring method was the most ideal. This was due to the fact that the error scoring method was able to differentiate between children with SLI and their age-matched controls for both aspect and passive sentences without the floor and ceiling effects of the other scoring methods, as it yielded the best diagnostic accuracy values. If the purpose of the test was to provide a qualitative profile of the child's productive ability, the core elements correct scoring method was more suitable. It was the only scoring method that mirrored the vulnerability of aspect markers and the resilience of passives in production tasks reported by Fletcher et al. (2005) and Leonard et al. (2006). It would have been interesting to know the diagnostic accuracy values for core aspect scores to compare with the error scoring method but those were not reported. It might be argued that a Sentence Repetition test should employ a combination of fine-grained quantitative score like the CELF scoring to identify individuals with language difficulties and a fine-grained qualitative score like the SIT (Seeff-Gabriel et al., 2008) to identify vulnerable grammatical elements that warrant further investigation.

Turning to the underlying processes tapped by Sentence Repetition, the weakness of aspect markers that act similarly to function words in English shows the vulnerability of morpho-syntax in children with SLI and supports the role language knowledge plays in Sentence Repetition. As to why aspect markers were weak and not passive sentences, there are no clear answers but it demonstrates that Sentence Repetition can provide qualitative evidence of difficulties and can inform theories of SLI.

Chapter Three: Linguistic Manipulation of Immediate Repetition Targets: Serial Recall and Sentence Repetition

3.1 Linguistic Manipulation of Serial Recall Tests

Serial recall tests are largely viewed as measures of VSTM and involve the immediate repetition of a string or list of verbal items. The to-be-recalled strings may consist of digits, letters, words, or nonwords. Serial recall and Sentence Repetition tests both require the immediate repetition of more than one item but they differ in the amount of linguistic information they carry. Serial recall lists are devoid of the morpho-syntactic relations found between sentence elements and are limited to list prosody. A further difference is the widely noted phenomenon that children and adults are able to repeat sentences that surpass the maximum length of lists they can accurately repeat in serial recall tests. Serial recall encompasses two types of test structures: span and fixed list length tests (Henry, 2011; Roodenrys & Stokes, 2001). Before addressing manipulation studies, the two types of serial recall tests are described with an emphasis on what we know so far about children's performance on span tests that included word, digit, and nonword subtests. This is followed by a listing of the types of linguistic manipulation that will be examined and the associated underlying assumptions, a table summary of the studies that will be presented, and the aim of the coming sections.

Span tests consist of lists of items that sequentially increase in length with an equal number of trials under each list length. Children are instructed to recall the exact items in correct order and testing is discontinued when a child exceeds the number of permitted errors within a list length. For example, a span test may consist of four trials under each list length and the threshold of error is set at two out of four trials; if the child repeats three trials incorrectly testing is discontinued. It is usually scored either by using a span score—equal to the maximum list length the child is able to repeat in correct order—or sometimes a trials score—equal to the total number of lists that a child is able to repeat in correct order across list lengths. This type of test is usually used to examine developmental change in VSTM capacity as well as comparing VSTM capacity between children with developmental disorders such as SLI (Archibald & Joanisse, 2009; Frizelle & Fletcher, 2015; Reichenbach, Bastian, Rohrbach, Gross, & Sarrar, 2016), dyslexia (Roodenrys & Stokes, 2001), and William's Syndrome (Carney et al., 2013) to Typically Developing controls. The most common type of span test is the digit span test, which is often incorporated in cognitive tests such as the Wechsler Intelligence Scale for Children-Third Edition (Wechsler, 1991) and the British Ability Scales II (Elliott, 1996).

Montgomery, Magimairaj, and Finney (2010) point out that there are two standardized tests of memory that are available for speech and language therapists: the Automated Working Memory Assessment (AWMA; Alloway, 2007) and the Working Memory Test Battery for

Children (WMTB-C; Pickering & Gathercole, 2001). Both tests include a VSTM component with three serial recall subtests comprising of different linguistic items: digit, word, and nonword. The primary purpose of a number of studies that administered these two standardized tests was to investigate group differences (age or language ability) and they did not statistically investigate the influence of item type on scores (Alloway, Gathercole, & Pickering, 2006; Frizelle & Fletcher, 2015; Injoque-Ricle, Calero, Alloway, & Burin, 2011; Nadler & Archibald, 2014). However, a consistent pattern emerged in the raw scores of the three subtests. Children were able to recall the greatest number of trials correct for Digit followed by Word and finally Nonword List Recall, and this was found to be true across different age groups, languages (English, Spanish), and language ability (SLI and controls). Nadler and Archibald (2014) administered the three subtests of AWMA to 178 Canadian children 5 to 9 years. The authors argued that the superiority of familiar items such as digits over nonwords indicated that long-term linguistic knowledge improves memory capacity. This was also found for a British sample of 708 children aged 4 to 11 years on the WMTB-C (Alloway et al., 2006) and a Spanish adaptation of the AWMA on a sample of 210 children aged 6 to 11 years (Injoque-Ricle et al., 2011). Finally, the pattern was not unique to Typically Developing children but also extended to English-speaking children with SLI. Frizelle and Fletcher (2015) found that while 32 children aged 7 years with SLI performed at a significantly reduced level in comparison to age-matched controls and younger Typically Developing controls (who were on average 2 years younger), the pattern of performance was similar in all three age groups. The Digit recall subtest of the WMTB-C was highest followed by Word List Recall and Nonword List Recall, respectively. To illustrate, in the SLI group the mean number of trials for the Digit recall subtest was 21.75 (SD = 3.12), for Word List Recall it was 13.41 (SD = 2.26), and for Nonword List Recall it was 9.56 (SD = 2.79). The difference between the studies that will be presented in this section and the studies mentioned above is that they purposefully manipulated various linguistic characteristics of items in serial recall tasks to examine how these influenced children's recall ability.

Fixed list tests, in contrast to span tests, consist of several trials of a constant number of items that varies depending on the age band of participants (usually just above span threshold for that age) and item type (longer list length for words in comparison to nonwords). The common scoring method used with this procedure is the percent of items recalled correctly, and usually the serial position of items is factored in. As the scoring method indicates, it is usually the test design of choice when serial position curves are examined. The design of the serial recall tests featured in this section was a mix of span and fixed list length and, at times, hybrids of both designs.

We will examine four aspects of linguistic manipulation at the word level in serial recall tests: lexicality (word vs. nonword), frequency (high vs. low), concreteness (concrete vs. abstract),

and imageability (high vs. low). The underlying assumption of the studies was that if children showed better recall ability for word over nonword strings, high over low frequency word strings, concrete over abstract word strings, or high over low imageability word strings the notion that long-term linguistic knowledge facilitates recall capacity would be supported, even in tests deemed to be assessments of VSTM. One possible alternative explanation for the pattern of findings was a difference in speech rate between the different linguistic conditions. For example, a greater number of words can be articulated within a given time frame in comparison to nonwords; this could account for the difference in recall ability rather than item type. Many studies featured here also examined speech rate to untangle its influence from that of long-term knowledge.

Table 3.1 provides an overview of the studies that will be presented in the subsequent sections. The table highlights the study sample, attributes of language that were manipulated, design of serial recall tasks, scoring method, and key findings. The aim of the following sections is to explore the pattern of performance according to the type of linguistic manipulation and compare the pattern of performance across different age groups, languages (English and French), and developmental disorders known to include VSTM difficulties (SLI and dyslexia) and those in which VSTM ability is intact (William's Syndrome and poor comprehenders).

Table 3.1 *Overview of Studies that Manipulated Linguistic Domains in Serial Recall Tests*

Study	Sample (mean age)	Stimuli highlighting linguistic manipulation	Scoring	Main findings
Roodenrys, Hulme, and Brown (1993)	English-speaking Typically Developing *n* = 24 equally divided between: • 5;7-6;3 (5;10) • 9;7-11;2 (10;4)	2x2 design • **Lexicality:** word & nonword • Item length: one-, two- & three-syllable • Closed item pool 8 for each condition • e.g.: bath, basket & butterfly gug, ballem & zegglepim • Controlled for frequency & phonotactic constrains • Span task with 4 lists at each length	**Span:** Maximum length with 4 out of 4 lists correctly repeated, plus .25 for every subsequent correct list	• Word > nonword • Older > younger • Short > long items • Lexicality effect same across age groups & word length (no sig. interaction) • Lexicality effect cannot be explained by speech rate
Gathercole, Pickering, Hall, and Peaker (2001)	English-speaking Typically Developing *n* = 16 • 8;3-8;11 (8;7)	2 x 2 design • **Lexicality:** word & nonword • List length: 4,5 & 6 items long • 120 each CVC words & nonwords e.g. jeep & neeb • 8 trials in each list length	Number of words/nonwords recalled correctly in correct serial position	• Word > nonword • Short > long lists • Sig. lexicality x list length due to poorer performance in long list lengths which reduced the extent of the lexicality effect
Majerus and Van der Linden (2003)	French-speaking *n* = 200 with 40 in each age group • 6 (6;5) • 8 (8;4) • 10 (10;5) • 13-16 (15;2) • 20-22 (20;8)	• **Lexicality:** CVC word & nonword o Matched for phonotactic frequency & controlled for word frequency • **Frequency:** bi-syllabic low & high frequency words o Frequency count < 200 & >10,000 • **Imageability:** low & high imageability o 1-6 rating scale: low < 3, high > 4 o Matched on length (1-3) syllables & frequency • Two lists of 108 items for each linguistic condition • Increasing length from 2 to 7 items with four trials under each length	Number of correctly recalled items in correct serial positions pooling over all sequence lengths	• Older > younger • Word > nonword independent of articulation rate • High > low frequency • High> low imageability • No sig. interaction with age for lexicality & frequency indicating the same pattern across age groups • Imageability showed a sig. interaction with age due to a lack of imageability effect in the adolescent group only
van der Lely and Howard (1993)	English-speaking • SLI *n* = 6 6;1-9;6 (7;2)	• **Lexicality:** 28 CVC each for words & nonwords • Controlled for word frequency and phonological similarity within lists	Number of correctly recalled lists with correct	• SLI = LM • Word > Nonword • No sig. interaction indicating both groups

	• LM *n* = 17 3;4-6;5 (5;2)	• Length of lists: 2-7 items with 4 trials in each list length	items in the correct serial position	equally benefit from lexical familiarity
Roodenrys and Stokes (2001)	English-speaking *n* = 48, 16 each • Dyslexia 7;9-9;7 (8;11) • AM 7;11-9;4 (8;7) • Reading Level matched 6-6;10 (6;3)	• **Lexicality:** • Closed set of 8 monosyllabic words and nonwords • Length of lists commenced at 2 items with sequential increase of length by 1 item • 4 lists in each length • Testing continued until participant incorrectly repeated two or more lists within an item length	**Span:** Maximum length with 4 out of 4 lists correctly repeated, plus .25 for every subsequent correct list	• AM > Dyslexia = reading level matched • Words > Nonwords independent of speech rate • No sig. interaction, groups equally affected by lexicality. • Neither group or lexicality could be explained by speech rate
Henry and Millar (1991)	English-speaking Typically Developing *n* = 24, equally divided between: • 5;1-5;10 (5;6) • 7;1-8;6 (7;8) Half boys & half girls in each age group	2 x 2 design • **Frequency:** 2 conditions ○ High frequency + familiar ○ Low frequency + unfamiliar • **Item length:** one- vs. three- four- syllables • Closed item pool, 15 in each condition • e.g. bed, policeman vs. plough, dowager • Starts at 2 word length with sequential increase • Each length contains 3 lists	**Span:** Maximum length with 2 out of 3 lists correctly repeated, plus .5 for if child repeats 1 correct out of 3 trials above span	• Older > younger • Gender not sig. • High frequency & familiar > Low frequency & unfamiliar • Short > long • Sig. familiarity x length interaction, with length effect more evident in unfamiliar words • No sig. age x familiarity • Findings cannot be explained by identification time or articulation rate
Coady, Mainela-Arnold, and Evans (2013)	English-speaking *n* = 32, 16 each • SLI 8;7-11;9 (10;2) • AM 8;5-12;3	• **Frequency:** high vs. low frequency CVC words • Pool of 320 words controlled for familiarity • e.g. car, rose vs. whif, shod • Starts at 2 word length with sequential increase until length of 6 items • 2 lists in each length and children attempted all lengths	Percent of words correctly repeated irrespective of order	• Collapsed across list length and frequency: AM > SLI • Collapsed across frequency: ○ Short > long list ○ Sig. list length x group SLI more influenced by list length • Collapsed across list length: ○ High > low frequency ○ No Sig. group x frequency ○ Effect of frequency disappeared when vocabulary was co-varied
Nation, Adams, Bowyer-	English-Speaking *n* = 32 equally divided • Poor Comprehenders	• **Concreteness:** Concrete vs. abstract • Closed item pool with 16 words in each condition e.g., tooth, plate vs. luck, pride	Response scored correct if the correct	• Poor = good comprehenders • Serial position sig. showing recency and primacy effects

Crane, and Snowling (1999)	(9.33) • Good Comprehenders (9.22) Matched for decoding skills	• Words matched for frequency • Fixed list length: 5 words and 3 nonwords • 16 trials for each condition	word is recalled in the correct serial position	• Concrete > abstract • Sig. group x concreteness with abstract lists more difficult for poor comprehenders • No sig. concreteness x serial position, indicating effect of concreteness present irrespective of word position • Concreteness effect independent of speech rate

Cain (2006)	English-Speaking
	n = 26 equally divided
	• Poor Comprehenders (9;7)
	• Good Comprehenders (9;6)
	
	Matched for word reading level, vocabulary knowledge & age
		• **Concreteness:** Same as above (Nation et al. 1999)	• Accuracy same as above.
			• **Error analysis:**
			  ○ Missing Word
			  ○ Position Error
			  ○ Intrusion Error
				• Poor = good comprehenders
				• Serial position sig. showing recency and primacy effects
				• Concrete > abstract
				• No sig. interactions
				• Position errors:
				• Concrete > abstract
				• No sig. group or group x word
				• Intrusion errors
				• Concrete words > abstract
				• Poor > good comprehenders
				• No sig. group x word
				• Missing errors more common for abstract words

Laing et al. (2005)	English-Speaking *n* = 42 equally divided • William's syndrome 10;11 – 52;1 (21;7) • Digit Span matched 5;1-40;5 (9;2) • Vocabulary matched 6;1-40;5 (10;9)	• **Concreteness:** Concrete vs. abstract • Closed item pool with 8 words in each condition e.g. key, lamp vs. joke, love • Matched for familiarity and frequency • List length: 2 to 8 items • 4 lists at each length • Testing stopped when all 4 lists at a particular list length were incorrectly repeated	**Span:** Maximum length with 2 out of 4 lists correctly repeated, plus .25 for every subsequent correct list	• No group effect • Concrete > abstract • No sig. group x concreteness • Concreteness effect independent of speech rate

3.1.1 Lexicality effect

Of the four linguistic attributes featured in this section, lexicality effect is the most widely studied in children. This is reflected in Table 3.1, with five of the 10 studies examining its effect. Roodenrys et al. (1993) compared the memory span of 6- and 10-year-olds on strings of items that were manipulated for lexical status (words vs. nonwords) and length (one- vs. two- vs. three-syllables). Items within a single list were similar with regards to item type and length. While the 10-year-olds obtained higher span scores and quicker speech rates, both groups showed superior recall of strings of words (e.g., bath, basket, and butterfly) over nonwords (e.g., gug, ballem, and zegglepim); this was irrespective of item length. The effect of lexicality and age remained significant even when speech rate was statistically partialed out as a covariate. Speech rate and span showed a similar linear relationship for word and nonwords in both age groups. There was a higher intercept for words in comparison to nonwords and for older in comparison to younger children. The older children showed a slightly greater advantage for words over nonwords. Taken together, the authors argued that the pattern of findings supports the contribution of established linguistic knowledge, with both age groups showing a benefit of lexical familiarity that could not be explained by speech rate. The authors suggested that the higher intercepts for speech rate/span in older children indicated that linguistic knowledge might also partly explain the developmental improvement in span scores.

Using a hybrid of fixed list length and span task design, Gathercole et al. (2001) arrived at the same conclusion, further supporting the contribution of established linguistic knowledge. Sixteen 8-year-old children were presented with lists of monosyllabic words and nonwords at three list lengths (four, five, and six). Results showed a reduction in recall accuracy as list length increased and that the percent of correctly recalled nonwords was less than half of the proportion of correctly recalled words. The findings also showed a reduction in the magnitude of the advantage of words over nonwords as list length increased, which the authors attributed to the extremely low levels of performance in the longer list lengths rather than a reduced lexicality effect. Majerus and Van der Linden (2003) extended the lexicality effect findings to French across a wide age range. Two hundred French-speaking participants aged 6 to 20 years (divided into five age groups) were presented with lists of monosyllabic words and nonwords. The number of correctly recalled items in correct serial position showed a developmental trend with older participants obtaining higher scores. A significantly higher number of words were recalled in comparison to nonwords. The advantage of words over nonwords was present in all five age groups investigated, indicating that participants benefited from lexical familiarity irrespective of age. In spite the quicker articulation rate for words, the lexicality effect and the effect of age remained significant when articulation rate was partialed out.

The superiority of words over nonwords was also evident in children with SLI and dyslexia using span recall tasks. van der Lely and Howard (1993) found that 7-year-olds with

SLI and their expressive and receptive language-matched controls recalled a significantly greater number of lists consisting of monosyllabic words in comparison to nonwords when language ability was covaried out of analysis. Recall ability was similar in magnitude in both groups with no significant interaction between group and list type. The authors argued that the findings indicated that lexical knowledge facilitated recall ability in both language groups. Roodenrys and Stokes (2001) compared the serial recall ability in three groups of 16 children: 9-year-olds with dyslexia, age-matched controls, and 6-year-old reading-ability-matched controls. Span scores were equal in magnitude in children with dyslexia and their reading-ability-matched controls. However, both groups obtained lower span scores in comparison to the age-matched control group. All three groups showed superior recall for word strings with an equal degree of word span advantage in all three groups. This was evident from the lack of significant interaction between group and list type. Consistent with studies discussed earlier, controlling speech rate did not eliminate the lexicality or group effect, which led the authors to conclude that long-term linguistic knowledge contributes to serial recall performance and increases in magnitude with age. To summarize, the studies provided concurring evidence that lexical knowledge facilitated children's recall ability irrespective of age and language spoken with similar patterns observed across groups of children with varying language, VSTM, or reading abilities. The lexicality effects could not be explained by the quicker speech rate of words and showed evidence that the contribution of linguistic knowledge increased with age.

3.1.2 Frequency effect

Word frequency is defined as the amount of usage of a word in spoken or written language (L. M. Miller & Roodenrys, 2009). Henry and Millar (1991) examined whether frequency (low vs. high) and length (one syllable vs. three/four syllable) of words influenced the recall ability of 5- and 7- year-old children. Unlike studies above that partialed out the effect of speech rate using analysis of co-variance, the influence of speech rate was directly examined in two experiments that differed in which word types were matched and how they measured speech rate. The first experiment matched 5- and 7-year-old participants on speech rate for short words (high and low frequency) and speech rate was calculated based on one repetition of each word. The second experiment matched the children in the two age groups on their speech rate for short and long words and calculated speech rate based on three repetitions of each word. The main reason behind including speech rate in the study design was to examine whether matching children on speech rate would cancel out the impact of age on span.

Both experiments showed that age (older > younger), frequency (high > low), and length (short > long) influenced span but differed in how frequency interacted with age. In the first experiment, both groups were similarly affected by word frequency with high frequency words (e.g., one syllable: bed, hat; three/four syllable: policeman, banana) being easier to recall than low frequency words (e.g., one syllable: debt, gill; three/four syllable: boundary, soprano). In the second experiment, only low frequency words differed significantly between the two age

groups. The advantage of high frequency words in both experiments irrespective of item length, the stimuli that were matched or how speech rate was measured further supports the contribution of established linguistic knowledge to serial recall tasks. The superiority of older children, even when they were matched with younger children on speech rate, led the authors to postulate that the developmental trend in span scores was partly due to older children being more familiar with words in general and can therefore rely more on their established linguistic knowledge. This is consistent with the explanation put forward by Roodenrys et al. (1993) regarding the superiority of older children when examining lexicality effects and speech rate.

Majerus and Van der Linden (2003) extended the findings of Henry and Millar (1991) to French-speaking participants. As with lexicality effect, the total number of high frequency words exceeded that of low frequency words, and the pattern of performance was consistent in all five age groups examined. Coady et al. (2013) found that the superiority of high frequency words also held true for English-speaking children with SLI and their 10-year-old age-matched peers despite lower scores of children with SLI. While the disparity between children with SLI and their age-matched controls remained when vocabulary score was covaried out, the frequency effect disappeared. The authors argued that this finding further supported the role of long-term knowledge in serial recall and postulated that the advantage of high frequency words was not due to the knowledge of individual words but rather due to the contribution of broader linguistic knowledge. To summarize, the frequency effect was found in English- and French-speaking participants as well as children with SLI. The pattern of performance was similar across age groups and different language ability groups even when speech rate was controlled for, further supporting the contribution of established linguistic knowledge to serial recall tests.

3.1.3 Concreteness and imageablity effects

Concreteness can be defined as the degree to which a word has a tangible referent (Coady et al., 2013). For example, “table” and “chair” are highly concrete words, while “joke” and “love” are abstract words. Imageability refers to the degree a word can conjure up a mental visualization (Friendly, Franklin, Hoffman, & Rubin, 1982). For example, “market” and “money” are high imageability words, while “remain” and “union” are low imageability words (Friendly et al., 1982). Lists of concrete words are easier to recall than lists of abstract words (Cain, 2006; Laing et al., 2005; Nation et al., 1999) and lists of high imageability words are easier to recall than low imageability words (Majerus & Van der Linden, 2003) further supporting the contribution of established linguistic knowledge to serial recall.

The focus of the concrete/abstract disparity studies has been largely on two groups of children known to have VSTM abilities that are comparable to their age-matched peers: children with specific reading comprehension difficulties and children with William’s syndrome (Carney et al., 2013; Oakhill, Yuill, & Parkin, 1986). Poor comprehenders are characterized by a weakness in semantic and syntactic abilities and since concreteness taps into children’s semantic ability, researchers were interested to see whether recall of abstract words

was disproportionately affected in poor comprehenders in comparison to controls. The picture is less clear in children with William's syndrome, with some studies pointing towards a weakness in semantic knowledge (e.g., Vicari, Carlesimo, Brizzolara, & Pezzini, 1996) while others characterize semantic knowledge as a relative strength (Tyler et al., 1997).

Nation et al. (1999) compared the recall ability of 16 children with specific reading comprehension difficulty to 16 children matched on chronological age, nonverbal ability, and decoding ability (reading nonwords). As expected, no group difference was found. Both groups showed superior recall for concrete words. However, a significant group x concreteness interaction showed that recall of abstract lists was disproportionately weaker in poor comprehenders. The pattern of performance remained the same even when speech rate was used as a covariate, showing that speech rate alone could not explain the pattern of findings and that poor comprehenders' weak semantic ability could be the reason for the weaker recall of abstract lists. Cain (2006) replicated the lack of group difference between poor comprehenders and controls on overall recall ability as well as the superior recall for concrete words. The findings of the two studies diverged with regard to a lack of difference in abstract recall between participants with William's Syndrome and controls. Cain (2006) postulated that the difference may be due to a more stringent matching criteria in her study, which included vocabulary knowledge or due to heterogeneity in the weakness profile exhibited by the poor comprehenders population with some, but not all, showing weakness in semantic ability.

Laing et al. (2005) compared the performance of 14 participants with William's syndrome aged 10 to 52 years to digit span and vocabulary matched controls on lists of abstract and concrete words. Span scores were comparable across the three groups and all obtained higher span scores for lists of concrete words in comparison to lists of abstract words with no group x list type interaction. The advantage of concrete lists remained even when speech rate was a covariate. Interestingly, upon examining speech rate (number of words per second) participants were able to repeat abstract words faster than concrete words. This was a surprising finding since speech rate and span are usually positively correlated. The authors argued that had abstract and concrete words been matched for speech rate, the advantage of concrete words would be even greater without providing a possible explanation for the pattern of findings.

Majerus and Van der Linden (2003) examined the imageability effect in five age groups of French speaking-participants aged 6 to 22 years. The number of words recalled increased with age, and high imageability word lists were recalled better than low imageability word lists in four of five age groups (adolescents being the only exception). To summarize, studies found significant concreteness and imageability effects largely independent of speech rate. Overall, the pattern was similar irrespective of presence or absence of developmental disorder or age groups investigated.

In conclusion, manipulation of all four linguistic attributes influenced recall showing significant lexicality effects (word > nonword), frequency effects (high > low frequency),

concreteness effects (concrete > abstract), and imageability effects (high > low imageability). The overall pattern of performance was similar irrespective of age, language spoken (English or French), or underlying developmental disorder with VSTM impairment (SLI and Dyslexia) or without VSTM impairment (poor comprehenders or William's syndrome). These results held whether the serial recall task was Span, Fixed List Recall, or a hybrid of both. Majerus and Van der Linden (2003) was the only study that compared the effects of more than one attribute, with lexicality showing the largest effect size followed by frequency with medium effect size and imageability with small effect size.

All the studies that investigated speech rate found that it was insufficient to explain the profile of performance irrespective of how they took count of speech rate, whether they statistically controlled for the effect of speech rate, or matched participants of different age groups on speech rate. Laing et al. (2005) found that speech rate would have predicted the opposite direction of findings with articulation rate of abstract words found to be quicker than concrete, demonstrating that the linguistic attribute of concreteness can supersede the influence of speech rate. Therefore, the findings indicated that even in the absence of morpho-syntactic relations, established linguistic knowledge facilitates serial recall.

Redintegration is postulated to be the cognitive mechanism underpinning the contribution of linguistic knowledge in serial recall (Hulme, Maughan, & Brown, 1991). Through this process, linguistic knowledge is called upon to reconstruct or clean up decaying traces of stimuli in VSTM. In the case of the lexicality effect, the decaying traces of words can be compared to a richer lexical and phonological store compared to nonwords. This is not to say that recall of nonwords is not facilitated by established linguistic knowledge, but may benefit to less of a degree than word recall (Gathercole et al., 2001). As detailed in Chapter 2, recall of single item nonwords comprising of attested syllables in Cantonese was found to be superior to that of nonwords with unattested syllables (Stokes et al., 2006). Although sub-lexical effects were not the focus of this section, it is notable that wordlikeness effects have been replicated in serial recall (Gathercole et al., 2001; Roodenrys & Stokes, 2001).

3.2 Syntactic Manipulation of Well-formed Sentences

Most of the Sentence Repetition tests featured thus far, like the CELF Recalling Sentences subtest (Semel et al., 2003), simultaneously manipulated grammatical complexity and length. By virtue of design and the use of purely quantitative scoring, tests like the Recalling Sentences subtest are unable to pinpoint whether certain verb-argument structures are more difficult to repeat than others (e.g., passive vs. active sentences), whether certain morpho-syntactic categories are more difficult to repeat than others (content vs. function words), and whether there are any differences within a category (noun vs. adjective vs. verb). Finally, they are unable to untangle whether length and grammatical complexity have independent effects and whether one has more influence on repetition than the other. There have been few exceptions with regard to test design, but the analyses did not investigate the different stimuli

categories. For example, the Redmond (2005) test kept the length constant and consisted of an equal number of passive and active sentences; however, they did not report the scores of the two sentence types separately to allow for a comparison between them. The studies featured in this section are characterized by the systemic manipulation of syntax. They can be broadly categorized into studies that investigated the effects of syntax while controlling for length and studies that simultaneously manipulated syntax and length. The underlying assumption of the manipulation of syntax is that a difference in Sentence Repetition scores based on syntactic condition would indicate that linguistic knowledge supports Sentence Repetition. For studies that manipulated length as well, the underlying assumption is that a difference between repetition scores of short and long sentences would indicate that Sentence Repetition is supported by VSTM too. The studies focusing solely on syntactic manipulation are presented first followed by studies that examined the influence of syntax and length.

3.2.1 Effects of syntax when length was controlled

Diessel and Tomasello (2005), Frizelle and Fletcher (2014), and Riches et al. (2010) investigated the influence of syntactic manipulation of relative clause constructions. All three studies found that Sentence Repetition scores were influenced by syntactic condition. Tables 3.2-3.4 provide an overview of study samples, sentence design, scoring with an example of stimulus sentences for each syntactic condition, and the key findings of each study according to scoring type. Scores were ether purely quantitative providing a cumulative error/accuracy score or qualitative indicating the most common error type.

Diessel and Tomasello (2005) investigated the influence of syntactic condition on the repetition scores of Typically Developing English-Speaking and German-Speaking 4-year olds (see Table 3.2). Two key features of relative clause constructions were manipulated: (1) the number of propositions (one or two) as determined by whether the relative clause was attached to a predicate nominal of a copular verb or the direct object of a transitive verb, and (2) the syntactic role of the post-modified noun in the relative clause via the relative pronoun. A total of six syntactic roles were investigated: subject relative with an intransitive verb, subject relative with a transitive verb, object relative, indirect object relative, oblique relative (post-modified noun functions as the object of a prepositional phrase), and genitive relative (the relativizer *whose* replaces a genitive noun). To help illustrate these are two examples from Diessel and Tomasello (2005):

1. This is the [boy] $_{\text{subject}}$ *who played in the garden yesterday* (The boy played in the garden yesterday.)
2. Mary fed the [cat] $_{\text{direct object}}$ *that the dog chased around the tree* (Mary fed the cat. The dog chased the cat around the tree)

In example 1, the sentence contains a single proposition and can be paraphrased using a single sentence (shown above between parentheses), the post modified noun (boy) is the predicate of the copular verb (is) in the main clause and acts as subject of the intransitive verb

play in the relative clause via the pronoun who. In example 2, the sentence contains two propositions and requires two sentences to paraphrase it; the post-modified noun (cat) is the direct object of the transitive verb (fed) in the main clause and acts as the direct object as well of the verb (chased) in the relative clause.

German relative clauses were manipulated in a similar fashion. German and English relative clauses are introduced by a relative pronoun but they differ in the amount of information provided by the relative pronoun (Diessel & Tomasello, 2005). The German relative pronoun is marked for gender, number, and case. The syntactic role of the relative clause can be gleaned from the case marking of the relative pronoun. For example, (den) is a masculine, singular accusative relative pronoun.

In spite of the differences in structure between English and German relative clause constructions, the influence of the number of propositions, the syntactic role of the relative clause, and the direction of syntactic role conversion when errors were made paralleled in both languages. Single propositions were easier to repeat than dual-propositional relative clauses. The pattern of syntactic role difficulty were as follows: subject relative clauses with an intransitive verb were easiest to repeat followed by subject transitive and direct object relative clause, respectively; genitive clauses were the most difficult to repeat. Direct object, indirect object, and oblique relatives were frequently converted to subject relatives (transitive or intransitive); conversions in the opposite direction were far less frequently observed. The findings diverged when it came to the type of errors committed, reflecting the morphological richness of German in comparison to English. English-speaking children committed word order errors while German-speaking children committed case marking errors. The authors postulated that while the findings in both English and German points towards the contribution of linguistic knowledge the type of linguistic knowledge varies between languages. English-speaking children relied primarily on word order knowledge; German-speaking children relied more on morphological knowledge, particularly case markings.

Table 3.2 *Summary of Diessel and Tomasello (2005)*

Sample English-speaking: 21 TD mean age 4;7
German-speaking: 24 TD mean age 4;7

Stimuli 2 x 2 design

- number of propositions/attachment:
  - single propositional: predicate nominal of a copular matrix
  - dual propositional: direct object of a transitive matrix
- syntactic role 6 levels: subject intransitive, subject transitive, direct object, indirect object, oblique, genitive
- genitive excluded due to floor effect in German

n = 24, 4 for each syntactic role the main clause of the sentence was distributed as:

- 3 with copular main clause with 2 declarative and one question
- 1 transitive main clause with direct object as head of the relative clause

Controlled for:

- length: words 11/12, syllables 13/14
- semantic factors: all nouns were animate

Example

Syntactic Role	**Attachment**	
	Single Proposition	**Dual Proposition**
Subject intransitive	There is the boy who played in the garden yesterday.	Peter saw the woman who sat on the bench this morning.
Subject transitive	This is the man who saw Peter on the bus this morning	Mary heard the dog that scared the little cat last night
Object direct	There is the cat that Mary fed in the kitchen last night	Mary fed the cat that the dog chased around the tree
Object indirect	This is the girl who Peter borrowed a football from.	Peter talked to the woman who Mary showed her bike to
Oblique	There is the horse that the little cat jumped on yesterday	Peter spoke to the man who Mary danced with last night
Genitive	This is the woman whose cat caught a mouse yesterday	Mary looked for the man whose cat Peter found in the house

Scoring Each sentence was awarded 1, 0.5 or 0

- 1: correct or changes that did not affect structure or content e.g. tense, number or definiteness.
- 0.5 minor lexical/grammatical errors that did not affect relative clause structure such as lexical substitution, omission of determiner or substitution of relativizer
- 0: incomplete, no response, structure or meaning of sentence changed or ungrammatical

Error Analysis: Direction of conversion errors and type

Key Findings

Quantitative

Syntactic complexity (Single vs. Dual Proposition):

- Relative clause attached to a predicate nominal in a copular matrix > relative clause attached to a direct object in a transitive matrix
- Same for English & German

Syntactic complexity (syntactic role of relative clause)

- In both languages
  - Subject intransitive > subject

Qualitative

Conversion:

- Direct object, indirect object and oblique relatives converted to subject relatives.
- Subject transitive less frequently converted to indirect object relatives.

Error Type:

	transitive o Subject transitive > object direct o Genitive most difficult Language & syntactic role: • No main effect of language (English = German) • No interaction (language x syntactic role) • Only oblique relatives differed significantly between languages, worse in German. \	• English: change in word order • German: substituting relative pronoun case marking or other case marked elements

Frizelle and Fletcher (2014) extended the findings of Diessel and Tomasello (2005) to school-aged children with SLI and their age-matched (AM) controls. Their ages ranged between 6;0 and 7;11 years. An additional younger Typically Developing (YTD) group (2 years younger, on average) was included in the study. As in the Diessel and Tomasello (2005) study, the number of propositions (one or two) and the syntactic roles of relative clause constructions were manipulated. The number of syntactic role conditions increased by two because of the subdivision of direct object relatives to include object relatives with an inanimate head and a personal pronoun as the subject of the relative clause and object relatives with an animate head and the subdivision of genitive relatives to include genitive subject and object (see Table 3.3).

Frizelle and Fletcher (2014) found that while children with SLI performed at a reduced level in comparison to AM and YTD groups in the different syntactic conditions, the profile of performance was broadly similar in all three groups. In line with Diessel and Tomasello (2005), single proposition relatives were easier to repeat than dual-propositional relatives. Dual-propositional relatives were disproportionately difficult in children with SLI. The average difference between single and dual-propositional phrases was greatest in children with SLI (36.8 points), which was three times greater than the average difference in the AM group and almost two times greater than the difference observed in the YTD group. With regard to the level of difficulty according to the syntactic role of relative clause constructions, findings again corresponded to that of Diessel and Tomasello (2005). Subject relatives with an intransitive verb were easiest to repeat followed by subject relatives with a transitive and direct object relatives when considered overall. When direct object relatives were limited to those with an inanimate head, the advantage of subject intransitive and transitive relatives disappeared with equal levels of repetition accuracy for inanimate object and subject verb intransitive relatives and superior repetition accuracy for object inanimate in comparison to subject transitive. Finally, indirect object, oblique, and genitive relative clauses were extremely difficult for children with SLI to repeat. Taken together, the findings of both studies suggest that established linguistic knowledge contributes to the repetition accuracy of children with SLI and Typically Developing children. The type of linguistic knowledge primarily tapped by the Sentence Repetition tests can very according to the language of participants, word order for English-

speaking participants, and morphological knowledge for German-speaking participants. The type of linguistic knowledge in English can also vary according to syntactic condition. Frizelle and Fletcher (2014) proposed that the difference in ease of repetition between subject and direct object relatives on one hand and indirect object, oblique and genitive in the other might be attributed to non-canonical word order (syntactic knowledge). However, they point out that structural configuration fails to explain the superiority of object inanimate over object animate relatives (same structural configuration: Object Subject Verb [OSV]), and superiority over subject transitive (easier configuration: Verb Subject Object [VSO]) and the lack of difference between subject intransitive and object inanimate (VSO vs. OSV). They postulate that in this case, lexical choice (semantic knowledge) overrides syntactic knowledge and processing load. Finally, VSTM cannot fully explain the pattern of findings since both studies controlled for length.

Table 3.3 *Summary of Frizelle and Fletcher (2014)*

Sample	English-speaking • 32 SLI mean age 6;10 • 32 AM mean age 6;11 • 20 TDY mean age 4;9 2 years younger not language matched		
Stimuli	2x2 design • number of propositions/attachment: o single propositional: predicate nominal of a copular matrix o dual propositional: direct object of a transitive matrix • syntactic role of the relative clause with 8 levels: subject intransitive, subject transitive, object indirect, object, oblique, genitive subject, genitive object *n* = 52, 14 conditions, 17 filler simple active sentences controlled for length: 10-13 syllables		
Example	**Syntactic Role**	**Attachment**	
		Single Proposition	**Dual Proposition**
	Subject intransitive	This is the bird that slept in the box all night.	The girl cleaned up the milk that spilt in the fridge.
	Subject transitive	There is the sheep that drank the water this morning.	Eddie met the girl who broke the window last week.
	Object animate	There is the boy that Emma helped in the kitchen.	The boy rode the horse that Anne put in the field.
	Object inanimate	There is the picture that you drew on the wall last week.	The girl ate the sweets that you brought to the party.
	Indirect object	There is the dog that the man kicked his football to.	Anne fed the baby who Emma sang a song to.
	Oblique	There is the tree that the car crashed into last night.	Anne painted the picture that the girl looked at today.
	Genitive subject	There is the girl whose juice spilt in the kitchen.	Anne saw the farmer whose cow fell in the shed.
	Genitive object	There is the girl whose toy Anne broke in the garden.	Emma met the girl whose bag Anne took to school.
Scoring	Syntactic Accuracy score: each sentence awarded a score from 10-0. For example: • 10 = accurate repetition • 9 = accurate syntax with lexical substitution • 8 = inflectional error Total Syntactic Accuracy: summation of the syntactic accuracy sores		
Key Findings	**Quantitative** Group (Total Syntactic Accuracy scores)		

- SLI < TDY < AM
- Percentage of responses that obtained a perfect score of 10 (Syntactic Accuracy)
    - 3 % SLI
    - 21% TDY
    - 51% AM

Syntactic complexity (single vs. dual proposition):

- Single > dual proposition
- Pattern similar in all 3 groups
- Greatest difference between sentence types seen in children with SLI (36.8 points) TDY (22.5 points) and AM (11.3 points)

Syntactic complexity (syntactic role):

- Pattern of performance similar in the three language groups
- Subject relatives were easiest > indirect object, genitive subject & genitive object
- Subject intransitive > subject transitive
- Intransitive subject > object overall
- Intransitive subject = object inanimate
- Object inanimate > subject transitive

Riches et al. (2010) examined the influence of syntactic condition on the repetition of 14 adolescents with SLI and compared it to 16 participants with combined Autism and Language Impairment (ALI) and 17 age-matched peers (Table 3.4). A total of four conditions were explored: two relating to the syntactic role of the relative clause (subject vs. object) and two relating to the position of an added adjective (main vs. relative clause). All the sentences were dual-propositional. Both clinical groups obtained comparable overall error rates that were higher than their age-matched controls. The overall pattern was similar in both clinical groups; sentences with object relatives and adjectives in the relative clause were more difficult to repeat. However, the pattern was more apparent in adolescents with SLI. This was reflected in three findings: (1) the influence of syntactic role and adjective position were only statistically significant in the SLI group and not the ALI group, (2) a significant group by syntactic role interaction, and (3) the frequency of object to subject relative transformations was significantly higher in participants with SLI. The influence of the syntactic role of relatives particularly in adolescents with SLI and the simplification of object relatives indicate that syntactic knowledge was tapped by Sentence Repetition. Since length was controlled in all conditions, VSTM cannot fully explain pattern of findings. As the authors pointed out, it is not immediately evident as to why participants with SLI were more influenced by syntactic condition. One possible explanation put forward was that the syntactic deficit was milder in participants with ALI; therefore, they were more able to access syntactic knowledge to facilitate repetition. With regard to the influence of adjective position independent of the syntactic role of the relative clause, a number of explanations were proposed: the longer syntactic dependency where the thematic role of the subject is not assigned until the main verb increasing the load on working memory, a combination of the semantic abstractness of adjectives and their serial position in the sense that the increased difficulty of adjectives had a domino effect on the accuracy of repetition of sentence elements that came after it or that adjectives in the relative clause are pragmatically unusual.

To conclude, when length was controlled the influence of syntactic condition was present irrespective of language typology (English vs. German), age of participants with participants as young as 4 years old to adolescents, and language ability (SLI vs. AM). The consistency of findings along with qualitative scoring showing that the direction of transformations tended to be from object relatives to subject relatives implicates the involvement of syntactic knowledge. As to why certain syntactic conditions were more difficult than others or the difference in type of errors observed in English or German, different explanations were put forward implicating different types of linguistic knowledge: morphological, syntactic, lexical, semantic, and pragmatic.

As detailed in previous sections on differential diagnosis, Sentence Repetition is a universal marker of language impairment. Riches et al. (2010) study showed that one way to overcome the lack of differentiation by Sentence Repetition for language impairment with different underlying developmental disorders can be improved by the syntactic manipulation of stimuli and the use of qualitative error analysis.

Table 3.4 *Summary of Riches et al. (2010)*

Sample	English-speaking • 14 SLI mean age 15;3 • 16 Autism +LI mean age 14;8 • 17 AM mean age 14;4		
Stimuli	2x2 design, all relative clause • attachment: subject/object • adjective position: main clause/relative clause *n* = 26, 6 in each condition controlled for: • length in phonemes • lexical frequency • plausibility • semantic properties of main clause		
Example	**Attachment**	**Adjective position**	
		Main clause	**Relative clause**
	Subject	The monster that killed the prince wore a bright green cloak	The boy that kicked the big old donkey wore a hat
	Object	The soldier that the criminal shot wore a bright green hat	The granny that the tall thin thief robbed wore some shoes
Scoring	• **Quantitative:** Levenshtein Distance in words (LDw): an algorithm which counts the minimum number of words that must be added, substituted or omitted to transform one sentence into another • **Qualitative**: Error based • Transform object relative to subject relative • Incomplete/omitted relative clause or incomplete/null response		
Key Findings	**Quantitative** • Group: o SLI = ALI > AM AM: low error rate, excluded from results below • Complexity (separate analysis within SLI & ALI groups): o Object relative > subject relative		**Qualitative** • Transformation of object to subject relative clause occurred significantly more in participants with SLI in comparison to ALI mean (16.4 vs. 6.78) • Transformation of subject to object relatives rarely

○ Adjective in relative clause > Adjective in main clause ○ Both significant in SLI only • Complexity (subject vs. object relative) and group (SLI vs. ALI) ○ Complexity: • Object > subject ○ Interaction (group x complexity) Object relative more vulnerable in SLI	occurred. • Incomplete/omitted relative clause and incomplete/null response was similar in both groups.

3.2.2 Effects of syntax when length was manipulated

Three studies simultaneously investigated the influence of syntactic manipulation and length on Sentence Repetition scores: Willis and Gathercole (2001), Wilsenach (2006) and Moll et al. (2015). Consistent with the findings of studies that solely manipulated syntax, all three studies found that Sentence Repetition scores were influenced by syntactic condition. The picture is less clear when it comes to the manipulation of sentence length and how to interpret the findings. The studies presented here will be discussed from two angles; first the findings of syntactic manipulation will be presented followed by the findings of length manipulation. Tables 3.5-3.8 summarize each study with a focus on study sample, sentence design, scoring method, example of stimulus sentences for each syntactic condition, and the key findings.

Willis and Gathercole (2001) investigated the repetition ability of 30 Typically Developing children 4 to 5 years old. The syntactic conditions comprised of six sentence types derived from the Test for the Reception of Grammar (TROG; Bishop, 1982) and contained the following constructions: in/on, above/below, reversible active, X-but-not-Y, embedded sentences, and relative clause. Not surprisingly, syntactic condition significantly impacted repetition accuracy with embedded sentences and relative clauses being the most difficult to repeat. Length was manipulated at the syllable level for content words; short sentences consisted of mono-syllabic content words, whereas long sentences consisted of two-three-syllable content words. Long sentences were significantly more difficult to repeat. However, the significant interaction between syntactic condition and length indicated that the influence of length was only true for two sentence types: those containing in/on and X-but-not-Y. The findings showed that even when length was manipulated, the contribution of linguistic knowledge superseded that of VSTM.

The authors explored possible reasons for the lack of influence of length on the remaining four sentence types. They argued that floor effects in embedded and relative clause sentences might have masked the effect of length. However, that still fails to explain the lack of difference observed for sentences that contained reversible passives and above/below. Another explanation was that the difference in syllable length between short and long sentences was not powerful enough. Long sentences contain an average of 40% more syllables than short sentences. The authors stated that 40% syllable length difference was far less than the 200-400% difference used in studies comparing span scores for short versus long word lists. This

highlights a design dilemma: what constitutes an acceptable difference in length between short and long sentences? Does the difference between sentences need to be as large as the difference in span/serial tasks to influence repetition scores? Would the same difference hold true for syllables and words in repetition tasks be it serial recall or Sentence Repetition? Furthermore, while the difference between short and long sentences was largely similar across sentence types, the starting point for syllable length in the short condition varied. For example, the mean number of syllables for short and long sentences containing in/on was 6.50 versus 9.75, but for relative clause was 9 versus 12.75, respectively. This highlights the difficultly of controlling multiple variables at once.

In a follow-up experiment to further examine the contribution of VSTM, 61 reception class children were screened on CNRep and a digit span test. A total of 13 participants with low VSTM ability and 13 with high VSTM ability were identified and asked to repeat 16 different sentence types derived from TROG (Bishop, 1982). As with the first experiment, sentence type significantly influenced repetition scores. Although participants with low VSTM ability performed at a reduced level, a significant group x sentence type interaction was found indicating that VSTM ability influenced just over half the sentence types, again showing that VSTM cannot fully explain pattern of findings with regard to sentence type for the low VSTM group. A major caveat of this experiment, as the authors point out, was that only VSTM and nonverbal IQ were accounted for while language ability was not tested. Therefore, the difference between group performance could be due to a number of different factors and not just VSTM ability.

Table 3.5 *Summary of Willis and Gathercole (2001) Experiment 1*

Sample
- English-speaking 30 Typically Developing mean age 4;6

Stimuli

2x2 design
- Sentence type: 6 types, from Test for Reception of Grammar (TROG) (Bishop, 1982): in/on, above/below, reversible passive, X-but-not-Y, embedded sentence, relative clause
- Length: short/long
- Length increased by increasing the number of syllables in nouns and adjectives when possible.
  - Short: 6.5 –9 syllables, one-syllable nouns and adjectives.
  - Long: 9.75-12.75 syllables, 2/3 syllable nouns and adjectives.
- *n* = 48 4 per condition

Example

Sentence Type	**Length**	
	Short	**Long**
In/on	The cup is in the box	The water is in the bottle
Above/below	The ball is above the cup	The banana is above the flower
Reversible passive	The fox is chased by the horse	The rabbit is chased by the donkey
X-but-not-Y	The box but not the chair is red	The butterfly but not the flower is red
Embedded sentence	The book the pen is on is red	The pencil the strawberry is on is yellow
Relative clause	The book is on the box that is red	The aeroplane is on the table that is broken

Scoring	All/none
Key Findings	Sentence type: ($\eta^2 = .59$) • Reversible, in/on > X-but-not Y, relative & embedded • X-but-not-Y, above/below > relative, embedded Length: ($\eta^2 = .26$) • Short > long Interaction (length x sentence type) was significant: • Length influenced scores on only 2 out of 6 sentence types: o In/on o X not Y

Table 3.6 *Summary of Willis and Gathercole (2001) Experiment 2*

Sample	61 English-speaking children in the age range of 4;8-5;8 were screened on the following VSTM tests: • Children's Test of Nonword Repetition (Gathercole & Baddeley, 1996) (standard score) • Auditory digit span (Gathercole, 1995) (standard score) The two standard scores were averaged to yield a composite phonological memory score for each child, two groups of children were identified based on their composite score: • 13 High VSTM • 13 Low VSTM The two groups were matched on age and nonverbal IQ
Stimuli	• 16 grammatical sentence types from TROG (Bishop, 1982), categories and examples below • *n* = 64, 4 per condition, length: 5-9.75 words

Example

Sentence Type	Example	**Sentence Type**	Example
Negative	The boy is not running	**In/On**	The cup is in the box
Three elements combined	The boy is jumping over the box	**Post modified subject**	The boy chasing the horse is fat
Plural personal Pronoun	They are sitting on the table	**X-but-not-Y**	The box but not the chair is red
Reversible active	The girl is pushing the horse	**Above/Below**	The ball is above the cup
Singular personal pronoun	She is sitting on the chair	**Not-only-X-but-also-Y**	Not only the bird but also the flower is blue
Plural noun Inflection	The cats look at the ball	**Relative clause**	The book is on the box that is red
Comparative adjective	The knife is longer than the pencil	**Neither-X-nor-Y**	Neither the dog nor the ball is brown
Reversible passive	The girl is chased by the horse	**Embedded sentences**	The shoe the comb is on is blue

Scoring	all/none
Key Findings	Group: ($\eta 2 = .43$) • High VSTM > Low VSTM Sentence Type: ($\eta 2 = .43$)

- Four most difficult: not only X but Y, relative, neither X nor Y and embedded

Interaction (group x sentence type) was significant:

- VSTM group influenced scores on 9 out 16 sentence types

Wilsenach (2006) compared the repetition ability of Dutch-speaking preschoolers with SLI, at familial risk of dyslexia, and Typically Developing age-matched controls. Syntax was manipulated by varying the transivity of verbs (transitive, intransitive, distransitive) and length was manipulated by adding an adjunct consisting of a prepositional phrase. The study is unique because the scoring system relied solely on qualitative scores targeting specific morphemes. The percent omissions of three closed class morphemes were calculated for each participant (auxiliary verb, prefix ge-, determiner) and a single open class morpheme (subject). Results showed that omission of closed class morphemes differentiated between groups with the highest omission scores observed in participants with SLI. Subject omission, on the other hand, was the same in all the three groups. Syntactic condition influenced the omission rate of auxiliary verb and determiner omission, with the distransitive condition being the most difficult to repeat. In the case of auxiliary verb omission, the influence of syntactic condition varied between groups as indicated by a significant group x syntactic condition interaction showing that only participants in the at-risk group significantly omitted auxiliary verb in the distransitive condition. Wilsenach (2006) argued that the influence of syntactic condition failed to reach significance for preschoolers with SLI because they have yet to master the auxiliary verb and find it difficult to repeat irrespective of syntactic condition. Length significantly influenced determiner and subject omission but not auxiliary or ge-omission. No interaction between group and length was found indicating that the pattern was the same for participants in the SLI, at-risk and control groups. Wilsenach (2006) was the first study to show that the influence of syntax and length was dependent on the type of score used to measure their impact.

Table 3.7 *Summary of Wilsenach (2006)*

Sample	Dutch-speaking • 21 SLI, mean age 4;3 • 36 at-risk of dyslexia, mean age 3;10 • 22 AM, mean age 3;11

Stimuli	2x2 design
	• complexity: intransitive/transitive/distransitive
	• length: short/long
	• length increased by adding a prepositional phrase
	• *n* = 18, 3 per condition
	• Controlled: overall sentence complexity
	  • structure NP+Aux V + remaining arguments
	  • familiar nouns
	  • monosyllabic nouns/prepositions

Example	None, examples provided in Dutch only.

Scoring	• **Quantitative:** none
	
	• **Qualitative**: Error based % omission of:
	  o auxiliary verb
	  o ge-
	  o determiner
	  o subject

Key Findings	**Qualitative** Group: • % omission of determiner, auxiliary verb & ge- was greatest in SLI group compared to controls and at-risk • % omission of subject same in all three groups Complexity: • % omission of auxiliary verb, determiner & subject was significantly influenced by complexity but not ge- • Distransitive was the most difficult syntactic condition • Auxiliary verb % omission showed a group x complexity interaction (only significant in at-risk group not SLI or controls) Length: • Significantly increased % omission of determiner & subject but not auxiliary verb & ge- • No length x group interaction

Building on Wilsenach's (2006) findings with Dutch-speaking preschoolers at familial risk for reading difficulties, Moll et al. (2015) compared the performance of children with dyslexia and Typical readers covering a wide age span: 6 to 12 years. The Dutch Sentence Repetition test was adapted to English and included two syntactic conditions (passive and active), and length was manipulated by adding adjectives. Different levels of scoring were used to gauge repetition performance. The first level consisted of the Total score and equalled the percentage of words repeated correctly irrespective of word type. The second and third levels of scoring were based on the SIT (Seeff-Gabriel et al., 2008) scoring system. Children with dyslexia performed at a reduced level in comparison to controls. Both syntactic condition and length influenced Total score, and long passive sentences were the most difficult to repeat. None of the interactions were significant, indicating that the pattern of performance was similar

in both reading groups and that syntax and length had independent effects on Total score. In addition to the Sentence Repetition test, two tests of VSTM—the Word List Recall subtest of the WMB-C (Pickering & Gathercole, 2001) and the Nonword Repetition Test (Dollaghan & Campbell, 1998)—were administered as well as a novel morphological awareness test to assess language ability. As with the Sentence Repetition test, the children with dyslexia performed poorer than controls on all three measures.

To gain a better understanding of the difference between the two reading groups and the underlying skills that contributed to their performance on the Sentence Repetition test, the main effects were re-examined using mixed-effects linear regression modelling but controlling for either VSTM or language ability and investigating the syntax and length effects on content and function words. Literacy group, VSTM ability (Word List Recall and Nonword Repetition), and morphological awareness predicted Total, Content, and Function word scores. Effects of syntax and length varied according to outcome score; both predicted Total score while only length predicted Content score and only syntax predicted Function score. For Total and Function score, literacy group effect disappeared when controlling for morphological awareness but remained when the two VSTM tests were controlled for. For Content score, literacy group effect disappeared when the VSTM tests and morphological awareness were controlled for. The authors argued that the disappearance of literacy group effect for Total and Function word scores despite that VSTM scores were reduced in children with dyslexia supports the greater contribution of linguistic knowledge to Sentence Repetition performance in comparison to VSTM. They also suggested that based on the study findings that, in general, sentence length largely influences content word retention while syntactic complexity primarily impacts function word retention. However, this interpretation is confounded by the design of the Sentence Repetition and the pattern of difficulty for the subtypes of content and function words. For content words, adjectives were the most difficult to repeat in comparison to nouns and verbs and they were only included in long sentences. For function words, prepositions were the most difficult to repeat and were only included in passive sentences. This would explain why length predicted content score and syntactic complexity predicted function score. Since Total score encompassed both content and function words it explains why length and syntactic complexity had independent effects. The question remains whether the independent effect would hold true had length and or complexity been manipulated differently.

Table 3.8 *Summary of Moll et al. (2015)*

Sample	English-speaking age range: 6-12 years in two groups • 40 with dyslexia mean age 9;3 • 57 Age Matched (AM) mean age 8;4		

Stimuli	2 x 2 design
	• complexity: low/ active, high/passive
	• length: short/long
	• length increased by adding adjectives
	  o short: 7-10 words, 8-13 syllables
	  o long: 10-13 words, 12-18 syllables
	• *n* = 20, 5 per condition,
	• all distransitive verb structure (Verb + 2 Objects)
	• controlled for:
	  o word frequency: high
	  o verb inflection: past tense "ed" for active, past participle for passive		

Example	**Complexity**	**Length**	
		Short	**Long**
	Low	A lady passed the man the paper.	A pretty woman passed the tall boy the crumpled magazine.
	High	The paper was passed by a lady to the man.	The crumpled magazine was passed by a pretty woman to the tall boy.

Scoring	• **Quantitative:** Level one: Total score % of words repeated correctly
	• **Qualitative**:
	  o Level two based on SIT (Seeff-Gabriel et al., 2008): Content % , Function % and Inflection (at ceiling not included)
	  o Level there:
	  o Content: Adjective, noun and verb
	  o Function: Preposition, pronoun and article		

Key Findings	**Quantitative** • Controlled for age and performance IQ • Level one: Total score (% of words repeated correctly) • Dyslexia < AM • Same pattern in both groups: long high complexity sentences < short low complexity sentences	**Qualitative** Dyslexia < AM in all comparisons Level two: SIT (Seeff-Gabriel, et al, 2008) • Content % > Function % in both groups • Group difference picked up more by function score • Length sig. predicted content score while complexity sig. predicted function score. Level three: • Content: Adjectives less than nouns and verbs • Function: Prepositions less than articles • Significant group x preposition score interaction o prepositions more vulnerable in Dyslexia group	

3.3 Linguistic Manipulation of Syntactically Simple Sentences

The previous section detailed how syntactic complexity influenced the repetition of well-formed sentences. There were hints of semantic influence with the finding that certain semantic relations were easier to repeat than others (animate vs. inanimate). The focus of this section will be on examining the influence of various domains of linguistic knowledge (syntax, semantics, prosody, and lexicality) on the repetition of syntactically simple sentences. The studies featured in this section systemically manipulated sentences by creating violations in the

different linguistic domains, essentially stripping away the information provided by a particular domain and examining the consequences. The underlying assumption is that if for example semantically implausible sentences (e.g., *The pretty milk sat on this pen*) were repeated with less accuracy than semantically well-formed sentences (e.g., *The happy boy sat on this chair*), which would indicate that children drew on their semantic knowledge to repeat simple sentences. The two examples illustrate how semantic plausibility of a sentence can be manipulated without disrupting sentence prosody, syntactic rules, or lexicality.

G. A. Miller and Isard (1963) were the first to see the potential of violating semantic and syntactic rules of sentences to examine whether semantic and syntactic knowledge aided the immediate repetition ability of English-speaking adults. A total of 150 sentences were constructed with 50 in each sentence type: Typical, Semantically Anomalous, and ungrammatical sentences. To help illustrate how sentences were constructed Table 3.9 provides five examples of each sentence type.

Table 3.9 *Stimuli Examples from G. A. Miller and Isard (1963, p. 220)*

Typical	Semantically Anomalous	Ungrammatical
• Gadgets simplify work around the house. • Accidents kill motorists on the highways. • Trains carry passengers across the country. • Bears steal honey from the hive. • Hunters shoot elephants between the eyes	• Gadgets kill passengers from the eyes. • Accidents carry honey between the house. • Trains steal elephants around the highways. • Bears shoot work on the country. • Hunters simplify motorists across the hive.	• Around accidents country honey the shoot. • On trains hive elephants the simplify. • Across bears eyes work the kill. • From hunters house motorists the carry. • Between gadgets highways passengers the steal.

All five Typical sentences used the same phrase structure. In creating, Semantically Anomalous and ungrammatical sentences, all content and function words were fair game with the exception of "the," which remained in the same position in every sentence. Anomalous sentences were created by jumbling the words that occurred in the same syntactic position across sentences. The first Anomalous sentence was created by taking the first word from the first Typical sentence followed by the second word from the second Typical sentence and so on. In the case of ungrammatical sentences, both syntactic and semantic rules were violated. Word order was permuted by commencing with a preposition and ending with a verb. The Preposition remained in the same order as Typical sentences while the remaining four words were jumbled across sentences in a similar fashion to Anomalous sentences. Each group of five sentences shared the same phrase structure and were manipulated using a similar method to the one presented here.

Results showed that Typical sentences (88.6%) were the easiest to repeat with complete accuracy followed by Semantically Anomalous sentences (79.3%) and ungrammatical

sentences (56.1%). This finding shows that adults draw on their semantic and syntactic knowledge to repeat sentences.

There have been relatively few studies that utilized the G. A. Miller and Isard (1963) approach to examine the contribution of different linguistic domains to the repetition of children. Most of the Sentence Repetition studies were conducted with English-speaking children and were largely focused on syntax and word order, since English is heavily reliant on word order.

3.3.1 Manipulation of syntax, semantics, prosody and lexicality

Table 3.10 provides an overview of studies presented in this section and highlights the sample, linguistic domain manipulated, scoring method used, and key findings. All studies converged on the influence of linguistic knowledge regardless of the linguistic domain examined be it syntax, semantics, prosody, or lexicality.

Three studies focused on the effect of syntactic manipulation via scrambling the word order of sentences and assessing its impact on the repetition of syntactically simple sentences: Bohannon (1975, 1976) and Love and Parker-Robinson (1972). Bohannon (1975) examined the ability of school-aged children in three grades (first, second, and fifth) to repeat 24 sentences in two syntactic conditions (typical vs. random) at six different words lengths (5, 7, 9, 11, 13, and 15 words). The following sentences are examples of stimuli (p. 445):

Typical word order (5 words): *The big dog ran outside* List 1

Random word order (5 words): *Ran dog the outside big* List 2

Syntax was found to significantly influence repetition with typical sentences repeated with twice as many words as scrambled sentences, 8.88 versus 4.04 mean words per sentence respectively, supporting the contribution of syntactic knowledge. Syntax significantly interacted with length, indicating that length had more of a detrimental effect on scrambled sentences than typical sentences. One possible explanation is that children can draw on their syntactic knowledge for typical sentences even when the load on their VSTM is increased in longer sentences. For short scrambled sentences, VSTM can compensate for the lack of support from syntactic knowledge; for long scrambled sentences, however, the support from VSTM and syntactic knowledge are compromised. Overall children's repetition scores improved with age. A significant interaction was found between length and grade, with older children repeating longer sentences better than younger children with less of a difference for short sentences. Syntax, unlike length, did not interact with age.

Bohannon (1976) extended the findings to younger children in kindergarten and a larger sample with 50 children in each grade. Unlike the first study, word length was held constant at five words per sentence. Again, scrambled sentences were more difficult to repeat than typical sentences (means 5.80 vs. 3.45 words per sentence, respectively). In contrast to the earlier study, there was a significant interaction between grade and syntactic condition. While typical sentences showed an increase in repetition scores, scores did not increase over age for

random sentences for kindergarten, first graders, second graders (mean scores equalled 3.27, 3.26 and 3.85 words per sentence, respectively) as opposed to Typical sentences (4.49, 5.98 and 6.63, respectively).

Love and Parker-Robinson (1972) manipulated syntax using a slightly different approach in an attempt to isolate the influence of grammaticality on function words (articles, prepositions, and auxiliary verbs) and inflections, and gauge whether one type of Grammatical Morpheme benefited more from accurate word order than the other. Overall, 4- and 6-year-old children were asked to repeat eight sentences in each of the following four string types (p. 312) the lexical status of content words was controlled for by replacing words in content word slots with nonwords in all four strings:

a) grammatical with function words:	*the zob is bixing the kiv*
b) grammatical without function words:	*zob bixing kiv*
c) ungrammatical with function words:	*the zob kiv the is bixing*
g) ungrammatical without function words:	*zob kiv bixing*

While 6-year-old children were able to correctly repeat more sentences than 4-year-old participants, the pattern of performance was the same in both age groups. Grammatical strings were easier to repeat than ungrammatical strings only when function words were present. When articles, prepositions, and auxiliary verbs were included, sentences with grammatical word order were correctly imitated more than scrambled sentences (a > c). However, in the absence of function words, the advantage of grammatical order disappeared even with the presence of inflections (b = d). In the case of grammatical sentences, children obtained higher scores when function words were present (a > b), in spite the fact that sentences in string (a) were twice as long as string (b).

The authors argued that the pattern of findings indicated that syntactic knowledge contributed to immediate repetition even when familiarity of content words was absent and that function words benefit from word order more than inflections. In the case of grammatical sentences with function words, VSTM failed to explain the superiority of longer sentences. The authors suggested that the addition of function words did not lead to an increase in VSTM load, and that children were able to use their familiarity with function words along with their syntactic knowledge to facilitate repetition.

Taken together all three studies show that for Typically Developing English-speaking children from kindergarten to fifth grade, scrambled sentences were repeated with less accuracy than typical sentences, supporting the contribution of syntactic knowledge to immediate repetition.

Table 3.10 *Summary of Studies that Linguistically Manipulated Syntactically Simple Sentence*

Study	Sample (mean age)	Linguistic Domain Manipulated					Outcome Measure	Main Findings
		Syntax	Semantic Plausibility	Prosody	Lexicality			
					Content	Function		
Bohannon (1975)	English-speaking 18 each grade • First grade (6;4) • Second grade (7;4) • Fifth grade (10;5)	✓	✗	✗	✗	✗	Number of words repeated in exact word order	• Typical > scrambled sentences • 5th grade = 2nd grade > 1st grade • Short > long • Length x syntax, grade x length and grade x syntax x length were significant
Bohannon (1976)	English-speaking 50 each grade • Kindergarten (5;9) • First grade (6;9) • Second grade (7;9)	✓	✗	✗	✗	✗	Number of words repeated regardless of word order	• Typical > scrambled sentences • Grade influence Typical sentences only: 1st grade = 2nd grade > kindergarten
Love and Parker-Robinson (1972)	English-speaking Typically Developing 12 in each age range: • 3;2-4;4 (3;9) • 5;2-7;2 (6;0)	✓	✗	✗	✗	✗	All/none	• 6-year-old > 4-year-old • Grammatical > ungrammatical only when function words were included • Grammatical with function > without function words
Bonvillian et al. (1979)	English-speaking 12 Typically Developing 3;4-4;4 (3;9)	✗	✗	✓	✗	✗	Number of morphemes deleted or incorrectly inserted	• Flat > normal intonation in long sentences • Presentation rate closest to children's own speaking rate led to the fewest errors
Akinsola (1986)	Bilingual (English & Yoruba) 12 Typically Developing 4;5-5;10 (5;0)	✗	✗	✓	✗	✗	Number of morphemes deleted or incorrect inserted	• Flat > normal intonation regardless of length • Presentation rate closest to children's own speaking rate led to the fewest errors
Polisenska et al. (2015)	50 Typically Developing in each language English-speaking: • 4-year-old (4;5) • 5-year-old (5;5) Czech-speaking:	✓	✓	✓	✓	✓	Span score: The highest target length at which the participant could successfully repeat 3 of the 4 targets in	• Same pattern in 2 age groups and language groups both interactions not significant: • age x linguistic condition (score: span) • language x linguistic condition (score: mean difference)

• 4-year-old (4;6) • 5-year-old (5;4)		a block. If 2 out of 4 a .5 was awarded for that length.	• All linguistic conditions showed sig. word span increase: • grammaticality > lexicality > semantic plausibility > prosody

Bonvillian et al. (1979) examined the effect of two aspects of prosody, intonation, and speaking rate (tempo), along with sentence length on the repetition of Typically Developing 3-year-old children. Sentences presented with flat intonation were repeated with less accuracy than sentences with normal intonation but this effect was limited to long sentences. The presentation rate that approximated the children's own speaking rate (two words per second) was repeated with greater accuracy than sentences presented with a faster or slower rate, one and three words per second, respectively; this occurred regardless of sentence length. Finally, short sentences were easier to repeat than long sentences.

Akinsola (1986) attempted to replicate the findings of Bonvillian et al. (1979) with older Yoruba- and English-speaking bilingual 5-year old children. The same sentences were presented with minor lexical substitutions due to lack of cultural familiarity. The presentation rate conditions were modified and included two-, three- and four-words per second, including the children's observed speaking rate in the spontaneous language samples (three words per second). Consistent with the earlier study, sentences presented at a rate closer to the children's own speaking rate were the easiest to imitate and short sentences were easier to imitate than long sentences. However, sentences with flat intonation were more difficult to repeat regardless of sentence length. The authors argued that although the sentences were presented in English, the typology of the children's mother tongue Yoruba crossed over to English and accounted for the stronger influence of intonation. Yoruba is a tonal language and variations in tone denote changes in meaning. Flat intonation not only violated prosody but also semantic features of the sentences. Together, the two studies show that prosody as reflected by intonation and tempo facilitate repetition. The degree of influence may vary based on language typology, with greater effects in tonal languages that rely on tone to convey word meaning.

Polisenska et al. (2015) was the only study that systematically manipulated all four linguistic domains and compared the influence of each domain on the immediate repetition of sentences. It was the first study to employ span as an outcome measure and the first to investigate whether the effects of the different linguistic conditions were comparable across two typologically different languages (English and Czech). Fifty Typically Developing English-speaking and 50 Typically Developing Czech-speaking children aged 4 to 5 years participated in the study. Children were asked to repeat sentences in a total of seven linguistic conditions ranging from well-formed sentences to sequences of nonwords. In order to calculate the maximum span for each linguistic condition, every condition commenced with a block of two-word sentences and successively increased length by one word until the final block of nine-word sentences. The following examples (A-G) illustrate the seven linguistic conditions for five-word sentences:

A) Well-formed sentence *He sent us a letter*

B) Well-formed sentence with list prosody *he, sent, us, a, letter*

C)	Semantically implausible sentence	*He sang us a kettle*
D)	Syntactically ill-formed pseudosentence with sentence prosody	*A sent he letter us*
E)	Pseudosentence with content words replaced by nonwords	*He /fɪnt/ us a /lɔpə/*
F)	Pseudosentence with function words replaced by nonwords	/vi/ *sent* /əʃʊ/ *letter*
G)	Pseudosentence with all lexical items replaced by nonwords	/vi fɪnt əʃʊ lɔpə/

A comparison of conditions showed that all four linguistic domains significantly impacted the performance of participants, but some domains had a greater impact than others. Although span scores increased with age, the pattern of performance within the two age groups was identical. Comparison across the two typologically different languages also revealed a largely similar pattern. Grammaticality (A vs. D) yielded the greatest difference in span scores with an increase in mean word span for English (4.35 to 8.1) and Czech (3.9 to 7.58), with a mean difference of 3.68 in both languages. It is striking, as the authors point out, that grammaticality affected both languages to a similar degree since Czech in contrast to English, has relatively free word order and violations were limited to permutations within phrases. Lexicality (D vs. G) produced the second largest difference in span scores, the mean word span increased in English (2.84 to 4.35) and Czech (2.54 to 3.9), with a mean difference of 1.51 and 1.36, respectively. Furthermore, pseudo-sentences with real function words (E) had a longer mean span than pseudo-sentences with real content words (F) with an advantage of over one word for English and about one for Czech. The similar degree of advantage (E vs. F) in both languages is interesting considering that the Czech is a morphologically rich language with a greater number of inflections in the Czech versus English stimuli for condition (E). Semantic plausibility (A vs. C) increased span by about one word for both languages. Prosodic structure provided the smallest jump in word span (.5 words) for both languages. Grammaticality and lexicality produced large effect sizes, while effect sizes for semantic plausibility and prosody were medium/small (Cohen, 1992).

The findings were in agreement with earlier studies from this section. The influence of syntax is consistent with previous research on children (Bohannon, 1975, 1976) and adults (G. A. Miller & Isard, 1963). The finding that familiarity with function words impacts span more than familiarity with content words is consistent with Love and Parker-Robinson's (1972) findings and further supports the role played by syntax and the morpho-syntactic relations as expressed by function words in Sentence Repetition. The impact of semantic plausibility and the reduced effects in comparison to syntax replicated the findings of G. A. Miller and Isard (1963) with adults. It also extends the influence of prosody to include list prosody in addition to the influence of intonation and tempo found by Bonvillian et al. (1979) and Akinsola (1986). Taken together, the studies suggest that irrespective of the age or language typology, linguistic knowledge contributes to Sentence Repetition with a privileged role for syntax and morpho-syntax in comparison to semantics or prosody. To the best of my knowledge, no studies

examined the underlying processes involved in the repetition of simple sentences by children with language impairment.

3.4 Literature Review Summary

Chapters 1 and 2 reviewed evidence on the clinical utility of SR as a measure of expressive language ability across a number of typologically diverse languages. In terms of psychometric properties, SR tests showed good levels of reliability and validity. For Typically Developing children, performance on SR tests improved as age increased. In children with language difficulties, performance was reduced in comparison to children without language difficulties. Therefore, it is a valid measure of development and language ability. Not only was SR able to distinguish between groups of children with different language abilities, but it was also the best individual clinical marker for SLI, showing high levels of sensitivity and specificity. This was found to be true irrespective of language typology (English, French, Cantonese), heterogeneity of SLI samples across studies, age of participants, and design of SR stimuli. Building on the extensive cross-linguistic research, SR holds promise as an assessment tool for Arabic-speaking preschool children, a language where there is a shortage of normative data and few, if any, language assessments. Having designed a novel test, a key aim of the current study was to examine its clinical utility and investigate whether findings from other languages could be replicated in Arabic. This is achieved by determining whether the newly developed test is psychometrically robust, sensitive to age and language ability of participants, and has acceptable levels of sensitivity and specificity.

It is important to determine the underlying skills tapped by SR in order to inform its use as a clinical assessment. Diagnostic accuracy studies featured in Chapter 2 relied mainly on correlational evidence and a polarized view of underlying skills with the assumption that span tasks measure memory while grammatical morpheme probes measure language. This yielded inconclusive results and highlighted the difficulty untangling the influence of memory and language on immediate repetition tests. An alternative way to address the question of underlying processes is to examine the impact of linguistic manipulation on children's repetition and determine what types of linguistic knowledge influence their performance; this was the focus of Chapter 3. In the case of span tests, lexicality, frequency, concreteness, and imageability were all found to influence the recall ability of children 6 years of age and older. This was found to be true regardless of language typology (English or French) and language ability (typical or children with SLI). For SR tests, the linguistic domains of syntax, semantics, prosody, and lexicality were all found to influence children's repetition irrespective of how syntax was manipulated (varying syntactic complexity or jumbling word order).

Taking inspiration from manipulation studies, two additional Saudi Arabic tests were created for this research in order to investigate skills underlying SR: (1) an adapted VSTM test based on the structure of the WMTB-C (Pickering & Gathercole, 2001) with three subtests: Digit Recall, Word List Recall and Nonword List Recall; and (2) an Anomalous Sentence

Repetition (ASR) test comprising sets of Semantically Anomalous and Syntactically Anomalous sentences. A comparison of the profile of performance on the VSTM test based on linguistic item would determine whether the influence of lexicality (nonword vs. word) and frequency (digit vs. word) could be replicated in Arabic and extended to children younger than 6 years of age. On the ASR test, a comparison of the profile of performance based on sentence type would determine whether the influence of semantic plausibility and grammaticality could be replicated in Arabic and extended to children with language difficulties.

Close inspection of SR tests used as assessments or research tools revealed important design issues that need to be taken into consideration. The Cantonese study (Stokes et al., 2006) highlighted the importance of adapting tests according to language typology and targeting structures that are known to be difficult for children within each language to prevent hindering the test's discrimination ability. Failure to take dialect into consideration can lead to misdiagnosing children who speak a different dialect as language impaired (Hemingway et. al., 1981). In terms of targets, most available tests such as the CELF-4 (Semel et. al., 2003) simultaneously manipulate length and grammatical complexity making it difficult to pinpoint which structures children have difficulty repeating. In terms of outcome measure, qualitative scoring methods are more discriminating and yield more information than a purely quantitative score such as all/none. Young children and children with language impairment found it difficult to repeat function words in comparison to content words (Brown & Fraser, 1963; Devescovi & Caselli, 2007; Seeff-Gabriel et al., 2008), indicating that repetition taps morpho-syntactic knowledge and that performance on Grammatical Morphemes may help identify children with difficulties, point to structures that may require further investigation, and guide targets for intervention.

To address these design issues the novel SR test was developed in the local Najdi dialect of Arabic and consisted of simple sentences that targeted specific morpho-syntactic structures known to be difficult for Arabic-speaking children with language impairment. Two scoring methods were used: a quantitative three point scoring system similar to the CELF-4 (Semel et. al., 2003), and a qualitative scoring method adapted from the SIT (Seeff-Gabriel et al., 2008), which yielded a Grammatical Morpheme and a Lexical Morpheme score. A comparison of Lexical and Grammatical Morpheme scores in the SR and ASR tests would indicate whether the impact of morpho-syntax on repetition could be replicated in Arabic. Also, it allows for a comparison of lexical and grammatical morphemes across sentence type in the ASR test to determine whether violations in semantic plausibility or grammaticality impact them differently.

Chapter Four: Methods

The study aimed to develop tests of Sentence Repetition (SR) and Verbal Short Term Memory (VSTM) and evaluate their use as possible assessment tools for Arabic-speaking Saudi children by investigating the performance of Typically Developing participants across the ages of 2:6 to 5:11 years and Language Concerns participants within the same age range. A further aim of the study was to investigate the underlying processes involved in SR, and more specifically the contribution of established linguistic knowledge, by comparing the performance of participants across different linguistic factors on the SR, VSTM, and ASR tests.

Three novel tests were developed: VSTM, a SR test, and an Anomalous Sentence Repetition test (ASR). The VSTM test was based on three subtests of the Working Memory Test Battery for Children (WMTB-C; Pickering & Gathercole, 2001) and includes Digit Recall, Word List Recall, and Nonword List Recall. The SR test consisted of well-formed sentences containing a range of Grammatical Morphemes with a focus on morphemes reported to be difficult for language impaired Saudi children. The ASR test consisted of Semantically and Syntactically Anomalous sentences, created by manipulating the lexical content and grammatical structure of the target sentences in the SR test.

A mixed research design was adopted. The between-subject variables were age of participants (2:6 to 5:11) with seven levels corresponding to 6-month intervals; gender with two levels (boys and girls); and language status with two levels (Typically Developing and Language Concerns). The within-subject variables were the performance of participants on the VSTM, SR, and ASR tests. For the SR and ASR tests, three measures of accuracy were obtained: Grammatical Morpheme score (the number of Grammatical Morphemes repeated correctly), Lexical Morpheme score (the number of Lexical Morphemes repeated correctly), and Total Sentence Accuracy, based on Clinical Evaluation of Language Fundamentals 4th edition (CELF–4; Semel et al., 2003) scoring method. For the VSTM test, four accuracy measures were obtained: a span score for each of the subtests (Digit Recall, Word List Recall, and Nonword List Recall) representing the highest number of items (digits, words, nonwords) repeated in correct order, and a Total span score that equalled the sum span score of the three subtests. In addition to accuracy scores, the frequencies of selected error types were obtained.

In keeping with the two key aims of the study the following questions were addressed:

Aim: To examine the clinical utility of VSTM and SR tests.

- What are the psychometric properties of the tests?
    - Do the tests have acceptable levels of reliability (inter-rater, test re-test, and internal consistency)?
    - Do the tests have acceptable levels of validity (construct and concurrent)?
    - Is there an effect of age on children's performance?
    - Is there an effect of language ability on children's performance?

- To what extents do the tests accurately classify individuals from the Typical sample and the children with Language Concerns (specificity and sensitivity)?

Aim: To evaluate the contribution of established linguistic knowledge to immediate repetition by comparing the following patterns.

- Is there an effect of morpheme type on SR performance (Lexical vs. Grammatical Morpheme scores)?
- Is there an effect of item type on VSTM span score? (Digit vs. Word vs. Nonword)
  - Is there a difference in the type of errors across subtests?
- Is there an effect of sentence type on ASR performance? (Typical vs. Semantically Anomalous vs. Syntactically Anomalous)
  - Is there an interaction between morpheme type and sentence type?
- If there are effects of linguistic factors (item type, morpheme type, sentence type), are the profiles of performance similar across different age groups of children in the Typical sample and across children with different levels of language ability or do they differ?

4.1 Participants

The Research Ethics Committee at City University granted ethical approval for the study. Participants were recruited from two nursery schools in Riyadh, one private and one public. The heads of both nurseries were sent invitation letters and were willing to distribute invitations and consent forms to parents (Appendix A). Teachers were informed of the study's purpose and selection criteria in a face-to-face meeting and were asked to identify children who fit the criteria by reviewing the class list and sending the provided invitation letter and consent form to parents. Parents were also asked to provide information regarding their education level. For children with language concerns, teachers were asked to identify those children who were not speaking or understanding at the level of their peers. When possible, parents were asked to confirm the teachers' concerns, this was possible for 12 of the 16 participants in the Language Concerns group. Teachers were asked the following question: "Are you concerned about this child's ability to speak or understand at the level of their peers?" No further specific criteria were given.

To confirm language status, teachers' concerns would ideally be checked against language measures other than those developed for the current study. However, there were no additional assessments available. Based on my clinical practice, the most common test used in clinical practice was an Arabic-translated version of the PLS (Zimmerman, Pond, & Steiner, 2009). There were no unified translations in the clinic and clinicians often translated the PLS on the spot. Additionally, the PLS is not culturally appropriate and did not account for the unique characteristics of Arabic. There were also no established protocols for informal tests or criterion reference tests (Shalaan, 2009).

The criteria for inclusion in the Typically Developing group were the following:

- aged 2:6 to 5:11;
- Saudi, with Arabic as a first language; children are exposed to English as part of their curriculum and sometimes through their caregivers and TV at home;
- no concerns expressed by teachers regarding language development;
- Typically Developing with no history of neurological or behavioural impairment;
- parents provided consent; and
- passed a hearing screening at 20 dB for the frequencies between 500-1000 Hz (for participants 3 years of age and older).

A total of 153 children were recruited. Background information on participants in the Typically Developing group can be found in Table 4.1. A number of children were referred for inclusion in the Typically Developing group but were not included in the study: three children refused to participate, five children moved to another school before commencement of testing, one child was absent on testing day and four children failed the hearing-screening test.

The criteria for inclusion in the Language Concerns group were:

- aged 2.6-5.11 years;
- Saudi, with Arabic as a first language;
- identified by teachers who expressed concerns about the child's language development and the concern was confirmed with parents when possible;
- no history of hearing loss, neurological impairment or severe behavioural problems;
- parents provided consent; and
- passed a hearing screening at 20 dB for the frequencies between 500-1000 Hz (for participants 3 years of age and older).

A total of 17 children were recruited. Background information on participants in the Language Concerns group can be found in Table 4.2. One participant was excluded because he refused to participate.

Table 4.1 *Background Information on Typically Developing Participants*

Age Group	Number of Participants	Mean age in months (years)	School (Private:Public)	Gender (Female:Male)
2;6 -2;11	20	31.85 (2.6)	(10:10)	(11:9)
3;0-3;5	20	38.8 (3.23)	(10:10)	(10:10)
3;6-3;11	20	45 (3.75)	(10:10)	(10:10)
4;0-4;5	20	51 (4.25)	(10:10)	(10:10)
4;6-4;11	20	55 (4.58)	(10:10)	(10:10)
5;0-5;5	20	62.8 (5.23)	(10:10)	(10:10)
5;6-5;11	20	67 (5.58)	(10:10)	(10:10)

Table 4.2 *Background Information on Language Concerns Participants*

Participant	Age in months	Gender	School
141	34	Female	Public
142	36	Male	Private
143	43	Female	Private
144	43	Female	Public

145	44	Male	Public
146	52	Male	Public
147	53	Male	Private
148	57	Male	Private
149	60	Male	Private
150	61	Male	Private
151	63	Female	Private
152	67	Male	Public
153	71	Male	Private
154	71	Male	Public
155	71	Male	Private
156	71	Male	Public

4.1.1 Demographics

Parents were asked to indicate the highest level of schooling achieved from the following categories: less than secondary school, secondary school, diploma, university degree, and graduate degree. Table 4.3 shows the breakdown of the educational level of mothers and fathers of the Typically Developing children and the Language Concerns group. As seen in Table 4.3, the majority of children in both groups (the Typically Developing children and the Language Concerns) had mothers who completed their university degree (73.6 % and 68.8% respectively). Similarly, most of the fathers of the Typically Developing children had a university degree. While, the educational level of fathers of the Language Concerns children was roughly equally distributed among the three categories: secondary school (31.3%), university degree (31.3%), and graduate degree (25%). Thus, it seems that in general most of the children in our sample came from highly educated backgrounds.

It is difficult to compare the educational level of parents in our sample to Saudi parents in Riyadh since the General Authority for Statistics in Saudi Arabia does not provide values for parents specifically. It describes the educational level for females and males who are 15 years of age and over according to their marital status: never married, married, divorced, and widowed (see Appendix B). In order to compare the level of education of parents in our sample to the Saudi population in Riyadh, the number of males and females who have never been married was subtracted from the totals provided at each educational level. Percentages were then calculated and compared to percentages in our sample (see Table 4.3).

As shown in Table 4.3, the educational level of parents in our sample seems to be higher than the Saudi population in Riyadh. While, the majority of parents in our study had a university degree, most Saudi females in Riyadh had less than a secondary degree and most males had a diploma. This might be partly due the differences in the age range between parents included in the study compared to the Saudi population. It should be noted, however, though our sample was skewed to highly educated parents it might be very comparable to the population seen at the speech and language clinics in Saudi. As noted by AlKadhi (2015), parents who seek speech and language therapy services tend to come from families with relatively high educational backgrounds.

4.2 Development and Piloting of Tests

The instruments used in the study needed to be administered to a substantial number of children across age groups at the development and pilot stages for several reasons. First, the researcher needed to build relationships with schools and teachers, in recruiting participants, and in running the novel tests with a wide age range of participants. In addition, extensive piloting was needed in order to

1. adapt the WMTB-C subtests for Arabic, investigate the appropriateness of items selected for the Word List Recall and Nonword List Recall subtests, and establish the maximum span needed for each subtest;
2. establish if reliable responses could be obtained on the VSTM test from participants who were younger than the standardization sample of the WMTB-C;
3. establish the best procedure to elicit responses from younger participants included in the study; and
4. establish at what age children were able to perform the ASR test.

Information on the number of participants in the development and pilot stages can be found in Table 4.4. The results of the development and pilot stages can be found in Appendix C.

Table 4.3 *Education Levels of the Parent Samples and Saudi Population*

Group		**< Secondary**		**Secondary**		**Diploma**		**University**		**Graduate**	
		N	**%**	***N***	**%**	***n***	**%**	***n***	**%**	***N***	**%**
Typically	Female	9	6.4	12	8.6	7	5	103	73.6	9	6.4
Developing	Male	4	2.9	18	12.9	9	6.4	78	55.7	31	22.1
Language	Female	0	0	4	25	1	6.3	11	68.8	0	0
Concerns	Male	1	6.3	5	31.3	0	0	5	31.3	4	25
Census[a]	Female	553,294	57.2	213,712	22.1	35,298	3.6	158,616	16.4	6,533	0.7
	Male	284,593	31.8	270,113	30.2	81,919	9.0	229,498	25.6	28,905	3.2

Note. Saudi Population (>15 Years, excluding those never married) in Riyadh Administrative Region

Table 4.4 *Summary of Participants in Development and Pilot Stages*

Development Stage	**Pilot Stage**
32 children were recruited (2;6-5;6)	30 Typically Developing (2;6-5;11)
4 were excluded because parents expressed concern about language development.	5 children in each six-month interval with the exception of the final age group which was an 11-month interval 5;0-5;11
4 children refused to participate	
Final sample included 26 children	

4.3 Assessments

4.3.1 Verbal Short-Term Memory test

The VSTM test will be discussed in several sections. The first section will describe WMTB-C, particularly the VSTM subtests with some critical observations. The second section will describe the process of developing the VSTM test and how it was presented in the main study, followed by sections describing the procedure and scoring.

4.3.1.1 Working Memory Test Battery for Children

The WMTB-C is one of the few standardized memory tests that are available for clinical use by speech and language therapists (Montgomery et al., 2010). The test was standardized on 734 children aged 4;7 to 15 years. The battery includes multiple subtests that tap into each of the three components of the working memory model: phonological loop, central executive, and visuo-spatial sketchpad (Baddeley, 1986; Baddeley & Hitch, 1974). The phonological loop component is concerned with the immediate recall of verbal information and consists of four subtests: Digit Recall, Word List Matching, Word List Recall, and Nonword List Recall.

The Digit Recall, Word List Recall, and Nonword List Recall subtests were found to exclusively tap into the phonological loop component of the working memory model, across all the age groups of the standardization sample. However, the Word List Matching subtest was found to tap into both the phonological loop and central executive components of the working memory model in children younger than 5;6 of the standardization sample (Pickering & Gathercole, 2001). Since the oldest participants in this study were 5;11 years old and the reason for including a VSTM test in the battery of the study was to investigate immediate serial recall related to the phonological loop only, the Word List Matching subtest was not adapted.

The Digit Recall, Word List Recall, and Nonword List Recall subtests of the WMTB-C use a serial-recall paradigm, which involves the auditory presentation of a sequence of digits, words, and nonwords to be repeated immediately in the correct order. Each subtest commences at a span of one item and increases in length by one item in consecutive spans. Within each span there are six trials of an equal number of items. There are six trials in each span to increase the robustness of each subtest and reduce the likelihood of participants achieving a span score by chance (Pickering & Gathercole, 2001).

Each subtest yields a span score, which is equal to the highest number of digits, words, or nonwords repeated in the correct order. Three practice trials are presented at the beginning of each subtest and range in span from one to three. The WMTB-C is presented in a fixed order and commences testing with Digit Recall because digits are highly familiar to children 4;7 years and above. Subsequent subtests range from familiar to nonsense (Pickering & Gathercole, 2001), starting with Word List Matching, then Word List Recall, and finally Nonword List Recall.

The Digit Recall subtest consists of digits ranging from one to nine. All digits are

monosyllabic apart from bisyllabic *seven*. They include three phonemes that are identified as late developing in English: /θ, s, r/ (Shriberg & Kwiatkowski, 1994). No digits occur more than once in a single trial but it was possible for digits to be repeated in a single span across trials. The maximum span for the Digit Recall subtest is nine.

The Word List Recall subtest consists of monosyllabic words with CVC structure. The pool of consonants included were /p, b, t, k, g, ʧ, ʤ, m, n, l/, with only /l/ identified as late developing by Shriberg and Kwiatkowski (1994). The selected words were identified as words that are common and likely to be familiar to young children in the prototype battery, but there was no reference provided of what measures were used to establish familiarity (Gathercole & Pickering, 2000). Words include nouns, adjectives, verbs, and function words such as the negative particle "not" and "but." Nouns are in singular form with the exception of "men" and include abstract nouns (e.g., "luck" and "doom"), emotionally charged words (e.g., "church" and "god"), ambiguities between content and function words (e.g., "can" and "back"), and the number "ten." Verbs include irregular past tense participle (e.g., "caught" and "torn"). No words are repeated in the same trial or span. Only two words are repeated in different spans: "pin" in spans 4 and 6, and "turn" in spans 4 and 5. Some trials contain words that rhyme (e.g., span 5 trial 6 dug pan bug man catch). Within the same trial phonemes are repeated across words (e.g., span 3 trial 4 park cod dip). The maximum span for the Word List Recall subtest is seven.

The Nonword List Recall subtest consists of monosyllabic nonwords with Consonant Vowel Consonant (CVC) structure. Nonwords were created from the same pool of phonemes as the words in the Word List Recall subtest. There was no mention if estimates of wordlikeness were obtained in the test manual or prototype battery. There is no repetition of nonwords across trials or spans. Within the same trial, phonemes are repeated across nonwords (e.g., span 4 trial 4 ped barp korp cheed). The maximum span for the Nonword List Recall subtest is six.

4.3.1.2 Arabic Verbal Short Term Memory test

Three subtests of the WMTB-C that tap into the phonological loop component of the working memory were adapted to Arabic: Digit Recall, Word List Recall, and Nonword List Recall. The test was extended to include two Word List subtests and two Nonword List subtests to investigate whether the lexical category of words influences span. Word List subtests were Word List Recall-Noun consisting of nouns only and Word List Recall-Mixed consisting of an equal mix of nouns, adjectives and verbs. Nonword List subtests were Nonword List-Noun and Nonword List-Mixed based on the lexical category they were derived from. The VSTM Word List Recall subtests used in the development stage can be found in Appendix D.

4.3.1.2.1 Digit recall

The Digit Recall subtest of the WMTB-C was translated into Arabic. Digits range from two syllables (1,2,5,6,7,9) to three syllables (3,4,8) as shown in Table 4.5. To date there are no

published studies of the development of speech sounds in Typically Developing Saudi children. A study investigating Jordanian Arabic (Amayreh & Dyson, 1998) identified the following sounds included in digits as late developing: /**ʔ,** θ, **ʕ** / . In contrast to English, the sounds /s, r/ —which are also included in digits—were identified as intermediate developing sounds in Jordanian Arabic. The Digit Recall subtest ranges in span from one to seven.

Table 4.5 *Transcription of Arabic Digits and Number of Syllables*

Digit	Transcription	Number of Syllables
1	/wa:.Hid/	2
2	/**ʔ**iθ.najn/	2
3	/θa.la:.θa/	3
4	/**ʔ**ar.ba.**ʕ**a:/	3
5	/Xam.sa:/	2
6	/sit.ta:/	2
7	/sab.**ʕ**a:/	2
8	/θa.man.ja:/	3
9	/tis.**ʕ**a:/	2

4.3.1.2.2 Word List Recall

The subtest included bi-syllabic words with CVC.CVC structure. Bi-syllabic words were chosen because this is the most common word length in Arabic. This is supported by Alsari (2015) who found that bi-syllabic words were the most frequent word length used by 72 Saudi children aged 3-5 years in their spontaneous speech and a story-retelling task in comparison to monosyllabic and multisyllabic words. Also, bi-syllabic words would most closely match the syllable length of digits. In the absence of familiarity measures and normative data on vocabulary development, words were selected from books aimed at 3-year-old children. Books were written in Modern Standard Arabic, so the regional Najdi dialect was taken into account; in instances where lexical variation occurred, words in Modern Standard Arabic were replaced with the common form used in the Najdi dialect (e.g., /t̲abi:b/ was replaced with /duk.to:r/ [doctor]). Words containing gemination (e.g., /sik.ki:n/ [knife]) and words containing clusters (e.g., /sta:.ra/ [curtain]) were avoided. The pool of consonants included all Arabic consonants with the exception of the sounds /q, d̲/ , which do not occur in the Najdi dialect. Translation of words from the WMTB-C Word List Recall subtest into Arabic was not an option because

1) some words were culturally inappropriate (e.g., “pig,” “nude,” “church,” “pork”); and
2) translated words violated the selection criteria:
   a) translated word is one syllable only (< 2; e.g., cheek /Xad/ mud /t̲i:n/);
   b) translated word is more than two syllables (e.g., map /xa.ri:t̲a/ bill /fa:.tu:.rah/; both words are three syllables); and
   c) translated word contains germination (e.g., man /ri ʤ. ʤa:l/ lip /**ʃ** if.fah/).

Two separate subtests were created to investigate whether the lexical category of words affected span. Both ranged in span from one to five. The maximum span in the Word List Recall subtest of the WMTB-C is seven. However, the WMTB-C was designed to test children as old as 15 years; the oldest participants in this study were 5;11 years, so spans were not required longer than five.

4.3.1.2.3 Word List Recall: Noun

Items in this subtest were singular nouns that met the above-mentioned criteria. Nouns included both animate and inanimate objects. Inanimate objects included a mix of masculine and feminine nouns. Feminine nouns that contained a feminine infliction were only included if they did not have a singular masculine counterpart. For example, /naml-a/ 'ant-f', removal of the feminine marker results in /naml/ 'ants,' which is the plural form and not a masculine counterpart. In addition, abstract nouns were avoided (e.g., /**ʔ**is.bu:**ʕ**/ [week]). Nouns were allocated to ensure that nouns did not rhyme within the same trial and that the same noun did not occur in two consecutive spans.

4.3.1.2.4 Word List Recall: Mixed

Items in this subtest included nouns, adjectives, and verbs that met the above-mentioned criteria. In addition, adjectives and verbs were in the simplest syntactic form to match them with nouns as far as possible and to avoid errors at the morpheme level. The simplest syntactic form of verbs is past third person, singular, masculine (e.g., /katab/ [write.pf3msg]); the following forms were avoided (e.g., /katabat/ [write.pf-3fsg]; /jiktib/ [imp3pmsg-write]). The simplest syntactic form of adjectives is singular masculine (e.g., /kibi:r/ 'big' vs. /kibi:ra/ 'big-f'). Nouns were selected from those in the Word List Recall-Noun subtest. Words were allocated to ensure that the six trials at each span had an equal number of each lexical category and that words within the same trial did not rhyme. There was no repetition of words.

4.3.1.2.5 Nonword List Recall

Arabic is a Semitic language similar to Hebrew. Words contain a consonantal root that carries the meaning and a vowel template. Ravid and Schiff (2006) identified three possible ways of creating nonwords in Hebrew:

1. combining a non-existent consonantal root with a non-existent vowel template;
2. combining a non-existent consonantal root with an existent vowel template;
3. combining an existent consonantal root with an existent vowel template.

The first method was avoided because it yielded nonwords that bared no phonotactic resemblance to real words. The latter two methods were selected to create non-words in Arabic. For example,

2. /θ-k-b/ a non-existent consonantal root + CuC.Ca as in /ruk.ba:/ (knee) ⇒ /θuk.ba/

3. /d-k-k/ (destroy a mountain) + CaCi:C as in /ħa.di:d/ (steel) ⇒/daki:k/

An online search engine containing the six most common Classical Arabic dictionaries

(Arab Scholar, 2016) was consulted to determine whether the selected consonantal roots existed and whether the combination of existent consonantal roots and patterns yielded a word. A consonantal root was considered existent if it was found in at least one of the dictionaries. Because of the possible lexical differences between Najdi Arabic and Classical Arabic, five adult native speakers of the Najdi dialect were asked to indicate whether they knew the nonwords.

Two subtests were created based on the real word subtests: (1) Nonword List Recall-Noun was derived from the nouns in the Word List Recall subtest, and (2) Nonword List Recall-Mixed derived from the nouns, adjectives, and verbs in the Word List Recall-Mixed. Hence, items in the two subtests were bi-syllabic nonwords with CVC structure. Nonwords were allocated to ensure that nonwords did not rhyme within the same trial. There was no repetition of nonwords. Both subtests ranged in span from one to four.

Results of the development stage (see Appendix C) indicated that the span score of participants was unaffected by whether the subtest consisted of a list of nouns only or a mix of verbs, adjectives, and nouns. It was therefore decided to reduce the number of lists and combine Word List Recall-Noun and Word List Recall-Mixed, as well as Nonword List Recall-Noun and Nonword List Recall-Mixed. A new pool of words/nonwords from both lists was compiled with the added criterion that they did not contain late developing sounds (as identified by Jordanian Arabic study, Amayreh & Dyson, 1998) other than those occurring in digits. Words/Nonwords were allocated following the same rules mentioned previously.

The final Nonword List Recall subtest contained 20 nonwords created by combining a non-existent consonantal root with an existing vowel pattern, and 40 nonwords created by combining an existing consonantal root with an existing vowel pattern. Of these 40 nonwords, four yielded a rare word in Classical Arabic but this was not a word in Najdi Arabic as verified by the five adult native speakers. The VSTM subtests used in the study can be found in Appendix E. Based on the results of the development and pilot stages no additional spans were needed in any of the subtests.

4.3.1.3 Procedure of the Verbal Short Term Memory test

The subtests were presented in a fixed order starting with Word List Recall, followed by Digit Recall, then Nonword List Recall. The order of presentation differed from that of the WMTB-C, which commences with Digit Recall, as most of the participants were younger than the WMTB-C standardization sample (which started at age 4:7) and had been less exposed to digits. Digit Recall was nevertheless included in the test because it is the most common form of VSTM test used in Saudi, and the Recall of Digits Forward subtest of the British Ability Scales (BAS) was standardized on children as young as 2;6 (Elliott, 1996). In addition, the use of multiple measures improves the reliability of the assessment and reduces the measurement error associated with use of individual measures (Pickering & Gathercole, 2001).

The following instruction was given to participants before each subtest: “I will say a list of words/numbers/silly words. I want you to listen carefully and then I want you to copy me and say them in the same order. Are you ready?" Three practice trials were presented before each subtest in ascending order from a span of one to three words, digits, or nonwords. If a participant did not respond or did not repeat any of the trials correctly in the first presentation, instructions were repeated with demonstration of correct responses, and practice trials were re-administered. After the second administration of the practice trial, testing commenced regardless of whether participants repeated them correctly or not. If a participant repeated early practice trials correctly, testing commenced at the highest span the participant was able to repeat. Practice trials did not count towards span score. If the participant repeated three trials correctly in a span the examiner moved on to the next span level. If the participant repeated fewer than three trials correctly the examiner reversed to a lower span. Testing was discontinued if three trials were repeated incorrectly in the same span.

4.3.1.4 Scoring of the Verbal Short Term Memory test

In scoring each trial allowances were given for any misarticulations consistent with the child’s speech. Span scores were obtained for each subtest and a VSTM Total score was calculated by adding the span scores of the three subtests.

4.3.1.4.1 Span score

Span score was the highest number of items (digits, words, nonwords) repeated with complete accuracy and in the correct serial position by a participant in 4 out of 6 trials. To illustrate with digits:

Span	**Tester**	**Child**	**Trial Score**
Span 2	6 2	6 2	1
	4 9	6 3	0
	9 1	9 1	1
	3 8	2 8	0
	7 4	7 4	1
	2 5	2 5	1
Span 3	4 8 3	4 8 5	0
	2 6 1	2 1 7	0
	7 4 3	9 4 5	0
	3 7 6	3 8 8	0
			2
			Span Score 2

A 0.5 was awarded if a participant accurately repeated items in three out of six trials. This differed from the scoring method used in the WMTB-C in which achievement at each span is all or nothing. The change was implemented to improve the test’s discrimination of VSTM capacity at this young age, and the test’s potential sensitivity to age (see Polišenská, 2011). For example:

Span	**Tester**	**Child**	**Trial Score**

Span 3	4 8 3	4 8 3	1
	2 6 1	2 6 3	0
	7 4 3	7 4 3	1
	3 7 6	3 7 6	1
	1 8 4	1 7 4	0
	6 9 4	6 2 5	0
Span 4	5 9 2 6	5 9 8 7	0
	3 1 7 4	3 1 1 7	0
	2 8 5 1	2 8 4 5	0
	9 6 2 7	9 6 1 7	0
			Span Score 2.5

Maximum Span Score for subtests:

Digit Recall: 7

Word List Recall: 5

Nonword List Recall: 4

4.3.1.4.2 Total Span score

Total Span score was the sum span of the three subtests. For example, if a participant obtained the following spans in each subtest:

Digit Recall: 4

Word Recall: 3

Nonword Recall: 2

Total Span: 9

4.3.1.4.3 Errors

Errors which occurred in the final span attempted by participants were examined for all subtests and classified as follows:

- Item errors
  - Omission: Target item does not occur in any position
    (e.g., Tester: 1 4 Child: 4)
  - Substitution: Target item replaced with an item that did not occur in the trial
    (e.g., Tester: 9 4 6 Child: 9 4 5)
  - Perseveration: Target item replaced with an item from the same trial
    (e.g., Tester: 9 4 6 Child: 9 4 4)
    In the Word Recall subtest, an error was also classified as perseveration when a target word was replaced with a word from the previous trial
    e.g., Tester Child
    walad/ 'boy' /walad/
    /giri:b/ 'near' /walad/
  - Unintelligible. Target item replaced with a speech segment that is unrecognizable
    (e.g., Tester 1 8 4 Child 1 # 4)

- Order errors
    - Migration at item level: Target item recalled in the incorrect serial position (e.g., Tester: 1 5 8 Child: 1 8 5)
    - Migration at phoneme/syllable level: Transposition of part of an item in a trial to another (e.g., Tester: / libi:r, da:fa:/ Child: /da:fi:r, da:fa:/)

In some instances, a single error constituted both an item and order error (e.g., Tester: 9 4 6; Child: 9 5 4). In this case one item error was calculated as a substitution in addition to a migration error.

4.3.2 Sentence Repetition test

This novel test consisted of 14 sentences sampling a wide range of basic sentence structures and key grammatical markers. The sentences were created specifically for Arabic rather than translated from any of the available tests in English in order to account for the morphological richness of Arabic.

Sentence Structure:

- Length:
    - 5-7 words (mean 5.57)
    - 6-11 Grammatical Morphemes (mean 8.36)
- Najdi Dialect (language spoken at home)
- Verb tense was equally divided between past and present tense and in third person form based on the finding of Abdalla and Crago (2008) that these are the most error-prone morphemes in the spontaneous speech of Hijazi Arabic-speaking Saudi children with SLI
- Equally divided between Verb Subject_ and Subject Verb_ with a range of verb complements and modifiers; in both word orders, the subject agreed with the verb in person, number, and gender and is marked as a prefix on the verb
- Morphemes included the following:
    - Lexical:
        - Verbs: familiar verbs including eat, run, drink, ask, love, sit, buy, wash, put, take, read, live, play, swim
        - Nouns: familiar nouns in the following lexical categories: animals, food, outdoors, toys, household objects and people. Also, some common names
        - Adjectives: tall, black, small, pretty, big, new
    - Grammatical: pronouns, prepositions, conjunctions, copula, and a range of noun and verb markers.

See Table 4.6 for a list of the lexical and Grammatical Morphemes and their distribution. The full set of sentences is presented in Table 4.7, and score sheets can be found in Appendix F.

Table 4.6 *Distribution of Morphemes in the Sentence Repetition Test*

Morpheme Type		Arabic	English	Total
Lexical				
Noun				34
Adjective				6
Verb				16
Grammatical				
Nominal	Determiner	ʔil- bissa	the-cat	22
		ʔis-so:da:	the-black	6
	Feminine marker	zara:f-a	giraffe-fsg	12
	Demonstrative	ha:ði:	this.fsg	2
		ha:ða:	this.msg	2
	Possessive	ʕarus-at -ha:	doll-fsg-her	6
		sidig-a	friend-his	
	Plural	asHa:b-a	friend.pl-his	2
Verbal	Tense	saʔal	ask.pf	10
		ka:n	was.pf	
		ti-ʃrab	imp3fsg-drink.imp	8
	Gender agreement	ti-ʃrab	imp3fsg-drink.imp	4
		ji-Hib	imp3msg-love.imp	4
		ʤalas-at	sit.pf-pf3sg	4
		ka:n-at	was.pf-pf3fsg	1
Adjectival	Gender agreement	ʔil-Hilw-a	the-pretty-fsg	3
		ʔis-so:da:	the-black.fsg	
Preposition		b-ʔil-Hali:b	with-the-milk	15
		ʕan ʔid-di:k	about the-hen	
		ʕala il-kursi:	on the-chair	
		wara: ʔil-fi:l	after the-elephant	
		f-il-ʁa:b-a	in-the-forest-fmsg	
		min	from the-store	
		maʕ	with friend.pl-his	
Pronoun		hij ti-Hib	she imp3fsg-love.imp	5
		hu: ji-ʃrab	he imp3msg-drink.imp	
		ʁasal-at-ha:	wash-pf3fsg-her	
		minn-a	rom-him	
Conjunction		w-ʔin-naml-a	and-the-ant-fsg	2
Copula		ka:n	was.pf	2

Table 4.7 *List of Sentences in the Sentence Repetition Test with English Gloss, in the Order Presented*

Item	Sentence
1	ʃa:f mHammad ʔXwan-a f-il-madras-a Saw.pf prop.n brother.pl-his in the school-fsg Mohammed saw his brothers in school
2	ʔil-walad saʔal sidi:g-a ʕan ʔil-Hafl-a the-boy ask.pf friend-his about the-party-fsg The boy asked his friend about the party

3 Hat-at ʔil-bint daftar-ha: ʕala: it-tawl-a
put-pf3fsg the-girl notebook-her on the-table-fsg
She put her notebook on the table

4 ʤu:d ʃar-at haða: il-galam min ʔil-maHal
prop.n buy-pf3fsg this.msg the-pen from the store
Jude bought this pen from the store

5 ʤalas ʔil-walad ʔit-twi:l ʕala haða: il-kursi:
sit.pf the-boy the-tall on this.msg the-chair
The tall boy sat on this chair

6 maha: ka:n-at ti-sbaH maʕ Xal-ha: il-kibi:r
prop.n was-pf3fsg imp3fsg-swim with uncle-her the-big
Maha was swimming with her uncle

7 nu:ra: ʁasal-at-ha: bi-ʔil-mo:j-a w-is-sabu:n
prop.n wash-pf3fsg-her with-the-water-f and-the-soap
Nora washed it with water and soap

8 ʤar-at ʔil-biss-a is-so:da wara: il-fi:l
run-pf3fsg the-cat-fsg the-black.fsg after the-elephant
The black cat ran after the elephant

9 ʔil-walad is-si:ʁi:r ʔaXað haði: il-ku:r-a minn-a
the-boy the-small take.pf this.fsg the-ball-fsg from-him
The small boy took this ball from him

10 ji-gra: hu: kita:b ʕan ʔid-di:k w-ʔin-naml-a
imp3msg-read he book about the-hen and-the-ant-fsg
He is reading a book about the hen and the ant

11 ka:n na:jif ji-rkið f-il-Hadi:g-a maʕ ʔasHa:b-a
was prop.n imp3msg-run in-the-garden-f with friend.pl-his
Nayif was running in the garden with his friends

12 hu: ji-Hib ji-ʃrab ʔil-Hali:b bi-il-fara:wl-a
He imp3msg-drink the-milk with-the-straberry-fsg
He likes to drink strawberry flavoured milk

13 ti-ʕi:ʃ ʔiz-zara:f-a il-Hilw-a fi: haði: il-ʁa:b-a
imp3fsg-live the-giraffe-fsg the-prett-f3sg this.fsg the-forest-fsg
The pretty giraffe lives in this forrest

14 hij ti-Hib ti-lʕab b-ʕaru:s-at-ha: il-ʤidi:d-a
she imp3fsg-love imp3fsg-play with-doll-fsg-her the-new-fsg
She loves to play with her new toy

4.3.2.1 Procedure of the Sentence Repetition test

The following instructions were given to participants: "I will say a sentence. I want you to listen carefully until I finish the sentence and then repeat the sentence exactly like I said it. Are you ready?" Two practice sentences were presented first to familiarize participants with the test. If a participant did not respond or gave an incorrect response, the examiner repeated the instructions with demonstration of correct response and practice sentences were re-

administered. If the participant repeated the first practice session correctly or after administration of the second practice sentence, irrespective of the participant's response, the 14 target sentences were presented. Sentences were presented in fixed order commencing with the sentence with the fewest Grammatical Morphemes and gradually increased in length. Sentences with the same number of Grammatical Morphemes were positioned randomly in the test. If the participant did not respond to a target sentence or requested a repetition, the sentence was presented again to allow one further opportunity to repeat. The word order Subject Verb__ and Verb Subject__ were alternated as shown in Table 4.6.

4.3.2.2 Scoring of the Sentence Repetition test

Practice sentences did not count towards the participant's score. Any misarticulations that were consistent with the participant's spontaneous speech were not considered errors. Any change of word order from Verb Subject_ to Subject Verb_ or vice versa were not considered errors.

Studies in both English and Cantonese (Redmond, 2005; Stokes et al., 2006) emphasized the importance of using a graded scoring method for SR when it is used to identify children with SLI rather than an all-or-none scoring method. Two scoring methods were adapted from English. The first scoring method was based on the scoring method used in the Sentence Imitation Test (SIT; Seeff-Gabriel, 2006). The SIT scoring is based on three morpho-syntactic categories: content words, function words, and inflections. In Arabic, there is no clear distinction between function words and inflections. Prepositions, for instance, can be function words, as in: **/ʕala: it-twl-a**/ 'on the table' or prefixes as in /**b-ʔil-Hali:b /** 'with-the-milk.' Also, there are grammatical categories that do not take the form of either a function word or inflection for example the vowel pattern that distinguishes number **/walad/** 'boy' vs. /**ʔawla:d/** 'boys.' Therefore, inflections, function words, and patterns were combined as the Grammatical Morpheme category, distinguished from the Lexical Morpheme category. The second scoring method adapted from English was the CELF-4 scoring method.

These two methods yielded the following scores:

1) Adapted from the SIT scoring method, the following scores were obtained:

Lexical Morpheme Score: the number of correctly repeated Lexical Morphemes, including verbs, nouns and adjectives. Maximum score: 56.

Grammatical Morpheme Score: the number of correctly repeated Grammatical Morphemes. See Table 4.5 for relevant categories. Maximum score: 117.

2) Adapted from the CELF scoring method, the following score was obtained:

Total Sentence Accuracy:

- 3: sentence repeated with complete accuracy.
- 2: one morpheme was not repeated correctly.
- 1: two or three morphemes were not repeated correctly.
- 0: four or more morphemes were not repeated correctly.

Maximum score: 42.

4.3.2.2.1 Errors

The type and number of errors were examined for both Grammatical and Lexical Morphemes. Errors were categorized as follows:

- Omission: target morpheme did not occur.
  - For example: Item 13

    ti-ʕi:ʃ ʔiz-zara:f-a il-Hilw-a fi: haði: il-ʁa:b-a

    imp3fsg-live the-giraffe-fsg the-prett-f3sg this.fsg the-forest-fsg
  - Response:

    ti-ʕi:ʃ ʔiz-zara:f-a fi: haði: il-ʁa:b-a

    imp3fsg-live the-giraffe-fsg ∅-∅-∅ this.fsg the-forest-fsg
  - Lexical Morpheme Error: **hilw 'pretty'** was omitted.
  - Grammatical Morpheme Error: **il 'the'** and **a 'fsg'** were omitted
- Substitution: target morpheme replaced with a morpheme that did not occur in the target sentence.
  - For example: Item 9

    ʔil-walad iṣ-ṣi:ʁi:r ʔaXað haði: il-ku:r-a minn-a

    the-boy the-small take.pf this.fsg the-ball-fsg from-him
  - Response:

    ʔil-walad iṣ-ṣi:ʁi:r ʔaXað haða: il-ku:r-a minn-a

    the-boy the-small take.pf this.msg the-ball-fsg from-him
  - Grammatical Morpheme Error: **haði: 'this.fsg'** was replaced with **haða: 'this.msg'**.
- Perseveration: target morpheme was substituted with a different morpheme from the target sentence or the sentence immediately preceding it.
  - For example: Item 10

    ji-gra: hu: kita:b ʕan ʔid-di:k w-ʔin-naml-a

    imp3msg-read he book about the-hen and-the-ant-fsg
  - Item 11

    ka:n na:jif ji-rkiḏ f-il-Hadi:g-a maʕ ʔaṣHa:b-a

    was prop.n imp3msg-run in-the-garden-f with friend.pl-his
  - Response to Item 11

    ka:n na:jif ji-rkiḏ maʕ ʔid-di:k maʕ ʔaṣHa:b-a

    was prop.n imp3msg-run with the-hen with friend.pl-his
  - Lexical Morpheme Error: **Hadi:g-a 'garden-fsg'** was substituted with **di:k 'hen'** (different item from previous sentence).
  - Grammatical Morpheme Error: **fi: 'in'** was substituted with **maʕ 'with'** (different item from the same sentence).

- Unintelligible: target morpheme replaced with a speech segment that is unrecognizable.
    - For example: Item 3

      Hat-at ʔil-bint daftar-ha: ʕala: it-tawl-a

      put-pf3fsg the-girl notebook-her on the-table-fsg
    - Response:

      Hat-at ʔil-bint daftar-ha: ʕala: #-###-#

      put-pf3fsg the-girl notebook-her on
    - Lexical Morpheme Error: **tawl 'table'** was unintelligible.
    - Grammatical Morpheme Error: **it- 'the-'** and **–a '-fsg'** were unintelligible.
- Refusal: participant did not attempt sentence containing the morpheme after two repetitions of the target sentence.
    - The following is an example of scoring a sentence for both accuracy and errors: Item 2

      saʕal ʔil-walad sidi;g-a ʕan ʔil-Hafl-a

      ask the-boy friend-his about the-party-f
    - Response:

      saʔal ∅-bint sidig-a ʕan ʔil-Hafl-a

      ask ∅-girl friend-his about the party

saʕal	Tense	ʔil-	Walad	sidig	-a	ʕan	ʔil-	Hafl	-a
ask	Pf	the-	Boy	friend	-his	About	The	party	-f
1	1	0	0	1	1	1	1	1	1

Lexical	Grammatical	TSA
3	**5**	**1**

TSA = Total Sentence Accuracy

	Omission	Substitution	Unintelligible	Refusal	Perseveration
Lexical	**0**	**1**	**0**	**0**	**0**
Grammatical	**1**	**0**	**0**	**0**	**0**

4.3.3 Anomalous Sentence Repetition test

This novel test consisted of 16 sentences: eight Semantically Anomalous and eight Syntactically Anomalous. The Anomalous sentences were created from eight Typical sentences in the SR test and contain the same target lexical and Grammatical Morphemes as shown in Table 4.8. As in the SR test, both the Semantically and Syntactically Anomalous sentences were 5 to 7 words in length, contained 6 to 11 morphemes per sentence and contained an equal number of Verb Subject_ and Subject Verb_ word order sentences.

Table 4.8 *Distribution of Morphemes in Semantically Anomalous Sentences, Syntactically Anomalous Sentences, and Typical Sentences*

Morpheme Type	Arabic	English	Total
Lexical			

Noun				18
Adjective				4
Verb				10
Grammatical				
Nominal	Determiner	ʔil- bissa	the-cat	18
		ʔis-so:da:	the-black	
	Feminine marker	zara:f-a	giraffe-fsg	8
	Demonstrative	ha:ði:	this.fsg	2
		ha:ða:	this.msg	
	Possessive	ʕarus-at -ha:	doll-fsg-her	3
		sidig-a:	friend-his	
Verbal	Tense	Saʔal	ask.pf	4
		Ti-ʃrab	imp3fsg-drink.imp	6
	Gender agreement	ti-ʃrab	imp3fsg-drink.imp	6
		ji-Hib	imp3fsg-love.imp	
		ʤalas-at	sit.pf-pf3sg	2
Adjectival	Gender agreement	ʔil-Hilw-a	the-pretty-fsg	3
		ʔis-so:da:	the-black.fsg	
Preposition		b-ʔil-Hali:b	with-the-milk	8
		ʕan ʔid-di:k	about the-hen	
		ʕala il-kursi:	on the-chair	
		wara: ʔil-fi:l	after the-elephant	
		f-il-ʁa:b-a	in-the-forest-fmsg	
Pronoun		hij ti-Hib	she imp3fsg-love.imp	3
		hu: ji-ʃrab	he imp3msg-drink.imp	
Conjunction		w-in-naml-a	and-the-ant-fsg	1

In the case of the Semantically Anomalous sentences, the syntactic and morphological structure of the sentences remained the same as in the Typical sentences but the Lexical Morphemes were shuffled between sentences to yield semantically odd sentences. For example:

Typical sentence: (Item 5 in the SR test)

ʤalas ʔil-walad ʔit-twi:l ʕala: haða: il-kursi:

sat.pf the-boy the-tall on this.msg the-chair

The tall boy sat on this chair

Semantically Anomalous sentence: (Item 2 in the ASR test)

ʤalas ʔil-Hali:b ʔil-Hilu ʕala: haða: il-fi:l

sat.pf the-milk the-pretty on the.msg the-elephant

The pretty milk sat on the elephant

In the case of Syntactically Anomalous sentences, the Lexical Morphemes were identical in both the Anomalous and Typical sentences but each sentence contained a total of three rule violations in its syntactic and morphological structure. The violations included a mismatch in subject-verb gender agreement, noun-adjective gender agreement, or demonstrative-noun gender agreement, inappropriate addition of determiner, incorrect preposition-noun order, verb order, noun-adjective order or conjunction-noun order. The type

of violations in each Anomalous sentence varied according to the affordance of the Typical sentence it was created from. For example:

Syntactically Anomalous sentence: (Item 12 in the ASR test)

Note. Items in red highlight the rule violations, items in green highlight the correct form

ʔit-twi:l ʔil-walad ʤalas-at ʕala: haði: il-kursi:

the-tall the-boy sat.pf-pf3fsg on this.fsg the-chair

Violated Rules:

Subject-Verb gender agreement:

ʔil-walad ʤalas-at the-boy sit.pf-pf3fsg

ʔil-walad ʤalas the-boy sit.pf

Demonstrative-Noun agreement:

kursi: haði: chair.m this.fsg

kursi: haða: chair.m this.msg

Noun-Adjective order:

ʔil-walad ʤalas ʕala: it-twi:l the-boy sit.pf on the-tall

ʔil-walad it-twi:l ʤalas ʕala: the-boy the-tall sit.pf on

The full set of Semantically Anomalous sentences is presented in Table 4.9. See Appendix G for score sheets and a breakdown of violations in each Syntactically Anomalous sentence.

Table 4.9 *List of Semantically Anomalous Sentences in the Order Presented*

Item	Sentence
1	ʔid-daftar saʔal zara:f-t-a ʕan ʔil-walad the-notebook ask.pf girrafe-fsg-his about the-boy The notebook asked the giraffe about the boy
2	ʤalas ʔil-Hali:b ʔil-Hilu ʕala: haða: ʔil-fi:l sat.pf the-milk the-pretty on this.msg the-elephant The pretty milk sat on this elephant
3	ʔit-ta:wl-a Hat-at sidi:g-ha ʕala: ʔid-di:k the-table-fsg put-pf3fsf friend-her on the-hen The table put her friend on the hen
4	ʤar-at ʔil-ʁa:b-a ʔit-twi:l-a wara: il-walad run-pf3fsg the-forrest-fsg the-tall-fsg after the-boy The tall forest ran after the boy
5	ji-gra: hu: kursi: ʕan ʔil-walad w-il-ʕaru:s-a imp3msg-read he chair about the-boy and-the-doll-fsg He reads a chair about the boy and the doll

6	hu: ji-Hib ji-ʃrab ʔil- kita:b b-il-naml-a he imp3msg-love imp3msg-drik the-book with-the-ant-fsg He loves to drink the ant flavoured book
7	ti-ʕi:ʃ ʔil-farawl-a il-ʤidi:d-a fi: haði: il-biss-a imp3fsg the-straberry-fsg the-new-fsg in this.fsg the-cat-fsg The new strawberry lives in this cat
8	Hij ti-Hib ti-lʕab b-Hafl-at-ha: ʔis-so:da she imp3fsg-love imp3fsg-play with-party-fsg-her the-black.fsg She loves to play with her black party

4.3.3.1 Procedures for the Anomalous Sentence Repetition test

The test was presented in a fixed order commencing with the easier Semantically Anomalous sentences followed by the Syntactically Anomalous sentences. Within each Anomalous sentence type, sentences were presented based on increasing the length in terms of the number of Grammatical Morphemes in each sentence. The order of presentation for sentences with the same Grammatical Morpheme length was selected randomly. The word order Subject Verb_ and Verb Subject_ were alternated.

The following instructions were given to participants: "I will say some funny sentences. I want you to listen carefully, wait until I stop, then I want you to repeat the sentences exactly like I said them. Are you ready?" Two practice sentences were presented before the 16 test sentences, one for each sentence type. The Semantically Anomalous practice sentence was presented first, followed by the Syntactically Anomalous practice sentence. If the participant did not respond to a practice sentence, requested a repetition or responded incorrectly, the examiner modelled the correct response, the instructions were repeated, and practice sentences were re-administered. After the second administration of the practice sentences, irrespective of the participant's response, the 16 test sentences were administered. As in the SR test, if the participant did not respond to a target sentence or requested a repetition, the sentence was presented again to allow one further opportunity to repeat.

4.3.3.2 Scoring for the Anomalous Sentence Repetition test

Practice sentences did not count toward scores. Recorded responses for test sentences were transcribed and scored offline using the same scoring system as the SR test. The maximum scores for each sentence type were:

Lexical Morpheme Score: 32.

Grammatical Morpheme Score: 69.

Total Sentence Accuracy: 24.

4.3.3.2.1 Errors

The same error categories were examined at morpheme level for both lexical and Grammatical Morphemes as in the SR test.

For example: Item (13) in the ASR test (Syntactically Anomalous):

ji-gra: hiy kita:b ʔid-di:k w-ʔin-naml-a ʕan

imp3msg-read.imp she book the-hen and-the-ant-fsg about

Response:

ji-gra: hiy kita:b ʔid-di:k ʕan

imp3msg-read.imp she book the-hen about

ji-	gra:	tense	hiy	pat	kita:b	w-	id-	di:k	ʕin-	naml	-a	ʕan
imp 3msg-	Read	imp	she		Book	And	The	hen	the	ant	-fsg	about
1	1	1	1	1	1	0	1	1	0	0	0	1
						omit			omit	omit	omit	

Lexical	Grammatical	TSA
3	6	0

	Omission	Substitution	Unintelligible	Refusal	Perseveration
Lexical	1	0	0	0	0
Grammatical	3	0	0	0	0

4.3.4 Nonverbal IQ test

To date, there are no published standardized tests available with Saudi norms. Most studies of Arabic-speaking children use adapted tests from English for both verbal and nonverbal scores with no standardization. For the purpose of this study, the Block Design and Object Assembly subtests of the Wechsler Preschool and Primary Scale of Intelligence-Third Edition (WPPSI-III ^{UK}; Wechsler, 2003) was chosen because

- none of the items in the Object Assembly subtest were culturally inappropriate; for example, it did not contain a puzzle of a pig;
- the same stimuli can be used for the entire age range of participants; and
- good levels of reliability and validity

In the absence of an official Arabic version available, instructions were translated to Arabic as suggested by the test manual. The same order of presentation in the WPPSI-III ^{UK} was followed with Block Design presented first, followed by Object Assembly.

4.4 General Procedure

Each participant was seen individually in a quiet area of the nursery. The assessment session lasted from 20 to 60 minutes. The session length varied depending on the age of the participant; older participants required longer testing sessions because they needed to complete more tasks than younger participants. In the case of the youngest participants (2;6-2;11), a warm-up task was used to assist them in understanding the concept of 'repetition' before the tests were administered. Participants were instructed to copy what the examiner did or said. The

warm-up consisted of three trials, a non-verbal trial (covering face with hands), an animal sound trial (moo), and a single word trial (mama). If the participant did not repeat the item correctly the examiner demonstrated the correct response and repeated the warm-up trial. None of the participants failed to respond to the warm-up task. The warm-up task was added after the development strange, where four children in the youngest age group refused to participate. The benefit of the warm-up task was evident in the pilot stage, where an increased level of participation was noted, none of the children refused to participate.

Because testing was conducted in an unfamiliar room to the participants, it was sometimes reassuring for them to include a classmate in the testing session. Only classmates who already completed their own testing session where allowed to join. They did not participate during testing and were given puzzles to play with.

All the participants were required to complete the VSTM test, the SR test, and the Nonverbal IQ test. Participants 4 years of age and older were asked to complete the ASR test as well. Tests were presented in a fixed order as shown in Figure 4.1. The order of presentation moved from easy to difficult, both within and across tasks. The ASR test was placed after the Nonverbal IQ test to give the participants a break from the verbal tasks and help maintain cooperation.

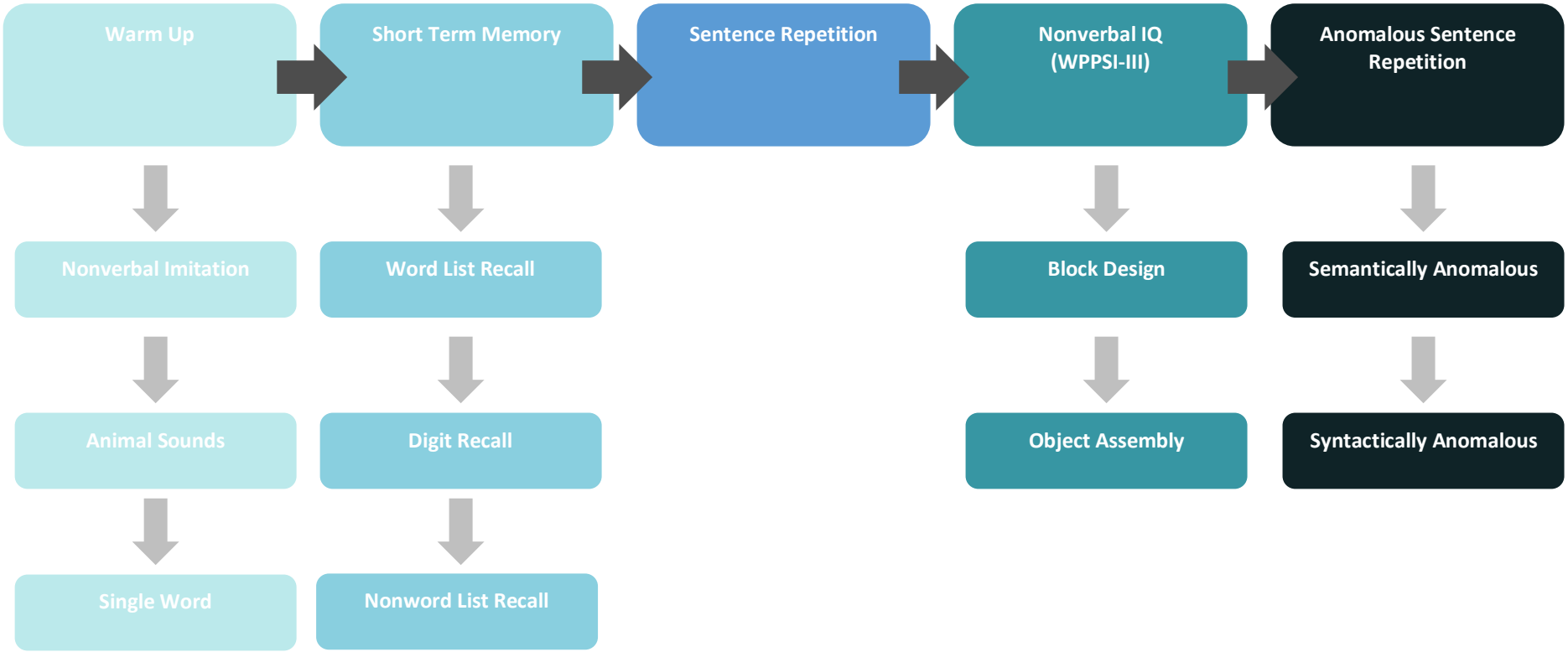


Figure 4.1. Order of presentation of test battery.

Tests were presented live rather than recorded. The evidence suggests that live presentation produces higher levels of compliance in young children and clinically referred children (Chiat & Roy, 2007; Roy & Chiat, 2004; Stokes & Klee, 2009). In a study investigating nonword repetition in participants aged between 28-31 months (Fisher, Hunt, Chambers, & Church, 2001), tape-recorded nonwords were presented via speaker. Out of the 53 participants, 29 gave no responses or irrelevant responses such as "I don't know" to the first 8 out of 32 nonwords and testing was discontinued. In addition, several clinical assessments and screening tools using SR are presented live by the clinician (Carrow, 1974; Gardner et al., 2006; Newcomer & Hammill, 1997; Seeff-Gabriel et al., 2010; Semel et al., 2003; Zimmerman, Steiner, & Pond, 1992). Accordingly, all tests were administered live to ensure high levels of participation and reduce the likelihood of non-response rates that would have reduced the representativeness of the sample and the usefulness of the tests as clinical assessment tools.

The type and frequency of the reinforcement used varied according to the age of participants. Younger participants responded well to puzzles as reinforcement. Older participants responded well to stickers and verbal praise and required reinforcement less frequently than younger participants to maintain cooperation. Reinforcement was provided irrespective of the accuracy of the participant's response. None of the participants included in the study gave a no response to any of the tasks.

All responses were recorded with an Olympus WS-650S digital recorder to allow later transcription. For the VSTM test, incorrect responses were orthographically transcribed for the Digit and Word List Recall subtests and phonetically transcribed for the Nonword List Recall subtest. For the SR and ASR tests responses were orthographically transcribed unless they were unintelligible in which case they were phonetically transcribed.

4.4.1 Reliability

Intra-rater and inter-rater reliability for all measures was obtained for 10% of data. Also, test-retest reliability for all measures except nonverbal IQ was obtained for 10% of data. Retest sessions were scheduled within 1 week of initial test.

Chapter Five: Results

The results chapter is divided into six main sections. The first section presents reliability results for the three experimental assessments: VSTM test, SR test, and ASR test. The second section presents validity results for the VSTM and SR tests. The third section presents results for Typically Developing participants according to age of participants. It includes performance on the Nonverbal IQ test, the influence of gender and school type, and performance VSTM, SR and ASR tests. For each experimental measure, descriptive statistics are presented first followed by inferential statistics. The fourth section presents results for participants with Language Concerns and compares these with Typically Developing participants on the three experimental assessments. It includes the rationale for matching between language status groups as a basis of this comparison, followed by descriptive and inferential statistics. The fifth section explores sensitivity and specificity values of the VSTM and SR tests along with profiles of performance on the experimental measures for participants within the Language Concerns group through z-scores. The sixth section compares the type of errors that occurred in the experimental assessments within and across groups.

5.1 Reliability

Reliability of a test can be defined as its ability to produce consistent results when the same entities are measured under different conditions (Field, 2009). In this study, two types of reliability were examined: scoring across two raters (inter-rater reliability) and scoring across time (test-retest reliability). A 10% sample of the data was selected to evaluate both forms of reliability: one female and one male participant were randomly selected from each age group in the Typically Developing group and one participant from the Language Concerns group, for a total of 15 participants for the VSTM and SR test, and eight participants on the ASR test because it was only administered to participants who were 4 years of age or older. Both types of reliability were checked on all assessments and included different participants.

The intraclass correlation coefficient (ICC) was selected to assess reliability. The ICC was preferred over the Pearson correlation coefficient (PCC). The ICC provides the average correlation between observers or times of testing and indicates the degree of agreement between them. The PCC, on the other hand, is based on regression analysis and indicates whether the relationship between two variables can be expressed via a straight line; for example, rater A consistently scores higher than rater B by 3 points (Streiner & Norman, 1995).

The measure used to report ICC is Cronbach's alpha (α). Acceptable values of α vary by study (Tavakol & Dennick, 2011). Landis and Koch (1977) report that values from .61 to .80

indicate substantial agreement, and from .81 to 1.00 indicate almost perfect agreement. Field (2009) reports an acceptable value of .70 to .80.

5.1.1 Inter-rater reliability

A second rater independently scored the performance of 15 participants for the VSTM and SR tests, and nine participants for the ASR test. The second rater was a native speaker of Arabic, a certified speech and language therapist in Saudi Arabia, and a PhD student in Language and Communication Science. She received training in the scoring system of the experimental assessments and was blind to group membership of participants. The ICC with 95% confidence interval of a mixed model and consistency type are summarized in Tables 5.1 to 5.3. Table 5.1 shows that the ICC values of accuracy measures between raters fell between .95 to 1.00, indicating near perfect agreement.

Table 5.1 *Interclass Correlation Coefficient between Rater 1 and Rater 2 Accuracy Scores*

Assessment		**ICC**
VSTM	Digit Recall Span	.97
	Word List Recall Span	1.00
	Nonword List Recall Span	.98
	Total	.99
SR	Lexical Morpheme Score	.99
	Grammatical Morpheme Score	.99
	Sentence Accuracy Score	.99
ASR Semantic	Lexical Morpheme Score	.97
	Grammatical Morpheme Score	.99
	Sentence Accuracy Score	.99
ASR Syntactic	Lexical Morpheme Score	.97
	Grammatical Morpheme Score	.98
	Sentence Accuracy Score	.95

Note. ASR = Anomalous Sentence Repetition; SR = Sentence Repetition; VSTM = Verbal Short Term Memory.

Table 5.2 shows that the ICC values for most of the error types fell in the category of near perfect agreement. Perseveration, unintelligible errors on all three subtests, and phoneme migration errors in Digit Recall and Word List Recall did not occur often enough to be able to obtain accurate ICC values. Perseveration errors in Nonword List Recall obtained an ICC value of 1.00 because it occurred only once and both scorers agreed on coding the error as perseveration.

Table 5.3 shows that the ICC values for most error types showed substantial to near perfect agreement between raters with values between .63 to .97. As in the VSTM test, it was not possible to obtain accurate ICC values for perseveration errors because they did not occur often enough.

Table 5.2 *Interclass Correlation Coefficients between Rater 1 and Rater 2 Error Classification for Verbal Short Term Memory Subtests*

Error		**DR**	**WLR**	**NLR**
Omission		.92	.98	.91
Migration	Item	.89	.95	n/a*
	Phoneme	n/a	n/a	.85
Perseveration		.77	1	1
Substitution		.87	.85	.85
Unintelligible		n/a	n/a	n/a

Note. DR = Digit Recall span; NLR = Nonword List Recall span; n/a = not applicable; WLR = Word List Recall Span.
* ICC was not computed because the scale had zero variance.

Table 5.3 *Interclass Correlation Coefficients between Rater 1 and Rater 2 Error Classification for Sentence Repetition and Anomalous Sentence Repetition Tests*

Error		**SR**	**ASR Semantic**	**ASR Syntactic**
Lex	Omission	.97	.63	.66
	Substitution	.85	.81	.97
	Perseveration	.69	.86	n/a
Gram	Omission	.92	.99	.90
	Substitution	.68	.89	.65
	Perseveration	.53	.32	n/a

Note. ASR = Anomalous Sentence Repetition; Gram = total Grammatical Morpheme score; Lex = total Lexical Morpheme score; n/a = not applicable; Semantic = Semantically Anomalous Sentences; Syntactic = Syntactically Anomalous Sentences; SR = Sentence Repetition.
* ICC was not computed because the scale had zero variance.

5.1.2 Test-retest reliability

The VSTM and SR tests were re-administered within a week to 15 participants. The ASR test was re-administered to 8 participants because the participant with Language Concerns was younger than 4 years. The same researcher scored the assessments at Times 1 and 2. The ICC with 95% confidence interval of a mixed model and consistency type are summarized in Tables 5.4-5 6.

Table 5.4 shows that the ICC values of accuracy measures for the two test times fell between .81 to .99, indicating almost perfect agreement with the exception of the Nonword List Recall subtest of the VSTM test. The low value of ICC (.28) may be a result of the narrow range of span scores for this subtest across age groups, span scores ranged from 1 to 3 (see Table 5.8). Given the narrow range of span scores for the Nonword List Recall subtest it was retained. Pickering and Gathercole (2001) also reported a low test-retest reliability for the Nonword List Recall subtest in comparison to the other two WMTB-C subtests. The test-retest reliability for

Nonword List Recall was .56 in comparison to.72 for Word List Recall and .81 for Digit List Recall.

Table 5.4 *Interclass Correlation Coefficients between Time 1 and Time 2 Accuracy Scores*

Assessment		**ICC**
VSTM	Digit Recall Span	.81
	Word List Recall Span	.90
	Nonword List Recall Span	.28
	Total	.92
SR	Lexical Morpheme Score	.98
	Grammatical Morpheme Score	.99
	Sentence Accuracy Score	.96
ASR Semantic	Lexical Morpheme Score	.94
	Grammatical Morpheme Score	.93
	Sentence Accuracy Score	.88
ASR Syntactic	Lexical Morpheme Score	.94
	Grammatical Morpheme Score	.87
	Sentence Accuracy Score	.96

Note. ASR = Anomalous Sentence Repetition; SR = Sentence Repetition; VSTM = Verbal Short Term Memory.

Table 5.5 shows that the ICC values for error types for VSTM subtest. With the exception of item migration for Digit Recall, substitution for Word List Recall and omission for Nonword List Recall, the ICC values show poor agreement between Time 1 and Time 2. The poor agreement between Time 1 and Time 2 can be attributed to several reasons. The ICC values indicate the consistency of errors at an individual level rather than a group level. Ten percent of the study sample was included in the analysis, not a 10% representative sample of each error type. However, the error types of most interest are item migration for Digit Recall, substitution for Word List Recall, phoneme migration for Nonword List Recall, and those values range for .50 to .77. Errors were analysed qualitatively only, looking at error trends in each subtest type for all participants irrespective of age.

Table 5.5 *Interclass Correlation Coefficients for Time 1 and Time 2 on Error Classification for Verbal Short Term Memory Subtests*

Error		**DR**	**WLR**	**NLR**
Omission		-	.38	.97
Migration	Item	.77	.47	n/a
	Phoneme	n/a	n/a	.50
Perseveration		-	-	n/a
Substitution		.37	.61	.11
Unintelligible		n/a	-	n/a

Note. DR = Digit Recall span; n/a = not applicable; NLR = Nonword List Recall span; WLR = Word List Recall Span.
*ICC was not computed because the scale had zero variance; -: value is negative due to a negative average covariance among items and violates reliability model assumptions.

Table 5.6 shows the values for error types on the SR and ASR. The ICC values on most tests fell between .68 to .97 indicating substantial to near perfect agreement. The only exception was substitution errors under Grammatical Morpheme in both SR and ASR.

Table 5.6 *Interclass Correlation Coefficients for Time 1 and Time 2 on Error Classification for Sentence Repetition and Anomalous Sentence Repetition Tests*

Error		**SR**	**ASR Semantic**	**ASR Syntactic**
Lex	Omission	.92	.79	.87
	Substitution	.73	.80	.68
	Perseveration	.70	.86	.80
Gram	Omission	.97	.92	.87
	Substitution	.55	.78	.38
	Perseveration	-	-	-

Note. ASR = Anomalous Sentence Repetition; Gram = total Grammatical Morpheme score; Lex = total Lexical Morpheme score; n/a = not applicable; Semantic = Semantically Anomalous Sentences; Syntactic = Syntactically Anomalous Sentences; SR = Sentence Repetition.
*- value is negative due to a negative average covariance among items and violates reliability model assumptions.

5.1.3 Internal consistency

Internal consistency can be defined as the degree of homogeneity between test items or subtests (Streiner & Norman, 1995). Field (2009) reported that values between .7 and .8 indicate good levels of reliability. Internal consistency was calculated for SR and ASR. Table 5.7 shows that Cronbach's alpha values for both tests fell between .86 to .97 indicating good levels of internal reliability for the two tests. In addition, split-half reliability which measures the correlation coefficient between odd and even items of a test ranged from .78 to .92 on both tests. Field (2009) reported that correlation coefficients between .7 to .8 were acceptable. These results support the internal reliability of both the SR and ASR.

Table 5.7 *Internal Consistency for Sentence Repetition and Anomalous Sentence Repetition Tests*

Test	**Items**	**Score**		**α**	***r***
Sentence Repetition	14	Morpheme	Lexical	.97	.92
			Grammatical	.97	.95
		Total Sentence Accuracy		.95	.88
Anomalous Sentence	16	Morpheme	Lexical	.86	.78
			Grammatical	.87	.79

Repetition	Total Sentence Accuracy	.87	.80

5.2 Validity

Validity refers to whether a test measures what it was supposed to measure (Field, 2009). Two forms of validity were measured in this study: construct validity and concurrent validity. Construct validity examines relations between performance on test items or subscales to determine if they reflect a single underlying variable. Construct validity can be assessed by looking at intercorrelations between test subscales. As shown in Table 5.8, strong correlations were found between the SR scores. Similarly, strong correlations were found between the VSTM subtests and Total span with the exception of the Nonword List Recall which showed weak correlation with the Word List Recall.

Concurrent validity refers to the degree to which scores of a new instrument correlate with those of an existing measure (Paul, 2007). Due to the lack of standardized tests in Arabic, validity for the VSTM and SR was assessed by looking at the relationship between the two tests since they both involve immediate repetition. As can be seen in Table 5.8, strong correlations were found between subscales of the two tests with the exception of the Nonword List Recall which showed weak correlations with the scores of the SR test.

Table 5.8 *Partial Correlation Controlling for Age in Months between the Verbal Short Term Memory and Sentence Repetition tests for Typical Sample (n = 140)*

Test		**Verbal Short Term Memory**				**Sentence Repetition**		
		DR	**WLR**	**NLR**	**Total**	**Lex**	**Gram**	**TSA**
Verbal Short Term Memory	DR		.65	**.50**	.91	.62	.60	.51
	WLR			**.33**	.83	.53	.57	.57
	NLR				.69	.33	.28	.38
	Total					.62	.62	.60
Sentence Repetition	Lex						.92	.67
	Gram							.75
	TSA							

Note. All correlations significant at $p < 0.001$ level (2-tailed) for all. Bolding for emphasis. DR = span score of the Digit Recall subtest; WLR = span score of the Word List Recall Subtest; NLR = span score of the Nonword List Recall subtest; Total = the total score of the three Verbal Short Term Memory subtests; Lex = the percent correct Lexical Morpheme score; Gram = the percent correct Grammatical Morpheme score; TSA = Total Sentence Accuracy raw score.

5.3 Typically Developing Participants

A general overview of the statistical methods used for analysis of the experimental tasks is provided in this section. In analysing the performance of the Typical Participants on each of the experimental tasks, descriptive statistics are presented first followed by inferential statistics. Descriptive statistics are presented through tables and graphs. The tables present measures of

central tendency (mean and median) and variability (range and SD). Boxplots and error bars are used to illustrate distributions.

Boxplots illustrate the distribution of the median scores. For illustration, Figure 5.1 shows parallel boxplots of the total repetition score for female and male participants on a SR test (Wallan, 2006). The line across each box represents the median score for each gender. The length of the box represents the inter-quartile range. Whiskers stemming from the box represent the minimum and maximum scores that are within a reasonable distance from the box (not more than 1.5 times the height of the box beyond either quartile) and encompass around 99% of the scores in normally distributed data (Pagano & Gauvreau, 2000). Values beyond the whiskers are considered outliers. Outliers represented by a circle are scores that extend more than 1.5 box lengths from the edge of the box. Extreme outliers are indicated with an asterisk and are scores that extend more than 3 box lengths from the edge of the box. Boxplots indicate the range of scores in the distance between the two whiskers; symmetry of data in the position of the median in the box and the distance of each whisker from the box; and finally, the tail length in the length of each whisker in comparison to the length of the box. We can tell from Figure 5.1 that the female and male participants obtained a similar range of scores on the SR test with no outliers (Wallan, 2006). Unlike the male participants, the female repetition scores lack symmetry and are positively skewed.

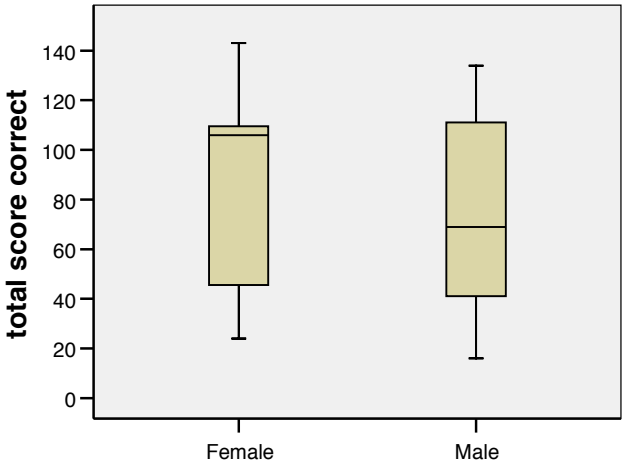


Figure 5.1. Boxplot showing performance on a Sentence Repetition test according to gender of participants (Wallan, 2006).

Error bars illustrate the distribution of the mean scores. For, illustration, Figure 5.2 shows parallel error bars for the Lexical, Preposition, and Overall Gender Agreement repetition scores of participants for each age group (Wallan, 2006). The circle in the centre of the error bars denotes the mean score for that age group and the whiskers represent their 95% confidence intervals. The figure shows that all three score types increase with age, and within each age group Lexical

repetition scores are the highest followed by Preposition and Overall Gender Agreement repetition scores.

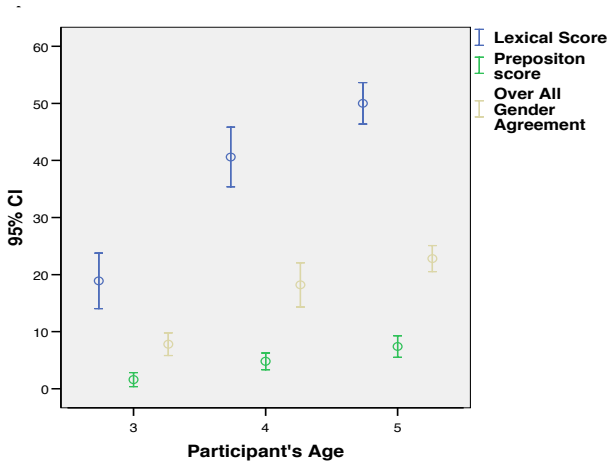


Figure 5.2. Error bar chart of Lexical, Preposition, and Overall Gender Agreement scores according to age (Wallan, 2006).

Inferential statistics indicate whether the independent manipulated variables lead to a difference in the dependent variable. For example, do the span scores of participants differ according to their age? The choice of inferential statistical analysis depends on the type and number of independent variables examined.

There are two types of independent variables: between subjects and within subjects. Between subjects independent variables are manipulated using different participants (Field, 2009); for example, participants from different age groups. Within subjects independent variables are manipulated using the same participants (Field, 2009). For example, in this study each participant provided three span scores for each subtest of the VSTM test. All analysis regarding the Typically Developing participants involved more than two levels of the independent variable (there are more than two age levels); therefore, analysis of variance (ANOVA) was used rather than a t-test.

ANOVA is a parametric test so a number of assumptions have to be met in order to run the analysis (Schiavetti & Metz, 2006). The first assumption is that the data must be normally distributed. The next is the homogeneity of variance. The third is that the data are measured at interval or ratio level. The fourth, the assumption of sphericity, is specific to analyses involving a within subject independent variable with three or more levels (Field, 2009; Schiavetti & Metz, 2006). To assess whether data are normally distributed, in addition to Boxplot figures, the Shapiro-Wilk test was applied, and favoured over the Kolmogrov-Smirnov test because it is more powerful in detecting deviations from normality (Field, 2009). To assess homogeneity of variance, Levene's

test was applied. Significant results in the two tests would indicate a deviation from normal distribution or homogeneity of variance respectively.

Although the data in the following section did not meet the assumptions of normal distribution and homogeneity of variance in several instances, a decision was made to run an ANOVA rather than non-parametric analysis. This is because the F-statistic can be robust and withstand violations of normality when group sizes are equal (Field, 2009). In this study, 20 participants were in each age group. Pring (2005) stated that "raw scores are seldom, if ever, normally distributed…even if one age had normally distributed raw scores, another would not" (p. 48). Parametric analysis is more powerful, and more sensitive to differences, than nonparametric analysis is (Schiavetti & Metz, 2006). Only one-way analysis can be run using nonparametric tests, and not factorial designs involving interactions. However due to violation of assumptions, all parametric analyses were supported with nonparametric analyses, reported in Appendix H. The effect size is reported using partial eta-squared (η_p^2). Cohen's (1992) guidelines are used in interpreting the strength of η_p^2 values (.01: small effect; .06: moderate effect; .14: large effect).

5.3.1 Background assessment: Nonverbal IQ scores

The Nonverbal IQ test was included in the test battery to investigate the nonverbal abilities of participants in both the Typically Developing and Language Concerns groups. Participants were included in the study regardless of their nonverbal IQ score. A further aim was to match participants with Language Concerns with participants in the Typically Developing group based on Nonverbal IQ scores. WPPSI-III ^{UK} (Wechsler, 2003) is a UK-normed test. In the absence of Saudi norms, the scaled scores of the British sample were used in order to compare the performance of the Saudi participants with the British sample. The scaled score of 10 represents the mean and median subtest score for each age group in the British sample and allows for a deviation of ≈ ±3 SD from the mean in each age group.

Table 5.9 shows the mean, median, SD, minimum, and maximum values for each age group and all the age groups combined for each subtest. The overall mean for both subtests and the SD fall within the accepted norms for the British sample. In taking a closer look at performance on the Block Design subtest, there is a wide range of scores with most of the participants falling within the accepted ± 3 SD from the mean, with some participants in the younger age groups performing above the average range with means ranging from 10.5 to 11.0 and SD values ranging from 2.29 to 2.48. Some participants in the older age groups scored below the average range. Means ranged from 9.9 to 11.3, and SD values ranged from 2.75 to 3.27.

The overall mean for the Object Assembly subtest was lower than the overall mean for the Block Design subtest (M = 8.72, M = 10.55, respectively). Participants performed below the mean subtest scaled score of 10 in all age groups, with means that ranged from 7.65 to 9.55. In taking a

closer look, 56% of participants scored equally on both subtests, nearly two-thirds of participants scored higher in the Block Design subtests, of which 10% scored substantially higher in the Block Design subtest with a difference of up to 6 points between the scaled scores. Only 17% of participants scored higher in the Object Assembly subtest. One possible explanation of the lower performance in the Object Assembly subtest is the lack of familiarity of individual items or the task itself. Due to this difference, it was decided that participants in the Language Concerns group were matched on age and the Block Design subtest only.

Table 5.9 *Descriptive Statistics for Block Design and Object Assembly of Wechsler Preschool and Primary Scale of Intelligence-Third Edition for Typically Developing Participants, According to Age*

Subtest	Age Group	Mean	Median	SD	Min	Max
Block Design	2.6 - 2.11	10.50	10	2.48	7	16
	3.0 - 3.5	11.00	11	2.29	7	15
	3.6 – 3.11	10.70	10	2.34	7	16
	4.0 – 4.5	9.90	9	3.02	6	17
	4.6 – 4.11	10.05	9	3.27	6	17
	5.0 – 5.5	10.40	11	3.08	6	16
	5.6 – 5.11	11.30	11.5	2.75	6	17
	Overall	10.55	10	2.75	6	17
Object Assembly	2.6 - 2.11	8.55	9	2.21	6	12
	3.0 - 3.5	8.45	8	2.21	5	13
	3.6 – 3.11	9.55	9.5	2.93	2	15
	4.0 – 4.5	7.65	7.5	1.98	3	11
	4.6 – 4.11	8.90	9	2.65	5	15
	5.0 – 5.5	9.15	9	2.16	5	12
	5.6 – 5.11	8.55	9	2.21	6	12
	Overall	8.72	9	2.41	2	15

** $p < 0.001$
* $p < 0.01$

5.3.2 Gender and school type

The aim of this section is to examine if gender and school type influenced performance on the three novel tests: VSTM, SR, and ASR. Since establishing the sensitivity of the novel assessments to age across a representative sample was a primary focus of this study, age groups were matched for gender and school type. Table 5.10 and 5.11 show the mean, median, and SD ages for girls/boys and children attending public/private schools within each 6-month age band. The distribution of age was very similar across genders and school types; therefore, age was not taken into account in examining the effect of gender or school type.

Table 5.10 *Distribution of Typically Developing Participants According to Age (6-month Age Bands) and Gender*

Age group	Girls				Boys			
	N	**Mean age, months**	**Median age, months**	**SD**	***n***	**Mean age, months**	**Median age, months**	**SD**
2;6 – 2;11	11	31.64	30.00	1.96	9	32.11	32.00	2.21
3;0 – 3;5	10	38.60	39.00	1.51	10	39.00	39.00	1.76
3;6 – 3;11	10	45.00	45.00	1.33	10	44.40	44.50	1.90
4;0 – 4;5	10	51.00	51.50	1.83	10	51.20	52.00	1.99
4;6 – 4;11	10	55.80	55.00	1.32	10	56.00	56.00	1.76
5;0 – 5;5	10	62.80	63.00	1.32	10	62.80	63.50	2.10
5;6 – 5;11	10	68.40	68.50	1.43	10	68.70	69.50	1.83
Total	71	50.20	51.00	12.41	69	50.87	52.00	12.18

Table 5.11 *Distribution of Typically Developing Participants According to Age (6-month Age Bands) and School Type*

Age group	Public				Private			
	N	**Mean age, months**	**Median age, months**	**SD**	***n***	**Mean age, months**	**Median age, months**	**SD**
2.6 – 2.11	10	31.80	30.00	2.35	10	31.90	32.50	1.79
3.0 – 3.5	10	38.80	39.00	1.55	10	38.80	39.00	1.75
3.6 – 3.11	10	44.30	44.50	1.57	10	45.10	45.50	1.66
4.0 – 4.5	10	51.20	52.00	2.04	10	51.00	52.00	1.76
4.6 – 4.11	10	55.90	55.00	1.45	10	55.90	56.00	1.66
5.0 – 5.5	10	62.00	61.50	1.70	10	63.60	64.00	1.35
5.6 – 5.11	10	68.10	68.00	1.66	10	69.00	69.50	1.49
Total	70	50.30	52.00	12.14	70	50.76	52.00	12.46

Tables 5.12 and 5.13 report the scores on all three tests for girls vs. boys and public vs. private school participants. Since most of the scores violated the assumption of normality (see results of Shapiro-Wilk test in Appendix H), the non-parametric Mann-Whitney test was used to compare performance of participants on all tests. Results revealed no significant gender or school effects on any of the tests. Consequently, gender and school type were not included in subsequent analyses. The lack of gender effect falls in line with previous research for VSTM (Alloway et al., 2006; Nadler & Archibald, 2014), SR (Gardner et al., 2006; Wallan, 2006), and ASR (Polišenská, 2011).

Table 5.12 *Descriptive Statistics for Typically Developing Participants on All Measures According to Gender*

Test		**Gender**	***n***	**Mean**	**Median**	**SD**
VSTM	Digit Recall	Girls	71	3.27	3.50	.90
		Boys	69	3.28	3.00	.88
	Word List Recall	Girls	71	2.87	3.00	.75
		Boys	69	3.05	3.00	.80
	Nonword List Recall	Girls	71	1.44	1.00	.49
		Boys	69	1.56	1.50	.52
	Total	Girls	71	7.58	7.50	1.97
		Boys	69	7.88	8.00	1.98
SR	Lexical Morpheme	Girls	71	41.73	48.00	12.86
		Boys	69	42.10	47.00	13.08
	Grammatical Morpheme	Girls	71	79.20	93.00	31.37
		Boys	69	78.62	92.00	31.50
	Total Sentence Accuracy	Girls	71	16.92	18.00	12.80
		Boys	69	16.68	17.00	12.83
ASR	Semantic Lexical Morpheme	Girls	40	26.53	28.00	4.03
		Boys	40	27.85	28.00	3.12
	Semantic Grammatical	Girls	40	55.98	56.50	7.60
	Morpheme	Boys	40	57.70	58.00	6.65
	Semantic Total Sentence	Girls	40	11.75	12.00	5.91
	Accuracy	Boys	40	13.23	13.00	5.59
	Syntactic Lexical Morpheme	Girls	40	24.33	24.00	4.53
		Boys	40	24.63	24.00	3.68
	Syntactic Grammatical	Girls	40	44.58	44.50	8.30
	Morpheme	Boys	40	44.75	43.50	7.06
	Syntactic Total Sentence	Girls	40	5.48	5.00	3.80
	Accuracy	Boys	40	5.08	4.00	3.72

Table 5.13 *Descriptive Statistics for Typically Developing Participants on All Measures According to School Type*

Test		School	*n*	Mean	Median	SD
VSTM	Digit Recall	Public	70	3.34	3.00	0.89
		Private	70	3.20	3.25	0.88
	Word List Recall	Public	70	3.04	3.00	0.76
		Private	70	2.89	3.00	0.79
	Nonword List Recall	Public	70	1.52	1.50	0.53
		Private	70	1.47	1.50	0.48
	Total	Public	70	7.90	7.75	2.01
		Private	70	7.56	8.00	1.94
SR	Lexical Morpheme	Public	70	43.57	48.00	11.19
		Private	70	40.26	47.00	14.34
	Grammatical Morpheme	Public	70	82.30	93.50	29.77
		Private	70	75.53	86.50	32.65
	Total Sentence Accuracy	Public	70	18.16	20.00	13.34
		Private	70	15.44	17.00	12.11
ASR	Semantic Lexical Morpheme	Public	40	27.55	28.00	3.56
		Private	40	26.83	28.00	3.73
	Semantic Grammatical	Public	40	57.13	59.50	7.42
	Morpheme	Private	40	56.55	57.00	6.70
	Semantic Total Sentence	Public	40	13.10	14.00	5.80
	Accuracy	Private	40	11.88	11.00	5.75
	Syntactic Lexical Morpheme	Public	40	24.63	24.50	4.21
		Private	40	24.33	24.00	4.03
	Syntactic Grammatical	Public	40	45.15	44.50	7.05
	Morpheme	Private	40	44.18	43.00	8.28
	Syntactic Total Sentence	Public	40	5.45	4.50	3.62
	Accuracy	Private	40	5.10	4.50	3.88

5.3.3 Verbal Short Term Memory test

The VSTM test consists of three subtests: Word List Recall; Digit Recall, and Nonword List Recall. Each subtest provided a span score representing the highest number of items (Digits, Words, Nonwords) repeated with complete accuracy and in the correct serial position by a participant in four of six trials. A score of 0.5 was awarded if a participant accurately repeated items in three of six trials. In addition, a Total Span score was calculated based on the sum of span scores of the three subtests. The results of the VSTM test are presented in two main sections: the first section presents a comparison of the performance of participants on the three subtests of the VSTM test according to the age of participants; the second section presents the performance of participants on the Total Span score according to age of participants. Within each section descriptive statistics are presented first followed by inferential statistics.

5.3.3.1 A comparison of Verbal Short Term Memory subtest span scores

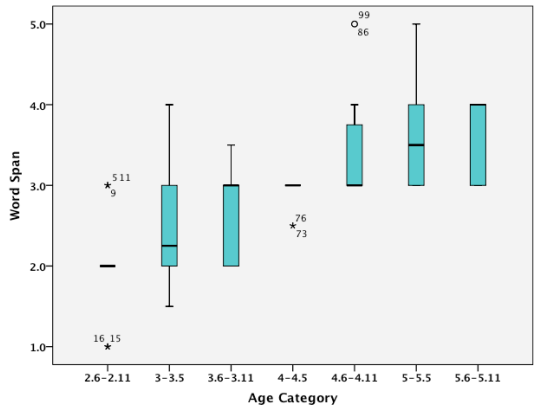
Descriptive statistics for each of the subtests of the VSTM test are presented in Table 5.14 and show the mean, median, maximum, and minimum span scores according to the age group of

participants. Figure 5.3 (a-c) illustrates parallel boxplots of span scores for the three subtests and illustrates the distribution and median span scores according to the age group of participants. Figure 5.3 (d) shows error bars of the mean span scores and the 95% confidence intervals of the three subtests according to the age group of participants and illustrates the relationship between the three subtests.

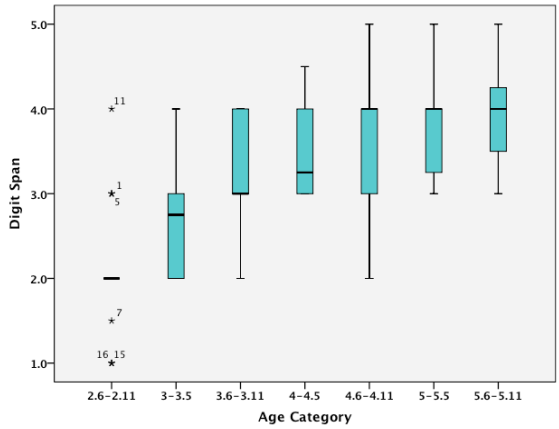
Table 5.14 *Descriptive Statistics for the Verbal Short Term Memory Subtests for Typically Developing Participants According to Age (6-month Age Bands)*

Subtest	Age Group	*N*	Mean	Median	SD	Min	Max
Word List Recall	2.6 – 2.11	20	2.00	2	0.56	1	3
	3.0 – 3.5	20	2.45	2.25	0.60	1.5	4
	3.6 – 3.11	20	2.68	3	0.49	2	3.5
	4.0 – 4.5	20	2.95	3	0.15	2.5	3
	4.6 – 4.11	20	3.43	3	0.65	3	5
	5.0 – 5.5	20	3.55	3.5	0.60	3	5
	5.6 – 5.11	20	3.68	4	0.47	3	4
Digit Recall	2.6 – 2.11	20	2.08	2	0.73	1	4
	3.0 – 3.5	20	2.68	2.75	0.65	2	4
	3.6 – 3.11	20	3.28	3	0.64	2	4
	4.0 – 4.5	20	3.48	3.25	0.53	3	4.5
	4.6 – 4.11	20	3.63	4	0.69	2	5
	5.0 – 5.5	20	3.83	4	0.59	3	5
	5.6 – 5.11	20	3.95	4	0.65	3	5
Nonword List Recall	2.6 – 2.11	20	1.00	1	1.00	1	1
	3.0 – 3.5	20	1.10	1	0.07	1	2
	3.6 – 3.11	20	1.45	1.25	0.48	1	2
	4.0 – 4.5	20	1.60	2	0.48	1	2
	4.6 – 4.11	20	1.50	1.50	0.49	1	2
	5.0 – 5.5	20	2.00	2	0.28	1.5	3
	5.6 – 5.11	20	1.83	2	0.47	1	2.5

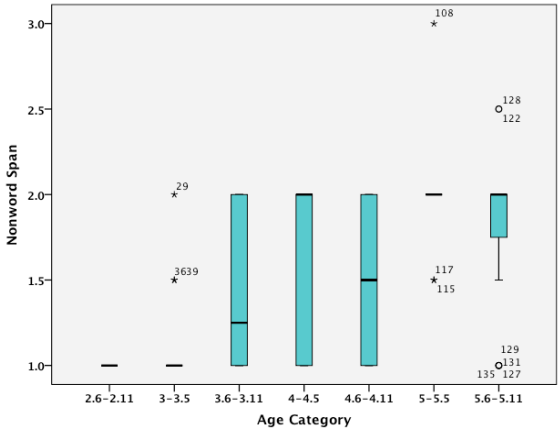
(a) Word List Recall



(b) Digit Recall



(c) Nonword List Recall



(d) Error bars for the three subtests

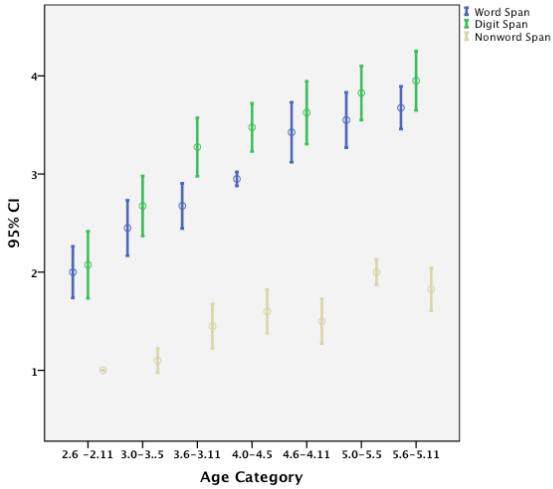


Figure 5.3. Boxplots and error bars showing Typically Developing group's performance on Verbal Short Term Memory subtests according to age.

The lack of normal distribution of subtest span scores is evident from Table 5.14 and Figure 5.3 (a-c). This is due to the narrow range of scores within age groups. For example, the youngest age group (2;6 to 2;11), both the mean and median was approximately a span of 2 for Word List and Digit Recall subtests. Any participant who scored above or below 2 is presented as an extreme outlier in Figure 5.3 (a). In the Word List Recall subtest, all of the participants in the age group 4;0 to 4;5 had a span of 3 except two participants who presented as extreme outliers. For Nonword List Recall, the mean and median span scores ranged from 1 to 2 across the seven age groups. With the exception of the youngest age group, there is no one single participant who was an extreme outlier in all three subtests.

The VSTM subtest span scores for Word List Recall and Digit Recall improved with age. The Nonword List Recall span scores showed the least improvement with age. The observed increase in span scores of adjacent 6-month age bands decreased as age increased. In comparing the Digit Recall and Word List Recall subtests, there was an uneven surge in span score development. In the Word List Recall subtest, the largest gain in mean span score occurred in the age group 4;6 to 4;11 where the mean Span score increased from 2.95 to 3.43. For the Digit Recall subtest, the largest gain in mean Span score occurred earlier, in the age group 3;6 to 3;11 where the mean Span score increased from 2.68 to 3.28.

Figure 5.3 (d) shows that Span scores for Digit Recall and Word List Recall are greater than Span scores for Nonword List Recall. The gap between the two subtests and Nonword List Recall grows as age increases. There is an overlap between the Span scores of Word List Recall and Digit Recall subtests across age groups. Digit Recall Span scores show a slight advantage over Word List Recall Span scores. This advantage is most evident in the two age groups 3;6 to 3;11 and 4;0 to 4;5.

To investigate the effects of age and subtest type on span scores a two-factor mixed design ANOVA was employed, with age as a between-subject variable with seven 6-month age bands and subtest type as a within-subject variable with three levels (Word List, Digit, and Nonword List Recall). Span score was the dependent variable.

The assumptions of normality and homogeneity of variance were violated. The Shapiro-Wilk test of normality was significant ($p < .05$) indicating non-normal distribution for span scores of the three subtests within all seven age groups (details in Appendix H); this finding was illustrated earlier in Figure 5.3 (a-c). Levene's test was significant for Word List Recall subtest $F(6,133) = 5.07$, $p < .001$ due to the small variance in span scores of the youngest age group (2;6 to 2;11) and for the Nonword List Recall subtest $F(5,114) = 11.42$, $p < .001$ due to the small variance of span scores across age groups. The assumption of sphericity was met; Mauchly's test was not significant, $\chi^2(2) = 2.25$, $p > .05$. Supporting non-parametric analysis can be found in Appendix H.

Results show a significant effect of age $F(6,133) = 31.05$, $p < .001$, $\eta_p^2 = .58$, with span scores increasing as age increased. A significant effect of subtest type was found $F(2,266) = 909.09$, $p < .001$, $\eta_p^2 = .87$. All effect sizes were large. To follow up the main effect of age the Games-Howell *post-hoc* test was applied. This test was chosen because the assumption of homogeneity of variance was violated (Field, 2009). Table 5.15 shows the mean difference between age groups, the significance of the difference, the standard error, and 95% confidence intervals. There was no significant difference between adjacent 6-month age groups and it extended to no significant difference in 1-year age bands for participants in the age groups 3;0 to 3;5, 3;6 to 3;11, and 4;6 to 4;11.

Table 5.15 *Comparison of Mean Difference in Span Scores between Age Groups*

Age Category Comparison		Mean Difference	Standard Error	95% CI Lower Bound	95% CI Upper Bound
2;6-2;11	3-3;5	-0.38	.14	-0.81	0.04
	3;6-3;11	-0.78**	.14	-1.21	-0.34
	4-4;5	-0.98**	.12	-1.35	-0.62
	4;6-4;11	-1.16**	.15	-1.63	-0.69
	5-5;5	-1.43**	.13	-1.83	-1.03
	5;6-5;11	-1.46**	.14	-1.89	-1.03
3-3;5	3;6-3;11	-0.39	.14	-0.84	0.06
	4-4;5	-0.60**	.12	-0.98	-0.22
	4;6-4;11	-0.78**	.15	-1.26	-0.29
	5-5;5	-1.05**	.13	-1.47	-0.64
	5;6-5;11	-1.08**	.14	-1.52	-0.63
3;6-3;11	4-4;5	-0.21	.12	-0.60	0.18
	4;6-4;11	-0.38	.16	-0.87	0.10
	5-5;5	-0.66**	.14	-1.08	-0.24
	5;6-5;11	-0.68*	.14	-1.13	-0.23
4-4;5	4;6-4;11	-0.18	.14	-0.61	0.26
	5-5;5	-0.45*	.11	-0.80	-0.10
	5;6-5;11	-0.48*	.12	-0.86	-0.09
4;6-4;11	5-5;5	-0.28	.15	-0.73	0.18
	5;6-5;11	-0.30	.16	-0.79	0.19
5-5;5	5;6-5;11	-0.03	.13	-0.44	0.39

** $p < 0.001$
*$p < 0.01$

To follow up the main effect of subtest type, the adjusted Bonferroni *post-hoc* test was applied. Table 5.16 shows the mean difference between subtest Span scores, the significance of the difference, the standard error, and 95% confidence intervals. A significant difference was found between all three subtests. Both Word List and Digit Recall Span scores were higher than Nonword List Span scores. Digit Recall Span scores were slightly higher than Word List Recall Span scores.

Table 5.16 *Comparison of Mean Difference in Span Scores between Verbal Short Term Memory Subtests*

Subtest Comparison	Mean Difference	Standard Error	95% CI Lower Bound	95% CI Upper Bound
Digit vs. Word	0.31*	0.04	0.21	0.41
Digit vs. Nonword	1.78*	0.05	1.66	1.89
Word vs. Nonword	1.46*	0.05	1.36	1.57

* $p < 0.001$ (adjusted for multiple comparisons)

A significant interaction effect (age x subtest) was found $F(12,266) = 6.68$, $p < .001$, $\eta_p^2 = .23$; this indicated that age influenced span scores of the three subtests differently. Figure 5.4 shows the age x subtest interaction graph. Across tests, there was a greater change in the Word List and Digit Recall span scores in comparison to Nonword List Recall. There was also a difference between how age affected span score with in Digit and Word List Recall. In the Digit Recall subtest, the three youngest age bands age bands showed a greater difference between span scores, while there was more of a spread between span scores of Word List Recall according to age. To follow-up the interaction, three one-way ANOVAs with seven levels for the 6-month age bands were employed for each subset. Results revealed that age was a significant factor for all three subtests. To explore the advantage of Digits over Words a paired t-test was conducted for each level and showed that the advantage was significant in only two age groups: 3;6 to 3;11 and 4;00 to 4;05. To explore the advantage of Words over Nonwords, paired t-test were employed for each age level and revealed that the gap was significant for all age levels.

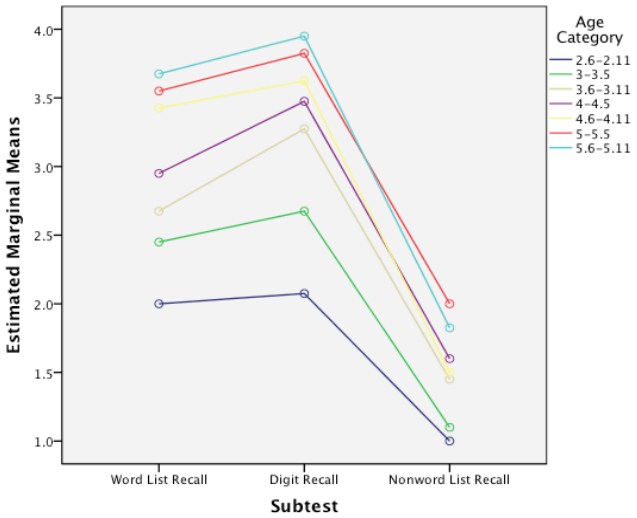


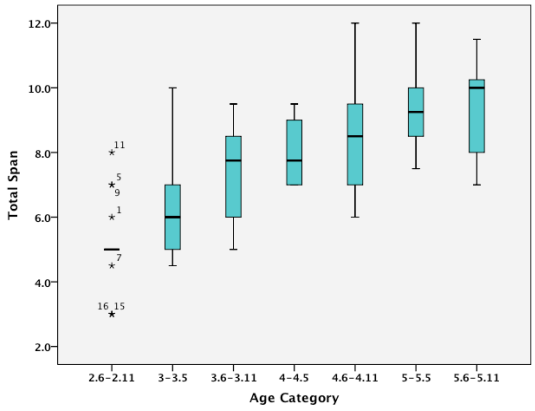
Figure 5.4. Age x subtest interaction graph.

5.3.3.2 A comparison of Total Span scores

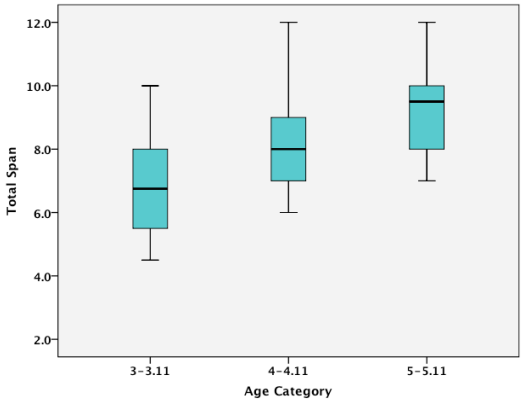
Figure 5.5 (a) shows the boxplots for median Total Span score of participants according to 6-month age bands. It illustrates the narrow range of Total Span scores in the youngest age group and the minimal increase in scores between adjacent 6-month age bands. Therefore, accordingly the youngest age group was omitted and the 6-month age bands were combined into 1-year age bands. Descriptive statistics for Total Span scores according to 1-year age bands are presented in Table 5.17 and show the mean, median, SD, minimum, and maximum Total Span scores. Figure 5.5 (b) shows the boxplots for median Total Span scores according to 1-year age bands. Both Table

5.17 and Figure 5.5 (b) show a wider range of Span scores in 1-year age bands in comparison to 6-month age bands, better symmetry of Total Span scores, an increase in between group differences and no outliers. Figure 5.5 (c) shows the error bars of mean Total Span score according to 1-year age bands and illustrates the increase in mean Total Span scores as age increased.

(a) Total Span score for 6-month age bands



(b) Total Span score for 1-year age bands



(c) Total Span score for 1-year age bands (error bars)

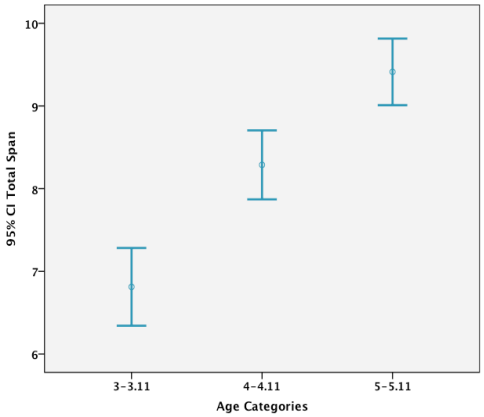


Figure 5.5. Boxplots and error bars of Total Span scores of Typically Developing participants according to age groups

Table 5.17 *Descriptive Statistics for Verbal Short Term Memory Total Scores of Typically Developing Participants According to Age (in 1-year age Bands)*

Age Group	*n*	Mean	Median	SD	Min	Max
3.0 – 3.11	40	6.81	6.75	1.47	4.5	10
4.0 – 4.11	40	8.29	8	1.31	6	12
5.0 – 5.11	40	9.41	9.5	1.26	7	12

To investigate the effect of age on Total Span scores, a one-way ANOVA was employed with age as a between-subjects variable with three levels. The assumption of normality was violated as indicated by the significant result of the Shapiro-Wilk test ($p < .05$) for Total Span scores of the middle age group (4;0-4;11). The assumption of homogeneity of variance was met. Supporting nonparametric analysis can be found in Appendix H.

Results revealed a significant effect of age $F(2,117) = 37.4$, $p < .001$, $\eta_p^2 = .39$ with Total Span scores increasing as age increased. The Bonferroni *post-hoc* test was applied to follow-up this finding. Table 5.18 shows the mean difference between Total Span scores according to age, the significance of the difference, the standard error, and 95% confidence intervals. Results of the *post-hoc* test show a significant difference between all three age groups.

Table 5.18 *Comparison of Mean Difference Total Span Scores between Age Groups*

Age Category Comparison	Mean Difference	Standard Error	95% CI	
			Lower Bound	Upper Bound
3-3;11 vs. 4-4;11	-1.48**	0.3	-2.21	-0.74
3-3;11 vs. 5-5;11	-2.60**	0.3	-3.33	-1.87
4-4;11 vs. 5-5;11	-1.13*	0.3	-1.86	-0.39

** p < 0.001
* p < 0.01

5.3.4 Sentence Repetition test

The SR test provided three scores: a Lexical Morpheme score equal to the number of correctly repeated Lexical Morphemes with a maximum score of 56; a Grammatical Morpheme score equal to the number of correctly repeated Grammatical Morphemes with a maximum score of 117; and a Total Sentence Accuracy score based on a three point scoring system similar to the (CELF-4; Semel et al., 2003) with a maximum score of 42.

The results of the SR test are presented in two main sections: the first section presents a comparison of Lexical and Grammatical Morpheme scores according to the age of participants; the second section presents the Total Sentence Accuracy score according to the age of participants. Within each section the descriptive statistics are presented first followed by inferential statistics.

5.3.4.1 A comparison of Lexical and Grammatical Morpheme scores

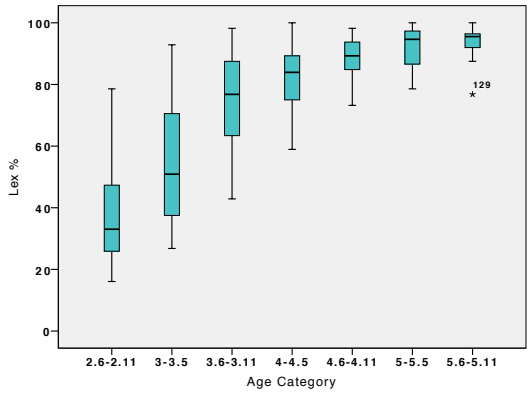
As the maximum scores for Lexical and Grammatical Morphemes differed, their raw scores were converted into percentages in order to allow for a comparison of scores on the two measures. Descriptive statistics for Lexical and Grammatical Morpheme scores according to the age of participants are presented in Table 5.19 and show the mean, median, maximum, and minimum scores for both measures according to age group. Figure 5.6 (a-b) illustrates the distribution of the median scores for Lexical and Grammatical Morphemes through boxplots for

each age group. Figure 5.7 provides error bars for the mean Lexical and Grammatical Morpheme scores for each age group together with their 95% confidence intervals and illustrates the increase in both scores with age and the relationship between the two types of Morpheme scores across the seven age groups.

Table 5.19 *Descriptive Statistics for the Sentence Repetition Test Morpheme Scores for Typically Developing Participants According to Age (in Percentages)*

Score	Age Group	*n*	Mean	Median	SD	Min	Max
	2.6 – 2.11	20	39.02	33.04	16.76	16.07	78.57
Lexical	3.0 – 3.5	20	53.93	50.89	18.83	26.79	92.86
Morpheme	3.6 – 3.11	20	74.20	76.79	16.44	42.86	98.21
	4.0 – 4.5	20	81.61	83.93	10.46	58.93	100.00
	4.6 – 4.11	20	88.75	89.29	6.15	73.21	98.21
	5.0 – 5.5	20	92.32	94.64	6.22	78.57	100.00
	5.6 – 5.11	20	94.11	95.54	5.25	76.79	100.00
	2.6 – 2.11	20	25.60	15.81	18.52	8.55	71.79
Grammatical	3.0 – 3.5	20	43.50	36.32	21.99	18.80	93.16
Morpheme	3.6 – 3.11	20	66.28	65.81	18.13	32.48	91.45
	4.0 – 4.5	20	75.13	76.50	11.61	47.86	89.74
	4.6 – 4.11	20	83.55	83.76	8.30	65.81	96.58
	5.0 – 5.5	20	86.24	87.18	8.64	68.38	98.29
	5.6 – 5.11	20	91.84	94.02	5.18	80.34	98.29

(a) Lexical Morpheme score (percentage)



(b) Grammatical Morpheme score (percentage)

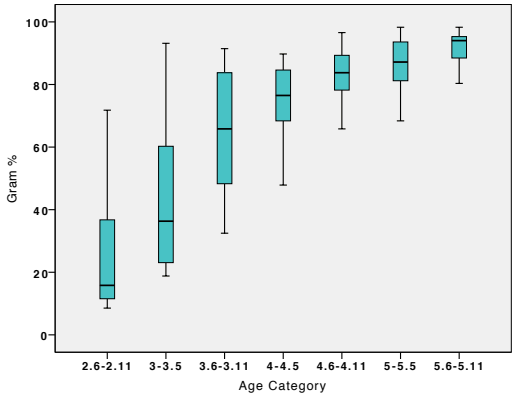


Figure 5.6. Boxplots showing Typically Developing group's performance on the Sentence Repetition test according to age

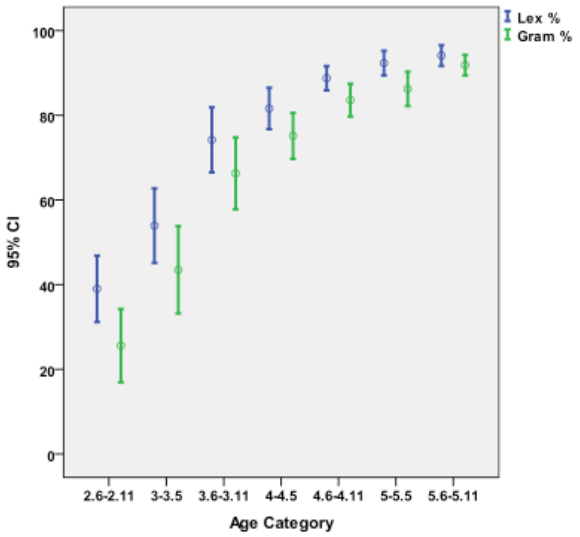


Figure 5.7. Error bars for Sentence Repetition scores of Typically Developing participants according to age groups (in percentage)

The median for Lexical and Grammatical Morpheme scores increased with age. The largest increase in both Morpheme scores occurred in the age 2;6 to 3;11. In contrast in the older age 4;0 to 5;11, the observed increase in scores between adjacent age bands lessened and ceiling effects were evident. Participant 129's Lexical Morpheme score emerged as an extreme outlier, substantially lower than the scores of other children in the oldest age group. The ranges of Lexical and Grammatical Morpheme scores were widest in the three youngest age groups. Participants in the oldest age group 5;6 to 5;11, showed the smallest range with participants performing close to ceiling. Scores deviated from normal distribution in some age groups as can be observed from the asymmetry in age groups 2;6 to 2;11, 5;0 to 5;5, and 5;6 to 5;11 for Lexical Morpheme scores, and 2;6 to 2;11, 3;0 to 3;5 and 5;6 to 5;11 for Grammatical Morpheme scores. Figure 5.7 shows that the mean scores for Lexical Morphemes were higher than that of Grammatical Morphemes, with almost no overlap in the youngest age group. The gap between the two scores gradually reduced as age increased until both percentage scores were close to ceiling in the oldest age group.

To investigate the effects of age and morpheme type, a two-factor mixed design ANOVA was employed, with age as a between-subject variable with seven 6-month age bands and morpheme type as a within-subject variable with two levels: Lexical and Grammatical; repetition score was the dependent variable. The two assumptions of parametric analysis were not met. As reported earlier, Lexical and Grammatical Morpheme scores were not normally distributed, which was further confirmed with significant results of Shapiro-Wilk test of normality in the youngest and two oldest age groups for the Lexical Morpheme scores and the two youngest and the oldest age groups for the Grammatical Morpheme scores (full details can be found in Appendix H). The

assumption of homogeneity of variance was also violated; Levene's test was significant for Lexical Morpheme score $F(6,133) = 13.44$, $p < .001$ and Grammatical Morpheme score $F(6,133) = 10.11$, $p < .001$. All parametric analyses were supported with non-parametric analyses are in Appendix H.

The main effect of age was significant $F(6,133) = 59.16$, $p < .001$, $\eta_p^2 = .73$ with scores increasing as age increased. The main effect of morpheme type was significant $F(1,133) = 193.8$, $p < .001$, $\eta_p^2 = .59$. Overall Grammatical Morpheme mean percentage scores were significantly lower than Lexical Morpheme mean percentage scores. The interaction of age and morpheme type was also significant. $F(6,133) = 6.69$, $p < .001$, $\eta_p^2 = .23$. Effect sizes were large (Cohen, 1992).

The main effect of age was followed-up with the Games-Howell *post-hoc* test. Results are summarized in Table 5.20 and show that the mean percentage scores of the age group 3;6 to 3;11 and the two youngest age groups differed significantly from all age groups except one adjacent 6-month age band either above or below. The mean percentage scores of the age group 4;0 to 4;5 and older did not differ from immediately adjacent 6-month age bands.

Table 5.20 *Comparison of Mean Difference in Morpheme Scores between Age Groups*

Age Category Comparison		**Mean Difference**	**Standard Error**	**95% CI**	
				Lower Bound	**Upper Bound**
2;6-2;11	3-3;5	-16.41	5.91	-34.82	2.01
	3;6-3;11	-37.93**	5.44	-54.84	-21.02
	4-4;5	-46.06**	4.57	-60.43	-31.69
	4;6-4;11	-53.84**	4.18	-67.24	-40.44
	5-5;5	-56.97**	4.20	-70.41	-43.53
	5;6-5;11	-60.66**	4.03	-73.73	-47.60
3-3;5	3;6-3;11	-21.52*	5.84	-39.74	-3.30
	4-4;5	-29.65**	5.05	-45.63	-13.67
	4;6-4;11	-37.43**	4.71	-52.58	-22.29
	5-5;5	-40.56**	4.72	-55.74	-25.38
	5;6-5;11	-44.26**	4.57	-59.12	-29.40
3;6-3;11	4-4;5	-8.13	4.48	-22.22	5.97
	4;6-4;11	-15.91*	4.09	-29	-2.82
	5-5;5	-19.04*	4.11	-32.17	-5.91
	5;6-5;11	-22.73**	3.94	-35.48	-9.99
4-4;5	4;6-4;11	-7.78	2.84	-16.71	1.15
	5-5;5	-10.91*	2.87	-19.91	-1.92
	5;6-5;11	-14.60**	2.62	-22.94	-6.27
4;6-4;11	5-5;5	-3.13	2.20	-9.98	3.72
	5;6-5;11	-6.82*	1.86	-12.66	-.99
5-5;5	5;6-5;11	-3.69	1.90	-9.65	2.27

** $p < 0.001$
* $p < 0.05$

The significant interaction effect indicates that age had a different influence on Lexical and Grammatical Morpheme scores. Figure 5.8 illustrates the interaction between age and morpheme

type. As was reported above, Lexical Morpheme scores were higher than Grammatical Morpheme scores. The gap between the two morphemes is most evident in the three youngest age groups and was reduced as age increased. To follow up the interaction, two one-way ANOVAs for Lexical and Grammatical Morpheme scores were employed, each with seven levels for the seven 6-month age bands. Results revealed a similar significance pattern as reported earlier for both Morpheme scores. In addition, a paired t-test was conducted for each age level to compare Lexical and Grammatical Morpheme scores. A Bonferroni correction of $\alpha = .007$ was applied. Results revealed that Lexical Morpheme score was significantly higher than Grammatical Morpheme score in all the age groups except the eldest.

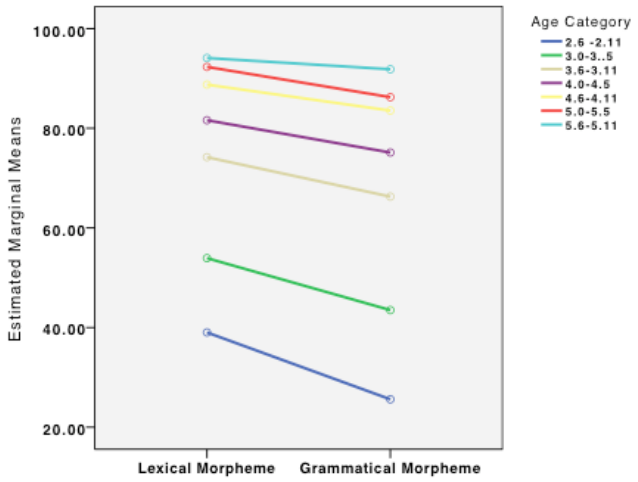


Figure 5.8. The interaction between age and morpheme type.

5.3.4.2 A comparison of Total Sentence Accuracy scores for age groups

Descriptive statistics are presented in Table 5.21 and show the mean, median, maximum, and minimum scores for Total Sentence Accuracy scores according to age group. Figure 5.9 illustrates the ranges of scores and distribution of the median scores for Total Sentence Accuracy through parallel boxplots for each age group.

Table 5.21 *Descriptive Statistics for the Sentence Repetition Test Total Sentence Accuracy Scores for Typically Developing Participants According to Age*

Score	Age Group	*n*	Mean	Median	SD	Min	Max
	2.6 – 2.11	20	1.25	0	2.43	0	10
Total	3.0 – 3.5	20	4.90	1	7.77	0	32
Sentence	3.6 – 3.11	20	13.75	13	9.87	1	30
Accuracy	4.0 – 4.5	20	16.40	16.50	8.38	3	29
	4.6 – 4.11	20	23.65	23	7.77	9	38
	5.0 – 5.5	20	26.25	26.50	8.71	12	40
	5.6 – 5.11	20	31.40	33	6.39	14	39

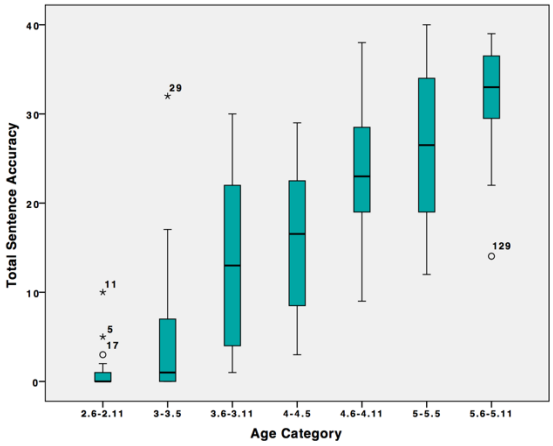


Figure 5.9. Boxplot showing Typically Developing group's Total Sentence Accuracy scores according to age of participants.

The median Total Sentence Accuracy scores increased as a function of age. The youngest age group exhibited a floor effect with a median score of 0. Of the 20 participants, nearly half obtained a score of 0 and only four participants scored higher than 1. Due to the narrow score range in the youngest age group, three participants emerged as outliers: two with extreme scores and all with scores above average for their age range. With the exception of the youngest age group, participants showed a wide range of scores for all age categories. Participant 29 appeared as an extreme outlier in the age group 3;0 to 3;5 for Total Sentence Accuracy with a score substantially higher than his/her age group and equivalent to the median score of children in the oldest age group, more than 2 years older than this child. Interestingly, the same child did not emerge as an outlier for Lexical and Grammatical morpheme percentage scores as can be seen in Figure 5.4 (a-b). In contrast the low Total Sentence Accuracy of Participant 129 in the oldest age group appeared as an outlier, was consistent with the low Lexical Morpheme percentage score (see Figure 5.4 (a)). Lack of symmetry was most apparent in the two youngest age groups.

Performance of participants on the Total Sentence Accuracy score was compared across the seven age groups. The two assumptions of parametric analysis were not met. As reported earlier Total Sentence Accuracy scores were not normally distributed, this was further confirmed with significant Shapiro-Wilk test results in the two youngest and oldest age groups (details can be found in Appendix H). The assumption of homogeneity of variance was also violated, Levene's test was significant for $F(6,133) = 5.94$, $p < .001$. Supporting non-parametric analysis can be found in Appendix H.

Results revealed a significant age effect $F(6,133) = 42.15$, $p < .001$, $\eta_p^2 = .66$, with scores increasing as age increased. Once again the effect size was large and of a similar order to the effect size for age effects on Morpheme percentage scores reported above.

The main effect of age was followed-up with the Games-Howell *post-hoc* test. Results are summarized in Table 5.22 and show a similar pattern of significance as the Morpheme percentage scores reported above.

Table 5.22 *Comparison of Mean Difference in Total Sentence Accuracy Score between Age Groups*

Age Category Comparison		**Mean Difference**	**Standard Error**	**95% CI**	
				Lower Bound	**Upper Bound**
2;6-2;11	3-3;5	-3.56	1.82	-9.53	2.23
	3;6-3;11	-12.50**	2.72	-19.88	-5.12
	4-4;5	-15.15**	1.95	-21.46	-8.84
	4;6-4;11	-22.40**	1.82	-28.27	-16.53
	5-5;5	-25.00**	2.02	-31.55	-18.45
	5;6-5;11	-30.15**	1.53	-35.05	-25.25
3-3;5	3;6-3;11	-8.85*	2.81	-17.61	-0.09
	4-4;5	-11.50*	2.56	-19.45	-3.55
	4;6-4;11	-18.75**	2.46	-26.40	-11.10
	5-5;5	-21.35**	2.61	-29.48	-13.22
	5;6-5;11	-26.50**	2.25	-33.52	-19.48
3;6-3;11	4-4;5	-2.65	2.89	-11.67	6.37
	4;6-4;11	-9.90*	2.81	-18.66	-1.14
	5-5;5	-12.50*	2.94	-21.67	-3.33
	5;6-5;11	-17.65**	2.63	-25.90	-9.40
4-4;5	4;6-4;11	-7.25	2.55	-15.20	0.70
	5-5;5	-9.85*	2.70	-18.26	-1.44
	5;6-5;11	-15.00**	2.36	-22.36	-7.64
4;6-4;11	5-5;5	-2.60	2.61	-10.71	5.53
	5;6-5;11	-7.75*	2.25	-14.76	-0.74
5-5;5	5;6-5;11	-5.15	2.42	-12.70	2.40

** $p < 0.001$
* $p < 0.05$

5.3.5 Anomalous Sentence Repetition test

Stimuli for the ASR test were created from eight of the twelve Typical sentences in the SR test. The ASR test consisted of two types of sentences: Semantically Anomalous and Syntactically Anomalous sentences. Each sentence type provided a Lexical Morpheme score equal to the number of correctly repeated Lexical Morphemes with a maximum score of 32, and a Grammatical Morpheme score equal to the number of correctly repeated Grammatical Morphemes with a maximum score of 69. The Lexical and Grammatical Morpheme scores of the two Anomalous sentence types were compared to each other and to the Lexical and Grammatical Morpheme scores of the Typical sentences they were created from.

The test was administered to participants who were 4 years of age or older, with 20 participants in each 6-month age band, for a total of 80 participants. As was the case in the analyses of the SR test, Lexical and Grammatical Morpheme scores were converted to percentage scores to allow for a comparison between the two types of morpheme scores. The results of the ASR test are presented in two main sections: the first section presents the descriptive statistics for Lexical and Grammatical morpheme percentage scores according to sentence type and age category of participants. The second section presents the related inferential statistics.

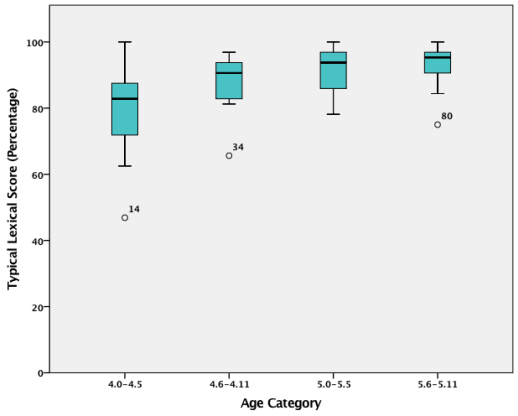
5.3.5.1 A comparison of morpheme and sentence type (descriptive statistics)

Descriptive statistics for Lexical Morpheme scores according to the age of participants and sentence type are presented in Table 5.23 and show the mean, median, maximum, and minimum scores for Typical, Semantically Anomalous, and Syntactically Anomalous sentences according to age group. Figure 5.9 (a-c) illustrates the distribution of the median scores and ranges of scores for Lexical Morpheme for each sentence type through boxplots for each age group. Figure 5.10 provides error bars together with their 95% confidence intervals for the mean Lexical Morpheme score for each of the four older age groups and illustrates the increase in Lexical Morpheme score with age and the relationship between the three sentence types.

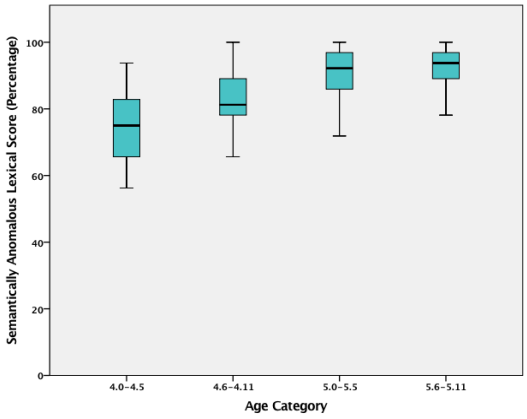
Table 5.23 *Descriptive Statistics for Lexical Morpheme Scores of Typically Developing Participants According to Age and Sentence Type (in Percentages)*

Sentence Type	Age Group	*n*	Mean	Median	SD	Min	Max
Typical	4.0 – 4.5	20	80.00	82.81	12.84	46.88	100
	4.6 – 4.11	20	88.44	90.63	7.86	65.63	96.88
	5.0 – 5.5	20	91.72	93.75	7.54	78.13	100
	5.6 – 5.11	20	93.91	95.31	6.04	75	100
Semantically Anomalous	4.0 – 4.5	20	73.91	75	10.7	56.25	93.75
	4.6 – 4.11	20	83.28	81.25	9.14	65.63	100
	5.0 – 5.5	20	89.84	92.19	9.17	71.88	100
	5.6 – 5.11	20	92.81	93.75	5.92	78.13	100
Syntactically Anomalous	4.0 – 4.5	20	64.53	65.63	9.25	46.88	81.25
	4.6 – 4.11	20	73.91	73.44	10.79	53.13	93.75
	5.0 – 5.5	20	83.13	85.94	11.35	62.5	96.88
	5.6 – 5.11	20	84.38	84.38	9.07	71.88	96.88

(a) Typical Sentences



(b) Semantically Anomalous



(c) Syntactically Anomalous

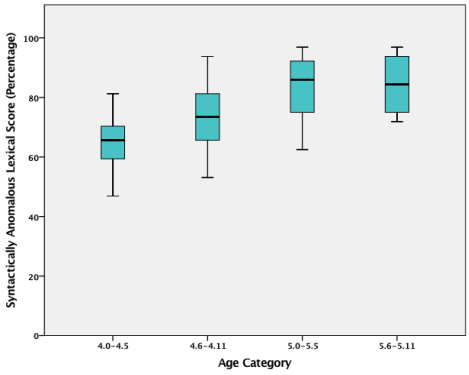


Figure 5.9. Boxplots showing Typically Developing group's Lexical Morpheme scores according to age and sentence type.

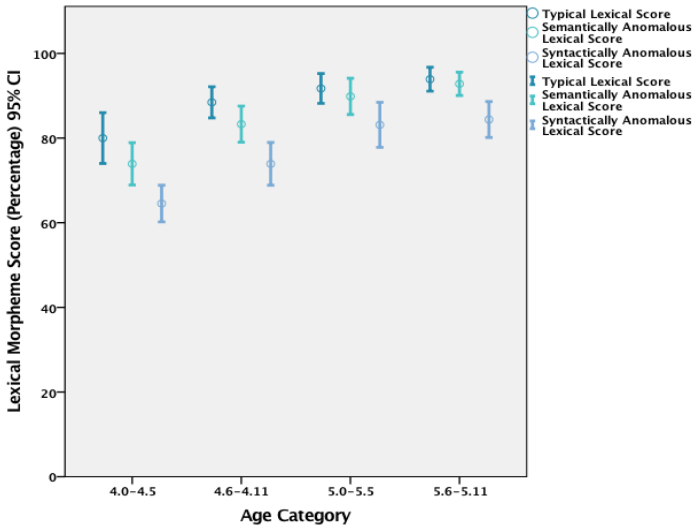


Figure 5.10. Error bars showing Typically Developing group's Lexical Morpheme scores according to age and sentence type.

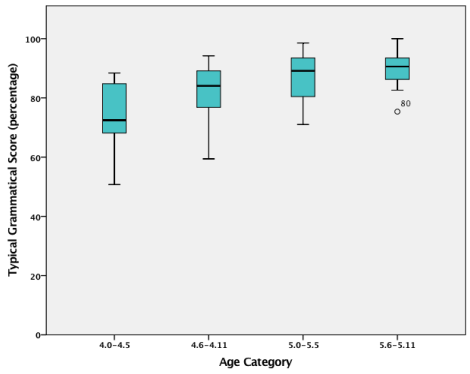
For Typical sentences, Lexical Morpheme scores showed a relatively narrow range of scores and small increase in median scores in the three oldest groups, with participants performing close to ceiling. For the Semantically and Syntactically Anomalous Sentence types, there was a more gradual increase in median scores for the three youngest age groups with this increase levelling off in the oldest age group. Median scores were close to ceiling in the oldest age group for Semantically Anomalous sentences but not Syntactically Anomalous sentences. Outliers were observed for Typical sentences but none were extreme. Figure 5.10 shows that there was a large overlap between mean Lexical Morpheme scores for Typical and Semantically Anomalous sentences with a slight advantage for Typical sentences. The advantage was most noticeable in the youngest age group; the gap was reduced as age increased. For Syntactically Anomalous sentences, mean Lexical Morpheme scores were lower than the score for Typical and Semantically Anomalous sentence types; there was almost no overlap between the Lexical Morpheme scores for Syntactically Anomalous sentences and the two other sentence types in any age group.

Descriptive statistics for Grammatical Morpheme scores according to the age of participants and sentence type are presented in Table 5.24 and show the mean, median, maximum, and minimum scores for Typical, Semantically Anomalous, and Syntactically Anomalous sentences according to age group. Figure 5.11 (a-c) illustrates the distribution of the median scores and ranges of scores for Grammatical Morphemes for each sentence type through boxplots for each of the four older age groups. Figure 5.12 provides error bars together with their 95% confidence intervals for the mean Grammatical Morpheme scores for each age group and illustrates the relationship between the three sentence types.

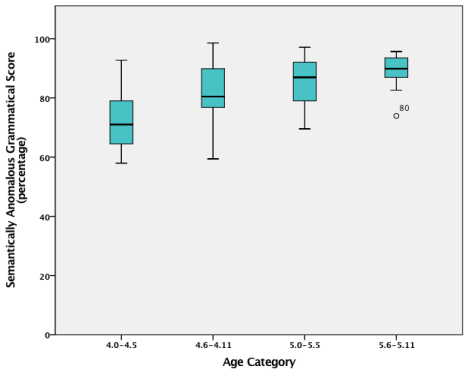
Table 5.24 *Descriptive Statistics for Grammatical Morpheme Scores of Typically Developing Participants According to Age and Sentence Type (in Percentages)*

Sentence Type	Age Group	*n*	Mean	Median	SD	Min	Max
Typical	4.0 – 4.5	20	74.64	72.46	11.04	50.72	88.41
	4.6 – 4.11	20	82.83	84.06	8.80	59.42	94.2
	5.0 – 5.5	20	87.75	89.13	7.65	71.01	98.55
	5.6 – 5.11	20	90.07	90.58	6.14	75.36	100
Semantically Anomalous	4.0 – 4.5	20	72.83	71.01	10.42	57.97	92.75
	4.6 – 4.11	20	82.25	80.43	9.68	59.42	98.55
	5.0 – 5.5	20	85.51	86.96	8.02	69.57	97.1
	5.6 – 5.11	20	88.91	89.86	5.33	73.91	95.65
Syntactically Anomalous	4.0 – 4.5	20	55.14	55.8	8.70	37.68	72.46
	4.6 – 4.11	20	64.06	63.77	9.60	43.48	85.51
	5.0 – 5.5	20	67.83	68.84	10.37	50.72	82.61
	5.6 – 5.11	20	71.88	71.01	8.68	57.97	85.51

a) Typical Sentences



b) Semantically Anomalous



(c) Syntactically Anomalous

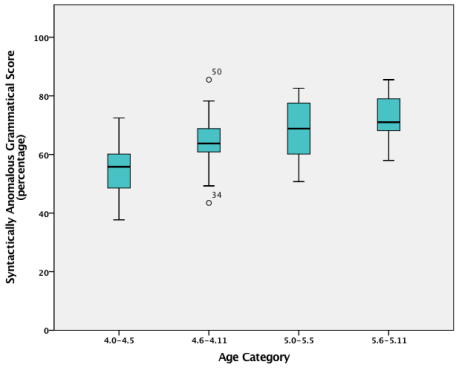


Figure 5.11. Boxplots showing Typically Developing group's Grammatical Morpheme scores according to age and sentence type.

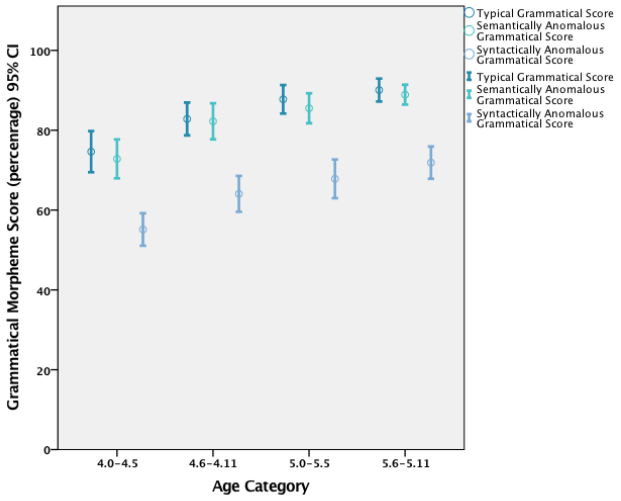


Figure 5.12. Error bars showing Typically Developing group's Grammatical Morpheme scores according to age and sentence type.

As can be seen from Table 5.24 and Figure 5.11 (a-c), for all sentence types Grammatical Morpheme scores showed a gradual increase in the first three age groups, then levelled off in the oldest age group, unlike the Lexical Morpheme scores which showed a narrower range of scores. Scores were close to ceiling in the oldest age group for Typical and Semantically Anomalous sentences but not Syntactically Anomalous sentences. A few outliers were present but none of them were extreme outliers. Figure 5.12 shows that mean scores for Grammatical Morphemes in Typical and Semantically Anomalous sentences overlapped with a slight advantage for Typical sentences. Scores for Syntactically Anomalous sentences were markedly lower than the other two sentence types with no overlap between them.

5.3.5.2 A comparison of morpheme and sentence type (inferential statistics)

To investigate the effects of age, morpheme, and sentence type, a three-factor mixed design ANOVA was employed with age as a between-subjects variable which included the older four 6-month age bands; morpheme type was a within-subject variable with two levels: Lexical and Grammatical. Also, sentence type was a within-subject variable with three levels: Typical, Semantically Anomalous, and Syntactically Anomalous. The dependent variable was repetition percentage score. The repetition scores were converted to percentages to allow for a comparison between Lexical and Grammatical morphemes since the maximum raw scores of the two morpheme types differed.

The three assumptions of parametric analysis were not met. The assumption of normal distribution was violated for Lexical Morpheme scores for a number of age groups across sentence types and Grammatical Morpheme score of the oldest age group for Semantically Anomalous sentences, as indicated by the significant results of Shapiro-Wilk test of normality (details can be

found in Appendix H). The assumption of homogeneity of variance was violated; Levene's test was significant for Lexical Morpheme scores of Typical sentences $F(3,76) = 4.03$, $p = .01$ and for Grammatical Morpheme scores of Typical sentences $F(3,76) = 3.22$, $p = .03$. The assumption of sphericity was also violated for the main effect of sentence type as indicated by the significant result of Mauchly's test of sphericity $\chi^2(2) = 9.42$, $p = .009$. Therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity $\varepsilon = .89$. All parametric analyses were supported with non-parametric analyses which are reported in Appendix H. Non-parametric analyses are only reported here if results were inconsistent with parametric analyses.

The main effect of age was significant $F(3,76) = 19.03$, $p < .001$, $\eta_p^2 = .43$ with scores increasing as age increased. The main effect of morpheme type was significant $F(1,76) = 185.27$, $p < .001$, $\eta_p^2 = .71$. Overall Grammatical Morpheme mean scores were significantly lower than Lexical Morpheme mean scores. The main effect of sentence type was significant $F(1.79, 135.95) = 228.74$, $p < .001$, $\eta_p^2 = .75$. The interaction of morpheme and sentence type was significant $F(2,152) = 65.85$, $p < .001$, $\eta_p^2 = .46$. The interaction of morpheme sentence type and age was significant $F(6,152) = 2.54$, $p = .02$, $\eta_p^2 = .09$. All effect sizes were large except the effects size of the three-way interaction between morpheme sentence type and age which was moderate. The remaining interactions were non-significant (morpheme and age, $F(3,76) = 1.66$, $p = .182$, $\eta_p^2 = .06$; sentence and age, $F(5.38,135.95) = 0.7$, $p = .64$, $\eta_p^2 = .03$).

The main effect of age was followed-up with the Games-Howell *post-hoc* test. Results are summarized in Table 5.25 and show that the mean score of the age group 4;0-4;5 differed significantly from the three older age groups. The mean score of the age group 4;6-4;11 differed from the youngest and oldest age group.

Table 5.25 *Comparison of Mean Difference in Scores between Age Groups*

Age Category	**Comparison**	**Mean Difference**	**Standard Error**	**95% CI**	
				Lower Bound	**Upper Bound**
4-4;5	4;6-4;11	-8.95*	2.66	-16.09	-1.82
	5-5;5	-14.12**	2.62	-21.16	-7.08
	5;6-5;11	-16.82**	2.33	-23.14	-10.50
4;6-4;11	5-5;5	-5.17	2.47	-11.81	1.47
	5;6-5;11	-7.87*	2.16	-13.71	-2.02
5-5;5	5;6-5;11	-2.70	2.12	-8.41	3.01

** $p < 0.001$
* $p < 0.01$

The main effect of sentence type was followed-up with the Bonferroni *post-hoc* test. Results are summarized in Table 5.26 and show that the mean score for the three sentence types

differed significantly from each other with repetition scores for Typical sentences being the highest followed by Semantically Anomalous sentences and Syntactically Anomalous sentences, respectively.

Table 5.26 *Comparison of Mean Difference in Scores between Sentence Types*

Sentence Comparison	**Mean Difference**	**Standard Error**	**95% CI**	
			Lower Bound	**Upper Bound**
Typical vs. Semantic	2.50 *	0.66	0.89	4.11
Typical vs. Syntactic	15.56**	0.89	13.38	17.75
Semantic vs. Syntactic	13.06**	0.79	11.16	14.97

The significant interaction between morpheme and sentence types indicates that Lexical and Grammatical Morpheme scores were influenced differently by sentence type. Figure 5.13 illustrates the interaction between morpheme and sentence type. To follow-up the interaction two one-way ANOVAs for Lexical and Grammatical Morphemes scores were employed, each with three levels for each sentence type. Results revealed a different profile for Lexical and Grammatical Morpheme scores. Lexical Morpheme score was significantly reduced in Semantically Anomalous sentences compared to Typical sentences and again in Syntactically Anomalous sentences compared Semantically Anomalous sentences. However, Grammatical Morpheme score did not significantly differ between Typical and Semantically Anomalous sentences but was reduced in Syntactically Anomalous sentences. In addition, a paired t-test was conducted for each sentence type to compare Lexical and Grammatical Morpheme scores. A Bonferroni correction of $\alpha = .017$ was applied. Results revealed that Lexical Morpheme score was significantly higher than Grammatical Morpheme score in all sentence types.

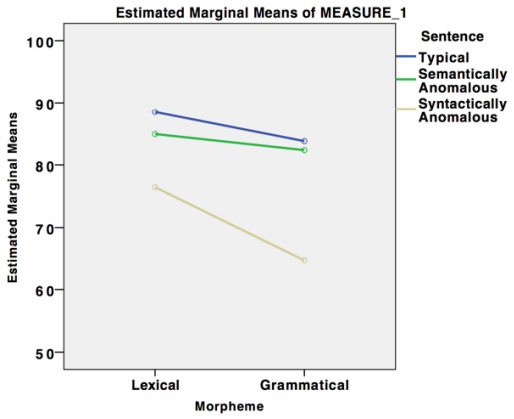


Figure 5.13. The interaction between Morpheme and Sentence type.

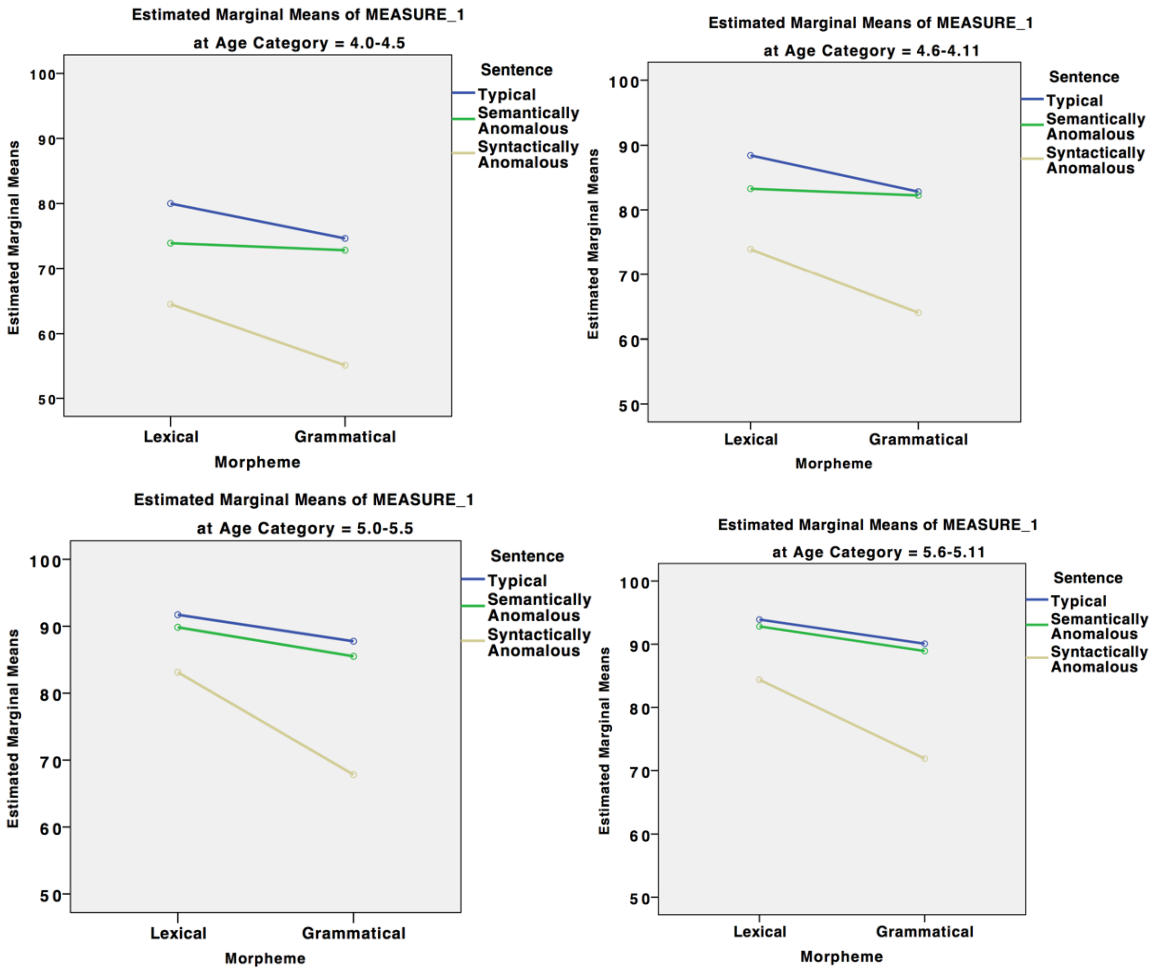


Figure 5.14. The interaction between Morpheme and Sentence type and age.

5.4 Participants with Language Concerns

The Language Concerns group consisted of 16 participants in the same age range as children in the Typically Developing sample, 2;6 to 5;11. In order to compare performance on three tests according to language status, 16 of the 140 participants in the Typically Developing sample were selected to match the 16 participants in the Language Concerns group age in months and Nonverbal IQ (Control group). Although both the Block Design and Object Assembly subtests of the WPPSI-III ^{UK} (Wechsler, 2003) were administered, only Block Design was included in matching (as previously detailed).

An independent *t*-test was conducted to compare the age in months and scaled Block Design scores of the Language Concerns and Control groups. Results indicated no significant difference between the Language Concerns (M = 55.88, SD = 13) and Control groups (M = 55.56, SD = 12.70) on age in months of participants, t (30) = -.07, r = .01. No significant difference was

found between the scaled Block Design scores for the Language Concerns (M = 8.56, SD = 3.20) and Control groups (M = 8.81, SD = 2.69), t (30) = .24, r = .04. Therefore, any difference between the group means of participants in the Language Concerns and Control groups on the three assessments cannot be attributed to age or IQ scores of participants.

In analysing the performance of the participants in the Language Concerns group, descriptive statistics are presented first, followed by inferential statistics.

5.4.1 Verbal Short Term Memory test

The VSTM test provided four span scores: Word List Recall, Digit Recall, Nonword List Recall, and a Total Span. The results of the VSTM test are presented in two main sections: the first presents a comparison of subtest span scores and the second a comparison of Total Span scores according to the language status of participants.

5.4.1.1 A comparison of Verbal Short Term Memory subtests according to language status

Table 5.26 shows the mean, median, maximum, and minimum Span scores according to language status of participants. Boxplots in Figure 5.15 (a) illustrate the distribution of the three subtest Span scores for each language group. Figure 5.15 (b) provides error bars showing the mean subtest Span scores for both language groups along with their 95% confidence intervals and illustrates the relationship between the three subtests in both language groups.

Table 5.26 *Descriptive Statistics for the Verbal Short Term Memory Subtests According to Language Status, Controls, and Language Concerns*

Span	Group	*n*	Mean	Median	SD	Min	Max
Digit	C	16	3.50	3.75	.97	1	5
	LC	16	2.50	3	.75	1	3.5
Word	C	16	3.06	3	.83	1	4.5
	LC	16	2.19	2	.68	1	3
Nonword	C	16	1.63	2	.47	1	2
	LC	16	1.16	1	.35	1	2

Note: C = Control; LC = Language Concerns.

a) Boxplots b) Error Bars

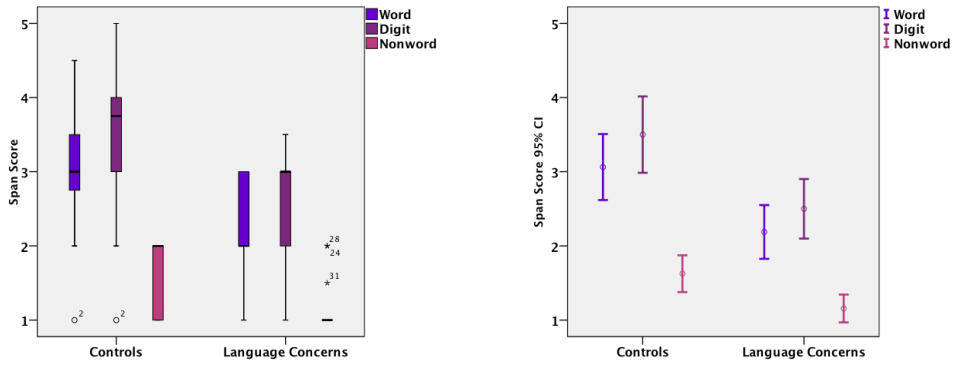


Figure 5.15. Verbal Short Term Memory subtest span scores according to language status of participants.

The median span scores for all three subtests were lower in the Language Concerns group. The boxplots show a lack of normal distribution of span scores for all three subtests in the Language Concerns group. This is most evident in the Nonword List Recall subtest, where there was a narrow range of scores and little variability: of the 16 participants, only three had a span score higher than 1 and all appeared as extreme outliers. For the Control group, span scores on the Word List Recall subtest appear to be normally distributed with only one participant who presented as an outlier but not extreme. Span scores for the Digit Recall subtest appear to be slightly positively skewed, with the same participant appearing as an outlier but not extreme. Figure 5.15 (b) shows that the relationship between the three subtest Span scores was similar in the two groups. The Span Score for Digit Recall was the highest followed by Word List Recall Span score and Nonword List Recall Span score respectively. None of the three subtest Span scores showed a greater disadvantage in the Language Concerns group.

To investigate the effects of language status on subtest type, a two-factor mixed design ANOVA was employed, with language status as the between-subject factor with two levels: Language Concerns and Control groups, and Subtest type as the within-subject variable with three levels (Word List, Digit, and Nonword List Recall). Span score was the dependent variable.

Three assumptions of parametric analysis were violated. The assumption of normality was violated as indicated by the lack of normal distribution illustrated above in Figure 5.15 (a) and the significant Shapiro-Wilk test results (details can be found in Appendix H). The assumption of homogeneity of variance was violated as indicated by the significant Levene's test result for Nonword List Recall $F(1,30) = 5.5$, $p = .026$. The assumption of sphericity was also violated for the main effect of subtest type as indicated by the significant result of Mauchly's test of sphericity $\chi^2(2) = 8.57$, $p = .014$. Therefore, the degrees of freedom were corrected using Greenhouse-Geisser

estimates of sphericity ε = .80. All parametric analyses were supported with non-parametric analyses which are reported in Appendix H.

Results show a significant effect of language status $F(1,30) = 13.21$, $p = .001$, $\eta_p^2 = .31$, with the overall mean span scores of participants in the Language Concerns group being significantly lower than the Control group. Results show a significant effect of subtest type $F(1.59,47.78) = 116.17$, $p < .001$, $\eta_p^2 = .80$. Both effect sizes were large (Cohen, 1992). No significant interaction was found between language status and subtest type $F(1.59,47.78) = 3.16$, $p = .062$, $\eta_p^2 = .1$, indicating that both language status groups were influenced similarly by subtest type, as was evident in Figure 5.15 (b).

To follow up the main effect of subtest type, a Bonferroni adjusted *post-hoc* test was applied. Table 5.27 shows the mean difference between Span scores, the significance of the difference, the standard error, and 95% confidence intervals. A significant difference was found between all three subtests with Digit Recall Span scores being the highest followed by Word List Recall and Nonword List Recall.

Table 5.27 *Comparison of Mean Difference in Span Scores between Verbal Short Term Memory Subtests*

Subtest Comparison	Mean Difference	Standard Error	95% CI Lower Bound	95% CI Upper Bound
Digit vs. Word	0.38*	0.08	0.17	0.58
Digit vs. Nonword	1.61*	0.13	1.28	1.94
Word vs. Nonword	1.23*	0.11	0.95	1.52

* p < 0.001 (adjusted for multiple comparisons)

5.4.1.2 A comparison of Total Span score according to language status

Table 5.28 shows the mean, median, maximum, and minimum Total Span scores according to language status. Figure 5.16 (a) illustrates the distribution of Total Span scores for each language group. Figure 5.16 (b) provides error bars showing the mean Total Span scores for both language groups along with their 95% confidence intervals.

Table 5.28 *Descriptive Statistics for Total Span Scores According to Language Status, Controls, and Language Concerns*

Score	**Group**	***n***	**Mean**	**Median**	**SD**	**Min**	**Max**
Total Span Score	C	16	8.19	8.75	2.06	3	11
	LC	16	5.84	6	1.55	3	8

Note: C = Control; LC = Language Concerns.

a) Boxplots b) Error Bars

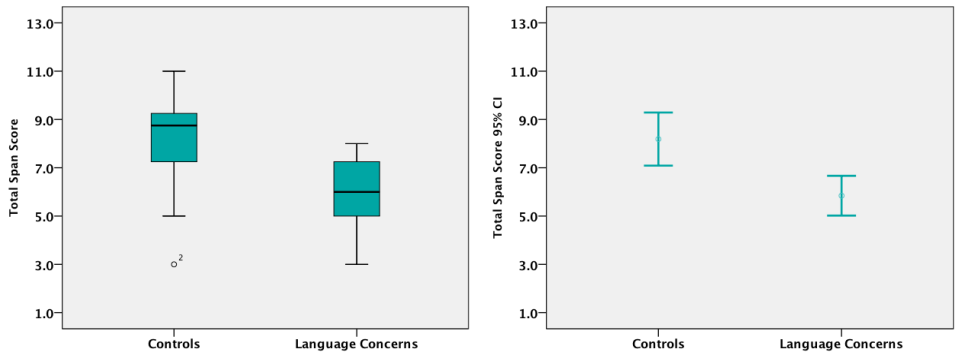


Figure 5.16. Total Span score according to language status of participants.

The median Total Span scores were lower in the Language Concerns group. Scores appeared to be normally distributed in the Language Concerns group and slightly positively skewed in the Control group. One non-extreme outlier appeared in the Control group, interestingly the same participant that appeared in the Word List Recall and Digit Recall subtests (see Figure 5.16 (a)). Figure 5.16 (b) shows that the mean span scores were lower in the Language Concerns group with no overlap between them.

To investigate the effect of language status on Total Span scores, an independent *t*-test was carried out. The assumptions of normality and homogeneity of variance were met. The mean Total Span scores for participants in the Language Concerns group were significantly lower than controls, $t\ (30) = 3.64$, $p < .001$, $r = .55$, with a large effect size (Cohen, 1992).

5.4.2 Sentence Repetition test

The SR test provided three scores: a Lexical Morpheme score, a Grammatical Morpheme score, and a Total Sentence Accuracy score. The results of the SR test are presented in two main sections: the first section presents a comparison of Lexical and Grammatical Morpheme scores and the second section presents a comparison of Total Sentence Accuracy scores according to the language status of participants.

5.4.2.1 A comparison of Lexical and Grammatical Morpheme scores according to language status

Lexical and Grammatical Morpheme raw scores were converted into percentages to allow for a comparison of the two scores. Table 5.29 shows the mean, median, maximum, and minimum scores for both morpheme types according to language group. Boxplots in Figure 5.17 (a) illustrate the distribution of the median scores for Lexical and Grammatical Morpheme scores. Figure 5.17

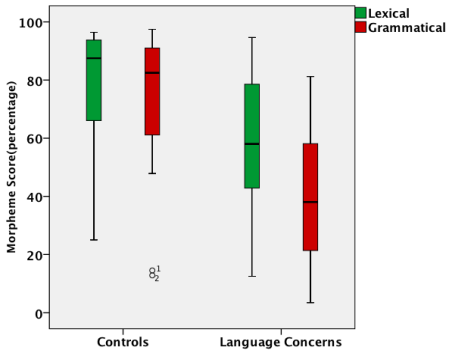
(b) provides error bars showing the mean Lexical and Grammatical morpheme scores for each language group along with their 95% confidence intervals.

Table 5.29 *Descriptive Statistics for the Sentence Repetition Test Morpheme Scores (in Percentages) According to Language Status, Controls, and Language Concerns*

Score	Group	*n*	Mean	Median	SD	Min	Max
Lexical Morpheme	C	16	77.68	87.50	23.16	25.00	96.43
	LC	16	56.70	58.04	24.10	12.50	94.64
Grammatical Morpheme	C	16	71.63	82.48	26.84	12.82	97.44
	LC	16	39.90	38.03	23.83	3.42	81.20

Note: C = Control; LC = Language Concerns.

a) Boxplots



b) Error bars

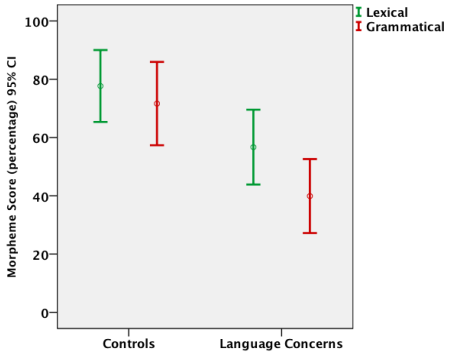


Figure 5.17. Morpheme scores according to language status of participants.

The median Lexical and Grammatical Morpheme scores were lower in the Language Concerns group in comparison to Controls. Although the medians of both Morpheme scores were lower in the Language Concerns group, the median Grammatical Morpheme score showed a greater disadvantage than the median Lexical Morpheme score. Lexical and Grammatical Morpheme scores in the Language Concerns group were normally distributed with no outliers. In the Control group, however, both Lexical and Grammatical Morpheme scores deviated from normal distribution and were positively skewed with a long left tail for Lexical Morpheme scores. Participant 1 and 2's Grammatical Morpheme scores appeared as outliers but neither of them were extreme. Figure 5.17 (b) shows that the mean Lexical and Grammatical Morpheme scores for participants in the Language Concerns group were less than the mean scores of Controls with Grammatical Morpheme scores, showing no overlap between the two groups. In both groups, Lexical Morpheme scores were higher than Grammatical Morpheme scores. However, the difference between Lexical and Grammatical Morpheme mean scores (16.8) was almost three times

greater in the Language Concerns group in comparison to Controls (6). However, a similar large discrepancy was observed in the youngest age group of the Typical sample (see Figure 5.5).

To investigate the effects of language status and morpheme type, a two-factor mixed design ANOVA was employed, with language status as the between-subject factor with two levels Language Concerns and Controls, and morpheme type was the within-subject variable with two levels: Lexical and Grammatical Morphemes. The dependent variable was percentage morphemes correct. One of the assumptions of parametric analysis was not met: as reported above, the Lexical and Grammatical Morpheme scores for the Control group were not normally distributed, and this was further confirmed with significant results of Shapiro-Wilk test of normality (details can be found in Appendix H). All supporting non-parametric analyses are reported in Appendix H.

The main effect of language status was significant $F(1,30) = 9.51$, $p = .004$, $\eta_p^2 = .24$ with participants with Language Concerns scoring lower than controls. The main effect of Morpheme Type was significant $F(1,30) = 61$, $p < .001$, $\eta_p^2 = .67$. The overall mean Grammatical Morpheme scores were significantly lower than the mean Lexical Morpheme scores. The interaction of language status and morpheme type was also significant $F(1,30) = 13.51$, $p = .001$, $\eta_p^2 = .31$. All these effect sizes were large (Cohen, 1992).

The significant interaction effect indicates that language status had a different influence on Lexical and Grammatical Morpheme scores. As observed above and seen from Figure 5.18, although Grammatical Morpheme scores were less than Lexical Morpheme scores in both language status groups, the Grammatical Morpheme was more vulnerable in participants with Language Concerns. To follow up this interaction a total of four *t*-tests were conducted with a Bonferroni correction of ($\alpha = .013$). Two independent *t*-tests were conducted to compare Lexical and Grammatical Morpheme scores in the two language status groups. Lexical Morpheme scores were significantly lower for participants with Language Concerns $t(30) = 2.51$, $p = .018$ $r = .42$ with a medium effect size (Cohen, 1992). Grammatical Morpheme scores were also significantly lower for participants with Language Concerns $t(30) = 3.54$, $p = .001$ $r = .54$, with a large effect size. Two dependent *t*-tests were conducted to compare Lexical and Grammatical Morpheme scores within each language status group. Grammatical Morpheme scores were significantly lower than Lexical Morpheme scores in the Language Concerns group $t(15) = 7.26$, $p < .001$ $r = .88$ and Controls $t(15) = 3.39$, $p = .004$ $r = .66$, both with large effect sizes.

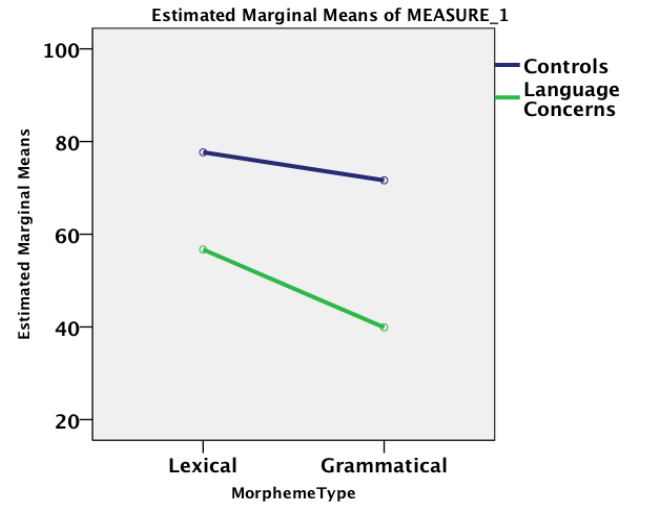


Figure 5.18. The interaction between morpheme type and language status of participants.

5.4.2.2 A comparison of Total Sentence Accuracy scores according to language status

Table 5.30 shows the mean, median, maximum, and minimum Total Sentence Accuracy scores according to language status. Boxplots in Figure 5.19 (a) illustrate the distribution of the Total Sentence Accuracy scores for each language group. Figure 5.19 (b) provides error bars showing the mean Total Sentence Accuracy scores for both language groups along with their 95% confidence intervals. Table 5.30 and Figure 5.19 (a) show that the median Total Sentence Accuracy score was lower in the Language Concerns group in comparison to Controls. Scores were normally distributed in the Control group. Scores in the Language Concerns group, however, lacked symmetry and were negatively skewed. There were no outliers in either group. Figure 5.19 (b) shows that the mean score for the Total Sentence Accuracy was higher in the Control group with no overlap between the two groups.

Table 5.30 *Descriptive Statistics for Total Sentence Accuracy Scores According to Language Status, Controls, and Language Concerns*

Score	Group	*n*	Mean	Median	SD	Min	Max
Total Sentence Accuracy	C	16	18.50	18.00	12.78	0	37
	LC	16	5.19	1.50	6.74	0	19

Note: C = Control; LC = Language Concerns.

a) Boxplots b) Error Bars

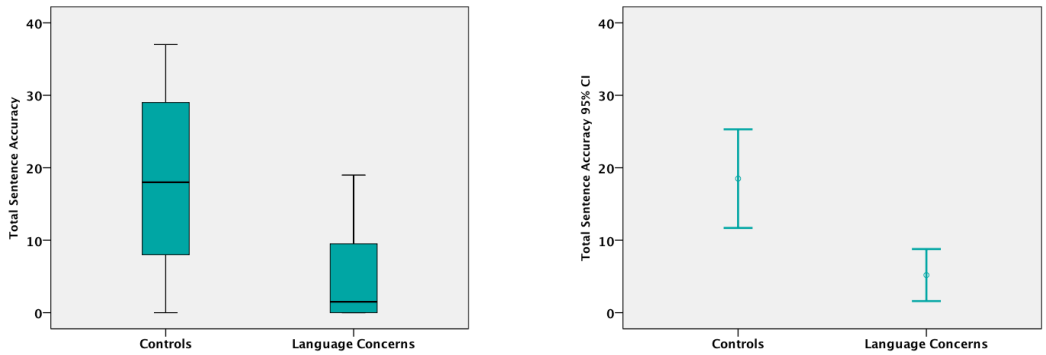


Figure 5.19. Sentence Accuracy scores according to language status of participants.

To investigate the effect of language status on Total Sentence Accuracy scores, an independent *t*-test was carried out. Two assumptions of parametric analysis were violated: as reported above, Total Sentence Accuracy scores for participants with Language Concerns were not normally distributed, and this was further confirmed with significant Shapiro-Wilk test results (details can be found in Appendix H). The assumption of homogeneity of variance was violated as indicated by the significant result of Levene's test $F(30) = 9.24$, $p = .005$; therefore, the degrees of freedom were adjusted. The Total Sentence Accuracy scores of the Language Concerns group were significantly lower than the Control group, $t(22.74) = 3.69$, $p < .001$ $r = .56$, with a large effect size (Cohen, 1992).

5.4.3 Anomalous Sentence Repetition test

The ASR test provided a Lexical and a Grammatical Morpheme score for Semantically Anomalous and Syntactically Anomalous sentences. The scores for the two Anomalous sentences were compared to the scores for the Typical sentences they were created from. Lexical and Grammatical Morpheme scores were converted to percentage scores to allow for comparison. The results are presented in two main sections: the first section presents the descriptive statistics for Lexical and Grammatical Morpheme scores according to sentence type and language status of participants. The second section presents the related inferential statistics.

5.4.3.1 A comparison of morpheme scores according to sentence type and language status

Table 5.31 shows the mean, median, maximum, and minimum Lexical and Grammatical Morpheme scores for Typical, Semantically Anomalous, and Syntactically Anomalous sentences according to language group. Boxplots in Figure 5.20 (a-b) illustrate the distribution of Lexical and Grammatical Morpheme scores for each sentence type. Figure 5.20 (c-d) provides error bars

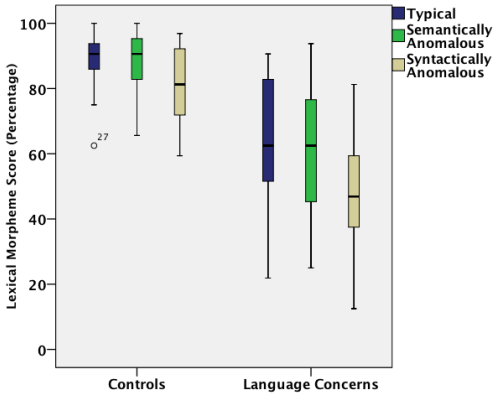
showing means together with their 95% confidence intervals for the mean Lexical and Grammatical Morpheme score for each language group and illustrates relations between the two Morpheme scores and language status and between the three sentence types.

Table 5.31 *Descriptive Statistics for the Anomalous Sentence Repetition Test Morpheme Scores According to Language Status and Sentence Type (in Percentages), Controls, and Language Concerns*

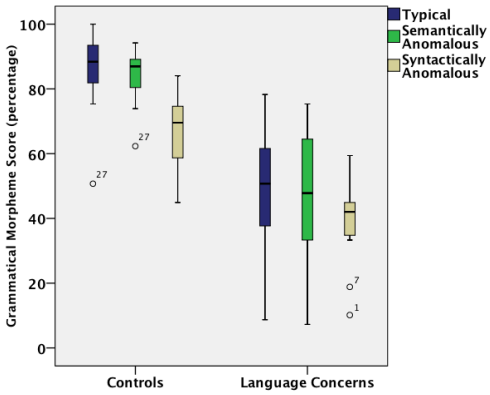
Sentence Type	**Morpheme Type**	**Group**	***N***	**Mean**	**Median**	**SD**	**Min**	**Max**
Typical	Lexical	C	11	88.07	90.63	10.81	62.50	100
		LC	11	65.34	62.50	21.40	21.88	90.63
	Grammatical	C	11	85.51	88.41	13.69	50.72	100
		LC	11	50.07	50.72	20.20	8.70	78.26
Semantically Anomalous	Lexical	C	11	88.07	90.63	10.53	65.63	100
		LC	11	59.09	62.50	21.44	25	93.75
	Grammatical	C	11	83.40	86.96	9.01	62.32	94.20
		LC	11	46.64	47.83	20.79	7.25	75.36
Syntactically Anomalous	Lexical	C	11	80.68	81.25	13.24	59.38	96.88
		LC	11	48.86	46.88	20.93	12.50	81.25
	Grammatical	C	11	66.93	69.57	12.24	44.93	84.06
		LC	11	38.21	42.03	13.65	10.14	59.42

Note: C = Control; LC = Language Concerns.

a) Boxplot: Lexical Morpheme Score



b) Boxplot: Grammatical Morpheme Score



c) Error bar: Lexical Morpheme Score d) Error bar: Grammatical Morpheme Score

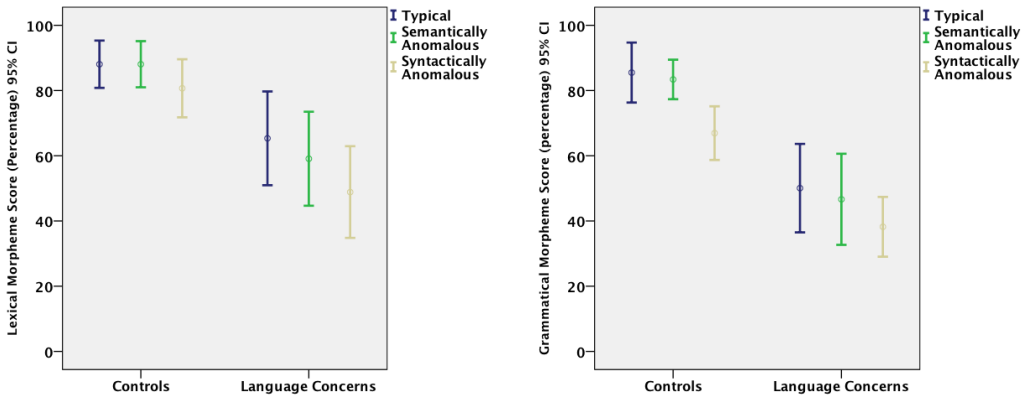


Figure 5.20. Morpheme scores according to language status and sentence type.

Table 5.31 and Figure 5.20 (a-b) show that the median Lexical and Grammatical Morpheme scores for the Language Concerns group were lower for all three sentence types. In the Language Concerns group, scores appeared to be normally distributed with a wide range of scores for both morpheme types. Grammatical Morpheme scores for Syntactically Anomalous sentences was the only exception with a narrower range of scores in comparison to the other two sentence types and where two non-extreme outliers appeared. The distribution of scores was narrower in the Controls group compared to the Language Concerns group. Participant number 27 appeared as a non-extreme outlier for Lexical Morpheme score in Typical sentences and Grammatical Morpheme score for Typical and Semantically Anomalous sentences. Figure 20 (c-d) shows that the mean Lexical Morpheme scores were higher than the mean Grammatical Morpheme scores in both language status groups and all three sentence types. However, this advantage was more marked in the Language Concerns group. In the Controls group, the mean Lexical and Grammatical Morpheme scores were most affected by Syntactically Anomalous sentences with little or no difference between mean scores for Typical and Semantically Anomalous sentences. In the Language Concerns group, the Lexical and Grammatical Morpheme scores were highest for Typical sentences followed by Semantically Anomalous sentences and Syntactically Anomalous sentences respectively. The mean Grammatical Morpheme score for Syntactically Anomalous sentences in the Language Concerns group was the most vulnerable.

5.4.3.2 A comparison of morpheme, sentence type and language status (inferential statistics)

To investigate the effects of language status, morpheme, and sentence type, a three-factor mixed design ANOVA was employed with language status as the between-subject variable.

Morpheme type was a within-subject variable with two levels: Lexical and Grammatical. Sentence type was also a within-subject variable with three levels: Typical, Semantically Anomalous, and Syntactically Anomalous. The dependent variable was the percentage morphemes correct.

The three assumptions of parametric analysis were not met. The assumption of normality was violated. As reported above, Lexical and Grammatical Morpheme scores for Typical Sentences in the Control group were not normally distributed. This was confirmed by the significant results of the Shapiro-Wilk test (details can be found in Appendix H). The assumption of homogeneity of variance was violated. Levene's test was significant for: Lexical Morpheme scores for Typical sentences $F(1,20) = 5.58$, $p = .029$, Lexical Morpheme scores for Semantically Anomalous sentences $F(1,20) = 5.41$, $p = .031$, and Grammatical Morpheme scores for Semantically Anomalous sentences $F(1,20) = 5.86$, $p = .025$. The assumption of sphericity was violated for the main effect of sentence type as indicated by the significant result of Mauchly's test of sphericity $\chi^2(2) = 7.96$, $p = .019$. Therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity $\varepsilon = .75$. All parametric analyses were supported with non-parametric analyses which are reported in Appendix H.

The main effect of language status was significant $F(1,20) = 24.14$, $p < .001$, $\eta_p^2 = .55$ with scores of participants in the Language Concerns group significantly lower than Controls. The main effect of morpheme type was significant $F(1,20) = 56.78$, $p < .001$, $\eta_p^2 = .74$. Overall Grammatical Morpheme mean scores were significantly lower than Lexical Morpheme mean scores. The main effect of sentence type was significant $F(1.49, 29.80) = 21.50$, $p < .001$, $\eta_p^2 = .52$.The interaction of morpheme and language status was significant $F(1,20) = 4.88$, $p = .039$, $\eta_p^2 = .20$. The three-way interaction of morpheme, sentence type, and language status was significant $F(2,40) = 8.05$, $p = .001$, $\eta_p^2 = .29$. All effect sizes were large (Cohen, 1992). The remaining interactions were non-significant (sentence and language status, $F(1.49,29.8) = 0.40$, $p = .68$, $\eta_p^2 = .02$; morpheme and sentence, $F(2,40) = 1.98$, $p = .155$, $\eta_p^2 = .09$).

The main effect of sentence type was followed-up with a Bonferroni adjusted *post-hoc* test. Results are summarized in Table 5.32 and show that the mean scores for Typical sentences and Semantically Anomalous were significantly higher than Syntactically Anomalous sentences but they did not differ significantly from each other.

Table 5.32 *Comparison of Mean Difference in Scores Between Sentence Types*

Sentence Comparison	**Mean Difference**	**Standard Error**	**95% CI**	
			Lower Bound	**Upper Bound**
Typical vs. Semantic	2.95	1.73	-1.58	7.48
Typical vs. Syntactic	13.58*	2.74	6.43	20.72
Semantic vs. Syntactic	10.63*	1.93	5.58	15.68

* $p < 0.001$

The significant interaction between morpheme type and language status (illustrated in Figure 5.21) indicates that Lexical and Grammatical Morpheme scores were influenced differently by the language status of participants. In both language status groups, the mean Lexical Morpheme score was higher than the mean Grammatical Morpheme score. However, the difference between the groups' mean Grammatical Morpheme scores was greater than the difference in mean Lexical Morpheme scores. This indicates that Grammatical Morphemes were more vulnerable in participants with Language Concerns, in line with the interaction found between morpheme type and language status in the results of the SR test. To follow up this interaction, a total of four *t*-tests were conducted with a Bonferroni correction of $\alpha = .013$. Two independent *t*-tests were conducted to compare the Lexical and Grammatical Morpheme scores between the two language status groups. The mean Lexical Morpheme score was significantly higher in the Control group $t(15.07) = 4.15$, $p = .001$, $r = .73$. The mean Grammatical Morpheme score was also significantly higher in the Control group $t(20) = 5.55$, $p < .001$, $r = .79$. Two paired *t*-tests were conducted to compare the mean Lexical and Grammatical Morpheme scores within each language status group. The mean Lexical Morpheme score was significantly higher than the mean Grammatical Morpheme score in the Control group $t(10) = 5.89$, $p < .001$, $r = .88$, and the Language Concerns group $t(10) = 5.46$, $p < .001$, $r = .86$. The effect sizes for all the *t*-tests were large (Cohen, 1992).

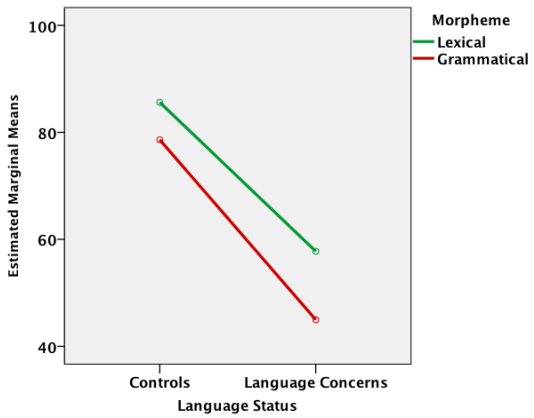


Figure 5.21. The interaction between morpheme and sentence type.

The significant three-way interaction indicates that the morpheme x sentence interaction varied according to the language status of participants. The interaction between morpheme and sentence type is illustrated in Figure 5.22 (a) for participants in the Control group and 5.22 (b) for participants in the Language Concerns group. Figure 5.22 (a) shows that for the Control group, mean Lexical Morpheme scores were the same for Typical and Semantically Anomalous sentences; both were greater than the mean Lexical Morpheme score for Syntactically Anomalous sentences. The mean Grammatical Morpheme score for Typical sentences showed a slight advantage over the mean Grammatical Morpheme score for Semantically Anomalous sentences, both were greater than the mean Grammatical Morpheme score for Syntactically Anomalous sentences. The advantage of the mean Lexical Morpheme score over the mean Grammatical Morpheme score is most evident for Syntactically Anomalous sentences. Figure 5.22 (b) shows that for participants in the Language Concerns group, both the mean Lexical and Grammatical Morpheme scores were higher for Typical sentences followed by Semantically Anomalous sentences and Syntactically Anomalous sentences. This pattern of performance for Lexical Morpheme scores across the three sentence types was similar to participants in the youngest age group of the Typical sample. The advantage of mean Lexical Morpheme scores over the mean Grammatical Morpheme scores was evident in all three sentence types. In both the Control and Language Concerns groups, the mean Grammatical Morpheme score for Syntactically Anomalous sentences was the most vulnerable; however, it was more vulnerable in the Language Concerns group.

a) Controls

b) Language Concerns

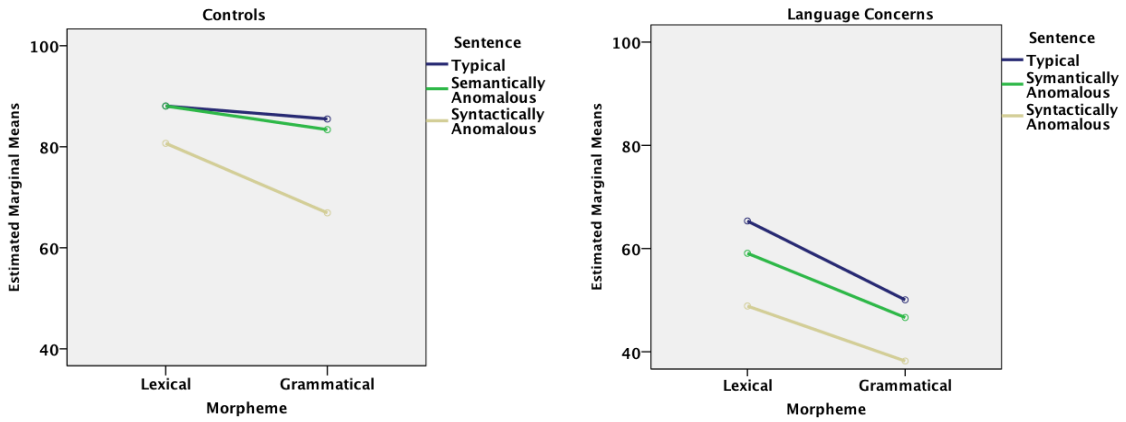


Figure 5.22. The interaction between morpheme type and sentence type according to language status.

5.5 Z-scores

Z-scores quantify the distance between raw scores and the mean score in terms of SD. They were calculated in order to evaluate the sensitivity and specificity measures for the VSTM and SR tests and gain a better understanding of the relationship between the VSTM, SR, and ASR tests. This was achieved by zooming in on the profile of performance across three tests for each participant in the Language Concerns group and examining whether at group level participants with Language Concerns exhibited consistent or varying profiles.

The first step was to explore the frequency distribution of scores in the Typically Developing sample. For each test, the percentage of children who scored in the following categories was calculated: above the normal range (1 SD or more above the mean); within the normal range (between 1 SD and -1 SD); between 1 to 2 SD below the mean; and at or below -2 SD. Results are presented in Tables 5.33 for the VSTM and SR tests and 5.34 for the ASR test. Table 5.34 also shows the expected percentage for each category of scores in a normal distribution. Descriptive statistics and normality testing indicated a lack of normal distribution in scores especially for performance on the VSTM test. The expected percentages were included to quantify how much scores veered off within each category of scores.

As can be seen from Table 5.33 and 5.34 most of the scores fell within normal limits. In the case of the VSTM and SR tests, Digit span, Lexical and Grammatical Morpheme scores largely corresponded with the expected percentages for a normal distribution. For Word span, scores between 1 and -1 SDs were over represented (due to narrow range of scores in 2;6 to 2;11 and 4;0 to 4;5) . For Nonword span and Total span children scoring more than 2 SD below the mean were under represented.

In the absence of standardized clinical assessments for Najdi Arabic, sensitivity and specificity values were judged against the concern of parents and teachers regarding children's language development. Sensitivity was calculated as the proportion of children in the Language Concerns group who scored at or below four commonly used clinical cutoffs 1, 1.25, 1.5 and 2 SDs below the mean (Conti-Ramsden, et al. 2001; Paul, 2007). Specificity was calculated based on the number of children in the entire Typical sample scoring above -1, -1.25, -1.5 and -2 SDs. Sensitivity and specificity values of the VSTM and SR test scores are given in Table 5.35. The "ideal" cut off for each test is highlighted in bold in Table 5.35 and was determined based on the combination of sensitivity and specificity exceeding the minimum requirement of Plant and Vance's 80% (1994). Priority is given for a higher sensitivity value. Although there is a trade-off between sensitivity and specificity, it is more important for the purpose of this study to avoid missing children who may require further assessment and services than falsely identifying children from the Typical sample as having possible language impairment. Examination of Table 5.35

shows that for the VSTM test Word, Digit and Total span (cumulative score of the three VSTM subtests) the best cutoffs ranged from 1 to 1.25 SDs below the mean. Nonword span failed to reach acceptable sensitivity values at any of the cutoffs. For the SR test, the ideal cut-off for all three scores ranged from 1 to 1.25 SDs below the mean. The Grammatical Morpheme score was the only score that exceeded 80% even in the more stringent cut-off of 2 SD below the mean.

The next step was to establish the individual profile of performance for children in the Language Concerns group. Tables 5.36 to 5.39 illustrate the individual performance on all three tests based on z-scores. On the left side of the tables, the raw scores corresponding to the z –scores of 0, -1, and -2 for each age group are presented. On the right side, the individual's z-scores are presented using a color-coding system: scores falling within the normal range are highlighted in green; scores falling between 1 to 2 SD below the mean are highlighted in amber; and scores of 2 or more SD below the mean are highlighted in red. Most children scored consistently low across the tests (amber or red range) further confirming that they have language difficulty. Z-scores for the Nonword subtest of the VSTM test and the Total Sentence Accuracy score of the SR test fell in the green range for children in the younger age groups, reflecting floor effects.

Just four of the 16 children showed a mismatch in performance across the tasks. Their profile of performance across the three tasks relative to their Typically Developing peers was examined to see if it could provide an indication on what's causing them to have difficulty with some but not all the immediate repetition tests. To allow for ease of comparison, the z-sores for the four participants on all measures are summarized in Table 5.40. It was only possible to re-assess participant 148 to establish the consistency of the mismatch over time, and his performance in Time 2 was in keeping with the profile exhibited in Time 1.

On the VSTM test, 148's Span scores were in line with peers. 150's Digit and Word Recall was in line with peers while Nonword Recall was in the red range surpassing 3 SD below group mean. 147 obtained span scores in the amber to red range with the greatest degree of impairment in Digit Recall which was 6 SD below the group mean. 153's span scores were all in the amber range around 1.5 SD below group mean.

On the Sentence Repetition test, 147 was the only participant with a Grammatical Morpheme score in line with peers. The remaining participants were mainly in the red range, in keeping with the vulnerability of Grammatical Morpheme in the Language Concerns group. For the Lexical Morpheme score, 147 and 153 were in line with peers while 148 and 150 obtained Lexical Morpheme scores that were more than 2 SD below the group mean of the Typically Developing group.

On the Anomalous Sentence Repetition test, again Grammatical Morpheme score was impaired in three out of the four participants across all conditions. 147's Grammatical morpheme

score was in line with peers for the Typical and Semantically Anomalous conditions but was just above 1.5 SD below the mean for Syntactically Anomalous sentences. For the Lexical Morpheme score, 153 was in line with peers in all conditions. 148 only showed difficulty with Typical sentences with Lexical Morpheme scores for Semantically and Syntactically Anomalous sentences within 1 SD of Typically Developing group's mean. 150 showed the opposite profile with impaired Lexical Morpheme scores for Syntactic and Semantically Anomalous sentences in the amber to red range and intact imitation of Typical sentences. 147 showed the greatest impairment in Lexical Morpheme score for Syntactically Anomalous sentences scoring more than 2 SD below peers and to less difficulty with Semantically Anomalous sentences with a score just above 1SD below mean of peers.

Taken together:

- 147 shows consistent difficulty with imitation tasks that provide less linguistic information, indicating a possible underlying weakness in memory.
- 148 shows the opposite profile with intact performance on the VSTM subtests, impaired SR scores relative to peers and an improvement in Lexical Morpheme scores in the ASR test the more it is stripped of linguistic information. Indicating that memory skills are in line with peers.
- 150 shows inconsistent poor memory partially explaining performance
- 153 shows consistently intact Lexical Morpheme across SR and ASR tests indicating a possible strength in lexical/semantic knowledge.

Table 5.33 *Percentage of Children Scoring Above, Within, and Below Normal Range on the Verbal Short Term Memory and Sentence Repetition Tests*

Category	**VSTM**				**SR**			**Expected Percentage**	
	Word	**Digit**	**Nonword** [a]	**Total** [a]	**Lex**	**Gram**	**TSA**		
≥ +2 SD	2.86	2.14	1.67	3.33	1.43	1.43	1.43	2.3	
≥ +1 SD to < +2 SD	3.57	19.29	17.5	10	12.14	16.43	12.86	13.6	
< +1 SD to > -1 SD	**80.71**	**62.14**	**62.5**	**70**	**71.43**	**67.14**	**71.43**	**68.2**	**84.1**
≤ - 1 SD to > -2 SD	11.43	14.29	18.33	16.67	12.14	12.14	13.57	13.6	
≤ - 2 SD	1.43	2.14	0	0	2.86	2.86	0.71	2.3	

Note. n = 140; Lex = Lexical Morpheme; Gram = Grammatical Morpheme; TSA = Total Sentence Accuracy.
[a] n = 120

Table 5.34 *Percentage of Children Scoring Above, Within, and Below Normal Range on the Anomalous Sentence Repetition Test*

Category	**Lex**			**Gram**		
	Typical	**Semantically Anomalous**	**Syntactically Anomalous**	**Typical**	**Semantically Anomalous**	**Syntactically Anomalous**
≥ +2 SD	0	0	0	0	0	1.25
≥ +1 SD to < +2 SD	17.5	17.5	18.75	18.75	17.5	18.75
< +1 SD to > -1 SD	**68.75**	**68.75**	**60**	**67.5**	**67.5**	**62.5**
≤ - 1 SD to > -2 SD	10	12.5	21.25	8.75	12.5	15
≤ - 2 SD	3.75	1.25	0	5	2.5	2.5

Note. n = 80; Lex = Lexical Morpheme; Gram = Grammatical Morpheme.

Table 5.35 *Sensitivity and specificity values for the VSTM and Sentence Repetition tests*

	Sensitivity				**Specificity**			
	-1	-1.25	-1.5	-2	-1	-1.25	-1.5	-2
VSTM								
Word	**87.5**	**87.5**	50	37.5	**87.14**	**87.14**	95.71	98.57
Digit	**87.5**	81.25	31.25	25	**83.57**	89.29	97.86	97.86
Nonword	56.25	56.25	37.5	18.75	81.67	**89.17**	95	100
Total	**93.33**	73.33	73.33	33.33	**83.33**	95	95	100
Sentence Repetition								
Lex	87.5	**87.5**	75	56.25	85	**87.86**	95.71	97.14
Gram	**93.75**	87.5	81.25	81.25	**85**	92.14	94.29	97.14
TSA	81.25	**81.25**	56.25	50	86.43	**92.86**	95.71	99.29

Lex = Lexical Morpheme; Gram = Grammatical Morpheme; TSA = Total Sentence Accuracy

Table 5.36 *Z-scores of Participants in the Language Concerns Group for the Verbal Short Term Memory Test*

Age Group	Raw Scores												Case	Z Scores			
	Word			Digit			Nonword			Total				Word	Digit	Nonword	Total
	0	-1	-2	0	-1	-2	0	-1	-2	0	-1	-2					
2;6-2;11	2	1.44	.88	2.08	1.35	.62	1	0	0				141	-1.79	-1.48	0	
3;0-3;5	2.45	1.85	1.25	2.68	2.03	1.38	1.10	1.03	.96	6.81	5.34	3.87	142	-1.58	-1.05	-1.43	-1.57
3;6-3;11	2.68	2.19	1.7	3.28	2.64	2	1.45	.97	.49				143	-1.38	-2.0	-.94	-1.23
													144	-1.38	-2.0	-.94	-1.23
													145	-1.38	-2.0	-.94	-1.23
4;0-4;5	2.95	2.8	2.65	3.48	2.95	2.42	1.60	1.12	.64	8.29	6.98	5.67	146	-13	-4.68	-1.25	-4.04
													147	-6.33	-1.85	-1.25	-2.13
4;6-4;11	3.43	2.78	2.13	3.63	2.94	2.25	1.50	1.01	.52				148	-.66	-.91	1.02	-.22
5;0-5;5	3.55	2.95	2.35	3.83	3.24	2.65	2	1.72	1.44	9.41	8.15	6.89	149	-2.58	-1.41	-3.57	-2.71
													150	-.92	-.56	-3.57	-1.52
													151	-2.58	-1.41	-3.57	-2.71
5;6-5;11	3.68	3.21	2.74	3.95	3.3	2.65	1.83	1.36	.89				152	-2.51	-1.46	.36	-1.52
													153	-1.45	-1.46	-1.77	-1.91
													154	-1.45	-1.46	-1.77	-1.91
													155	-1.45	-1.46	-.70	-1.52
													156	-3.57	-1.46	-1.77	-2.71

Table 5.37 *Z-scores of Participants in the Language Concerns Group for the Sentence Repetition Test*

Age Group	Raw Scores									Case	Z scores		
	TSA max=42			Lex max=56			Gram max=117				TSA	Lex	Gram
	0	-1	-2	0	-1	-2	0	-1	-2				
2;6-2;11	1.25	-	-	21.85	12.46	3.07	29.95	8.28	-	141	-.51	-1.58	-1.2
3;0-3;5	4.9	-	-	30.2	19.65	9.1	50.9	25.17	-	142	-.63	-1.35	-1.31
3;6-3;11	13.75	3.88	-	41.55	32.34	23.13	77.55	56.33	35.11	143	-1.39	-1.8	-2.52
										144	-1.29	-2.01	-2.05
										145	-1.39	-1.69	-2.43
4;0-4;5	16.4	8.02	-	45.7	39.84	33.98	87.9	74.32	60.74	146	-1.96	-5.75	-5.66
										147	.31	.73	.52
4;6-4;11	23.65	15.88	8.11	49.7	46.28	42.86	97.75	88.04	78.33	148	-1.37	-1.37	-2.45
5;0-5;5	26.25	17.54	8.83	51.7	48.22	44.74	100.9	90.78	80.66	149	-2.67	-5.66	-5.92
										150	-2.78	-2.21	-3.84
										151	-2.9	-5.37	-5.23
5;6-5;11	31.4	25.01	18.62	52.7	49.76	46.82	107.45	101.39	95.33	152	-3.97	-5	-8.99
										153	-2.88	.1	-5.19
										154	-2.1	-2.96	-2.38
										155	-4.76	-6.7	-11.96
										156	-3.97	-8.7	-8.82

Note. Lex = Lexical Morpheme; Gram = Grammatical Morpheme; TSA = Total Sentence Accuracy.

Table 5.38 *Z-scores of Participants in the Language Concerns Group for the Lexical Morpheme Score of the Anomalous Sentence Repetition Test*

Age Group	Raw Lexical Morpheme Scores max=32									Case	Z Scores		
	Typical			Semantically Anomalous			Syntactically Anomalous				Typical	Semantically Anomalous	Syntactically Anomalous
	0	-1	-2	0	-1	-2	0	-1	-2				
4;0-4;5	25.6	21.48	17.36	23.65	20.23	16.81	20.65	17.69	14.73	146	-4.51	-4.58	-5.63
										147	.58	-1.07	-2.25
4;6-4;11	28.3	25.78	23.26	26.65	23.72	20.79	23.65	20.2	16.75	148	-1.31	-.9	.39
5;0-5;5	29.35	26.94	24.53	28.75	25.81	22.87	26.6	22.97	19.34	149	-4.71	-6.38	-4.3
										150	-.98	-1.28	-2.37
										151	-5.95	-5.02	-2.09
5;6-5;11	30.05	28.12	26.19	29.7	27.8	25.9	27	24.1	21.2	152	-5.21	-8.79	-6.55
										153	-.54	.16	-.34
										154	-2.1	-2.47	-2.76
										155	-5.21	-7.21	-4.83
										156	-7.8	-5.11	-4.14

Table 5.39 *Z-scores of Participants in the Language Concerns Group for the Grammatical Morpheme Score of the Anomalous Sentence Repetition Test*

Age Group	Raw Grammatical Morpheme Scores max=69									Case	Z Scores		
	Typical			Semantically Anomalous			Syntactically Anomalous				Typical	Semantically Anomalous	Syntactically Anomalous
	0	-1	-2	0	-1	-2	0	-1	-2				
4;0-4;5	51.5	43.88	36.26	50.25	43.06	35.87	38.05	32.05	26.05	146	-5.97	-6.29	-5.18
										147	.33	-.73	-1.51
4;6-4;11	57.15	51.08	45.01	56.75	50.07	43.39	44.2	37.58	30.96	148	-2.99	-1.91	-1.69
5;0-5;5	60.55	55.27	49.99	59	53.45	47.9	46.8	39.64	32.48	149	-6.73	-5.95	-2.35
										150	-4.27	-4.68	-2.49
										151	-6.35	-4.68	-2.07
5;6-5;11	62.15	57.92	53.69	61.35	57.67	53.99	49.6	43.61	37.62	152	-7.13	-12.32	-6.11
										153	-3.82	-2.54	-1.44
										154	-2.16	-4.44	-3.61
										155	-8.78	-11.24	-4.44
										156	-6.42	-7.16	-4.11

Table 5.40 *Summary of Children Showing a Mismatch in Performance*

Case	STM				SR			ASR					
									Lex			Gram	
	Digit	Word	Nonword	Total	TSA	Lex	Gram	Typ	Sem	Syn	Typ	Sem	Syn
147	-6.33	-1.85	-1.25	-2.13	.73	.52	**.31**	**.58**	**-1.07**	**-2.25**	**.33**	**-.73**	**-1.51**
148	-.66	-.91	1.02	-.22	-1.37	-2.45	**-1.37**	**-1.31**	**-.9**	**.39**	**-2.99**	**-1.91**	**-1.69**
150	-.92	-.56	-3.57	-1.52	-2.78	-2.21	-3.84	-.98	-1.28	-2.37	-4.27	-4.68	-2.49
153	-1.45	-1.46	-1.77	-1.91	-2.88	.1	-5.19	-.54	.16	-.34	-3.82	-2.54	-1.44
148 T_2	**-.91**	**-.66**	**1.02**	**-.98**	**-2.25**	**-5.12**	**-2.53**	**-2.5**	**.46**	**-.77**	**-5.46**	**-4.15**	**-3.18**

Note. Lex = Lexical Morpheme; Gram = Grammatical Morpheme; Typ = Typical Sentences; Sem = Semantically Anomalous sentences; Syn = Syntactically Anomalous sentences

5.6 Error Analysis

The errors for the VSTM test was investigated in order to identify possible error patterns, with regards to frequency and type of errors, according to age and language status groups. The error patterns were also compared across subtests in the VSTM test to investigate whether error patterns provided insight into the underlying processes involved in each test.

5.6.1 Verbal Short Term Memory Errors

For each subtest, the errors that occurred in three trials above the span of participants were coded as omission, substitution, perseveration, unintelligible, and order, which was further broken down into item and phoneme migration. Errors made by Typically Developing children were first pooled for each of the seven age groups, and for each of the three subtests. The distribution of errors was determined for each pool by calculating the percentage of occurrence of each error category within the pool. A similar pattern emerged for each subtest in most, if not all, the age groups. Therefore, participant errors in the three subtests were combined, regardless of the age group of participants. Table 5.41 provides a breakdown of the proportion of errors for the three subtests in Typically Developing participants. Pie charts in Figure 5.23 (a-c) illustrate the distribution of errors.

Omission was the most common error type for Typical Participants in the Digit Recall and Word List Recall subtests, while phoneme migration was the most common error for Nonword List Recall. In Digit Recall, the second most common error type was item migration, followed by substitution. Unintelligible and phoneme migration errors did not occur at all. In the Word List Recall, an opposite trend appeared in the proportion of errors, with substitution errors substantially more common than item migration. In Nonword List Recall, substitution errors closely followed phoneme migration errors, with omission errors a distant third.

The errors for all 16 participants with Language Concerns were pooled, and the percentage occurrence of each error type was calculated (see Table 5.42 and Figure 5.23 (a-c)).

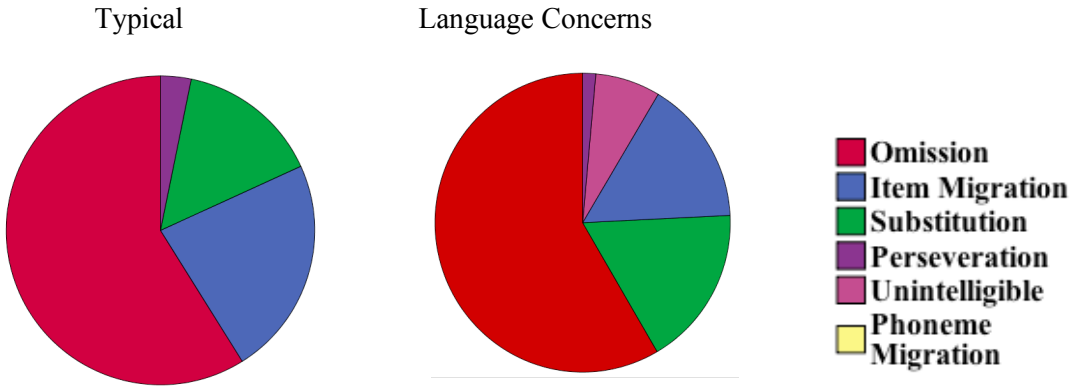
Table 5.41 *Proportion of Errors in the Three Subtests of the Verbal Short Term Memory Test for Typically Developing Participants*

Subtest	**Total Error**	**Migration**		**Omission**	**Substitution**	**Perseveration**	**Unintelligible**
		Item	**Phoneme**				
Digit Recall	781	22.91	0	58.90	14.98	3.20	0
Word List Recall	764	8.24	0.65	60.21	25.65	3.40	1.83
Nonword List Recall	711	1.12	45.15	11.53	38.68	0.84	2.67

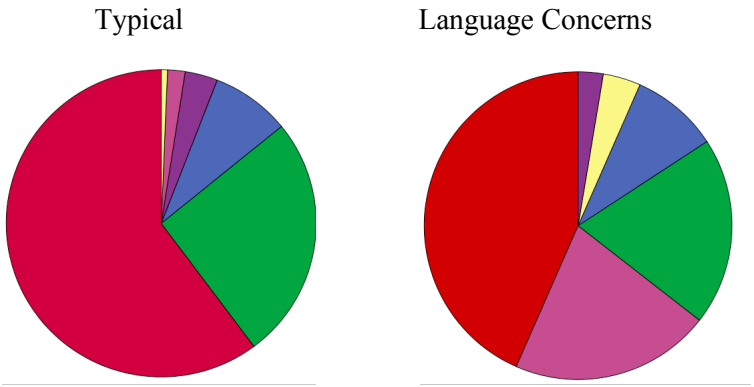
Table 5.42 *Proportion of Errors in the Three Subtests of the Verbal Short Memory Test for the Language Concerns Group*

Subtest	**Total Error**	**Migration**		**Omission**	**Substitution**	**Perseveration**	**Unintelligible**
		Item	**Phoneme**				
Digit Recall	70	15.71	0	58.57	17.14	1.43	7.14
Word List Recall	76	9.21	3.94	43.42	19.73	2.63	21.05
Nonword List Recall	77	0	37.66	12.98	36.36	0	12.98

a) Digit Recall



b) Word List Recall



c) Nonword List Recall

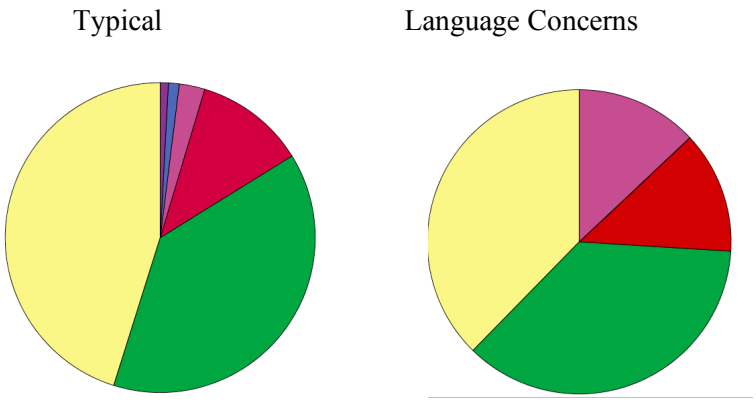


Figure 5.23 Proportion of errors in the three subtests of the Verbal Short Term Memory Test in each language status group.

There were similarities between the Typically Developing participants and those with Language Concerns. Perseveration errors rarely occurred in any of the subtests, and phoneme migration errors did not occur in Digit Recall. The pattern of predominant errors showed a similar trend, with omission as the most common error type in Digit and Word List Recall while phoneme migration was the most common in Nonword List Recall. Substitution errors were slightly higher than item migration in Digit Recall, the reverse of the pattern in the Typical group. The most striking difference between groups was the greater proportion of unintelligible errors in the Language Concerns group for all three subtests, with unintelligible errors being even more common than substitution errors in Word List Recall.

In both groups, whenever substitution errors occurred in Digit Recall and Word List Recall, the test item was replaced with an item from the same linguistic category; for example, digits were always substituted by other digits and never with other real words. In Nonword List Recall, however, nonwords were substituted with either real words that where phonologically similar to the nonword item or another nonword that could not be categorized in any of the other error types.

Chapter Six: Discussion

This thesis examined the immediate repetition abilities of Typically Developing Najdi Arabic-speaking children between the ages 2;6 to 5;11 and children with Language Concerns. Three immediate repetition tests were either developed or adapted: VSTM, SR, and ASR tests. The first stage of the study involved extensive piloting to ensure that the adapted VSTM test was linguistically and culturally appropriate and to establish the best procedure to elicit responses from young children.

A key aim was to examine the clinical utility of the tests. Results show that the levels of test-retest reliability, internal consistency, and inter-rater reliability were generally high. In the absence of standardised language measures for Saudi children, the relationship found between the VSTM and SR tests was taken as a measure of their concurrent validity. The relationship found between subtests of the VSTM test and between outcome measures of the SR test support their construct validity. All three tests were sensitive to the age and language status of participants, confirming evidence for the validity of these measures of development and language ability. Taken together, the findings indicate that the tests are psychometrically robust.

A further aim of the study was to throw light on the nature of the underlying skills involved in the VSTM and SR tests in order to inform their use as clinical assessments. This was achieved by comparing the pattern of performance across different linguistic factors. For Typically Developing participants, (1) Digit span was higher than Word span, and Word span was higher than Nonword span; (2) Lexical Morphemes were repeated more accurately than Grammatical Morphemes; and (3) Typical sentences were easier to recall than Semantically Anomalous sentences followed by Syntactically Anomalous sentences. While the levels of performance were reduced in children with Language Concerns, the pattern of performance was similar to Typically Developing participants. This pattern of findings indicates that rather than merely parroting, Typically Developing children and children with Language Concerns draw on their linguistic knowledge to perform immediate repetition tests.

The chapter commences with a discussion of the key findings relating to: the psychometric properties of the VSTM and SR tests; levels and patterns of performance in the Typical sample; and levels and patterns of performance in the Language Concerns sample. Clinical and theoretical implications are then considered, followed by limitations and future directions.

6.1 Main Findings

6.1.1 Reliability and replication

The inter-rater reliability and test-retest reliability levels of the VSTM, SR, and ASR tests largely fell in the near perfect agreement range (Landis & Koch, 1977), indicating that the three tests were stable across time and raters, and are reliable assessments for Najdi Arabic-speaking preschool children. The high reliability levels for the SR test are consistent with the levels reported in several studies (Carrow, 1974; Devescovi & Caselli, 2007; Gardner et al., 2006; Wallan, 2006). The reliability of the VSTM were largely in agreement with those reported for the WMTB-C (Pickering & Gathercole, 2001). The inter-rater reliability levels for the ASR subtests were consistent with the substantial to near perfect agreement reported for the corresponding conditions in (Polišenská, 2011). Finally, high levels of internal consistency were found for the SR and ASR tests.

If reliability of a test is defined as its ability to produce consistent results under different conditions (Field, 2009), one could argue that replication also falls under reliability. AlKadhi (2012) administered the SR test as part of the assessment battery in the first phase of her study to Typically Developing children recruited from different nurseries than the ones approached in this study. The current study and AlKadhi (2012) were conducted in Riyadh and overlapped in two 6-month age bands (2;6-2;11 and 3-3;5). Table 6.1 provides the mean scores and SD for Lexical Morpheme, Grammatical Morpheme, and Total Sentence Accuracy scores for the overlapping age groups in the two studies. We can see from Table 6.1, that the scores were in agreement, further strengthening evidence of the test's stability.

Table 6.1 *Descriptive Statistics for the Sentence Repetition Test Scores in current study and AlKadhi (2012)*

Age	***N***	**Morpheme Score**				**Total Sentence Accuracy Score**	
		Lexical max: 56		**Grammatical max: 117**		**max: 42**	
		M	**SD**	**M**	**SD**	**M**	**SD**
2;6-2;11	42	21.67	12.63	30.02	22.45	1.88	3.68
	20	21.85	9.39	29.95	21.67	1.25	2.45
3;0-3;5	22	31.14	10.53	49.68	24.97	4.41	5.36
	20	30.20	10.55	50.90	25.73	4.9	7.77

Note. AlKadhi (2012) (black font); this study (red font)

6.1.2 Validity

In the absence of standardized language assessments for Saudi Arabic and since all the tests in the study involve immediate repetition, concurrent validity was evaluated by examining the relationship between VSTM and SR tests for Typically Developing participants. Partial correlations between the tests were all significant and moderate to strong (Cohen, 1992).

Further supporting the SR test's concurrent validity, AlKadhi (2012) found significant moderate to strong correlations between the SR test and the Arabic Picture Vocabulary test (Shaalan, 2010) while partialing out age in months (see Table 6.2). This suggests that both tests are informative about children's language skills and is in keeping with partial correlations found between SR and receptive vocabulary reported in previous studies (e.g. $r = .59$ in Cantonese-speaking TDY group of Stokes et al. (2006); $r = .55$ in French-speaking 4;6 to 5;6 Typically Developing children of Thordardottir et al. (2010))

Table 6.2 *Partial Correlation (Controlling for Age in Months) between Sentence Repetition Scores and the Arabic Picture Vocabulary Test (AlKadhi, 2012)*

Test	Lex	Gram	TSA	APVT
Lex				
Gram	.95**			
TSA	.71**	.84**		
APVT	.54**	.53**	.37*	

Note. Lex = raw Lexical Morpheme score; Gram = raw Grammatical Morpheme score; TSA = Total Sentence Accuracy Score; APVT = Arabic Picture Vocabulary Test
** Correlation is significant at $p < .001$ level (2-tailed)
* Correlation is significant at $p < .01$ level (2-tailed)

6.1.3 Levels and patterns of performance in the Typically Developing sample

All three tests show a significant improvement in the level of performance as age increased with large effect sizes, indicating that the tests can tap into developmental change and have potential to identify children who are unable to repeat at age equivalent levels.

The finding that VSTM improves with age is well documented. This is illustrated in Table 6.3 which reports the estimated span scores for a number of studies that included children between the ages of 2 to 7 years old in their sample and shows that older children obtained higher span scores than younger children within each study, this was true irrespective of language (English, French, Norwegian) or item type (Word, Digit, Nonword). Due to the lack of variation in span scores of the youngest age group in the current study, it is unlikely that it will be an informative test for children younger than 3 years of age. This was not due to children's inability to perform the task since compliance was not an issue.

The sensitivity of the SR to age replicates the findings of Wallan (2006) and extends it to a wider age group. Not only were Lexical and Grammatical Morpheme scores lower on average in younger children, they also showed a large variability in scores within each of the three youngest age groups (2;6 to 3;11). This suggests that the test taps into a period of rapid growth and maps onto variability of scores observed in younger children of the SIT's normative sample (Seeff-Gabriel et al., 2008). The variability on both Morpheme scores reduced as the age of participants increased with participants close to ceiling in the three oldest age groups 4;6 to 5;11. Again this was consistent with the SIT's normative sample (Seeff-Gabriel et al., 2008).

The Total Sentence Accuracy score was also sensitive to age extending the findings of Shalaan's (2009) CELF score in older Gulf Arabic-speaking children to younger children. Finally, improvement in Lexical and Grammatical Morphemes scores in the ASR test is consistent with the significant increase in quantitative scores observed across age in English and Czech-speaking children (Bohannon, 1975, 1976; Polisenska et al., 2015).

For pattern of performance, all three tests provide evidence that supports the contribution of linguistic knowledge to the repetition of Typically Developing children. Children's performance across the subtests of the VSTM differed both quantitatively and qualitatively.

There was a significant effect of item type: span score was highest for Digit Recall, followed by Word List Recall which was higher than Nonword List Recall. The superiority of Digit Recall in comparison to Word List Recall is consistent with the pattern of performance observed in English-speaking children between the ages of 4 to 7 years (DeMarie & Ferron, 2003; Dempster, 1981; Pickering & Gathercole, 2001) (see Table 6.3). A number of possible explanations have been put forward to account for the advantage of Digit Recall over Word List Recall: digits are drawn from a small pool of items of the same semantic category (Hulme & Roodenrys, 1995); digits are highly familiar and largely phonologically distinguishable from each other (Gathercole & Pickering, 2000; Hulme & Roodenrys, 1995); and overall digits have a higher frequency and frequency of co-occurrence rates (Jones & Macken, 2015). We know from studies that manipulated frequency in serial recall tasks that children are better at recalling lists of high frequency items compared to lists of low frequency items (see section 3.1.2 Frequency effect).

The significant effect of lexical status on a span test is in line with findings for 5-year-old children in English (Gathercole et al., 2001) and 6-year-old French-speaking children (Majerus & Van der Linden, 2003) and extends this to children as young as 2;6 years of age. The superior recall of words shows that children draw on their lexical-semantic knowledge to support serial recall.

Table 6.3 *Developmental Differences in Verbal Span Tasks*

Study	Language	Test	2	3	4	5	6	7	Reported Age Effects
Hulme, Thomson, Muir, and Lawrence (1984)[b]	English	Word short			2			3	Yes sig.
		Word medium			1.5			2.2	
		Word long			.7			1.3	
Hitch, Halliday, Dodd, and Littler (1989)	English	Word short			3.09	3.38		4.11	Yes sig.
		Word long			2.67	2.83		3.39	
Henry (1991)	English	Word				3.2		3.7	Yes sig.
Ottem, Lian, and Karlsen (2007)[c]	Norwegian	Word		3.03	3.4	3.56	3.5	3.63	Yes sig.
Noël (2009)	French	Word			3.56	3.87			Yes sig.
Gathercole and Adams (1993)	English	Digit		2.78					--------
DeMarie and Ferron (2003)	English	Digit				4.15	4.4		Not reported
		Word				3.19	3.73		
Dempster (1981)[a,b]	English	Digit	2.33			4.30		5	Review study
		Word	2.75			4		4.25	not reported
Pickering and Gathercole (2001)[b]	English	Digit			3.2	3.7	4.2	4.4	Yes sig.
		Word			2.2	2.4	2.6	2.9	
		Nonword			1.5	1.6	2	1.9	
Current study	Arabic	Digit	2.08 2.68 3.28	3.48 3.63	3.83 3.95				Yes sig.
		Word	2 2.45 2.68	2.95 3.43	3.55 3.68				
		Nonword	1 1.1 1.45	1.6 1.5	2 1.83				

[a] estimated values derived from a 10 study review; [b] exact values not reported in text or tables estimate calculated from figures; [c] only span for phonologically distinct words reported in table.

The significant item type by age group interaction showed that the effect of item type varied with age. The advantage of Digits over Words and Words over Nonwords was examined more closely. While Digit span was higher than Word span across the seven age groups, the magnitude of the advantage was only significant in two age groups 3;6-3;11and 4;0-4;5 with a mean difference of .60 and .56 respectively. The absence of a significant difference in children younger than 3;6 may be attributed to the children's lack of familiarity or knowledge of digits. Although the children's knowledge of digits was not directly assessed in the current study, (Almoammer et al., 2013) examined the comprehension of numbers one to five in 84 Najdi-Arabic speaking children between the ages of 36 to 60 months. A Give-A-Number task was administered and showed that children between 36 to 53 months have not fully acquired number word meaning, with 20% of children 36 to 41 months not knowing the meaning of any numbers (see Figure 6.1). The lack of significant difference in Digit span compared to Word span in children older than 4;5 years may be a result of the growth spurt for Word span at the 4;6 to 4;11age group.

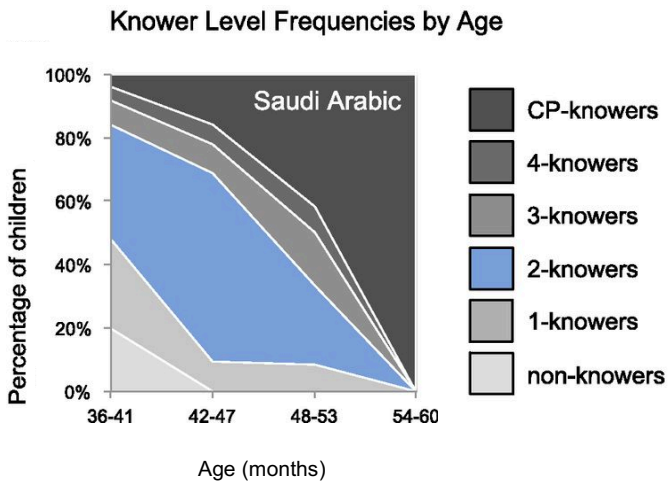


Figure 6.1 Digit comprehension levels of Najdi Arabic-Speaking children aged 24-60 months (*Note.* CP, Cardinal principal); adapted from "Grammatical morphology as a source of early number word meanings" by Almoammer et al. (2013) *Proceedings of the National Academy of Sciences, 110*(46), 18448-18453.

In contrast, Word Recall was superior to Nonword Recall in all age groups and the mean difference was larger in magnitude. Just over 97% of participants showed an advantage of Words over Nonwords. The mean difference varied in magnitude with the key age range being 4;6 to 4;11. Children who were 4;5 years and younger obtained Word span scores that were roughly one item longer than Nonword spans, while children 4;6 years and older were able to repeat Word Lists that were 1.55 to 1.93 items longer than Nonword Lists. The superior recall of words even in the youngest age group is consistent with Roy and Chiat's (2004) finding that 2-year-old children recall more Words than Nonwords in single item recall. They

argued that this shows that even young children benefit from lexical familiarity in spite their limited exposure to words compared to their older counterparts.

The proportion of errors, which occurred in lists that exceeded children's span, revealed that performance on the three subtests was qualitatively different. While omissions were the most common error type made in Digit and Word List Recall they differed in the proportion of item migration and substitution errors. For Word List Recall, substitution errors occurred three times more than item migration errors, whereas an opposite trend was observed in Digit Recall with item migration errors occurring 1.5 times more than substitution errors. The dominance of item migration errors in Digit Recall and substitution errors in Word List Recall is broadly similar to the findings of Burkholder and Pisoni (2004) and Gathercole et al. (2001). Gathercole et al. (2001) attribute the difference in profile to the type of lists the two stimuli items are drawn from: digits are drawn from a highly familiar and closed item pool while words were drawn from an open item pool with varying levels of familiarity.

For Nonword List Recall, omission errors were markedly reduced due to children attempting shorter lists in comparison to Digit and Word List Recall. Omission errors were five times less likely in comparison to other subtests. Interestingly, phoneme migration errors accounted for 45% of errors. This type of error almost never occurred in Digit (0%) or Word List Recall (0.65%). The occurrence of migration errors at the phoneme level rather than item level for Nonword List Recall is consistent with the findings of Gathercole, Frankish, Pickering, and Peaker (1999) and Treiman (1995). Gathercole et al. (1999) found that when children were asked to repeat a list of CVC Nonwords, the majority of errors were comprised of partially correct repetitions that matched the test item in one or two phonemes. Treiman (1995) analysed children's errors even further and found that the substituted phonemes were not random and usually occurred in other CVC combinations from the same list. One explanation put forward is that Nonwords lack the semantic glue that binds phonemes together leading to sub-lexical errors (Patterson, Graham, & Hodges, 1994). Taken together, the qualitative differences further support the contribution of linguistic knowledge to VSTM tests.

On the SR test, Lexical Morpheme scores were significantly higher than Grammatical Morpheme scores. The superiority of Lexical Morphemes was most evident in children between the ages of 2;6 to 3;11 and reduced in magnitude as the age of participants increased until children performed close to ceiling in the oldest age group (5;6-5;11) on both scores. This profile of performance replicates the findings of Wallan (2006) and extends it to a wider age group. It is also in agreement with cross-linguistic repetition studies in two typologically distinct languages (English: the standardization sample of the SIT (Seeff-Gabriel et al., 2008); Italian: (Devescovi & Caselli, 2007)). The greater vulnerability of Grammatical Morphemes in the repetition of younger children may be due to the fact that they are still emerging and not fully acquired.

The telegraphic nature of imitations exhibited by children in the youngest age groups parallels the profile of performance reported by Dashash and Safi (2008) using a parent-reported measure of early vocabulary (Arabic adaptation of the MacArthur-Bates Communicative Development Inventory). The expressive vocabulary of 353 children between 16 to 30 months acquiring Hijazi Arabic (a dialect spoken in the western region of Saudi Arabia) showed that nouns matched predicates in frequency and both were three times more frequent than free Grammatical Morphemes Although the current study looked at both free and bound Grammatical Morphemes, the qualitative similarities between children's imitations and the expressive vocabulary of children speaking a similar dialect of Arabic enhances the test's validity and shows that it is informative about children's language ability.

Turning now to the ASR test, Lexical Morphemes were more likely to be retained than Grammatical Morphemes, as was the case in the SR test. In line with cross-linguistic evidence in adults and 4-5 year old children from English and Czech (G. A. Miller & Isard, 1963; Polisenska et al., 2015) the type of sentence had an impact on recall ability. Typical sentences were the easiest to recall followed by Semantically Anomalous sentences, which were intermediate in difficulty between Typical and Syntactically Anomalous sentences. Syntactically Anomalous sentences posed the greatest difficulty for children. This profile of performance suggests that Arabic-speaking 4-5;11 year old children draw on their semantic and syntactic knowledge during immediate recall of sentences with a possible greater role for syntactic knowledge. However, it is important to note that when morpho-syntactic relations are disrupted inevitably both semantic relations/plausibility and prosody are disrupted as well. Therefore, the lower scores for Syntactically Anomalous sentences could be a result of the combined effect of the disruption of prosody, semantics, and syntax.

To the best of our knowledge, this is the first study to compare Lexical and Grammatical Morpheme scores when examining the influence of violations in semantics and syntax. Interaction analysis revealed that the effect of sentence type varied according to Morpheme type and the interaction between sentence type and Morpheme type also varied according to age group. In the case of Lexical Morpheme score, scores significantly dropped for Semantically Anomalous sentences compared to Typical sentences in the two younger age groups (4;0-4;5 & 4;6-4;11) and almost disappeared in the oldest age groups (5;0-5;5 and 5;6 to 5;11). For the Grammatical Morpheme score, the pattern was consistent across age groups: scores did not significantly differ between Typical and Semantically Anomalous sentences but Grammatical Morphemes in both sentences types were recalled with better accuracy than Syntactically Anomalous sentences. The key difference between the two Morpheme types lies in the Semantically Anomalous condition, which suggests that semantic knowledge has a greater impact on Lexical Morpheme scores as opposed to Grammatical Morpheme scores. The difference in the profile of performance for Lexical Morpheme score across sentence type and age group might be explained by children's mastery of Grammatical Morphemes. For younger

children, the Grammatical Morphemes that are targeted are still emerging and children relay more on their semantic knowledge whereas for older children Grammatical Morphemes are consolidated and children can rely on them to pivot Lexical Morphemes even in the absence of semantic relations. This explanation is consistent with Polisenska and colleague's (2015) finding that 5 year-old children benefit more from familiarity of function words in comparison to content words.

Taken together, all three tests show improvements with age and show evidence that suggests children draw on their linguistic knowledge during immediate repetition tests rather than merely parrot test stimuli. Nonword List Recall (VSTM) and Syntactically Anomalous (ASR) were consistently difficult across age groups. However, the profile of performance with regard to: Digit vs. Word List Recall (VSTM); Lexical vs. Grammatical Morpheme (SR); and Lexical Morpheme in Semantically Anomalous vs. Typical sentence (ASR) change with age this change may be explained by an increase in knowledge base rather than just maturation.

6.1.4 Levels and patterns of performance in Language Concerns sample

Levels of performance on all three tests were markedly lower in children with Language Concerns in comparison to their age and nonverbal IQ matched peers suggesting that the VSTM and SR tests show potential as clinical markers of language impairment in Arabic-speaking Saudi preschoolers. This finding is broadly in keeping with a growing body of cross-linguistic studies which found significant differences between children with SLI and Typically Developing children on immediate repetition tests (VSTM: e.g. English: Archibald and Joanisse (2009); Frizelle and Fletcher (2015); van der Lely and Howard (1993); German: Reichenbach et al. (2016); Dutch: Wilsenach (2006); Sentence Repetition: e.g. Gulf Arabic: Shaalan (2010); English: Conti-Ramsden et al. (2001); Cantonese: Stokes et al. (2006); French: Leclercq et al. (2014); Turkish: Topbaş and Güven (2009)).

Studies mentioned above and the current study differ in the nature of the sample of children with language impairment, preventing exact comparison. In the current study children with language difficulties were recruited based on teacher concerns which was confirmed by parents when possible regarding the progression of language development rather than low performance on a standardized language measure accompanied by a mismatch between nonverbal IQ and language ability. Due to the lack of standardized language and nonverbal IQ measures in Arabic and the recruitment process of the current study, it is not possible to establish whether the children meet the criteria for SLI. Recently the very strict criteria for the diagnosis of SLI have been brought into question (Bishop, 2014; Reilly et al., 2014). In order to overcome the lack of agreement on terminology and criteria a group of 57 international experts across a number of disciplines introduced the term Developmental Language Disorder (DLD) (Bishop, Snowling, Thompson, Greenhalgh, & Catalise Consortium, 2017), which focuses more on functional criteria that are associated with poor prognosis. In retrospect, the Language

Concerns group in the current study is more in line with the term DLD as defined, than with SLI.

Results showed that sensitivity and specificity levels for the VSTM and SR tests exceeded Plant and Vance's (1994) minimum guideline of 80%, further demonstrating that they have potential as clinical markers. The Nonword List Recall subtest was the only exception, identifying only around half of the children in the Language Concerns group. This may be explained by the limited range of Nonword span scores in the Typical sample. The most useful cut-off scores which differentiated between children with Language Concerns and the Typical sample ranged between 1 to 1.25 SDs below the mean. This corresponds to the range of cut-offs (-1 to -1.2 SDs) used by clinicians to identify English-speaking children with language impairment (Records & Tomblin, 1994). The best individual marker was the SR test's Grammatical Morpheme score with relatively high sensitivity (93.75) and specificity (82) at 1 SD below the mean. At 2 SD below the mean, it was the only score that maintained a sensitivity level above 80 while the other SR scores were reduced to a sensitivity level in the 50s. The ability of the Grammatical Morpheme score to maintain good levels of sensitivity even at the more stringent cut-off of -2 SD is consistent with Leclercq et al. (2014) finding that optimal cut-off points for scores in the morpho-syntactic category were -1.7 to -2 SD in comparison to more lax cut-offs scores for lexico-semantic scores followed by global scores. This finding highlights the benefit of using a qualitative score rather than just relying on a purely quantitative/global score and is in keeping with morphosyntactic skills being a key area of impairment in children with DLD (Bishop et al., 2017).

Finally, performance reduction was also observed in the ASR test. Unlike the VSTM and SR tests, the ASR test was included to better understand the underlying processes involved in immediate repetition rather than using the test as a clinical measure. To the best of our knowledge, this is the first study to administer sentences that were manipulated by creating violations in semantic or syntactic rules to children with language difficulties.

Contrary to the levels of performance, the pattern of performance across the three tests were largely similar between children with Language Concerns and their age and nonverbal IQ matched controls. This indicates that even children in the Language Concerns group benefitted from the contribution of linguistic knowledge in support of immediate repetition tests. The two instances in which they diverged, as was evident from significant interactions found in the SR and ASR tests, were in degree of difference between the two groups of children and not direction. Grammatical Morphemes were more vulnerable than Lexical Morphemes in both tests. Also, the Lexical Morpheme score was reduced in Semantically Anomalous sentences compared to Typical sentences in the Language Concerns group only and not controls. Interestingly, these interactions were in line with performance of younger children in the larger Typical sample, suggesting that the performance of children in the Language Concerns group was delayed rather than disordered or atypical. The similarities between children in the

Language Concerns group and the younger children in the Typical sample fits well with the findings of a recent cross-linguistic review by Leonard (2014) which examined the profile of weakness exhibited by children with SLI in morpho-syntax and Nonword Repetition. Leonard (2014) concluded that while the children differed on the particular aspects of each language they found difficult, a common theme emerged:

> ***"All areas of special weakness correspond to details of language that are relatively difficult for typically developing children to acquire. These areas seem to represent the fault lines in each of the languages, detectable but of no particular concern under ordinary conditions. However, in the case of children with limited language skills, they spell serious trouble" (p. 4).***

On the VSTM, span scores and the distribution of errors varied across subtests. In line with the larger Typical sample, there was a significant effect of item type. Span score was highest for Digit Recall followed by Word List Recall which was higher than Nonword List Recall. The advantage of Digit Recall over the other subtests suggests that children in the Language Concerns group also benefit from a greater familiarity with digits. The finding that item order errors were 1.7 times more likely in Digit Recall compared to Word List Recall further supports this. The reduced Nonword span is consistent with the lexicality effect found in older English-speaking children with SLI (van der Lely & Howard, 1993) and extends it to children younger than seven years of age. This indicates that children in the Language Concerns group also benefit from lexical familiarity. Another difference between the Nonword List Recall test and the other subtests was the higher proportion of phoneme migration errors. It accounted for 38% of errors in Nonword List Recall, 4% of errors in Word List Recall and never occurred in Digit Recall. The one striking difference between the frequency of errors between the Language Concerns group and children in the Typical sample is the overall higher rate of unintelligible errors most notably in the Word List Recall. The higher rate of unintelligible errors might be explained by the fact that children in the Language Concerns group have difficulty with speech sounds or it might be a by-product of recruitment process, with parents/teachers more attuned to expressive language difficulties and expressive phonology (Glascoe, 1997).

On the SR test, Lexical Morphemes were repeated with greater accuracy than Grammatical Morphemes in both language ability groups. The greater vulnerability of Grammatical Morphemes in the Language Concerns group is in keeping with cross-linguistic evidence in SR (English: Chiat and Roy, 2008; Seeff-Gabriel et al., 2010; Czech: Smolik and Vavru, 2014; French: Leclercq et al., 2014) and cross-modality in Deaf Children with SLI (Marshall et al., 2015). It is also in line with production studies of Arabic-speaking children with Language Impairment through elicitation techniques (Abdalla, Aljenaie & Mahfoudhi, 2013; Faquih, 2014) or spontaneous language sampling. Faquih (2014) found that 14 Hijazi Arabic-speaking children with language difficulty aged (3;0-6;11) produced bound pronouns

with less accuracy compared to 38 Typically Developing children in the same age range. Abdalla et al. (2013) compared the production of plural noun inflection in 12 Kuwaiti Arabic-speaking children with SLI aged 3;7 to 6;2 and 12 age-matched controls and results showed that children with SLI had greater difficulty producing plural noun inflections than their controls. Finally, the weakness in morpho-syntax is in agreement with spontaneous language of Hijazi Arabic-speaking children with SLI aged 4;0-5;3 (Abdalla & Crago, 2008). Children with SLI significantly differed from MLU and age-matched controls in the correct use of verb tense and subject-verb agreement with a particular weakness in 3rd person agreement.

On the ASR test, the profile of performance extended the influence of semantic plausibility and grammaticality to children with language difficulties. While Typical sentences were repeated with less accuracy than Semantically Anomalous sentences, the difference was not statistically significant. This could be due to the poor performance of children in the Language Concerns group even in Typical sentences. Both Typical and Semantically Anomalous sentences were repeated with better accuracy than Syntactically Anomalous sentences. As previously mentioned, the significant 3-way interaction (morpheme x sentence x language group) showed that the Lexical Morpheme score was reduced in Semantically Anomalous sentence in the Language Concerns group but not controls. The similarity between children in the Language Concerns group and younger children in the Typical sample noted above may be a result of both groups of children not having fully acquired the grammatical morphemes targeted. The disruption of semantic plausibility coupled with impaired grammatical knowledge puts Lexical Morphemes at a higher risk of omission. However, age-matched controls have the benefit of using their grammatical knowledge to override the absence of semantic relations between lexical morphemes. In the Language Concerns group, this might be compounded by a reduction in phonological storage capacity and a reduction in speech sound skills as is evident from the reduced levels of performance on the VSTM test and higher occurrences of unintelligible errors.

To conclude, all three tests were sensitive to language ability. The VSTM and SR tests showed good levels of sensitivity and specificity. Overall the profile of performance indicated that children in the Language Concerns group were also able to benefit from established linguistic knowledge during immediate repetition and were delayed rather than atypical. Grammatical Morphemes were a locus of weakness in the repetition of children with Language Concerns. The similarities between the findings of the current research and cross-linguistic studies as well as studies in different dialects of Arabic with regards to the vulnerability of grammatical morphemes strengthens the SR test's validity.

6.2 Clinical Implications

The current study emerged from the need to develop clinical assessments that are suitable to administer to Najdi Arabic-speaking preschool children as well as inform their use as assessments by determining the underlying skills involved. This is the first study that

explored the clinical utility of a novel SR test and an adapted VSTM test for Najdi Arabic-speaking children aged 2;6 to 6 years. The psychometric properties of both measures are robust, sensitive to age and language ability. It's important to acknowledge that it is not sufficient to rely on either or both tests to diagnose children with language difficulties (de Villiers & de Villiers, 2010; Dockrell & Marshall 2015).

The profile of performance across the different linguistic factors suggests that the tests are informative about children's language. Children showed high compliance rates on both tests, and the tests were quick to administer and easy to score. Both tests yielded high levels of sensitivity and specificity, exceeding the minimum guideline of 80% (Plante & Vance, 1994). The normative data and z-scores provide important diagnostic information, supported by the findings of the sensitivity and specificity results. This is of great diagnostic value since it provides clinicians with objective measures to identify children with language difficulties and lays the foundation for creating standardized assessments for Arabic-speaking preschool children.

It is important to consider the age of participants in relation to test targets and scoring method to maximize the information provided from each test. The VSTM test is informative commencing at the age of 3 years onwards. This extends the age range below 4;7 years, the youngest age group included in the WMTB-C standardization sample (Pickering & Gatherchole, 2001). For children younger than 3 years, single item recall tests should be included. In English, single item recall has shown to be informative from 2 years onwards according to the PSRep (Roy & Chiat, 2004). Both Digit Recall and Word List Recall should be administered to balance out the possible effects of item familiarity and how familiarity effects can change with age. Combining the Span score of the three subtests yielded the highest levels of sensitivity (93.33%) and specificity (83.33%) at the threshed of -1 SD. On the SR test, the test was informative from 2;6 years of age onwards. The global TSA score was informative from 3;6 years onwards due to floor effects in younger children.

While both tests differentiated between children in the Language Concerns group and controls the SR test is more informative about children's linguistic ability. Falling below the cut-off scores on the VSTM test would suggest difficulties with phonological storage, lack of familiarity with test items in the case of digits or words, and possibly speech sound difficulties if there were high rates of unintelligible errors. The careful selection of test targets on the SR test allowed for a precise sampling of a range of morpho-syntactic structures that are known to be difficult for Arabic-speaking children with language impairment. Also, the use of a qualitative scoring system allowed for identification of possible areas of weakness that warranted further investigation and were possible targets for therapy.

SR test:

Level 1: (1) TSA score: If performance is in line with peers (above -1.25 SD) we can infer that the child's lexical-semantic and morpho-syntactic knowledge are intact. If performance falls

more than 1.25 SDs below the mean, this would suggest that the child most likely has difficulty with lexical-semantic knowledge and/or morpho-syntactic knowledge and Lexical and Grammatical Morpheme scores should be calculated.

Level 2: (1) Lexical Morpheme score: If performance is in line with peers (above -1.25 SD) we can infer that the child's lexical-semantic knowledge is an area of strength. If performance falls more than 1.25 SDs below the mean, this would suggest that the child most likely has difficulty with lexical-semantic knowledge and scores should be broken down to subcategories in Level 3.

(2) Grammatical Morpheme score: If performance is in line with peers (above -1 SD) we can infer that the child's morpho-syntactic knowledge is an area of strength. If performance falls more than 1 SD below the mean, this would suggest that the child most likely has difficulty with morpho-syntactic knowledge and scores should be broken down to subcategories in Level 3.

Level 3: (1) Lexical: Adjective, Noun, Verb.

(2) Grammatical: Verb tense (perfect/imperfect), preposition, pronoun.

Breaking down scoring in Level 3 can highlight possible areas of strengths/weaknesses within a category as well as highlight possible error patterns, which may guide further assessment and targets for intervention. For example:

- Participant 152 age (5.6) both Lexical and Grammatical Morpheme scores were below cut-off (-5 SD; -8.99 SD)
    - 5/6 Adjectives and 28/34 nouns were correctly imitated while only 6/16 verbs were imitated and omission was the dominant error type
    - 5/6 possessives, 4/5 pronouns were correctly imitated, Verb Gender Agreement were correct for the 6 verbs that were imitated, 1/4 demonstrative, 5/15 prepositions, Verb Tense 3/10 perfect and 3/8 imperfect and omission was the dominant error type.
- Participant 153 age (5.9) Lexical Morpheme score in line with peers. Grammatical Morpheme score well below cut-off (-5.19 SD)
    - 6/6 possessives, 5/5 pronouns, Verb Tense 7/10 perfect, 8/8 imperfect, and 11/13 Verb Gender Agreement were correctly imitated, 0/2 copula, 7/15 prepositions, 2/4 demonstrative, 8/22 determiner and omission were the dominant error type

Examples of systematic error types:

- Participant 148 age (4.7) both Lexical and Grammatical Morpheme scores were below cut-off (-1.37 SD; -2.45 SD)
    - Verb Tense: 9/10 perfect, 1/8 imperfect, Verb Gender Agreement 5/5 for perfect tense, 1/8 imperfect
        - 7/7 of the imperfect verbs were substituted with a tenseless imperative verb e.g.

- Target: ji-Hib imp3msg-love.imp
- Imitation: Hib love.2msg.imperative

- Participant 151 age (5.2) both Lexical and Grammatical Morpheme scores were below cut-off (-5.37 SD; -5.23 SD)
  - 3/15 Preposition
    - 12/12 of the prepositions were substituted with the preposition in e.g.
      - Target: b-ʕaru:s-at-ha: with-doll-fsg-her
      - Imitation: fi: ʕaru:s-at-ha: in doll-fsg-her

6.3 Theoretical Implications: Immediate Repetition - Memory, Language or Both?

The second aim of the study was to examine the contribution of linguistic knowledge to tests of immediate repetition in order to inform their use as clinical assessments. This was achieved by looking at the overall pattern of performance within and across the three immediate repetition tests and examining whether the profile of performance changed with age and language ability.

6.3.1 Implications of linguistic effects on repetition performance in Typically Developing children

The pattern of performance across different linguistic factors within each test suggests that children are not merely parroting when performing the VSTM, SR and ASR tests but rather that they draw on their linguistic knowledge. On the VSTM test, the longer span for words and digits compared to nonwords suggests that children's long-term lexical knowledge facilitates recall. The overall superiority of Digit span compared to Word span coupled with the higher frequency of item order errors when recalling digits indicates that children benefit from a familiarity with digits. On the SR and ASR tests, Lexical Morphemes were recalled more accurately than Grammatical Morphemes suggesting that morpho-syntactic knowledge aids recall. On the ASR test, children recalled Syntactically Anomalous sentences with the least accuracy further supporting the contribution of morpho-syntactic knowledge. Semantically Anomalous sentences were recalled with more accuracy than Syntactically Anomalous sentences but with less accuracy than Typical sentences, indicating that children draw on their semantic knowledge when repeating sentences but that morpho-syntactic knowledge plays a more privileged role. Recall of Lexical Morphemes was influenced by semantic plausibility while recall of Grammatical Morphemes was not.

A distinction has long been made between serial recall tests such as digit span and single item recall tests such as the PSRep (Seeff-Gabriel et al., 2008). Although both types of tests along with SR are tests of immediate repetition, they are largely studied by different groups of researchers with serial recall tests being the focus of cognitive psychologists while single item recall and SR tests are the focus of psycholinguists. This distinction is reflected in the theoretical underpinning of the WMTB-C (Pickering & Gathercole, 2001) with the Digit Recall, Word List Recall and Nonword List Recall subtests designed to tap the phonological

loop component of the working memory model (Baddeley, 1986; Baddeley & Hitch, 1974) and collectively called Phonological Short Term Memory. The VSTM test used in the current study was adapted from the structure of the WMTB-C (Pickering & Gathercole, 2001). In hindsight, it would have been more appropriate to group together Digit Recall, Word List Recall and Nonword List Recall under the term 'Verbal Span' test rather than 'VSTM'. The VSTM label is misleading, as it is a theoretically charged label that undercuts the contribution of established linguistic knowledge.

Essentially all three immediate repetition tests (Verbal Span, SR, ASR) assess verbal recall and could be called VSTM, the difference between them being the degree of contribution each receives from memory and language, which is difficult to disentangle. This difficulty was evident whether the question of underlying processes was addressed via correlational analysis or by via manipulating linguistic factors and properties of VSTM such as length. For example, Everitt (2009), found that digit span and SR tests significantly correlated with single item recall and PLS-3 implicating VSTM and established linguistic knowledge for both tests but SR showed more consistent and stronger correlations with the PLS-3 indicating that established linguistic knowledge is tapped more by SR than Digit span. In manipulation studies, linguistic factors and length were found to have independent effects on Verbal Span (Gathercole et al., 2001; Henry & Millar, 1991; Roodenrys et al., 1993) and SR performance (Moll et al., 2015; Willis & Gathercole, 2001;Wilsenach, 2006) implicating VSTM and linguistic knowledge in both tasks. In Verbal Span tasks, Roodenrys et al. (1993) found that the lexicality effect was present irrespective of item length. On the other hand, Henry and Millar (1991) found that the item length effect was more evident in lists of low frequency words. In SR, length was found to influence certain sentence structures and not others (Willis & Gathercole, 2001) or certain qualitative scores but not others (Wilsenach, 2006; Moll et al., 2015) and the influence of sentence structure was present even in short sentences (Willis & Gathercole, 2001). Riches (2012) suggests that the role of VSTM and established linguistic knowledge is not determined by the length of sentences but rather they both efficiently work together at all sentence lengths. Polisenska et al. (2015) used a unique approach in addressing the interaction of memory and language during immediate repetition. Instead of comparing the profile of performance across long and short sentences, Polisenska et al., (2015) followed the procedure of the WMTB-C and presented successively longer sentences within each of the manipulated linguistic domains and span was used as an outcome measure. Results showed that children's capacity varied according to their knowledge of the sentences to be recalled.

Rather than adopting a polarised view of the underlying skills as being either VSTM or language, we can assume each task receives support from both VSTM and language (Archibald & Joanisse, 2009; Everitt, 2009; Frizelle & Fletcher, 2015; Polisenska et al., 2015; Poll et al. 2010) with the degree of contribution from each varying depending on the task. For example, on the SR test there is more opportunity for support from the linguistic knowledge end of the

continuum in the form of the morpho-syntactic frames in sentences and the semantic relationship between words compared to the unrelated words and absence of grammatical morphemes in the Verbal Span test. We know from the profile of performance on the SR and ASR tests that semantic knowledge and morpho-syntactic knowledge both contribute to recall accuracy. The ASR test falls somewhere between the Verbal Span test (closer to the VSTM end) and the SR test (closer to the established linguistic knowledge end) depending on the linguistic condition. Within the ASR test, Semantically Anomalous sentences are more similar to the SR test since they still have intact morpho-syntactic frames while Syntactically Anomalous sentences are more similar to the Verbal Span test due to the disruption of the morpho-syntactic frames, semantic relationships and prosody.

6.3.2 Implications of linguistic effects across age groups

The profile of performance across some linguistic factors showed a change with age as indicated by significant interactions between age group and linguistic effects. The observed change in both cases was in degree and not direction. On the Verbal Span test, the effect of lexicality increased with age, with the average gap between Word and Nonword Span increased from 1 to 2 items. The difference in degree of improvement with age might be explained by the fact that children's receptive and expressive vocabulary expands with age and the meaning of words and their phonological templates are more readily available to support recall of words. Henry and Millar (1991) and Roodenrys et al. (1993) both suggest that developmental trends in span scores could be explained by an increase in the availability of long-term linguistic representations. In the case of nonwords, the same level of support from established linguistic knowledge is not available. Nonwords in the current study were created from the same pool of phonemes with either an existing consonantal root or vowel template suggesting possible support at a sub-lexical level. This may explain why age effects were also observed for Nonword span but to a reduced degree. Nonword Span increased from a mean Span of 1 to 2 while Word Span increased from a mean of 2 to 3.68.

The effect of item familiarity (Digit > Word Span) was greatest in children between the ages of 3;6 and 4;5. The reduced effect for children younger than 3;6 might be explained by a lack of familiarity with digits. This is in agreement with Almoammer et al. (2013)'s finding that Najdi-Arabic speaking children younger than 41 months have not fully acquired the meaning of numbers 1 to 5. For children older than 4;5 the lack of significance might be due to a growth spurt in word knowledge. Al-Sa'bi (2007) reported that vocabulary development milestones of Jordanian Arabic were similar to the millstones observed for English-speaking children with expressive vocabulary, doubling from 1000 items at the age of 3 years to 2000 items by the age of 4 years (as cited in Faquih, 2014), this roughly corresponds to when the growth spurt in Word span was observed. Taken together, the change in the degree of lexicality and familiarity effects with age might be explained by a difference in degree of support from linguistic knowledge with age for the different linguistic items (Henry & Millar, 1991;

Roodenrys et al., 1993). Both item specific knowledge and general knowledge have been suggested to explain developmental trends in span. In an attempt to untangle the influence of age from knowledge, Chi (1978) compared children who were experts in chess and adult novices on a chess reconstruction task and Digit span. Children were able to recall larger sequences of chess position while adults recalled longer lists of digits. This pattern of findings was replicated in a larger scale study by (W. Schneider, Gruber, Gold, & Opwis, 1993) and suggests that item specific knowledge can reverse expected developmental trends. Ottem, Lian and Karlsen (2007) investigated whether the observed increase in Word Span in 123 typically developing participants between the ages of 3 to 6 years could be explained by improvement in language ability. When receptive grammar and receptive vocabulary scores were entered as a covariate, there was no significant effect of age on span. The authors argued that this indicated that increase in span was due to improvement in receptive grammar and vocabulary and could not be just explained by an increase in VSTM capacity.

On the SR test, the advantage of Lexical over Grammatical Morpheme reduced with age but remained statistically significant in all age groups except in children between the ages of 5;6 to 5;11 who were performing close to ceiling. This profile parallels findings of SR studies cross-linguistically (Carrow, 1974; Devescovi & Caselli, 2007; Seeff-Gabriel et al., 2008) as well as a study of expressive vocabulary in Hijazi Arabic (Dashash & Safi, 2008) which reported that children between the ages of 16 to 30 months produced Lexical Morphemes (nouns and predicates) three times more often than free Grammatical Morphemes suggesting that the SR test is sensitive to morpho-syntactic development. On the ASR test, the influence of sentence type (Typical > Semantically Anomalous > Syntactically Anomalous) was consistent across age groups while both Lexical and Grammatical Morpheme scores improved with age. The lack of change in the linguistic effects of sentence type across age is consistent with Polisenska et al.'s (2015) finding and extends it to Arabic. In view of the typological differences between Arabic, English and Czech, the similarities between the two studies indicate that the relative contribution of semantic and syntactic knowledge to SR performance across age groups of 4 and 5 year olds is universal and not language specific. By using qualitative scores (Lexical and Grammatical Morphemes) rather than span score (Polisenska et al., 2015), the findings of the current study suggest that while the effect of sentence type is consistent across age, the effect of sentence type on morpheme type varied with age. For children younger than 5 years of age, semantic plausibility had an effect on the recall of Lexical Morphemes but this almost disappeared in children older than 5 years of age. On the other hand, recall of Grammatical Morphemes largely remained consistent across age groups and only showed a marked drop in recall accuracy when grammaticality of sentences was disrupted. If we look at the profile of performance of the ASR test in relation to the profile of performance on the SR across age we know that older children are better at recalling Grammatical Morphemes. Also, Polisenska et al. (2015) showed that across age, familiarity

with function words aids recall more than familiarity with content words, further supporting the possible contribution of morpho-syntactic knowledge. Therefore, when semantic plausibility is disrupted, older children may be able to rely on their knowledge of Grammatical Morphemes and morpho-syntactic frames available in the sentence even when Lexical Morphemes are not in predictable slots within a sentence. For younger children, the lack of morpho-syntactic knowledge coupled with the disruption of semantic relations in a sentence has more of a detrimental effect on recall of Lexical Morphemes and suggests that when morpho-syntactic knowledge is compromised, children draw on their semantic knowledge to recall Lexical Morphemes. Taken together, Verbal Span tests are not pure measures of VSTM, the change observed in profile of performance across age for the Verbal Span, SR and ASR tests relates to change in language ability and shows that every verbal recall task will benefit from any level of linguistic knowledge that is relevant to the recall target and indicates that all three tests are capable of identifying atypical language abilities across the age groups examined.

6.3.3 Implications of linguistic effects across language ability groups

The effects of linguistic manipulation in children with Language Concerns as a group largely mirrored their age and nonverbal IQ matched controls. Both language ability groups showed effects of item lexicality and familiarity on the Verbal Span test, effects of Morpheme type on the SR test and effects of semantic plausibility and grammaticality on the ASR test. When performance varied across language ability groups the profile was similar to younger children in the Typical sample as discussed above. On the SR test, the greater vulnerability of Grammatical Morphemes was consistent with production studies of Arabic-speaking children with DLD (Abdalla & Crago, 2008; Abdalla et al. 2013; Faquih, 2014) showing that it is sensitive to the well-documented weakness in morpho-syntax exhibited by children with DLD cross-linguistically (Bishop et al., 2017; Leonard, 2014). On the ASR test, Lexical Morphemes were reduced when semantic plausibility was manipulated but not for controls. The majority of children in the Language Concerns group were low across-the-board on the Verbal Span, SR and ASR tests relative to peers as reflected in their Z-scores. However, they showed similar profiles of performance to peers on the Verbal Span, and where they diverged on the SR and ASR tests, they were similar to the younger Typical sample suggesting that they were able to draw on the same underlying skills when performing tests of immediate repetition. Since children in the Language Concerns group were sensitive to linguistic effects as well and in light of the argument presented above that all three tests of immediate repetition require support from both memory and language to varying degrees, it is not possible to pinpoint whether the underlying nature of their difficulty is due to a weakness in memory or language. The interdependence of memory and language in the performance of children with language difficulties and the similarities between children with language difficulties and younger typically developing children is in agreement with Frizelle and Fletcher's (2015) finding that children with SLI and children in the YTD group showed the same profile of performance

across sentences of varying grammatical complexity that were matched on length and also showed the same profile of associations with Digit span across these sentences types. The authors argued that when syntactic knowledge is not available to support recall, children with SLI and younger typically developing children call more heavily on their VSTM ability to repeat sentences.

Only 4 out of the 16 children showed a mismatch in their performance across tasks. Two of these participants, 147 and 148, showed an extreme mismatch between performance on the Verbal Span and SR tests and closer consideration of the profile of performance can shed more light on the interdependence of memory and language. While rare, there have been few instances where mismatches between the performance on Verbal span or single item recall and language assessment have been reported in the literature. Archibald and Joanisse (2009) reported that 4 out of 9 participants had impaired Verbal Span scores but attained language scores at or above the mean. Chiat and Roy (2007) also reported that few children showed poor performance on the PSRep but had language scores in line with peers as well as the opposite profile. In the current study Participant 147's performance on the SR test was in line with peers while performance on all the Verbal Span subtests was impaired. Participant 148 showed the opposite profile, performance on the Verbal Span test was in line with peers while performance on the SR test was impaired. If we take a polarised view of the underlying skills involved in immediate repetition critiqued above it would seem that 147 has a memory deficit while 148 has a language deficit. Yet despite their respective impairments relative to peers both showed lexicality effects on Verbal span tests and higher recall for Lexical Morphemes compared to Grammatical Morphemes on the SR test indicating that both participants drew on their established linguistic knowledge for Verbal Span and SR tests.

To summarize, while the study is not unique in conclusion, it is first study to explore the effects of linguistic manipulation on immediate repetition tests in Arabic and sheds new light on underlying processes involved.

- The degree of contribution from memory and language can vary according to test.
- The contribution of the different types of linguistic knowledge can change with age.
- The profile of performance of children in the Language Concerns at group level are similar in profile to the young Typically Developing participants suggesting the difference in the profile of performance compared to peers stems from a language processing difficulty.
- Even with the rare extreme mismatch between Verbal Span and SR tests children showed that they can benefit from established linguistic knowledge.

6.4 Limitations and Future Directions

Several limitations need to be considered when interpreting the study findings. The study sample is relatively small with both the Typical sample and Language Concerns groups not fully representing the populations they are drawn from. A robust diagnostic assessment requires evidence of valid and reliable discrimination between fully representative samples.

While the absence of fully representative samples hinders the generalizability of the study findings, it is deemed acceptable in the first phase of establishing a novel measure (Sackett & Haynes, 2002). Participants in the study were recruited based on teacher concern; this may have introduced selection bias. Teachers may have been inclined to select children with moderate to severe language impairment and parents of children with severe language difficulties might have been more motivated to agree to participate in the study. Therefore, the study sample may have more children with severe language impairment than the broader population and this may have inflated sensitivity values. Based on parent education, the sample was skewed towards higher levels of education. Teachers were not asked to record how many parents were approached and how many parents did not agree to take part in the study so it was not possible to estimate response rate to the invitation letters. The availability of assessments for Arabic-speaking children with language impairment remains an issue today. In a recent survey of 101 speech and language therapists in Saudi Arabia, Khoja (2017) reported that 85% used self-translated or adapted English tests to assess language impairment and that the most commonly translated assessment was the PLS (50%).

Future studies could implement a population-based design to better reflect the heterogeneity in terms of language ability and also employ a cross-sectional developmental trajectories approach with Language Impaired children rather than matching (Carney et al., 2013; McKean, Letts, & Howard, 2013). To better understand the underlying skills tapped by the VSTM and SR tests, the relationship between the profile of performance on the two tests and performance on receptive and expressive assessments targeting similar types of knowledge in different age and language ability groups could be examined. For example, including an elicitation test that targets the same Grammatical Morphemes and a Grammaticality Judgment test.

6.5 Conclusion

The current study emerged from the need to provide speech and language therapists with clinically viable assessments for Arabic-speaking preschool children and contribute to the understanding of underlying processes involved in immediate repetition across languages. The significance and novelty of the study lies in the range of immediate repetition tests employed, the design of each test with carefully selected targets, and scoring methods that took into account the rich morphology of Arabic and its heavy reliance on inflections, and finally that it allows for a comparison of the profile of performance across different age and language ability groups. The study results are consistent with a growing body of cross-linguistic evidence demonstrating that SR and Verbal Span tests are sensitive to developmental change and language difficulties and that they are informative about children's language processing abilities. More specifically, it highlights (1) the role of lexical knowledge, morphosyntactic knowledge, and semantic knowledge in immediate repetition, (2) how the contribution from each type of linguistic knowledge is not static but can change with age, and (3) the similarities

between the profile of performance in children with Language Concerns and younger Typically Developing children. It emphasizes the need for research to move away from broad questions such as whether children draw on their memory or language skills when performing immediate repetition to unpacking how the contribution of each type of linguistic knowledge changes with age. Further insight can be gained by examining the profile of performance in immediate repetition tests in relation to receptive and expressive tasks targeting similar types of knowledge structures.

References

Abdalla, F., & Crago, M. (2008). Verb morphology deficits in Arabic-speaking children with specific language impairment. *Applied Psycholinguistics, 29*(2), 315-340. doi:10.1017/S0142716408080156

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont Research Center for Children, Youth, and Families.

Akinsola, E. F. (1986). Effect of Rate, Intonation, and Sentence Length on Nigerian Children's Imitation of Adults' Utterances. *The Journal of Psychology, 120*(5), 439-446. doi:10.1080/00223980.1986.9915475

AlAbdulkarim, L. (2015). The role of speech-language pathologists and audiologists in the schools in Saudi Arabia. *International Journal of Health and Economic Development, 1*(2), 62-69.

Alcock, K. J., Rimba, K., & Newton, C. R. (2012). Early production of the passive in two Eastern Bantu languages. *First Lang, 32*(4), 459-478. doi:10.1177/0142723711419328

AlKadhi, A. (2012). *An Investigation of Relations between Sociocognitive Skills, Motor Imitation and Language Abilities in Young Saudi Children.* City University, London.

AlKadhi, A. (2015). *Assessing early sociocognitive and language skills in young Saudi children.* City University London, London, UK.

Allen, S. E. M., & Crago, M. B. (1996). Early passive acquisition in Inuktitut. *Journal of Child Language, 23*(01), 129-155. doi:10.1017/s0305000900010126

Alloway, T. P. (2007). *The Automated Working Memory Assessment*. London, UK: Pearson Assessment.

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: are they separable? *Child Development, 77*(6), 1698-1716. doi:10.1111/j.1467-8624.2006.00968.x

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., Žaucer, R., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences, 110*(46), 18448-18453. doi:10.1073/pnas.1313652110

Alsari, N. a. M. (2015). *The development of the Arabic lexical neighbourhood test.* (Ph.D.), University College London. Retrieved from http://discovery.ucl.ac.uk/1469481/

Al-Sa'bi, Y. (2007). The Jordanian expressive vocabulary test study. (Unpublished Ph.D). Howard University: Washington D.C., USA.

Amayreh, M. M., & Dyson, A. T. (1998). The acquisition of Arabic consonants. *Journal of Speech, Language, and Hearing Research, 41*(3), 642-653. doi:10.1044/jslhr.4103.642

American Speech-Language-Hearing Association. (2012). ASHA's Recommended Revisions to the DSM-5. Retrieved from http://www.asha.org/uploadedFiles/DSM-5-Final-Comments.pdf

Arab Scholar. (2016). Retrieved from http://www.baheth.info/index.jsp?page=/web/includes/start.jsp

Archibald, L. M., & Gathercole, S. E. (2006). Nonword repetition: a comparison of tests. *Journal of Speech, Language, and Hearing Research, 49*(5), 970-983. doi:10.1044/1092-4388(2006/070)

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research, 52*(4), 899-914. doi:10.1044/1092-4388(2009/08-0099)

Armon-Lotem, S., Haman, E., López, K. J., Smoczynska, M., Yatsushiro, K., Szczerbinski, M., . . . Lely, H. (2016). A large-scale cross-linguistic investigation of the acquisition of passive. *Language Acquisition, 23*(1), 27-56. doi:10.1080/10489223.2015.1047095

Baddeley, A. D. (1986). *Working Memory*. Oxford, UK: Clarendon Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47-89). New York, NY: Academic Press.

Bishop, D. V. M. (1982). *Test for Reception of Grammar*. Manchester, UK: Published by author at Manchester University.
Bishop, D. V. M. (1998). Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 39*(6), 879-891.
Bishop, D. V. M. (2014). Ten questions about terminology for children with unexplained language problems. *International Journal of Language and Communication Disorders, 49*(4), 381-415. doi:10.1111/1460-6984.12101
Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*(4), 391-403. doi:DOI 10.1111/j.1469-7610.1996.tb01420.x
Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Catalise Consortium. (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*. doi:10.1111/jcpp.12721
Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology, 6*(3), 380-420. doi:10.1016/0010-0285(74)90018-8
Bohannon, J. N. (1975). The Relationship between Syntax Discrimination and Sentence Imitation in Children. *Child Development, 46*(2), 444-451. doi:10.2307/1128140
Bohannon, J. N. (1976). Normal and Scrambled Grammar in Discrimination, Imitation, and Comprehension. *Child Development, 47*(3), 669-681. doi:DOI 10.1111/j.1467-8624.1976.tb02230.x
Bonvillian, J. D., Raeburn, V. P., & Horan, E. A. (1979). Talking to children: the effects of rate, intonation, and length on children's sentence imitation. *Journal of Child Language, 6*(3), 459-467.
Botting, N. (2005). Non-verbal cognitive development and language impairment. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 46*(3), 317-326. doi:10.1111/j.1469-7610.2004.00355.x
Botting, N., & Conti-Ramsden, G. (2003). Autism, primary pragmatic difficulties, and specific language impairment: can we distinguish them using psycholinguistic markers? *Developmental Medicine and Child Neurology, 45*(8), 515-524. doi:10.1111/j.1469-8749.2003.tb00951.x
Botting, N., & Conti-Ramsden, G. (2007). Autism, primary pragmatic difficulties, and specific language impairment: can we distinguish them using psycholinguistic markers? *Developmental Medicine and Child Neurology, 45*(8), 515-524. doi:10.1111/j.1469-8749.2003.tb00951.x
Boulianne, L., & Labelle, M. (2006, 2006). *Working version of a French adaptation of the digit span subtest of the CELF–4 and the Following Directions subtest*. Unpublished manuscript. Montreal, Quebec, Canada.
Brown, R. (1973). *A first language: The early stages* (Vol. xx). Oxford, England: Harvard U. Press.
Brown, R., & Bellugi, U. (1964). Three Processes in the Child's Acquisition of Syntax. *Harvard Educational Review, 34*(2), 133-151.
Brown, R., & Fraser, C. (1963). The acquisition of syntax. In C. N. Cofer & B. Musgrave (Eds.), *Verbal Behavior and Learning: Problems and Processes* (pp. 158-201). New York, NY: McGraw-Hill.
Burgemeister, B., Blum, L., & Lorge, I. (1972). *The Columbia Mental Maturity Scale*. New York, NY: Harcourt Brace Jovanovich.
Burkholder, R., & Pisoni, D. (2004). Digit span recall error analysis in pediatric cochlear implant users. *International Congress Series / Excerpta Medica, 1273*, 312-315.
Cain, K. (2006). Individual differences in children's memory and reading comprehension: an investigation of semantic and inhibitory deficits. *Memory, 14*(5), 553-569. doi:10.1080/09658210600624481
Carney, D. P., Henry, L. A., Messer, D. J., Danielsson, H., Brown, J. H., & Ronnberg, J. (2013). Using developmental trajectories to examine verbal and visuospatial short-term

memory development in children and adolescents with Williams and Down syndromes. *Research in Developmental Disabilities, 34*(10), 3421-3432. doi:10.1016/j.ridd.2013.07.012

Carrow, E. (1974). A test using elicited imitations in assessing grammatical structure in children. *Journal of Speech and Hearing Disorders, 39*(4), 437-444. doi:10.1044/jshd.3904.437

Carrow-Woolfolk, E. (1985). *Test for Auditory Comprehension of Language-Revised*. Allen, TX: DLM Teaching Resources.

Catts, H. W. (1993). The Relationship Between Speech-Language Impairments and Reading Disabilities. *Journal of Speech, Language, and Hearing Research, 36*(5), 948-958. doi:10.1044/jshr.3605.948

Chevrie-Muller, C., Maillart, C., Simon, A. M., & Fournier, S. (2010). *Batterie langage oral, langage e´crit, me´moire, attention [Battery for oral language, writing, memory, attention* (2nd edition ed.). Montreuil, France: Les Éditions du Centre de Psychologie Appliquée.

Chi, M. T. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73-96). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Chiat, S., Armon-Lotem, S., Marinis, T., Polišenská, K., Roy, P., & Seeff-Gabriel, B. (2013). The potential of sentence imitation tasks for assessment of language abilities in sequential bilingual children. In V. C. Mueller-Gathercole (Ed.), *Issues in the Assessment of Bilinguals* (pp. 56-89). Bristol, UK: Multilingual Matters.

Chiat, S., & Roy, P. (2007). The preschool repetition test: an evaluation of performance in typically developing and clinically referred children. *Journal of Speech, Language, and Hearing Research, 50*(2), 429-443. doi:10.1044/1092-4388(2007/030)

Chiat, S., & Roy, P. (2008). Early phonological and sociocognitive skills as predictors of later language and social communication outcomes. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 49*(6), 635-645. doi:10.1111/j.1469-7610.2008.01881.x

Chomsky, N. (1965). *Aspects of the theory of syntax*. Oxford, England: M.I.T. Press.

Coady, J. A., Mainela-Arnold, E., & Evans, J. L. (2013). Phonological and lexical effects in verbal recall by children with specific language impairments. *International Journal of Language and Communication Disorders, 48*(2), 144-159. doi:10.1111/1460-6984.12005

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

Conti-Ramsden, G., & Botting, N. (2008). Emotional health in adolescents with and without a history of specific language impairment (SLI). *Journal of Child Psychology and Psychiatry and Allied Disciplines, 49*(5), 516-525. doi:10.1111/j.1469-7610.2007.01858.x

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(6), 741-748.

Conti-Ramsden, G., St Clair, M. C., Pickles, A., & Durkin, K. (2012). Developmental trajectories of verbal and nonverbal skills in individuals with a history of specific language impairment: from childhood to adolescence. *Journal of Speech, Language, and Hearing Research, 55*(6), 1716-1735. doi:10.1044/1092-4388(2012/10-0182)

Cornelius, S. A. (1974). *A comparison of the elicited language inventory with the developmental sentence scoring procedure assessing language disorders in children.* (M.Sc.), University of Texas, Austin, TX.

Courcy, A. (2000). *Conscience phonologique et apprentissage de la lecture (Phonological awareness and reading acquisition).* (Ph.D.), Université de Montréa, Montreal, Quebec, Canada.

Dailey, K., & Boxx, J. R. (1979). A Comparison of Three Imitative Tests of Expressive Language and a Spontaneous Language Sample. *Language, Speech, and Hearing Services in Schools, 10*(1), 6-13. doi:10.1044/0161-1461.1001.06

Dale, P. S. (1976). *Language Development: Structure and Function*. New York, NY: Holt Rinehart and Winston.

Dashash, N., & Safi, S. (2008). *Investigating Lexical Development of Hijazi Infants & Toddlers Using the Arabic Version of the MCDI*. Paper presented at the ASHA Convention, Chicago, IL.

de Bree, E. H. (2007). *Dyslexia and phonology: A study of the phonological abilities of Dutch children at-risk of dyslexia.* Utrecht, NL. Retrieved from https://books.google.co.uk/books?id=sT4mAQAAIAAJ

de Villiers, P. A., & de Villiers, J. G. (2010). Assessment of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(2), 230-244.

Deevy, P., Weil, L. W., Leonard, L. B., & Goffman, L. (2010). Extending Use of the NRT to Preschool-Age Children With and Without Specific Language Impairment. *Language Speech and Hearing Services in Schools, 41*(3), 277-288. doi:10.1044/0161-1461(2009/08-0096)

DeMarie, D., & Ferron, J. (2003). Capacity, strategies, and metamemory: Tests of a three-factor model of memory development. *Journal of Experimental Child Psychology, 84*(3), 167-193. doi:10.1016/S0022-0965(03)00004-3

Dempster, F. N. (1981). Memory span: Sources of individual and developmental differences. *Psychological Bulletin, 89*(1), 63-100. doi:10.1037/0033-2909.89.1.63

Demuth, K. (1990). Subject, topic and Sesotho passive. *Journal of Child Language, 17*(1), 67-84.

Devescovi, A., & Caselli, M. C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language and Communication Disorders, 42*(2), 187-208. doi:10.1080/13682820601030686

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language, 81*(4), 882-906. doi:DOI 10.1353/lan.2005.0169

Dispaldro, M., Leonard, L. B., & Deevy, P. (2013). Real-word and nonword repetition in Italian-speaking children with specific language impairment: a study of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 56*(1), 323-336. doi:10.1044/1092-4388(2012/11-0304)

Dockrell, J. E. (2001). Assessing Language Skills in Preschool Children. *Child and Adolescent Mental Health, 6*(2), 74-85. doi:10.1111/1475-3588.00325

Dockrell, J. E., & Marshall, C. R. (2015). Measurement Issues: Assessing language skills in young children. *Child and Adolescent Mental Health, 20*(2), 116-125. doi:10.1111/camh.12072

Dollaghan, C. (2007). *The Handbook for Evidence-Based Practice in Communication Disorders* (1 edition ed.). Baltimore, MD: Brookes Publishing.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136-1146. doi:10.1044/jslhr.4105.1136

Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test—III*. Circle Pines, MN: American Guidance Service.

Dunn, L., Thériault-Whalen, C., & Dunn, L. (1993). *Échelle de vocabulaire en images Peabody: Adaptation française du Peabody Picture Vocabulary Test*. Toronto, ON: PsyCan.

Ebbels, S. (2014). Introducing the SLI debate. *International Journal of Language and Communication Disorders, 49*(4), 377-380. doi:10.1111/1460-6984.12119

Elin Thordardottir, E., & Gagné, A. (2006). *Limites et possibilités de la narration comme outil de dépistage du trouble spécifique de la lecture (Limitations and possibili- ties of narration as a screening tool for specific reading difficulty)*. Paper presented at the the Annual Conference of the Ordre de Orthophonistes er Audiologistes du Québec, Gatineau, Quebec, Canada.

Elliott, C. D. (1996). *British Abilities Scale* (Second Edition ed.). Windsor, UK: NFER-Nelson.

Elliott, C. D. (1997). *The British Ability Scales II* (2nd edition ed.). Windsor, England: NEFR-Nelson.

Ervin-Tripp, S. M. (1964). Imitation and structural change in children's language'. In Lenneberg, E. H. (ed.), Cambridge, Mass.: M.I.T.Press. In E. H. Lenneberg (Ed.), *New Directions in the Study of Language*. Cambridge, MA: M.I.T. Press.

Everitt, A. (2009). *Speech and language therapy in preschool children : assessing the problems.* (Ph.D.), University of Aberdeen. Retrieved from http://digitool.abdn.ac.uk/webclient/DeliveryManager?pid=53351

Everitt, A., Hannaford, P., & Conti-Ramsden, G. (2013). Markers for persistent specific expressive language delay in 3-4-year-olds. *International Journal of Language and Communication Disorders, 48*(5), 534-553. doi:10.1111/1460-6984.12028

Faquih, N. O. (2014). *Production of Arabic bound pronouns by typically developing children and by children with language disorders.* (Unpublished doctoral dissertation), Howard University, Washington, DC.

Ferguson, C. A. (1959). Diglossia. *WORD: Journal of the International Linguistic Association, 15*(2), 325-340. doi:10.1080/00437956.1959.11659702

Field A. (2009). *Discovering Statistics Using SPSS* (Third Edition edition ed.). Los Angeles, CA: Sage Publications.

Fisher, C., Hunt, C., Chambers, K., & Church, B. (2001). Abstraction and specificity in preschoolers' representations of novel spoken words. *Journal of Memory and Language, 45*(4), 665-687. doi:10.1006/jmla.2001.2794

Fletcher, P., Leonard, L. B., Stokes, S. F., & Wong, A. M. (2005). The expression of aspect in Cantonese-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 48*(3), 621-634. doi:10.1044/1092-4388(2005/043)

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation, 14*(4), 375-399. doi:10.3758/BF03203275

Frizelle, P., & Fletcher, P. (2014). Relative clause constructions in children with specific language impairment. *International Journal of Language and Communication Disorders, 49*(2), 255-264. doi:10.1111/1460-6984.12070

Frizelle, P., & Fletcher, P. (2015). The role of memory in processing relative clauses in children with specific language impairment. *American Journal of Speech-Language Pathology, 24*(1), 47-59. doi:10.1044/2014_AJSLP-13-0153

Fujiki, M., & Brinton, B. (1983). Sampling reliability in elicited imitation. *Journal of Speech and Hearing Disorders, 48*(1), 85-89. doi:10.1044/jshd.4801.85

Fujiki, M., Spackman, M. P., Brinton, B., & Hall, A. (2004). The relationship of language and emotion regulation skills to reticence in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(3), 637-646. doi:10.1044/1092-4388(2004/049)

Fujiki, M., & Willbrand, M. L. (1982). A Comparison of Four Informal Methods of Language Evaluation. *Language, Speech, and Hearing Services in Schools, 13*(1), 42-52.

Gagné, A., & Elin Thordardottir, E. (2006). *La petite histoire des jeunes conteurs: Étude du discours narratif chez les enfants québecois francophones âgés entre 4 et 6 ans (The story of young narrators: Study of the narrative discourse of Quebec francophone children between 4 and 6 years of age)*. Paper presented at the the annual meeting of the Association Francophone Pour le Savoi, Montreal, Quebec, Canada.

Gardner, H., Froud, K., McClelland, A., & van der Lely, H. K. (2006). Development of the Grammar and Phonology Screening (GAPS) test to assess key markers of specific language and literacy difficulties in young children. *International Journal of Language and Communication Disorders, 41*(5), 513-540. doi:10.1080/13682820500442644

Gathercole, S. E. (1995). The assessment of phonological memory skills in preschool children. *British Journal of Educational Psychology, 65*(2), 155-164. doi:10.1111/j.2044-8279.1995.tb01139.x

Gathercole, S. E., & Adams, A. (1993). Phonological working memory in very young children. *Developmental Psychology, 29*(4), 770-778. doi:10.1037/0012-1649.29.4.770

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*(3), 336-360. doi:10.1016/0749-596X(90)90004-J

Gathercole, S. E., & Baddeley, A. D. (1996). *Nonword Memory Test. Bristol, England*. Bristol, UK: University of Bristol.

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 25*(1), 84-95.

Gathercole, S. E., & Pickering, S. J. (2000). Assessment of working memory in six- and seven-year-old children. *Journal of Educational Psychology, 92*(2), 377-390. doi:Doi 10.1037//0022-0663.92.2.377

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *Quarterly Journal of Experimental Psychology. A: Human Experimental Psychology, 54*(1), 1-30. doi:10.1080/02724980042000002

General Authority for Statistics of Kingdom of Saudi Arabia. (2010). *Detailed results of Riyadh (general population and housing census)* Retrieved from https://www.stats.gov.sa/sites/default/files/en-riyadh-pulation-by-gender-govnernorate-nationality_0.pdf

Gilliam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed, Inc.

Glascoe, F. P. (1997). Parents' concerns about children's development: Prescreening technique or screening test? *Pediatrics, 99*(4), 522-528. doi:10.1542/peds.99.4.522

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the Nonword Repetition Performance of Children With and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech Language and Hearing Research, 50*(1), 177. doi:10.1044/1092-4388(2007/015)

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (1994). *Test of Adolescent and Adult Language* (3rd ed ed.). Austin, TX: Pro-Ed, Inc.

Hemingway, B. L., Montague, J. C., & Bradley, R. H. (1981). Preliminary Data on Revision of a Sentence Repetition Test for Language Screening with Black First Grade Children. *Language, Speech, and Hearing Services in Schools, 12*(3), 153-159. doi:10.1044/0161-1461.1203.153

Henry, L. A. (1991). Development of auditory memory span: The role of rehearsal. *British Journal of Developmental Psychology, 9*(4), 493-511. doi:10.1111/j.2044-835X.1991.tb00892.x

Henry, L. A. (2011). *The Development of Working Memory in Children*. Los Angeles, CA: Sage Publications.

Henry, L. A., & Millar, S. (1991). Memory span increase with age: A test of two hypotheses. *Journal of Experimental Child Psychology, 51*(3), 459-484. doi:10.1016/0022-0965(91)90088-a

Hesketh, A., & Conti-Ramsden, G. (2013). Memory and language in middle childhood in individuals with a history of specific language impairment. *PloS One, 8*(2), e56314. doi:10.1371/journal.pone.0056314

Hitch, G. J., Halliday, M. S., Dodd, A., & Littler, J. E. (1989). Development of rehearsal in short-term memory: Differences between pictorial and spoken stimuli. *British Journal of Developmental Psychology, 7*(4), 347-362. doi:10.1111/j.2044-835X.1989.tb00811.x

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language, 30*, 685-701.

Hulme, C., & Roodenrys, S. (1995). Practitioner review: verbal working memory development and its disorders. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 36*(3), 373-398.

Hulme, C., Thomson, N., Muir, C., & Lawrence, A. (1984). Speech rate and the development of short-term memory span. *Journal of Experimental Child Psychology, 38*(2), 241-253. doi:10.1016/0022-0965(84)90124-3

Injoque-Ricle, I., Calero, A. D., Alloway, T. P., & Burin, D. I. (2011). Assessing working memory in Spanish-speaking children: Automated Working Memory Assessment battery adaptation. *Learning and Individual Differences, 21*(1), 78-84. doi:10.1016/j.lindif.2010.09.012

Jespersen, O. (1922). *Language, Its Nature, Development and Origin*. London, UK: Gorge Allen and Unwin.

Johnson, C. J., Beitchman, J. H., Young, A., Escobar, M., Atkinson, L., Wilson, B., . . . Wang, M. (1999). Fourteen-year follow-up of children with and without speech language impairments: Speech language stability and outcomes. *Journal of Speech Language and Hearing Research, 42*(3), 744-760.

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition, 144*, 1-13. doi:10.1016/j.cognition.2015.07.009

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13*(2), 161-176. doi:10.1177/026565909701300204

Khoja, M. A. (2017). A survey of formal and informal assessment procedures used by speech-language pathologists in Saudi Arabia. *Speech, Language and Hearing*. doi: 10.1080/2050571X.2017.1407620

Khomsi, A. (2001). *Evaluation du langage oral (ELO). [Oral language assessment]*. Paris, France: Les Éditions du Centre de Psychologie Appliquée.

Komeili, M., & Marshall, C. R. (2013). Sentence repetition as a measure of morphosyntax in monolingual and bilingual children. *Clinical Linguistics & Phonetics, 27*(2), 152-162. doi:10.3109/02699206.2012.751625

Laing, E., Grant, J., Thomas, M., Parmigiani, C., Ewing, S., & Karmiloff-Smith, A. (2005). Love is ... An abstract word: The influence of lexical semantics on verbal short-term memory in Williams syndrome. *Cortex, 41*(2), 169-179. doi:Doi 10.1016/S0010-9452(08)70891-8

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling Developmental Language Difficulties From School Entry Into Adulthood: Literacy, Mental Health, and Employment Outcomes. *Journal of Speech Language and Hearing Research, 52*(6), 1401-1416. doi:10.1044/1092-4388(2009/08-0142)

Leclercq, A. L., Quemart, P., Magis, D., & Maillart, C. (2014). The sentence repetition task: a powerful diagnostic tool for French children with specific language impairment. *Research in Developmental Disabilities, 35*(12), 3423-3430. doi:10.1016/j.ridd.2014.08.026

Lee, K. Y. S., Lee, L. W. T., & Cheung, P. S. P. (1996). *Hong Kong Cantonese Receptive Vocabulary Test*. Hong Kong: Child Assessment Services.

Lee, L. L., & Canter, S. M. (1971). Developmental sentence scoring: a clinical procedure for estimating syntactic development in children's spontaneous speech. *Journal of Speech and Hearing Disorders, 36*(3), 315-340.

Leonard, L. B. (1989). Language learnability and specific language impairment in children. *Applied Psycholinguistics, 10*(2), 179-202. doi:10.1017/S0142716400008511

Leonard, L. B. (2014). Specific Language Impairment Across Languages. *Child Development Perspectives, 8*(1), 1-5. doi:10.1111/cdep.12053

Leonard, L. B., Wong, A. M., Deevy, P., Stokes, S. F., & Fletcher, P. (2006). The Production of Passives by Children with Specific Language Impairment Acquiring English or Cantonese. *Applied Psycholinguistics, 27*(2), 267-299. doi:10.1017/S0142716406060280

Love, J. M., & Parker-Robinson, C. (1972). Children's Imitation of Grammatical and Ungrammatical Sentences. *Child Development, 43*(2), 309. doi:10.2307/1127538

Lukacs, A., Leonard, L. B., Kas, B., & Pléh, C. (2009). The Use of Tense and Agreement by Hungarian-Speaking Children with Language Impairment. *Journal of Speech,*

Language, and Hearing Research, 52(1), 98-117. doi:10.1044/1092-4388(2008/07-0183)

Majerus, S., & Van der Linden, M. (2003). Long-term memory effects on verbal short-term memory: A replication study. *British Journal of Developmental Psychology, 21*(2), 303-310. doi:Doi 10.1348/026151003765264101

Marchman, V. A., Wulfeck, B., & Weismer, S. E. (1999). Morphological Productivity in Children With Normal Language and SLIA Study of the English Past Tense. *Journal of Speech, Language, and Hearing Research, 42*(1), 206-219. doi:10.1044/jslhr.4201.206

Marinis, T., & Armon-Lotem, S. (2015). Sentence repetition. In S. Armon-Lotem, J. de Jong, & N. Mier (Eds.), *Methods for Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment* (pp. 95-124). Bristol, UK: Multilingual Matters.

Marshall, C. R., Mason, K., Rowley, K., Herman, R., Atkinson, J., Woll, B., & Morgan, G. (2015). Sentence repetition in deaf children with specific language impairment in British sign language. *Language Learning and Development, 11*(3), 237-251. doi:10.1080/15475441.2014.917557

Marshall, C. R., & Morgan, G. (2015). Investigating sign language development, delay, and disorder in deaf children *The Oxford Handbook of Deaf Studies in Language* (Vol. 3, pp. 311-324). Oxford, UK: Oxford University Press.

Mayers, A. (2013). *Introduction to Statistics and SPSS in Psychology* (01 edition ed.). Harlow, UK: Pearson.

McKean, C., Letts, C., & Howard, D. (2013). Developmental change is key to understanding primary language impairment: the case of phonotactic probability and nonword repetition. *Journal of Speech, Language, and Hearing Research, 56*(5), 1579-1594. doi:10.1044/1092-4388(2013/12-0066)

Menyuk, P. (1963). A preliminary evaluation of grammatical capacity in Children. *Journal of Verbal Learning and Verbal Behavior, 2*(5-6), 429-439. doi:10.1016/s0022-5371(63)80044-4

Menyuk, P. (1964). Comparison of Grammar of Children with Functionally Deviant and Normal Speech. *Journal of Speech and Hearing Research, 7*(2), 109-121. doi:10.1044/jshr.0702.109

Menyuk, P., & Looney, P. L. (1972). Relationships among Components of the Grammar in Language Disorder. *Journal of Speech, Language, and Hearing Research, 15*(2), 395-406. doi:10.1044/jshr.1502.395

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2*(3), 217-228. doi:Doi 10.1016/S0022-5371(63)80087-0

Miller, L. M., & Roodenrys, S. (2009). The interaction of word frequency and concreteness in immediate serial recall. *Memory & Cognition, 37*(6), 850-865. doi:10.3758/MC.37.6.850

Miller, W., & Ervin, S. (1964). The development of grammar in child language. In U. Bellugi & R. Brown (Eds.), *The acquisition of language. Monographs of the Society of Research in Child Development.* (Vol. 29, pp. 9-34).

Moerk, E. L. (1977). Processes and products of imitation: Additional evidence that imitation is progressive. *Journal of Psycholinguistic Research, 6*(3), 187-202. doi:10.1007/BF01068019

Moll, K., Hulme, C., Nag, S., & Snowling, M. J. (2015). Sentence repetition as a marker of language skills in children with dyslexia. *Applied Psycholinguistics, 36*(2), 203-221. doi:10.1017/S0142716413000209

Montgomery, J. W., Magimairaj, B. M., & Finney, M. C. (2010). Working memory and specific language impairment: an update on the relation and perspectives on assessment and treatment. *American Journal of Speech-Language Pathology, 19*(1), 78-94. doi:10.1044/1058-0360(2009/09-0028)

Nadler, R. T., & Archibald, L. M. D. (2014). The assessment of verbal and visuospatial working memory with school age canadian children. *ResearchGate, 38*(3), 262-279.

Naglieri, J. A. (2003). *Naglieri Nonverbal Ability Test— Individual Administration*. San Antonio, TX: Harcourt Assessment.
Nation, K., Adams, J. W., Bowyer-Crane, C. A., & Snowling, M. J. (1999). Working memory deficits in poor comprehenders reflect underlying language impairments. *Journal of Experimental Child Psychology, 73*(2), 139-158. doi:10.1006/jecp.1999.2498
Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development-Primary* (3rd ed.). Austin, TX: Pro-Ed.
Noël, M.-P. (2009). Counting on working memory when learning to count and to add: a preschool study. *Developmental Psychology, 45*(6), 1630-1643. doi:10.1037/a0016224
Oakhill, J., Yuill, N., & Parkin, A. (1986). On the nature of the difference between skilled and less-skilled comprehenders. *Journal of Research in Reading, 9*(2), 80-91. doi:10.1111/j.1467-9817.1986.tb00115.x
Ottem, E. J., Lian, A., & Karlsen, P. J. (2007). Reasons for the growth of traditional memory span across age. *European Journal of Cognitive Psychology, 19*(2), 233-270. doi:10.1080/09541440600684653
Pagano, M., & Gauvreau, K. (2000). *Principles of Biostatistics* (2nd Revised edition edition ed.). Pacific Grove, CA: Brooks/Cole.
Patterson, K., Graham, N., & Hodges, J. R. (1994). The impact of semantic memory loss on phonological representations. *Journal of Cognitive Neuroscience, 6*(1), 57-69. doi:10.1162/jocn.1994.6.1.57
Paul, R. (2007). *Language Disorders from Infancy Through Adolescence: Assessment & Intervention* (Third edition ed.). Saint Louis, MO: Mosby Elsevier.
Pickering, S. J., & Gathercole, S. E. (2001). *Working Memory Test Battery for Children*. London, UK: Harcourt Assessment.
Plante, E., & Vance, R. (1994). Selection of Preschool Language Tests A Data-Based Approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15-24. doi:10.1044/0161-1461.2501.15
Polišenská, K. (2011). *The influence of linguistic structure on memory span: Repetition tasks as a measure of language ability.* City University of London, London, UK. Retrieved from https://books.google.co.uk/books?id=sT4mAQAAIAAJ
Polisenska, K., Chiat, S., & Roy, P. (2015). Sentence repetition: what does the task measure? *International Journal of Language and Communication Disorders, 50*(1), 106-118. doi:10.1111/1460-6984.12126
Poll, G. H., Betz, S. K., & Miller, C. A. (2010). Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language, and Hearing Research, 53*(2), 414-429. doi:10.1044/1092-4388(2009/08-0016)
Pring, T. (2005). *Research Methods in Communication Disorders* (1 edition ed.). London, UK: John Wiley & Sons.
Prutting, C. A., & Connolly, J. E. (1976). Imitation: a closer look. *Journal of Speech and Hearing Disorders, 41*(3), 412-422.
Ramer, A. L. (1976). The function of imitation in child language. *Journal of Speech and Hearing Research, 19*(4), 700-717. doi:10.1044/jshr.1904.700
Ratner, N. B. (2000). Elicited Imitation and other methods for the analysis of trade-offs between speech and language skills in children. In L. Menn & N. B. Ratner (Eds.), *Methods for Studying Language Production* (pp. 291-311). Mahwah, NJ: Lawrence Erlbaum Associates.
Ravid, D., & Schiff, R. (2006). Roots and patterns in Hebrew language development: evidence from written morphological analogies. *Reading and Writing, 19*(8), 789-818. doi:10.1007/s11145-006-9004-3
Records, N. L., & Tomblin, J. B. (1994). Clinical decision making: Describing the decision rules of practicing speech-language pathologists. *Journal of Speech and Hearing Research, 37*(1), 144-156.
Redmond, S. M. (2005). Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics & Phonetics, 19*(2), 109-127. doi:10.1080/02699200410001669870

Redmond, S. M., Thompson, H. L., & Goldstein, S. (2011). Psycholinguistic profiling differentiates specific language impairment from typical development and from attention-deficit/hyperactivity disorder. *Journal of Speech Language and Hearing Research, 54*(1), 99-117. doi:10.1044/1092-4388(2010/10-0010)

Rees, N. S. (1975). Imitation and language development: issues and clinical implications. *Journal of Speech and Hearing Disorders, 40*(3), 339-350.

Reichenbach, K., Bastian, L., Rohrbach, S., Gross, M., & Sarrar, L. (2016). Cognitive functions in preschool children with specific language impairment. *International Journal of Pediatric Otorhinolaryngology, 86*, 22-26. doi:10.1016/j.ijporl.2016.04.011

Reilly, S., Tomblin, B., Law, J., McKean, C., Mensah, F. K., Morgan, A., . . . Wake, M. (2014). Specific language impairment: a convenient label for whom? *International Journal of Language and Communication Disorders, 49*(4), 416-451. doi:10.1111/1460-6984.12102

Renfrew, C. (1997). *Action Picture Test*. New York, NY: Harcourt Assessment.

Reynell, J., & Huntley, M. (1987). *Reynell Developmental Language Scales: Cantonese Version*. Windsor, UK: NEFR-Nelson.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*(6), 1239-1257.

Rice, M. L., & Wexler, K. (2001). *Rice/Wexler Test of Early Grammatical Impairment*. San Antonio, TX: Psychological Corporation.

Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: an investigation of complex syntax. *International Journal of Language and Communication Disorders, 45*(1), 47-60. doi:10.3109/13682820802647676

Riches, N. G., (2010). Sentence repetition in children with specific language impairment: An investigation of underlying mechanisms. *International Journal of Language and Communication Disorders*, 47(5):499-510. doi: 10.1111/j.1460-6984.2012.00158.x

Roid, G. H., & Miller, L. J. (1995). *Leiter International Performance Scale—Revised*. Wood Dale, IL: Stoelting Co.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale—Revised*. Wood Dale, IL: Stoelting Co.

Roodenrys, S., Hulme, C., & Brown, G. (1993). The development of short-term memory span: separable effects of speech rate and long-term memory. *Journal of Experimental Child Psychology, 56*(3), 431-442. doi:10.1006/jecp.1993.1043

Roodenrys, S., & Stokes, J. (2001). Serial recall and nonword repetition in reading disabled children. *Reading and Writing, 14*(5-6), 379-394. doi:Doi 10.1023/A:1011123406884

Roy, P., & Chiat, S. (2004). A Prosodically Controlled Word and Nonword Repetition Task for 2- to 4-Year-OldsEvidence From Typically Developing Children. *Journal of Speech, Language, and Hearing Research, 47*(1), 223-234. doi:10.1044/1092-4388(2004/019)

Roy, P., & Chiat, S. (2013). Language and socioeconomic disadvantage: Teasing apart delay and deprivation from disorder. In C. R. Marshall (Ed.), *Current Issues in Developmental Disorders* (pp. 125-150). Hove, UK: Psychology Press.

Royle, P., & Elin Thordardottir, E. (2003, 2003). *Le grand déménagement [French adaptation of the Recalling Sentences in Context subtest of the CELF–P]*. Unpublished research tool. McGill University, Montreal, Quebec, Canada.

Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire*. Los Angeles, CA: Western Psychological Services.

Sackett, D. L., & Haynes, R. B. (2002). The architecture of diagnostic research. *BMJ, 324*(7336), 539-541.

Sahlén, B., Reuterskiöld-Wagner, C., Nettelbladt, U., & Radeborg, K. (1999). Non-word repetition in children with language impairment--pitfalls and possibilities. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists, 34*(3), 337-352.

Schiavetti, N., & Metz, D. E. (2006). *Evaluating Research in Communicative Disorders*. Boston, MA: Allyn & Bacon.

Schneider, P., Dubé, R., & Hayward, D. (2002). *The Edmonton Narrative Norms Instrument*: Retrieved from University of Alberta Faculty of Rehabilitation Medicine: http://www.rehabmed.ualberta.ca/ /spa /enni.

Schneider, W., Gruber, H., Gold, A., & Opwis, K. (1993). Chess expertise and memory for chess positions in children and adults. *Journal of Experimental Child Psychology, 56*(3), 328-349.

Schopler, E., Reichler, R. J., DeVellis, R. F., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of Autism and Developmental Disorders, 10*(1), 91-103.

Seeff-Gabriel, B. (2006). *An investigation of sentence-level abilities in children with different types of speech disorder.* (Ph.D.), University of London. Retrieved from http://discovery.ucl.ac.uk/1445053/

Seeff-Gabriel, B., Chiat, S., & Dodd, B. (2010). Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language and Communication Disorders, 45*(6), 691-702. doi:10.3109/13682820903509432

Seeff-Gabriel, B., Chiat, S., & Roy, P. (2008). *Early Repetition Battery*. London, UK: Pearson Assessment.

Semel, E., Wiig, E., & Secord, W. (1994). *Clinical Evaluation of Language Fundamentals-Revised*. San Antonio, TX: Psychological Corporation.

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals* (Fourth ed.). San Antonio, TX: Psychological Corporation.

Semel, E., Wiig, E., & Secord, W. (2004). *Clinical Evaluation of Language Fundamentals—Fourth Edition Screening Test*. San Antonio, TX: Harcourt Assessment.

Semel, E., Wiig, E., & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals* (4th Edition ed.). San Antonio, TX: The Psychological Corporation.

Shaalan, S. (2009). Considerations for Developing and Adapting Language and Literacy Assessments in Arabic-Speaking Countries. In E. L. Grigorenko (Ed.), *Multicultural Psychoeducational Assessment* (pp. 287-314). New York, NY: Springer Publishing Company.

Shaalan, S. (2010). *Investigating grammatical complexity in Gulf Arabic speaking children with specific language impairment (SLI).* (Ph.D.), University College London, London, UK. Retrieved from http://eprints.ucl.ac.uk/20472/

Shakespeare, W. (trans. 2000). *Hamlet* (Reprinted edition edition ed. Vol. Act 1). New York: Dover Publications.

Shriberg, L. D., & Kwiatkowski, J. (1994). Developmental phonological disorders. I: A clinical profile. *Journal of Speech and Hearing Research, 37*(5), 1100-1126.

Simkin, Z., & Conti-Ramsden, G. (2001). Non-word repetition and grammatical morphology: normative data for children in their final year of primary school. *International Journal of Language and Communication Disorders, 36*(3), 395-404. doi:10.1080/13682820110045856

Smolik, F., & Vavru, P. (2014). Sentence imitation as a marker of SLI in Czech: disproportionate impairment of verbs and clitics. *Journal of Speech, Language, and Hearing Research, 57*(3), 837-849. doi:10.1044/2014_JSLHR-L-12-0384

Snowling, M. J. (2014). SLI—not just a semantic issue. Commentary on Reilly, S., Tomblin, B., Law, J., McKean, C., Mensah, F. K., Morgan, A., Goldfeld, S., Nicholson, J. M. and Wake, M., 2014, Specific language impairment: a convenient label for whom? *International Journal of Language and Communication Disorders, 49*(4), 416-451. doi:10.1111/1460-6984.12102

Stokes, S. F., & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 50*(4), 498-505. doi:10.1111/j.1469-7610.2008.01991.x

Stokes, S. F., & So, L. K. H. (1997). Classifier use by language-disordered and age-matched Cantonese-speaking children. *Asia Pacific Journal of Speech, Language and Hearing, 2*(2), 83-101. doi:10.1179/136132897805577413

Stokes, S. F., Wong, A. M. Y., Fletcher, P., & Leonard, L. B. (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech Language and Hearing Research, 49*(2), 219-236. doi:10.1044/1092-4388(2006/019)

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (Second edition ed.). Oxford, England: Oxford University Press.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. doi:10.5116/ijme.4dfb.8dfd

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. doi:10.1016/j.jcomdis.2012.08.002

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., . . . Chilingaryan, G. (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech Language and Hearing Research, 54*(2), 580-597. doi:10.1044/1092-4388(2010/09-0196)

Thordardottir, E., Keheyia, E., Lessard, N., Sutton, A., & Trudeau, N. (2010). Typical Performance on Tests of Language Knowledge and Language Processing of French-Speaking 5-Year-Olds. *Canadian Journal of Speech-Language Pathology and Audiology, 34*(1), 5-16.

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.

Topbaş, S., & Güven, O. S. (2009). *Sentence Repetition in Assessing SLI in Turkish Children [poster presentation]*. Paper presented at the American Speech and Hearing Association Annual Convention, New Orleans, LA. Poster retrieved from

Treiman, R. (1995). Errors in short-term memory for speech: a developmental study. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 21*(5), 1197-1208.

Tyler, L. K., Karmiloff-Smith, A., Voice, J. K., Stevens, T., Grant, J., Udwin, O., . . . Howlin, P. (1997). Do individuals with Williams syndrome have bizarre semantics? Evidence for lexical organization using an on-line task. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 33*(3), 515-527.

van der Lely, H. K., & Howard, D. (1993). Children with specific language impairment: linguistic impairment or short-term memory deficit? *Journal of Speech and Hearing Research, 36*(6), 1193-1207. doi:10.1044/jshr.3606.1193

Vicari, S., Carlesimo, G., Brizzolara, D., & Pezzini, G. (1996). Short-term memory in children with Williams syndrome: a reduced contribution of lexical--semantic knowledge to word span. *Neuropsychologia, 34*(9), 919-925. doi:10.1016/0028-3932(96)00007-3

Wallan, A. (2006). *Arabic sentence repetition: an investigation of performance of typically developing Saudi children on a sentence repetition task.* (M.Sc.), University College London, London. Available from ucl-primo.com

Wechsler, D. (1991). *The Wechsler Intelligence Scale for Children—Third Edition*. San Antonio, TX: Pycological Corporation.

Wechsler, D. (1992). *Wechsler Intelligence Scale for Children Third Editition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Preschool and Primary Scale of Intelligence™ - Third Edition (WPPSI-III UK)*. London, UK: Psychological Corporation.

Wechsler, D. (2005). *Wechsler Intelligence Scale for children* (4th edition ed.). Paris, France: Les Éditions du Centre de Psychologie Appliquée.

Werner, E., & Kresheck, J. D. (1981). Variability in Scores, Structures, and Errors on Three Measures of Expressive Language. *Language, Speech, and Hearing Services in Schools, 12*(2), 82-89. doi:10.1044/0161-1461.1202.82

Wiig, E., Secord, W., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals-Preschool*. San Antonio, TX: The Psychological Corporation.

Wiig, E., Secord, W., & Semel, E. (2000). *Clinical Evaluation of Language Fundamentals UK-Preschool UK Edition. The Psychological Corporation, Sidcup, UK* (UK edition ed.). Sidcup, UK: Psychological Corporation.

Williams, K. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Willis, C. S., & Gathercole, S. E. (2001). Phonological short-term memory contributions to sentence processing in young children. *Memory, 9*(4-6), 349-363. doi:10.1080/09658210143000155

Wilsenach, A. C. (2006). Syntactic processing in developmental dyslexia and in Specific Language Impairment : a study on the acquisition of the past participle construction in Dutch. *LOT, 128*.

Zimmerman, I. L., Pond, R. E., & Steiner, V. G. (2009). *Preschool Language Scale* (Fourth Edition UK ed.). London, UK: Psychological Corporation.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1992). *Preschool Language Scale* (Third ed.). San Antonio, TX: Psychological Corporation.

Zimmerman, I. L., Steiner, V. G., Pond, R. E., Boucher, J., & Lewis, V. (1997). *Preschool Language Scale-3 (UK)*. London, UK: Psychological Corporation.

APPENDIX A: INVITATION LETTERS, CONSENT, QUESTIONNAIRE

A.1. Invitation Letter to Heads of Nursery

City University London

Invitation for your nursery to participate in a research study

Project title: An investigation of sentence repetition as a measure of linguistic processing in both typically developing and language impaired Saudi children

Investigators: Shula Chiat, Penny Roy & Ashwag Wallan
Department of Language and Communication Science
City University, Northampton Square, London EC1V 0HB
Telephone: 020-7040-8238 E-mail: [redacted]

Secretary of Ethics Committee: Naomi Hammond
Academic Registrar's Office, City University
Telephone: 020-7040-8106
E-mail: [redacted]

Dear Nursery Manager

I am a doctoral student in Language and Communication Science at City University. As part of my studies, I am carrying out a research project in which I am investigating the use of an Arabic sentence repetition task as a language assessment.

The greatest difficulty facing clinicians in Saudi is the lack of informative assessment tools for language impairment in Arabic. Sentence repetition has been found to be helpful in identifying children with language impairment in both English and Cantonese speaking children; my study aims to investigate if this is also true for Arabic. Furthermore, my study aims to investigate the underlying skills involved in sentence repetition by comparing the children's performance on the sentence repetition task with a short term memory test and nonverbal IQ tasks. I am writing to ask if you would consider participating in this research.

In order to carry out my study, I hope to see approximately 140 typically developing children, and 14 children whose parents or teachers express concern about their language development and are aged 2;6-5;11years. The session will last about an hour. During the session a short term memory test (consisting of lists of numbers, words and made-up words) and a sentence repetition task will be presented, the children will be asked to repeat what they hear. In addition, they will be asked to perform two nonverbal IQ tasks involving blocks and puzzles. These tasks will be carried out at the child's pace, taking breaks as appropriate and spreading the tasks over two sessions if necessary. The sessions will be audio recorded so that I can write down the child's responses after the session. The audio recording will be destroyed at the end of the study (July, 2012).

Only children whose parents agree to their participation and who are themselves willing to participate would be included. Also, sessions in the nursery will be planned around regular activities such as circle time, play and meal breaks without interrupting the child's classes.

If you have any further concerns or questions, please do not hesitate to contact my supervisor, Professor Shula Chiat, contact details above.

Many thanks for giving this your consideration.

Yours sincerely,
Ashwag Wallan

Doctoral student in Language and Communication Science at City University

A.2. Invitation Letter To Parents

City University London

PARENT INFORMATION SHEET

Project title: An investigation of sentence repetition as a measure of linguistic processing in both typically developing and language impaired Saudi children

Investigators: Professor Shula Chiat, Dr. Penny Roy and Miss Ashwag Al-Wallan
Department of Language and Communication Science
City University, Northampton Square
London, EC1V 0HB
Telephone: 020-7040-8238

I am a doctoral student in Language and Communication Science at City University. As part of my studies, I am carrying out a research project in which I am investigating the use of an Arabic sentence repetition task as a language assessment.

The greatest difficulty facing clinicians in Saudi is the lack of informative assessment tools for language impairment in Arabic. Sentence repetition has been found to be helpful in identifying children with language impairment in both English and Cantonese speaking children; my study aims to investigate if this is also true for Arabic. Furthermore, my study aims to investigate the underlying skills involved in sentence repetition by comparing the children's performance on the sentence repetition task with a short term memory test.

In order to carry out my study, I hope to see approximately 140 typically developing children, and 14 children whose parents or teachers expressed concerns about their language development and are aged 2;6-5;11 years. The testing session will last about an hour. During the session a short term memory test (consisting of lists of numbers, words and made-up words) and a sentence repetition task will be presented and children will be asked to repeat what they hear. In addition, they will be asked to perform two Nonverbal IQ tasks involving blocks and puzzles. These tasks will be carried out at the child's pace, taking breaks as appropriate and spreading the tasks over two sessions if necessary. The sessions will be audio recorded so that I can write down the child's responses after the session. The audio recording will be destroyed at the end of the study (July, 2012).

The results of each child's assessments will be anonymous, identified by number only. However, I will be happy to give your child's results to or to your child's teacher or therapist should you so wish. You will also be welcome to read the final report of the study. There is no direct benefit to your child but we hope that findings of this study will help assess children who have language impairment.

If you are willing for your child to participate in this study, I would be grateful if you could fill in the attached consent form and questioner. Please return it to your child's teacher. Your child does not have to participate in this study and you may withdraw them at any time even after approval.

If you have any further concerns or questions, please do not hesitate to contact my supervisor, Professor Shula Chiat, contact details above.

Many thanks for giving this your consideration.

Yours sincerely,
Ashwag Al-Wallan
Doctoral student in Language and Communication Science at City University

If there is an aspect of the study which concerns you, you may make a complaint by contacting the Secretary to the Research Ethics Committee by:
Phone: 004420 7040 8106.
Address: Dr Naomi Hammond, Secretary to Senate Ethical Committee, Academic Development and Services, City University, Northampton Square, London, EC1V 0HB
Email[redacted]

A.3. Consent Form

Informed Consent Form

Title of Project:	An investigation of sentence repetition as a measure of linguistic processing in both typically developing and language impaired Saudi children
Investigators:	Professor Shula Chiat, Dr. Penny Roy and Miss Ashwag Al-Wallan

	YES	NO
Have you read the Parent Information Sheet?		
Have you had the opportunity to ask questions and discuss the study?		
Have you received satisfactory answers to all your questions?		
Have you received enough information about the study?		
Do you agree to your child participating in this study?		
Do you give permission to audio record the session with your child and keep the recording until the end of the study (November, 2011)?		
Do you understand that you are free to withdraw your child from the study without penalty at any stage?		
Do you agree with the publication of the results of this study in an appropriate outlet/s?		
Do you give permission for any assessments of your child to be made available to your child's teacher or speech and language therapist?		

Participant's Name: ..... (please print)

Participant's Age:.....

Parent's/Guardian's Name .....

Your relationship to participant: .....

Signature of Parent/Guardian: .....Date:.....

A.4. Parental Questionnaire

[1] Child's order in the family..... Number of Siblings.....

[2] What language do you use to communicate with your child at home? *(please circle one)*

a. Arabic only
b. Arabic and other languages

If answer is (b) please specify the other languages used

[3] Does your child have a medical or neurological diagnosis? *(please circle one)*

Yes No

If yes, please specify:

[4] Mother's education level .....

Father's education level.....

APPENDIX B: PARENT DEMOGRAPHICS

B.1. Saudi Population (15 Years and Over) by Marital Status, Gender, and Educational Status (Riyadh Administrative region); adapted from General Authority for Statistics of Kingdom of Saudi Arabia (2010)

Gender	Marital Status	Educational Status									
		Illiterate	Read & Write	Primary	Intermediate	Secondary / Equiv.	Dip. LT University	University	Master / High Dip.	Ph. D.	Total
Male		51884	56398	169228	329767	520527	125378	265109	20033	10790	1549114
	Never Married	6846	19572	74067	222199	250414	43459	35611	1632	286	654086
	Married	42541	35672	92766	104828	266680	80971	227019	18070	7534	876081
	Divorced	875	634	1987	2459	3175	860	2270	294	2946	15500
	Widowed	1622	520	408	281	258	88	209	37	24	3447
Female		154137	133131	201080	317615	398074	40749	191488	5468	2078	1443820
	Never Married	6308	12084	60957	173320	184362	5451	32872	766	247	476367
	Married	83134	86177	107436	122099	205370	33280	152589	4419	1642	796146
	Divorced	18725	8589	11669	15516	5345	1300	4709	230	128	66211
	Widowed	45970	26281	21018	6680	2997	718	1318	53	61	105096
Total		206021	189529	370308	647382	918601	166127	456597	25501	12868	2992934
	Never Married	13154	31656	135024	395519	434776	48910	68483	2398	533	1130453
	Married	125675	121849	200202	226927	472050	114251	379608	22489	9176	1672227
	Divorced	19600	9223	13656	17975	8520	2160	6979	524	3074	81711
	Widowed	47592	26801	21426	6961	3255	806	1527	90	85	108543

APPENDIX C: DEVELOPMENT STAGE AND PILOT RESULTS

C.1. Development Stage Results

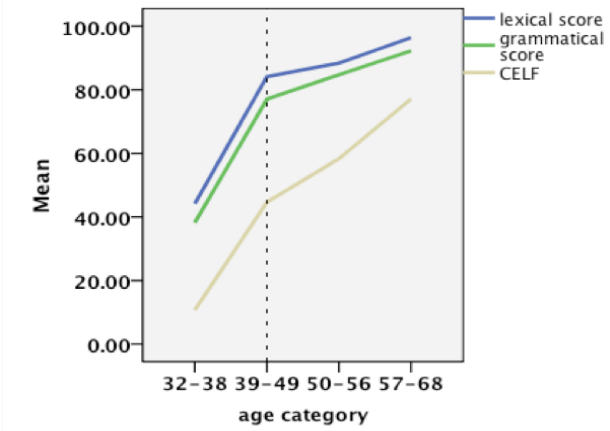


Figure B1. *Sentence Repetition Scores as a Function of Age: Development Stage*

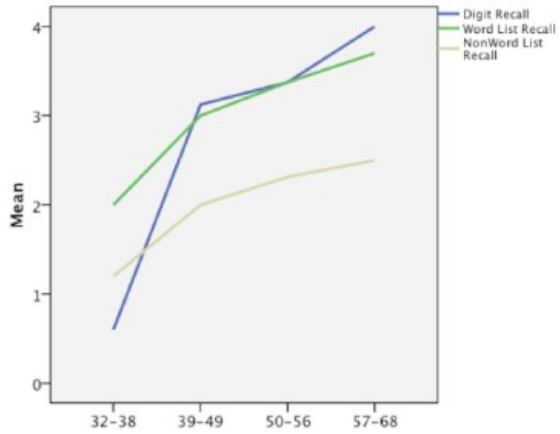


Figure B2. *Verbal Short Term Memory Subtest Span Scores as a Function of Age: Development Stage*

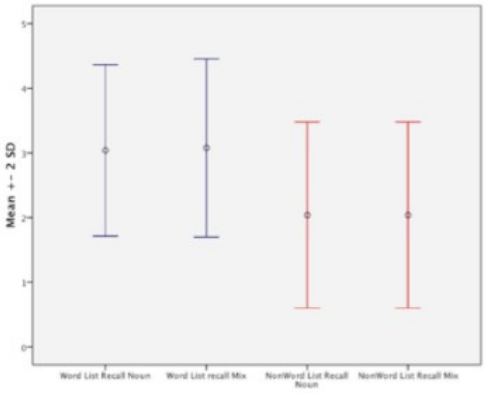


Figure B3. *Mean Span Scores of participants on STM subtests: Word List Recall Noun, Word List Recall Mixed, Nonword List Recall Noun, Nonword List Recall Mixed: Development Stage*

C.2. Pilot Results

Table B1. Sentence Repetition Test: Lexical Morpheme Scores, Grammatical Morpheme Scores, Total Sentence Accuracy Scores (CELF) and Age: Pilot Stage

SR Score	Age Category						p value
	2;6-3;5		3;6-4;5		4;6-5;11		
	M	*SD*	*M*	*SD*	*M*	*SD*	
Lexical	55.18	21.78	68.39	26.12	93.57	7.91	.001***
Grammatical	41.54	25.79	56.92	29.82	91.34	7.60	<.001***
Total Sentence	5.40	7.62	13.00	13.56	31.60	7.73	<.001***

Note. Parametric data reported and supported by nonparametric data.

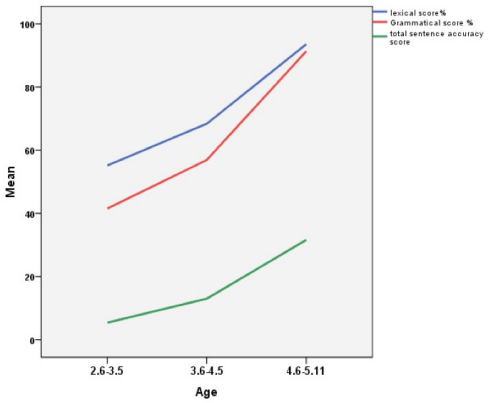


Figure B4. *Sentence Repetition Scores as a function of age: Pilot Stage*

Table B2. Verbal Short Term Memory Subtest Span Scores and Age: Pilot Stage

STM Subtest Span	Age Category						p value
	2;6-3;5		3;6-4;5		4;6-5;11		
	M	*SD*	*M*	*SD*	*M*	*SD*	
Word	2.5	.53	2.9	.57	3.7	.68	<.001***
Digit	2.4	.52	3.2	.37	4.2	.79	<.001***
Nonword	1.7	.48	2.1	.32	2.2	.42	.028*

Note. Parametric data reported and supported by nonparametric data.

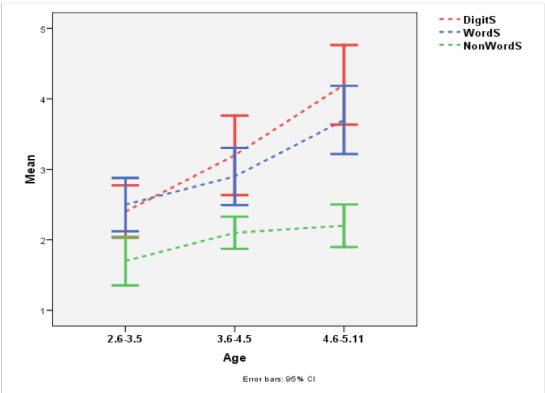


Figure B5. Error bars showing pilot stage participant's VSTM subtest span scores as a function of age

C.2.1. Correlation reseults between VSTM subtests: Pilot Stage

Word & Digit r(27) = .73, $p < .001$
Word & Nonword r(27) = .43, $p < .001$
Nonword & Digit r(27) = .44, p =.018

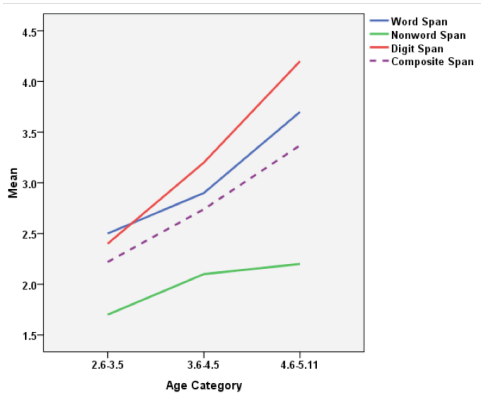


Figure B6. *Verbal Short Term Memory Subtest Span Scores and Composite Span Score: Pilot Stage*

C.2.2. Correlation results betweenVSTM Total Span score and Total Sentence Accuracy Score (CELF): Pilot Stage

Total Sentence Accuracy (CELF) & Verbal Short Term Memory composite span
$r(27) = .61, p < .001$

APPENDIX D: DEVELOPMENT STAGE VERBAL SHORT TERM MEMORY WORD RECALL SUBTEST

D.1. Word List Recall-Noun

Word List Recall-Noun

Name: Test day:

Span: Errors:

	Practice List			Score (1 or 0)
P1	ماما			
	/ma.ma:/			
	mom			
P2	صديق	مكان		
	/ʂi.di:g/	/ma.ka:n/		
	friend	place		
P3	رحله	بابا	مرمى	
	/riH.la/	/ba.ba:/	/mar.ma:/	
	trip	dad	goalpost	

Span	List	Score (1 or 0)
1	ولد	
	/wa.lad/	
	boy	
	كتاب	
	/ki.ta:b/	
	book	
	نحلة	
	/naH.la/	
	bee	
	سماء	
	/sa.ma:/	
	sky	
	لبن	
	/la.ban/	
	buttermilk	
	مفتاح	
	/mif.ta:H/	
	key	

Span	List		Score (1 or 0)
2	لمبة	خروف	
	/lam.ba/	/Xa.ru:f/	
	light	sheep	
	دواء	جبل	
	/da.wa:/	/ʤa.bal/	
	medicine	mountain	
	ليمون	أسد	
	/laj.mu:n/	/ʔa.sad/	
	lemon	lion	
	ملك	وادي	
	/ma.lik/	/wa:di:/	
	king	valley	
	كرة	نجمة	
	/ku:.ra:/	/naʤ.ma/	
	ball	star	
	دودة	حبل	
	/du:.da/	/Ha.bil/	
	worm	rope	

Span	List			Score (1 or 0)
3	لسان	نبتة	منديل	
	/li.sa:n/	/nab.ta/	/man.di:l/	
	tongue	plant	tissue	
	درج	هلال	مكتب	
	/da.raʤ/	/hi.la:l/	/mak.tab/	
	stairs	crescent	office	
	جناح	قلم	لوحة	
	/ʤa.na:H/	/ga.lam/	/lu:Ha/	
	wing	pen	sign	
	ركبة	حليب	أرنب	
	/ruk.ba/	/Ha.li:b/	/ʔar.nab/	
	knee	milk	rabbit	
	جمل	دفتر	خاتم	
	/ʤa.mal/	/daf.tar/	/Xa:.tam/	
	camel	notebook	ring	
	قدر	حفلة	كوكب	
	/gi.dir/	/Haf.la/	/kaw.kab/	
	pot	party	planet	

Span	List				Score (1 or 0)
4	نمر	لبنة	صابون	مجلس	
	/ni.mir/	/lab.na/	/ṣa:.bu:n/	/madʒ.lis/	
	tiger	yogurt /labneh	soap	living room	
	بطن	تمساح	أمير	قهوة	
	/ba.ṭin/	/tim.sa:H/	/ʔa.mi:r/	/gah.wah/	
	stomach	alligator	prince	coffee	
	قمر	خيمة	شوكة	جسم	
	/ga.mar/	/Xe:.ma/	/ʃo:.ka/	/dʒi.sim/	
	moon	tent	fork	body	
	خشم	مسمار	فوطة	نملة	
	/xa.ʃim/	/mis.ma:r/	/fu:.ṭah/	/nam.la/	
	nose	nail	towel	ant	
	جرس	عسل	سرير	معجون	
	/dʒa.ras/	/ʕa.sal/	/si.ri:r/	/maʕ.dʒu:n/	
	bell	honey	bed	toothpaste	
	طالب	محل	شاهي	لعبة	
	/ṭa:.lib/	/ma.Hal/	/ʃa:.hi:/	/li?.ba/	
	student	store	tea	toy	

Span	List					Score (1 or 0)
5	كرسي	شمعة	إذن	مسبح	جدار	
	/kur.si:/	/ʃam.ʕa/	/ʔi.ðin/	/mas.baH/	/dʒi.da:r/	
	chair	candle	ear	pool	wall	
	مقص	صحن	غزال	ساعة	رمل	
	/ma.gaṣ/	/ṣa.Han/	/ɣa.za:l/	/sa:.ʕa/	/ra.mil/	
	scissor	plate	gazelle	watch	sand	
	وردة	مسجد	قدر	خشب	فستان	
	/war.da/	/mas.dʒid/	/gi.dir/	/Xa: ʃab/	/fis.ta:n/	
	flower	mosque	pot	wood	dress	
	سيكل	فرشة	موزة	دكتور	ثعلب	
	/saj.kal/	/fir.ʃa/	/mo:.za:/	/dik.tu:r/	/ɵaʔ.lab/	
	bike	brush	banana	doctor	wolf	
	حديد	جزمة	قطار	إصبع	غابة	
	/Ha.di:d/	/dʒaz.ma:/	/gi.ṭa:r/	/ʔiṣ.baʕ/	/ɣa:.ba/	
	steel	shoe	train	finger	forest	
	بحر	صاحب	نخلة	عصفور	طاولة	
	/ba.Har/	/ṣa:Hib/	/naX.la/	/ʕaṣ.fu:r/	/ṭa:w.la:/	
	sea	friend	palm tree	bird	table	

D.2. Word List Recall-Mixed

Word List Recall-Mixed

Name: Test day:

Span: Errors:

	Practice List			Score (1 or 0)
P1	بابا			
	/ba.ba:/			
	dad			
P2	بدأ	حالي		
	/ba.da:/	/Ha.li:/		
	start	sweet		
P3	حامض	ماما	مكان	
	/Ha:.mið/	/ma:.ma/	/ma.ka:n/	
	sour	mom	place	

Span	List	Score (1 or 0)
1	أكل	
	/ʔa.kal/	
	ate	
	مفتاح	
	/mif.ta:H/	
	key	
	قوي	
	/gu.wi:/	
	strong	
	نملة	
	/nam.la/	
	ant	
	كامل	
	/ka:.mil/	
	whole	
	درس	
	/da.ras/	
	studied	

Span	List		Score (1 or 0)
2	وردة	سبح	
	/war.da/	/sa.baH/	
	flower	swim	
	خاتم	مالح	
	/Xa:.tam/	/ma:liH/	
	ring	salty	
	كتب	سماء	
	/ka.tab/	/sa.ma:/	
	wrote	sky	
	خائف	لبنة	
	/Xa:.jif/	/lab.na/	
	scared	yogurt /labneh	
	طاولة	سمع	
	/ṭa:w.la/	/sa.maʕ/	
	table	hear	
	مسك	جائع	
	/ma.sak/	/ʤa:.jiʕ/	
	held	hungry	

Span	List			Score (1 or 0)
3	ساعة	ثقيل	كبر	
	/sa:.ʔa/	/θi.gi:l/	/ka.bar/	
	watch	heavy	grew	
	حلو	وقف	غابة	
	/Hi.lu/	/wa.gaf/	/ɣa:.ba/	
	pretty	stood	forest	
	ليس	خيمة	بارد	
	/la.bas/	/Xe:.ma/	/ba:rid/	
	wore	tent	cold	
	قطار	غني	شرب	
	/gi.ṭa:r/	/ɣa.ni/	/ʃa.rab/	
	train	rich	drank	
	أحمر	ساعد	لبن	
	/ʔaH.mar/	/sa:.ʔad/	/la.ban/	
	red	helped	buttermilk	
	فتح	نحلة	طويل	
	/fa.taH/	/naH.la/	/ṭu.wi:l/	
	opened	bee	tall	

Span	List					Score (1 or 0)
5	ليمون	فرح	كثير	قلم	مشى	
	/laj.mu:n/	/fa.raH/	/ki.ei:r/	/ga.lam/	/ma.ʃa:/	
	lemon	became happy	a lot	pen	walked	
	رمى	كوكب	واسع	حفلة	سمين	
	/ra.ma:/	/kaw.kab/	/wa:.siʕ/	/Haf.la/	/si.mi:n/	
	threw	planet	wide	party	fat	
	مكتب	خشن	فقير	رسم	شارع	
	/mak.tab/	/Xi.ʃin/	/fa.qi:r/	/ra.sam/	/ʃa:.riʕ/	
	office	rough	poor	drew	road	
	أبيض	خرج	نظيف	بنى	فوطة	
	/ʔab. jað/	/xa.raʤ/	/ni.ð̞i:f/	/ba.na:/	/fu:.ṭa/	
	white	exited	clean	built	towel	
	وقف	شجاع	موزة	جناح	أخذ	
	/wa.gaf/	/ʃu.ʤa:ʕ/	/mo:.za:/	/ʤa.na:H/	/ʔa..Xað/	
	stood	brave	banana	wing	took	
	حكى	غالي	نادى	كريم	لسان	
	/Ha.ka:/	/ɣa.li:/	/na:.da:/	/ka.ri:m/	/li.sa:n/	
	told	expensive	called out	giving	tongue	

Span	List				Score (1 or 0)
4	لمبة	أصفر	بكى	جديد	
	/lam.ba/	/ʔaṣ.far/	/ba.ka:/	/ʤi.di:d/	
	light	yellow	cried	new	
	جلس	دودة	هلال	قرأ	
	/ʤa.las/	/du:.da/	/hi.la:l/	/ga.ra:/	
	sat	worm	crescent	read	
	بعيد	مسح	شرى	ناعم	
	/bi.ʕi:d/	/ma.saH/	/ʃa.ra:/	/na:.ʕim/	
	far	wiped	bought	soft	
	ذكي	منديل	أسود	ضحك	
	/ð̞a.ki:/	/man.di:l/	/ʔas.wad/	/ð̞a.Hak/	
	smart	tissue	black	laughed	
	وادي	سأل	لوحة	كبير	
	/wa:di:/	/sa.ʔal/	/lo:Ha/	/ki.bi:r/	
	valley	asked	sign/painting	big	
	جرى	نبتة	كتاب	قليل	
	/ʤa.ra:/	/nab.ta/	/ki.ta:b/	/gi.li:l/	
	ran	plant	book	few/ little bit	

APPENDIX E: VERBAL SHORT TERM MEMORY SUBTESTS

E.1. Digit Recall

Name: Test day:

Span: Errors:

	Practice List	Response	Score (1 or 0)
P1	5		
P2	3 4		
P3	5 2 8		

	List	Response	Score (1 or 0)		List	Response	Score (1 or 0)
1	4			4	2 8 1 4		
	6				6 2 8 4		
	2				9 6 2 4		
	7				8 1 6 2		
	1				6 3 5 9		
	9				5 3 8 2		
2	2 8			5	8 1 3 9 2		
	5 3				3 5 8 2 6		
	4 6				2 9 7 3 1		
	8 1				4 6 3 1 9		
	9 2				5 8 1 3 6		
	1 3				7 1 3 6 2		
3	8 1 3			6	5 2 1 7 9 3		
	6 3 7				2 8 6 3 7 1		
	2 6 8				4 6 3 7 1 9		
	1 8 2				6 2 9 7 3 1		
	7 1 9				3 5 8 2 6 9		
	4 6 2				1 9 5 8 2 4		

7	8 3 5 2 9 7 1		
	7 9 2 6 3 5 8		
	8 5 2 9 6 3 1		
	9 6 2 8 1 4 7		
	3 1 8 2 6 9 5		
	5 3 7 1 9 6 4		

E.2. Word List Recall

Name: Test day:

Span: Errors:

	Practice List			Score (1 or 0)
P1	ماما			
	/ma.ma:/			
	mom			
P2	بدأ	مكان		
	/ba.da:/	/ma.ka:n/		
	start	place		
P3	رحله	بابا	مرمى	
	/riH.la/	/ba.ba:/	/mar.ma:/	
	trip	dad	goalpost	

Span	List	Score (1 or 0)
1	ولد	
	/wa.lad/	
	boy	
	قريب	
	/gi.ri:b/	
	near/close	
	اكل	
	/?a.kal/	
	ate	
	سماء	
	/sa.ma:/	
	sky	
	حبل	
	/Ha.bil/	
	rope	
	مسبح	
	/mas.baH/	
	pool	

Span	List		Score (1 or 0)
2	لية	خروف	
	/lam.ba/	/Xa.ru:f/	
	light	sheep	
	نبتة	بارد	
	/nab.ta/	/ba:.rid/	
	plant	cold	
	حليب	مسك	
	/Ha.li:b/	/ma.sak/	
	milk	held	
	منديل	كبر	
	/man.di:l/	/ka.bar/	
	tissue	grew	
	نحلة	خاتم	
	/naH.la/	/Xa:.tam/	
	bee	ring	
	مليان	سرير	
	/mal.ja:n/	/si.ri:r/	
	full	bed	

Span	List			Score (1 or 0)
3	سمين	مفتاح	كتب	
	/si.mi:n/	/mif.ta:H/	/ka.tab/	
	fat	key	wrote	
	لبن	سريع	رمل	
	/la.ban/	/si.riʕ/	/ra.mil/	
	buttermilk	fast	sand	
	مسمار	ركب	عاقل	
	/mis.ma:r/	/ra.kab/	/ʕa:.gil//	
	nail	rode	polite/smart	
	نملة	مالح	دفتر	
	/nam.la/	/ma:.liH/	/daf.tar/	
	ant	salty	notebook	
	كتاب	ليمون	سمع	
	/ki.ta:b/	/laj.mu:n/	/sa.maʕ/	
	book	lemon	heard	
	دودة	بحر	قليل	
	/du:.da/	/ba.Har/	/gi.li:l/	
	worm	sea	few	

Span	List				Score (1 or 0)
4	بكى	أحمر	ملعب	لسان	
	/ba.ka:/	/ʔaH.mar/	/malʕab/	/li.sa:n/	
	cried	red	stadium/play area	tongue	
	سبح	وردة	أسد	قديم	
	/sa.baH/	/war.da/	/ʔa.sad/	/gi.di:m/	
	swim	flower	lion	old	
	حفله	بعيد	دواء	ليس	
	/Haf.la/	/bi.ʕi:d/	/da.wa:/	/la.baʂ/	
	party	far	medicine	wore	
	محل	خائف	رمى	فستان	
	/ma.Hal/	/Xa:.jif/	/ra.ma:/	/fis.ta:n/	
	store	scared	threw	dress	
	قهوة	أمير	ناعم	سأل	
	/gah.wah/	/ʔa.mi:r/	/na:.ʕim/	/sa.ʔal/	
	coffee	prince	soft	asked	
	تمساح	نحيف	دخل	كرسي	
	/tim.sa:H/	/ni.Hi:f/	/da.Xal/	/kur.si:/	
	alligator	thin	entered	chair	

Span	List					Score (1 or 0)
5	مكتب	ثقيل	ابره	وادي	رسم	
	/mak.tab/	/θi.gi:l/	/ʔib.ra/	/wa:di:/	/ra.sam/	
	office	heavy	needle	valley	drew	
	خيمة	مسح	أرنب	ساعة	كثير	
	/Xe:.ma/	/ma.saH/	/ʔar.nab/	/sa:.ʕa/	/ki. θi:r/	
	tent	erased	rabbit	watch	a lot	
	قوي	لوحة	كبير	نادى	عسل	
	/gu.wi:/	/lo:Ha/	/ki.bi:r/	/na:.da:/	/ʕa.sal/	
	strong	sign/painting	big	called out	honey	
	خفيف	مسكين	وقف	لعبة	دكتور	
	/Xa.fi:f/	/mis.ki:n/	/wa.gaf/	/li?.ba/	/dik.tu:r/	
	light	poor	stood	toy	doctor	
	رحله	نمر	ساعد	كوكب	خلق	
	/riH.la/	/ni.mir/	/sa:.ʕad/	/kaw.kab/	/Ha.lag/	
	trip	tiger	helped	planet	earring	
	ملك	ركبة	سيكل	فرحان	درس	
	/ma.lik/	/ruk.ba/	/saj.kal/	/far.Ha:n/	/da.ras/	
	king	knee	bike	happy	studied	

E.3. Nonword List Recall

Name: Test day:

Span: Errors:

	Practice List			Score (1 or 0)
P1	نانا			
	/na.na:/			
P2	بنل	حتير		
	/ba.nal/	/Ha.ti:r/		
P3	مكديل	فافا	بحف	
	/mik.di:l/	/fa.fa:/	/ba.Haf/	

Span	List	Score (1 or 0)
1	كويل	
	/ki.wi:l/	
	لد	
	/la.mad/	
	عالب	
	/ʕa:.lib/	
	حنل	
	/Ha.nal/	
	نارس	
	/na:.ris/	
	فاتم	
	/fa:.tam/	

Span	List		Score (1 or 0)
2	لبير	ومد	
	/li.bi:r/	/wa.mad/	
	بلف	ثائم	
	/ba.laf/	/θa:.tam	
	مالع	دافى	
	/ma:.liʕ/	/da:.fa:/	
	قامن	حكل	
	/fa:.min/	/Ha.kal/	
	مفى	لاتب	
	/ma.fa:/	/la:.nib/	
	دفل	بمار	
	/da.fal/	/bi.ma:r/	

Span	List			Score (1 or 0)
3	فمح	آكتب	خالم	
	/fa.maH/	/ʔak.nab/	/Xa:.lam	
	سفى	دوال	محكن	
	/sa.fa:/	/di.wa:l/	/miH.ki:n/	
	آنود	رايس	لك	
	/ʔan.wad/	/ra:.bis/	/la.mak/	
	ثكبه	ديفه	مالت	
	/θuk.ba/	/di:.fa/	/ma:.lit/	
	ناوح	بفله	دلاه	
	/na:.wiH/	/baf.la/	/da.la:/	
	حبد	كمه	يافل	
	/Ha.bid/	/ka.ma/	/ja:.fil/	

Span	List				Score (1 or 0)
4	موعه	حكن	فايخ	أبيم	
	/mo:ʕa/	/Ha.kin/	/fa:.jiX/	/ʔab.jam/	
	خويه	لحك	يارع	فبده	
	/Xo:.ba/	/la.Hak/	/ja:.riʕ/	/fib.da/	
	متل	ارن	دكيك	حابه	
	/ma.tal/	/ʔi.rin//	/da.ki:k/	/Ha:.ba/	
	لبع	خدى	يلن	كامه	
	/la.baʕ/	/Xa.da:/	/ba.lin/	/ka:.mih/	
	سافى	وامي	اكن	خاوب	
	/sa:.fa:/	/wa:.mi:/	/ʔi.kin/	/Xa:.wab/	
	تمواح	بسيد	لاثق	مند	
	/tim.wa:H/	/bi.si:d/	/la:.θig/	/mi.nid/	

APPENDIX F: SENTENCE REPETITION TESTS

1									
ʃaːf	tense	mHammad	ʔXwan	pat	-a	f-	il-	madras	-a
saw	pf	prop.n	brother	plural	his	in	the-	school	-fsg

2									
ʔil-	walad	saʔal	tense	ʂidiːg	-a	ʕan	ʔil-	Hafl	-a
the-	boy	ask	pf	friend	-his	about	the-	party	-fsg

3										
Haʈ	-at	tense	ʔil-	bint	daftar	-haː	ʕalaː	iʈ-	ʈaːwl	-a
put	-pf3fsg	pf	the-	girl	notebook	-her	on	the-	table	-fsg

4										
dʒuːd	ʃar	tense	-at	haðaː	pattern	il-	galam	min	ʔil-	maHal
prop-n	buy	pf	-pf3fsg	this	msg	the-	pen	from	the-	store

5										
dʒalas	tense	ʔil-	walad	ʔiʈ-	ʈwiːl	ʕala	haðaː	patter	il-	kursiː
sit	pf	the-	boy	the	tall	on	this	msg	the-	chair

6											
mahaː	kaːn	tense	-at	ti-	sbaH	tense	maʕ	Xal	-haː	il-	kibiːr
PropN	was	pf	-pf3fsg	imp3fsg-	swim	imp	with	brother	-her	the	big

7											
nuːraː	ʁasal	-at	tense	-haː	bi-	ʔil-	moːj	-a	w-	iʂ-	ʂabuːn
prop-n	wash	-pf3fsg	pf	-her	with	the-	water	-fsg	and-	the-	soap

8											
dʒar	-at	tense	ʔil-	biss	-a	is-	soːda	pattern	waraː	il-	fiːl
run	-pf3fsg	pf	the-	cat	-fsg	the-	black	fsg	after	the-	elephant

9												
ʔil-	walad	iʂ-	ʂiːniːr	ʔaXað	tense	haðiː	pattern	il-	kuːr	-a	minn	-a
the-	boy	the-	small	take	pf	this	fsg	the-	ball	-fsg	from	-him

10													
ji-	graː	tense	huː	pattern	kitaːb	ʕan	ʔid-	diːk	w-	in-	naml	-a
imp3msg-	read	imp	pro	msg	book	about	the-	hen	and	the-	ant	-fsg

11													
kaːn	tense	naːjif	ji-	rkiʂ	tense	f-	il-	Hadig	-a	maʕ	ʔaʂHaːb	pat	-a
was	pf	prop.n	imp 3sg-	run	imp	in-	the-	garden	-f	with	friend	plural	-his

12														
huː	pattern	ji-	Hib	tense	ji-	ʃrab	tense	ʔil-	Haliːb	bi-	il-	faraːwl	-a
he	msg	imp3 msg-	love	imp	imp3 msg-	drink	imp	the-	milk	with-	the-	straw-berry	fsg

13															
ti-	ʕiːʃ	tense	ʔiz-	zaraːf	-a	il-	Hilw	-a	fiː	haðiː	pattern	il-	naːb	-a
imp 3fsg-	live	imp	the-	girrafe	-fsg	the	pretty	-fsg	in	this	fsg	the-	forest	-fsg

14															
hij	pattern	ti-	Hib	tense	ti-	lʕab	tense	h-	ʕaruːs	-at	-haː	il-	dʒidiːd	-a
he	fsg	imp 3fsg-	love	imp	imp 3fsg-	play	imp	with-	doll	-fsg	-her	the-	new	fsg

	lexical		grammatical		CELF	
1	4		6		3	
2	4		6		3	
3	4		7		3	
4	4		7		3	
5	4		7		3	
6	4		8		3	
7	4		8		3	
8	4		8		3	
9	4		9		3	
10	4		9		3	
11	4		10		3	
12	4		10		3	
13	4		11		3	
14	4		11		3	
	56		117		42	

	Omission	Substitution	Refusal	Unintelligible	Perseveration
lexical					
grammatical					

APPENDIX G: ANONMALOUS SENTENCE REPETITION TEST

G.1. Typical sentences that were used to derive the Anomalous sentences and their score sheet.

2. The boy asked his friend about the party.									
ʔil-	walad	saʔal	tense	ʃidi:g	-a	ʕan	ʔil-	Hafl	-a
the-	boy	ask	pf	friend	-his	about	the-	party	-fsg

5. The tall boy sat on this chair.										
dʒalas	tense	ʔil-	walad	ʔiʈ-	ʈwi:l	ʕala:	haða:	pattern	il-	kursi:
sit	pf	the-	boy	the	tall	on	this	msg	the-	chair

3. The girl put her notebook on the table.										
Haʈ	-at	tense	ʔil-	bint	daftar	-ha:	ʕala:	iʈ-	ʈa:wl	-a
put	-pf3fsg	pf	the-	girl	notebook	-her	on	the-	table	-fsg

8. The black cat ran after the elephant.											
dʒar	-at	tense	ʔil-	biss	-a	is-	so:da	pattern	wara:	il-	fi:l
run	-pf3fsg	pf	the-	cat	-fsg	the-	black	fsg	after	the-	elephant

10. He reads a book about the hen and the ant.												
ji-	gra:	tense	hu:	pattern	kita:b	ʕan	ʔid-	di:k	w-	in-	naml	-a
imp3msg-	read	imp	he	msg	book	about	the-	hen	and	the-	ant	-fsg

12													
hu:	pattern	ji-	Hib	tense	ji-	ʃrab	tense	ʔil-	Hali:b	bi-	il-	fara:wl	-a
he	msg	imp3 msg-	love	imp	imp3 msg-	drink	imp	the-	milk	with-	the-	straw-berry	fsg

13 The pretty giraffe lives in the forest.														
ti-	ʕi:ʃ	tense	ʔiz-	zara:f	-a	il-	Hilw	-a	fi:	haði:	pattern	il-	ʁa:b	-a
imp 3fsg-	live	imp	the-	girrafe	-fsg	the	pretty	-fsg	in	this	fsg	the-	forest	-fsg

14														
hij	pattern	ti-	Hib	tense	ti-	ʕab	tense	b-	ʕaru:s	-at	-ha:	il-	dʒidi:d	-a
he	fsg	imp 3fsg-	love	imp	imp 3fsg-	play	imp	with-	doll	-fsg	-her	the-	new	-fsg

G.1.1. Summary Score Sheet for typical sentences

	lexical		grammatical		CELF	
2	*4*		*6*		*3*	
5	*4*		*7*		*3*	
3	*4*		*7*		*3*	
8	*4*		*8*		*3*	
10	*4*		*9*		*3*	
12	*4*		*10*		*3*	
13	*4*		*11*		*3*	
14	*4*		*11*		*3*	
	32		*69*		*24*	

	Omission	Substitution	Refusal	Unintelligible	Perseveration
lexical					
grammatical					

G.2. Semantically Anomalous sentences and their score sheets.

1. The notebook asked his giraffe about the boy									
ʔid-	daftar	saʔal	tense	zara:f	-t	-a	ʕan	ʔil-	walad
the-	notebook	ask	pf	girrafe	-fsg	-his	about	the-	boy

2. The pretty milk sat on this elephant										
dʒalas	tense	ʔil-	Hali:b	ʔil-	Hilu	ʕala:	haða:	pat	ʔil-	fi:l
sit	pf	the-	boy	the	tall	on	this	msg	the-	elephant

3. The table put her friend on the hen										
ʔit-	ʈa:wl	-a	Haʈ	-at	tense	ʂidi:g	-ha:	ʕala:	ʔid-	di:k
the-	table	-fsg	put	-pf3fsg	pf	friend	-her	on	the-	hen

4. the tall forest ran after the boy												
dʒar	-at	tense	ʔil-	ʁa:b	-a	ʔiʈ-	ʈwi:l	-a	wara:	il-	walad
run	-pf3fsg	pf	the-	forest	-fsg	the-	tall	fsg	after	the-	boy

5. He reads a chair about the boy and the doll												
ji-	gra:	tense	hu:	pat	kursi:	ʕan	ʔil-	walad	w-	il-	ʕaru:s	-a
imp3msg-	read	imp	he	m	chair	about	the-	boy	and	the-	doll	-fsg

6. He loves to drink the book with the ant													
hu:	pat	ji-	Hib	tense	ji-	ʃrab	tense	ʔil-	kita:b	b-	il-	naml	-a
he	m	imp3msg-	love	imp	imp3msg-	drink	imp	the-	book	with-	the-	ant	-fsg

7. The new strawberry lives in this cat.														
ti-	ʕi:ʃ	tense	ʔil-	fara:wl	-a	il-	dʒidi:d	-a	fi:	haði:	pat	il-	biss	-a
imp3fsg-	live	imp	the-	straw-berry	fsg	the-	new	-fsg	in	this	fsg	the-	cat	-fsg

8. She loves to play with her black party														
hij	pat	ti-	Hib	tense	ti-	ʁab	tense	b-	Hafl	-at	-ha:	ʔis-	so:da	pat
she	f	imp3fsg-	love	imp	imp3fsg-	play	imp	with-	part	-fsg	-her	the-	black	fsg

G.2.1. Summary Score Sheet for Semantically Anomalous sentences

	lexical		grammatical		CELF	
1	*4*		*6*		*3*	
2	*4*		*7*		*3*	
3	*4*		*7*		*3*	
4	*4*		*8*		*3*	
5	*4*		*9*		*3*	
6	*4*		*10*		*3*	
7	*4*		*11*		*3*	
8	*4*		*11*		*3*	
	32		*69*		*24*	

	Omission	Substitution	Refusal	Unintelligible	Perseveration
lexical					
grammatical					

G.3. Syntactically Anomalous sentences and their score sheets.

Note. Sentences are numbered according to their appearance in the ASR test.

9. The girl put her notebook the table on									
ʔil-	bint	Hat̪	tense	ʔid-	daftar	-ha:	t̪a:wl	-a	ʕala:
the-	girl	put	pf	the-	notebook	-her	table	-fsg	on

Violated Rules: (Items in red highlight the rule violations & items in green highlight the correct form)

- Subject-Verb gender agreement:
  ʔil-bint Hat-Ø the-girl put.pf-Ø
  ʔil-bint Hat-at the-girl put.pf-pf3fsg
- Determiner addition:
  ʔil-dftar-ha the-notebook-her
  dftar-ha notebook-her
- Preposition-Noun order:
  ʔit̪-t̪awl-a ʕala: the-table-fsg on
  ʕala: ʔit̪-t̪awl-a on the-table-fsg

10. The boy asked his friend the party about										
saʔal	-at	ʔil-	walad	tense	ʔiʒ-	s̪idi:g	-a:	Hafl	-a	ʕan
ask	-pf3fsg	the-	boy	pf	the-	friend	-his	party	fsg	about

Violated Rules:
- Subject-Verb gender agreement:
  ʔil-walad saʔal-at the-boy ask.pf-pf3fsg
  ʔil-walad saʔal the-boy ask.pf
- Determiner addition:
  ʔil-s̪idi:g-a: the-friend-his
  s̪idig-a: friend-his
- Preposition-Noun order:
  ʔil-Hafl-a ʕan the-party-fsg about
  ʕan ʔil-Hafl-a about the-party-fsg

11. ran the black the cat the elephant after										
ʤara:	tense	ʔil-	ʔaswad	pattern	ʔil-	biss	-a	ʔil-	fi:l	wara:
run	pf	the-	black	msg	the-	cat	-fsg	the-	elephant	after

Violated Rules:
- Subject-Verb gender agreement:
  ʤara:-Ø ʔil-biss-a run.pf-Ø the-cat-fsg
  ʤara:-at ʔil-biss-a run.pf-pf3fsg the-cat-fsg
- Noun-Adjective gender agreement: *(pattern)*
  ʔil-biss-a ʔil-ʔaswad cat-fsg black.msg
  ʔil-biss-a ʔis-so:da: cat-fsg black.fsg
- Preposition-Noun order:
  ʔil-fi:l wara: the-elephant after
  wara: il-fi:l after the-elephant

12. The boy tall sat on this chair.											
ʔiʒ-	t̪wi:l	ʔil-	walad	ʤalas	-at	tense	ʕala:	haði:	pat	ʔil-	kursi:
the	tall	the-	tall	sit	-pf3fsg	pf	on	this	fsg	the-	chair

Violated Rules:
- Subject-Verb gender agreement:
  ʔil-walad ʤalas-at the-boy sit.pf-pf3fsg
  ʔil-walad ʤalas the-boy sit.pf
- Demonstrative-Noun agreement:
  haði: ʔil-kursi: this.fsg chair.m
  haða: ʔil-kursi: this.msg chair.m
- Noun-Adjective order:
  ʔil-walad ʤalas ʕala it̪-t̪wi:l the-boy sit.pf on the-tall
  ʔil-walad it̪-t̪wi:l ʤalas ʕala the-boy the-tall sit.pf on

13. She reads a book and the hen the ant about.												
ji-	gra:	tense	hiy	pat	kita:b	w-	id-	di:k	ʕin-	naml	-a	ʕan
imp3msg-	read	imp	she	f	book	and	the-	hen	the-	ant	-fsg	about

Violated Rules:
- Subject-Verb gender agreement:
  hiy ji-gra: she imp3msg-read.imp
  hu: ji-gra: he imp3msg-read.imp
- Conjunction-Noun order:
  w-ʔid-di:k ʔin-naml-a and-the-hen the-ant-fsg
  ʔid-di:k w-ʔin-naml-a the-hen and-the-ant-fsg
- Preposition-Noun order:
  ʔid-di:k w-ʔin-naml-a ʕan about the-hen and-the-ant-fsg
  ʕan ʔid-di:k w-ʔin-naml-a about the-hen and-the-ant.fsg

14. He drinks loves with the milk the strawberry													
hu:	pat	ti-	ʃrab	tense	ji-	Hib	tense	b-	il-	Hali:b	ʔil-	fara:wl	-a
he	m	imp3sg-	drink	imp	imp3msg-	love	imp	with-	the-	milk	the-	straw-berry	-fsg

Violated Rules:
- Subject-Verb gender agreement:
  hu: ti-ʃrab he imp3sg-drink.imp
  hu: ji-ʃrab he imp3msg-drink.imp
- Verb order:
  imp3msg-drink.imp imp3msg-love.imp
  imp3msg-love.imp imp3msg-drink.imp
- Preposition-Noun order:
  b-ʔil-Hali:b ʔil-fara:wl-a with-the-milk the-strawberry-fsg
  ʔil-Hali:b b-il-fara:wl-a the-milk with-the-strawberry-fsg

15. The pretty giraffe lives in this forest.														
ji-	ʕi:f	tense	ʔil-	Hilw	-a	iʒ-	ʒara:f	-a	fi:	haða:	pat	il-	ʁa:b	-a
imp3msg-	live	imp	the	pretty	-fsg	the-	girrafe	fsg	in	this	msg	the-	forest	-fsg

Violated Rules:
- Subject-Verb gender agreement:
  ji-ʕi:f ʔiz-zara:f-a imp3msg-live.imp the-giraffe-fsg
  ti-ʕi:f ʔiz-zara:f-a imp3fsg-live.imp the-giraffe-fsg
- Demonstrative-Noun agreement:
  ha:ða: il-ʁa:b-a this.msg the-forest-fsg
  ha:ði: il-ʁa:b-a this.fsg the-forest-fsg
- Noun-Adjective order:
  ʔil-Hilw-a iʒ-ʒara:f-a the-pretty-fsg the-giraffe-fsg
  ʔiz-zara:f-a il-Hilw-a the-giraffe-fsg the-pretty-fsg

16. She plays loves with the new her doll														
Hij	pat	ti-	ʃaʕab	tense	ji-	Hib	tense	b-	il-	ʤidi:d	-a	ʕarus	-at	-ha:
she	f	imp3fsg-	play	imp	imp3msg-	love	imp	with-	the-	new	-fsg	doll	fsg	-her

Violated Rules:
- Subject-Verb gender agreement:
  hij ji-Hib she imp3msg-love.imp
  hij ti-Hib she imp3fsg-love.imp
- Verb order:
  imp3fsg-play.imp imp3fsg-love.imp
  imp3fsg-love.imp imp3fsg-play.imp
- Noun-Adjective order:
  b-il-ʤidi:d-a ʕarus-at-ha: with-the-new-fsg doll-fsg-her
  b-ʕarus-at-h:a idʒ-ʤidi:d-a with-doll-fsg-her the-new-fsg

G.3.1. Summary Score Sheet for Syntactically Anomalous sentence

	lexical		grammatical		CELF	
9	4		6		3	
10	4		7		3	
11	4		7		3	
12	4		8		3	
13	4		9		3	
14	4		10		3	
15	4		11		3	
16	4		11		3	
	32		69		24	

	Omission	Substitution	Refusal	Unintelligible	Perseveration
lexical					
grammatical					

APPENDIX H: TESTS OF NORMALITY AND SUPPLEMENTARY NONPARAMETRIC ANALYSIS

H.1. Verbal Short Term Memory Test

Table G1. *Normality Test (subtest Span Score)*

Tests of Normality[b,c]

	Age Category	Shapiro-Wilk Statistic	df	Sig.
STM Word	2;6-2;11	.736	20	.000
	3;0-3;5	.854	20	.006
	3;6-3;11	.766	20	.000
	4;0-4;5	.351	20	.000
	4;6-4;11	.694	20	.000
	5;0-5;5	.822	20	.002
	5;6-5;11	.628	20	.000
STM Digit	2;6-2;11	.822	20	.002
	3;0-3;5	.849	20	.005
	3;6-3;11	.804	20	.001
	4;0-4;5	.763	20	.000
	4;6-4;11	.894	20	.032
	5;0-5;5	.790	20	.001
	5;6-5;11	.883	20	.020
STM Nonword	3;0-3;5	.447	20	.000
	3;6-3;11	.698	20	.000
	4;0-4;5	.688	20	.000
	4;6-4;11	.701	20	.000
	5;0-5;5	.506	20	.000
	5;6-5;11	.733	20	.000

b. Group = TD
c. STM Nonword is constant when Age Category = 1. It has been omitted.

H.1.1. Nonparametric Analysis (Subtest Span Score)

To investigate the effect of subtest type (three levels: Digit; Word; Nonword) on Span score; a Freedman's ANOVA was employed. Results indicated that the Span score was significantly influenced by subtest type $\chi^2(2) = 246.67$, $p < .001$. Wilcoxson tests were used to follow up this finding. A Bonferroni correction was applied ($\alpha = .0167$). Word Span was significantly less than Digit Span $Z = -5.97$, $p < .001$, $r = -.36$, medium effect size according to Cohen's (1988) guidelines. Nonword Span was significantly less than Word Span $Z = -10.32$, $p < .001$, $r = -.62$ and Digit Span $Z = -10.34$, $p < .001$, $r = -.62$, both were large effect sizes.

To investigate the effect of age (seven 6-month age bands) on Span score, a Kruskal-Wallis test was employed for each subtest. Results revealed that all three subtests were significantly affected by age with large effect sizes. For the Word List Recall subtest, $H(6)= 82.27$, $p < .001$. Jonckheere's test revealed a significant trend in the data: as age increased Word Span increased, $J = 6717$, $z = 9.75$, $r = .82$. For the Digit Recall subtest, $H(6)= 63.55$, $p < .001$. Jonckheere's test revealed a significant trend in the data: as age increased Word Span increased, $J = 6358.5$, $z = 8.17$, $r = .69$. . For the Nonword List Recall subtest, $H(6)= 60.95$, $p < .001$. Jonckheere's test revealed a significant trend in the data: as age increased Word Span increased, $J = 6034.5$, $z = 7.41$, $r = .63$.

Table G2. *Normality Test (Total Span Score)*

Tests of Normality[a]

	Age Categories STM	Shapiro-Wilk Statistic	df	Sig.
STM Total Span	3-3.11	.946	40	.053
	4-4.11	.909	40	.004
	5-5.11	.957	40	.128

H.1.2. Nonparametric Analysis (Total Span Score)

To investigate the effects of age on Total Span score a Kruskal-Wallis test was employed. Results revealed that Total Span score was significantly affected by age. $H(2)= 46.23$, $p < .001$. Jonckheere's test revealed a significant trend in the data: as age increased Total Span score increased, $J = 3859$, $z = 7.08$, $r = .65$.

H.2. Sentence Repetition Test

Table G3. *Normality Test (Morpheme Score)*

Tests of Normality[a]

	Age Category	Shapiro-Wilk Statistic	df	Sig.
Lex %	2.6-2.11	.892	20	.029
	3-3.5	.948	20	.342
	3.6-3.11	.950	20	.362
	4-4.5	.960	20	.544
	4.6-4.11	.959	20	.525
	5-5.5	.889	20	.026
	5.6-5.11	.808	20	.001
Gram %	2.6-2.11	.840	20	.004
	3-3.5	.907	20	.056
	3.6-3.11	.925	20	.126
	4-4.5	.932	20	.167
	4.6-4.11	.974	20	.841
	5-5.5	.954	20	.435
	5.6-5.11	.906	20	.054

H.2.1. Nonparametric Analysis (Morpheme Score)

To investigate the effects of Morpheme type (Lexical and Grammatical) on repetition score a Wilcoxon signed-rank test was employed. Results revealed that Grammatical Morpheme score was significantly lower than Lexical Morpheme score, with a large effect size. $z = -9.21$, $p < .001$, $r = -.55$.

To investigate the effects of age (7 six-month age bands) on Lexical Morpheme score a Kruskal-Wallis test was employed for each age category. Results revealed a significant effect of age $H(6)= 92.24$, $p < .001$. Jonckheere's test revealed a significant trend in the data with a large effect size: as age increased Lexical Morpheme score increased, $J = 7113.5$, $z = 10.62$, $r = .90$.

To investigate the effects of age (7 six-month age bands) on Grammatical Morpheme score a Kruskal-Wallis test was employed for each age category. Results revealed a significant effect of age $H(6)= 93.07$, $p < .001$. Jonckheere's test revealed a significant trend in the data with a large effect size: as age increased Grammatical Morpheme score increased, $J = 7162.5$, $z = 10.79$, $r = .91$.

Table G4. *Normality Test (Total Sentence Accuracy Score)*

Tests of Normality

	Age Category	Shapiro-Wilk Statistic	df	Sig.
TSA	2.6-2.11	.579	20	.000
	3-3.5	.664	20	.000
	3.6-3.11	.911	20	.066
	4-4.5	.933	20	.177
	4.6-4.11	.980	20	.930
	5-5.5	.955	20	.450
	5.6-5.11	.901	20	.043

H.2.2. Nonparametric Analysis (Total Sentence Accuracy Score)

To investigate the effects of age on Total Sentence Accuracy score a Kruskal-Wallis test was employed. Results revealed that Total Sentence Accuracy score was significantly

affected by age. $H(6)= 92.75$, $p < .001$. Jonckheere's test revealed a significant trend in the data with a large effect size: as age increased Total Sentence Accuracy score increased, $J = 7130$, $z = 10.59$, $r = .90$.

H.3. Anomalous Sentence Repetition Test

Table G 5. *Normality Test (Morpheme Score by Sentence Type)*

Tests of Normality[a]

	Age Category	Shapiro-Wilk Statistic	df	Sig.
Typical Lexical Score	4.0-4.5	.939	20	.226
	4.6-4.11	.847	20	**.005**
	5.0-5.5	.869	20	**.011**
	5.6-5.11	.819	20	**.002**
Semantically Anomalous Lexical Score	4.0-4.5	.958	20	.504
	4.6-4.11	.954	20	.427
	5.0-5.5	.890	20	**.027**
	5.6-5.11	.924	20	.119
Syntactically Anomalous Lexical Score	4.0-4.5	.963	20	.603
	4.6-4.11	.979	20	.914
	5.0-5.5	.900	20	**.041**
	5.6-5.11	.892	20	**.029**
Typical Grammatical Score	4.0-4.5	.920	20	.099
	4.6-4.11	.924	20	.119
	5.0-5.5	.940	20	.242
	5.6-5.11	.970	20	.756
Semantically Anomalous Grammatical Score	4.0-4.5	.939	20	.225
	4.6-4.11	.943	20	.271
	5.0-5.5	.951	20	.375
	5.6-5.11	.907	20	**.056**
Syntactically Anomalous Grammatical Score	4.0-4.5	.991	20	.999
	4.6-4.11	.959	20	.531
	5.0-5.5	.928	20	.144
	5.6-5.11	.934	20	.182

a. Group = TD

H.3.1. Nonparametric Analysis

To investigate the effect of sentence type (three levels: Typical; Semantically Anomalous; Syntactically Anomalous) on Lexical Morpheme score; a Freedman's ANOVA was employed. Results indicated that Lexical Morpheme score was significantly influenced by sentence type $\chi^2(2) = 75.74$, $p < .001$. Wilcoxson tests were used to follow up this finding. A Bonferroni correction was applied ($\alpha = .0167$). Lexical Morpheme score for Semantically Anomalous sentences was significantly less than Typical sentences $Z = -3.69$, $p < .001$, $r = -.29$, approaching a medium effect size. Lexical Morpheme score for Syntactically Anomalous sentences was significantly less than Typical sentences $Z = -7.05$, $p < .001$, $r = -.56$ and Semantically Anomalous sentences $Z = -6.78$, $p < .001$, $r = -.52$, both were large effect sizes. To investigate the effect of sentence type (three levels: Typical; Semantically Anomalous; Syntactically Anomalous) on Grammatical Morpheme score; a Freedman's ANOVA was employed. Results indicated that Grammatical Morpheme score was significantly influenced by sentence type $\chi^2(2) = 121.67$ $p < .001$. Wilcoxson tests were used to follow up this finding. A Bonferroni correction was applied ($\alpha = .0167$). Grammatical Morpheme score for Semantically Anomalous sentences was significantly less than Typical sentences $Z = -1.88$, $p < .001$, $r = -.15$, a small effect size. Grammatical Morpheme score for Syntactically Anomalous sentences was significantly less than Typical sentences $Z = -7.77$, $p < .001$, $r = -.61$ and Semantically Anomalous sentences $Z = -7.76$, $p < .001$, $r = -.61$, both were large effect sizes.

To investigate the effects of Morpheme type (Lexical and Grammatical) on repetition score for each sentence type, a Wilcoxon signed-rank test was employed. Results revealed that Grammatical Morpheme score was significantly lower than Lexical Morpheme score in all of the three sentence types. For Typical sentences $z = -5.67$, $p < .001$, $r = -.49$, approaching a large effect size. For Semantically Anomalous sentences, $z = -3.74$, $p < .001$, $r = -.30$, a medium effect size. For Syntactically Anomalous sentences, $z = -7.68$, $p < .001$, $r = -.61$, a large effect size.

To investigate the effects of age (with four 6-month age bands) on Lexical Morpheme score, a Kruskal-Wallis test was employed for each sentence type. Results revealed that the

Lexical Morpheme score for all three sentences types was significantly affected by age with large effect sizes. For Typical sentences, $H(3)= 20.35$, $p < .001$. For Semantically Anomalous sentences, $H(3)= 31.04$, $p < .001$. For Syntactically Anomalous sentences, $H(3)= 31.07$, $p < .001$. Jonckheere's test revealed a significant trend in the data: as age increased the median Lexical Morpheme score increased for all three sentence types. For Typical Sentences [$J = 1721.5$, $z = 4.53$, $r = .51$], for Semantically Anomalous sentences [$J = 1851.5$, $z = 5.63$, $r = .63$] and for Syntactically Anomalous sentences [$J = 1829$, $z = 5.43$, $r = .61$].

To investigate the effects of age (with four 6-month age bands) on Grammatical Morpheme score, a Kruskal-Wallis test was employed as well, for each sentence type. Results revealed that the Grammatical Morpheme score for all three sentences types was significantly affected by age with large effect sizes. For Typical sentences, $H(3)= 25.16$, $p < .001$. For Semantically Anomalous sentences, $H(3)= 23.8$, $p < .001$. For Syntactically Anomalous sentences, $H(3)= 24.42$, $p < .001$. Jonckheere's test revealed a significant trend in the data: as age increased the median Grammatical Morpheme score increased for all three sentence types. For Typical Sentences [$J = 1783.5$, $z = 5.03$, $r = .56$], for Semantically Anomalous sentences [$J = 1769.5$, $z = 4.9$, $r = .55$] and for Syntactically Anomalous sentences [$J = 1770.5$, $z = 4.91$, $r = .55$].

Table G6. *Normality Test (Morpheme Score)*

		Shapiro-Wilk		
		Statistic	df	Sig.
Lexical Morpheme %	Controls	.767	16	.001
	Language Concerns	.970	16	.843
Grammatical Morpheme %	Controls	.828	16	.007
	Language Concerns	.963	16	.719

To investigate the effects of Morpheme type (Lexical and Grammatical) on repetition score a Wilcoxon signed-rank test was employed. Results revealed that Grammatical Morpheme score was significantly lower than Lexical Morpheme score, with a large effect size. $z = - 4.49$, $p < .001$, $r = -.56$.

To investigate the effects of language group on Lexical Morpheme score, a Mann-Whitney U test was employed. Results indicated a significant difference with a medium effect size; the Lexical Morpheme score of participants in the Language Concerns group was lower. $U = 60.5$, $p =.01$, $r = -.45$. Grammatical Morpheme score was also significantly lower in the Language Concerns group with a large effect size. $U = 46$, $p =.001$, $r = -.55$.

Table G7. *Normality Test (Total Sentence Accuracy Score)*

		Shapiro-Wilk		
		Statistic	df	Sig.
Total Sentence Accuracy	Controls	.938	16	.321
	Language Concerns	.766	16	.001

To investigate the effects of language group on Total Sentence Accuracy score, a Mann-Whitney U test was employed. Results indicated a significant difference with a large effect size; the Total Accuracy Score of participants in the Language Concerns group was lower. $U = 50.5$, $p =.003$, $r = -.52$.

Tests of Normality

	LI_Comp	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Lex %	TD	.261	16	.005	.767	16	.001
	LI	.131	16	.200*	.970	16	.843
Gram %	TD	.191	16	.123	.828	16	.007
	LI	.100	16	.200*	.963	16	.719
CELF	TD	.108	16	.200*	.938	16	.321
	LI	.252	16	.008	.766	16	.001
STM Word	TD	.220	16	.037	.934	16	.282
	LI	.234	16	.020	.856	16	.016
STM Digit	TD	.198	16	.095	.888	16	.052
	LI	.309	16	.000	.822	16	.005
STM Nonword	TD	.352	16	.000	.697	16	.000
	LI	.484	16	.000	.496	16	.000
STM total	TD	.158	16	.200*	.917	16	.153
	LI	.148	16	.200*	.929	16	.231

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Mann-Whitney Test

Ranks

	LI_Comp	N	Mean Rank	Sum of Ranks
STM Word	TD	16	21.34	341.50
	LI	16	11.66	186.50
	Total	32		
STM Digit	TD	16	21.78	348.50
	LI	16	11.22	179.50
	Total	32		
STM Nonword	TD	16	20.66	330.50
	LI	16	12.34	197.50
	Total	32		
STM Total	TD	16	21.88	350.00
	LI	16	11.13	178.00
	Total	32		

Test Statistics[b]

	STM Word	STM Digit	STM Nonword	STM Total
Mann-Whitney U	50.500	43.500	61.500	42.000
Wilcoxon W	186.500	179.500	197.500	178.000
Z	-3.010	-3.273	-2.835	-3.259
Asymp. Sig. (2-tailed)	.003	.001	.005	.001
Exact Sig. [2*(1-tailed Sig.)]	.003[a]	.001[a]	.011[a]	.001[a]

a. Not corrected for ties.
b. Grouping Variable: LI_Comp

Friedman Test

Ranks

	Mean Rank
STM Digit	**2.69**
STM Word	**2.22**
STM Nonword	**1.09**

Test Statistics[a]

N	32
Chi-Square	52.846
df	2
Asymp. Sig.	.000

a. Friedman Test

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
STM Digit - STM Word	Negative Ranks	1[a]	4.50	4.50
	Positive Ranks	16[b]	9.28	148.50
	Ties	15[c]		
	Total	32		
STM Digit - STM Nonword	Negative Ranks	0[d]	.00	.00
	Positive Ranks	29[e]	15.00	435.00
	Ties	3[f]		
	Total	32		
STM Word - STM Nonword	Negative Ranks	0[g]	.00	.00
	Positive Ranks	29[h]	15.00	435.00
	Ties	3[i]		
	Total	32		

a. STM Digit < STM Word
b. STM Digit > STM Word
c. STM Digit = STM Word
d. STM Digit < STM Nonword
e. STM Digit > STM Nonword
f. STM Digit = STM Nonword
g. STM Word < STM Nonword
h. STM Word > STM Nonword
i. STM Word = STM Nonword

Test Statistics[b]

	STM Digit - STM Word	STM Digit - STM Nonword	STM Word - STM Nonword
Z	-3.510[a]	-4.757[a]	-4.758[a]
Asymp. Sig. (2-tailed)	.000	.000	.000

a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

Mann-Whitney Test

Ranks

	LI_Comp	N	Mean Rank	Sum of Ranks
STM Word	TD	16	21.34	341.50
	LI	16	11.66	186.50
	Total	32		
STM Digit	TD	16	21.78	348.50
	LI	16	11.22	179.50
	Total	32		
STM Nonword	TD	16	20.66	330.50
	LI	16	12.34	197.50
	Total	32		
STM Total	TD	16	21.88	350.00
	LI	16	11.13	178.00
	Total	32		

Test Statistics[b]

	STM Word	STM Digit	STM Nonword	STM Total
Mann-Whitney U	50.500	43.500	61.500	42.000
Wilcoxon W	186.500	179.500	197.500	178.000
Z	-3.010	-3.273	-2.835	-3.259
Asymp. Sig. (2-tailed)	.003	.001	.005	.001
Exact Sig. [2*(1-tailed Sig.)]	.003[a]	.001[a]	.011[a]	.001[a]

a. Not corrected for ties.
b. Grouping Variable: LI_Comp

Tests of Normality

	LIcomp	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
Typical Lexical Score	TD	.230	11	.108	.848	11	.041
	LI	.179	11	.200*	.926	11	.371
Typical Grammatical Score	TD	.185	11	.200*	.843	11	.035
	LI	.156	11	.200*	.954	11	.693
Typical Total Sentence Accuracy Score	TD	.139	11	.200*	.949	11	.628
	LI	.241	11	.074	.887	11	.129
Semantically Anomalous Lexical Score	TD	.160	11	.200*	.925	11	.358
	LI	.135	11	.200*	.969	11	.876
Semantically Anomalous Grammatical Score	TD	.199	11	.200*	.884	11	.116
	LI	.159	11	.200*	.954	11	.694
Semantically Anomalous Total Sentence Accuracy Score	TD	.165	11	.200*	.956	11	.727
	LI	.232	11	.099	.783	11	.006
Syntactically Anomalous Lexical Score	TD	.137	11	.200*	.933	11	.445
	LI	.126	11	.200*	.974	11	.925
Syntactically Anomalous Grammatical Score	TD	.131	11	.200*	.971	11	.899
	LI	.205	11	.200*	.916	11	.284
Syntactically Anomalous Total Sentence Accuracy Score	TD	.230	11	.109	.925	11	.358
	LI	.274	11	.020	.757	11	.003

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
Typical Lexical Score	22	76.7045	20.22307	21.88	100.00
Semantically Anomalous Lexical Score	22	73.5795	22.17539	25.00	100.00
Syntactically Anomalous Lexical Score	22	64.7727	23.60667	12.50	96.88

Descriptive Statistics

	Percentiles 25th	50th (Median)	75th
Typical Lexical Score	62.5000	84.3750	91.4063
Semantically Anomalous Lexical Score	60.1563	78.1250	93.7500
Syntactically Anomalous Lexical Score	46.0938	67.1875	82.0313

Friedman Test

Ranks

	Mean Rank
Typical Lexical Score	2.34
Semantically Anomalous Lexical Score	2.18
Syntactically Anomalous Lexical Score	1.48

Test Statistics[a]

N	22
Chi-Square	10.487
df	2
Asymp. Sig.	.005

a. Friedman Test

Friedman Test

Ranks

	Mean Rank
Typical Grammatical Score	2.59
Semantically Anomalous Grammatical Score	2.14
Syntactically Anomalous Grammatical Score	1.27

Test Statistics[a]

N	22
Chi-Square	20.186
df	2
Asymp. Sig.	.000

a. Friedman Test

Test Statistics[b]

i. Syntactically Anomalous Lexical Score = Semantically Anomalous Lexical Score

	Semantically Anomalous Lexical Score - Typical Lexical Score	Syntactically Anomalous Lexical Score - Typical Lexical Score	Syntactically Anomalous Lexical Score - Semantically Anomalous Lexical Score
Z	-1.085[a]	-3.082[a]	-3.209[a]
Asymp. Sig. (2-tailed)	.278	.002	.001

a. Based on positive ranks.
b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	Semantically Anomalous Grammatical Score - Typical Grammatical Score	Syntactically Anomalous Grammatical Score - Typical Grammatical Score	Syntactically Anomalous Grammatical Score - Semantically Anomalous Grammatical Score
Z	-1.740[a]	-3.686[a]	-3.670[a]
Asymp. Sig. (2-tailed)	.082	.000	.000

a. Based on positive ranks.
b. Wilcoxon Signed Ranks Test

Test Statistics[b]

	Typical Grammatical Score - Typical Lexical Score	Semantically Anomalous Grammatical Score - Semantically Anomalous Lexical Score	Syntactically Anomalous Grammatical Score - Syntactically Anomalous Lexical Score
Z	-3.328[a]	-3.458[a]	-3.847[a]
Asymp. Sig. (2-tailed)	.001	.001	.000

a. Based on positive ranks.
b. Wilcoxon Signed Ranks Test

Mann-Whitney Test

Ranks

	LIcomp	N	Mean Rank	Sum of Ranks
Typical Lexical Score	TD	11	15.55	171.00
	LI	11	7.45	82.00
	Total	22		
Semantically Anomalous Lexical Score	TD	11	16.00	176.00
	LI	11	7.00	77.00
	Total	22		
Syntactically Anomalous Lexical Score	TD	11	16.00	176.00
	LI	11	7.00	77.00
	Total	22		
Typical Grammatical Score	TD	11	16.32	179.50
	LI	11	6.68	73.50
	Total	22		
Semantically Anomalous Grammatical Score	TD	11	16.55	182.00
	LI	11	6.45	71.00
	Total	22		
Syntactically Anomalous Grammatical Score	TD	11	16.50	181.50
	LI	11	6.50	71.50
	Total	22		

Test Statistics[b]

Test Statistics[b]

	Typical Lexical Score	Semantically Anomalous Lexical Score	Syntactically Anomalous Lexical Score	Typical Grammatical Score	Semantically Anomalous Grammatical Score	Syntactically Anomalous Grammatical Score
Mann-Whitney U	16.000	11.000	11.000	7.500	5.000	5.500
Wilcoxon W	82.000	77.000	77.000	73.500	71.000	71.500
Z	-2.935	-3.261	-3.258	-3.483	-3.649	-3.614
Asymp. Sig. (2-tailed)	.003	.001	.001	.000	.000	.000
Exact Sig. [2*(1-tailed Sig.)]	.002[a]	.001[a]	.001[a]	.000[a]	.000[a]	.000[a]

a. Not corrected for ties.
b. Grouping Variable: LIcomp