

Language and Literacy Development in Children Learning English as an Additional Language: a Longitudinal Cohort and Vocabulary Intervention Study

Christopher Dixon

A thesis submitted for the degree of
Doctor of Philosophy

Department of Human Communication Sciences
University of Sheffield



The
University
Of
Sheffield.

March, 2018

Abstract

Children learning English as an Additional Language (EAL) are a growing but understudied population of learners in English primary schools. As EAL learners vary in their amount of exposure to English, they often begin formal education with relatively lower levels of English language proficiency than their monolingual peers. Little is known about the English language and literacy developmental trajectories of EAL learners in England, and particularly, the extent to which the two groups of learners converge or diverge over time. Additionally, no studies to date have assessed the efficacy of explicit, targeted vocabulary instruction in this group of learners in the run up to the end of primary school.

The present study comprised a longitudinal cohort study of 48 EAL learners and 33 monolingual peers who were assessed at three time points between Year 4 (age 8-9) and Year 5 (age 9-10) on a battery of English language and literacy measures. All EAL learners had received English-medium education since at least Year 1 (age 5-6). Relative to their monolingual peers, EAL learners showed strengths in rapid naming, single-word reading efficiency, and spelling, but weaknesses in vocabulary knowledge, expressive syntax, and passage reading accuracy. Where they exhibited weaknesses, EAL learners generally did not make sufficient progress in order to close gaps with their monolingual peers.

A subgroup of nine EAL learners with English vocabulary weaknesses also participated in short-term vocabulary intervention. Working one-to-one with speech and language therapy students, children showed significant gains in receptive and productive knowledge of target vocabulary which were maintained six months later. Together, results indicate that regular classroom instruction may be insufficient for EAL learners to close gaps with their monolingual peers in certain domains of oral language, but that targeted vocabulary instruction may be an effective means of achieving this end.

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Silke Fricke and Dr. Jenny Thomson, for their ongoing support and constructive comments on earlier drafts of this thesis. I would also particularly like to thank Dr. Fricke for her support in my application to join the very first cohort of PhD scholars at Cumberland Lodge in 2014.

I have met and worked with many wonderful people along the way. I would like to thank the schools, teachers, parents, and pupils who participated in the study. I would also like to thank the Human Communication Sciences department and DevRes research group for opportunities to discuss my work, as well as my fellow PhD students for their conversation and support.

Finally, I would like to thank my parents, who sparked my interest in language by bringing me up in a country where my mother tongue was but one of many.

Contents

Abstract	iii
Acknowledgements	v
Contents	vii
List of Tables	xiii
List of Figures	xv
List of Abbreviations	xvii
Overview	1
1 Bilingual Language Development and Learning EAL in England	3
1.1 The Education of Bilingual Children	4
1.2 Two Types of Language Proficiency	4
1.3 English as an Additional Language in England	5
1.3.1 EAL Population Characteristics in England	5
1.3.2 Learning EAL in English Schools	6
1.3.3 Educational Achievement of EAL Pupils	8
1.3.4 Home Language and Literacy Experiences of Pupils Learning EAL	9
1.4 Summary	10
2 Literature Review I: Language and Literacy in Monolingual and Bilingual Development	13
2.1 Oral Language	13
2.1.1 Linguistic Input	14
2.1.2 Vocabulary Knowledge	14
2.1.2.1 Breadth, Depth and Measurement of Vocabulary Knowledge	15
2.1.2.2 Developmental Changes in Vocabulary Knowledge	16
2.1.2.3 Vocabulary Development in Mono- and Bilingual Children	17
2.1.3 Syntactic Knowledge	23
2.1.3.1 Syntactic Development in Mono- and Bilingual Children	23
2.1.3.2 Syntactic Knowledge and Oral Narrative	24
2.1.3.3 Syntax and Oral Narrative in Mono- and Bilingual Children	25
2.1.4 Listening Comprehension	26
2.1.4.1 Listening Comprehension in Mono- and Bilingual Children	26
2.1.5 Summary of Oral Language Development in Mono- and Bilingual Children	28
2.2 Literacy	28
2.2.1 The Simple View of Reading	29
2.2.1.1 The Role of Vocabulary in the SVR	30

2.2.1.2	The Role of Fluency in the SVR	31
2.2.1.3	The SVR in the Reading Acquisition of Bilingual Learners	31
2.2.2	Constraints on Reading Skills	32
2.2.3	Lower-Level Skills in Reading Development	33
2.2.3.1	Phonological Awareness and Orthographic Knowledge	34
2.2.3.2	Rapid Automatised Naming	36
2.2.3.3	Word Recognition	38
2.2.3.4	Summary of Lower-Level Reading Skills in Mono- and Bilingual Development	41
2.2.4	Higher-Level Skills in Reading Development	41
2.2.4.1	Cognitive Architecture and Mental Models	42
2.2.4.2	Inference Generation	43
2.2.4.3	Prior Knowledge	44
2.2.4.4	Passage Reading Skill in Mono- and Bilingual Development	45
2.2.5	Writing	48
2.2.5.1	Writing Development in Mono- and Bilingual Children	49
2.3	The Contribution of Oral Language to Reading Comprehension	50
2.3.1	The Role of Vocabulary, Syntax, and Listening Comprehension in Reading Comprehension	50
2.3.2	The Role of Oral Language in the Reading Comprehension of Mono- and Bilingual Learners	53
2.4	Summary of Literature Review I and Aims of Longitudinal Study	54
3	Methods I: Longitudinal Study	57
3.1	Design and Recruitment	57
3.2	Ethical Considerations	58
3.3	Participants	59
3.4	Measures	60
3.4.1	Non-Verbal Reasoning and Memory	60
3.4.1.1	WISC-IV Matrix Reasoning	60
3.4.1.2	CELF-IV Number Repetition	61
3.4.2	Vocabulary Measures	61
3.4.2.1	BPVS-III	61
3.4.2.2	CELF-IV Expressive Vocabulary	62
3.4.2.3	WISC-IV Vocabulary	62
3.4.3	Other Oral Language Measures	63
3.4.3.1	CELF-IV Understanding Spoken Paragraphs	63
3.4.3.2	CELF-IV Formulated Sentences	63
3.4.4	Oral Narrative Retell	64
3.4.4.1	Peter and the Cat	64
3.4.5	Phonological Processing Measures	66
3.4.5.1	PhAB Spoonerisms	66
3.4.5.2	CTOPP Rapid Naming	66

3.4.6	Literacy Measures	67
3.4.6.1	TOWRE-II	67
3.4.6.2	YARC-II Primary	67
3.4.6.3	Writing: Bespoke Task	68
3.5	Questionnaires	69
3.6	Procedure	71
4	Results and Discussion I: Longitudinal Study	73
4.1	Preliminary Considerations and Background Measures	73
4.1.1	Attrition and Missing Data	73
4.1.2	Use of Raw and Standardised Scores	74
4.1.3	Effect Sizes	75
4.2	Background Measures	75
4.3	Descriptive Statistics	75
4.3.1	Descriptives for Vocabulary Measures	76
4.3.2	Descriptives for Oral Language Measures	76
4.3.3	Descriptives for Oral Narrative Measures	79
4.3.4	Descriptives for Phonological Processing Measures	81
4.3.5	Descriptives for Literacy Measures	83
4.3.6	Summary of Trends in Descriptive Statistics	89
4.4	Linear Mixed Modelling	89
4.4.1	Goodness-of-fit and Model Building	90
4.4.2	General Procedure and Model Assumptions	92
4.4.3	Linear Mixed Modelling of Vocabulary Measures	95
4.4.3.1	Receptive Vocabulary	97
4.4.3.2	Expressive Vocabulary	98
4.4.3.3	Vocabulary Depth	99
4.4.4	Linear Mixed Modelling of Other Oral Language Measures	100
4.4.4.1	Listening Comprehension	102
4.4.4.2	Expressive Grammar	103
4.4.5	Linear Mixed Modelling of Oral Narrative Measures	104
4.4.5.1	Total Utterances	107
4.4.5.2	Mean Length of Utterance in Words (MLUw)	108
4.4.5.3	Lexical Diversity	109
4.4.5.4	Oral Narrative Error Analysis	110
4.4.6	Linear Mixed Modelling of Phonological Processing Measures	114
4.4.6.1	Rapid Automatised Naming	117
4.4.6.2	Spoonerisms	119
4.4.7	Linear Mixed Modelling of Literacy Measures	120
4.4.7.1	Single-Word Reading Efficiency	120
4.4.7.2	Passage Reading Rate, Accuracy, and Comprehension	123
4.4.7.3	Writing	131
4.4.8	Summary of Linear Mixed Effects Modelling	139

4.5	Discussion	142
4.5.1	Vocabulary Knowledge	143
4.5.1.1	Group Differences in Vocabulary Knowledge at t1	143
4.5.1.2	Group Differences in Trajectories for Vocabulary Knowledge	145
4.5.2	Oral Language	147
4.5.2.1	Group Differences in Other Oral Language Measures at t1	147
4.5.2.2	Group Differences in Trajectories for Other Oral Language Measures	149
4.5.3	Oral Narrative	151
4.5.3.1	Group Differences in Oral Narrative Measures at t1	151
4.5.3.2	Group Differences in Trajectories for Oral Narrative	152
4.5.4	Phonological Processing	153
4.5.4.1	Group Differences in Phonological Processing at t1	154
4.5.4.2	Group Differences in Trajectories for Phonological Processing	155
4.5.5	Single-Word Reading Efficiency	156
4.5.5.1	Group Differences in Single-Word Decoding at t1	156
4.5.5.2	Group Differences in Trajectories for Single-Word Decoding	157
4.5.6	Passage Reading	158
4.5.6.1	Group Differences in Passage Reading Measures at t1	158
4.5.6.2	Group Differences in Trajectories for Passage Reading Measures	161
4.5.7	Writing	163
4.5.7.1	Group Differences in Writing at t1	163
4.5.7.2	Group Differences in Trajectories for Writing	164
4.5.8	Summary	165
5	Literature Review II: Vocabulary Instruction for Mono- and Bilingual Learners	167
5.1	Incidental Learning of Word Meanings	167
5.2	Explicit Learning of Word Meanings: Definitional and Contextual Methods	169
5.3	Key Considerations in Vocabulary Instruction with Monolingual Children	171
5.3.1	Word Selection	171
5.3.2	Contextual Affordances	172
5.3.3	Depth of Processing and Active Engagement	173
5.3.4	Dosage and Multiple Exposures	176
5.3.5	Child-friendly Definitions	178
5.4	Summary	178
5.5	Oral Language and Vocabulary Instruction of Bilingual Learners	179
5.5.1	Intervention Studies with Younger Bilingual Learners	179
5.5.2	Intervention Studies with Older Bilingual Learners	182
5.6	Summary of Literature Review II and Aims of Vocabulary Intervention Study	184
6	Methods II: Vocabulary Intervention Study	187
6.1	Design and Recruitment	187
6.2	Ethical Considerations	188
6.3	Participants	188
6.4	Timeline and Delivery of Intervention	189

6.5	Target Word Selection	190
6.6	Structure of Intervention Sessions	191
6.7	Pilot Study	193
6.8	Recruitment and Training of Student Coordinators	194
6.9	Measures and Analytical Strategy of the Intervention Study	195
6.9.1	Primary Outcome Measure of Word Learning	195
6.9.1.1	Inter-rater Reliability	197
6.9.2	Transfer Measures	198
6.9.3	Timing and Administration of Measures	198
6.9.4	Multiple Case Series Design Analytical Strategy	199
6.10	Implementation Fidelity	200
7	Results and Discussion II: Vocabulary Intervention Study	201
7.1	Fidelity of Implementation	201
7.1.1	Dosage	202
7.1.2	Attention and Engagement	203
7.1.3	Mind Map Activities	204
7.1.4	Coordinator Observations and Feedback Session	204
7.1.5	Summary of Implementation Fidelity	205
7.2	Results of the Vocabulary Intervention Study	206
7.2.1	Group and Individual Trajectories for Taught and Untaught Words	211
7.2.1.1	Incidence of Non-Responses	213
7.2.2	Group and Individual Trajectories in Receptive and Productive Knowledge	214
7.2.3	Further Analysis of Cases BA and JG	217
7.2.4	Transfer to Standardised Assessments	220
7.3	Discussion of Results	225
7.3.1	Efficacy of Intervention Teaching	226
7.3.2	Acquisition of Receptive and Expressive Knowledge	230
7.3.3	Individual Growth Trajectories	233
7.3.4	Transfer of Teaching onto Standardised Assessments	234
7.3.5	Conclusions	235
8	General Discussion	237
8.1	Language and Literacy Development in Children Learning EAL	237
8.1.1	Schooling Closes Some but not All Gaps	237
8.1.2	Developmental Trajectories can be Altered with Targeted Instruction	241
8.2	Strengths, Limitations and Future Directions	243
8.2.1	Design and Statistical Framework	243
8.2.2	Choice of Measures	244
8.2.3	Selection Criteria and Retention of Intervention Participants in the Longitudinal Study	246
8.2.4	Design and Analytical Strategy of the Vocabulary Intervention	247
8.2.5	Selection of Target Words	248
8.3	Educational Implications	249

8.4 Conclusion	250
Appendices	251
References	291

List of Tables

3.1	School Characteristics at t1	59
3.2	Summary of Measures by Type and Time Point	60
3.3	Results of Child Language Balance Questionnaire	70
4.1	Use of Raw and Standardised Scores Across Measures	74
4.2	t1-t3 Descriptive Statistics for Vocabulary Measures	77
4.3	t1-t3 Descriptive Statistics for Oral Language Measures	78
4.4	t1-t3 Descriptive Statistics for Oral Narrative Measures	80
4.5	t1-t3 Descriptive Statistics for Phonological Processing Measures	82
4.6	t1-t3 Descriptive Statistics for Single-Word Reading Efficiency Measures	84
4.7	t1-t3 Descriptive Statistics for Passage Reading Measures	85
4.8	t1-t3 Descriptive Statistics for Writing Measures	88
4.9	Step-Up Linear Mixed Model Building Process	92
4.10	Linear Mixed Modelling of Vocabulary Measures	96
4.11	Linear Mixed Modelling of Oral Language Measures	101
4.12	Linear Mixed Modelling of Oral Narrative Measures	106
4.13	Linear Mixed Modelling of Oral Narrative Errors	112
4.14	Linear Mixed Modelling of Phonological Processing Measures	116
4.15	Linear Mixed Modelling of Single-Word Reading Efficiency	121
4.16	Linear Mixed Modelling of Passage Reading Measures (Standard Scores)	125
4.17	Linear Mixed Modelling of Passage Reading Measures (Passage 3 only)	126
4.18	Linear Mixed Modelling of Written Narrative Measures	132
4.19	Examples of T-Units in the Bespoke Writing Task	135
4.20	LMMs for Written Narrative Error Types	137
4.21	Summary of Intercepts and Slopes for Mono and EAL Group on All Variables	140
4.22	Comparison of Monolingual-EAL Group Discrepancies on the BPVS	144
4.23	Summary of Group Coefficients in Passage Reading Measures	159
6.1	Intervention Session Structure Outline	191
6.2	Inter-Rater Reliability for Bespoke Word Knowledge Assessment	198
7.1	Fidelity of Implementation of Intervention Teaching	202
7.2	Examples from Mind-Map Activities among the 9 Intervention Participants	205
7.3	Progress of Intervention Participants in Total, Word, and Sentence Score for Taught Words	208
7.4	Progress of Intervention Participants in Total, Word, and Sentence Score for Untaught Words	209
7.5	Raw Scores and Percentage Change of the 9 Children on Total, Word, and Sentence Score	210
7.6	Incidence of Non-Responses Across Time Points (Taught and Untaught Words)	213

7.7 Individual Children’s BPVS, CELF EV, WISC VC, and CELF FS Standardised Scores from t1 of the Longitudinal Study	218
7.8 CELF Observational Rating Scale Scores for 6 Intervention Participants	219
7.9 Progress on Standardised Measures Across the Study for all Intervention Participants	221
7.10 Individual Trajectories in Standardised Measures between t1, Baseline, Posttest, and t3	222

List of Figures

1.1	BICS and CALP	5
1.2	Percentage Enrolment of Primary School Pupils in England with EAL in 2015	6
3.1	Timeline of the Longitudinal Study	58
4.1	Graphical Explanation of Linear Mixed Modelling	91
4.2	Linear Mixed Modelling of t1-t3 Receptive Vocabulary	97
4.3	Linear Mixed Modelling of t1-t3 Expressive Vocabulary	98
4.4	Linear Mixed Modelling of t1-t3 Vocabulary Depth	99
4.5	Linear Mixed Modelling of t1-t3 Listening Comprehension	102
4.6	Linear Mixed Modelling of t1-t3 Expressive Grammar	103
4.7	Linear Mixed Modelling of t1-t3 Expressive Grammar (Common Items)	104
4.8	Linear Mixed Modelling of t1-t3 Total Utterances in Oral Narrative	107
4.9	Linear Mixed Modelling of t1-t3 Mean Length of Utterance in Words	108
4.10	Linear Mixed Modelling of t1-t3 Lexical Diversity (Root TTR)	109
4.11	Linear Mixed Modelling of t1-t3 Error Rate in Oral Narrative	110
4.12	Linear Mixed Modelling of t1-t3 Morphosyntactic Error Rate in Oral Narrative	113
4.13	Linear Mixed Modelling of t1-t3 Semantic Error Rate in Oral Narrative	114
4.14	Linear Mixed Modelling of t1-t3 Rapid Automatised Letter Naming	117
4.15	Linear Mixed Modelling of t1-t3 Rapid Automatised Digit Naming	118
4.16	Linear Mixed Modelling of t1-t3 Spoonerisms	119
4.17	Linear Mixed Modelling of t1-t3 Sight-Word Reading Efficiency	122
4.18	Linear Mixed Modelling of t1-t3 Phonemic Decoding Efficiency	123
4.19	Linear Mixed Modelling of t1-t3 Passage Reading Rate	127
4.20	Linear Mixed Modelling of t1-t3 Passage Reading Accuracy	128
4.21	Linear Mixed Modelling of t1-t3 Passage Reading Comprehension	130
4.22	Linear Mixed Modelling of t1-t3 Total T-Units in Writing	133
4.23	Linear Mixed Modelling of t1-t3 Mean Length of T-Unit in Words	134
4.24	Linear Mixed Modelling of t1-t3 Errors in Writing	136
4.25	Linear Mixed Modelling of t1-t3 Spelling Error Rate in Writing	138
4.26	Linear Mixed Modelling of t1-t3 Morphosyntactic Errors in Writing	139
6.1	Participant Flow throughout the Intervention Study	188
6.2	Intervention Study Timeline in Relation to Longitudinal Study	199
7.1	Coordinators' Mean Ratings of Children's Engagement and Attention Throughout the Intervention	203
7.2	Group and Individual Trajectories in Total Score for Taught and Untaught Words	212
7.3	Group and Individual Trajectories in Receptive Productive Knowledge of Taught Words	215
7.4	Intervention Participants' Progress on CELF FS and WISC VC (Raw Scores)	223

7.5 Intervention Participants' Progress on CELF FS and WISC VC (Scaled Scores) . . 225

List of Abbreviations

AoA	Age of Acquisition
BPVS	British Picture Vocabulary Scale-III
CELF	Clinical Evaluation of Language Fundamentals-IV
CTOPP	Comprehensive Test of Phonological Processing
DfE	Department for Education
DfES	Department for Education and Skills
EAL	English as an Additional Language
ELL	English Language Learner
ESCAL	Every Sheffield Child Articulate and Literate team
GCSE	General Certificate of Secondary Education
GLD	Good Level of Development
KS2	Key Stage 2
LMM	Linear Mixed Modelling
MCDI	MacArthur Communicative Development Inventory
MLU / MLT	Mean Length of Utterance / Mean Length of T-Unit
MWP	Multi-Word Phrase
NARA	Neale Analysis of Reading Ability
NDW	Number of Different Words
NQT	Newly Qualified Teacher
PA	Phonological Awareness
PPVT	Peabody Picture Vocabulary Test
SD	Standard Deviation
SVR	Simple View of Reading
TROG	Test for Reception of Grammar
TTR	Type/Token Ratio
WISC	Wechsler Intelligence Scale for Children
YARC	York Assessment of Reading Comprehension Primary

Overview

This thesis concerns the development of language and literacy skills in children who are learning English as an Additional Language (EAL) in primary school in England. The proportion of children with EAL in English primary schools has grown steadily in recent years, roughly doubling from 11% in 2004 to 20.6% in 2017 (Department for Education and Skills [DfES], 2004; Department for Education [DfE], 2017). 'EAL' in England describes a highly heterogeneous group of learners who are estimated to speak upwards of 300 individual languages (CILT, 2005), and who tend to underperform in relation to their monolingual peers on national assessments of reading and writing in primary school (Strand, Malmberg & Hall, 2015). Many pupils learning EAL face the dual challenge of acquiring English language proficiency while mastering curriculum content, and because monolingual children also continue to develop their understanding, speaking, reading, and writing skills in English, many EAL learners are said to be 'aiming at a moving target' (Lesaux, 2015; NALDIC, 1999).

The first aim of this study is to follow a cohort of 8 to 10 year-old primary school children learning EAL and their monolingual peers over time in various aspects of English oral language and literacy skill in order to examine and compare developmental trajectories. A growing base of research literature from the U.K. consistently points to the significantly lower English vocabulary knowledge of EAL learners in relation to their monolingual peers: thus, the second aim of this study is to design, deliver, and evaluate a low-intensity vocabulary intervention programme for a small subgroup of EAL learners from the longitudinal cohort study.

The thesis is organised as follows: Chapter 1 introduces the study of bilingualism and the make-up and educational attainment of EAL learners in England; Chapter 2 reviews literature on language and literacy development in mono- and bilingual learners; Chapters 3 and 4 cover the methods, results, and discussion of the longitudinal cohort study; Chapter 5 reviews literature on vocabulary acquisition and instruction in mono- and bilingual learners; Chapters 6 and 7 cover the methods, results, and discussion of the vocabulary intervention study; and finally Chapter 8 discusses overarching themes, strengths, limitations, and educational implications of both studies.

Chapter 1

Bilingual Language Development and Learning English as an Additional Language in England

While no one definition of bilingualism exists, it may be broadly considered as the ability to understand or communicate in two languages (Baetens Beardsmore, 1982; Ng & Wigglesworth, 2007). Bilingualism can be said to describe a dimensional rather than discrete phenomenon, with bilinguals varying in the proficiency of their second or additional language according to age of acquisition, patterns of language use, and educational and societal demands. For instance, it is common to find bilinguals with varying competencies across the four skills of understanding, speaking, reading, and writing in each of their languages (Romaine, 1995).

In England, the term 'English as an Additional Language' (EAL) is most commonly used to describe children who are exposed to a 'first language' other than English during 'early development' and who continue to use this language in the home or community setting (DfES, 2007). Unfortunately, such a definition is rather vague in nature and makes no assumptions as to relative proficiency in each language or patterns of acquisition, and thus acts as a 'catch-all' for all bilingual learners in the U.K.

Although definitions in the bilingual development literature differ greatly as to exact cut-off points, 'simultaneous' bilinguals are generally considered to be those children who begin to acquire their second or additional language in infancy or toddlerhood and to acquire proficiency in both languages roughly in tandem, while 'sequential' or 'successive' bilinguals begin to acquire it after this period, potentially already being fluent in one language before beginning to acquire another (de Houwer, 2009; Edwards, 2004; Gathercole et al., 2014; Lesaux, 2015; Paradis Genesee & Crago, 2010). The DfES (2007) definition of EAL makes no such distinction.

Within the international literature, bilingual learners are referred to as, for example, 'English Language Learners', 'Limited English Proficient', 'Language Minority Learners', or simply 'Bilinguals' (Carlo et al., 2004; Mancilla-Martinez, 2010; Silverman et al., 2014). As a result, caution is warranted in the interpretation and synthesis of results across different studies, given that there is not necessarily a one-to-one correspondence between the label used and the linguistic history and proficiency of the individuals contained therein. A number of studies make use of the terms L1 (first language) and L2 (second language): the acquisition of an L2 is said to occur after that of the L1 and thus is found to be a qualitatively different process (Johnson & Newport, 1989). The terms L1 and L2 will not be used to describe participants in the present study so as to avoid assumptions concerning the relative timing of exposure to children's different languages which, again, does not form part of the DfES (2007) definition of EAL. Additionally, in this thesis the term 'target language' will be used to refer to the majority language, typically being the language of instruction in school (for example, English in the case of children learning EAL in England).

This chapter begins with a brief overview of educational provision for bilingual children, followed by the introduction of a crucial distinction between two types of language proficiency and their development in this population of learners. The final section of the chapter will introduce the demographics, educational status, attainment, and language learning experiences of learners with EAL in England specifically.

1.1 The Education of Bilingual Children

There are a number of different approaches to the education of bilingual children, varying in the degree to which they recognise and foster development of the first or home language. Baker (2006) draws distinctions between 'monolingual'; 'weak'; and 'strong' forms of bilingual education. Typically, monolingual and weak forms of bilingual education aim to assimilate pupils into the majority language and culture, with all instruction being delivered in the target language either from the very beginning of formal education, or after a short period of instruction in the home language (a transition process). Strong forms and dual-language programs, on the other hand, explicitly promote bilingualism by providing equitable instruction in each language. In Canada, for instance, immersion programs introduce children to a second or additional language they may otherwise not experience outside of the home to a level of exposure that results in a high level of linguistic competence. In immersion programs, the school curriculum is delivered in the additional language by bilingual teachers, with dedicated first language instruction being introduced some time later (Swain & Johnson, 1997).

There is evidence to suggest that recognising and promoting bilingual children's first or home language may result in higher educational attainment than for those who receive all instruction through the target language only (Duran, Roseth & Hoffman, 2010; Thomas & Collier, 2002). However, proponents of Content and Language Integrated Learning (CLIL), in which instruction is delivered solely in the target language, emphasise the advantage of opportunities for the simultaneous acquisition of content and linguistic knowledge (for a review of CLIL, see Dalton-Puffer, 2011).

1.2 Two Types of Language Proficiency

Cummins (1979; 1981a) distinguishes between two types of language proficiency in terms of basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP). These skills are conceptualised within a Cartesian space, along one axis ranging from cognitively demanding to cognitively undemanding, and on another from context-embedded to context-reduced (see Figure 1.1 below).

BICS and CALP exist in opposing quadrants. BICS relies heavily on context, is cognitively undemanding, and is characteristic of face-to-face communication in everyday situations. CALP, on the other hand, relies upon abstraction and displaced reference, and is characteristic of reading comprehension tasks, for example, which make higher demands on vocabulary knowledge and inferencing without the help of real-time context. While BICS is typically mastered by bilingual learners within around two years of exposure to the target language, CALP has a longer developmental trajectory of between five to seven years (Collier, 1987, 1989; Cummins, 1981b; Demie,

1.3. English as an Additional Language in England

2013; Hakuta, Butler & Witt, 2000; Thomas & Collier, 2002). CALP becomes increasingly important over a child's academic career, especially for the transition from learning to read to reading to learn, when higher demands come to be placed on children's reading comprehension skills (Chall, Jacobs & Baldwin, 1990).

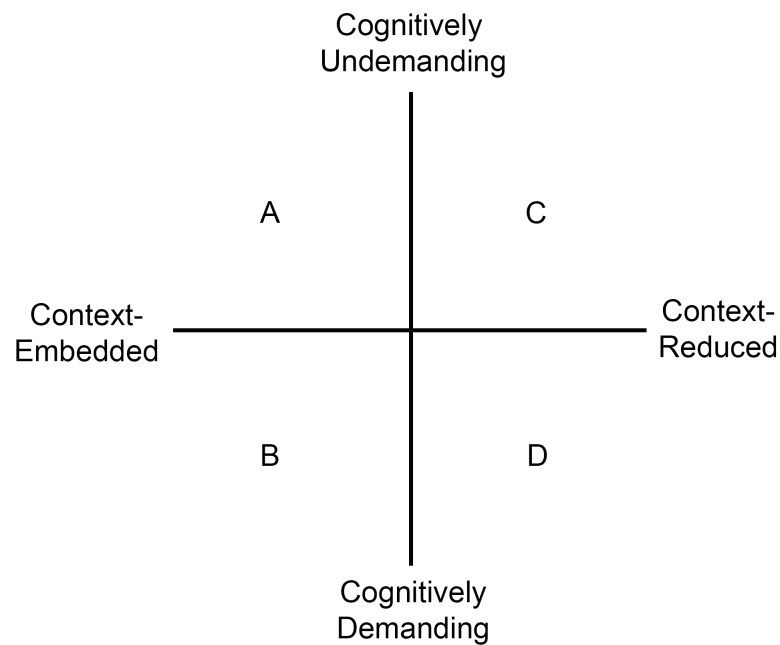


Figure 1.1: BICS and CALP (adapted from Cummins, 1981). BICS and CALP are represented by quadrants A and D, respectively.

1.3 English as an Additional Language in England

The U.K. is home to speakers of a wide variety of languages, but differs importantly from a number of other countries with similarly diverse populations. The officially recognised status of languages such as Gaelic in Scotland and Welsh in Wales means that the U.K. as a whole cannot truly be described as a monolingual nation. Therefore, this review will be restricted to learners of EAL in English schools, which almost exclusively employ a 'monolingual' form of language education.

1.3.1 EAL Population Characteristics in England

The U.K. mimics a trend seen in many other countries of a steadily increasing proportion of school pupils whose first language differs from that of the majority culture. As of 2015 (the year in which recruitment in the current study began), over 693,000 children, or 19.4% of primary school pupils in England, did not speak English as their first language (DfE, 2015a). By 2017, this proportion had increased by 1.2 percentage points to 20.6% (DfE, 2017). The 2011 U.K. census identified Polish, Panjabi, Urdu, Bengali, and Gujarati as the top five most commonly spoken 'other' main languages in England and Wales (ONS, 2013), although it is estimated that at least 300 distinct languages are spoken by primary school pupils in England (CILT, 2005). Geographically, children

learning EAL are unevenly distributed, with the highest concentrations of ethnolinguistic minority communities¹ in the areas of Inner and Outer London, Yorkshire and the Humber, and the West Midlands (see Figure 1.2 below). In a small number of areas it is not uncommon for children learning EAL to comprise the majority of schools' enrolment (e.g. often upwards of 80%). In stark contrast, however, EAL learners comprise up to 5% of all pupils in over half of England's primary schools, and 1% or fewer in nearly one quarter of schools (Strand et al., 2015).

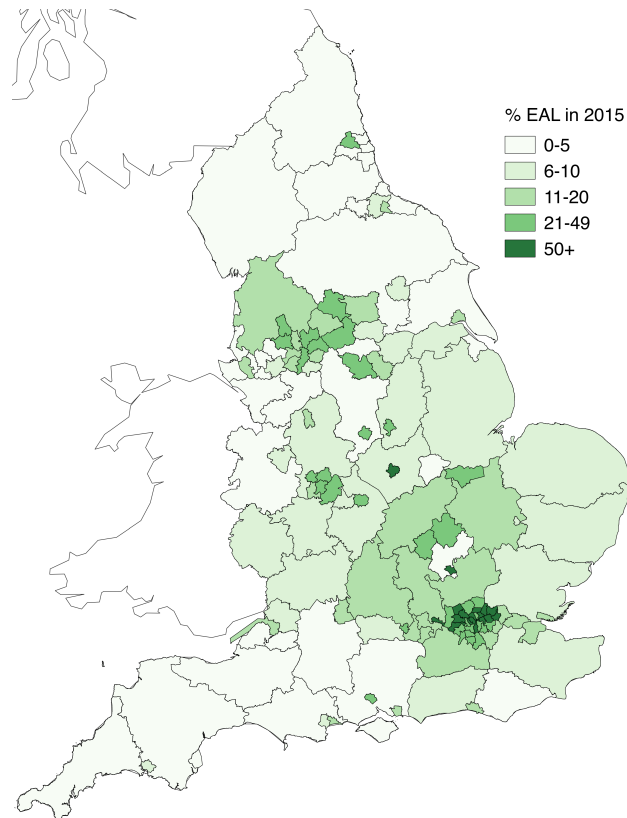


Figure 1.2: Percentage enrolment of pupils with EAL in English primary schools by Local Authority (map created using data from DfE, 2015b)

1.3.2 Learning EAL in English Schools

In England, all primary school mainstream classroom instruction takes place in English (Tsimpli, 2017). However, some schools do employ permanent or peripatetic bilingual support staff for the purpose of inducting and supporting new arrivals, a policy previously endorsed and promoted by the schools inspectorate OFSTED (2005). Where specific instruction (as opposed to support) in minority languages is available, this tends to take place outside of mainstream classes or school hours, particularly in complementary schools (Blackledge & Creese, 2010; CILT, 2005; Wardman, 2013).

The educational status of EAL learners in England is somewhat attributable to the recommendations of the Swann Report, which advocated inclusive education of all pupils in the state-

¹An ethnolinguistic community is defined as one in which members do not speak the same language and are not of the same ethnic or cultural group as the majority culture (Paradis, Genesee & Crago, 2010).

1.3. English as an Additional Language in England

maintained education sector (Swann, 1985). The recommendations were laudable in promoting social integration, but ultimately decreased the visibility of EAL in the curriculum. As a result, EAL may now be considered only a 'supra-subject phenomenon' (Leung, 2001). It is for this reason that EAL learners do not receive any dedicated language instruction in school, but rather are expected to acquire English through regular classroom instruction and engagement with the National Curriculum (Cameron & Besser, 2004).

As EAL has no concrete status in the National Curriculum, and concentrations of EAL learners are by no means uniform across the country, responsibility for EAL provision is devolved from central government to Local Education Authorities (LEAs) to take on as they best see fit (Costley, 2014). The latest iteration of England's National Curriculum in 2013 dedicated only 94 words to learners of EAL and offered no concrete guidance on EAL pedagogy or assessment (DfE, 2013). Nationally, student teachers and newly qualified teachers (NQTs) receive inconsistent, and in only few cases accredited, training in EAL pedagogy (NALDIC, 2014). Indeed, it is not uncommon for student teachers to receive only one hour of instruction on the subject throughout their initial teacher training (D. Excell, personal communication, February 2, 2015). As a result, many NQTs feel inadequately prepared for the task of teaching and assessing children whose first language is not English (Cajkler & Hall, 2009).

The diversity of EAL learners' cultural and linguistic experiences is not well captured by the binary EAL/non-EAL label currently used in the English school system: not only is this label problematic in the sheer number of different languages it subsumes, but also in the fact that it says nothing of children's level of English language proficiency, resulting in a situation in which "the bilingual child of a French banker is grouped together with a Somali refugee who may not speak English at all" (EEF, 2015. p.1). Indeed, no explicit mention of English language proficiency is made in the definition of EAL found in government documents (e.g. DfES, 2007).

Despite this, alternative forms of English language assessment for pupils learning EAL have been devised. The Stages of English (Ellis, Hester & Barrs, 1990), for instance, lists four broad stages ranging from 'new to English', to 'fully fluent user of English'. Additionally, the QCA (2000) assessment tool and NASSEA (2001, 2015) EAL Assessment Systems provide extensions to traditional National Curriculum descriptors of speaking, listening, reading and writing to capture levels of linguistic proficiency particularly for children in the 'new to English' category. It should be noted, however, that use of such assessment frameworks is optional and intended only for monitoring purposes. One study using attainment data of 940 pupils in one London local authority found that children learning EAL took an average of six years to reach the 'fully fluent' stage of English proficiency (Demie, 2013). However, what is most interesting about these findings is that pupils spent less time in the early stages and more in the later stages, supporting the protracted developmental nature of CALP (Cummins, 1981; Thomas & Collier, 2002).

As of September 2016, primary maintained schools in England (i.e. those that are funded by the government and adhere to the National Curriculum) have been required to implement a new 'proficiency in English' framework by recording the English language proficiency of pupils learning EAL against a 5-point scale, ranging from 'new to English' to 'fluent' (DfE, 2015c). Similar to the Stages of English, this framework provides descriptors of each stage including competence in speaking, listening, reading, and writing, as well as the typical amount of support pupils may need to access the curriculum; however, at the time of the present study it is too early to assess the

validity and utility of this scale and to what extent it represents an improvement over use of the EAL/non-EAL label for pedagogical or assessment purposes.

1.3.3 Educational Achievement of EAL Pupils

Before a discussion of the educational achievement of EAL pupils in England, it should be pointed out that EAL learners, just like their non-EAL peers, fall along a spectrum of low and high attainment, and that knowledge of a home language other than English does not necessarily relate to educational outcomes in either direction².

A recent analysis of the National Pupil Database in the U.K. revealed that when evaluated as a group, children learning EAL are disadvantaged in relation to their monolingual peers in some aspects of their educational attainment (Strand et al., 2015). From Early Years Foundation Stage (age 5) to Key Stage 2 in primary school (KS2; age 7-11), children learning EAL are particularly behind in their reading but less so in their mathematical ability and 'Grammar, Punctuation, and Spelling', as defined by the National Curriculum (DfE, 2013). By the end of secondary school (age 15-16), an achievement gap is still visible in terms of the proportion of pupils with EAL who achieve A*-C grades in GCSE English (odds ratio = 0.83; where < 1 indicates lower probability for the EAL group of attaining these grades and > 1 indicates higher probability). However, it is also interesting to note that by this point pupils with EAL begin to outperform their monolingual peers in mathematics and modern foreign languages (odds ratios = 1.03 and 1.90, respectively), suggesting that even by age 16, the specific learning needs of pupils with EAL continue to lie in the area of English language and literacy skills.

Some research on EAL attainment has attempted to disaggregate pupils according to language proficiency and exposure to English. Using data from the Longitudinal Study of Young People in England, Strand et al. (2015) were able to compare school attainment at age 14 (a composite of English, Science, and Maths) amongst three groups of pupils: those with English as their only language (English Only; n=11,878); those with English as their main language alongside a different home language (English Main; n=2,100); and those with another language as their main language (Other Main; n=976). Additionally, pupils were disaggregated by their length of residence in the U.K., ranging from birth to 14 years. In general, there were clear trends for pupils who were born or who had arrived in the U.K. at an early age to perform more highly than those who had arrived later, and also for the English Main group to perform very similarly to the English Only group. In contrast, Other Main pupils significantly and consistently underperformed in relation to the other two groups regardless of length of residence. This pattern had shifted slightly by age 16, whereby among pupils born in the U.K., English Main pupils were now outperforming monolingual English speakers (English Only), although not significantly so. While the Other Main group continued to underperform in relation to the other groups, there was clear evidence of a closing of the gap, with these learners now scoring on average -0.20 SD in relation to the mean achievement of all groups (an improvement from -0.40 SD at age 14). Thus, although

²Contrary to often reported monolingual advantages in educational attainment, there is evidence for certain advantages among bilingual learners who display relatively high degrees of proficiency in both languages. Particularly, these individuals have been found to outperform monolingual peers on measures of executive functioning (e.g. attentional shifting and inhibition; Adesope, Lavin, Thompson & Ungerleider, 2010; Barac & Bialystok, 2011; although see Gathercole et al., (2014) and Paap & Greenberg (2013) for criticisms and contrary findings).

1.3. English as an Additional Language in England

these data are limited to the secondary schooling phase and do not provide a pure indication of language and literacy development alone, they do highlight the effect of different language exposure patterns on the attainment of children with EAL, and provide evidence of a general closing of the gap in attainment over time.

Whiteside, Gooch and Norbury (2017) considered the independent contribution of language proficiency and EAL status on social, emotional, and behavioural difficulties, as well as educational attainment in reception and Year 2. The study reports data from teacher questionnaires on 7,267 reception children in England (age 3 to 5), including scores on the Children's Communication Checklist-2 (Bishop, 2003a) as a measure of perceived language proficiency, information about whether children made a good level of development (GLD) as defined by the Early Years Foundation Stage (DfE, 2014), whether they were on target by Year 2, and the extent to which poorly performing children in reception caught up by Year 2. Results showed that when group status (i.e. EAL/non-EAL) was used as an independent variable, the picture was rather negative for children with EAL, who were found to have significantly lower language proficiency in reception and lower likelihood of achieving a GLD in reception. However, when language proficiency was added as a predictor in a hierarchical regression analysis, this pattern changed, showing that, for all children, lower language proficiency was associated with poorer outcomes, but that EAL status was somewhat of a protective factor. Particularly, children with EAL were more likely than their non-EAL peers to be on target in Year 2, and more likely to catch up between reception and Year 2. This study illustrates the importance of not only EAL status, but also general language proficiency in consideration of developmental trajectories, and has strength in also considering the language skills of a monolingual comparison group.

1.3.4 Home Language and Literacy Experiences of Pupils Learning EAL

England is an increasingly multicultural country – a melting pot of different cultures, ethnicities, religions, and languages (Crouch & Stonehouse, 2016). Although recent changes in migratory patterns mean that Polish is now the most commonly spoken language in England after English (ONS, 2013; Sumption & Somerville, 2010), a great deal of research has been conducted on what were the previously most populous ethnolinguistic minority groups – particularly speakers of Punjabi, Urdu, Gujarati, and Bengali.

Children of South Asian heritage often occupy an interesting learning space between language, literacy, culture, and religion: in particular, they may acquire the home language orally as their first language, but often do not learn to read and write in this language (common examples include Mirpur Punjabi and Sylheti; Anwar, 1998; Gregory, 1996; Rosowsky, 2001, 2010). Thus, many children from these communities may acquire literacy for the first time in their second or additional language which they have not yet fully mastered (i.e. English), contrary to monolingual children who begin to acquire oral English from birth and then proceed to learn to read and write in that same language at or even before school entry. However, this is not to suggest that these learners do not experience literacy practices at all before beginning formal education, as many attend mosque regularly from a young age and learn to read Qur'anic Arabic (Parke & Drury, 2001; Rosowsky, 2001). Hirst (1998) conducted interviews with South Asian (predominantly Pakistani) bilingual families in the U.K. with children aged between 2 and 4 years of age. The study revealed

that many children were receiving exposure to three or four languages in the home, including Punjabi as the language of the family, Urdu for written correspondence with relatives, English as the language of the majority culture, and Arabic as the language of the Qur'an. Observations indicated prevalence of "a rich print and linguistic environment" (p.423) for the majority of the children, with storytelling and shared book reading often taking place. Although a number of parents professed limited proficiency in English, many communicated high aspirations for their children's language development and educational attainment.

Beech and Keys (1997) administered a Language Preference Questionnaire to forty 8 year-old South Asian EAL primary school pupils in order to ascertain relative balance in exposure between the children's two languages. Results showed clear trends for higher exposure to and use of the home language at home with parents and other relatives, but higher exposure to English in terms of media consumption, talking with siblings and friends at school (even if they spoke the same home language), and use of English as the language of thought when completing school work. This pattern coincides with the findings of a small-scale ethnographic study by Parke and Drury (2001), who conducted interviews with the parents of three young Pahari-speaking children (ages 3;6 to 4;4) during their transition to nursery school. The study revealed a strict separation in parents' perceived roles of the home environment and the school in terms of supporting language development (also see Garton & Pratt, 2009; Gregory, 1996): at home, children were socialised and highly immersed in Pahari language and culture, tending to play exclusively with Pahari-speaking siblings or extended family, and some making regular trips to Pakistan. Indeed, parents professed a strong desire to nurture their children's home language development before "the change from Pahari to English" when children would begin nursery school (p.123). Interviews with teachers revealed that all three children began nursery with well-developed oral language skills in Pahari but with little or no productive English.

These studies provide interesting insights into some young EAL learners' home language environments and patterns of exposure to different languages. On a national level, too, many ethnolinguistic minority communities in England share certain characteristics which may have impacts upon language development. In particular, many such communities are highly concentrated in urban, socio-economically deprived, residentially segregated areas (Lothers & Lothers, 2012; Sumption & Somerville, 2010; for a full review see Casey, 2016). Therefore, it is important to consider that a number of children acquiring EAL may face additional barriers in their acquisition of English, which may or may not be shared with their monolingual peers.

1.4 Summary

England is home to a growing number of primary school children who are acquiring EAL. The term 'EAL' represents a country-specific instantiation of bilingualism, although the vague and all-encompassing nature of this label is somewhat problematic. While changes are afoot to better categorise the language proficiency of pupils with EAL (i.e. proficiency in English descriptors), this system is by no means well-established and it is too early to ascertain whether such a system does indeed provide advantages over the previously binary classification.

EAL learners in England who attend state-maintained primary schools enter into a form of 'monolingual' education in which the national curriculum and all its associated assessment pro-

1.4. Summary

cedures are delivered entirely in English. While there is evidence for the underperformance of bilingual children on national assessments of reading and writing from the earliest stages of formal education, the magnitude of this discrepancy does decrease gradually over time (Strand et al., 2015).

Research from certain ethnolinguistic communities in England shows that families are eager to transmit their heritage language and culture to their children. In some cases, they are not able to provide English language support, and defer to schools to fulfil this purpose. However, such observations must be qualified with a degree of uncertainty due to the relatively little amount of research on EAL learners in England, and by the high degree of heterogeneity in terms of patterns of linguistic exposure.

Given that EAL learners are likely to begin their formal education with lower levels of English language proficiency than their monolingual peers, and given that studies suggest a period of five to seven years for bilingual learners to catch up with their monolingual peers in academic language proficiency, a key aim of this study is to examine to what extent discrepancies continue to exist in various English language and literacy skills towards the end of primary education, after four years of English-medium formal instruction.

The following chapter will discuss language and literacy development in mono- and bilingual children in more detail, and will introduce research questions associated with the first aim of the study.

Chapter 2

Literature Review I: Language and Literacy in Monolingual and Bilingual Development

Oral language is a cornerstone of human development. In typically developing children, phonology, semantics, vocabulary, grammar, and pragmatics develop naturally and with little or no effort in the presence of sufficient linguistic exposure; literacy, on the other hand, is typically acquired only with a great deal of conscious effort and explicit instruction (Cain & Oakhill, 2007; Perfetti, 1985). While literacy does have a set of specific skills such as conceptions of print, letter knowledge, and phoneme-grapheme correspondence rules (Whitehurst & Lonigan, 1998), many foundational skills of learning to read are 'parasitic' or dependent upon oral language skills (Nation & Angell, 2006). Evidence of the strong link between oral language and literacy skill is taken from findings that children who struggle with the ability to extract meaning from text also tend to have weaknesses in their oral language skills, including in vocabulary, syntax, and listening comprehension (Adloff & Catts, 2015; Catts, Adlof, & Weismer, 2006; Nation, Clarke, Marshall, & Durand, 2004; Nation & Snowling, 2004; Stothard & Hulme, 1996). In other words, oral language is a crucial foundation upon which literacy relies.

This chapter will begin with a discussion of oral language development in mono- and bilingual children (including simultaneous and sequential), before going on to discuss literacy development and the specific role played by oral language skills in reading comprehension. It should be noted from the outset that much of the international literature discussed here provides no readily available comparison with EAL learners in England, although the small pool of U.K.-based studies will be discussed when applicable.

2.1 Oral Language

This section will begin by considering the role of linguistic exposure or input in oral language development, and will subsequently examine the nature, development, and growth of three domains of oral language including vocabulary knowledge, syntactic knowledge, and listening comprehension in mono- and bilingual children¹. As much of the bilingualism literature has focused on development of vocabulary knowledge, the review will focus primarily on this domain.

¹For ease of comparison between studies, the effect size of standardised monolingual-bilingual group differences will be reported, where available and appropriate, in Cohen's *d*, where 0.2 is considered a small effect, 0.5 is medium, and 0.8 is large (Cohen, 1988).

2.1.1 Linguistic Input

Quantity and quality of linguistic input are strong predictors of language ability in both mono- and bilingual development (Cattani et al., 2014; Hart & Risley, 1995; Hoff, Core, Place, Rumiche & Parra, 2012; Huttenlocher et al., 1991; Jia & Aaronson, 2003; Thordardottir, 2011). By definition, bilingualism necessitates division of linguistic exposure between two languages, with the result that bilingual children do not receive one hundred percent of exposure in the target language. A large literature reports on the linguistic developmental trajectories of children and adults who immigrate to a foreign country and subsequently begin the path to second language acquisition (L2A). These studies support the general finding that older children and adults experience a faster rate of L2A, but individuals with an early age of exposure to the second language experience higher end-state proficiency (Krashen et al., 1979; Long, 1990). Such maturational constraints are suggestive of a sensitive period in L2A, although this does not exert equal influence on all linguistic domains: for instance, a much earlier age of exposure is typically needed to attain native-like competency in phonology than in morphology or syntax (Johnson & Newport, 1989).

2.1.2 Vocabulary Knowledge

The extremely large number of words in a language and their multidimensional shades of meaning make vocabulary acquisition a lifelong task. Individuals vary widely as to the size and quality of their word knowledge, especially as a result of reading experience and domain-specific knowledge. Nagy and Scott (2000) discuss five key characteristics of word knowledge:

1. *Incrementality*. Over time word knowledge becomes increasingly detailed and specified in small steps.
2. *Polysemy*. Many words have more than one meaning, and words in such networks may be more or less closely related. Polysemy is especially prevalent in figurative language.
3. *Multidimensionality*. Aside from meaning, knowledge of a word also includes its behaviour and occurrence with other words, its spoken and written forms, associations, and so on (e.g. Nation, 2001). Dimensions are independent from one another such that knowing one particular aspect does not guarantee knowledge of another.
4. *Interrelatedness*. Words cannot be conceptualised as 'isolated units of knowledge' but rather share properties. For example, knowing 'mammal' implies (at least an implicit) understanding of 'whale'.
5. *Heterogeneity*. Different types of words require different types of knowledge. For example, compare the knowledge required for function words such as 'the' and 'if', with the nouns 'hypotenuse' and 'ion', which require a high level of, in this case, scientific understanding.

Word knowledge is complex, and so too therefore are attempts to measure it. There is no one agreed definition of what it means to 'know' a word, although it may be said to incorporate understanding of spoken and written forms, meaning, grammatical functions, associations with other words, and stylistic constraints, each aspect of which has a receptive and productive component (Nation, 2001). The kind of vocabulary that is employed in communication is to some extent a

2.1. Oral Language

function of modality: the highly context-embedded nature of oral language interaction reduces demands on vocabulary knowledge, while the opposite is true of written language, which is relatively context-reduced, or 'decontextualised' (Cummins, 1981) and thus employs a more diverse and precise range of vocabulary (Kamil & Herbert, 2005; Perfetti, 1985).

The following subsections begin with a brief discussion of important distinctions between different types of vocabulary knowledge and how this relates to measurement, as well as some general principles in the acquisition of that knowledge. Following this, the discussion will turn to research on bilingual vocabulary development specifically, considering relevant studies with EAL learners in England.

2.1.2.1 Breadth, Depth and Measurement of Vocabulary Knowledge

The multidimensional nature of vocabulary knowledge requires a multiplicity of assessment instruments in order to investigate its acquisition and use. Vocabulary knowledge may be conceptualised in terms of *breadth*, i.e. size of vocabulary or total number of words known, or *depth*, i.e. quality or interconnectedness of word knowledge (Anderson & Freebody, 1981). Further distinctions exist according to the method of vocabulary measurement. Measures requiring recognition of vocabulary are said to be *receptive*, while those requiring recall and use are said to be *expressive* (Laufer & Goldstein, 2004; Schmitt, 2010). Melka (1997) discusses the receptive-productive difference as one of degree or mastery only; that is, while an incipient lexical representation is sufficient for the purposes of recognition and imitation, a deeper level of knowledge is required for comprehension and production. These two forms of knowledge lend themselves to different methods of assessment, some brief examples of which will now be discussed.

Measures of breadth may include, for example, questionnaires to ascertain whether an individual feels that they know a word well enough *to be able* to give a definition (e.g. the Vocabulary Size Test; Mears & Jones, 1990), forced-choice tasks which require a target word to be identified from within a set of distractors (e.g. British Picture Vocabulary Scale III; Dunn, Dunn & NFER, 2009), and expressive naming tasks in which individuals are required to provide a target word verbally, typically with the aid of a visual cue (e.g. the Clinical Evaluation of Language Fundamentals Expressive [CELF] IV Vocabulary subtest; Semel, Wiig & Secord, 2006). On the other hand, measures of depth often include definition tasks (in which examinees are actually *prompted* to give a definition), cloze tasks, and oral interviews. Examples include the Wechsler Intelligence Scale for Children (WISC IV) Vocabulary depth subtest, a productive measure in which individuals are asked to produce verbal definitions of words and are awarded points for synonyms, examples, and explanations, and the Word Associate Test (Read, 1993), a receptive multiple-choice measure in which individuals are presented with stimuli and potential matches which are related either paradigmatically (e.g. *team* and *group*, which are synonymous), syntagmatically (e.g. *team* and *scientists*, as in the collocation *a team of scientists*), or analytically (*team* and *together*, as together represents one aspect of the stimulus word likely to be found in a dictionary definition). As can be seen, performance on such measures allows a more in-depth view of the degree and nature of connectedness of an individual's word knowledge, often as it relates to knowledge of other words.

Vocabulary breadth and depth are found to be both highly interrelated and also independent (Schmitt, 2014; Tannenbaum, Torgesen & Wagner, 2006). Lexical knowledge is organised in

networks of semantic relations, and therefore the ability to define a given word will often depend upon knowledge of other words in its network (Vermeer, 2001). Breadth and depth also have a reciprocal relationship in developmental and instructional aspects, as the acquisition of new words serves to create more fine-grained distinctions in lexical knowledge (Carey, 1978; Gardner, 2013; Hadley, Dickinson, Hirsh-Pasek, Golinkoff & Nesbitt, 2016). One example of the simultaneous relationship and separability between these constructs is found in a study by Tannenbaum et al. (2006), who set out to evaluate the most parsimonious structure of vocabulary knowledge regarding breadth, depth, and fluency. A sample of 203 third-graders (age 7-8) were administered a battery of six vocabulary measures. Confirmatory factor analysis best supported a two-factor model of Breadth and Depth/Fluency, which provided the best fit to the data. It should be noted that despite the apparent separability of these two factors, there was a strong and statistically significant correlation between them ($r = .88$).

2.1.2.2 Developmental Changes in Vocabulary Knowledge

Although estimates vary and children exhibit considerable individual differences in word learning, vocabulary knowledge during the school years is said to grow at a rate of approximately 2,500 to 3,000 words per year, or roughly at a rate of seven words per day (Beck & McKeown, 1991). While much of this word learning is due to exposure to oral language (Huttenlocher et al., 1991), the role of explicit instruction is also important. Biemiller and Slonim (2001) estimated vocabulary size and growth rates in terms of root words, i.e. not including derived, inflected, or compound forms and found a mean root word vocabulary size of just under 6,800 words by the end of Grade 4 (age 9-10). Crucially, this study offered two important findings: firstly, that between Grades 3 to 5, children gain an average of around 1,000 root words per year, or three root words per day, which is within the scope of instruction; and secondly, that there is evidence for a common developmental sequence of vocabulary acquisition, making it possible to predict which words children are likely to know.

Children's vocabulary knowledge development is also characterised by qualitative changes. Early lexical representations are considered 'partial' in nature – e.g. particularly constrained by perceptual characteristics – and over time begin to approximate adult models of word knowledge with the gradual addition of semantic features (Clark, 1973; Carter, 2012; Hadley et al., 2016). Furthermore, there is evidence of a gradual shift from syntagmatic to paradigmatic word relations as word knowledge becomes more abstract, potentially coinciding with the acquisition of literacy (Anglin, 1993; Cronin, 2002; Russel & Saadeh, 1962).

In summary, distinctions drawn between breadth, depth, receptive, and productive forms of vocabulary knowledge serve to promote methodological convenience, but also to allow investigation of multiple aspects of word knowledge and acquisition. Much work has focused on vocabulary acquisition in monolingual populations, which shows a steady increase during early schooling. The following section will consider the nature of vocabulary knowledge and acquisition in bilingual learners, beginning in early development (infancy and toddlerhood) and then moving on to later development (primary and secondary school years).

2.1.2.3 Vocabulary Development in Monolingual and Bilingual Children

Quantity and quality of language input have been shown to play an important role in the vocabulary development of monolingual children (e.g. Hart & Risley, 1995; Huttenlocher, Haight, Bryk, Seltzer & Lyons, 1991) and this, too, has important implications for bilingual language development. Children in bilingual environments receive split exposure between languages and therefore often exhibit smaller vocabularies in each of their languages separately than monolingual children do in their one and only language (Paradis et al., 2010). Furthermore, bilinguals use words in each language less frequently than do monolinguals, resulting in weaker links between semantic and phonological representations (the weaker links hypothesis; Gollan, Montoya, Cera & Sandoval, 2008). Early studies took a deficit view of bilingual language development (Hakuta & Diaz, 1985) but failed to take into account bilingual children's lexical knowledge from both languages. The following section describes more recent work which supports parity between monolingual and bilingual children in total vocabulary size even prior to two years of age.

2.1.2.3.1 Early Vocabulary Development

Many studies of early bilingual vocabulary development have sought to contrast vocabulary size in each of a child's languages, with a common finding that bilingual children possess a similarly sized conceptual lexicon to their monolingual peers when both languages are taken into consideration. Pearson, Fernandez and Oller (1993) asked mothers of 25 Spanish-English bilingual and 35 English monolingual children in Florida to complete inventories of their toddlers' total productive vocabularies using the MacArthur Communicative Development Inventory (MCDI, 1989). Mothers of bilingual children completed English and Spanish versions of the MCDI by indicating which words their children used productively and spontaneously. The scores from the MCDI were summed to give a measure of total vocabulary, and then compared to scores of monolingual children. The bilingual group showed a lower absolute productive vocabulary size in each of their languages relative to monolingual speakers of each language, but strikingly, did not differ from the monolingual children in total vocabulary size. Although generalisation is limited by the relatively small sample size of this study, similar results have been obtained from other cross-linguistic samples of bilingual toddlers (e.g. de Houwer, Bornstein & Putnick, 2014; Hoff et al., 2012; Junker & Stockman, 2002).

There is evidence that the monolingual vocabulary size advantage continues to be seen in preschool and kindergarten-age children. Leseman (2000) recruited Turkish-Dutch bilingual and low-SES Dutch monolingual 3 year-olds and followed their vocabulary development until kindergarten entry. While the two groups of children performed on a par with respect to first language vocabulary knowledge, the Turkish-speaking children fell far short of monolingual levels of Dutch receptive ($d = 2.19$) and productive ($d = 2.33$) vocabulary knowledge by age 4;2. Similarly, a study in Miami of 3 to 5 year-old monolingual English and bilingual Spanish-English children found a monolingual advantage in English on the Peabody Picture Vocabulary Scale (PPVT; Dunn & Dunn, 1997) of around 1.5 SD (Fernández, Pearson, Umbel & Oller, 1992).

Similar findings of English vocabulary differences between young EAL learners and their monolingual peers have been found in studies conducted in England. For example, in the language intervention study of Dockrell, Stuart and King (2010; discussed further in Section 5.5.1),

96 EAL learners with an average age of 3.5 years were assessed on the British Ability Scales (Elliott, Smith & McCulloch, 1997), including the productive Naming Vocabulary subtest, prior to and proceeding a 15-week packaged oral language intervention. Results showed that, even at posttest after receiving the intervention, the EAL learners were still significantly underperforming in relation to their monolingual peers in English productive vocabulary knowledge (a large effect size of $d = 1.5$; Cohen, 1988). Additionally, an investigative study of early language development by Basit, Hughes, Iqbal and Cooper (2015) found first language status to be a significant predictor of delay in language comprehension and production skills of 3 to 4 year-old children in English nursery schools. Although this study did not examine vocabulary explicitly, the primary outcome measure utilised in the study, the New Reynell Developmental Language Scales (Edwards, Letts & Sinka, 2011), did incorporate subtests of vocabulary knowledge. Interestingly, although non-English first language status was associated with a higher risk of language delay, children who spoke Asian languages (e.g. Bengali, Kurdish, Punjabi, and Urdu) were found to be less delayed in relation to their monolingual English-speaking peers ($d = 0.77$) compared to those children who spoke 'Other' languages (e.g. Polish, Mandarin, and Czech; $d = 1.34$), potentially suggesting that linguistic as well as acculturation factors may contribute to early language development in young children learning EAL in England.

In summary, the findings of studies on early bilingual language development reveal that there is some evidence for parity between the vocabulary knowledge of monolingual and bilingual toddlers, provided that both languages are considered. However, much of this work has been carried out in the U.S. with more homogeneous populations (e.g. Spanish-English speakers) and thus may lack direct applicability to bilingual learners elsewhere (see Murphy & Unthiah, 2015 for a similar criticism relating to intervention research). Studies in England suggest that, even before school entry, EAL learners are at risk of delay in their English language comprehension and production skills, which may require intervention to bring up to the level of monolingual peers.

2.1.2.3.2 Later Vocabulary Development

The lower vocabulary knowledge of bilingual learners in the target language has also been evidenced in older children, both internationally (e.g. Bialystok, Luk, Peets & Yang, 2010; Droop & Verhoeven, 2003; Geva & Farnia, 2012; Verhallen & Schoonen, 1998), and in a small number of studies conducted in the U.K. (e.g. Babayiğit, 2014a; Burgoyne, Kelly, Whiteley & Spooner, 2009; Burgoyne, Whiteley & Hutchinson, 2011a; Cameron, 2002; Hutchinson, Whiteley, Smith and Connors, 2003; Mahon & Crutchley, 2006).

In a longitudinal study in England, Hutchinson et al. (2003) followed the development of children learning EAL ($n=43$) and their monolingual peers ($n=43$) from Year 2 to the end of Year 4 (ages 6 to 8). All children were assessed on a battery of language and literacy measures in English including the Test of Word Knowledge (Wiig & Secord, 1992), yielding both receptive and expressive composite scores. The results revealed a '2 year developmental lag' in English vocabulary knowledge of children learning EAL, with the largest lag in Year 3 ($d = 2.17$). The developmental portrait had changed by Year 4, however, with EAL children catching up in their receptive knowledge but falling behind further on their expressive knowledge.

2.1. Oral Language

In a later study, Mahon and Crutchley (2006) assessed the receptive vocabulary knowledge of 97 EAL and 69 monolingual children in one-year age bands between 4 and 9 years of age using the BPVS-II. In accordance with the results of Hutchinson et al. (2003), there was a consistent monolingual advantage in receptive vocabulary knowledge in English at all ages. The two groups of children began to approximate each other more closely over time, with children learning EAL making a great deal of progress between ages 6 and 8, at which time the EAL group was performing at just under 0.5 SD below the monolingual group. Although reference to scaled scores may be questionable in samples of bilingual learners, results showed that the majority of EAL learners in this study were performing within the normal range in reference to a monolingual norming population, with 85% obtaining standard scores of between 85 and 115 on the BPVS-II.

Burgoyne et al. (2011a) followed a cohort of 39 EAL and 39 monolingual learners from primary schools in England on a range of language and literacy measures in Years 3 to 4, including the Receptive and Expressive One-Word Picture Vocabulary Test (Brownell, 2000). Results were supportive of a significant monolingual vocabulary knowledge advantage in English, the magnitude of which varied across the two time points of the study. Specifically, while the groups converged slightly in expressive vocabulary ($d = 0.98$ to 0.80), the opposite pattern applied to receptive vocabulary, where the gap widened over time ($d = 0.75$ to 0.98).

In a cross-sectional study, Babayiğit (2014a) assessed the oral language and reading skills of 56 monolingual and 69 EAL learners in England (ages 9-10) who were matched on their amount of English-medium instruction (a minimum of four years). Children with EAL spoke a total of 15 different languages (the sample also included three trilingual speakers) and all had limited experience of reading instruction in the home language. Results showed that, despite their equal amount of instruction, EAL learners were performing well below the level of their monolingual peers in their receptive vocabulary knowledge as measured by the BPVS-II ($d = 1.12$).

There is also some evidence to suggest that English vocabulary weaknesses are to be found at later educational stages. Cameron (2002) recruited a sample of 84 monolingual and 63 EAL secondary school students (ages 13-15) who were administered the Levels Test (Nation, 1990), an assessment which measures word knowledge at various frequency bands, for example, knowledge of the most frequently occurring 1,000 words, 3,000 words, and so on, with each band becoming progressively less frequent and therefore more difficult. Results showed that after 10 years of English-medium instruction, EAL students still displayed significant lags in their English vocabulary knowledge relative to their monolingual peers, particularly at the 3,000 and 5,000 word frequency levels which are considered crucial for adequate comprehension in reading (Nation & Waring, 1997). The results of this study complement previously discussed findings in indicating that word frequency may be an important variable to consider in comparisons of vocabulary knowledge among the two groups.

Finally, as alluded to above, monolingual advantages in target language vocabulary knowledge are also reported in the international literature. In a large study of 772 mono- and 966 bilingual children in Canada, Bialystok, Luk, Peets and Yang (2010) found a significant and enduring monolingual advantage in receptive vocabulary knowledge in English as measured by the PPVT. This pattern remained across the age groups, and increased slightly in magnitude between ages 8 and 10. With a similarly aged cohort, Droop and Verhoeven (2003) conducted a longitudinal study of monolingual Dutch ($n=163$) and bilingual Turkish-Dutch ($n=82$) and Moroccan-Dutch

($n=60$) primary school children in the Netherlands. In this study, the monolingual group was split into higher and lower socio-economic status (SES) based on school statistics in order to provide a more appropriate comparison with the two bilingual groups which were also categorised as low-SES. In terms of receptive vocabulary knowledge, results showed a large advantage of the low-SES monolingual Dutch group at the start of Grade 3 (age 8) relative to both the Turkish-Dutch ($d = 2.62$) and Moroccan-Dutch ($d = 1.93$) bilingual groups. A similar monolingual advantage applied to expressive vocabulary at this time, although to a relatively reduced degree ($d = 1.85$ for Turkish-Dutch, and $d = 1.38$ for Moroccan-Dutch). The magnitude of this monolingual advantage decreased with age for receptive but not expressive vocabulary, exactly the opposite pattern to that reported by Burgoyne et al. (2011a), underlining the presence of discrepancies in the literature.

In summary, monolingual advantages in very early vocabulary development appear to be retained at later stages, including into secondary education. This group discrepancy thus appears to be an enduring and well-established one, although the variety of language exposure patterns of EAL learners inevitably leads to a great deal of heterogeneity in this population of learners (Cline & Shamsi, 2000).

2.1.2.3.3 Vocabulary Depth Knowledge in Monolingual and Bilingual Children

Most studies in the literature tend to focus on *discrete* measures of vocabulary knowledge (i.e. whether words are known or not), but some work has also assessed depth of knowledge. Verhallen and Schoonen (1993) conducted a vocabulary depth interview with 40 monolingual Dutch and 40 bilingual Turkish-Dutch 9 and 10 year-olds who were asked questions about stimulus words such as 'what can you do with it?' and then asked to use the word in a sentence. Responses were coded as paradigmatic, syntagmatic, or subjective. The bilingual children produced fewer 'meaning aspects' than the monolingual children overall, suggesting that their lexical knowledge was not as varied or interconnected as that of their monolingual peers. Additionally, while the monolingual children tended to produce more paradigmatic meaning aspects (e.g. taxonomical or superordinate categories), the bilingual children produced more syntagmatic ones (e.g. collocations or associations). The conclusions of this study are supported by Keith & Nicoladis (2013) who analysed the errors made by 20 monolingual English and 20 bilingual English-French 8 year-olds in a picture-naming task. They found that bilinguals produced more syntagmatic or 'schematic' responses (e.g. *cord* for the target 'electric outlet') compared to paradigmatic or 'categorical' responses (e.g. *hawk* for the target 'eagle'). Importantly, this group difference ceased to be significant once English PPVT scores were entered as a covariate for the bilingual group, suggesting that for bilingual children, the syntagmatic-paradigmatic shift is constrained to a greater degree by vocabulary breadth knowledge.

In Vermeer (2001), 50 monolingual and bilingual 5 year-olds were asked to define the meaning of 27 words in a breadth task, and subsequently to provide associations for 10 words in a depth task. From this latter measure, association networks were constructed for stimulus words and were assessed on a 0-3 point scale. While results generally pointed to a monolingual advantage, there were no differences in the association networks provided by the two groups. Crucially, Vermeer awarded points for non-verbal and exemplar responses during the tasks, which may have

2.1. Oral Language

benefited the performance of the bilingual children. In contrast to Vermeer's (2001) bespoke scoring approach, other studies using standardised measures do report monolingual advantages in vocabulary depth. For example, in a growth modelling study of 198 Norwegian monolingual and 90 Urdu-Norwegian bilingual 7 year-old children, Lervåg & Aukrust (2010) administered two measures of vocabulary depth knowledge, including subtests from the WISC-III and the Danish Ability Scales (Elliott, 1996). At the first time point, shortly after the onset of formal reading instruction, the monolingual Norwegian group showed large advantages on both measures of depth ($d = 1.35$ and 1.59 , respectively). In their 2003 study, Droop and Verhoeven administered an expressive vocabulary definitions measure to 8 year-old monolingual and bilingual children in the Netherlands. Results similarly showed large monolingual advantages in vocabulary depth knowledge, although somewhat smaller than those found for receptive vocabulary breadth, ranging from $d = 1.38$ to 1.85 .

More recent work has examined knowledge of multi-word phrases (MWP). Smith and Murphy (2014) designed the Multi-Word Phrase Test (MPT) in order to assess MWP knowledge among 108 children with and without EAL in English primary school Years 3 to 5. A MWP is defined as having the structure *verb + object*, such as 'break the ice' or 'pay attention'. The MPT follows a cloze procedure style in requiring test-takers to fill in a gap in a sentence by combining a verb and an object from a list of possibilities. Results of this study revealed significant group differences on background measures of vocabulary breadth (BPVS-II) and depth (Test of Word Knowledge) at every time point in favour of monolingual children. While scores on the MWP correlated moderately with these measures, significant differences in MWP mean scores were apparent only from Year 4 (age 8-9) onwards. The results of this study draw attention to the development of figurative language, which is known to correlate with reading comprehension performance, especially for bilingual learners (Oakhill, Cain & Nesi, 2016; Palmer, Shackelford, Miller & Leclere, 2006).

In summary, some studies support the existence of qualitative differences between the organisation of lexical knowledge in monolingual and bilingual children, although as shown by the results of Vermeer (2001), this finding may be influenced by the choice of vocabulary measure employed. At the time of writing, studies of language development in children learning EAL in England have not assessed vocabulary depth and as a result, it is unknown whether the consistent weaknesses of EAL learners in vocabulary breadth measures apply similarly to vocabulary depth knowledge.

2.1.2.3.4 Growth in Vocabulary Knowledge

Given that bilingual learners are likely to possess lower levels of target language vocabulary knowledge prior to and after school entry, other work has examined growth in word knowledge over time. There is some evidence to suggest that bilingual children experience faster rates of vocabulary growth than their monolingual peers. In a follow-up to their 2012 study, Hoff, Rumiche, Burrige and Ribot (2013) examined the developmental trajectory of expressive vocabulary in monolingual English and bilingual Spanish-English children up to the age of 4 years. In terms of total vocabulary growth, as measured by the MCDI, it was bilingual children with two native Spanish-speaking parents who started on the lowest intercept (i.e. level of knowledge at the first time point) and subsequently experienced a significantly faster rate of growth from 22 to 48 months. However, this result appears to be due to a relatively large increase in Spanish and not

English vocabulary knowledge. Contrary findings are reported by Leseman (2000), who found that Turkish-Dutch bilingual children exhibited a slower rate of vocabulary growth than their monolingual peers between ages 3;2 to 3;8.

Simos, Sideridis, Mouzaki and Chatzidaki (2014) assessed the receptive vocabulary knowledge of monolingual Greek and bilingual Albanian-Greek 6 to 9 year-olds. All children were assessed at five time points bi-annually on a series of language and literacy tasks including a Greek-adapted version of the PPVT. In a hierarchical linear model, vocabulary scores functioned as level-1 predictors, while language group, nonverbal ability, and parental education served as level-2 predictors. When accounting for background variables in this way, the study found that the Albanian-Greek children exhibited significantly steeper slopes, i.e. a higher rate of vocabulary growth over time: while this pattern did result in some convergence between the vocabulary knowledge of the two groups, a monolingual advantage was still evident by the end of the study ($d = 0.71$).

In an investigation of growth in vocabulary breadth and depth, Karlsen, Lyster and Lervåg (2017) assessed a sample of 191 monolingual Norwegian and 66 bilingual Urdu / Punjabi-Norwegian 5 year-olds on translated versions of the BPVS-II and the word definition subtest of the Wechsler Preschool and Primary Scale of Intelligence. All children were assessed at two time points across the transition from kindergarten to first grade. At the first time point (t1), the study found a significant monolingual group advantage on both measures of vocabulary knowledge, and due to the very similar trajectories of the two groups over time, this discrepancy remained in place by the second time point (t2), although to a slightly diminished degree (breadth: $d = 2.23$ to 1.80 ; depth: $d = 1.35$ to 1.06). Interestingly, t1 vocabulary breadth knowledge was modestly and similarly predictive of t2 depth knowledge for both groups of children.

2.1.2.3.5 Summary of Vocabulary Development

As the development of vocabulary is crucially dependent upon linguistic input, and as bilingual children necessarily receive split exposure between their languages, vocabulary continues to be a variable of high interest in bilingual development. Bilingual children often attain a similarly sized conceptual vocabulary as their monolingual peers when word knowledge from both languages is taken into account; from this it follows that their monolingual peers will continue to have the advantage in vocabulary knowledge of the target language of instruction, by virtue of a higher degree of input in that language both in and outside of school.

A review of the literature reveals great variability in the magnitude of the monolingual vocabulary advantage, with relatively large effect sizes (averaging around or above $d = 1$, but sometimes reaching much higher) for groups of children in the 7 to 10 year-old age range and who are matched on SES. Most studies report only group differences in receptive vocabulary knowledge, although there is evidence for a similar-sized effect in expressive word knowledge. Currently, there remains inconsistency in the literature regarding patterns of convergence and divergence between the two groups in receptive and expressive vocabulary, and currently there is no research on vocabulary depth knowledge in EAL learners in England.

Despite – or perhaps as a result of – a lower initial level of target language vocabulary knowledge, bilingual learners are often found to acquire vocabulary at a faster rate than their mono-

2.1. Oral Language

lingual peers, particularly in early stages of development around or prior to the onset of formal instruction. However, there is evidence that even this relatively faster rate of progress is not sufficient in order to close the gap with monolingual peers, and it is interesting to note that up to and even after ten years of education, there still exist gaps in target language vocabulary knowledge between bilingual and monolingual children.

2.1.3 Syntactic Knowledge

Syntactic knowledge represents a set of mentally-instantiated rules or constraints relating to linguistic form and meaning; using only a finite number of constituents, such rules allow for the comprehension and production of an infinite number of phrases and sentences, and are generally acquired through linguistic exposure in the natural course of language acquisition (Guasti, 2002). Syntactic development typically begins in infancy with the production of single words at around 12 months of age, followed around a year later by multi-word phrases, with the addition in the following years of more complex constructions such as *wh*-questions, inversion, and relative clauses (O'Grady, 1997). Syntactic knowledge is considered an aspect of oral language (Adlof & Catts, 2015) and plays a crucial role in aspects of literacy performance, particularly reading comprehension (discussed further in Section 2.3). A number of standardised assessments exist for the measurement of syntactic knowledge, including the oral narrative retell procedure which will be discussed specifically in Section 2.1.3.2.

2.1.3.1 Syntactic Development in Monolingual and Bilingual Children

Much like the conclusions of research into early vocabulary development, differences between monolingual and bilingual syntactic development tend to be quantitative rather than qualitative in nature (Unsworth, 2013). In the case of simultaneous bilingualism, there is evidence for a great deal of parity in the syntactic development in bilinguals' two languages as compared with that of monolinguals: specifically, bilingual infants pass through the same single- and multi-word phases as monolingual infants and ultimately attain the same level of syntactic competence (de Houwer, 1995; Meisel, 2011; Paradis & Genesee, 1996). Where differences between monolinguals and bilinguals are found, these are often as a result of differing amounts of exposure to the two languages. For instance, in their longitudinal study, Hoff et al. (2012) assessed the syntactic development of monolingual English- and simultaneous bilingual Spanish-English toddlers at three time points between ages 1;10 and 2;6. Initially, results showed a clear advantage in the direction of the monolingual group in grammatical complexity and mean length of utterance (MLU). However, when disaggregated according to degree of exposure to English, it was found that the performance of the 'English-dominant' and 'balanced' bilingual groups was indistinguishable from that of the monolinguals. In a similar fashion, Thordardottir (2015) compared the syntactic development of 3 to 5 year-old English/French monolingual and bilingual children grouped by amount exposure to each language, i.e. as being entirely monolingual, receiving more exposure to one language than another, or receiving equal amounts of exposure to both languages. For children with the least amount of exposure to English, there was a trend for lower accuracy rates in syntactic variables in English such as contracted verb forms (e.g. copulas and auxiliaries), tense, and third-person *-s*, although accuracy rates were generally fairly high across all groups. Although

these studies speak only to syntactic development in very young bilingual children, they do provide further evidence for the role of linguistic exposure in this aspect of linguistic proficiency.

Studies in England similarly offer evidence for advantages of monolingual children over their EAL learning peers in syntactic knowledge. In a randomised-controlled trial of 80 monolingual and 80 EAL learners with weak oral language skills, Bowyer-Crane, Fricke, Schaefer, Lervåg and Hulme (2017) found significant monolingual group advantages at primary school entry (age 4-5) in CELF Sentence Structure as well as the Information and Grammar scores of the Renfrew Action Picture Test (RAPT; Renfrew, 2003). By the end of Year 1 after two years of formal instruction, the monolingual group continued to significantly outperform the EAL group on all three measures: while the monolingual group advantage in Sentence Structure remained fairly stable, EAL learners appeared to close the gap to a larger extent on both measures of the RAPT by t2 (Information: $d = 0.39$; Grammar: $d = 0.49$). It is noteworthy that despite both groups having been selected due to oral language weaknesses, monolingual children still outperformed EAL learners on measures of syntactic knowledge and production after two years of formal instruction, potentially as a result of different amounts of exposure to English outside of school.

Syntactic knowledge has also been measured in older EAL learners. In her study of 9 and 10 year-old primary school children in England, Babayiğit (2014a) found a large and statistically significant monolingual advantage in comparison to an EAL group in performance on the Recalling Sentences subtest of the CELF-IV, a productive measure of syntax in which examinees repeat back increasingly long and complex sentences. Similarly, Hutchinson et al. (2003) found large and statistically significant advantages of a monolingual group in primary school Years 2 to 4 (ages 6 to 10) in relation to a group of EAL learners on performance on the Test for Reception of Grammar (Bishop, 2003b), a receptive measure of syntax in which examinees are required to identify illustrations which correspond to sentences spoken by an examiner. It should be noted that these studies focus on performance in terms of accuracy and error rates, as opposed to studies of earlier development which focus on presence or use of particular syntactic features. Nevertheless, this work does suggest a sustained monolingual advantage in both receptive and productive aspects of syntactic knowledge from the very start of formal education through to the period approaching the end of primary school in England.

2.1.3.2 Syntactic Knowledge and Oral Narrative

A narrative is a visually or orally presented sequence of interrelated events (Toolan, 2001). Oral narrative tasks, in which examinees are asked to tell or retell a narrative, typically with the aid of prompts, offer a rich source of information about children's spoken language skills, including syntactic knowledge. Although narratives vary widely in their specific content and style, there is evidence for a common underlying story structure shared across languages and cultures (Mandler, Scribner, Cole & DeForest, 1980). Oral narrative ability has received research attention due to its relation with literacy skill (Cain, 2003; Paris & Paris, 2003; Scarborough, 1990) and ability to distinguish typical from disordered language development (Allen, Kertoy, Sherblom & Pettit, 1994; Liles & Purcell, 1987).

Measures of oral narrative allow investigation into the productive use of vocabulary and syntax. At the sentence or microstructural level, individual strands within a narrative including characters and events are woven together with cohesive devices such as reference and pronominalisation,

2.1. Oral Language

coordinating and subordinating conjunctions, and vocabulary (Heilmann, Miller, Nockerts & Dunaway, 2010). Computer transcription programs such as Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012) and Child Language Analysis (CLAN; MacWhinney, 2000) readily provide measures of syntactic and lexical complexity in oral narratives, including mean length of utterance either in words (MLUw) or morphemes (MLUm), total number of different words (NDW), total number of utterances, type/token ratio (TTR), and proportion of grammatically acceptable utterances (see Section 3.4.4 for explanations of these metrics). At a thematic or macrostructural level, utterances can be classified according to their function within the narrative, ultimately contributing towards story structure and coherence. A number of different macrostructure nomenclatures have been proposed, with one of the most influential being that of the story grammar model (Mandler & Johnson, 1977; Stein and Glenn, 1975). Within this model, a narrative consists of an episode system of settings, initiating events, internal responses, plan sequences, attempts, and resolutions. Although the focus of the present study will be on narrative microstructure as a method of assessing syntax in speech, it is common in the literature for studies to explicitly contrast macro- and microstructure.

2.1.3.3 Syntax and Oral Narrative in Monolingual and Bilingual Children

In terms of narrative comprehension, children become sensitive to story structure and conventions from an early age often as a result of experiences with storytelling and book reading activities (Lynch et al., 2008; Stein & Albro, 1997). Similarly, studies support a common progression over time towards longer and more structurally organised narrative production in monolingual (Appelbee, 1978; Feagans & Short, 1984; McCabe & Rollins, 1994; Peterson & Dodsworth, 1991) and bilingual children (Muñoz, Gillam, Peña & Gulley-Faehnle; Ukrainetz et al., 2005). While narrative macrostructure is generally acquired early, microstructural elements of narrative are intrinsically tied to the acquisition of complex syntax and vocabulary, and thus have a longer developmental trajectory (Berman, 1988).

Studies that directly compare the narrative ability of bilingual children with that of their monolingual peers generally find parity in macro- but not microstructural development. Pearson (2002) compared the oral narrative ability of 80 monolingual English and 160 bilingual Spanish-English children in Grades 2 and 5, who were asked to retell the Frog Story (Mayer, 1969). Children's narratives were coded in terms of *story score* (i.e. macrostructure), as well as *language score* (including morphosyntactic accuracy and use of complex syntax). While language status did not correlate significantly with story scores, monolingual children performed significantly higher in both aspects of language score in Grade 2, but not in Grade 5 where the monolingual group maintained its advantage only in morphosyntactic accuracy. As a cross-sectional study, conclusions about developmental trajectories must be cautious, although it would appear that in this study initial weaknesses in morphosyntactic accuracy were perhaps harder to overcome than differences in use of complex syntax. Similar results were obtained in a study by Hipfner-Boucher et al. (2014), who assessed the oral narrative performance of 3 to 5-year old English monolingual and mixed-language bilingual children on the Renfrew Bus Story (Renfrew et al., 1994). Controlling for age and phonological short-term memory, it was found that bilingual children's retellings in English were characterised by significantly smaller average MLU (in words), lower lexical diversity (NDW), and fewer grammatically acceptable utterances.

As studies of bilingual language development attest, children acquiring two or more languages necessarily receive less exposure to each language, with the effects being most prevalent in group differences in vocabulary and syntactic knowledge. It would appear that this pattern of development plays a part in oral narrative development, and particularly in, but not limited to the realm of, microstructure. To the author's knowledge, no published studies have examined oral narrative development in EAL learners in England.

The review turns next to listening comprehension, the third and final skill within the domain of oral language to be considered.

2.1.4 Listening Comprehension

Listening comprehension refers to an individual's ability to understand and answer questions about aurally presented information, such as instructions or narrative passages (Hogan, Adlof & Alonzo, 2014). Listening comprehension entails not only understanding of the phonological and semantic form of individual words, but also the construction of a mental model in order to incorporate the various propositions, events, and referents within an aurally-presented passage with background knowledge (see the Construction-Integration model of Kintsch & Van Dijk, 1978 in Section 2.2.4.1.1). Listening comprehension develops in children prior to formal literacy instruction (Hogan et al., 2014), but also plays a crucial role in theoretical models of reading such as the Simple View of Reading (Gough & Tunmer, 1986, introduced fully in Section 2.2.1), in which it is conceptualised within a factor of linguistic comprehension. Alongside word-level or decoding skills, this general comprehension factor parsimoniously accounts for variance in children's ability to comprehend written passages, with reading comprehension becoming more highly dependent on linguistic comprehension over time (Gough, Hoover & Petersen, 1996; Tunmer & Chapman, 2012). That listening comprehension taps into a general discourse comprehension ability is supported by findings that children with specific difficulties in reading comprehension also exhibit difficulties in listening comprehension tasks (Catts et al., 2006; Nation, Cocksey, Taylor, Bishop, 2010; Stothard & Hulme, 1996).

Questions in listening comprehension measures tap both literal and inferential understanding, and may be administered in a receptive (e.g. a multiple choice or cloze procedure) or an expressive fashion (i.e. answering questions verbally). Although listening comprehension has been relatively little researched (Hogan et al., 2014; McKendry & Murphy, 2011), the following section will discuss relevant work on monolingual-bilingual group differences and development in this domain, including the specific effect of administration format of listening comprehension assessments on the performance of bilingual children.

2.1.4.1 Listening Comprehension in Monolingual and Bilingual Children

As in other domains of oral language discussed above, both international and U.K.-based studies report advantages of monolingual learners in relation to their bilingual peers in listening comprehension performance. In a meta-analysis of 124 independent effect sizes across 51 studies, Melby-Lervåg and Lervåg (2014) provide robust evidence for a large and significant bilingual language comprehension deficit, with a mean effect size of $d = -1.12$. Note, however, that in this study, language comprehension was defined as children's performance on vocabulary, oral cloze,

2.1. Oral Language

or listening comprehension measures, and therefore does not provide a pure comparison of listening comprehension skill between the two groups. However, this pattern is supported by other studies in the international literature. For example, in their longitudinal study of mono- and bilingual learners in Canada, Geva and Farnia (2012) report a significant monolingual advantage in Grade 5 (age 9-10) of $d = 0.45$ on the Understanding Spoken Passages (USP) subtest of the CELF-IV (see Section 3.4.3.1 for a full description of this measure). Similarly, in their longitudinal study of low-SES monolingual Dutch and bilingual Moroccan-Dutch and Turkish-Dutch 8 year-olds in the Netherlands, Droop and Verhoeven (2003) found significant and large monolingual advantages on a measure of oral text comprehension ($d = 0.82$ to 1.90) in which children were required to answer questions about orally presented stories, interviews, and conversations.

Similar findings are also reported by studies of EAL learners and their monolingual peers in England. For instance, statistically significant and moderate-to-large monolingual advantages in listening comprehension performance have been found in samples of 8 to 10 year-old children on measures including the USP subtest of the CELF-IV (Babayigit, 2014a), as well as a tape-recorded version of the Neale Analysis of Reading Ability (Neale, 1997), henceforth NARA (Burgoyne et al., 2009, 2011a; Hutchinson et al., 2003). A small number of studies report similar growth in listening comprehension skill in both mono- and bilingual learners, meaning that monolingual advantages tend to remain in place over time. Assessing EAL learners and their monolingual peers in Years 3 and 4 (ages 6 to 8), Burgoyne et al. (2011a) found significant main effects of group and time – in that EAL learners scored significantly lower, and both groups made progress between the two time points – but no significant interaction effect, as the EAL group performed consistently below the level of the monolingual group. However, such a significant interaction term was found across the three time points in Hutchinson et al. (2003), where monolingual children made a significantly faster rate of progress than their EAL learning peers between primary school Years 2 to 3, but a similar rate by Year 4. In this study, the monolingual advantage in listening comprehension increased in magnitude over time from $d = 1.28$ to 1.60 . Again, similar findings are reported amongst samples of bilingual learners in other countries wherein, for the most part, monolingual group advantages remain or increase in magnitude over time (Droop & Verhoeven, 2003; Geva & Farnia, 2012).

Listening comprehension assessments differ in the demands they place upon oral language skills. McKendry and Murphy (2011) investigated the effect of administration procedure on the listening comprehension performance of a sample of 128 monolingual and EAL learners in primary school Years 2 to 4. As in previous studies (e.g. Burgoyne et al., 2009), written passages of the NARA were tape-recorded and administered auditorially in both forced-choice and open-ended formats, alongside the similarly forced-choice listening comprehension subtest of the WIAT-II (Wechsler, 2005). Results indicated significant monolingual advantages across all three measures of listening comprehension, with EAL learners performing most poorly in relation to their monolingual peers on the open-ended format of the NARA, and least poorly on the forced-choice format. While it should be noted that neither of the measures used in this study were designed for use with bilingual populations, results do suggest that groups of children with lower oral English language skills – such as EAL learners – may be disproportionately disadvantaged on open-ended listening comprehension test formats, of which the USP subtest of the CELF is one example.

In summary, listening comprehension is a domain of oral language which plays an important role in the understanding of both oral and written language, although it has been less studied in bilingual populations. With their typically higher level of target language proficiency, monolingual children are shown to outperform their bilingual peers on listening comprehension tasks: such a finding follows logically from the lower vocabulary and syntactic knowledge of bilingual learners in relation to their monolingual peers – domains which must be drawn upon for the appropriate construction of a mental model representing propositions within an orally-presented passage (Kintsch, 1988). Again, it follows that growth in listening comprehension skill is dependent on component skills of vocabulary and grammar, and indeed, longitudinal work shows that bilingual learners tend to underperform in relation to their monolingual peers over time, with little evidence of closing this gap throughout the primary school period.

2.1.5 Summary of Oral Language Development in Monolingual and Bilingual Children

Research reviewed thus far supports the view that bilingual learners experience challenges in relation to their monolingual peers in various aspects of their oral language development in the target language, including receptive and expressive vocabulary, syntactic knowledge, and listening comprehension. Due to the similar developmental trajectories between mono- and bilingual children, early deficits in these domains are unlikely to diminish substantially over time as a result of normal participation in classroom teaching, and even where bilingual learners are found to make a faster rate of progress, this is often insufficient to close the gap. Such patterns are supported by studies of EAL learners and their monolingual peers in English primary schools, often with the additional effect of medium to high levels of social deprivation.

As discussed in Section 2.3 in the final part of this chapter, oral language skills form a critical foundation for later literacy skills: all children vary in the linguistic resources they bring to the task of literacy instruction, although this is likely to be a more difficult task for bilingual learners, many of whom receive relatively limited exposure to the target language prior to or during formal education. The following section will begin a discussion of components and constructs of literacy, focusing initially on lower- and higher-order skills involved in reading, then moving on to discuss writing development.

2.2 Literacy

The ultimate goal of reading is comprehension – to understand a linguistic message encoded in symbols. In order to understand such a message, a reader must possess minimally some ability to decode written to oral language, and some ability to construct a mental model of the text incorporating background knowledge, as referred to above in relation to listening comprehension skill (Perfetti, 1985). Reading is a relatively recently contrived cognitive activity, and is therefore acquired with a great deal of effort and in many cases, struggle (Snow, Burns & Griffin, 1998). Superficial differences between languages such as the mapping of phonemes to graphemes (see below) dictate the exact route of reading acquisition, but all writing systems pertain to the same underlying principle in their graphic representation of spoken language (Fischer, 2001). It follows,

2.2. Literacy

therefore, that a child who has not yet mastered her language orally may have difficulty becoming literate in that language. The first step in the reading process is perception and transformation of visual units to linguistic units. Once this has been achieved, general language processing is applied to the resulting representation, involving semantic and syntactic parsing and discourse processes (Perfetti, 1999; Rayner & Pollatsek, 1989).

Reading for meaning is a highly complex, strategic, and interactive process that inevitably makes recourse to a common underlying linguistic system. Crucially, the acquisition of reading is a function not only of cognitive and linguistic variables such as decoding, linguistic comprehension, working memory, vocabulary, and so on, but also of psychological-ecological factors such as motivation, teacher expectations, home environment, and cultural and linguistic diversity (Joshi & Aaron, 2000). Additional factors in the case of literacy acquisition in bilingual learners include the possibility of crosslinguistic transfer, extent of similarity between the scripts of each language, the linguistic resources that children bring from the first language in the task of acquiring literacy in a second, and societal pressures to acquire literacy in the target language (Koda, 2007; Skutnabb-Kangas, 1981).

Writing systems may be characterised according to the exact way in which they map spoken to written language units. For example, children acquiring literacy in *alphabetic* orthographies must learn to map graphemes to phonemes, while those acquiring literacy in alpha-syllabaries must learn to decode at the level of the syllable. However, certain languages pose challenges to the traditional writing system nomenclature, for instance Mandarin orthography, in which characters are considered to be simultaneously syllabic and morphemic (Perfetti, 2003). Additionally, orthographies differ according to the consistency of these mappings, which has been shown empirically to impact on word recognition processes across a range of languages (Frost, Bentin & Katz, 1987; Seymour, Aro & Erskine, 2003; Rao, Vaid, Srinivasan & Chen, 2011; Ziegler et al., 2010). For example, a consistent or transparent orthography such as that of Finnish employs an approximate one-to-one relation between phonemes and written units, while an inconsistent or opaque orthography such as that of English and French employs a number of one-to-many and many-to-one such relations.

In what follows, Gough and Tunmer's (1986) Simple View of Reading (SVR) and Paris's (2005) Constraints on Reading Skills will be introduced as frameworks within which to consider literacy development in the present study. The SVR is a framework benefitting from the empirical validation of numerous studies with both mono- and bilingual learners, and has been highly influential in reading research. Paris's (2005) Constraints on Reading, on the other hand, is lesser-known, but lends itself well to the study of literacy development among samples of children for whom amount of exposure to the target language is variable, i.e. EAL learners who necessarily receive less than 100 per cent of their linguistic exposure in English. After an introduction to both frameworks, the literature review will turn to the role of lower- and higher-level skills in reading, considering development in both monolingual and bilingual populations of learners.

2.2.1 The Simple View of Reading

Multiple component skills contribute towards the ability to read. Within the SVR (Gough & Tunmer, 1986; Hoover & Gough, 1990), a framework originally formulated in order to account for the

reading development in bilingual children, these various skills load on two primary components. Decoding (*D*) refers to the ability to convert written to spoken language, and linguistic comprehension (*LC*) refers to the process by which sentences and discourse are interpreted on the basis of lexical information (Gough & Tunmer, 1986, p.7). Both *D* and *LC* are necessary components for reading comprehension (*RC*) with neither alone being sufficient. A multiplicative relationship is proposed between the two components, where ability in each ranges from 0 (no ability) to 1 (perfect ability). This allows for the fact that with zero decoding ability, there will be no reading comprehension, no matter how good *LC*, and conversely with zero linguistic comprehension ability, there will be no reading comprehension, no matter how good a reader's level of *D*.

The SVR makes a number of testable predictions regarding reading ability and the relationship between *D* and *LC*. The framework predicts that reading difficulty may be the result of: poor decoding in the presence of adequate comprehension (i.e. dyslexia); poor comprehension in the presence of adequate decoding (i.e. a 'poor comprehender'); or a combination of the two. From this it follows that individual differences in *D* and *LC* should predict *RC*. Small but significant numbers of children present with specific difficulties in either *D* or *LC* (Catts, Adlof, & Weismer, 2006; Justice, Mashburn & Petscher, 2013).

The SVR also has support for the description of typical reading development. Large-scale studies employing structural equation modelling find that latent variables of *D* and *LC* account for up to 90% of variance in *RC* (e.g. Tunmer & Chapman, 2012; Language and Reading Research Consortium, [LARRC] 2015). However, this figure is lower in studies which measure *D* and *LC* with only one variable (e.g. Joshi & Aaron, 2000; Tilstra, McMaster, Van Den Broek, Kendeou & Rapp, 2009).

Over time, the cognitive demands of reading change as texts increase in complexity. After children have mastered the basic mechanics of literacy, good comprehension becomes more highly contingent upon higher-level skills, including vocabulary knowledge and grammatical awareness. The SVR captures this developmental shift: Gough and colleagues have shown that early on, *D* has more influence than *LC* on reading ability – at least in English – but that over time this relationship gradually reverses such that by young adulthood, reading ability is more dependent upon *LC* (Gough, Hoover & Petersen, 1996; Hoover & Gough, 1990). There is good empirical evidence for this framework, with studies finding that this correlational shift occurs as early as the third year of formal reading instruction (LARRC, 2015; Verhoeven & van Leeuwe, 2008; Vellutino, Tunmer, Jaccard & Chen, 2007).

2.2.1.1 The Role of Vocabulary in the SVR

The parsimonious nature of the SVR is both its strength and its weakness. As indicated previously, there is empirical support for the contribution and predictive power of individual differences in *D* and *LC* in reading performance. However, measures of fluency and vocabulary have been found to account for unique variance over and above that of *D* and *LC* combined. Ouellette and Beers (2010) found that vocabulary breadth, as measured by the Peabody Picture Vocabulary Test (PPVT), accounted for unique variance in *RC* amongst a group of fifty-six 12 to 13 year-old students when entered in a multiple regression equation after measures of phonological awareness and listening comprehension. A similar pattern was found in the results of Tilstra et al. (2009), in which a measure of vocabulary depth accounted for an increasing proportion of variance in *RC*

2.2. Literacy

across Grades 4, 7, and 9. Again, however, methodological considerations may go some way to explaining these results, as shown by more recent larger-scale studies (LARRC, 2015; Tunmer & Chapman, 2012). Both these studies recruited large numbers of children in Grades 1 to 3 and employed a battery of measures including the PPVT. The major advantage of structural equation modelling is its ability to incorporate latent variables and account better for measurement error (Bowen & Guo, 2012). Neither of the structural equation models supported a direct path from vocabulary to RC, but rather an indirect effect through D. Nevertheless, Tunmer and Chapman (2012) suggest that vocabulary may be best considered a component of LC rather than D or some variable in between, as vocabulary loaded more highly on LC in their study. This argument falls in with Gough and Tunmer's (1986) original formulation of the SVR in which LC is highly contingent upon lexical information.

2.2.1.2 The Role of Fluency in the SVR

A similar controversy underlies the role of reading fluency, defined as a competency which allows text to be "effortlessly, smoothly, and automatically understood" (Schreiber, 1980, p. 177). There is empirical support for the addition of a fluency component to the SVR. Within a small sample of Grade 3 children, Joshi and Aaron (2000) found that a nonword decoding measure and cloze listening comprehension task accounted for 46% of variance in RC and that the addition of rapid letter naming – as a measure of reading fluency – accounted for an additional 10% (rapid automatized naming tasks are discussed in Section 2.2.3.2). Despite providing a seemingly more parsimonious account of reading, however, it should be noted at the outset that the measures employed in this study only accounted for only 46% of variance in reading ability overall – a considerably lower figure than that reported by other studies utilising latent variables, causing one to question the adequacy of such measures. Be that as it may, the unique influence of fluency is still supported by larger scale and more methodologically rigorous studies. In both the LARRC (2015) and Tilstra et al. (2009) studies, for instance, measures of fluency (Test of Word Reading Efficiency, Torgesen, Wagner & Rashotte, 1999, and CBM Oral Reading, Deno, 1985, respectively) did make a unique contribution to RC, and there was some evidence that the influence of fluency became more important over time.

2.2.1.3 The SVR in the Reading Acquisition of Bilingual Learners

Although the SVR is supported across many studies of *monolingual* children, the framework was in fact originally formulated for the description of reading development in bilingual learners (Gough & Tunmer, 1986). In an early study, Hoover and Gough (1990) applied the SVR framework to a sample of 264 bilingual Spanish-English children from Hispanic communities in the U.S. Participants across school Grades 1 to 4 were administered measures of single-word decoding and reading and listening comprehension. Regression analyses indicated that D and LC measures accounted for significant amounts of variance in children's RC performance, with correlations rising from $r = .72$ to $.85$ from the youngest to the oldest year groups, a pattern analogous to that found in studies of monolingual participants only (e.g. LAARC, 2015; Tilstra et al., 2009; Tunmer & Chapman, 2011).

In a cross-lagged longitudinal study, Verhoeven and van Leeuwe (2012) explicitly compared the predictive power of the SVR framework across a group of 1,292 mono- and 394 bilingual learners of Dutch, who were assessed on measures of D, LC, and RC between primary school Grades 1 to 6. Although bilingual learners remained behind their monolingual peers in their performance across the three measures, results revealed very similar goodness of model fit for both groups, with RC being accounted for by D and LC, and the division of labour shifting over time as D became a weaker predictor of RC across grades, again, for both groups. Similar results are reported by Bonifacci and Tobia (2017) among a sample of primary school-age language-minority learners acquiring the transparent orthography of Italian, in which the RC performance was similarly accounted for by independent clusters of variables representing D and DC, and in Burgoyne et al. (2011a) in which listening comprehension, reading accuracy, and vocabulary knowledge equally predicted the reading comprehension performance of 7 to 8 year-old EAL learners and their monolingual peers. Finally, among a sample of 135 bilingual Spanish-English 10 year-olds in dual language education programmes in the U.S, Proctor, August, Carlo and Snow (2005) employed structural equation modelling procedures to assess the contributions of measures of D and LC to reading performance. Again, as reported in samples of monolingual children, variance in RC was accounted for directly by children's alphabetic knowledge (pseudoword decoding), listening comprehension, and both directly and indirectly through vocabulary knowledge. Although this latter finding stands in opposition to the Tunmer and Chapman (2012) study in which vocabulary contributed only indirectly to RC, it does serve to underline the importance of word knowledge within the SVR for bilingual learners as well. Indeed, in this study, vocabulary correlated strongly and significantly with both listening comprehension ($r = .85$) and reading comprehension ($r = .73$), suggesting that for bilingual learners too, vocabulary forms an important component of a general *linguistic comprehension capacity*.

In summary, the SVR is a highly influential and empirically supported framework with which to consider children's reading acquisition. Although bilingual learners are typically found to possess relatively lower oral language proficiency than their monolingual peers, the multiplicative relationship between D and LC appears to be equally predictive of reading performance in both groups of children, and as a result, the SVR will be adopted in the present study as a theoretical framework from which to consider the reading development of EAL learners.

2.2.2 Constraints on Reading Skills

According to Paris (2005), traditional reading research has failed to take account of the differing nature of component skills of decoding and linguistic comprehension, treating all of them in a fairly uniform manner. Paris instead presents an alternative framework for interpreting the conceptual and developmental nature of reading skills, arguing that these lie along a continuum ranging from most to least 'constrained'. More highly constrained skills are limited in *scope*, are typically acquired early on and to a high degree of *mastery*, are fairly *universal* to all skilled readers, and have a low degree of *codependency* with other skills. Less highly constrained skills, on the other hand, are less limited in scope, mastery, and universality, and have a higher degree of codependency with other skills. For example, letter knowledge would be considered to be highly constrained because it involves the early and rapid acquisition of a small, finite set of items to a high degree

of mastery, and this knowledge is shared by all skilled readers (of alphabetic orthographies). Additionally, fluent and accurate word recognition (discussed in Section 2.2.3.3) may be considered constrained to a degree due to its dependence on letter knowledge. Vocabulary, on the other hand, would be considered to be a much less highly constrained skill, as development of word knowledge continues across the lifespan, and total mastery is not achievable (it is impossible to know all of the words in a language). Less highly constrained skills have a larger range of influence and are more codependent on other skills: for example, comprehension is dependent upon decoding ability, metacognitive processes, vocabulary, and so on.

These qualitative differences between reading skills result in different patterns of acquisition and individual differences. Because less highly constrained skills involve a high degree of mastery, performance reaches ceiling level after a relatively brief period of acquisition (which has implications for expected patterns of growth). Performance on less highly constrained skills, however, lies along a normal distribution. Crucially, these facts have implications expectations of children's developmental trajectories and their determinants (e.g. constrained skills are likely to plateau, while unconstrained ones may grow at a relatively steadier rate). The next section will consider the nature and development of lower-level reading skills in mono- and bilingual learners, followed by higher-level skills, and finally writing development.

2.2.3 Lower-Level Skills in Reading Development

Written language differs considerably from spoken language in its design and availability of context (Perfetti, 1985). As a result, children must acquire concepts of print, such as the awareness of correspondence between written and spoken language, and that words may be further decomposed into units such as syllables and phonemes (Snow et al., 1998; Whitehurst & Lonigan, 1998). As well as this conceptual understanding, children in the early stages of learning to read in English must acquire certain skills to a full or high degree of mastery, including knowledge of the 26 letters of the alphabet, and the various relationships between these graphemes and the phonemes they represent in speech (Paris, 2005; Lesaux et al., 2008). However, the decoding of text into spoken language is a necessary but not sufficient condition of skilled reading; in addition, such a process needs to be automatised and efficient in order to free up cognitive resources dedicated to higher-order skills such as inferencing and comprehension (Perfetti, 1985; Wolf & Katzir-Cohen, 2001).

This section will focus on lower-level reading skills (i.e. those related to the rapid and/or accurate conversion of written to spoken language, as distinct from higher-level reading skills discussed in Section 2.2.4 relating to *comprehension* of written language). Here, 'decoding' will be taken to refer to the slower letter-by-letter reading strategy employed in unskilled or non-word reading (Gough & Tunmer, 1986), whereas 'sight-word reading' will be taken to refer to the faster, more automatised reading characteristic of skilled readers and real-word reading (also see discussion of the Dual-Route Model in Section 2.2.3.3). The overall term 'word recognition' will be used to refer to the general process of reading single words aloud (Aaron et al., 1999). Specifically, this section will consider the development of phonological awareness and rapid automatised naming, as well as research comparing the development of these skills in monolingual and bilingual learners.

2.2.3.1 Phonological Awareness and Orthographic Knowledge

Broadly defined, phonological awareness (PA) is the ability to reflect on and manipulate the phonological structure of language independent of meaning – it refers to a range of measurable abilities such as the identification, insertion, and deletion of phonological units such as syllables and phonemes in real or nonsense words (Goswami & Bryant, 1990; Snow et al., 1998; Wagner & Torgesen, 1987). Over time, children's PA shows a developmental progression from large to small units, as the ability to identify syllables and rimes typically precedes that of individual phonemes (Goswami, 2000; Liberman, Shankweiler, Fischer & Carter, 1974; Nunes, Bryant & Barros, 2012; Wagner, Torgesen & Rashotte, 1994). PA is a precursor skill of phonological decoding, defined as the ability to convert letter strings into words in spoken language (Kirby et al., 2008; Wagner & Torgesen, 1987). For phonological decoding to occur successfully, beginning readers must acquire the alphabetic principle, or an understanding of the correspondence between graphemes in written language and phonemes in spoken language (i.e. that the grapheme 'p' represents the phoneme /p/; Adams, 1990).

According to Share (1995), phonological recoding – the conversion of written symbols to spoken language – acts as a self-teaching mechanism, allowing the nascent reader to cope with an ever-increasing amount of unfamiliar words. The theory proposes that “word-specific and general orthographic knowledge” are acquired as a result of repeated exposures to and successful decoding of text (p.155). Additionally, because orthographic information is quickly acquired, this process becomes highly 'lexicalised' as readers become sensitive to commonly-occurring orthographic patterns such as *-ment* or *-ing* which may be processed non-phonologically. Within the self-teaching hypothesis, phonology is of primary importance early on in the establishment of grapheme-phoneme correspondence rules; however, orthography also makes an independent contribution to decoding, as beginning readers exhibit individual differences in the storage and retrieval of orthographic knowledge.

A wealth of empirical work from longitudinal and intervention studies suggests that PA is one of the most important skills in the development of word recognition in English and other alphabetic orthographies (Caravolas et al., 2012; Goswami, 2000; Liberman & Shankweiler, 1985; Muter, Hulme, Snowling & Stevenson, 2004). Strong evidence of this causal relationship is reported in a now classic study by Bradley and Bryant (1983), in which four groups of 4 and 5 year-old children received differing types of reading instruction over a period of four years. On measures of word recognition (Schonell Reading Test; Schonell & Goodacre, 1971) and passage reading (NARA), children who were taught to categorise words by their initial, medial, and final phonemes significantly outperformed comparison groups who were taught non-phonological categorisation strategies or had no instruction at all, suggesting that explicit awareness at the phoneme level specifically served to improve children's reading skills. More recent studies also support the notion that PA can be explicitly taught and that this can improve the reading skills of children with reading difficulties (Hulme, Bowyer-Crane, Carroll, Duff & Snowling, 2012; Torgesen et al., 2001).

Despite the fact that the English language employs an alphabetic script, a host of historical changes and lexical borrowings mean that English orthography is rather inconsistent in its grapheme-to-phoneme mappings (Kessler & Treiman, 2003), and as a result, a wholly grapheme-to-phoneme decoding strategy is often inappropriate for the purposes of accurate and rapid word

2.2. Literacy

recognition in English. In a longitudinal study mapping the literacy development of children in Scotland, Duncan, Seymour and Hill (2000) found that in the first year of reading instruction, children were better able to identify commonalities between words when shared segments were phonemes (e.g. FACE – FOOD) than when they were rimes (e.g. BOAT – GOAT). However, when assessed again one year later, children had improved significantly in their shared rime detection performance, indicating a higher degree of awareness of larger orthographic units. Thus, in light of earlier discussion concerning a large-to-small progression in PA, this study provides some evidence of a small-to-large unit progression in orthographic awareness in children acquiring English literacy.

2.2.3.1.1 Phonological Awareness and Orthographic Knowledge in Mono- and Bilingual Development

Bilingual learners are often found to perform similarly or more highly than their monolingual peers in tasks which tap awareness of phonological structure (August & Shanahan, 2008; Chiappe & Siegel, 1999; Geva & Yaghouh Zadeh, 2006; Oller, Pearson & Cobo-Lewis, 2007; Melby-Lervåg & Lervåg, 2014; Lesaux, 2015). Robust evidence of a relatively small difference between the two groups in PA performance comes from a meta-analysis of 51 studies by Melby-Lervåg and Lervåg (2014) which found only a small and non-significant average effect size of $d = -0.08$, in contrast to much larger and significant monolingual advantages in the domains of reading and language comprehension. Furthermore, PA is found to be similarly predictive of word recognition and reading comprehension in both populations of learners when entered into multiple regression analyses (Geva & Farnia, 2012; Lesaux & Siegel 2003), and average and poor readers are discriminated by their PA abilities rather than language status (e.g. Da Fontoura & Siegel, 1995).

Longitudinal studies provide additional evidence of the similarity between mono- and bilingual learners on assessments of PA. Lesaux, Rupp and Siegel (2007) followed a large cohort of 689 monolingual and 135 English Language Learners (ELLs) from mixed language backgrounds between kindergarten (age 5) and Grade 4 (age 9) in Canada. At both time points the two groups differed minimally and non-significantly on measures of PA including identification of rhymes, syllables, and phonemes, suggesting not only cross-sectional similarities, but also close resemblance in progress over time. Such a pattern is found in other longitudinal studies of similarly aged children, including Geva and Farnia (2012), Lesaux and Siegel (2003), Geva, Yaghouh Zadeh and Schuster (2000), and additionally in samples of older children up to age 12 (Jean & Geva, 2009).

Much of this work lacks direct comparison with EAL learners in the U.K. context, due to the measures typically employed in test batteries and the often homogeneous populations of bilingual learners, although there is some evidence to suggest parity between the two groups. Frederickson and Frith (1998) assessed the PA skills of 50 bilingual Sylheti-English 10 to 11 year-olds against those of their monolingual English-speaking peers using the Spoonerisms subtest of the Phonological Assessment Battery (PhAB; Frederickson, Frith & Reason, 1997; see Section 3.4.5.1 for a description of this measure). Results provided further evidence of similarity in the PA performance of the two groups, with a slight but non-significant monolingual advantage in spoonerism performance.

Unlike domains such as vocabulary and grammatical knowledge, PA has been found to transfer across languages, particularly where there is a degree of overlap between the phonological inventories and structures of each language (Durgonoğlu, Nagy & Hancin-Bhatt, 1993; Dickinson, McCabe, Clarke-Chiarelli & Wolf, 2004; Wang, Park & Lee, 2006). Additionally, there is some evidence that the acquisition of a second orthography may also enhance PA, and that this may be a function of orthographic transparency. For instance, in a study by Murphy, Macaro, Alba and Cipolla (2014), 7 to 9 year-old English-speaking monolingual children who received 15 hours of literacy instruction in Italian (a more transparent orthography) significantly outperformed peers who received instruction in French (a less transparent orthography) or no training at all, on measures of PA. In the case of EAL learners in the U.K., this transfer may occur for children who attend complementary language schools or madrasas, where engagement and instruction – in Latin or non-Latin scripts – may serve to enhance metalinguistic, and particularly phonological, awareness (Rosowsky, 2001).

In summary, parity in the skills and developmental trajectory of PA in mono- and bilingual children is supported by cross-sectional and longitudinal studies employing measures of phoneme identification and manipulation, and spoonerism performance. Such work often reports similar or slightly higher performance of bilingual children in relation to their monolingual peers, suggesting that such lower-level literacy skills may represent an area of relative strength for bilingual learners.

2.2.3.2 Rapid Automatised Naming

In writing, spoken language is encoded on the page or surface by converting speech sounds into graphemes; reading, therefore, requires an ability to convert these visually presented symbols back into spoken language, in a process termed phonological recoding in lexical access (Wagner & Torgesen, 1987). Rapid decoding allows the reader to combine graphemes into an orthographic string, leading to a look-up process in the mental lexicon (cf. sight-word reading; Adams, 1990; Aaron et al., 1999; Bowers & Newby-Clark, 2002). Additionally, in skilled reading, the highly automatised nature of phonological recoding frees up cognitive resources for higher-level skills involved in comprehension (Fuchs, Fuchs, Hosp & Jenkins, 2001; LaBerge & Samuels, 1974; Perfetti, 1985), and also serves to promote greater access to and enjoyment of reading experiences (Grabe, 2009). Alongside PA, fluency is recognised as a key and independent determinant of reading ability (Araújo et al., 2015; Fuchs et al., 2001), with skilled readers being able to decode text quickly, accurately, and with appropriate expression (NLP, 2000).

Rapid automatised naming (RAN) measures an individual's ability to name a series of repeating items such as letters, digits, colours, or objects as quickly as possible (Denckla & Rudel, 1976; Norton & Wolf, 2012). Note that RAN, therefore, is not equivalent to a measure of *text*-reading fluency, which is considered a separate construct (Kim & Wagner, 2015). There is mounting evidence of the significant contribution of RAN to reading skill independent of PA, although its exact role is not yet fully understood (Kirby, Parrilla & Pfeiffer, 2003; Manis, Doi & Badha, 2000; Roman, Kirby, Parrilla, Wade-Woolley & Deacon, 2009; Warmington & Hulme, 2012; Wolf & Bowers, 1993). A meta-analysis by Araújo et al. (2015) investigated relationships between RAN and reading ability in 5 to 11 year-old children across 151 studies. Moderator variables included RAN task stimuli, type of reading ability measure, grade level, and orthographic consistency. Results indicated a moderate and statistically significant correlation of $r = .43$ between RAN and reading

2.2. Literacy

ability, including with both word- and nonword-reading measures. The relatively more important role of RAN in orthographic processing is supported by other studies in which RAN tends to account for more variance in tasks requiring orthographic choice and recognition (e.g. Manis et al., 2000; Roman et al., 2009). Other findings of the meta-analysis included a significant correlation between RAN and reading comprehension ($r = .39$), RAN's relatively stronger relationship with reading fluency than reading accuracy measures, stronger relationships between RAN of letters and digits than that of colours and pictures, and RAN's relatively stronger relationship with reading measures in opaque orthographies, such as that of English. In terms of development over time, grade level was found not to significantly influence the magnitude of the RAN-reading relationship, suggesting a fairly stable correlation between the two constructs over time. Thus, results of this meta-analysis generally support the correlation of RAN and reading ability and also the apparently similar role played by RAN in real- and nonword-reading tasks.

However, there is some disagreement in the literature concerning the exact role of RAN as a concurrent and longitudinal predictor of reading ability. Although the meta-analysis of Araújo et al. (2015) found no significant change in the magnitude of the relationship between RAN and reading ability according to grade level, individual studies do report changing relationships over time. Kirby et al. (2003) assessed the contribution of RAN (of pictures) and PA to word-reading ability in an unselected sample of 161 kindergarten children who were followed until Grade 5 (age 10-11). The two predictor variables showed opposite developmental trajectories, with PA accounting for most variance between kindergarten and Grade 2, and RAN becoming a significant predictor only after this point. In contrast, a study by Wagner et al. (1997) utilising structural equation modelling found that a latent RAN variable (digits and letters) accounted for significant variance in word reading ability in early not but later stages of reading acquisition, as the autoregressive effect of prior word reading ability began to account for more variance over time. This study also found that the contribution of RAN decreased in magnitude once letter-name knowledge was included in the model, mimicking the results of Manis et al. (2000) and suggesting that RAN may play a relatively more important role in orthographic processing.

In summary, there is growing evidence for the significant contribution of RAN to constructs of reading ability, including in word recognition and passage reading measures. Although its exact mechanism is yet to be fully understood, studies suggest that, in alphabetic and opaque orthographies such as English, RAN may play a relatively larger role in orthographic processing than PA (Araújo et al., 2015; Kirby, Desrochers, Roth & Lai, 2008), particularly in the *retrieval* of phonological information (Warmington & Hulme, 2012). Studies supporting the increasingly important role of RAN throughout reading development are consistent with the self-teaching hypothesis of Share (1995), as readers come to rely more heavily on orthographic knowledge over time.

2.2.3.2.1 Rapid Automatised Naming in Mono- and Bilingual Development

There is some research to suggest that, similar to performance on PA tasks, bilingual learners exhibit advantages relative to their monolingual peers in rapid naming skill. For instance, Geva and Farnia (2012) tracked the development of 390 mixed-language ELL and 149 monolingual English-speaking children in Canada between Grade 2 (age 7-8) and Grade 5 (10-11). At both time points, the ELL group significantly outperformed the monolingual group in RAN of letters,

with the magnitude of this advantage increasing slightly over time as bilingual children were able to name stimuli increasingly more quickly than their monolingual peers. Although U.K.-based studies often do not include measures of RAN, there is some evidence for the higher RAN performance of EAL learners in England, with Sylheti-speaking EAL learners in Frederickson and Frith (1998) significantly outperforming their monolingual peers in RAN of digits.

Whether RAN may be considered a relative strength of bilingual learners may depend upon on children's stage of literacy acquisition, amount of exposure to the target language, and type of RAN task. For instance, in contrast to the findings discussed above, young mixed-language bilingual learners in the beginning phases of formal literacy instruction and in some cases with a minimum of only four months of exposure to the target language have been found to underperform in relation to their monolingual peers across different types of RAN tasks, including RAN of digits, letters, and objects (Chiappe & Siegel 1999; Geva, Yaghoub Zadeh, & Schuster, 2000; Lesaux & Siegel, 2003). However, there is also work supporting the relatively faster progress made by bilingual learners on RAN tasks. For instance, Lesaux and Siegel (2003) found such a pattern in RAN of letters between kindergarten and Grade 2, while Jean and Geva (2009) report significantly faster progress of a group of bilingual learners relative to their monolingual peers between Grades 5 and 6.

In summary, although RAN has been less thoroughly investigated than PA in studies comparing the development of mono- and bilingual learners, there is some evidence for a relative strength of bilingual children in their rapid naming of letters and digits but not objects. Furthermore, this relative advantage may be developmentally constrained, as groups of bilingual learners tend to perform on a par or below the level of monolingual children in the earliest stages of literacy acquisition, but outperform them later as a result of a faster rate of progress over time.

2.2.3.3 Word Recognition

As discussed above, beginning readers must 'crack the code' by becoming aware of the correspondence between written symbols and oral language. Over time, children are exposed to an 'orthographic avalanche' in the form of a large volume of unfamiliar words (Share, 1995), and therefore cannot rely exclusively on rote association or simple grapheme-phoneme correspondence rules, especially in the acquisition of an opaque orthography such as that used in English (Kessler & Treiman, 2003). An overview of some models of word reading is provided below, beginning with dual-route and connectionist models. Subsequently, two particularly influential developmental phase theories of word reading will be introduced, and finally, studies investigating the development of this skill in mono- and bilingual children will be discussed.

In the dual-route model of Coltheart, Curtis, Atkins and Haller (1993), skilled word recognition is achieved through one of two paths: a lexical route by which real words are looked up in an orthographic lexicon (cf. sight-word reading), and a non-lexical route by which nonwords are decoded via grapheme-phoneme correspondence rules (cf. decoding). The model has been shown to provide a good approximation of adult performance in various psycholinguistic tasks, including latency for the reading of irregular words, caused by competition between the two routes of the model (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001), and is supported by neuroimaging research showing functional separation of the lexical and non-lexical routes in the brain (Jobard, Crivello & Tzourio-Mazoyer, 2003). Although predicated on skilled adult reading, the dual-route

2.2. Literacy

model also has developmental applications. For instance, the reading difficulties of some children can be classified as phonological or surface dyslexia².

A connectionist approach, on the other hand, models word recognition as a set of “cooperative and competitive interactions among large numbers of simple neuron-like processing units” (Plaut, 2005, p.25). Such connections are distributed across many units and are weighted through a process of learning. Models such as those of Plaut, McClelland, Seidenberg and Patterson (1996) are populated by groups of input and output units (e.g. those which encode orthography, phonology, or semantics), and groups of ‘hidden’ units which mediate between the two. Connectionist models have been able to emulate the learning process of word recognition. The parallel-distributed processing model of Harm and Seidenberg (2004) was able to approximate the development of division of labour over time: while activation along a phonological pathway accounted for a large degree of performance in word reading accuracy rates early on in the model’s training, this effect levelled over time, with the orthography to semantics pathway becoming more important.

While much research in this area has attempted to model skilled word recognition from a monolingual perspective, the connectionist Bilingual Activation Model (Dijkstra & van Heuven, 1998) explicitly models reading accuracy in the presence of two languages. Again, a full discussion of this work is beyond the scope of this thesis, however the model is able to account for slower response times in naming tasks where there exist a high number of phonologically or orthographically similar neighbours in a second language. For example, in a lexical decision task, response times increased for the French word *gens*, which has as one of its English neighbours *guns* (Bijeljac-babic, Biardeau & Grainger, 1997; van Heuven, Dijkstra & Grainger, 1998).

In contrast to dual-route and connectionist approaches, developmental models seek to describe the phases in which word reading develops and evolves in children acquiring the ability to read – in particular, such models generally converge in describing the reading process before and after acquisition of the alphabetic principle. Discussion for the present purposes will be limited to two particularly influential models; namely those of Frith (1985) and Ehri (1995).

Frith’s (1985) phase model consists of three distinct strategies in the development of word recognition. In the earliest phase of reading, a *logographic* strategy allows children to instantly access the pronunciations of familiar words based on visual cues. Later, growing PA and knowledge of grapheme-phoneme correspondence rules results in an *alphabetic* strategy, allowing for the decoding of unfamiliar and nonsense words. Finally, an *orthographic* strategy allows rapid conversion of words into orthographic units (e.g. morphemic units such as *-ing* or *-ment*). Although strategies within this model are said to adhere to a strict sequential order, ‘breakthrough’ to the following phase occurs only when strategies are merged (e.g. recognition of larger orthographic units cannot be attained without the ability to analyse words into their constituent parts; at the same time however, grapheme-phoneme conversion is necessary when confronted with unfamiliar or nonsense words).

Ehri’s (1995) model contributes to that of Frith (1985) with the addition of a fourth phase. Much like in Frith’s (1985) model, children in the *pre-alphabetic* phase utilise salient visual cues in order

²Phonological dyslexia refers to particular difficulties with the reading of nonwords (reliance on the lexical route; e.g. where *zint* cannot be looked up in a mental lexicon), while surface dyslexia is characterised by particular difficulty in reading orthographically irregular words (reliance on the non-lexical route; e.g. *island* → */izland/*, Coltheart, 2005). depending on selective impairment in the lexical or non-lexical route, respectively (Coltheart, 2005).

to sight-read words. However, unlike in Frith's model, children are then said to progress onto a *partial* alphabetic, and then a *full* alphabetic phase, allowing for the finding that children often make some grapheme-phoneme correspondences before they are able to fully decode unfamiliar words. Finally, in the *consolidated alphabetic* phase, readers map larger written- and spoken-language units such as morphemes and syllable clusters.

In summary, models of skilled and nascent word reading serve to illustrate the processes by which phonological recoding occurs and the changing divisions of labour between phonological, semantic, and orthographic information over time according to the familiarity and consistency of written words. The review now turns to studies of word recognition skill and development in mono- and bilingual learners.

2.2.3.3.1 Word Recognition in Mono- and Bilingual Development

Similar to other lower-level reading skills discussed above, word recognition is often identified as an area of relative strength for bilingual learners, with similar or significantly higher performance in relation to their monolingual peers (August & Shanahan, 2008; Cline & Shamsi, 2000; Oller et al., 2007). Such a pattern has also been found in samples of EAL learners in England. Bowyer-Crane et al. (2017) present data from a large randomised controlled trial in England of 80 EAL and 80 monolingual learners recruited at school entry (mean age 4;7) who exhibited weak English oral language skills. Children were assessed on a battery of language and literacy measures across two time points, including the Early Word Recognition subtest of the York Analysis of Reading Comprehension (YARC) Early Reading (GL Assessment, 2011) in Reception year, and the Diagnostic Test of Word Reading Processes (Forum for Research in Literacy and Language, 2012) in Year 1, after two years of formal literacy instruction. EAL learners significantly outperformed their monolingual peers on both word-reading measures at both time points, suggesting early-emerging strengths of EAL learners in this aspect of reading. Such bilingual advantages have also been found in older samples of EAL learners in England, for instance, 7 to 8 year-olds in Burgoyne et al. (2009; 2011a) who also significantly outperformed their monolingual peers in word recognition skill (WRAT-3). Together, these results hint at the early emergence of strengths for bilingual learners in word recognition – including those children with weaknesses in oral language – and there is some evidence that such advantages are maintained until a later educational stage.

However, the robustness of a bilingual advantage in word recognition skill is questioned by a number of studies which find either equivalent performance between mono- and bilingual learners, or slight and non-significant bilingual advantages across measures including the WRAT-3 (Chiappe & Siegel, 1999; Geva & Farnia, 2012; Geva et al., 2000; Lesaux, Rupp & Siegel, 2007; Jonejan, Verhoeven & Siegel, 2007; Lesaux & Siegel, 2003), the Test of Word Reading Efficiency (Lervåg & Aukrust, 2010), and the Single Word Reading Test (Babayiğit, 2015). Additionally, a meta-analysis of 79 studies by Melby-Lervåg and Lervåg (2014) indicated a small but statistically significant monolingual advantage in word recognition ($d = -0.12$). Interestingly, this effect was mediated by geographical location, with bilingual learners in Canada outperforming their monolingual peers, but the opposite pattern for studies from Europe and the U.S, hinting at the important influence of the role of educational and cultural factors.

2.2. Literacy

Studies report mixed results concerning mono- and bilingual learners' progress in word recognition skills over time, with some reporting very similar rates of progress (e.g. Burgoyne et al., 2011a; Geva et al., 2000; Limbird, Maluch, Rjosk, Stanat & Merkens, 2014), and others reporting divergence over time, with bilingual learners overtaking their monolingual peers (e.g. Droop & Verhoeven, 2003; Geva & Farnia, 2012). For instance, in their longitudinal study of low-SES monolingual Dutch and bilingual Turkish-Dutch and Moroccan-Dutch children, Droop and Verhoeven (2003) report initially similar performance between the groups in word recognition skill at the beginning of Grade 3, but a widening bilingual group advantage by the end of Grade 4, particularly in orthographically simple real words, but also in more complex polysyllabic real words. Finally, in their study of children between Grades 2 and 5 in Canada, Geva and Farnia (2012) found a trend for a slightly faster rate of development in bilingual learners' word recognition performance on the Woodcock Reading and Mastery Test (WRMT; Woodcock, 1987), although the performance of the two groups was not statistically significantly different by Grade 5.

2.2.3.4 Summary of Lower-Level Reading Skills in Mono- and Bilingual Development

In contrast to target language vocabulary and syntax, bilingual learners often exhibit strengths relative to their monolingual peers in aspects of lower-level reading skills including PA, RAN, and word recognition. In some cases, letter knowledge and PA may transfer across children's languages, and proficiency in another orthography may serve to further enhance PA (for example, the case of many EAL learners in England who attend mosque and learn to read Qur'anic Arabic; Hirst, 1998; Rosowsky, 2001). The literature is more equivocal with regard to developmental trajectories in lower-level reading skills, with some showing similar rates of development between mono- and bilingual learners, and others indicating relatively faster growth.

The strengths of many bilingual learners in word recognition may be seen as facilitative of their reading development; however, with reference to the SVR (Gough & Tunmer, 1986), even perfect decoding ability will not result in a good level of reading *comprehension*; indeed, reading comprehension weaknesses may still be expected in populations of bilingual children, such as EAL learners in England, who exhibit weaknesses in aspects of English oral language. The next section will consider higher-level skills in reading development, particularly in terms of reading comprehension and its contributory skills.

2.2.4 Higher-Level Skills in Reading Development

The ultimate aim of reading is comprehension. After successful word recognition, relationships with actions, agents, and their intentions are encapsulated in a mental model of a passage (Graesser et al., 1994; Kintsch, 1988). Children bring both decoding and linguistic comprehension skills to the task of reading from the beginning of reading instruction; however, in line with predictions of the SVR (Gough & Tunmer, 1986; Section 2.2.1), the relationship between these skills changes over time such that decoding explains less variance, and linguistic comprehension explains more variance in reading comprehension skill. Therefore, it follows that as children progress through formal education, reading tasks will come to place higher demands on domains of oral language such as vocabulary, syntax, and listening comprehension.

After a brief overview of cognitive architecture and mental models, the current review will turn towards higher-order skills involved in reading: as processes involved in skilled reading comprehension are numerous, such a review is necessarily brief, focusing on inference generation and use of prior knowledge. Following this, the review will turn towards research investigating differences in reading comprehension performance and development in mono- and bilingual learners. Section 2.2.5 will consider writing development.

2.2.4.1 Cognitive Architecture and Mental Models

Comprehension is underpinned by an array of different cognitive and linguistic resources, leading to the observation that “virtually everything that logically can be identified as a component of comprehension has been identified as a source of comprehension failure” (Perfetti, Marron & Foltz, 1996, p.140). Broadly, in order to comprehend a text, a reader must build a mental representation of the information presented within that text and often make inferences or ‘fill in gaps’ where information is presupposed but not necessarily explicitly stated. Two particularly influential models of comprehension are discussed briefly below.

2.2.4.1.1 The Construction-Integration Model

The Construction-Integration model (Kintsch & van Dijk, 1978; Kintsch, 1988) describes a two-part process during reading comprehension. Firstly during the construction phase, a representation is formed of the words and propositions as they appear in the text. This *textbase* is formed of microstructural elements such as the interrelationships between arguments, and macrostructural elements such as global themes or topics. Secondly, this textual representation is integrated with the reader’s background knowledge and personal experiences to form a *situation model*. The high flexibility of the construction phase allows for the activation of a large array of propositions and their interconnections. In a connectionist fashion, a resultant over-specified set of activated knowledge is then subject to pruning, after which only the most closely associated items in the knowledge net remain activated (Gernsbacher & Foertsch, 1999). As a result, inference generation is said to be a rather passive and uncontrolled process, in contrast to other accounts that view this process as active and goal-oriented (e.g. Graesser, Singer & Trabasso, 1994). For deep and rich comprehension to occur, a reader must go beyond the literal information presented in the text by making recourse to information that is missing but implied or that depends on prior knowledge of a particular topic, genre, or vocabulary. It follows that difficulties in reading comprehension may arise due to either poor integration between propositions and referents within the textbase, to limitations in knowledge required to go beyond the literal presentation of the text (including vocabulary knowledge), or to lack of integration between the textbase and the situation model.

2.2.4.1.2 Structure Building Framework

Gernsbacher (1990) proposes a general model of discourse comprehension processing in which comprehenders are driven to construct mental representations that are coherent and hierarchical. The theory holds that mental structures are established according to first-mentioned information such as first sentences, settings, characters, and so on. Any subsequently presented information

2.2. Literacy

that is consistent with prior information in terms of temporality, spatiality, causality, and reference, will then be mapped to the existing structure. In contrast, inconsistent or unrelated information will initiate a shift, and result in the construction of a new (sub-)structure.

Other important mechanisms in the construction of a coherent mental structure include suppression and enhancement. Under Gernsbacher's theory, comprehenders should activate and retain only information that is relevant for the process of coherent structure building and discard or suppress irrelevant information. In one psycholinguistic experiment, Gernsbacher, Varner and Faust (1990) asked young adults to read short sentences ending in either an ambiguous word (e.g. *spade*) or unambiguous word (e.g. *shovel*), and then to decide whether a target word matched the meaning of the sentence (e.g. *ace*). The key variable of interest was participants' response times, and in particular, the amount of interference caused by ambiguous words. It was found that while skilled comprehenders initially showed a significant amount of interference, this had decreased dramatically after a short (750ms) interval, due to suppression and isolation of the appropriate word meaning. Less-skilled comprehenders, on the other hand, failed to show any reduction in interference at all. These results point towards a poor suppression mechanism as a causal factor in comprehension difficulties: since less-skilled comprehenders are less likely to suppress irrelevant information, they may shift to new structures unnecessarily.

In summary, mental models of comprehension describe individuals' continual search for meaning and how different skills and sources of information serve to build up a representation of discourse. Crucially, comprehension processes are brought about not only through a mental representation of a text itself, but also its incorporation with prior knowledge. As a result, reading comprehension tasks may pose difficulties for many EAL learners who possess potentially fewer linguistic or cultural resources to draw upon when reading for meaning. The review will now turn towards research on the skills and knowledge that are crucial predictors of the ability to comprehend written language. Although such research has typically been conducted on monolingual populations, studies involving bilingual learners will be discussed where appropriate.

2.2.4.2 Inference Generation

The ability to generate inferences is considered to be the hallmark of skilled reading comprehension. Elaborative inferences are those that embellish or add information to the text, while cohesive or coherence references establish links within the text and are considered to be both necessary and sufficient for adequate comprehension (Garnham, 1982; Hulme & Snowling, 2009; Yuill & Oakhill, 1991). There is debate in the literature as to what kinds of inferences and how many inferences are generated, as well as the point at which they are generated, during reading. On one hand, a constructionist approach states that inference generation is strategic and based on the reader's previous experiences and goals, which may be more or less highly specified (Graesser et al., 1994). In contrast, a memory-based approach suggests that information presented within a text will activate the reader's background knowledge regardless of any goals she may have; analogous to Kintsch's (1988) construction phase, this results in a passive but fast process of activation of an associative net of knowledge in long-term memory (Gerrig & McKoon, 1998). According to the constructionist approach, certain types of inferences are more likely to be generated on-line (i.e. during the process of reading), such as referential, thematic, and case structure role assign-

ment (e.g. the agent of an action), while other types such as causal consequence and author's intent are more likely to occur off-line due to their costly processing nature (Graesser et al., 1994).

Inference generation is found by a number of studies to be an important and in some cases causal factor in skilled reading comprehension. Cain and Oakhill (1999) assessed inferencing ability and reading comprehension among groups of skilled, less-skilled, and younger comprehension-age matched (CAM) monolingual English-speaking children. After reading short passages, participants were asked questions requiring literal, gap-filling, and text-connecting information while the text was obscured. The results of the study provide important insights into the role of inferencing in comprehension: while the groups did not differ on questions requiring literal information, the CAM group outperformed the less-skilled group in answering text-connecting questions, suggesting that failure to generate inferences is a cause rather than consequence of reading comprehension skill. An interesting observation of inferencing studies is that children who are less skilled than their peers are found, in fact, to be capable of *generating* inferences and to possess the knowledge required to make inferences (Barnes, Dennis & Haefele-Kalvaitis, 1996; Cain, Oakhill, Barnes & Bryant, 2001; Cain & Oakhill, 1999). Failure to generate inferences, therefore, may be indicative of a differential strategy use during reading.

Relatively little work has investigated reading comprehension processes among bilingual learners who are acquiring literacy in a second or additional language. However, when faced with the decontextualised language of written passages, all children, regardless of language learning background, are necessarily required to engage in inferencing and to make connections where they are not explicit in the text, often by incorporating their background knowledge. Similarly, given that the SVR is shown by some work to be equally predictive of the reading comprehension performance of both groups of children (e.g. Bonifacci & Tobia, 2017; Verhoeven and van Leeuwe, 2012), it stands to reason theoretically that both will be ultimately constrained by the same underlying skills, namely decoding and linguistic comprehension. A recent cross-sectional study by D'Angelo and Chen (2017) confirms this supposition in suggesting that inferencing too plays an important role in the reading performance of bilingual children. In this study, 62 monolingual English and 83 bilingual English-French 10 and 11 year-olds in Canada were grouped as good, average, or poor comprehenders. Crucially, poor comprehenders in both language groups performed significantly below the good and average comprehenders on a task requiring inferencing based on passages of the Gates-MacGinitie, suggesting that this skill is integral to adequate reading comprehension processes in mono- as well as bilingual learners.

2.2.4.3 Prior Knowledge

Studies have examined the roles of availability and accessibility of knowledge in the process of reading comprehension. Barnes et al. (1996) attempted to control for the confounding effects of background knowledge. In their study, 6 to 15 year-old typically-developing monolingual children were taught a novel knowledge-base consisting of twenty facts about the fictional planet 'Gan' such as 'the frogs on Gan glow in the dark'. After ensuring a minimum threshold of correctly retained information, children were presented with 10 one-paragraph passages and asked questions which tapped elaborative inferences (embellishment of content not necessary for comprehension), coherence inferences (bridging gaps to illustrate understanding of tacit information), recall of literal information, and interpretation of similes. With age as a between-subjects factor,

2.2. Literacy

it was found that even after controlling for availability of knowledge, older children comprehended to a higher level and made more inferences than younger children.

The results of this study have been replicated elsewhere. Cain et al. (2001) employed the 'Gan' paradigm amongst a group of skilled ($n=13$) and less-skilled ($n=13$) comprehenders aged 7-8 years. The two groups of children were matched on chronological age and word reading accuracy. Similar to the poor comprehender profile discussed above, in this study, 'less-skilled' comprehenders were selected as those who exhibited lower-than-expected reading comprehension as predicted by their word reading ability. The skilled comprehenders generated significantly more elaborative inferences ($d = 0.96$), and there was also a pattern for a higher occurrence of coherence inferences ($d = 1.3$). Although the skilled comprehenders acquired and retained the knowledge-base at a better rate than their less-skilled peers, this was found not to account for differences in inference generation or comprehension skill. The results of these studies suggest that while background knowledge is necessary for comprehension, it may not be sufficient. It is striking that in both studies, children who failed to score highly on inferencing ability nevertheless did recall the requisite knowledge-base facts, but failed to integrate them in order to maintain coherence. The results also accord with suggestions made by the Structure Building framework of Gernsbacher (1990) in that while less-skilled comprehenders adequately processed all necessary background information, they did not engage in mechanisms of suppression and enhancement in order to sift relevant from irrelevant information necessary for inference-making.

The results of a study by Burgoyne, Whiteley and Hutchinson (2011b) suggest that this conclusion may be warranted similarly for children learning EAL. In this study, also employing the 'Gan' paradigm, sixteen 8 to 9 year-old children learning EAL were paired with monolingual peers matched on chronological age and reading accuracy. Results indicated that although both groups of children acquired and recalled the knowledge-base to criterion level, children learning EAL performed more poorly overall on comprehension ($d = -1.14$) and on questions tapping literal information and similes. One interesting observation of the study was that EAL children did make both elaborative and coherence inferences despite their relatively lower reading comprehension performance. Burgoyne et al. (2011b) also found large discrepancies in receptive and expressive vocabulary knowledge between the groups, and offer anecdotal evidence that children in the EAL group tended to adopt a key-word strategy when answering questions, in contrast to monolingual children who often answered from memory.

In summary, these results suggest that alongside lower levels of oral language proficiency, EAL learners may experience additional difficulties in the extraction and integration of various sources of information in texts. While the artificiality of the 'Gan' paradigm may be criticised, it is also the case that EAL learners underperform in relation to their monolingual peers on traditional measures of reading comprehension which do not rely on a pre-taught knowledge base (discussed further below), suggesting the presence of a general reading comprehension weakness in this population of learners.

2.2.4.4 Passage Reading Skill in Monolingual and Bilingual Development

This section will discuss studies comparing mono- and bilingual learners on measures of passage reading, in which examinees are typically required to read aloud short passages and answer questions tapping literal understanding and inference generation. Although this section focuses

on higher-level skills in passage reading comprehension, reference to passage reading accuracy will also be made, as progress on passage reading assessments is often determined by the number of accuracy errors made.

Melby-Lervåg and Lervåg (2014) conducted a meta-analysis of 57 studies involving comparisons between mono- and bilingual learners on passage reading measures. Overall, the analysis found a statistically significant monolingual advantage in reading performance ($d = -0.62$), which was particularly more pronounced in measures requiring the reading of full passages ($d = -0.78$) as opposed to sentences only ($d = -0.43$). A similar pattern has been found to apply across a range of different bilingual learner populations in different countries, including among a homogeneous sample of Urdu-Norwegian bilingual children at school entry (age 7) in Norway (Lervåg & Aukrust, 2010), predominantly Spanish-English 5th graders (age 9-10) in the U.S (Geva & Farnia, 2012), and Turkish-Dutch and Moroccan-Dutch bilingual 8 to 10 year-olds in the Netherlands (Droop & Verhoeven, 2003). In addition, the reporting of this finding across a range of different passage reading measures (e.g. NARA, WRMT, Gates-MacGinitie, and bespoke measures) provides further support for the existence of a monolingual group advantage.

Similarly, reading comprehension advantages of monolingual learners over their EAL learning peers in England have been found across a range of assessments, including the Suffolk Reading Scale (Beech & Keys, 1997), the NARA (Babayiğit 2014a; Burgoyne et al., 2009, 2011a; Frederickson & Frith, 1998; Hutchinson et al., 2003; Rosowsky, 2001; Stuart, 2004), and the YARC (Babayiğit, 2015; Bowyer-Crane et al., 2017). Additionally, such a pattern is found regardless of matching between the two groups of children on amount of English-medium educational experience. For instance, studies by Frederickson and Frith (1998) and Babayiğit (2014a) both report a monolingual passage reading comprehension advantage on the NARA relative to EAL learning peers who had also received the same amount of instruction in English ($d = 0.95$ and 0.85 , respectively).

In a cross-sectional study of 46 monolingual and 46 EAL learners in primary school Year 3 (age 7-8), Burgoyne et al. (2009) exposed one weakness of the NARA, a commonly used assessment in such studies. As progress on the NARA is determined solely by the number of *accuracy* errors, children with strong decoding skills are able to attempt more passages, and thereby attempt more comprehension questions. Indeed, as the EAL learners in this study significantly outperformed their monolingual peers in passage reading accuracy, they were able to attempt more passages and therefore obtained comprehension scores similar to those of the monolingual group. However, a significant monolingual reading comprehension advantage was obtained ($d = 0.49$) when accuracy scores were entered as a covariate in an ANCOVA. These results accord with those of other studies, in that weaknesses of EAL learners in passage reading comprehension are often found in the presence of relative strengths in reading accuracy (Babayiğit, 2014a; Bowyer-Crane et al., 2017; Burgoyne et al., 2011a; Frederickson & Frith, 1998; Hutchinson et al., 2003; Rosowsky, 2001), bearing resemblance to profiles of monolingual 'poor comprehenders' (Catts et al., 2006). The Burgoyne et al. (2009) study does highlight a significant issue with the NARA, however, and it may be questioned to what extent this pattern is present in other studies utilising this measure, which did not covary passage reading accuracy scores (e.g. Hutchinson et al., 2003; Babayiğit 2014a), potentially underestimating monolingual-EAL group differences in reading comprehension skill.

2.2. Literacy

Some longitudinal work has also assessed the development of passage reading skill in bilingual learners. For instance, when assessed over an 18-month period, Urdu / Norwegian bilingual children in Lervåg and Aukrust (2010) exhibited lower passage reading comprehension initially at school entry (age 7), but also made a slower rate of progress over time than their monolingual peers, particularly on the NARA. Similarly, across the two-year testing period of Droop and Verhoeven (2003), monolingual Dutch children maintained an advantage in their 'text coherence' performance which remained fairly static over three time points for the Turkish-Dutch group (around $d = 0.4$) but which did decrease in magnitude for the Moroccan-Dutch group ($d = 0.96$ to 0.73).

Again, such a pattern is confirmed by studies of EAL learners in England. In their longitudinal study of EAL and monolingual learners from primary school Years 2 to 4 (ages 6 to 8;11), Hutchinson et al. (2003) found consistent and significant monolingual group advantages on passage reading comprehension using the NARA. Although both groups of children made significant progress from year to year, EAL learners were not able to close gaps in reading comprehension performance by Year 4 ($d = 0.77$). Finally, somewhat similar results are supplied by Burgoyne et al. (2011a), who followed the reading development of EAL and monolingual learners between Years 3 and 4 (ages 7 to 9). In this study, as in those described above, in both years EAL learners significantly underperformed their monolingual peers on passage reading comprehension on the NARA. However, instead of the group convergence found in other work, this gap actually increased over time from $d = 0.77$ to 1.31 , perhaps as demands on the children's reading comprehension skills increased.

Taken together, both sets of international and U.K.-based studies discussed here suggest that passage reading comprehension – but not passage reading accuracy – remains a relatively challenging domain for 7 to 10 year-old bilingual learners from the point of school entry up until later stages of primary education. Analogous to other domains of oral language such as vocabulary, syntax, and listening comprehension (Section 2.1), even where there is evidence of convergence between the groups, this is not sufficient to close the gap over time as a result of regular classroom instruction. Adopting the theoretical standpoint of the SVR (Gough & Tunmer, 1986), this pattern may be understood as a result of EAL learners' relatively lower levels of oral language proficiency in English. Given the multiplicative relationship between decoding (D) and linguistic comprehension (LC) in the prediction of reading comprehension (RC), EAL learners' relative strengths in D are not able to compensate for weaknesses in LC, as no matter how high the ability to decode words fluently, reading comprehension will ultimately be constrained by the ability to understand what has been read, a skill highly dependent on LC. As discussed in Chapter 1, a number of EAL learners may receive little exposure to English prior to school entry, and thus their underperformance relative to their monolingual peers at this point, who have received *only* exposure to English, may be expected. Furthermore, a constraints on reading framework (Paris, 2005) is able to account for EAL learners' (at least) equivalent performance in more constrained skills of reading such as letter knowledge and word recognition, which differ qualitatively and quantitatively in their pattern of acquisition (e.g. scope and mastery; Section 2.2.2) to the more unconstrained skills such as vocabulary and syntax which underpin linguistic comprehension. The final section will now go on to discuss mono- and bilingual development in writing.

2.2.5 Writing

Written language differs from spoken language in terms of its physical and social design: as opposed to the shared context and co-constructed meaning of spoken language, writing is necessarily decontextualised and one-way in its direction of communication (Perfetti, 1985). Moreover, written language has a range of concepts and conventions that make the process of writing more than the simple transcription of spoken language: it can be distinguished from speaking both by what it lacks – for example the encoding of suprasegmental features such as prosody – and by what it introduces – for example punctuation and boundaries between words and sentences (Olsen, 1994; Perfetti, 1985). This section will discuss some seminal research on writing development in monolingual children, before going on to consider comparative studies with bilingual learners. Similar to oral narrative discussed in Section 2.1.3.2, writing will be considered in this thesis as a vehicle through which to study children's spelling and expressive syntax skills.

A number of studies have investigated the process of written composition in an effort to model the writing process (Abbott, Berninger & Fayol, 2010). An early study by Juel, Griffith and Gough (1986) is often cited as the first to produce a 'simple view of writing', consisting of two sets of skills in written composition, namely spelling and the ability to generate ideas (or 'ideation'). Earlier models of writing tended to downplay the contribution of transcription factors (handwriting and spelling) as these models were based on adult writing in which mastery and automatising of transcription are already achieved; however, transcription has been shown to be a significant predictor of quality of early writing skill in children (Berninger et al., 1992; Hayes, 2012). In the work of Berninger (2000), the writing process is modelled as a triangle framework with text generation at its apex, and transcription skills and executive functions on each of its vertices. Executive functions incorporate processes of, for example, planning, reviewing, revising, and self-regulation.

As in studies of reading ability, writing is considered to consist of multiple components which are differentially related to oral language skills. In Juel et al.'s (1986) seminal study, a sample of primary school-age children was followed between Grades 1 and 2 and assessed on a battery of reading and writing measures. Both spelling and ideation were found to be significant predictors of writing quality, but interestingly, the balance shifted over time such that spelling became less correlated and ideation became more correlated with writing quality, analogous to the shift found in studies of the SVR relating to the developmental shift between decoding and comprehension (Hoover & Gough, 1990). In a large longitudinal study involving students from Grades 1 to 7, Abbot et al. (2010) found consistent and stable relationships over time between individual differences in spelling and written composition as measured by the WIAT II Written Language subtests (Psychological Corporation, 2001). Interestingly, significant relationships were more likely to be found in the direction of word-level skill (particularly spelling) to text-level composition. This study supports the contribution of word- and text-level factors in written composition, but did not find evidence of a developmental shift in the correlation between and importance of each skill over time.

Written expression has also been found to be a sensitive index of language difficulties (Bishop & Clarkson, 2003; Cragg & Nation, 2006). Cragg and Nation (2006) investigated group differences in written composition and spelling between 10 year-old children with or without specific compre-

hension difficulties (a typical 'poor comprehender profile' defined as performance of at least -1 SD on NARA reading comprehension but adequate decoding skill). The two groups of children did not differ in terms of spelling ability or in length or complexity of written narratives, but the compositions of the poor comprehenders were significantly less well-structured and contained fewer ideas. These results are supportive of an early study by Juel (1988) showing a strong connection between difficulties in reading and writing; specifically, that poor readers tend to become poor writers. There is a clear parallel between poor comprehenders and poor writers in that these groups of children do not appear to have difficulties in the mechanical aspects of the task (decoding in the case of reading, and transcription or spelling in the case of writing), but rather in the other, cognitive processes of comprehension or ideation.

2.2.5.1 Writing Development in Monolingual and Bilingual Children

The writing of monolingual and bilingual children is ultimately constrained by the same relationship between transcription and composition factors (Babayiğit, 2014b; Silverman et al., 2015). Depending on the specific language combination in question, and where children are literate in both of their languages, there may be crosslinguistic transfer effects and particular benefits of biliteracy, for example in understanding of conventions of print or grapheme-phoneme correspondence (Bouchereau & Gort, 2012; Koda, 2007). Silverman et al. (2015) investigated relationships between various components of oral and written language in a sample of monolingual English and bilingual Spanish-English speaking students aged 8 to 11 years. Participants were asked to write a story using a picture prompt, which was scored in terms of contextual conventions (spelling and grammar) and story composition (organisation, characters, theme, and vocabulary). In addition, all participants were assessed on measures of expressive grammar, morphology, and vocabulary. Results showed advantages of the monolingual group in both contextual conventions and story composition, although it was the bilingual group that showed an advantage in single-word spelling.

There is a small amount of research on the writing abilities of EAL learners in the U.K. Babayiğit (2014b) recruited 94 monolingual and 72 EAL learners in Year 5 (average age 9;7), with the children in the EAL group having received a minimum of 3 years of English-medium instruction. The WIAT-II written expression subtest served as the primary measure of writing, in which examinees are allowed 20 minutes to write two paragraphs based on given prompts. This assessment yields a score for writing quality based on organisational, lexical, and holistic assessment. Other measures in the study included single-word spelling and spelling error rate. Results shared similarity with those of Silverman et al. (2015) in that the monolingual group obtained a higher writing quality score ($d = 0.69$); however, both groups performed similarly in both single-word spelling error rate in prose writing. Babayiğit (2014b) also utilised structural equation modelling to determine the relative contributions of component skills to the writing quality of both groups of children. A composite latent variable of 'verbal' skills including vocabulary, working memory, and semantic fluency contributed significantly to writing quality to very similar degrees in both groups ($\beta = .33$ and $.32$ for mono- and EAL, respectively), with a similar pattern applying to the role played by word-level skills in writing (e.g. spelling and single-word reading). As a result, this study provides evidence to suggest that the writing quality of mono- and EAL learners is constrained to an equal degree in both groups by children's verbal and word-level writing skills.

Finally, an in-depth study of writing in EAL is provided by Cameron and Besser (2004), who used U.K. national curriculum criteria to assess the written composition of 'advanced bilingual learners' in English schools, i.e. those with a minimum of five years of residence in the U.K. A total of 264 pupils with and without EAL in KS2 (age 8–11) and KS4 (age 14–16) were asked to write short expository and narrative texts which were analysed for composition quality and use of language. Of particular interest is the analysis concerning differences in the mechanics and syntax in the writing of the pupils with and without EAL: overall, pupils with EAL made a higher number of agreement errors, used fewer 'advanced subordinators' (e.g. *as soon as*, *while*, *until*), omitted more prepositions, and used shorter verb phrases, but made fewer spelling errors than their monolingual peers. Particularly, omission of prepositions and errors in the use of articles were found to be salient and unique features of writing in the EAL group, present only to a far lesser degree in the writing of monolingual pupils. Finally, it is also interesting to note that the number of syntactic errors had decreased considerably in the writing of pupils with EAL by KS4.

In summary, these studies provide evidence for a particular profile of strengths and weaknesses in the writing of EAL learners in relation to their monolingual peers. Namely, EAL learners appear to possess similar or better spelling ability than their monolingual peers, but their lower levels of vocabulary and grammar knowledge result in lower ratings of written narrative composition. Evidence of this is provided by studies showing that the writing of both monolingual and EAL learners is equally constrained by transcription and composition factors, and studies finding lower levels of English oral language knowledge of EAL learners (Babayigit, 2014b; Silverman et al., 2015).

2.3 The Contribution of Oral Language to Reading Comprehension

Section 2.1 discussed the development of vocabulary, syntax, and listening comprehension in mono- and bilingual learners. Although this thesis does not explicitly investigate the predictive role of oral language in reading development, a brief examination of the literature in this area is warranted in order to demonstrate the inherent connections between the two domains, and the importance of oral language development particularly for EAL learners. This brief review will focus on the contribution of oral language proficiency to passage reading comprehension performance specifically, as this has been found to be an area of difficulty for EAL learners (Section 2.2.4.4). The following sections will discuss the roles of vocabulary, syntax, and listening comprehension in passage reading comprehension performance, firstly from a theoretical perspective based on studies with monolingual learners, and secondly from a comparative perspective based on studies contrasting the role of these skills in both mono- and bilingual groups of children.

2.3.1 The Role of Vocabulary, Syntax, and Listening Comprehension in Reading Comprehension

Generally, as noted in the introduction to this chapter, the role of oral language proficiency in reading comprehension is supported by studies showing that children identified as poor comprehenders also perform poorly on measures tapping vocabulary, syntax, and listening comprehension

2.3. The Contribution of Oral Language to Reading Comprehension

(Adlof & Catts, 2015; Clarke et al., 2014; Nation et al., 2010; Stothard & Hulme, 1996). Other than stating their generally important status in predicting reading ability, until this point the literature review has not considered the specific roles of these oral language skills in bringing about reading comprehension. This section will briefly discuss the roles played by vocabulary, syntax, and listening comprehension in reading comprehension.

There has long been a recognised link between size of vocabulary and ability to comprehend text. The causal status of this relationship has received attention and it has been debated whether it is word knowledge in and of itself that brings about comprehension, or whether it is merely a proxy for a different skill set, such as verbal aptitude or general conceptual knowledge (Anderson & Freebody, 1981). Vocabulary knowledge has been described as the ‘critical link between decoding and comprehension’ (Joshi, 2005, p.209), as words are the basic building blocks of the larger propositional structures which form the textbase (Kintsch, 1988). As explicated by the triangle framework (Plaut et al., 1996) and the lexical quality hypothesis (Perfetti, 2007), lexical items contain phonological, semantic, and orthographic representations which may be more or less highly specified. Thus, word recognition and comprehension skill both tap into vocabulary knowledge, leading Perfetti and Stafura (2014) to label the lexicon as a ‘pressure point in the [reading] system’ (p.26). A weakness in vocabulary knowledge, and specifically in the semantic pathway, is therefore likely to have a deleterious effect upon reading comprehension.

For sufficient comprehension to occur, it is necessary for readers to comprehend a certain proportion of the lexical material in any given text (Carver, 1994; Freebody & Anderson, 1983), with figures ranging from 95 to 98 per cent (Laufer, 1989; Schmitt, Jiang & Grabe, 2011). There is burgeoning evidence that individual differences in vocabulary knowledge correlate with and predict performance in reading comprehension tasks (Cunningham & Stanovich, 1997; Muter, Snowling & Hulme, 2004; Ouellette, 2006; Ricketts, Nation & Bishop, 2007; Roth, Speece & Cooper, 2002; Vellutino, Scanlon, Small & Tanzman, 1991).

Cain and Oakhill (2014) looked specifically at the contribution of vocabulary knowledge in the ability to comprehend literal information and to draw inferences from text. They found among a sample of 10 to 11 year-old children that vocabulary depth was significantly predictive of the ability to draw global inferences when entered in a multiple regression equation after age, word reading accuracy, and vocabulary breadth. Indeed, the ability to draw inferences depends on procedural or schematic knowledge of the world, and such a finding is therefore supportive of Anderson and Freebody’s (1981) *knowledge hypothesis*, on account of vocabulary knowledge acting as a proxy for general conceptual or world knowledge.

The role of syntactic knowledge in predicting reading comprehension has also been explored, albeit to a lesser extent. As introduced in Section 2.2.4.1.1, a key process in reading comprehension is the establishment of a textbase (Kintsch, 1988) which involves the integration of individual words and phrases into propositional and hierarchical structures. Readers must use syntactic knowledge to draw together connections between propositions and referents in order to maintain coherence. Over the course of an academic career, children are likely to encounter texts which make use of increasingly complex grammatical constructions; it follows therefore that syntactic

knowledge³ will become an increasingly important factor in reading skill over time. There is evidence that knowledge of syntax significantly predicts and correlates with reading comprehension performance in children, often after controlling for other variables such as vocabulary and working memory (Bowey, 2005; Cain, 2007; Cutting & Scarborough, 2006; Goff, Pratt & Ong, 2005; Muter et al., 2004).

Finally, in line with observations regarding the role of vocabulary and syntax, listening comprehension is also found to make a significant and unique contribution to reading comprehension performance in samples of monolingual children (Lepola, Lynch, Laakkonen, Silvén & Niemi, 2012; Ouellette & Beers, 2010; Tilstra et al., 2009; Vellutino et al., 2007; Verhoeven & van Leeuwe, 2008). In line with predictions of the SVR, in their longitudinal study of children between Grades 1 and 6, Ouellette and Beers (2010) found a changing role of listening comprehension over time. In Grade 1, listening comprehension accounted for a small but significant 2.5% of variance in reading comprehension, with this rising to 5.8% in Grade 6 when entered into a regression analysis after phonological awareness and decoding. In a much larger-scale study, Verhoeven and van Leeuwe (2008) followed a large cohort of 2,143 monolingual Dutch-speaking children who were assessed on word decoding, vocabulary, and listening and reading comprehension each year between school entry and Grade 6 in the Netherlands (although reading comprehension was measured only from Time 2). The authors examined cross-lagged associations in order to measure longitudinal relationships between variables after taking account of autoregressive effects. Again, in line with predictions of the SVR, results indicated reciprocal relationships between listening and reading comprehension between Grades 1 and 5; however, the introduction of vocabulary into the final model altered this pattern substantially, as listening comprehension ceased to covary significantly with reading comprehension after Grade 3, and entered into a reciprocal relationship with vocabulary itself across Grades 2 to 6. Thus, while reading comprehension depended on listening comprehension in the early stages of reading acquisition, vocabulary contributed more in later stages. This finding is commensurate with Ouellette and Beers (2010), who found a significant contribution of vocabulary breadth to later but not earlier reading comprehension performance. Taken together, these results suggest that listening comprehension does play a central part in predicting reading comprehension, but that some dissociation occurs in later stages of reading acquisition, in which measures of vocabulary capture more variance in reading comprehension.

In summary, vocabulary, syntax, and listening comprehension are shown to play important roles in monolingual children's reading comprehension performance. The ability to understand ideas, propositions, and relationships in written language is highly dependent upon the establishment of a mental model to represent the text, and the availability of vocabulary and syntactic knowledge. While this review has focused on only the three oral language skills discussed in Section 2.1, it should be noted that a range of other capabilities contribute towards reading comprehension, including story schema knowledge and the ability to monitor ongoing comprehension (Cain et al., 2004; Oakhill et al., 2003).

³A key distinction must be maintained between syntactic *knowledge*, i.e. understanding of rules pertaining to number, tense, agreement, thematic roles, and so on, and syntactic *awareness*, i.e. a metalinguistic skill requiring reflection on linguistic structure (Cain, 2007).

2.3.2 The Role of Oral Language in the Reading Comprehension of Mono- and Bilingual Learners

In general, studies tend to converge on the finding that vocabulary, particularly, plays a relatively more important role for the reading comprehension of bilingual learners than that of their monolingual peers. In their longitudinal study of mono- and bilingual learners of Norwegian, Lervåg and Aukrust (2010) found that vocabulary measured at school entry (age 7) was a unique and significant predictor not only of all children's initial reading comprehension performance (intercept), but also of their rate of growth over time (slope); additionally, for the bilingual group, vocabulary was a stronger predictor of intercept and slope on a translated version of the NARA. Droop and Verhoeven (2003) investigated the influence of vocabulary and morphosyntax on the reading comprehension of monolingual Dutch and bilingual Turkish-Dutch and Moroccan-Dutch children across three years. Analogous to the results of Verhoeven and van Leeuwe (2008) discussed above, this allowed the examination of cross-lagged longitudinal relationships between variables: while Time 2 vocabulary correlated significantly with later reading comprehension for both groups, the strength of this association was higher for the bilingual group, for whom the effect of vocabulary operated on reading comprehension directly and indirectly through listening comprehension, which itself was a stronger and more consistent predictor of reading comprehension for the bilingual group. Interestingly, morphosyntax did not play any role for bilingual learners, correlating only with listening and not reading comprehension. A similar pattern is reported by Geva and Farnia (2012), who regressed the reading comprehension scores of mono- and bilingual fifth graders (age 10-11) on a range of oral language measures. This study found that previously and concurrently measured syntactic skill (CELF-IV Formulated Sentences), as well as listening comprehension (CELF-IV Understanding Spoken Paragraphs) failed to predict reading comprehension of ELLs, but conversely did predict that of their monolingual peers. In an opposite pattern, however, vocabulary knowledge was found to be a significant predictor of reading comprehension in the ELL group only. Taken together, results of such studies suggest that oral language proficiency is important in predicting the reading comprehension of both groups of children, although results do hint at a relatively more important role for vocabulary knowledge among samples of bilingual learners.

Oral language skills are also shown to be a crucial predictor of reading comprehension in children learning EAL in England. In her study of 56 monolingual and 69 EAL-learning 9 and 10 year-olds, Babayiğit (2014a) regressed reading comprehension scores (NARA) on a range of oral language and reading measures, as well as dummy-coded language status (monolingual or EAL). The model yielded significant interaction terms between language status and vocabulary ($\beta = .28$), as well as language status and morphosyntax ($\beta = .27$), pointing to the relatively more important role of vocabulary knowledge in the reading comprehension performance of EAL learners. However, the robustness of this finding is questioned by a subsequent study employing the large sample technique of structural equation modelling (Babayiğit, 2015). In this follow-up study, a group of similarly-aged mono- and EAL learners were compared in terms of strength of association between a latent oral language variable (including receptive vocabulary, sentence repetition, and verbal working memory) and reading comprehension as measured by the YARC. While the EAL group exhibited a trend for a stronger association than the monolingual comparison group (r

= .86 and .72, respectively), this difference was not statistically significant. Taking a slightly different approach, Burgoyne et al. (2009) assessed the specific roles of receptive and expressive vocabulary knowledge in reading comprehension. Results from a regression analysis in this study found that, while both types of word knowledge accounted for similar amounts of total variance in the reading comprehension performance of mono- and EAL-learners, expressive vocabulary was uniquely predictive in the EAL group. Finally, in their longitudinal study, Hutchinson et al. (2003) regressed Year 4 reading comprehension scores on the NARA on a range of oral language measures administered in Year 2 with groups of mono- and EAL learners. Although models accounted for adequate amounts of reading comprehension variance in both groups, neither listening comprehension nor receptive grammar was found to be a significant predictor for either group when entered after a Year 2 reading comprehension autoregressor. However, similar to the results of Burgoyne et al. (2009), expressive vocabulary was found to be a significant predictor of reading comprehension in the EAL group only. The findings of these two latter studies may be explained from the perspective that measures requiring an expressive response format are considered a more sensitive measure of oral language proficiency (cf. McKendry & Murphy, 2011 in Section 2.1.4.1) and thus expressive vocabulary scores may have explained more variance than other assessments requiring receptive knowledge or non-expressive response formats.

In summary, this section has explored some of the mechanisms through which oral language skills bring about comprehension. A key theme emerging from this brief review is similarity between mono- and bilingual learners in the roles that vocabulary, syntax, and listening comprehension play in reading comprehension performance, but differences in the relative strength of these factors (particularly in vocabulary knowledge). As noted in Section 2.2.4.4, the task of extracting meaning from text does not differ markedly between the two groups of learners; however, given that EAL learners are more likely to experience difficulties in domains of oral language than in decoding, it follows that the reading comprehension of these children will depend to a greater extent on their oral language proficiency.

2.4 Summary of Literature Review I and Aims of Longitudinal Study

Building upon a foundation of oral language, children begin the process of literacy acquisition. After early mastery of lower-level reading skills, oral language comes to the fore as a particularly important determinant of the ability to extract meaning from text. In the case of typical development, monolingual children who are exposed to the target language from birth will have a more or less robust foundation upon which to build literacy skills. In contrast, children who acquire more than one language – especially in successive fashion – face the dual task of becoming literate in a language they may or may not yet have mastered orally, and of making a faster rate of progress in order to catch up with their monolingual peers who have a head start (NALDIC, 1999). Furthermore, the ability to access curriculum content requires an increasingly high level of CALP throughout schooling (Cummins, 1981), which makes higher demands on general language comprehension skills, including vocabulary, syntax, and listening comprehension.

2.4. Summary of Literature Review I and Aims of Longitudinal Study

Despite considerable heterogeneity, children learning EAL in England on average do not perform as highly as their monolingual peers on national assessments of reading and writing (Strand et al., 2015). Studies suggest that it takes upwards of five years for bilingual children to begin to approximate their monolingual peers in their higher-order, decontextualised language skills (Collier, 1987; Demie, 2013). Therefore, it is of interest whether, after four years of formal education in English, children learning EAL in English primary schools begin to perform on a par in relation to their monolingual peers on the oral language skills which ultimately feed into literacy skill. As reflected in the National Curriculum (DfE, 2013), primary school Years 3 and 4 (when pupils are aged 7-9) mark a transition from learning to read to reading to learn (Chall et al., 1990), where pedagogy shifts from an emphasis on decoding and fluency to vocabulary knowledge and sentence structure.

Longitudinal studies of EAL learners in England are rare, and where conducted, often do not match EAL learners with their monolingual peers on amount of English-medium instruction received (e.g. Hutchinson et al., 2003; Burgoyne et al., 2011a). Thus, a key question is whether and to what extent children with EAL are still 'catching up' with their monolingual peers in language and literacy skills after both groups of children have received an equal amount of formal instruction in English. The major research questions addressed in the first part of this thesis (longitudinal cohort study) concern the developmental trajectory of language and literacy skills in children who are learning EAL and their monolingual peers in the run up to the end of primary school. Specific questions include:

1. **What are the similarities and differences in the language and literacy skills of children learning EAL and their monolingual peers at the beginning of Year 4 (t1)?** That is, at t1, after a minimum of four years of English language instruction, how do the two groups compare in their English vocabulary knowledge, listening comprehension, expressive grammar, oral narrative, phonological processing, single-word reading efficiency, passage reading, and writing skills?
2. **What are the developmental trajectories of the two groups of children in language and literacy skills between Year 4 (t1) and Year 5 (t3)?** That is, do the two groups make comparable rates of progress over time, and additionally, where apparent, do discrepancies in performance change in magnitude over time such that the groups converge or diverge?

Chapter 3

Methods I: Longitudinal Study

This chapter will provide details concerning the methodology of the longitudinal cohort study, including design and recruitment of participants, measures employed, language questionnaires, and general procedure.

3.1 Design and Recruitment

The study adopted a longitudinal design consisting of three time points over 18 months in order to follow developmental trajectories of children's language and literacy skills between the beginning of Year 4 and middle of Year 5 (see Figure 3.1 overleaf). The study was carried out in Sheffield in collaboration with Sheffield City Council and in particular, members of the ESCAL (Every Sheffield Child Articulate and Literate) team who identified schools with a mixed enrolment of monolingual and bilingual children that were likely to engage with the project. At the time of recruitment in 2015, 22.1% of primary school pupils in Sheffield were classified as learning EAL, closely resembling the national average of 19.4%.

After ethical clearance was granted from the University of Sheffield Human Communication Sciences department (see Section 3.2 below for more information regarding ethical considerations), schools were approached in writing and over the telephone for participation in the study. Out of the 22 primary schools shortlisted by ESCAL, nine agreed to take part in the project, with one school subsequently withdrawing. The final eight schools differed as to their proportion of EAL learners, ranging from 10.5% to 91.6%, and in the proportion of pupils receiving Free School Meals (FSM), a widely-used metric of social disadvantage, ranging from 22.5% to 53.6% (DfE, 2015; Table 3.1).

School staff were asked to identify children in Year 4 (age 8-9 years) as potential participants in the study. Participation criteria were as follows: firstly, no statement of Special Educational Needs or receipt of additional intervention from a speech and language therapist or educational psychologist (this decision was taken as a result of the study's focus on typical language development); secondly, for children with EAL, receipt of English-medium education since at least Year 1 (age 5-6; in order to account for a potentially confounding effect of unequal amounts of English language instruction¹). The intention of the design was to recruit roughly equal numbers of monolingual and EAL learners in order to maximise comparability across the groups, however this was not always possible (see Table 3.1 for a breakdown of population statistics of recruited schools).

¹More information concerning the educational experience of EAL learners is discussed in reference to the parental language questionnaire in Section 3.5

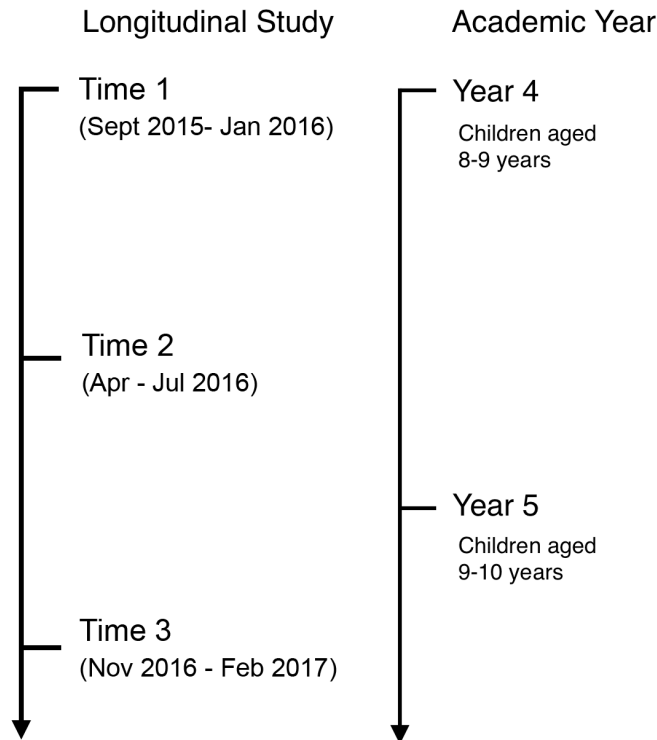


Figure 3.1: Timeline of the Longitudinal Study Relative to School Academic Year

Once pupils were identified by school staff as eligible to take part, information sheets and consent forms were sent home to parents/carers via participating schools (see Appendix 3.1 on page 253). Those who consented for their children to take part were asked to return the signed consent form to the school where they were collected by the researcher. Parents/carers were also asked to give consent to be contacted about the possibility of their child taking part in a sub-project (intervention study, Chapter 6). Parents/carers were informed that this phase of the study was not mandatory and children's participation in the intervention study would not affect their taking part in the longitudinal study. Recruitment began officially in the summer term of the 2014/15 academic year but continued into the autumn and winter of the 2015/16 academic year.

3.2 Ethical Considerations

Ethical clearance for the longitudinal study was granted in March 2015 by the University of Sheffield Human Communication Sciences department. This included considerations related to recruitment, length and wording of consent forms, and assessment battery procedures (see Section 3.6). Recruitment was facilitated for the most part by schools' dedicated EAL coordinators. Year 4 class teachers were provided with consent forms and asked to distribute these to pupils who met participation criteria. Parents who agreed for their children to participate were asked to return signed consent forms to the school, where they were collected by the researcher (in some cases, teachers discussed the project in person with children's parents.) As well as written informed consent from children's parents, verbal assent was also received from all children before testing began.

3.3. Participants

Table 3.1: School Characteristics at t1

School	School Size	Children recruited for present study (n)		School characteristics at t1		
		Monolingual	EAL	EAL (%)	IDACI (decile)	FSM (%)
1	469	2	2	10.5	0.57 (1)	53.6
2	519	15	4	22.5	0.34 (2)	22.5
3	439	1	8	77.7	0.42 (1)	30.8
4	316	2	4	83.9	0.33 (2)	41.1
5	254	0	10	82.0	0.30 (2)	35.1
6	444	6	4	17.5	0.39 (1)	40.3
7	275	2	11	91.6	0.29 (3)	29.8
8	553	5	5	13.6	0.42 (1)	35.0
		Totals		National Averages		
		33	48	19.4	0.17	16.5

Note: School size indicates total number of children on roll at time of recruitment; IDACI = Income Deprivation Affecting Children Index based on school postcode, where a decile of 1 represents the highest and 10 represents the lowest level of deprivation; FSM = Percentage of pupils eligible for Free School Meals (DfE, 2015a).

3.3 Participants

All children were recruited from eight primary schools in the city of Sheffield. At Time 1 (t1) the sample consisted of 33 monolingual children (14 male; mean age 8;8, SD = 3.4 months) and 48 children learning EAL (22 male; mean age 8;8, SD = 3.3 months). The two groups of children did not differ significantly in age ($t(79) = -.49, p = .630, r = .043$) or in gender distribution ($\chi^2(1) = .092, p = .760$) at t1. Both metrics of deprivation presented in Table 3.1 (IDACI and FSM)² indicate a moderate to high levels of social deprivation in participating schools relative to the national average.

Children's EAL status was ascertained from school records as well as from child and parent questionnaires. All parental questionnaires returned by parents of children with EAL indicated the presence of at least one language other than English in the home environment (see Section 3.5 for a summary of questionnaire results), although it should be noted that this is based only on the 36 out of 48 questionnaires that were returned (a return rate of 75%). Children in the EAL group spoke a total of 15 languages, including Punjabi (n=11), Arabic (n=10), Urdu (n=7), Bengali (n=5), Farsi (n=2), Polish (n=2), Turkish (n=2), Amharic (n=1), German (n=1), Hungarian (n=1), Nepali (n=1), Pushto (n=1), Somali (n=1), Thai (n=1) and Tigrinya (n=1).

²IDACI scores are based on government statistics collected in 2015 (DfE, n.d). The rank is based on the postcode of each school as opposed to the home address of each participating child, as this information was not collected. See Basit, Hughes, Iqbal and Cooper (2015) for a similar strategy for measurement of socio-economic status.

3.4 Measures

This section outlines the measures that were employed in the longitudinal study (summarised in Table 3.2 below). Measures were selected according to language and literacy constructs of interest and developmental appropriateness, as well as for comparison with studies of similar designs and research questions (e.g. Babayiğit, 2014a; Bowyer-Crane et al., 2017). Below is a brief description of each measure, including norming and reliability information where available (note that standardisation of reliability across measures is not possible here, as manuals report different types of coefficients, e.g. Cronbach's α , split-half, test-retest, confidence bands, and so on).

Table 3.2: Summary of Measures by Type and Time Point

Variable Group	t1	t2	t3
Non-Verbal Reasoning and Memory	WISC-IV MR CELF-IV NR	CELF-IV NR	CELF-IV NR
Vocabulary	BPVS-III CELF-IV EV WISC-IV VC	BPVS-III CELF-IV EV WISC-IV VC	BPVS-III CELF-IV EV WISC-IV VC
Other Oral Language	CELF-IV USP CELF-IV FS	CELF-IV USP CELF-IV FS	CELF-IV USP CELF-IV FS
Oral Narrative	Peter and the Cat	Peter and the Cat	Peter and the Cat
Phonological Processing	PhAB Spoonerisms CTOPP RDN & RLN	PhAB Spoonerisms CTOPP RDN & RLN	PhAB Spoonerisms CTOPP RDN & RLN
Literacy	TOWRE-2 YARC Writing task	TOWRE-2 YARC Writing task	TOWRE-2 YARC Writing task

Note: WISC-IV = Wechsler Intelligence Scale for Children-IV; MR = Matrix Reasoning; VC = Vocabulary (definitions); CELF-IV = Clinical Evaluation of Language Fundamentals-IV; NR = Number Repetition; EV = Expressive Vocabulary; USP = Understanding Spoken Paragraphs; FS = Formulated Sentences; BPVS-III = British Picture Vocabulary Scale-III; PhAB = Phonological Assessment Battery; CTOPP = Comprehensive Test of Phonological Processing; RDN = Rapid Digit Naming; RLN = Rapid Letter Naming; TOWRE-2 = Test of Word Reading Efficiency-II; YARC = York Assessment of Reading Comprehension-Primary.

3.4.1 Non-Verbal Reasoning and Memory

3.4.1.1 Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV^{UK}; Wechsler, 2003) – Matrix Reasoning

Measures of non-verbal reasoning and memory were included in the battery to ensure fair comparison between the two groups of children and to help rule out differences in performance on

3.4. Measures

language and literacy measures due to other cognitive factors. Matrix Reasoning is a measure of perceptual reasoning, and is considered a reliable estimate of general intellectual nonverbal ability. In this subtest, examinees are presented with increasingly difficult, partially complete sequences of coloured geometric shapes and are asked to select the appropriate pattern from a choice of five to complete the sequence. Testing proceeds backwards in the event of two consecutive incorrect scores until two consecutive correct answers are given. Testing discontinues after four incorrect responses in any consecutive sequence of five. The WISC-IV is normed on a nationally representative U.K. population of 780 children aged 6;0 to 16;11 years. The Matrix Reasoning subtest has a high split-half reliability coefficient for the age 8-9 year groups of $r = .92$ and an overall average of $.89$ across all age groups.

3.4.1.2 Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-IV; Semel, Wiig, & Secord, 2006) – Number Repetition

The Number Repetition subtest of the CELF-IV was employed as a measure of working memory. The Number Repetition subtest of the CELF-IV requires examinees to recall and recite random digit sequences of increasing length, e.g. 6-2-5-8. Each verbally presented item comprises two parts of equal length, a and b: testing discontinues when incorrect responses are given for both parts of an item. In the first part of the task, examinees recite digits in the same order that they are presented (Numbers Forwards; a measure of short-term memory) while in the second, they are required to do so in the reverse order (Numbers Backwards; a measure of working memory).

The CELF-IV UK is normed on a nationally representative U.K. population of 871 individuals aged between 5;0 and 16;11. Although proportions of participants for whom English is an additional language are not reported, individuals from ethnic minorities comprised 12.4% of this sample, and assessments were administered only to children who could “speak and understand English” (Semel et al., 2006, p.203). Although internal consistency reliability for Number Repetition is not reported for the 8;0 – 8;11 age group, a ‘best estimate’ coefficient α of $.81$ can be derived by averaging consistency coefficients for adjacent age groups for Number Repetition Total, comprising Numbers Forwards and Numbers Backwards (Semel et al., 2006, p. 219).

3.4.2 Vocabulary Measures

Assessments in the test battery included a measure of receptive vocabulary knowledge (British Picture Vocabulary Scale-III), expressive vocabulary knowledge (CELF-IV Expressive Vocabulary), and vocabulary depth knowledge (WISC-IV Vocabulary). Each assessment is detailed below.

3.4.2.1 British Picture Vocabulary Scale-III (BPVS-III; Dunn, Dunn & NFER, 2009)

The BPVS-III is a measure of receptive vocabulary breadth knowledge in which examinees are presented with a series of test plates containing four colour pictures and are asked to identify the picture corresponding to a stimulus word spoken by the examiner. All stimuli were selected for their similarity within a U.K. context, appropriateness for all ethnic groups, and historical relevance.

During administration, a basal set is established when the examinee makes no more than two errors within the first set of 12 items, from which point testing continues forwards. Testing discontinues once a total of 8 errors has been made within any set of 12 responses (ceiling set). The starting point is staggered according to chronological age: however, if examinees make more than two errors within the first set administered, testing proceeds backwards until a basal set is established.

The BPVS-III is normed on a nationally representative U.K. population of 3,238 individuals aged between 3 and 16 years. Although this norming population does include children with EAL, this number is very small ($n=45$; 1.39%) and no information is provided about the English language abilities of these children. The manual reports a slightly below average standard score for this group of learners within the norming population, at 93.2. Reliability for the BPVS-III is reported in terms of confidence bands for seven separate raw score bands, due to the fact that not all examinees in the standardisation process attempted exactly the same items. The resulting 95% confidence bands are asymmetric in nature and range from 5 to 13 for lower bands and from 4 to 13 for upper bands, suggesting relatively more uncertainty about the performance of particularly low- and high-ability examinees. No statistics are reported for test-retest reliability or internal consistency. However, the PPVT-4, the assessment on which the BPVS-III is based, reports a high split-half reliability coefficient of $r = .94$ (Dunn & Dunn, 2007).

3.4.2.2 CELF-IV Expressive Vocabulary

The Expressive Vocabulary subtest of the CELF-IV measures an examinee's ability to name illustrations depicting objects, people, and actions in a range of categories including animals, occupations, sports, science, communication, shapes, and so on. During administration of this measure, the examinee is presented with a colour illustration and asked 'what is this?' For items referring to part/whole objects, the examiner points to the object to be named, for example, a stamp on an envelope or a branch on a tree.

Responses that are semantically accurate are awarded the maximum of two points, while 1-point responses are those that accurately describe the activity or object but do not include the necessary vocabulary, for example 'instrument' for the target item 'saxophone'. Acceptable 1-point responses are provided for 11 of the 27 items on the subtest. Semantically inappropriate responses receive a score of 0 and testing discontinues after 7 consecutive scores of 0. The CELF-IV Expressive Vocabulary subtest has an internal reliability coefficient of $\alpha = .80$ for the 8;0-8;11 and 9;0-9;11 age bands, with an average coefficient of $\alpha = .83$ across all age groups.

3.4.2.3 WISC-IV^{UK} – Vocabulary

The Vocabulary subtest of WISC-IV^{UK} is designed to measure depth of word knowledge, abstract thinking, and long-term memory. The subtest consists of 36 verbally-administered items of increasing difficulty, with the starting point staggered according to chronological age; additionally, for examinees aged 9 years and above, written words accompany verbal items. Target words include nouns, verbs, and adjectives such as *hat* and *umbrella*, *obey* and *island*, and *dilatory* and *aberration*. Examinees are asked two types of questions: What is a x ? and What does x mean? Responses that make use of synonyms, describe major uses, or provide definitive features are

3.4. Measures

awarded the maximum score of 2 points. Incorrect responses on either of the first two items administered require testing to reverse until two perfect scores are achieved on two consecutive items. Testing discontinues after five consecutive scores of zero. Feedback is provided only on the first two items administered, and prompts may be given for vague or unclear responses, such as 'What do you mean?', 'Tell me more about it', or in the case of nonverbal responses, 'Yes, but what is it called?' When an examinee gives multiple responses, the best of these is counted for scoring purposes unless the examinee provides a response that indicates a fundamental misunderstanding of the word, in which case the response is scored 0, for example '*Ancient* means something very old. . . and it's magic'. All responses were audio recorded and transcribed for scoring.

The WISC-IV reports split-half reliability for the Vocabulary subtest of $r = .86$ to $.90$ across the 8-10 year bands, as well as test-retest reliability of $r = .91$ in the 8-9 year group, and $r = .93$ in the 10-11 year group, suggesting a high level of internal reliability overall for all age groups pertinent to the present study.

3.4.3 Other Oral Language Measures

Other oral language assessments in the battery included the CELF-IV Understanding Spoken Paragraphs (USP) subtest as a measure of listening comprehension, and the CELF-IV Formulated Sentences (FS) subtest as a measure of expressive grammar. Both assessments are detailed below.

3.4.3.1 CELF-IV Understanding Spoken Paragraphs

In the CELF-IV USP subtest, examinees are firstly presented with a trial paragraph in which they are required to listen to a verbally presented passage and then answer five questions tapping literal and inferential information as well as predictions. After this training phase, examinees repeat this procedure for a further three test passages, with a maximum total score of 15. There are no discontinuation criteria. All passages were recorded by the researcher at an even reading rate and played back to each participant using an Olympus digital voice recorder.

The USP subtest utilises one set of passages for 7 to 8 year-olds, and a different set for 9 to 11 year-olds. As a result, the transition in passage difficulty between the two age bands is rather abrupt in comparison to other assessments which utilise the same stimuli for all participants but stipulate discontinuation criteria. The effects of this age threshold are discussed further in Chapter 4. The CELF-IV USP subtest has a test-retest reliability coefficient of $r = .74$ for the 8;0 to 8;11 and 9;0 to 9;11 age bands, which decreases to $.64$ for the 10;0 to 10;11 age band.

3.4.3.2 CELF-IV Formulated Sentences

In the Formulated Sentences (FS) subtest of the CELF-IV, examinees are required to formulate syntactically and semantically appropriate sentences about pictures using target stimulus words or phrases of increasing difficulty, for instance *children*, *quickly*, *although*, and *however*. A score of 2 is awarded if a sentence is complete, meaningful, and contains no errors; a score of 1 is awarded for one or two deviations in semantics or syntax; a score of 0 is awarded for incomplete sentences, failure to use the target word, sentences with more than two syntactic or semantic

errors, and sentences that, despite use of the correct grammatical structure, are not meaningful or logical. After training on trial items, testing discontinues after five consecutive scores of 0. All responses were audio recorded and transcribed for scoring. Importantly, examinees attempt slightly different subsets of stimuli according to their age: children aged 8;11 or under discontinue at item 24, while those aged 9;00 and above begin at item eight and additionally attempt items 25 to 28. These last four items are phrases, including *as soon as*, *in order to*, *even though*, and *as a consequence*, and are not accompanied by pictures. In order to maintain comparability over time, children's performance on the 17 commonly attempted items at each time point was also examined.

The internal consistency reliability (Cronbach's α) of the FS subtest by age group is as follows: 8;0 to 8;11 age band, .80; 9;0 to 9;11 age band, .81; 10;0 – 10;11 age band, .76. As examinees' responses on the FS subtest often require judgements in scoring, for example as to whether to award partial or full credit, the CELF-IV manual also reports inter-scoring agreement rate, which is high at 90%.

3.4.4 Oral Narrative Retell

One measure of oral narrative retell was included in the test battery as an assessment of spoken language productivity (total number of utterances, mean length of utterance, diversity of vocabulary employed in speech), as well as morphosyntactic accuracy.

3.4.4.1 Peter and the Cat (Leitão & Allan, 2003) - Oral Narrative Retell

Peter and the Cat is an oral language sampling instrument intended for use with children aged between five and nine years. The assessment is not norm-referenced, but provides a descriptive profile of a child's narrative ability according to standardised assessment criteria. During administration of Peter and the Cat, the examiner reads the story from a script while the examinee looks at the illustrated colour booklet, following the sequence with each picture. The examinee is then asked to retell the story in his/her own words using the pictures. Prompting by the examiner is minimal and non-specific, avoiding leading prompts such as 'What happened next?'

At t1 the Peter and the Cat manual was used to score children's micro- and macrostructure according to a three-point scale for each subcategory. However, given the lack of detailed scoring guidance and concerns around reliability of scoring, the decision was taken not to incorporate micro- and macrostructure scores calculated from the manual, but rather to use alternative measures of microstructure only (see below).

Language sample variables include measures of productivity such as number of utterances, number of words (tokens; N), and number of different words (types; V). They also provide measures of utterance length or complexity, known as Mean Length of Utterance (MLU). There has been debate in the literature as to the comparability of MLU in words (MLUw) and MLU in morphemes (MLUm). Although the two measures correlate very highly with one another for children up to the age of 9 years (Rice et al., 2010), the use of MLUm may be more problematic due to 'arbitrary' decisions concerning the productive status of each morpheme: for example, a decision would have to be made as to whether *wanna* functions productively as one or two morphemes in a child's speech (Parker & Brorson, 2005). Given the high correlation between the two measures,

3.4. Measures

the time-consuming nature of calculating MLUm, and the relatively more developed speech of the children in this study (aged 8 to 10), the decision was made to employ MLUw as a measure of utterance length.

Children's utterances were transcribed into C-units, defined as a main clause and any of its subordinating clauses³. For example, utterance (1) below is considered a single C-unit with one main clause (*there was once...*) joined by a subordinate clause (*who...*). Since the subordinate clause cannot stand on its own, it is considered together with its main clause to form one C-unit. On the other hand, utterance (2) is considered to comprise two separate C-units joined by a coordinating conjunction (*and*). A second clause with an elided subject (represented by [...]) in utterance (3) is considered to be dependent on the first clause, and together both are considered to comprise a single C-unit.

1. There was once a boy called Peter who loved animals (1 C-unit)
2. There was once a boy called Peter | and he loved animals (2 C-units)
3. Peter loved animals and [...] wanted to help them (1 C-unit)

Language samples also provide measures of lexical diversity. Typically, the ratio between types and tokens (N/V) is taken as a measure of diversity. As a ratio, this ranges from 0 to 1, where a higher score indicates a higher level of diversity. However, TTR is shown to be flawed due to its over-reliance on language sample length (Malvern et al., 2004; Vermeer, 2000). Repetition is a part of natural speech, and thus TTR decreases with increasing sample length since words are more likely to be reused. Various corrections to TTR have been proposed, including the Root TTR (Guiraud, 1959), calculated as V/\sqrt{N} , which is less dependent on sample length by virtue of taking the square root of the total number of words. In an explicit comparison of various lexical diversity measures, Vermeer (2000) showed Root TTR to be significantly correlated with standardised measures of receptive and expressive vocabulary, and to be sensitive to growth in lexical diversity over time, unlike TTR. Thus, Root TTR was taken as a measure of lexical diversity in the present study.

All oral narratives were audio recorded and transcribed using CLAN software and CHAT conventions, including semantic and morphosyntactic errors (MacWhinney, 2000). Presentation of the raw number of errors for each retelling is inappropriate due to the differing length of children's retellings: alternatively, error rates are presented, calculated as the total number of raw errors divided by the total number of utterances. For the purposes of error marking in oral narrative data in the present study, semantic errors related to incorrect use of personal (*he, she*, etc.) and relative pronouns (*who, which*), errors relating to phrasal vocabulary (e.g. *thankful to, she scolded at him*), double comparatives (*more louder*), and related but incorrect choice of vocabulary (e.g. *a very long tree; the man was washing his garden*). Morphosyntactic error marking included subject-verb agreement (*they was...*), missing past tense (*he shouted as loud as he can*), over-regularisation (*he holded on to the tree*), and missing subjects, objects, and determiners.

³An early analogy is found in Hunt's (1966) T-unit. C-units in oral language sample analysis are distinguished from T-units in writing analysis in their ability to accommodate commonly found communicative strategies unique to oral communication, such as repetition, ellipsis, recasting, and so on (MacWhinney, 2000).

3.4.5 Phonological Processing Measures

Phonological processing refers to “the use of phonological information (i.e. the sounds of one’s language) in processing written and oral language” (Wagner & Torgesen, 1987, p.192). Two measures in the present study were categorised as phonological processing, including the Spoonerisms subtest of the Phonological Assessment Battery, and the Rapid Digit Naming and Rapid Letter Naming subtests of the Comprehensive Test of Phonological Processing, discussed below in turn. Although the categorisation of RAN as phonological processing is contested (Wolf, Miller & Donnelly, 2001), the decision was made to retain it under this category, since RAN does involve processing of phonological information at some level (Bowey, 2005).

3.4.5.1 Phonological Assessment Battery (PhAB; Frederickson, Frith & Reason, 1997) – Spoonerisms

The Spoonerisms subtest of the PhAB measures children’s ability to substitute individual phonemes and syllables in real and pseudowords. After 3 initial practice items, Part 1 comprises a partial spoonerisms task in which examinees are asked to substitute individual phonemes in words, e.g. *fun* with a *b* gives *bun* (a score of 0 or 1). Part 2 requires substitution between two words, e.g. *fed man* gives *med fan* (a score of 0, 1 or 2 as 1 point is awarded if only one of the words is correct). Children are allowed a maximum of 20 seconds for each item and while each part has no discontinuation criteria, children over the age of 7 who do not score on part 1 do not progress to part 2. The PhAB is standardised on a population of 628 children in primary school Years 1 to 9 in the U.K. (age 6 to 14), with children learning EAL comprising 3.6% of this sample. The Spoonerisms subtest has a high internal consistency reliability coefficient of $\alpha = .93$ for the 8;0 to 9;11 age band.

3.4.5.2 Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen & Rashotte, 1999): Rapid Digit Naming (RDN) and Rapid Letter Naming (RLN)

The rapid naming subtests of the CTOPP assess speed of letter and digit naming. Examinees are presented initially with practice items and asked to read the numbers 2, 7, 4, 5, 3, 8 and letters *a, t, s, k, c, n* as quickly as possible. Each subtest consists of two parallel forms with a matrix of four rows of 9 items each, which examinees are asked to read aloud, from left to right, as quickly as possible. Testing discontinues if examinees are unable to name all of the items on the practice sheet and if more than four errors are made on either form; errors include incorrect pronunciation and omission of items. Upon skipping a line, the first item is scored as incorrect and the examinee is redirected to the appropriate place. Scores for each subtest are calculated by summing the time taken in seconds to name stimuli on both forms. The scaled scores of each subtest (with a mean of 10 and standard deviation of 3) may be summed to form a rapid naming composite score.

The CTOPP is normed on a sample of 1,656 individuals across thirty U.S. states aged between 5 and 24 years, with rapid naming subtests being appropriate for those aged 7 years and over. Average content sampling coefficients across all age bands for RDN and RLN are .87 and .82, respectively. A subgroup of 30 individuals aged 7 to 18 years was assessed on two occa-

3.4. Measures

sions no more than two weeks apart. Test-retest coefficients of $r = .80$ and $.72$ for digit and letter naming, respectively, indicate an acceptable level of temporal reliability.

3.4.6 Literacy Measures

The decision was made to focus primarily on passage reading skills due to the age of the children in the study (Year 4; age 8-9 at t1) as well as the literacy demands of the national curriculum in this particular educational phase (i.e. a marked transition in focus from word- to passage-level reading skills).

3.4.6.1 Test of Word Reading Efficiency (TOWRE-II; Torgesen, Wagner & Rashotte, 1999)

The TOWRE-2 is a timed assessment in which examinees are asked to read aloud lists of real words (Sight-Word Efficiency) and non-words (Phonemic Decoding Efficiency) of increasing difficulty, for instance, *is, up, cat, ... morning, resolve, describe, ... calculated, alternative, collective* in sight-word reading, and *ip, ga, ko, ... prain, zint, bloot, ... strotalanted, prilingdorfent, chun-fendilt* in phonemic decoding. As such, the TOWRE is a measure both of the ability to recognise and produce real words varying in frequency and regularity of spelling, as well as to apply grapho-phonemic knowledge (Tarar, Meisinger & Dickens, 2015). Practice lists of 8 words are presented initially in order to encourage familiarity with the test format. Examinees are asked to read as many words as they can, as quickly as possible, within 45 seconds; after hesitations of up to 3 seconds, the examinee is prompted to continue onto the next item.

The TOWRE-2 is normed on a nationally representative U.S. population of 1,717 individuals aged between 6;00 and 24;11 years. The sample is representative of the U.S. school-aged population in terms of gender, geographical distribution, parental education and income, Hispanic status, and prevalence of learning difficulties. The assessment consists of four parallel forms, A, B, C, and D, which are shown to have delayed test-retest reliability coefficients ranging from $r = .89$ to $.93$, and very high inter-scorer reliability coefficients of $r = .99$ for sight-word efficiency, phonemic decoding efficiency, and total word reading efficiency. Form A only was selected for use in the study.

3.4.6.2 York Assessment of Reading Comprehension Primary 2nd Edition (YARC; GL Assessment, 2011)

The YARC Primary is an individually-administered, graded reading comprehension assessment for primary school pupils aged 5 to 11 years. The assessment consists of an interactive Beginner passage, as well as parallel forms (A and B) of 6 test passages which examinees are asked to read aloud. Form A only was used in the present study. Passages are a mixture of fiction and non-fiction text and increase in difficulty in accordance with year groups and requirements of the national curriculum. The YARC yields raw, ability, and standard scores (mean=100, SD=15) for passage reading rate, accuracy, and comprehension. It is recommended that testing begin at the level appropriate for the child's school year (e.g. Level 4 for Year 4), although it is advisable to begin on the previous level if testing takes place within the first half of the academic year. After

the beginner passage, an examinee then continues backwards if he or she scores 4 or fewer marks for comprehension or reaches near the limit of 20 reading accuracy errors, or forwards if either of these criteria is not met. Examinees progress through the assessment until they reach discontinuation criteria or complete the sixth and final passage. Scoring is calculated using the two highest, consecutively completed passages. Given that children in the present study were assessed at the beginning of Year 4, Level 3 was administered as a beginner passage.

At t1, children showed great variability in the passages they read, potentially creating difficulties for fair comparison between the monolingual and EAL groups over time. Thus, the decision was taken at subsequent time points to assess all children on the same two passages (2 and 3 at t2 and 3 and 4 at t3), then allowing them to continue forwards or backwards until discontinuation criteria were met or until they reached the final passage. Other than the choice of the minimum passages to be read at time points 1 and 2, the YARC was administered fully in accordance with instructions in the manual. Analyses in Chapter 4 will make reference both to YARC standard scores for rate, accuracy, and comprehension, as well as raw scores derived from passage 3 alone. Passage 3 was chosen for the analysis of raw scores in order to allow within- and between-subjects comparison over time holding passage constant. As stated above, all children read this passage at t2 and t3, and the majority read it at t1 (n=75 out of 81 children).

During administration of the YARC, the examiner provides assistance for decoding difficulties but does not provide correct answers for comprehension questions. Questions tap both literal and inferential comprehension skill, and prompts such as 'can you tell me more?' are given for vague or incomplete answers. Generally, answers are considered correct if the gist matches exemplar responses in the answer key of the manual, however some answers require explicit mention of certain key words. Questions that require two-part responses are marked correct if and only if both parts of the response are given, as no half marks are permitted.

The YARC was designed and standardised using a nationally representative U.K. primary school population of 1,376 children, 14% of whom did not speak English as a first language. The test is designed to be gender- and culture-neutral. The Accuracy and Rate components of the YARC have the highest reliability, with correlations between parallel forms of .75 to .93 for accuracy and .91 to .95 for rate. The Comprehension component, on the other hand, has a lower range of reliability of between $r = .48$ and $.77$ across all levels and alternative forms. However, reliability calculated from pair passages (e.g. scores of performance on Level 1A and 2A, and so on) is higher, with coefficients between $r = .71$ and $.84$. Taken together with the varied nature of passage content and types of comprehension questions, then, there is evidence for good reliability and validity of the YARC Primary.

3.4.6.3 Writing: Bespoke Task

For ease of comparison with oral narrative language measures such as number of utterances and MLU, a bespoke writing measure was employed in which examinees were asked to produce a short piece of writing based on topic prompts such as 'my favourite place' and 'my favourite thing to do'. Before writing, examinees were prompted by the examiner to think of examples, and then given a maximum of five minutes to write as much as possible. Examinees were prompted to pay attention to accuracy of spelling and grammar.

3.5. Questionnaires

Similar to the Peter and the Cat oral narrative measure, children's writing samples were transcribed manually into T-units, defined as a main clause plus its subordinate clauses (Bishop & Clarkson, 2002; Hunt, 1970). As well as total number of T-units, additional variables included the mean length of T-unit in words (MLTw), and the mean number of spelling and morphosyntactic errors per T-unit. Words were categorised as spelling errors simply if they did not conform to correct spelling, and as in analysis of oral narrative data, morphosyntactic errors included those in agreement (e.g. *there was lots of exsiting [sic] things*), past-tense morphology and over-regularisation (e.g. *when I saw the Queen I taked a photo with her*), as well as omission of obligatory elements such as pronouns, prepositions, or determiners (e.g. *I got Xbox, or however I wanted to go Japan*).

3.5 Questionnaires

Bespoke background questionnaires were administered to parents/carers of all participating children, and EAL pupils also completed a language preference questionnaire at t1 and a language balance questionnaire at t2. As the aim of such data collection was to provide contextual information regarding children's home and language experiences, data from the questionnaires were not linked to performance on any of the language and literacy measures described above.

Design of the parental questionnaire was based in part on the Language Preference Questionnaire of Beech and Keys (1997) and the English exposure questionnaire of Cattani et al. (2014), particularly in assessing the balance of linguistic exposure in children's home environments. While the questionnaire of Beech and Keys (1997) is very brief and contains only open-ended format questions, that of Cattani et al. (2014) is extremely quantitative and fine-grained in nature, with the aim of correlating linguistic exposure with language skill. Instead, the aim of the parental questionnaire in the present study was to strike a balance between these two extremes in terms of length and level of detail.

The parental questionnaire asked questions concerning: background information such as how long the child had been in the UK; maternal and paternal education and employment status; English language ability and prevalence of familial language or literacy difficulties; and language use in the home, such as which languages the child hears and speaks most often, levels of literacy ability in and receipt of formal instruction in the home language, which family members speak to the child in the home language, and prevalence of various media in the home language (e.g. television, books, internet). Summaries of the questionnaires will now be presented, although note that the response rate for parents/carers of monolingual children was 63.6%, and 75% for those of EAL learners. See Appendices 3.2 and 3.3 from page 254 for full numerical results.

Summary of parental questionnaire. Parental questionnaire results indicated that all monolingual children and a large majority of the children with EAL (92%) were born in the U.K., and that the majority of children in both groups had begun some form of formal education in English by age three to four (EAL: 89%, Mono: 100%). All parents of children categorised as learning EAL by their schools indicated presence of a language other than English in the home. The majority of parents of children with EAL had been educated outside of the U.K., with 58% of mothers and 67% of fathers indicating 'good' or 'excellent' oral proficiency in English and similar proportions for English literacy skill. There was a clear trend for moderate to high exposure to English in the homes of children with EAL, with 64% of respondents indicating that English was spoken 'most'

or 'all of the time'. This pattern also extended to media consumption, with a predominance of exposure to English through different media, including books (97%), television (75%), and the internet (97%). Finally, there was a fairly even split in the number of children with EAL who were said to have some literacy ability in the home language (54%): of these children, only 33% had received any formal instruction in the language, which tended to be in the form of attendance at Mosque.

Summary of the child language preference questionnaire. At t1 children were also asked which language(s) they speak at home and with whom, their preferred language, and level of literacy ability and extent of formal instruction in the home language. In general, children's answers mirrored those of their parents, with only a small degree of discrepancy in reporting of the home language. Examples of discrepancies included cases in which two or more languages other than English were spoken in the home, and variations on spelling/pronunciation, e.g. Nepalese/Nepali, Bengali/Bangali, and Pashto/Pushto. Ninety-four per cent of children spoke just one language other than English in the home, and many indicated predominance of the use of (62%) and preference for English (63%). Children showed relatively even splits between preferences for English or home language use with their parents (57% speaking English 'most' or 'all of the time' with their mothers, and 51% with their fathers), but clear preferences for speaking English with siblings and friends (88% and 97% 'most' or 'all of the time', respectively) (see Appendix 3.3).

Summary of the child language balance questionnaire. At t2 children were asked to rate their perceived proficiency in speaking, understanding, reading, and writing in each of their languages. Children were asked, for example, 'How well can you speak English?' and 'How well can you speak your other language?' Children were asked to indicate their answers using a likert scale from 1 (little or no ability) to 9 (high ability). Results of the language questionnaire are presented in Table 3.3. It appears that while children with EAL generally perceived themselves to have higher competence in English, the largest differences between languages were in ratings of reading and writing – indeed, only around half of the children indicated some literacy ability in the home language (n=25; 53%).

Table 3.3: Results of Child Language Balance Questionnaire

	English		Home Language		Difference
	Mean (SD)	Min-Max	Mean (SD)	Min-Max	
Speaking	8.11 (1.37)	4-9	5.41 (2.24)	4-9	2.70
Understanding	8.18 (1.45)	4-9	6.02 (2.44)	1-9	2.16
Reading	8.09 (1.57)	4-9	3.18 (2.97)	1-9	4.91
Writing	8.39 (1.24)	5-9	2.45 (2.43)	1-9	5.94

Note: A higher score for 'English' or 'Home Language' indicates a higher degree of self-reported proficiency in that language; Difference = mean score in English minus mean score in Home Language.

3.6 Procedure

At each time point, children were assessed one-to-one on school premises outside of classrooms by the researcher. Testing took place over two separate sessions on different days, lasting up to a maximum of 45 minutes each. Efforts were made to minimise the interval between each session: the mean interval at each time point was as follows: t1 (6.1 days); t2 (10.5 days); t3 (3.3 days). Similarly, efforts were made to maintain equidistance between time points. The average length between the first (t1-t2) and final (t2-t3) sets of time points was 6.9 and 6.8 months, respectively. The order of assessments within the test battery was balanced approximately according to task type, duration, and difficulty in order to promote children's interest and attention. At t2 and t3 the same order was retained with the exception of removal of nonverbal reasoning at t2.

Chapter 4

Results and Discussion I: Longitudinal Study

This chapter presents the results of the longitudinal cohort study. All children were assessed on a battery of language and literacy measures at three time points just under 7 months apart from each other; at the beginning of Year 4 (t1), end of Year 4 (t2), and middle of Year 5 (t3). The specific research questions addressed in this chapter are:

1. What are the similarities and differences in the language and literacy skills of children learning EAL and their monolingual peers at the beginning of Year 4 (t1)?
2. What are the developmental trajectories of the two groups of children in language and literacy skills between Year 4 (t1) and Year 5 (t3)?

The structure of this chapter is as follows: after some preliminary considerations regarding missing data and background measures, tables of descriptive statistics will be presented for each variable across the three time points, grouped thematically into vocabulary, oral language, phonological processing, and literacy. The aim of this is to familiarise the reader with the structure of and general trends within the data. Next, the linear mixed model statistical framework will be introduced and inferential analyses will be run in order to determine the statistical significance of the effects of time and group status upon children's performance. Finally, findings from all analyses will be discussed.

4.1 Preliminary Considerations and Background Measures

Before analysis of trends observed in descriptive statistics, some preliminary considerations with regards to attrition and missing data, use of raw and standardised (standard and scaled) scores, and effect size calculation warrant brief discussion.

4.1.1 Attrition and Missing Data

At t1, a total of 33 monolingual children and 48 children with EAL were recruited onto the study. By t3, one child from the monolingual group and three children from the EAL group no longer remained in the study, resulting in a sample of 32 monolingual children and 45 children with EAL and a total attrition rate of 4.9%¹. In addition to overall attrition, missing data also arose due to some children not being available for testing. Including attrition and absences, the total missing data was 3.46% across all three time points (monolingual group = 1.09%, EAL group = 3.65%).

¹At t2 two children from the EAL group had moved to different schools. At t3, one child from the monolingual group and one child from the EAL group had moved to different schools.

An advantage of the linear mixed model framework which was used for inferential analyses in this study (Section 4.4) is the ability to handle unbalanced and missing data (Muth et al., 2016). Therefore, the decision was made to use all available data at each time point (as opposed to using data for only the 74 children for whom data existed at all three time points). Note that the exact sample size for each group, time point, and measure is included in descriptives tables.

4.1.2 Use of Raw and Standardised Scores

For consistency and for issues concerning the appropriateness of using standardised scores in samples of bilingual learners, the decision was made to use only raw scores in descriptive and inferential analyses (i.e. linear mixed modelling) where possible - see Babayiğit (2014a), Farnia and Geva (2012), and Jean and Geva (2009) for similar justifications. However, standardised scores were also utilised in the following cases: (1) where EAL learners are included in the norming sample of an assessment (this included only the YARC and the BPVS, although the decision was made not to use standardised scores for the BPVS as EAL learners represent only 1.39% (n=45) of the norming sample for this assessment, compared to 14% (n=192) for the YARC); and (2) where children passed age thresholds and were therefore not all attempting the same passages (in particular, this applied to CELF USP, for which standardised [scaled] scores were therefore used). A similar situation applied to CELF FS, where children attempted slightly different subsets of stimuli according to their age (see Section 3.4.3.2). However, since these were not completely different stimuli, raw scores were used in terms of children's total scores on the subtest, as well as children's scores on a subset of 17 items which were attempted at all three time points regardless of children's ages. A breakdown of each variable and the type of score (raw or standardised) used in descriptive and inferential analyses is provided in Table 4.1 below.

Table 4.1: Use of Raw and Standardised Scores Across Measures

Variable Group	Measure	Data reported in descriptive statistics	Data used in linear modelling
Vocabulary	BPVS-III	Raw and Standard	Raw
	CELF-IV EV	Raw and Scaled	Raw
	WISC-IV VC	Raw and Scaled	Raw
Oral Language	CELF-IV FS	Raw and Scaled	Raw
	CELF-IV USP	Raw and Scaled	Scaled
	Peter and the Cat	Raw	Raw
Phonological Processing	CTOPP RLN	Raw and Scaled	Raw
	CTOPP RDN	Raw and Scaled	Raw
	PhAB Spoonerisms	Raw	Raw
Literacy	TOWRE Sight-Word Efficiency	Raw and Standard	Raw
	TOWRE Phonemic Decoding	Raw and Standard	Raw
	YARC Primary	Raw and Standard	Raw and Standard
	Bespoke Writing Task	Raw	Raw

4.1.3 Effect Sizes

Where appropriate, between-group effect sizes at each time point (and their 95% confidence intervals) are presented in Hedges' g (Hedges, 1981), a 'D-family' measure of effect size which corrects for bias introduced due to unbalanced sample size (Peng & Chen, 2014). Interpretation of g follows that of Cohen's d , whereby 0.2 is considered a small effect, 0.5 is medium, and over 0.8 is large (Cohen, 1988). In all analyses, the monolingual group served as the comparison group, and thus a positive effect size is interpreted as a higher score for the monolingual group, and a negative effect size as a higher score for the EAL group. Additionally, for independent T-tests, r is reported as a measure of effect size, where 0.1 is interpreted as a small effect, 0.3 as a medium effect, and 0.5 as a large effect (Cohen, 1988). Cronbach's α is presented as a measure of internal reliability. An average reliability coefficient of .63 was obtained for all measures in the test battery, although this was higher for norm-referenced assessments (.75) than oral and written narrative measures (.48)². Internal reliability is reported individually for each variable at each time point in descriptive tables.

4.2 Background Measures

At t1, all children were assessed on WISC Matrix Reasoning (MR) and CELF Number Repetition (NR) as background measures in order to ensure fair comparison across the two groups. Descriptive statistics for background measures are presented in Appendix 4.1 on page 262. At t1, the two groups did not differ significantly in nonverbal reasoning (WISC MR, $t(79) = -0.59$, $p = .556$, $r = .06$) or on short-term memory (CELF Number Repetition Forwards, $t(79) = 1.43$, $p = .157$, $r = .16$). However, the monolingual group showed a small but statistically significant advantage in CELF NR Total score (sum of NR Forwards and Backwards; $t(79) = 2.42$, $p = .018$, $r = .27$). This latter group difference had decreased in magnitude by t2, and had reversed direction slightly by t3, whereby the EAL group began to outperform the monolingual group, albeit not significantly so ($t(74) = -0.51$, $p = .613$, $r = .06$). Thus, neither group appeared to be advantaged relative to the other, supporting the appropriateness of group comparisons on language and literacy measures of interest.

4.3 Descriptive Statistics

Tables 4.2 to 4.8 present the mean, standard deviation, range, and internal reliability (α) of each variable, as well as the effect size (Hedge's g) and 95% confidence intervals of group differences. Note that g and α are presented only for variables used in linear mixed models in Section 4.4. Consistent colour-coding is applied in descriptives tables and graphs, with monolingual in green and EAL in blue. Line graphs are also presented in Appendix 4.2 on page 262.

²An alpha value of .70 is often considered as a minimum standard of internal reliability (Henson, 2001).

4.3.1 Descriptives for Vocabulary Measures

As shown in Table 4.2, children's progress in vocabulary knowledge tended to show an upward progression throughout the duration of the study. For all vocabulary variables, group discrepancies in performance were apparent at t1, with the monolingual group on average scoring higher than the EAL group, particularly in terms of expressive vocabulary. In general, the two groups made very similar rates of progress between t1 and t2, although a different pattern emerged between t2 and t3, whereby the monolingual group plateaued, but the EAL group continued to make a faster rate of progress. As a result of this deceleration, the monolingual advantages found at t1 decreased in magnitude by t3, particularly for vocabulary depth. All data approximated a normal distribution with few if any outliers, however CELF EV scores were somewhat negatively skewed at t3 when a number of children scored full or near full marks. Note, however, that this pattern did not apply to the BPVS, and thus the plateauing of the monolingual group was not a reflection of children performing near the limit of the possible scoring range. All three vocabulary measures indicated a high degree of internal reliability of $\geq .81$.

4.3.2 Descriptives for Oral Language Measures

Descriptive statistics for performance on the two other oral language measures – listening comprehension (CELF USP) and expressive grammar (CELF FS) – are shown in Table 4.3. Due to the fact that children attempted different passages on the USP subtest, listening comprehension scores are presented firstly in scaled scores, with raw scores for reference only. Similarly, children attempted slightly different subsets of items on the FS subtest depending on their age; for this reason, a separate analysis was carried out utilising scores from the 17 items that all children attempted at all time points. For the most part, groups showed upward trajectories in their performance on both oral language measures over time, although one exception was listening comprehension performance of the monolingual group which firstly accelerated and then decelerated to a below-t1 level. The two measures of expressive grammar generally agreed with one another, showing the upward trajectory of both groups, as well as a relatively larger decrease in the effect size of the group difference between t1 and t2 and an increase in the difference by t3. Scores on CELF USP and FS showed lower internal reliability than vocabulary measures, but approximated normal distributions. Interestingly, a considerable number of children from both groups obtained very low scaled scores at t2 ($n=26 \leq 6$), when a number of children passed from the age 7-8 to the age 9-11 CELF USP age threshold. This somewhat abrupt age threshold change was compounded by the fact that, of the 52 children who turned nine between t1 and t2, the average age was 9 years 3 months, well towards the lower range of the age band, potentially accounting for the elevated number of low scores on this subtest at t2.

Table 4.2: t1-t3 Descriptive Statistics for Vocabulary Measures

	Time	Data	Mono			EAL			Effect Size of Group Difference			
			N	Mean (SD)	Range	N	Mean (SD)	Range	α	g	95% CI	
BPVS (Max = 168)	1	Raw	33	103.00 (15.93)	80-149	48	94.81 (15.72)	71-143	.86	0.51	0.05 to 0.98	
		Std.		87.80 (15.30)	69-130		80.13 (10.62)	69-111				
	2	Raw	33	113.18 (14.04)	95-148	47	105.83 (15.09)	74-141	.90	0.50	0.03 to 0.96	
		Std.		87.00 (13.85)	69-122		80.70 (10.99)	69-114				
	3	Raw	32	116.78 (13.77)	90-147	44	110.84 (16.76)	71-156	.88	0.38	-0.10 to 0.85	
		Std.		86.34 (11.60)	69-116		81.70 (12.50)	69-122				
	CELF EV (Max = 54)	1	Raw	33	37.85 (5.44)	28-51	48	34.00 (7.51)	16-51	.81	0.56	0.10 to 1.03
			Std.		8.91 (2.45)	4-15		7.35 (2.81)	2-15			
		2	Raw	33	41.64 (5.88)	32-54	45	37.16 (6.56)	24-51	.82	0.71	0.23 to 1.18
Std.				9.42 (2.73)	5-15		7.33 (2.53)	3-13				
3		Raw	32	42.06 (4.81)	32-50	44	39.50 (6.39)	26-51	.80	0.44	-0.04 to 0.91	
		Std.		-	-		-	-				
WISC VC (Max = 68)		1	Raw	33	26.73 (6.37)	17-48	48	25.04 (5.89)	13-41	.90	0.27	-0.18 to 0.73
			Std.		9.70 (2.58)	5-18		8.90 (2.40)	4-15			
		2	Raw	33	28.85 (7.23)	18-51	45	27.16 (5.91)	18-44	.88	0.30	-0.17 to 0.76
	Std.			9.94 (2.97)	5-19		8.93 (2.28)	5-14				
	3	Raw	32	29.69 (6.24)	21-46	44	29.86 (6.26)	17-43	.82	-0.03	-0.50 to 0.44	
		Std.		9.19 (2.47)	5-15		9.11 (2.34)	5-14				

Note: Std. = Standardised Score (Standard or Scaled); BPVS = British Picture Vocabulary Scale III; CELF EV = Clinical Evaluation of Language Fundamentals IV; Expressive Vocabulary; WISC VC = Wechsler Intelligence Scale for Children IV Vocabulary (depth); standard scores and standard deviations in italics (CELF EV scaled scores not available above the 9;00-9;11 age band for t3); α = internal reliability (Cronbach alpha)

Table 4.3: t1-t3 Descriptive Statistics for Oral Language Measures

	Time	Data	Mono			EAL			Effect Size of			
			N	Mean (SD)	Range	N	Mean (SD)	Range	α	Group Difference g	95% CI	
CELLF USP (Max = 15); Std. scores above	1	Std.	33	8.00 (2.89)	3-13	48	7.00 (2.40)	1-12	.69	0.38	-0.08 to 0.84	
		Raw		9.36 (3.01)	3-14		8.29 (2.67)	1-13				
	2	Std.	33	8.30 (2.87)	4-14	47	7.30 (2.42)	4-13	.67	0.38	-0.08 to 0.84	
		Raw		9.06 (3.27)	4-15		7.81 (2.86)	4-14				
	3	Std.	32	7.84 (2.84)	2-14	44	8.05 (2.23)	4-15	.56	-0.08	-0.55 to 0.39	
		Raw		8.50 (3.26)	2-15		8.77 (2.66)	4-15				
	CELLF FS (Max = 56)	1	Raw	33	35.58 (5.37)	26-46	48	31.98 (6.99)	18-50	.60	0.56	0.09 to 1.02
			Std.		7.39 (2.26)	4-11		5.71 (2.58)	1-12			
		2	Raw	33	40.76 (6.67)	22-55	45	39.42 (5.83)	28-51	.62	0.21	-0.25 to 0.68
Std.				8.42 (2.56)	2-15		7.58 (2.40)	3-13				
3		Raw	32	45.63 (6.75)	29-56	44	43.40 (5.01)	31-55	.67	0.40	-0.08 to 0.87	
		Std.		9.75 (2.39)	3-15		8.50 (2.23)	4-14				
CELLF FS Common 17 Items (Max = 34)		1	Raw	33	22.64 (4.48)	15-30	48	18.60 (5.47)	9-31	.60	0.78	0.31 to 1.26
			Raw	33	23.91 (4.52)	11-33	45	22.07 (4.37)	12-29	.61	0.41	-0.06 to 0.88
		2	Raw	32	26.41 (5.27)	15-34	44	24.27 (3.66)	17-33	.62	0.48	0.00 to 0.96
	Raw											
	3	Raw										
		Raw										

Note: Std. = Standardised (Scaled) Score; CELLF USP = Clinical Evaluation of Language Fundamentals IV Understanding Spoken Paragraphs subtest; CELLF FS = Formulated Sentences subtest; reversal of order for CELLF USP, with scaled scores above and raw scores below; α = internal reliability (Cronbach alpha)

4.3.3 Descriptives for Oral Narrative Measures

Table 4.4 presents descriptive statistics for data obtained from the Peter and the Cat oral narrative retell task. Overall, while children in the EAL group tended to produce a higher number of utterances, these were shorter and contained more errors than those of their monolingual peers. The two groups showed similar levels of lexical diversity in their retellings, suggesting that both groups utilised a similar range of vocabulary when retelling the narrative. Oral narrative measures displayed approximately normal distributions aside from total, morphosyntactic, and semantic error rates which were all extremely positively skewed due to the majority of children who made no or very few errors in their spoken language retellings (this did not affect appropriateness of data for linear mixed modelling; see Sections 4.4.2 and 4.4.7). Internal reliability estimates for oral narrative measures ranged from a low of .28 (morphosyntactic error rate) to a high of .62 (utterance rate).

In terms of progress over time, it is interesting that both groups showed a downward trend in the total number of utterances produced. The pattern for a higher number of total utterances by the EAL group was maintained at each time point and the magnitude of this group difference did not change considerably over time. On the other hand, group differences in utterance length (MLUw) showed that children in the monolingual group were on average producing longer utterances. Progress in lexical diversity mirrored that of other oral language variables, particularly between t2 and t3 with the monolingual group decelerating and the EAL group continuing to make a steady rate of progress. Finally, in terms of error rates, it is interesting to note that both groups showed a trend firstly to decrease (between t1 and t2) but then to increase again by t3. As reflected in the ranges for utterance error rate in Table 4.4, a number of children at each time point did not make any errors at all in their speech.

Errors were recorded as semantic or morphosyntactic (see Section 3.4.4.1 for examples). As shown in Table 4.4, rates of each error type were low, with a high degree of variability, and the EAL group consistently made a higher rate of semantic and morphosyntactic errors. Both groups showed similar trajectories, with a reduction in semantic errors between t1 and t2 but an increase in both error types between t2 and t3. At t1 the most common types of morphosyntactic errors by a clear margin were over-regularisation, agreement errors, and missing past tense (comprising 31%, 29%, and 12% of total morphosyntactic errors, respectively). Incidence of each of these error types was overwhelmingly accounted for by children in the EAL group, apart from agreement errors which were evenly split across the groups. This pattern of errors was very similar across all time points with the exception that at t3, incidence of missing past tense errors had increased, now accounting for 43% of all morphosyntactic errors, again predominantly in the EAL group.

Table 4.4: t1-t3 Descriptive Statistics for Oral Narrative Measures

	Time	Mono			EAL			Effect Size of		
		N	Mean (SD)	Range	N	Mean (SD)	Range	α	Group Difference g	95% CI
Utterances	1	32	21.78 (3.85)	16-36	48	24.04 (5.64)	13-47	.60	-0.45	-0.91 to 0.02
	2	33	21.15 (4.72)	13-34	46	22.91 (3.85)	16-32	.61	-0.41	-0.88 to 0.05
	3	32	20.66 (4.61)	15-34	44	22.61 (3.91)	15-34	.62	-0.46	-0.93 to 0.02
MLUw	1	32	7.67 (1.34)	5.86-11.06	48	7.07 (0.99)	5.11-9.04	.48	0.52	0.06 to 0.99
	2	33	8.42 (1.19)	5.24-10.85	46	7.60 (1.12)	5.57-11.16	.39	0.71	0.23 to 1.18
	3	32	8.87 (1.17)	6.00-11.53	44	8.24 (1.20)	6.70-12.37	.47	0.53	0.05 to 1.00
Root TTR	1	32	6.28 (0.52)	5.18-7.68	48	6.28 (0.56)	4.71-7.63	.59	0.01	-0.45 to 0.47
	2	33	6.65 (0.57)	5.65-7.71	45	6.47 (0.48)	5.66-7.51	.59	0.36	-0.11 to 0.82
	3	32	6.77 (0.58)	5.30-7.76	45	6.77 (0.52)	5.10-7.57	.54	0.00	-0.47 to 0.47
Total Utterance Error Rate	1	32	0.07 (0.10)	0.00-0.38	48	0.13 (0.08)	0.00-0.32	.57	-0.64	-1.11 to -0.17
	2	33	0.06 (0.06)	0.00-0.18	46	0.11 (0.08)	0.00-0.29	.43	-0.65	-1.12 to -0.17
	3	32	0.08 (0.09)	0.00-0.33	44	0.12 (0.10)	0.00-0.42	.59	-0.41	-0.89 to 0.06
Morphosyntactic Error Rate	1	32	0.04 (0.06)	0.00-0.28	48	0.05 (0.05)	0.00-0.18	.35	-0.23	-0.69 to 0.23
	2	33	0.03 (0.04)	0.00-0.12	46	0.05 (0.05)	0.00-0.21	.28	-0.46	-0.92 to 0.01
	3	32	0.04 (0.07)	0.00-0.24	44	0.06 (0.08)	0.00-0.37	.46	-0.28	-0.75 to 0.19
Semantic Error Rate	1	32	0.03 (0.06)	0.00-0.29	48	0.04 (0.04)	0.00-0.13	.32	-0.21	-0.67 to 0.25
	2	33	0.01 (0.03)	0.00-0.12	46	0.03 (0.04)	0.00-0.11	.36	-0.48	-0.95 to -0.01
	3	32	0.02 (0.03)	0.00-0.13	44	0.03 (0.04)	0.00-0.14	.36	-0.30	-0.77 to 0.17

Note: Utterances = total number of C-units; MLUw = mean length of utterance in words; Root TTR = lexical diversity; Total Utterance Rate = morphosyntactic and semantic errors combined; α = internal reliability (Cronbach alpha)

4.3.4 Descriptives for Phonological Processing Measures

Table 4.5 presents descriptive statistics for phonological processing measures, including the Rapid Digit Naming and Rapid Letter Naming subtests of the CTOPP and the Spoonerisms subtest of the PhAB. It was interesting to note during testing, particularly at t1, that a number of children struggled with phonological processing tasks: on RAN tasks some children had difficulty naming letters in the appropriate manner (e.g. /keɪ/ vs. 'kicking k') or switched between different pronunciations, and performance on the Spoonerisms task was extremely variable, with roughly equal proportions of children struggling and excelling. Indeed, approximately normal distributions were observed for all three measures, with a slight exception for PhAB which displayed a negative skew at t3 as a number of children scored very highly. Overall, internal reliability for phonological processing measures was acceptable, ranging from .58 to .82.

Performance on both RAN subtests is measured in seconds taken to decode both Form A and B, where a lower score indicates faster fluency in naming. As expected based on previous work (e.g. Jean & Geva, 2009), a trend for an EAL group advantage was found in both RAN measures at t1, particularly for RAN of letters. Children's progress on the two measures over time was unexpectedly asymmetric: while children showed a steep deceleration in RAN of digits (i.e. became faster), the opposite pattern was observed in RAN of letters, in which both groups accelerated between t1 and t2 (i.e. became slower; see graphs for Models 10 and 11 in Appendix 4.2 on page 266). Similarly, interesting patterns were observed for progress between t2 and t3. For RAN of digits, the monolingual group plateaued in performance, whereas the EAL group continued to decelerate, although this deceleration was much less steep than that seen between t1 and t2. For RAN of letters, both groups showed deceleration between t2 and t3, showing a tendency to name letters more quickly.

In contrast to performance on RAN measures, a higher score on the Spoonerisms subtest of the PhAB is taken to indicate higher performance, namely in the ability to extract and replace phonemes or intra-syllabic structures. Overall, there was a consistent trend for a monolingual advantage in spoonerisms, established by t1 and decreasing in magnitude only slightly by t3 due to the very similar trajectories of both groups.

Table 4.5: t1-t3 Descriptive Statistics for Phonological Processing Measures

	Time	Data	Mono			EAL			α	Effect Size of Group Difference <i>g</i>	95% CI
			N	Mean (SD)	Range	N	Mean (SD)	Range			
RAN Letters	1	Raw	33	39.52 (9.81)	25-58	48	35.92 (7.73)	21-58	.66	0.41	-0.05 to 0.87
		Std.		<i>10.30 (2.34)</i>	<i>7-15</i>		<i>11.08 (2.17)</i>	<i>7-17</i>			
	2	Raw	33	45.12 (11.13)	29-79	47	41.28 (8.49)	21-62	.82	0.39	-0.07 to 0.86
		Std.		<i>9.03 (2.21)</i>	<i>5-13</i>		<i>9.77 (2.27)</i>	<i>6-19</i>			
	3	Raw	32	42.63 (11.40)	28-85	44	37.82 (6.98)	22-57	.73	0.52	0.05 to 1.00
		Std.		<i>9.13 (2.37)</i>	<i>5-14</i>		<i>9.73 (2.45)</i>	<i>2-18</i>			
RAN Digits	1	Raw	33	46.21 (12.65)	24-70	48	44.50 (11.08)	26-93	.58	0.14	-0.31 to 0.60
		Std.		<i>9.36 (2.40)</i>	<i>6-18</i>		<i>9.48 (1.94)</i>	<i>5-16</i>			
	2	Raw	33	33.24 (6.51)	22-43	47	31.74 (5.67)	19-46	.78	0.25	-0.21 to 0.71
		Std.		<i>11.09 (2.28)</i>	<i>8-15</i>		<i>11.51 (1.91)</i>	<i>8-17</i>			
	3	Raw	32	33.03 (7.28)	21-52	44	29.02 (5.70)	18-43	.71	0.62	0.14 to 1.10
		Std.		<i>10.75 (2.41)</i>	<i>7-16</i>		<i>11.95 (2.16)</i>	<i>7-18</i>			
PhAB Spoonerisms (Max = 30)	1	Raw	33	17.06 (6.69)	4-30	48	14.33 (6.36)	1-25	.75	0.42	-0.04 to 0.88
		Raw	33	19.70 (4.81)	10-27	46	17.65 (4.94)	7-27	.76	0.41	-0.05 to 0.88
		Raw	32	21.56 (5.58)	9-29	44	19.70 (5.17)	8-28	.73	0.34	-0.13 to 0.82
		Raw									
		Raw									
		Raw									

Note: Std. = Standardised (Scaled) Score; RAN Composite = RAN Digits and RAN Letters; RAN Digits = Rapid Digit Naming subtest of Comprehensive Test of Phonological Processing (CTOPP); RAN Letters = Rapid Letter Naming subtest of CTOPP; PhAB = Phonological Assessment Battery; standardised scores and their standard deviations in italics; standardised scores unavailable for PhAB; α = internal reliability (Cronbach alpha)

4.3.5 Descriptives for Literacy Measures

Descriptive statistics are presented for single- and pseudoword-decoding from the TOWRE (Table 4.6), passage reading rate, accuracy, and comprehension from YARC Primary (Table 4.7), and for written narrative on the bespoke writing task, including total number and average length of T-units, and spelling and morphosyntactic error rates (Table 4.8). As above, summaries of descriptive statistics from each time point for each assessment will be discussed in turn.

Scores derived from performance on the TOWRE indicate the total number of words read correctly in the time limit of 45 seconds, and thus a higher score is indicative of higher word-reading efficiency. Both subtests of the TOWRE showed similar patterns in terms of group differences and progress over time. At t1, the EAL group showed a trend for higher efficiency in relation to the monolingual group, the magnitude of which effect was slightly larger for word- than pseudoword-reading. Both groups made a steady and similar rate of progress over time. Exceptions to this trend were found between t2 and t3, where the monolingual group experienced a slight deceleration in single-word fluency, and the EAL group experienced a discernible acceleration in pseudoword fluency. As a result, group differences became slightly larger in magnitude by t3, with the EAL group maintaining a trend for faster fluency on both subtests. Performance on both TOWRE subtests approximated a normal distribution with no floor or ceiling effects, as well as an acceptable level of internal reliability of between .76 and .89.

In terms of passage reading performance, three primary variables were obtained through administration of the YARC, including passage reading rate, accuracy, and comprehension, each of which will be discussed in turn. Means and standard deviations are presented in standard scores derived from the two highest passages that each child attempted, as well as in raw scores on passage 3 (discussed separately below). Firstly, in terms of passage reading rate, group performance was largely comparable at t1, with both groups scoring on average slightly above the population mean of 100. While the monolingual group showed very little change between t1 and t2, and a slight deceleration by t3, the EAL group showed a small but consistent upward trend between each of the three time points. Group differences at t1 and t2 were negligible ($g = 0.07$ and 0.00 , respectively), but due to the differing directions of the groups' trajectories by t3, this difference had widened to $g = -0.21$. Despite this, both groups were performing on average 3 to 6 standard scores above the population mean of 100. Secondly, a more unexpected pattern was observed in terms of passage reading accuracy. Again, while both groups scored within the average range, there was a small monolingual advantage at t1. However, due to a deceleration of the monolingual group and an acceleration of the EAL group, this discrepancy in performance was far less pronounced at t2 and t3, hinting at an interaction effect. Thirdly, in terms of passage reading comprehension performance, the monolingual group also showed a trend for higher performance at t1. Both groups showed an upward trajectory in performance between t1 and t2, followed by a deceleration by t3. Standard scores for all three variables followed approximate normal distributions. Internal reliability was relatively higher for rate and accuracy (.73 to .91) than for comprehension (.51 to .57).

Table 4.6: t1-t3 Descriptive Statistics for Single-Word Reading Efficiency Measures

	Time	Data	N	Mono		Range	N	EAL		Range	α	Effect Size of Group Difference	
				Mean (SD)	SD			Mean (SD)	SD			<i>g</i>	95% CI
TOWRE Sight Word Efficiency (Max = 108)	1	Raw	33	63.00 (9.41)	9.41	41-84	48	65.19 (8.32)	8.32	47-87	.76	-0.25	-0.70 to 0.21
		Std.		104.12 (11.47)	11.47	76-131		106.08 (11.28)	11.28	76-134			
	2	Raw	33	67.36 (9.60)	9.60	42-91	45	69.71 (8.60)	8.60	53-92	.85	-0.26	-0.71 to 0.21
		Std.		101.58 (13.46)	13.46	71-134		104.09 (11.77)	11.77	81-135			
	3	Raw	32	70.16 (7.67)	7.67	52-92	44	73.57 (8.92)	8.92	57-97	.83	-0.40	-0.87 to 0.07
		Std.		102.00 (10.84)	10.84	79-135		106.59 (12.68)	12.68	85-138			
TOWRE Phonemic Decoding Efficiency (Max = 66)	1	Raw	33	33.55 (10.92)	10.92	16-62	48	34.60 (9.58)	9.58	18-57	.83	-0.10	-0.56 to 0.35
		Std.		104.52 (14.17)	14.17	81-145		105.35 (11.89)	11.89	81-131			
	2	Raw	33	36.03 (9.47)	9.47	22-60	45	37.16 (8.91)	8.91	21-57	.89	-0.12	-0.58 to 0.34
		Std.		103.64 (12.53)	12.53	85-140		104.44 (10.88)	10.88	84-130			
	3	Raw	32	39.00 (9.39)	9.39	23-58	44	41.18 (9.06)	9.06	24-57	.83	-0.23	-0.71 to 0.24
		Std.		104.91 (12.32)	12.32	83-131		107.43 (11.95)	11.95	84-130			

Note: Std. = Standardised (Standard) Score; TOWRE = Test of Word Reading Efficiency; SW = Sight-Word Efficiency; PD = Phonemic Decoding Efficiency (pseudowords); standard scores and their standard deviations in italics; α = internal reliability (Cronbach alpha)

Table 4.7: t1-t3 Descriptive Statistics for Passage Reading Measures

	Time	Data	Mono			EAL			Effect Size of Group Difference		
			N	Mean (SD)	Range	N	Mean (SD)	Range	α	g 95% CI	
YARC Rate	1	Std.	33	104.09 (12.49)	79-130	48	103.27 (11.52)	82-130	.88	0.07	-0.39 to 0.52
		P3	32	105.00 (34.41)	52-188	44	109.16 (40.91)	51-230	.72	-0.11	-0.58 to 0.36
	2	Std.	33	104.97 (12.80)	79-131	46	104.93 (11.39)	87-131	.91	0.00	-0.46 to 0.46
		P3	33	87.67 (30.14)	52-183	46	85.41 (23.46)	43-141	.84	0.08	-0.38 to 0.54
	3	Std.	32	103.63 (12.60)	72-131	44	106.20 (11.78)	86-131	.89	-0.21	-0.68 to 0.26
		P3	32	78.72 (21.79)	50-144	44	78.00 (18.30)	37-119	.78	0.04	-0.43 to 0.51
YARC Accuracy	1	Std.	33	104.76 (11.10)	86-130	48	98.48 (8.83)	84-118	.77	0.63	0.17 to 1.10
		P3	32	3.69 (2.71)	0-12	44	4.38 (2.88)	1-15	.62	-0.16	-0.63 to 0.30
	2	Std.	33	102.21 (9.57)	83-124	46	100.57 (10.14)	83-122	.79	0.16	-0.30 to 0.63
		P3	33	3.69 (2.71)	0-12	46	4.38 (2.88)	1-15	.59	-0.16	-0.63 to 0.30
	3	Std.	32	101.69 (10.73)	82-127	44	101.18 (10.62)	82-122	.73	0.05	-0.42 to 0.52
		P3	32	2.63 (2.49)	0-11	44	3.59 (3.04)	0-14	.55	-0.34	-0.81 to 0.13
YARC Comprehension	1	Std.	33	100.39 (9.08)	89-127	48	97.81 (7.63)	83-117	.51	0.31	-0.15 to 0.77
		P3	32	5.00 (1.80)	2-8	44	4.27 (1.78)	1-8	.57	0.40	-0.07 to 0.88
	2	Std.	33	102.21 (8.74)	89-128	46	100.52 (6.50)	88-117	.57	0.22	-0.24 to 0.68
		P3	33	5.97 (1.40)	2-8	46	5.35 (1.42)	3-8	.53	0.48	0.01 to 0.95
	3	Std.	32	100.34 (8.98)	88-126	44	99.45 (6.98)	83-115	.55	0.11	-0.36 to 0.58
		P3	32	6.44 (1.08)	3-8	44	6.18 (1.30)	3-8	.61	0.21	-0.26 to 0.68

Note: Std. = Standardised (Standard) Score; YARC = York Assessment of Reading Comprehension Primary 2nd Edition; standardised scores and their standard deviations above, with passage 3 raw scores below; P3 raw rate measured in seconds taken to read passage; P3 raw accuracy measured as total number of questions answered correctly; P3 raw comprehension measured as total number of comprehension questions answered correctly; α = internal reliability (Cronbach alpha)

Passage 3 raw score data generally supported the trends found in that of the standard scores. In terms of reading rate, both groups on average read the passage out loud in a very similar amount of time, as measured in total seconds taken, across all time points. Group differences in passage 3 reading rate did not change markedly over time, remaining in the range of $g = 0.11$ to 0.08 . In terms of passage reading accuracy, a higher score on passage 3 indicates a higher number of total accuracy errors (mispronunciations, omissions, substitutions, etc.). Here, across all time points, children in the EAL group made on average slightly more accuracy errors ($g = 0.16$ at $t1$ and $t2$), and although both groups made fewer errors over time, the magnitude of this group difference increased at $t3$ ($g = 0.34$). Finally in terms of reading comprehension, passage 3 raw scores tended to agree with standard scores in showing a trend for more comprehension questions answered correctly by the monolingual group at each time point. In contrast to standard scores which showed a decrease in the magnitude of group differences over time, passage 3 raw scores show a widening of this discrepancy between $t1$ and $t2$ from $g = 0.40$ to 0.48 , and finally a decrease by $t3$ ($g = 0.21$). In general, group discrepancies in comprehension performance on passage 3 alone appeared larger in magnitude at each time point than for standard scores.

Distributions for passage 3 rate and accuracy were slightly positively skewed, reflecting a tendency for children to read passages quickly and make few accuracy errors on average; passage 3 accuracy, on the other hand, showed negative skew (max score = 8) as children tended to answer most questions correctly. Similar to standard scores, internal reliability for passage 3 raw scores was highest for passage reading rate (.72 to .84), but relatively lower for accuracy (.55 to .62) and comprehension (.53 to .57).

Finally, descriptive statistics for written narrative variables are presented in Table 4.8. In general, extra caution should be exercised in the interpretation of data relating to the written narrative measure, given the trend for children to write very little (compare, for instance, the grand mean of 22.73 C-units in the oral narrative task to 5.89 T-units in the written narrative task). Despite these limitations, patterns in written narrative variables bear some similarity to those in oral narrative. The EAL group tended to produce a higher number of T-units than the monolingual group at each time point, with this group difference first decreasing and then increasing in magnitude. At the same time, however, it was the monolingual group that tended to produce longer T-units at each time point (i.e. a higher MLTw). Contrary to patterns observed in many other variables, between $t2$ and $t3$ the EAL group plateaued in MLTw while the monolingual group continued to make a steady rate of progress. In terms of error rate per T-unit (morphosyntactic and spelling errors combined), the two groups approximated one another fairly closely over time. Again, a similar pattern was observed in which the EAL group plateaued between $t2$ and $t3$, whereas the monolingual group continued to show reductions in the rate of errors produced. Written narrative variables showed approximately normal distributions with the exception of total, morphosyntactic, and semantic error rate in which distributions were heavily positively skewed as many children made no errors. Internal reliability of written narrative measures was generally low, ranging from .22 to .77.

Classification of written errors into spelling and morphosyntactic types presents an interesting picture. In terms of spelling errors, the two groups showed fairly similar patterns, and there was a trend for children in the EAL group to make fewer errors at $t1$ and $t2$, but not at $t3$. However, the EAL group showed the opposite pattern in morphosyntactic error rate at $t1$, producing many more errors per T-unit than the monolingual group. In general, errors in tense morphology were very rare

4.3. Descriptive Statistics

in both groups of children, whereas errors in agreement and formal lexical devices (e.g. omitted or erroneous pronouns, prepositions, copula verbs, etc.) were far more frequently occurring in the EAL group³. Both spelling and morphosyntactic error rates were observed to decrease for both groups over time.

³Of the 17 agreement errors made at t1, 12 (71%) were made by EAL learners, and of the 33 errors in the use of formal lexical devices, 21 (94%) were made by EAL learners.

Table 4.8: t1-t3 Descriptive Statistics for Writing Measures

	Time	Mono			EAL			Effect Size of Group Difference	
		N	Mean (SD)	Range	N	Mean (SD)	Range	α	g 95% CI
T-Units	1	33	5.45 (2.31)	2-12	48	6.48 (2.54)	3-13	.47	-0.41 -0.87 to 0.05
	2	33	5.45 (2.12)	2-10	47	6.00 (2.20)	2-13	.47	-0.25 -0.71 to 0.21
	3	32	5.13 (2.06)	2-10	45	6.32 (2.41)	2-15	.42	-0.52 -0.93 to 0.02
MLTw	1	33	9.75 (2.83)	5.80-16.75	48	9.18 (2.29)	5.83-16.00	.22	0.22 -0.24 to 0.68
	2	33	10.82 (3.41)	7.25-21.00	47	10.13 (3.23)	5.38-17.50	.31	0.21 0.26 to 0.67
	3	32	11.59 (4.40)	5.00-21.50	45	10.12 (3.24)	4.60-19.50	.28	0.38 -0.09 to 0.86
Total Errors per T-Unit	1	33	0.88 (0.86)	0.00-4.00	48	0.82 (0.51)	0.14-2.60	.75	0.09 -0.37 to 0.54
	2	33	0.70 (0.66)	0.00-2.67	45	0.67 (0.60)	0.00-2.29	.64	0.05 -0.42 to 0.51
	3	32	0.49 (0.54)	0.00-2.33	45	0.60 (0.56)	0.00-2.25	.64	-0.21 -0.68 to 0.26
Spelling Errors per T-Unit	1	33	0.83 (0.84)	0.00-4.00	48	0.66 (0.47)	0.00-2.40	.77	0.26 -0.20 to 0.72
	2	33	0.64 (0.61)	0.00-2.67	45	0.54 (0.55)	0.00-2.00	.72	0.17 -0.30 to 0.63
	3	32	0.45 (0.53)	0.00-2.33	45	0.51 (0.56)	0.00-2.25	.60	-0.11 -0.58 to 0.36
Syntactic Errors per T-Unit	1	33	0.05 (0.10)	0.00-0.40	48	0.16 (0.15)	0.00-0.60	.29	-0.81 -1.29 to -0.34
	2	33	0.06 (0.12)	0.00-0.50	45	0.13 (0.19)	0.00-0.75	.29	-0.42 -0.89 to 0.05
	3	32	0.04 (0.11)	0.00-0.50	45	0.09 (0.14)	0.00-0.67	.39	-0.41 -0.89 to 0.06

Note: T-Units = total number of T-units; MLTw = mean length of T-unit in words; Errors per T-Unit = total errors including spelling and syntactic errors; α = internal reliability (Cronbach alpha)

4.3.6 Summary of Trends in Descriptive Statistics

The research questions of the longitudinal cohort study concern the prevalence of group similarities and differences in performance at t1 as well as developmental trajectories over time. From the descriptive statistics presented thus far, a number of trends are observable. In general, after four years of formal English-medium instruction at t1, EAL learners showed a trend for lower levels of receptive and expressive vocabulary knowledge as well as vocabulary depth, a higher rate of morphosyntactic errors and shorter utterances in both speech and writing, and lower accuracy and comprehension in passage reading (although still within the normal range). On the other hand, trends for EAL group advantages were found in fluency tasks requiring the rapid naming of digits and letters, and reading efficiency of words and pseudowords. Both groups tended to make similar rates of progress over time, meaning that in some cases, group discrepancies at t1 were still evident by the end of the study at t3. Both groups tended to make steady rates of progress between t1 and t2, after which point one or both groups decelerated (for example in expressive vocabulary, vocabulary depth, listening comprehension, rapid digit naming, and reading comprehension). This may have been an artefact of the organisation of the academic year, in particular the intervening summer holiday after the end of t2.

As a qualification to these results, caution should be exercised when considering the monolingual comparison group: as demonstrated by standardised scores in descriptive tables, this group tended to perform within the lower average range on a number of measures. In other words, convergence between the EAL and monolingual group over the course of the study may appear encouraging, but it should be borne in mind that the language skills of this comparison group are were lower than expected relative to norming populations of assessments.

4.4 Linear Mixed Modelling

A linear mixed modelling (LMM) framework was employed as an inferential analytical strategy for the investigation of group trajectories in language and literacy skills between t1 and t3. As an extension of the general linear model (see Equation 4.1 below), LMMs allow increased flexibility and accuracy in the modelling of repeated measures data. Repeated measurements of the same individuals over time are likely to be correlated, violating the assumption of independence in traditional statistical procedures such as repeated measures (RM) ANOVA (Wainwright, Leatherdale & Dubin, 2007; West, Welch & Galecki, 2007). While correction procedures are available for RM ANOVA (e.g. Greenhouse & Geisser, 1959; Huynh & Feldt, 1976), the explicit modelling of non-independency between repeated measurements often leads to more efficient estimation of model parameters and smaller standard errors (Burton et al., 1998; Gibbons, Hedeker & DuToit, 2010; Osborne, 2008; West et al., 2007). Additional advantages of LMMs include the ability to model time continuously rather than categorically (Muth et al., 2016), and better handling of unbalanced and missing data (i.e. where ANOVA would typically require equal group sizes and would employ listwise deletion to any subjects with missing data).

$$Y_{ij} = \beta_0 + \beta_1 + u_i + \epsilon_{ij} \quad (4.1)$$

LMMs incorporate both fixed and random effects (hence 'mixed' model). Variables such as group, age, or time may be considered fixed effects in a LMM if explicit interest lies in the relationship between the levels of these factors (β) and the outcome variable (Y). In contrast, subject is considered a random effect if the goal of the study does not concern explicit comparison amongst individuals, who represent a random sample from a larger population (West et al., 2007). The inclusion of random effects (μ_i and ϵ_{ij} in Equation 4.1) allows for the estimation of an individual intercept and slope for each subject, thus minimising unexplained variance and accounting for within-subject correlation among residuals (Gibbons et al., 2010). Random effect variance components in LMMs include: $\text{Var}(\mu_{0j})$, variance of within-subject intercepts; $\text{Var}(\mu_{1j})$, variance of within-subject slopes, and; $\text{Cov}(\mu_{0j}, \mu_{1j})$, covariance of intercepts and slopes (Bryk & Raudenbush, 1992). Rather than estimate each slope and intercept for each subject, the model estimates a single variance, thus preserving degrees of freedom and statistical power. A graphical explanation of linear mixed modelling is presented in Figure 4.1 overleaf.

4.4.1 Goodness-of-fit and Model Building

LMMs employ maximum likelihood (ML) estimation, an iterative process with the aim of optimising a likelihood function in order to provide a best fit to the observed data (West et al., 2007). A likelihood-ratio test is used for purposes of model comparison, as the log-likelihood follows a χ^2 distribution. When comparing a reduced, intercept-only model with a full, random-slope model complete with fixed effects (e.g. *time* and *group*), the likelihood ratio statistic is derived by multiplying -2 times the log-likelihood (-2LL). While the -2LL is itself uninterpretable, a smaller figure is considered to indicate better model fit (West et al., 2007). ML estimation is ideal for the purposes of model comparison as it calculates likelihood values based on raw values of the dependent variable. However, Restricted Maximum Likelihood (REML) estimation is recommended once a final model has been chosen, as this produces unbiased estimates of covariance parameters by virtue of utilising residuals rather than raw scores (West et al., 2007).

Additional measures of model fit for LMMs include Akaike's Information Criterion (AIC), a statistic which adds a penalty for the inclusion of any fixed effect that does not improve model fit. Similar to the -2LL, AIC is itself uninterpretable, but a smaller statistic indicates better fit. In traditional linear regression modelling, R^2 provides a standardised measure of proportion of explained variance: as traditional R^2 is not appropriate for LMMs due to partitioning of fixed and random effects (Singer & Willett, 2003), Nakagawa and Schielzeth (2013) propose a pseudo R^2 statistic for linear mixed models, deriving both R^2 -marginal (proportion of variance explained at level-1, i.e. within-subjects) and R^2 -conditional (a measure of both within- and between-subject variance explained by the model).

In the present analyses, a step-up model building strategy was employed (see Table 4.9 below). The step-up strategy begins by specifying an 'empty' or 'unconditional' growth model (UGM) containing only an intercept and random effects, representing level-1 *within-subject* change over time. Next, level-2 fixed effects are added to the model, representing *between-subject* change over time as a function of *group* and *time* (the 'full' growth model; FGM). A step-up strategy is recommended in multilevel modelling, as any fixed effects that do not add meaningfully to a model by virtue of improving fit are removed in the interests of parsimony and to avoid overfitting (Nezlek,

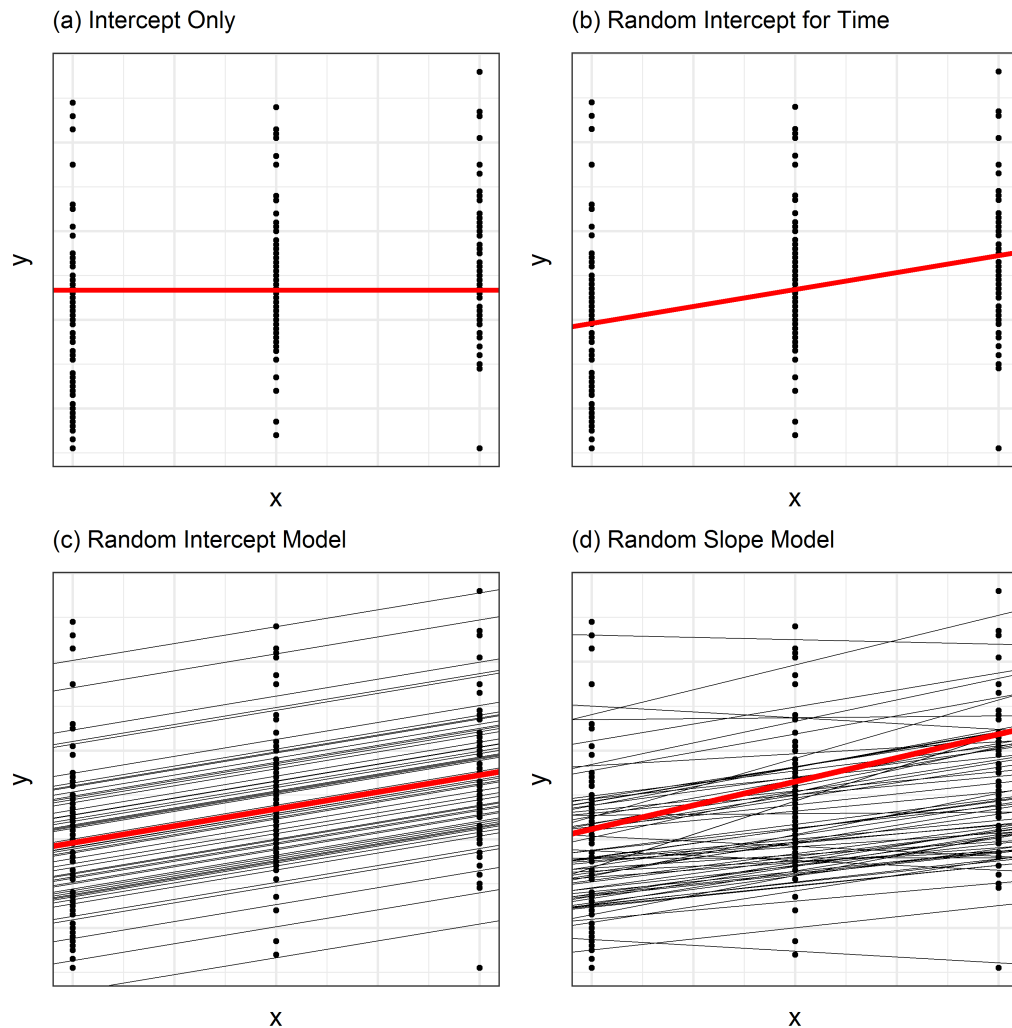


Figure 4.1: Graphical Explanation of Linear Mixed Modelling. Note: Panels represent repeated measurements of a continuous outcome variable (y) across three time points (x). Together, panels (a) and (b) represent traditional linear regression modelling, whereas panels (c) and (d) build on this model by incorporating additional random effects. Panel (a) is an intercept-only model with the intercept (red line) representing the grand mean of y . Panel (b) models y as a function of x , i.e. rather than being fixed to the grand mean value, the slope is allowed to vary over time. Panel (c) includes a hypothetical random intercept for each subject (thin black lines); here, the intercept is free to vary while its slope is fixed and identical to that of the grand mean introduced in panel (b). In panel (d) both the intercept and the slope of each subject are free to vary, and do not necessarily conform to those of the grand mean. The random effect of μ is the difference between the overall grand mean trajectory and that of each subject, while the error (ϵ) is the distance between each subject's actual data points and their slope.

2013; Singer & Willett, 2003; West et al., 2007). However, in the present analysis the fixed effects of *time* and *group* were maintained in models even when their inclusion did not improve model fit⁴. This decision was taken in order to be able to examine developmental trajectories which, even if not statistically significant, shed light on the developmental patterns of variables under investigation.

Table 4.9: Step-Up Linear Mixed Model Building Process

Step	Comments
1	Model 1: Random intercept only (UGM)
2	Model 2: Addition of random slopes.
3	Likelihood ratio test to establish whether the addition of random slopes improves model fit (criterion for retention = $\chi^2(df) p \leq .05$). If $p \geq .05$, revert to Model 1.
4	Model 3: Addition of fixed effects (<i>time</i> and <i>group</i>).
5	Model 4: Addition of <i>time</i> × <i>group</i> interaction.
6	Likelihood ratio test to establish whether Models 3 and 4 provide a statistically significant improvement in model fit over Model 1 or 2.
7	Model 5: Remove any non-significant fixed effects (if any) and re-run model with REML (FGM).
8	Calculate reduction in AIC and residual variance between UGM and FGM; calculate pseudo R ² , assess model diagnostics.

Note: UGM = unconditional growth model; FGM = full growth model; ICC = intraclass coefficient; REML = restricted maximum likelihood estimation; AIC = Akaike's information criterion.

4.4.2 General Procedure and Model Assumptions

In the linear mixed modelling analyses that follow, UGMs consisted solely of the grand mean of the continuous outcome variable plus random intercepts and/or slopes for subject. FGMs consisted additionally of the fixed effects of *time*, *group* and a *time* × *group* interaction, where the latter was found to add meaningfully to the model. With three repeated measurements, models were restricted to linear trajectories only, as additional measurements would be required for the fitting of non-linear or quadratic growth terms (Singer & Willett, 2003; Law et al., 2008). All linear mixed models were generated using the `lme4` software package (Bates, Maechler, Bolker & Walker, 2015) in R (R Core Development Team, 2017).

Reported statistics from LMMs include fixed effect coefficients (β) and their standard error (SE), *t*-statistics, variance estimates for the random effects of intercept and slope, residual terms for random effects, and AIC. Statistical significance of fixed effects is reported using the Kenward and Rogers (1997) scaled Wald Z-statistic and Satterthwaite approximation of degrees of freedom, which takes into account variation in estimation of the variance-covariance matrix and is found to provide more accurate estimates in small-sample studies (Verbeke & Molenberghs, 2009; available in the `lmerTest` package of Kuznetsova, Brockhoff & Christensen, 2016). Model

⁴Note that this occurred in only five instances (Models 15 to 17 for YARC passage reading rate, accuracy, and comprehension [all utilising scaled scores], and Models 20.1 and 20.2 for morphosyntactic and semantic error rate in writing, respectively)

4.4. Linear Mixed Modelling

fit is reported with AIC statistics for UGMs and FGMs, where a reduction in AIC is interpreted as an improvement in model fit. Raw AIC values are reported in tables, whereas change in AIC for each step in the model building process is reported in-text⁵. The intraclass correlation (ICC) coefficient is reported for the FGM as a measure of consistency across subjects over time, where a high ICC is interpreted as a high level of within-subject consistency and as justification for the inclusion of random effects in the model (Bliese & Ployhart, 2002; Burton et al., 1998). Marginal and conditional pseudo R^2 statistics are also reported, using the `MuMIn` package in R (Bartón, 2015). The `lme4` package also allows for the separate estimation of intercepts and slopes by group: these analyses were conducted for each model in order to answer questions regarding group differences in developmental trajectories (reported in-text).

Basic checks for univariate normality included assessment of Q-Q plots, boxplots, and histograms, and calculation of the proportion of data points with a z-score of ≥ 1.96 , 2.58, and 3.29 (given the assumption of a normal distribution, no more than 5% of z-scores should lie above 1.96; Field, 2012). Justification for the removal of outliers is provided when applicable. As a regression framework, linear mixed modelling is subject to certain underlying distributional assumptions; specifically, residuals (μ) and estimated random effects are assumed to be normally distributed (i.e. centered at zero), and to have constant variance across covariates (Pinheiro & Bates, 2000). The `lme4` package provides a number of exploratory data analysis tools and graphing capabilities for the investigation of LMM assumptions. For each fitted model, the following plots were generated and interpreted: histograms of all residuals; boxplots of residual variances by subject; boxplots of residual variances disaggregated by *group* and *time*; scatterplots of fitted versus observed values as an indication of models' accuracy in explaining the data; and normal plots for estimated random effects (Best Linear Unbiased Predictors). Pearson standardised residuals were utilised in order to compare plots across different models.

Traditional tests for the presence of highly influential observations (i.e. those with high leverage) include Mahalanobis distance and Cook's D . However, such tests are deemed not to be appropriate for LMMs due to their hierarchical structure, correlated errors, and inclusion of random effects (Bannerjee & Frees, 1997; Nieuwenhuis, Grotenhuis & Pelzer, 2012). Alternatively, case-deletion diagnostics provide one solution by iteratively deleting each subject (or observation) and then refitting the model in order to observe changes in coefficients and model fit (West et al., 2007). This procedure was carried out for each FGM using the `influence.ME` package in R (Nieuwenhuis et al., 2012), and particular attention was paid to any changes in the t -value of the fixed effect of *group*. Given the repeated-measures nature of the data, preference was given to the removal of individual influential observations rather than whole subjects in order to preserve data⁶.

ICCs of all UGMs were positive, ranging from 0.12 to 0.88 (mean = 0.47), which may be interpreted as a justification for the inclusion of random effects, indicating a degree of non-

⁵For example, Δ AIC is read as 'change in AIC' between different models.

⁶As alluded to in Section 4.4, a major advantage of linear mixed modelling is the flexible handling of missing data. Just as LMM models are not required to resort to listwise deletion in the face of *missing* data, they similarly have the option not to do so in the case of *influential* data. For example, in the vector [5, 6, 15], only the third data point represents an outlier, and LMM has the option to remove only this data point, rather than the whole vector. Where possible, this strategy was applied in the present study. In some cases, however, only whole subjects and not individual observations were found to be highly influential, in which case these subjects were removed.

independence, i.e. within-subject consistency over time (Bliese & Ployhart, 2002). Due to space limitations, ICCs are reported for FGMs only. Order of analysis will follow the categorisations of dependent variables as presented in Table 4.1 on page 74, beginning with vocabulary and oral language measures, and then moving on to phonological processing and literacy measures. Each analysis will be structured identically, beginning with a conceptual summary of findings and then detailing the model fitting process. Graphical representations of group trajectories are provided in figures in order to aid interpretation. Shaded areas on figures represent standard error (in keeping with descriptive tables, green represents monolingual and blue represents EAL) and where available, population norming means of assessments are depicted in equal-dashed lines (i.e. $y = 10$ [scaled score] or 100 [standard score]). A summary of group-specific intercepts and slopes for all models is provided in Table 4.21 before the Discussion.

4.4.3 Linear Mixed Modelling of Vocabulary Measures

Linear modelling of group differences and developmental trajectories in English vocabulary knowledge (Models 1 to 3, Table 4.4.3) confirms trends observed in the interpretation of descriptive statistics in Section 4.2. The models present a situation in which children with EAL underperform in relation to their monolingual peers at t1, and due to their very similar rate of progress to their monolingual peers, do not 'close the gap' over time; in other words, differences in performance established by t1 do not change to a large degree over time. Indeed, even where there is evidence of such convergence between the groups (i.e. in CELF EV), this appears to be due to a deceleration of the monolingual group rather than a higher slope of the EAL group (note that this was not due to the monolingual group reaching a ceiling level on any assessment). Finally, it is interesting to note that group differences were least apparent in vocabulary depth (WISC VC), suggesting that the two groups possess qualitatively similar knowledge of the same subset of words and that more prevalent differences are to be found in breadth of knowledge (BPVS; CELF EV).

Table 4.10: Linear Mixed Modelling of Vocabulary Measures

	Model 1: BPVS		Model 2: CELF EV		Model 3: WISC VC	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects						
Intercept	86.65 (3.63)	23.87 **	35.56 (1.17)	30.26 **	24.76 (1.07)	23.15 **
Time	7.66 (0.46)	16.63 **	2.49 (0.24)	10.25 **	2.08 (0.18)	11.56 **
Group: EAL	-7.94 (3.35)	-2.37 *	-3.77 (1.32)	-2.85 **	-1.86 (1.36)	-1.37
Time × Group	-	-	-	-	-	-
Random Effects						
Intercept Variance	201.76		41.00		28.81	
Slope Variance	1.82		0.19		0.10	
Residual Variance	28.28		8.75		4.79	
Change in Residual (%)	-68.81		-41.98		-47.90	
UGM AIC	1908.60		1471.30		1391.40	
FGM AIC	1729.84		1384.48		1295.93	
ICC	0.88		0.82		0.86	
Pseudo R ² -marginal	0.19		0.16		0.09	
Pseudo R ² -conditional	0.90		0.82		0.89	
Cases; Obs.	81; 237		81; 235		81; 235	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed);
Raw scores utilised in all models

4.4.3.1 Receptive Vocabulary (BPVS)

BPVS raw data for both groups at each time point approximated a normal distribution, although there were consistent outliers: aggregated data across time points revealed 6.75% of observations with a z-score of ≥ 1.96 , and 0.42% ≥ 2.58 , slightly higher than expected given the assumption of normally distributed data (Field, 2012). Despite the presence of potential outliers, case-deletion diagnostics did not indicate the presence of any highly influential observations which significantly altered t -values of any fixed effects, and thus all available data were used in the model. Mixed model assumptions were satisfied, as residuals were approximately normally distributed and centered at zero, and although there was a fairly high degree of variability, this was fairly constant across groups and time points. Estimated random effects approximated a normal distribution. Details of model fitting follow below.

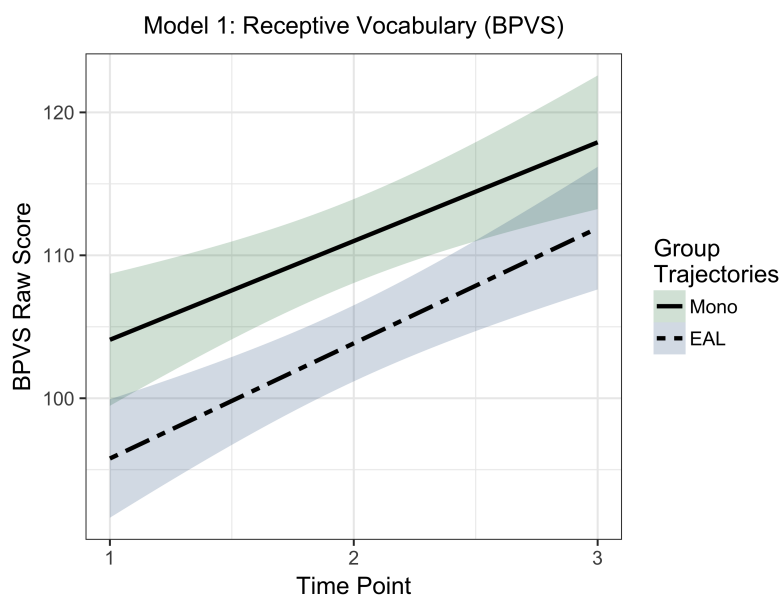


Figure 4.2: Linear Mixed Modelling of t1-t3 Receptive Vocabulary

The addition of a random slope term to the UGM resulted in improved model fit ($\Delta\text{AIC} = -36.51$, $\chi^2(2) = 40.51$, $p < .001$). Next, the addition of fixed effects of *time* and *group* continued to improve model fit to a large degree ($\Delta\text{AIC} = -121.63$, $\chi^2(2) = 125.63$, $p < .001$), but not the *time* \times *group* interaction ($\Delta\text{AIC} = 1.93$, $\chi^2(1) = 0.07$, $p = .789$) which resulted in an increase in AIC. Thus the FGM, fitted using REML estimation, consisted of a random intercept and slope, and the fixed effects of *time* and *group* only: *time* was a highly statistically significant predictor of performance, with all children correctly identifying an additional 7.66 words on the BPVS on average for each subsequent time point ($p < .001$). Growth in receptive vocabulary knowledge differed significantly across groups, with children in the EAL group correctly identifying on average 7.94 fewer words than their monolingual peers ($p = .020$). The final model represented a good fit to the data (pseudo R^2 -conditional = 0.90), and indicated a fairly high level of within-subject consistency (ICC = 0.86), suggesting that children tended to maintain their positions over time relative to their position at t1.

The two groups differed to a larger degree in initial levels of knowledge at t1 than in rate of progress over time (Intercepts: Mono = 96.43; EAL = 88.02; Slopes: Mono = 7.47; EAL = 7.75; Figure 4.2 – note that graphs indicate linear trajectories of each group from t1 to t3; shaded area represents standard error). Thus, the model supported trends in descriptive statistics for a subtle closing of the gap in receptive vocabulary knowledge between the two groups (i.e. indicated by the slightly steeper slope of the EAL group and the relatively smaller effect size of the group difference at t3, $g = 0.38$). Note, however, that although the group performance discrepancy at t1 was reduced by the end of the study, a $time \times group$ interaction term did not explain additional variance in this model.

4.4.3.2 Expressive Vocabulary (CELF EV)

Raw expressive vocabulary (CELF EV) data did not indicate any significant deviations from univariate normality for either group at any of the three time points, and z-scores of ≥ 1.96 accounted for 4.25% of the data. Residuals clustered around zero, with slightly higher variance in the EAL group, and random effects approximated a normal distribution. Case-deletion diagnostics did not indicate the presence of any highly influential subjects or observations, and thus all available data were utilised.

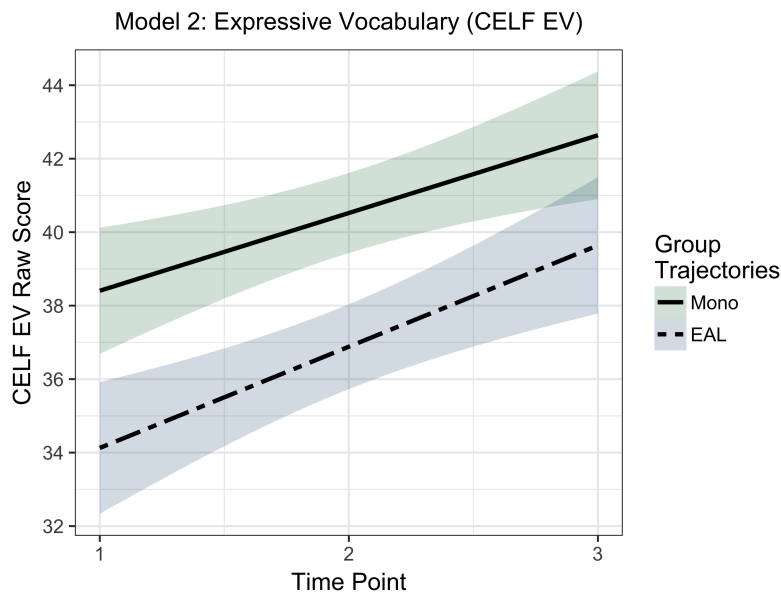


Figure 4.3: Linear Mixed Modelling of t1-t3 Expressive Vocabulary

The addition of a random slope term to the UGM resulted in improved model fit ($\Delta AIC = -13.18$, $\chi^2(2) = 17.18$, $p < .001$). The fixed effects of $time$ and $group$ also improved model fit to a moderate degree ($\Delta AIC = -71.31$, $\chi^2(2) = 75.31$, $p < .001$), but not the $time \times group$ interaction term ($\Delta AIC = 1.45$, $\chi^2(1) = 0.55$, $p = .458$), which was consequently removed. $Time$ was a significant predictor of children's performance in expressive vocabulary, with an additional average of 2.49 words correctly identified at each subsequent time point ($p < .001$). Group differences were also apparent, as children with EAL identified an average of 3.77 fewer words than their monolingual peers ($p = .006$). The final model represented a good fit to the data ($R^2_{conditional} = 0.82$),

4.4. Linear Mixed Modelling

and an ICC of 0.82 indicated a relatively high level of within-subject consistency in performance over time.

Again, as with receptive vocabulary, the two groups differed more in initial knowledge at t1 than in rate of progress over time. However, there was a trend for slightly faster progress in the EAL group, resulting in a reduction in between-group effect size at t3 (Intercepts: Mono = 36.07; EAL = 31.43; Slopes: Mono = 2.28; EAL = 2.64; Figure 4.3). However, it is important to bear in mind that, as noted in discussion of descriptive statistics, there was evidence for non-linearity in the groups' progress: while both groups made similar rates of progress between t1 and t2, the monolingual group plateaued somewhat between the final two time points, a pattern which did not apply to the EAL group.

4.4.3.3 Vocabulary Depth (WISC VC)

Although extreme observations (z -scores ≥ 1.96) accounted for only 4.68% of WISC VC data, there were three particularly high-scoring children in the monolingual group (1.7% of z -scores ≥ 2.58 and 0.43% ≥ 3.29). However, case-deletion diagnostics revealed that deletion of these observations did not result in significantly different parameter estimates. Thus, although such cases may be considered outliers, they did not exert any undue influence on the fit of the model and thus were retained in the dataset (Nieuwenhuis et al., 2012; West et al., 2007). Residuals displayed a normal distribution, being centered at zero for both groups at each time point. Random effects also approximated a normal distribution, although with a small number of high intercept terms.

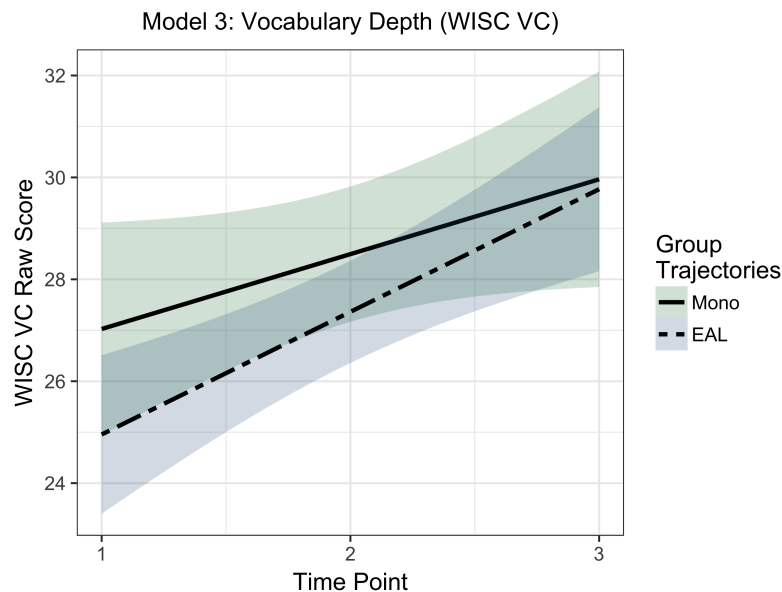


Figure 4.4: Linear Mixed Modelling of t1-t3 Vocabulary Depth

The addition of a random slope model to the UGM resulted in significantly improved model fit ($\Delta\text{AIC} = -16.54$, $\chi^2(2) = 20.54$, $p < .001$). The fixed effects of *time* and *group* additionally improved fit ($\Delta\text{AIC} = -77.11$, $\chi^2(2) = 81.11$, $p < .001$) but not the *time* \times *group* interaction ($\Delta\text{AIC} = 0.30$, $\chi^2(1) = 1.70$, $p = .192$), which was dropped from the model. For each subsequent time

point, children scored an average of 2.08 points higher in vocabulary depth: while this could have been as a result of deepening of lexical knowledge, children tended to give similar answers across repeated measurements, and growth on this assessment is therefore more likely to reflect the acquisition of novel vocabulary. Although there was a tendency for the monolingual group to perform higher than the EAL group ($\beta = -1.86$), this group difference did not achieve statistical significance ($p = .174$). Similar to Models 1 and 2, the final model for vocabulary depth represented a good fit to the data ($R^2_{\text{conditional}} = 0.89$) and indicated a high degree of within-subject consistency over time ($ICC = 0.86$).

The two groups differed both in their initial levels of knowledge and progress over time (Intercepts: Mono = 25.07; EAL = 22.64; Slopes: Mono = 1.84; EAL = 2.29; Figure 4.4). Thus linear modelling supports trends in descriptive statistics which show convergence between the two groups between t1 ($g = 0.27$) and t3 ($g = -0.03$).

4.4.4 Linear Mixed Modelling of Other Oral Language Measures

Other oral language measures included listening comprehension (CELF USP) and expressive grammar (CELF FS; Models 5 and 6 in Table 4.3). Firstly, although there was a significant effect of *time* for listening comprehension performance, this appeared to be accounted for by the EAL group alone; in other words, the EAL group continued to make a positive rate of progress over time, in contrast to the monolingual group whose slope was near horizontal, indicating similar performance at each time point. Therefore, the group discrepancy in listening comprehension performance at t1 was no longer evident by t3. Secondly in terms of expressive grammar performance, linear modelling confirmed trends evident in descriptive statistics (Table 4.3) in which the monolingual group obtained a higher score at t1 and, due to the similar developmental trajectories of both groups, maintained this advantage by t3. Despite this trend, the effect of *group* was non-significant in both models. Details of model fitting procedures follow below.

Table 4.11: Linear Mixed Modelling of Oral Language Measures

	Model 4: CELF USP		Model 5: CELF FS (Raw)		Model 5.1: CELF FS (17)	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects						
Intercept	7.53 (0.47)	16.13 **	29.89 (1.13)	26.43 **	19.40 (0.89)	21.81 **
Time	0.28 (0.14)	2.04 *	5.38 (0.34)	15.73 **	2.43 (0.27)	8.90 **
Group: EAL	-0.67 (0.50)	-1.35	-2.15 (1.14)	-1.89	-2.60 (0.86)	-3.01 **
Time \times Group			-	-		
Random Effects						
Intercept Variance	3.81		22.85		18.88	
Slope Variance	-		0.05		0.79	
Residual Variance	2.88		17.84		10.03	
Change in Residual (%)	-2.80%		-61.84		-39.83	
UGM AIC	1059.90		1620.30		1418.60	
FGM AIC	1060.59		1447.16		1345.94	
ICC	0.57		0.56		0.65	
Pseudo R ² -marginal	0.02		0.36		0.21	
Pseudo R ² -conditional	0.58		0.69		0.64	
Cases; Obs.	81; 237		80; 232		81; 232	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed)
 Scaled scores used in Model 4; raw scores in Model 5; 17 commonly attempted items in Model 5.1

4.4.4.1 Listening Comprehension (CELF USP)

CELF USP data were broadly normally distributed across groups and time points, although with slightly more extreme observations than would be expected under a normal distribution (6.33% of z-scores ≥ 1.96 ; Field, 2012). As described in Section 4.3.2, at t2 distributions were rather positively skewed in both groups due to a relatively large number of low scores. This was likely a result of a number of children passing between the 8-9 and 9-11 age thresholds of the CELF USP. Since children were attempting different passages, the decision was made to use scaled scores rather than raw scores in the LMM. Analysis revealed no threats to model assumptions, as residuals centered around zero and variance was constant across groups and time points. Additionally, random effects approximated a normal distribution and case-deletion diagnostics did not reveal any highly influential subjects or observations.

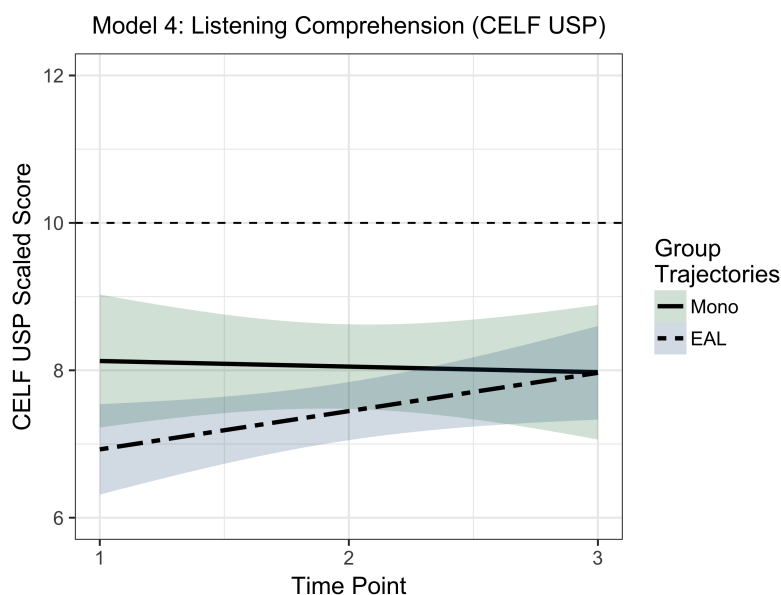


Figure 4.5: Linear Mixed Modelling of t1-t3 Listening Comprehension

The addition of a random slope term to the UGM did not result in improved model fit ($\Delta AIC = 2.03$, $\chi^2(2) = 1.97$, $p = .374$) and was therefore dropped from the model. The fixed effects of *time* and *group*, however, did contribute towards explaining variance in the intercept-only model ($\Delta AIC = -2.04$, $\chi^2(2) = 6.04$, $p = .048$), although not the *time* \times *group* interaction which merely approached statistical significance ($\Delta AIC = -1.64$, $\chi^2(1) = 3.65$, $p = .056$), thus failing to confirm the trend for the opposite directions of each group's trajectory. Of the fixed effects, *time* was a significant predictor of performance ($p = .043$) but not *group* ($p = .180$). The final model represented only a modest fit to the data (R^2 -conditional = 0.58), and children's performance over time was somewhat less consistent than that observed in previous models (ICC = 0.57).

The two groups showed divergence in their performance at t1 as well as their rate and direction of progress over time (Intercepts: Mono = 8.13; EAL = 6.43; Slopes: Mono = -0.02; EAL = 0.50; Figure 4.5). The relatively steeper slope of the EAL group suggested some degree of closing of the gap over time, with only a small group difference in scaled scores by t3 ($g = -0.08$; Table 4.3).

4.4.4.2 Expressive Grammar (CELF FS)

CELF FS data approximated a normal distribution in terms of z-scores (4.74% ≥ 1.96 ; 0.43% ≥ 2.58). Similar to the CELF USP, examinees on the FS subtest attempt slightly different subsets of stimuli depending on their age (see Section 3.4.3.2). However, all examinees in the present study attempted a set of 17 items at each time point, and therefore a separate analysis is presented for the groups' progress on these items below.

Unlike in the case of CELF USP, distributions for CELF FS were not as heavy-tailed at t2. Descriptive statistics (Table 4.3) revealed a clear time trend for both groups of children, with a slight deceleration in the EAL group between t2 and t3 (again, not due to reaching the ceiling of the subtest scoring range). Despite slightly larger residual variance for the EAL group, residuals centered around zero, and random effects approximated a normal distribution. Case-deletion diagnostics did reveal one subject from the monolingual group to be significantly influential on the group parameter estimate in the final model. Deletion of this subject altered the *t*-value for the *group* parameter estimate in the final model. Deletion of this subject altered the *t*-value for the *group* parameter estimate from -2.16 to -1.89 and subsequently altered the *p*-value associated with this fixed effect from *p* = .034 to .063, suggesting a high degree of leverage. Thus, the decision was taken not to include this child in the analysis of the expressive grammar data (Nieuwenhuis et al., 2012).

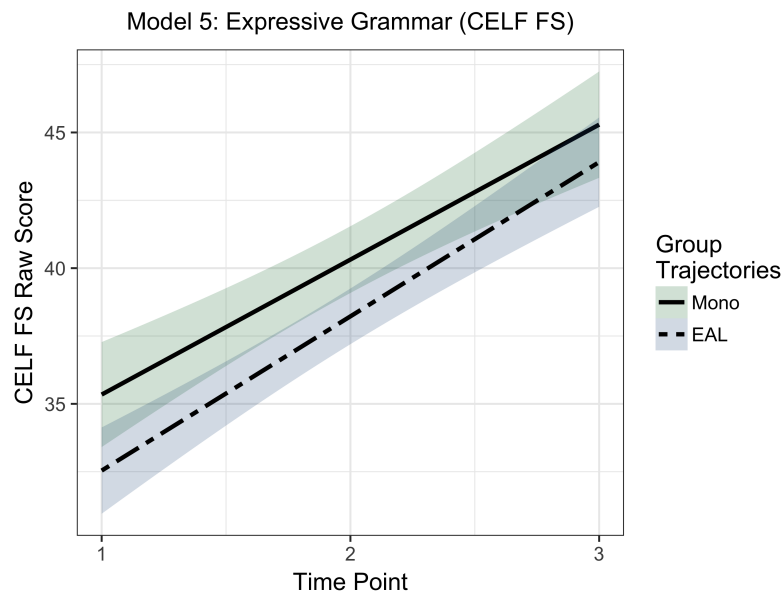


Figure 4.6: Linear Mixed Modelling of t1-t3 Expressive Grammar

The addition of a random slope term improved model fit ($\Delta\text{AIC} = -30.51$, $\chi^2(2) = 34.51$, $p < .001$), as did the inclusion of fixed effects of *time* and *group* to a substantial degree ($\Delta\text{AIC} = -114.62$, $\chi^2(2) = 118.62$, $p < .001$). The *time* \times *group* interaction did not result in improved fit and was removed from the model ($\Delta\text{AIC} = 1.34$, $\chi^2(1) = 0.66$, $p = .415$). For each subsequent time point, children scored an average of 5.38 points higher ($p < .001$), and children with EAL tended to score on average 2.14 points lower than their monolingual peers, although this pattern was not statistically significant ($p = .063$). A modest level of within-subject consistency was evidenced by an ICC of 0.56, and the final model represented a modest fit to the data ($R^2\text{-conditional} = 0.69$).

Children in the monolingual group scored higher at t1, but the children in the EAL group generally made a faster rate of progress over time, again demonstrating a modest degree of closing of the gap (Intercepts: Mono = 30.23; EAL = 26.93; Slopes: Mono = 5.09; EAL = 5.63; Figure 4.6). As shown in the descriptives (Table 4.3), between t1 and t2 the two groups appeared to converge in performance somewhat ($g = 0.56$ to 0.21), but then to diverge once more between t2 and t3 ($g = 0.40$).

Next, a separate analysis for the 17 commonly attempted items on the CELF FS at each time point was conducted (items 8 to 24; see Model 5.1 in Table 4.11 and Figure 4.7 above). Raw data approximated a normal distribution (5.96% of z-scores ≥ 1.96), and residuals and random effects centered around zero. Case deletion diagnostics did not reveal the presence of any highly influential subjects or observations.

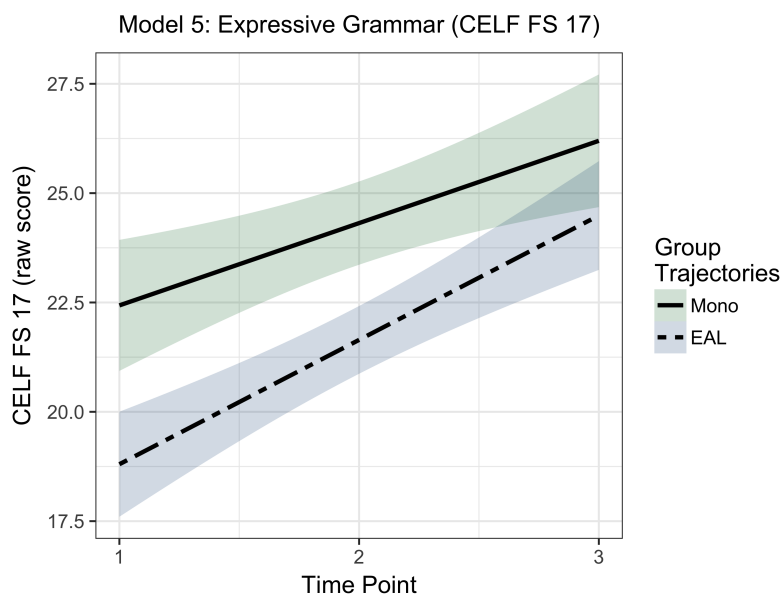


Figure 4.7: Linear Mixed Modelling of t1-t3 Expressive Grammar (Common Items)

The addition of a random slope significantly improved model fit ($\Delta\text{AIC} = -11.82$, $\chi^2(2) = 15.82$, $p < .001$), as did fixed effects of *time* and *group* ($\Delta\text{AIC} = -59.97$, $\chi^2(2) = 63.97$, $p < .001$). However, a *time* \times *group* interaction term did not improve fit ($\Delta\text{AIC} = 0.46$, $\chi^2(2) = 2.46$, $p = .116$) and was removed from the model. Similar to Model 5 for CELF FS raw scores, in the FGM for CELF FS commonly attempted items, all children made significant progress over time ($\beta = 2.43$, $p < .001$), but in contrast to Model 5, monolingual children significantly outperformed their monolingual peers ($\beta = -2.60$, $p = .003$). Again, similar to measures discussed thus far, EAL learners began at a lower intercept at t1, but made a faster rate of progress over time relative to their monolingual peers (Intercepts: Mono = 20.44; EAL = 15.99; Slopes: Mono = 1.96; EAL = 2.82).

4.4.5 Linear Mixed Modelling of Oral Narrative Measures

Oral narrative measures from the Peter and the Cat retell task, including total number of utterances (C-units), mean length of utterance in words (MLUw), lexical diversity (Root TTR), and

4.4. Linear Mixed Modelling

error rates are represented in Models 6 to 9 in Table 4.12. Additional models for morphosyntactic and semantic errors are presented in Table 4.13.

Linear mixed modelling confirmed trends found in descriptive statistics (Table 4.4). Firstly, children in the EAL group produced significantly more utterances than their monolingual peers, although this group effect was small in practical terms, equating to a difference of fewer than two utterances (see Model 6, Table 4.12). Secondly, the greater verbal productivity of the EAL group was complemented by the significantly longer MLU of the monolingual group. Unfortunately, the relatively short length of the Peter and the Cat story places limitations on the confidence of this finding, as had the story been longer, children may have produced more of a range of sentence types. Thirdly, the retellings of the two groups did not differ significantly in their lexical diversity, suggesting that children from both groups tended to employ the same range of vocabulary. Finally, the utterances of children with EAL contained significantly more errors, on average, than those of their monolingual peers. This was accounted for in fairly equal measure by semantic and morphosyntactic error types.

In order to provide context to the results presented here, the original Peter and the Cat narrative was transcribed and analysed in CLAN (MacWhinney, 2000) as if it were a retelling. This yielded a total of 25 utterances, 260 tokens (total words produced), 138 types (number of different words), a mean length of utterance in words (MLUw) of 10.4, and a lexical diversity ratio (Root TTR) of 8.56. Discussion of model fitting procedures below.

Table 4.12: Linear Mixed Modelling of Oral Narrative Measures

	Model 6: Utterances		Model 7: MLUw		Model 8: Lexical Diversity		Model 9: Total errors per utterance	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects								
Intercept	22.56 (0.84)	26.73 **	7.14 (0.21)	33.38 **	6.09 (0.10)	60.61 **	0.076 (0.016)	4.85 **
Time	-0.67 (0.34)	-1.99 *	0.59 (0.07)	7.59 **	0.24 (0.03)	7.44 **	-0.003 (0.005)	-0.63
Group: EAL	1.96 (0.66)	3.00 **	-0.68 (0.19)	-3.53 **	-0.07 (0.09)	-0.69	0.050 (0.015)	3.28 **
Time × Group	-	-	-	-	-	-	-	-
Random Effects								
Intercept Variance	2.27		0.41		0.21		0.003	
Slope Variance	-		-		0.01		-	
Residual Variance	17.82		0.92		0.13		0.004	
Change in Residual (%)	-1.84		-26.28		-35.00		0.00	
UGM AIC	1385.10		779.60		377.90		-509.90	
FGM AIC	1376.01		729.22		345.52		-493.52	
ICC	0.11		0.31		0.61		0.41	
Pseudo R ² -marginal	0.06		0.21		0.12		0.08	
Pseudo R ² -conditional	0.16		0.45		0.58		0.45	
Cases; Obs.	81; 235		81; 235		81; 235		81; 235	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed);

Raw scores used in all models

4.4.5.1 Total Utterances

In general, total utterances approximated a normal distribution across all time points, with 4.68% of z-scores ≥ 1.96 . One potential outlier was identified in the EAL group: this child produced 47 utterances at t1, (as compared to the group mean of 24.04; a z-score of 4.68), but 26 utterances at t2 and 27 utterances at t3. However, case-deletion diagnostics did not reveal this observation or any others in the dataset to be highly influential subjects or observations, and thus all data were retained in the LMM. Residuals tended to centre around zero with constant variance across groups and time points, and random effects approximated a normal distribution.

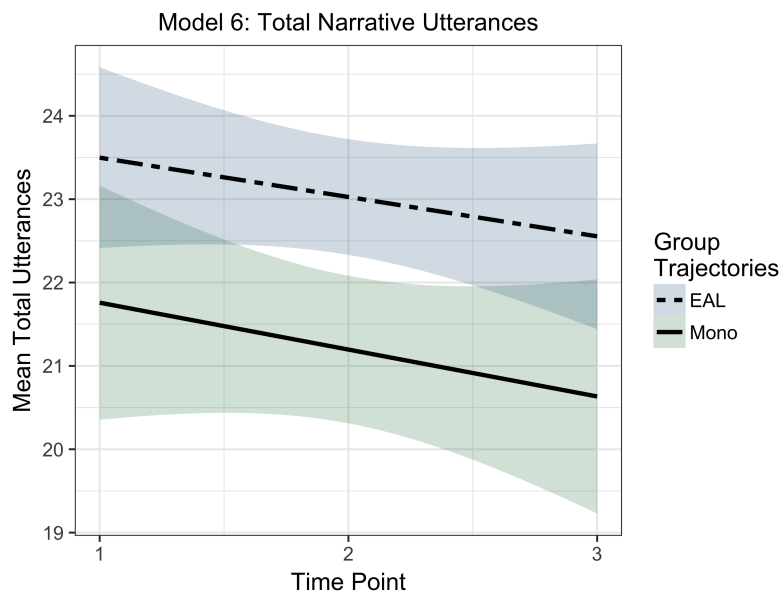


Figure 4.8: Linear Mixed Modelling of t1-t3 Total Utterances in Oral Narrative

The addition of a random slope term to the UGM did not improve model fit ($\Delta\text{AIC} = 3.72$, $\chi^2(2) = 0.28$, $p = .869$) and was consequently removed. While the fixed effects of *time* and *group* did contribute meaningfully to the intercept-only model ($\Delta\text{AIC} = -8.81$, $\chi^2(2) = 12.81$, $p = .002$), the *time* \times *group* interaction did not ($\Delta\text{AIC} = 1.94$, $\chi^2(1) = 0.06$, $p = .809$). Closer inspection of fixed effects revealed the significant contribution of both *time* ($p = .049$) and *group* ($p = .004$), with the EAL group producing on average 1.96 more utterances than the monolingual group. Unfortunately the final model represented a poor fit to the data ($R^2\text{-conditional} = 0.16$) with only a low degree of within-subject consistency ($\text{ICC} = 0.11$), resulting in a relatively larger estimate of standard error (see Table 4.12).

Both groups experienced a slight downward slope in total utterances produced over time, potentially (Intercepts: Mono = 22.36; EAL = 24.66; Slopes: Mono = -0.58; EAL = -0.74; Figure 4.8). At each time point children in the EAL group tended to produce more utterances than their monolingual peers, and thus the magnitude of group differences did not change appreciably over time (g between -0.45 and -0.46). Descriptive statistics in Table 4.4 show that children tended to produce only slightly fewer utterances on average (between 20.66 and 24.04 across all time points) than the original Peter and the Cat story contains ($n=25$).

4.4.5.2 Mean Length of Utterance in Words (MLUw)

MLUw data approximated a normal distribution, although there were a slightly larger than expected number of outliers (z-scores: 5.53% ≥ 1.96 ; 0.85% ≥ 2.58 ; 0.43% ≥ 3.29). One explanation for the relatively larger number of outliers in MLUw data is the generally low number of utterances (a mean of 22.26), meaning that the presence of merely a few longer utterances in a relatively limited pool of utterances may have inflated MLUw somewhat for some children. Indeed, there were consistent and statistically significant negative relationships between total utterances and MLUw (average $r = -0.47$ across time points, all $p < .001$). Residuals and random effects approximated a normal distribution and appeared to have constant variance across groups and time points. Case-deletion diagnostics did not reveal any highly influential subjects or observations.

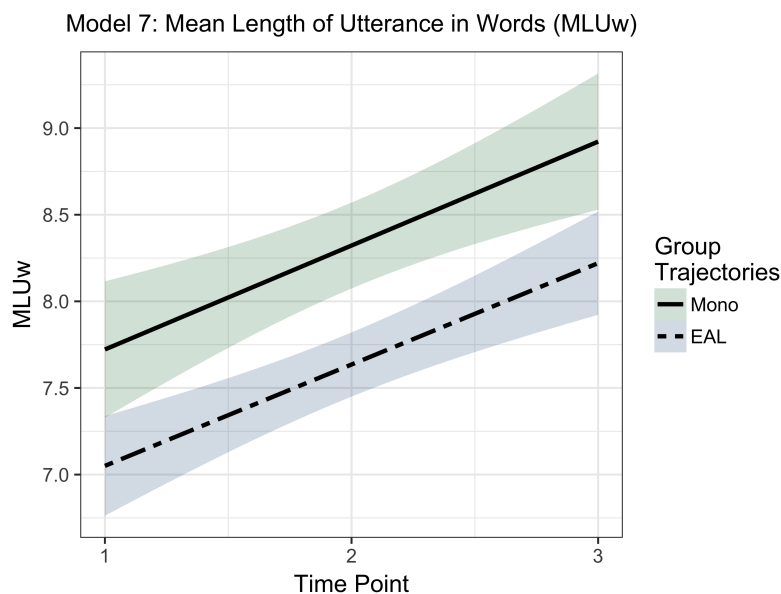


Figure 4.9: Linear Mixed Modelling of t1-t3 Mean Length of Utterance in Words

The addition of a random slope term to the UGM did not improve model fit ($\Delta AIC = 0.82$, $\chi^2(2) = 3.18$, $p = .204$), although the inclusion of fixed effects of *time* and *group* did ($\Delta AIC = -58.85$, $\chi^2(2) = 58.85$, $p < .001$). The addition of a *time* \times *group* interaction term, however, did not improve upon this model ($\Delta AIC = 1.99$, $\chi^2(1) = 0.01$, $p = .941$). The fixed effect of *time* was a significant predictor of performance, with children's utterances increasing by an average length of 0.59 words at each subsequent time point ($p < .001$). Model 7 for MLUw also contained a significant effect of *group*, albeit in the opposite direction as that for total utterances in Model 6: the MLUw of children in the EAL group was on average 0.68 words shorter than that of children in the monolingual group ($p < .001$). Thus, although children with EAL tended to produce more utterances, these utterances also tended to be shorter than those of the monolingual children. In contrast to Model 6 for total utterances, the final model for MLUw represented a better, although still modest, fit to the data (pseudo R^2 -conditional = 0.45), with a modest degree of within-subject consistency over time (ICC = 0.31). The trajectories of the two groups were fairly similar in terms of progress over time (Intercepts: Mono = 7.13; EAL = 6.48; Slopes: Mono = 0.60; EAL = 0.58; Figure 4.9), suggesting that by the end of the study the EAL group had not made progress in

4.4. Linear Mixed Modelling

closing the gap in performance observed at t1. As shown by descriptive statistics (Table 4.4), between t1 and t2 the two groups tended to diverge, and subsequently converged again by t3.

4.4.5.3 Lexical Diversity

Lexical diversity of language samples was measured through Root TTR (Guiraud, 1959). Root TTR data approximated a normal distribution, with 5.53% of z-scores ≥ 1.96 . Residuals centered around zero with constant variance across time points and groups. Random effects also approximated a normal distribution, and case-deletion diagnostics did not reveal the presence of any highly influential subjects or observations.

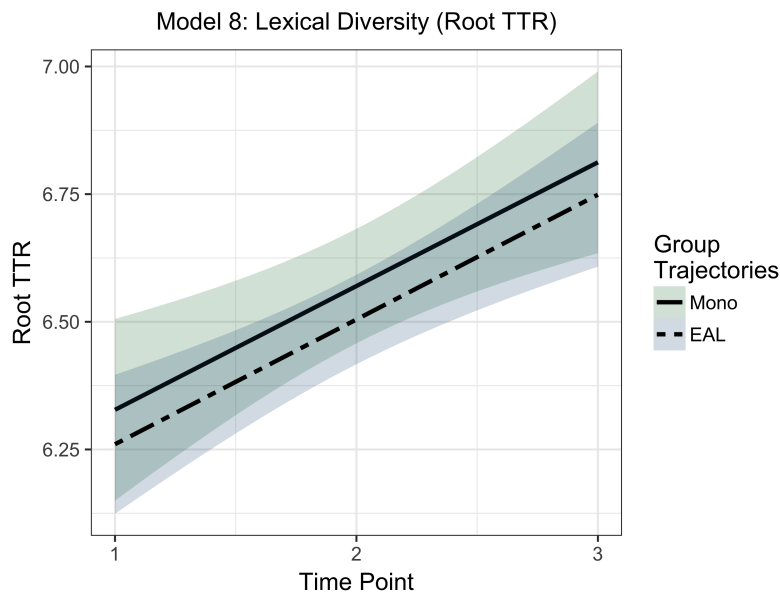


Figure 4.10: Linear Mixed Modelling of t1-t3 Lexical Diversity (Root TTR)

The addition of a random slope term to the UGM resulted in improved model fit ($\Delta AIC = -5.21$, $\chi^2(2) = 9.21$, $p = .010$). The inclusion of fixed effects of *time* and *group* additionally resulted in improved fit ($\Delta AIC = -39.30$, $\chi^2(2) = 43.30$, $p < .001$), but not the *time* \times *group* interaction term ($\Delta AIC = 2.00$, $\chi^2(1) = 0.00$, $p = .999$), which was subsequently removed from the model. Children's lexical diversity increased by an average of 0.24 for each subsequent time point ($p < .001$). Although this figure is itself uninterpretable, just as with raw TTR, a higher value indicates a higher degree of lexical diversity (Malvern et al., 2004). The fixed effect of *group* suggested a very small difference in favour of the monolingual group ($\beta = -0.07$), although this was not a significant predictor of development over time ($p = .495$). The final model represented a modest fit to the data (pseudo R^2 -conditional = 0.58) with a similarly modest degree of within-subject variability over time (ICC = 0.61).

Groups differed minimally in performance at t1 and made very similar rates of progress over time (Intercepts: Mono = 6.09; EAL = 6.02; Slopes: Mono = 0.24; EAL = 0.24; Figure 4.10), suggesting that on average, the lexical diversity of children's retellings did not change markedly over time. Descriptive statistics indicated that group differences became apparent only at t2 ($g = 0.36$) and no longer remained by t3 ($g = 0.00$).

4.4.5.4 Oral Narrative Error Analysis

Transcription of language samples in CLAN (MacWhinney, 2000) also allows for the marking of number of utterances containing at least one error, as well as specific types of morphosyntactic and semantic errors (see Section 3.4.4.1 for examples). Here, error *rates* are presented, calculated as the total number of raw errors divided by the total number of utterances. Narrative error rate data showed positive skewness for both groups at all time points due to the majority of children who made no or few errors. Despite this, residuals and random effects did follow a normal distribution and variance was fairly constant across groups and time points. Some outliers were present in the data (1.7% of z-scores ≥ 2.58), although case-deletion diagnostics did not reveal the presence of any highly influential subjects or observations.

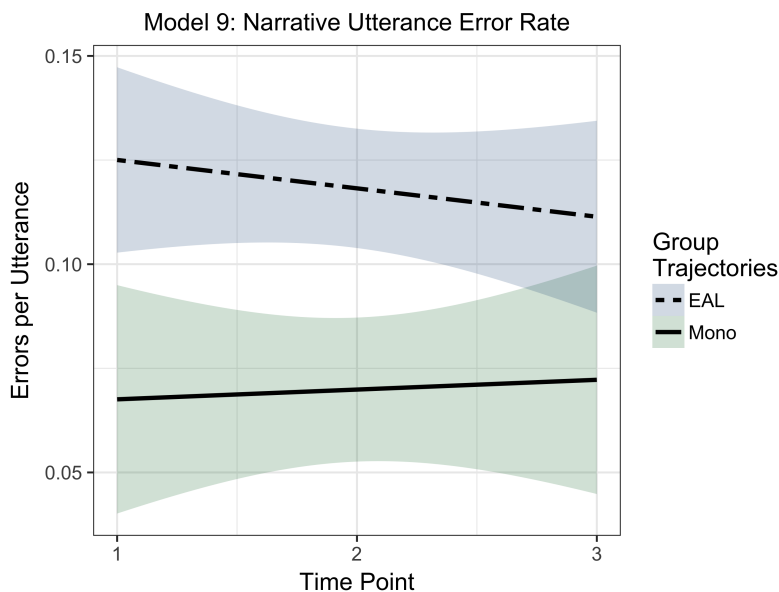


Figure 4.11: Linear Mixed Modelling of t1-t3 Error Rate in Oral Narrative

The addition of a random slope term did not improve model fit ($\Delta\text{AIC} = 3.89$, $\chi^2(2) = 0.11$, $p = .948$), however the fixed effects of *time* and *group* did make a significant contribution to the intercept-only model ($\Delta\text{AIC} = -6.80$, $\chi^2(2) = 10.80$, $p < .001$). The inclusion of a *time* \times *group* interaction did not improve fit and was removed from the model ($\Delta\text{AIC} = 1.43$, $\chi^2(1) = 0.57$, $p = .449$). While children did appear to make fewer errors over time, this effect was very small in magnitude and did not reach significance ($\beta = -0.003$, $p = .532$). Children in the EAL group made on average 0.05 more errors per utterance than their monolingual peers ($p = .002$). The final model represented a modest fit to the data (pseudo R^2 -conditional = 0.45) with a modest level of within-subject consistency over time (ICC = 0.41).

The two groups differed with respect to their initial error rates at t1 and in their direction and rate of change over time (Intercepts: Mono = 0.066; EAL = 0.132; Slopes: Mono = 0.002; EAL = -0.007; Figure 4.11). Although children in the EAL group did make fewer errors per utterances over time, this change was not of sufficient magnitude to result in a closing of the gap in performance.

Using the same procedure, additional LMMs were run for the two oral narrative error types, namely semantic and morphosyntactic (presented in Table 4.13 below; also see Section 3.4.4.1

4.4. Linear Mixed Modelling

for examples). In general, LMMs for oral narrative error types provided a fairly poor fit to the data (pseudo R^2 -conditional = 0.28 and 0.27 for Models 9.1 and 9.2, respectively. Additionally, children showed a low degree of within-subject consistency (ICCs of 0.26).

Table 4.13: Linear Mixed Modelling of Oral Narrative Errors

	Model 9.1		Model 9.2	
	Morphosyntactic Error Rate Estimate (SE)	t	Semantic Error Rate Estimate (SE)	t
Fixed Effects	Intercept	0.025 (0.011)	0.026 (0.007)	3.63 **
	Time	0.005 (0.004)	-0.004 (0.003)	-1.35
	Group: EAL	0.019 (0.010)	0.012 (0.006)	1.98 †
	Time × Group	-	-	-
Random Effects	Intercept Variance	0.001	0.000	
	Slope Variance	-	-	
	Residual Variance	0.003	0.001	
	Change in Residual (%)	0.00	0.00	
	UGM AIC	-671.60	-865.40	
	FGM AIC	-647.52	-839.16	
	ICC	0.26	0.26	
	Pseudo R ² -marginal	0.03	0.03	
	Pseudo R ² -conditional	0.28	0.27	
	Cases; Obs.	81; 235	81; 235	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed); † t-statistic approaches statistical significance.

Raw scores used in all models. Estimates rounded to 3 d.p. for ease of interpretation.

4.4. Linear Mixed Modelling

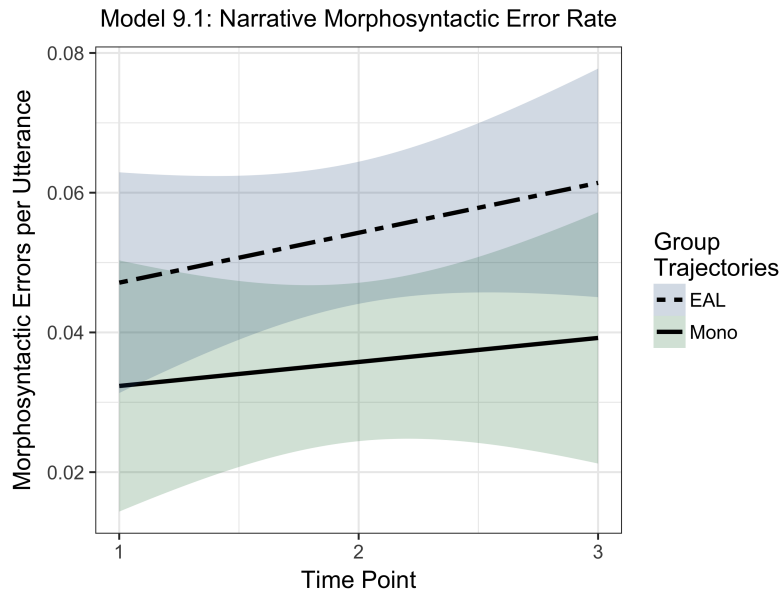


Figure 4.12: Linear Mixed Modelling of t1-t3 Morphosyntactic Error Rate in Oral Narrative

Oral Narrative Error Analysis: Morphosyntactic Errors. Residuals for morphosyntactic error rate generally followed a normal distribution but did display somewhat of a ‘fan effect’, suggesting non-constant variance. This was likely due to a high degree of positive skew in the raw data due to the majority of children who made few or no errors across all time points. Case-deletion diagnostics revealed a large number of significantly influential observations, but these were found to an equal degree in both groups at all time points, and thus retained in the dataset (cf. recommendations in Pinheiro & Bates, 2000, p.180). In terms of model fitting, the addition of a random slope improved model fit ($\Delta\text{AIC} = -3.81$, $\chi^2(2) = 7.81$, $p = .020$), but caused the model to fail to converge when fixed effects were added in Step 3. Even without a random slope term, the addition of fixed effects did not significantly improve model fit ($\Delta\text{AIC} = 1.51$, $\chi^2(2) = 5.51$, $p = .064$) and neither did a *time* \times *group* interaction ($\Delta\text{AIC} = 1.76$, $\chi^2(1) = 0.24$, $p = .623$). In the FGM, *time* was not a significant predictor ($\beta = 0.005$, $p = .184$), while *group* merely approached statistical significance ($\beta = 0.019$, $p = .055$). In terms of developmental trajectories, EAL learners tended to start on a slightly higher intercept at t1, but also made a slightly higher rate of errors at each subsequent time point (Intercepts: Mono = 0.030; EAL = 0.040; Slopes = Mono: 0.003; EAL: 0.007; Figure 4.12). Given the large number of children who made no errors at all, Model 9.1 represented a fairly poor fit to the data ($R^2\text{-conditional} = 0.28$).

Oral Narrative Error Analysis: Semantic Errors. Residuals were slightly positively skewed but followed a normal distribution. As in model 9.1, case-deletion diagnostics revealed a large number of influential observations which were found in both groups at all time points; as above, all observations were retained. The addition of a random slope term did not improve model fit ($\Delta\text{AIC} = 0.19$, $\chi^2(2) = 4.19$, $p = .123$), and fit was improved only marginally by the inclusion of fixed effects ($\Delta\text{AIC} = -1.85$, $\chi^2(2) = 5.85$, $p = .053$). A *time* \times *group* interaction similarly did not improve fit ($\Delta\text{AIC} = 2.00$, $\chi^2(1) = 0.002$, $p = .968$). In the FGM, *time* was not a significant predictor ($\beta = -0.004$, $p = .178$), and *group* again merely approached statistical significance ($\beta = 0.012$, $p = .051$). Similar to morphosyntactic error rate, EAL learners began on a slightly higher intercept at t1, but

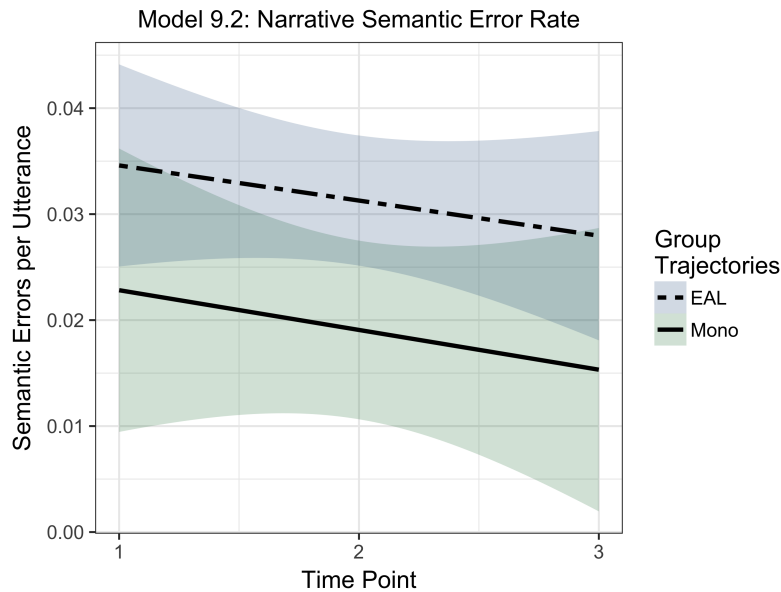


Figure 4.13: Linear Mixed Modelling of t1-t3 Semantic Error Rate in Oral Narrative

this time the two groups mirrored one another in their rate of fewer errors over time (Intercepts = Mono: 0.026; EAL: 0.038; Slopes = Mono: -0.004; EAL: -0.003; Figure 4.13). Similarly to Model 9.1, Model 9.2 also represented a fairly poor fit to the data (R^2 -conditional = 0.27).

Overall, models indicated a trend for EAL learners to make a higher rate of errors in spoken language while retelling a narrative. To some degree, this accorded with the EAL group's relatively lower expressive grammar (CELF FS) scores; however, the low number of errors made by children in both groups posed challenges for statistical modelling of oral narrative data.

4.4.6 Linear Mixed Modelling of Phonological Processing Measures

LMMs for the three phonological processing measures are presented in Table 4.14 (Models 10 to 12). Note that, as performance on RAN subtests is measured in seconds taken to name all stimuli, a higher score indicates slower naming speed, while a lower score indicates faster naming speed. The effect of *time* was significant for both RAN of letters and RAN of digits, although these trajectories differed in direction: while children became faster in their rapid naming of digits over time, they showed the opposite trend to become slower in their rapid naming of letters. Although group effects for RAN models did not reach statistical significance, LMMs confirmed a trend found in the descriptive statistics (Table 4.5) for an advantage of the EAL group in relation to their monolingual peers on both measures, with the magnitude of this advantage increasing in size over time, particularly for RAN of digits. Note, however, that descriptive statistics also report a tendency for non-linear performance, for example, with a much steeper slope between t1 and t2 for RAN of digits.

In contrast to RAN data, children's performance on the PhAB Spoonerisms subtest showed a much more consistent trajectory over time, corresponding to higher and more accurate performance. The effect of *group* was a significant predictor of performance, with monolingual children scoring higher than their bilingual peers. Again, the combination of a group difference at t1 and

4.4. Linear Mixed Modelling

similar rates of progress meant that the monolingual group maintained its advantage over time. Discussion of modelling procedures for RAN measures and Spoonerisms will follow below.

Table 4.14: Linear Mixed Modelling of Phonological Processing Measures

	Model 10: RAN Letters		Model 11: RAN Digits		Model 12: Spoonerisms	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects						
Intercept	39.88 (1.60)	24.88 **	51.72 (1.68)	30.70 **	14.47 (1.08)	13.38 **
Time	1.01 (0.42)	2.39 *	-7.29 (0.52)	-14.12 **	2.49 (0.27)	9.24 **
Group: EAL	-3.45 (1.78)	1.94	-2.32 (1.36)	-2.05	-2.31 (1.10)	-2.10 *
Time × Group	-	-	-	-	-	-
Random Effects						
Intercept Variance	52.52		101.05		38.84	
Slope Variance	-		4.83		1.21	
Residual Variance	27.36		31.49		8.90	
Change in Residual (%)	-2.94		-64.47		-44.85	
UGM AIC	1615.40		1747.30		1457.9	
FGM AIC	1605.26		1595.96		1378.74	
ICC	0.66		0.76		0.81	
Pseudo R ² -marginal	0.04		0.35		0.15	
Pseudo R ² -conditional	0.67		0.69		0.76	
Cases: Obs.	81; 236		81; 234		81; 236	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed);
Raw scores used in all models

4.4.6.1 Rapid Automatised Naming (CTOPP)

RAN of Letters. Inspection of data did reveal a small number of extreme observations (0.84% of z-scores ≥ 3.29), and influence diagnostics identified five particularly high-scoring subjects in the monolingual group whose scores were exerting leverage on model parameters. However, re-running of the model without these five children did not alter statistical significance of the *group* fixed effect *t*-value. Diagnostics did reveal one particularly influential observation which did exert such influence: this represented a score within the monolingual group at t3 (this child took 85 seconds to read all 36 letter stimuli, a z-score of 5.44). Thus, the decision was made to remove this single observation only, the effect of which was to alter the *group t*-value from -2.14 to -1.94. In the final dataset, random effects approximated a normal distribution, and residuals exhibited fairly constant variance across groups and time points.

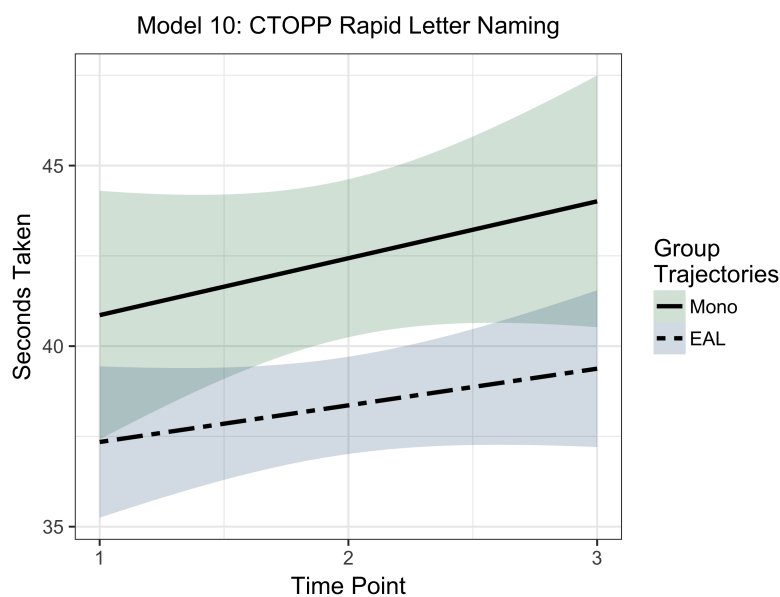


Figure 4.14: Linear Mixed Modelling of t1-t3 Rapid Automatised Letter Naming

The inclusion of a random slope term did not significantly improve model fit ($\Delta\text{AIC} = 3.78$, $\chi^2(2) = 0.22$, $p = .894$) and was consequently removed from the model. The addition of fixed effects of *time* and *group* did improve fit to a small degree ($\Delta\text{AIC} = -5.46$, $\chi^2(2) = 9.46$, $p = .009$) but not the inclusion of a *time* \times *group* interaction ($\Delta\text{AIC} = 1.76$, $\chi^2(1) = 0.24$, $p = .623$). Thus, in the final model, *time* was a significant predictor of performance with children taking an additional 1.04 seconds to name all items at each subsequent time point ($p = .018$), and although the EAL group showed a trend for faster naming performance than the monolingual group, this difference did not reach significance ($p = .056$). The final model represented a modest to good fit to the data (pseudo R^2 -conditional = 0.67) with a modest degree of within-subject consistency over time (ICC = 0.66).

The two groups showed similar trajectories, although the naming speed of the EAL group increased at a faster rate (i.e. children in this group became slower) than that of the monolingual group (Intercepts: Mono = 40.37; EAL = 36.09; Slopes: Mono = 0.76; EAL = 1.18; Figure 4.14). Therefore, in terms of Rapid Letter Naming performance, all children increased in the amount of

time it took them to name letters and the relatively slower performance of the monolingual group did not converge with the faster performance of the EAL group; indeed, descriptive statistics show that the groups diverged between t2 and t3 ($g = 0.39$ to 0.52).

RAN of Digits. Data for RAN of digits approximated a normal distribution (4.64% of z-scores ≥ 1.96), however influence diagnostics did reveal two observations within the monolingual group that were exerting a high degree of influence on the fixed effect of *group*, and these were consequently removed (Nieuwenhuis et al., 2012). Additionally, one child in the EAL group took 93 seconds to name all 36 digits at t1 (grand mean = 45.18 seconds; representing an extreme outlier with a z-score of 4.09) and this observation was therefore also removed⁷. Inspection of the resulting data revealed a slightly positively skewed distribution, as many children read items very quickly. Random effects and residuals approximated a normal distribution, with constancy of residuals across groups and time points.

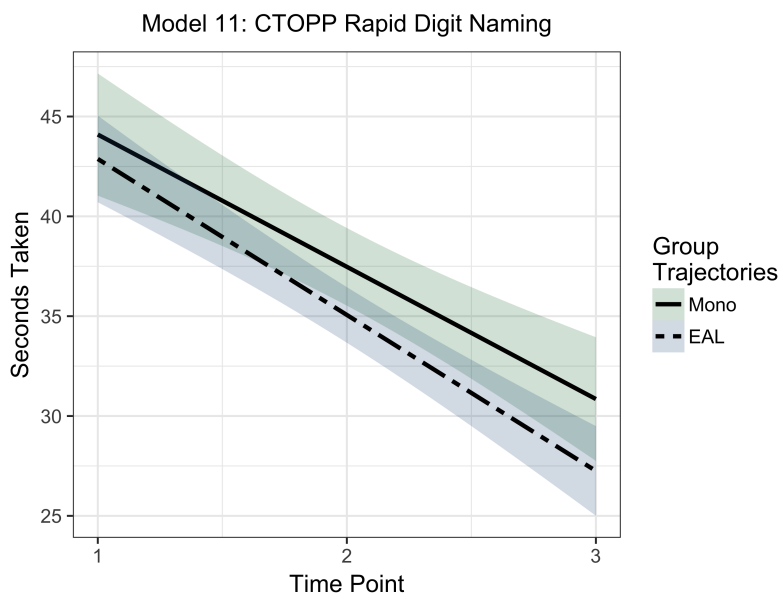


Figure 4.15: Linear Mixed Modelling of t1-t3 Rapid Automatised Digit Naming

The inclusion of a random slope term resulted in significantly improved model fit ($\Delta AIC = -44.71$, $\chi^2(2) = 48.71$, $p < .001$), as did the addition of fixed effects of *time* and *group* ($\Delta AIC = -102.68$, $\chi^2(2) = 106.68$, $p < .001$). The inclusion of a *time* \times *group* interaction, however, did not improve model fit and was removed from the final model ($\Delta AIC = 1.98$, $\chi^2(1) = 0.03$, $p = .879$). *Time* was a significant predictor of performance, with children taking on average 7.29 fewer seconds at each subsequent time point to name digit stimuli, suggesting a higher rate of automaticity. Despite a trend for faster performance of children in the EAL group, this effect did not achieve statistical significance ($p = .093$). The final model for Rapid Digit Naming similarly represented a modest to good fit to the data (pseudo R^2 -conditional = 0.69) and children showed a slightly higher degree of consistency in performance over time (ICC = 0.81).

The two groups showed very similar trajectories of performance over time, with slightly faster progress in the EAL group (Intercepts: Mono = 51.46; EAL = 49.52; Slopes: Mono = -7.18; EAL = -7.33; Figure 4.15). Descriptive statistics indicated a degree of nonlinearity in performance over

⁷Removal of this observation alone resulted in a reduction in residual variance of -1.86.

4.4. Linear Mixed Modelling

time, with a fairly large decrease in time taken to name stimuli between t1 and t2, followed by a levelling off by t3, particularly for the monolingual group. As a result, by the end of the study the EAL group had increased its advantage relative to t1 ($g = 0.62$ at t3).

4.4.6.2 Spoonerisms

PhAB Spoonerisms data approximated a normal distribution (4.24% of z-scores ≥ 1.96) with heavier-than-expected tails as a result of unusually low and unusually high scores. Residuals were centered around zero across groups and time points, and random effects were also normally distributed. Case-deletion diagnostics did reveal a number of highly influential subjects ($n=10$) but no highly influential individual observations. However, since these subjects represented extremely high- as well as extremely low-scoring children in both groups, the decision was made to analyse all available data (i.e. although heavy-tailed, the distributions were fairly symmetrical, meaning that fixed effects should not change substantially, cf. recommendations made in Pinheiro & Bates, 2000, p.180).

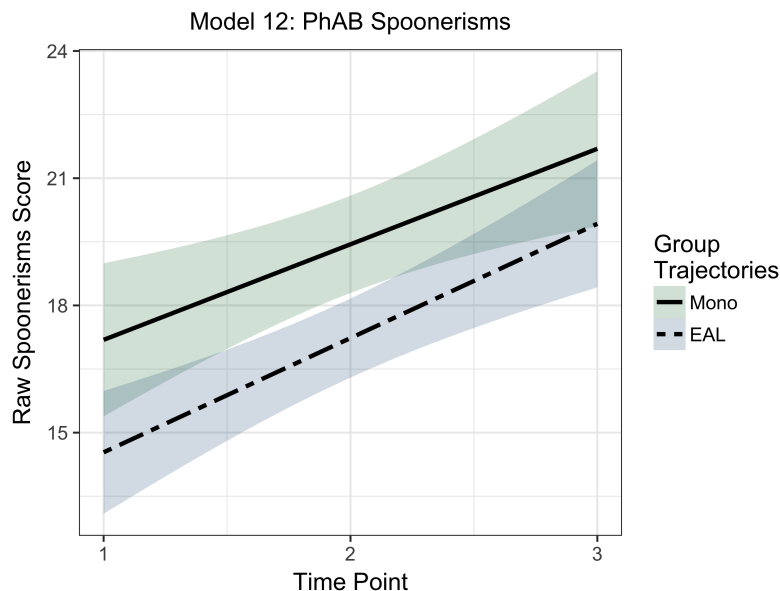


Figure 4.16: Linear Mixed Modelling of t1-t3 Spoonerisms

The inclusion of a random slope term in the UGM resulted in significantly improved model fit ($\Delta AIC = -18.6$, $\chi^2(2) = 22.61$, $p < .001$) as did the addition of fixed effects of *time* and *group* ($\Delta AIC = -58.74$, $\chi^2(2) = 62.74$, $p < .001$). However, the *time* \times *group* interaction did not improve fit ($\Delta AIC = 1.83$, $\chi^2(1) = 0.17$, $p = .678$), and was removed. In the FGM, *time* was a significant predictor of performance, with children scoring an additional 2.49 points at each subsequent time point ($p < .001$). The effect of *group* was also significant, with monolingual children scoring on average 2.31 points higher than their EAL peers ($p = .039$). The final model represented a good fit to the data (pseudo R^2 -conditional = 0.76) with a fairly high degree of within-subject consistency over time (ICC = 0.81).

The two groups had similar trajectories but differed mainly in initial skill at t1 (Intercepts: Mono = 14.80; EAL = 11.93; Slopes: Mono = 2.35; EAL = 2.58; Figure 4.16). Thus, the slightly steeper

slope of the EAL group resulted in a small degree of convergence between the groups across the three time points, from $g = 0.42$ at t1 to $g = 0.38$ at t3.

4.4.7 Linear Mixed Modelling of Literacy Measures

Children were assessed on three measures of literacy, including single-word reading efficiency (words and pseudowords), passage reading (rate, accuracy, and comprehension), and written narrative (total and mean length of T-unit, error rates). In the following section, a conceptual summary for development over the three time points of the monolingual and EAL group, followed by detailed model fitting procedures will be provided for each measure in turn.

4.4.7.1 Single-Word Reading Efficiency (TOWRE)

Linear mixed modelling of single word reading efficiency performance by the two groups of children across the three time points paints a very similar picture for both subtests of the TOWRE (SW, sight word efficiency [words] and PD, phonemic decoding efficiency [pseudowords]): in general, children made significant progress over time such that they are able to read a larger number of words correctly within the given time limit. Group effects appeared to play little part in predicting this performance. Both models explained a high proportion of variance, and children showed the greatest deal of variability with respect to their initial skills at t1 (see Table 4.15 overleaf). Interestingly, the monolingual group made a slightly faster rate of progress in PD efficiency than the EAL group, whereas the groups resembled one another more closely in progress on SW efficiency. Model fitting procedures for performance on both SW and PD subtests follow below.

Table 4.15: Linear Mixed Modelling of Single-Word Reading Efficiency

	Model 13: TOWRE SW		Model 14: TOWRE PD	
	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects				
Intercept	59.21 (1.57)	36.67 **	30.50 (1.77)	17.21 **
Time	3.93 (0.36)	10.80 **	2.97 (0.35)	8.45 **
Group: EAL	2.11 (1.84)	1.14	0.70 (2.06)	0.34
Time × Group	-	-	-	-
Random Effects				
Intercept Variance	69.41		107.87	
Slope Variance	3.16		4.36	
Residual Variance	14.09		10.34	
Change in Residual (%)	-56.64		-55.81	
UGM AIC	1634.30		1601.80	
FGM AIC	1534.55		1523.47	
ICC	0.83		0.91	
Pseudo R ² -marginal	0.13		0.06	
Pseudo R ² -conditional	0.84		0.90	
Cases; Obs.	81; 235		81; 235	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed); Raw scores utilised in all models

TOWRE Sight Word (SW) Efficiency. Children's scores on TOWRE SW subtest approximated a normal distribution, although there were a number of very high- and very low scores (7.23% of z-scores ≥ 1.96 and 1.28% ≥ 2.58). As a result, case-deletion diagnostics did not identify any highly influential subjects or observations, and thus all available data were used in the analysis. Residuals were approximately normally distributed, with constant variance for both groups across all time points. Random effects also approximated a normal distribution.

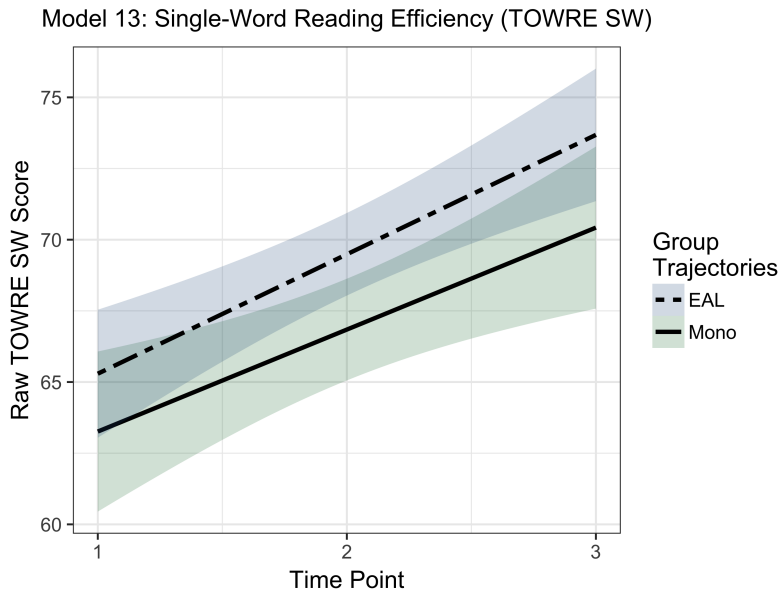


Figure 4.17: Linear Mixed Modelling of t1-t3 Sight-Word Reading Efficiency

The inclusion of a random slope term improved model fit ($\Delta\text{AIC} = -25.57$, $\chi^2(2) = 29.57$, $p < .001$) as did the fixed effects of *time* and *group* ($\Delta\text{AIC} = -69.67$, $\chi^2(2) = 73.66$, $p < .001$). The *time* \times *group* interaction, however, did not result in improved fit ($\Delta\text{AIC} = 1.20$, $\chi^2(1) = 0.004$, $p = .947$) and was removed from the model. On average, children read an additional 3.93 words correctly at each subsequent time point ($p < .001$): despite EAL learners reading more words correctly, however, this effect was not significant ($\beta = 2.11$; $p = .257$). The final model represented a good fit to the data (pseudo R^2 -conditional = 0.84), with a high degree of within-subject consistency over time (ICC = 0.83).

Both groups exhibited similar trajectories in both initial skill at t1 and rate of progress over time (Intercepts: Mono = 59.27; EAL = 61.29; Slopes: Mono = 3.89; EAL = 3.94; Figure 4.17). In this case, closing of the gap of the EAL group was not applicable due to its higher performance at t1; as indicated by descriptive statistics (Table 4.5), the relatively higher performance of the EAL group remained stable between t1 and t2, and had increased by t3 ($g = 0.40$).

TOWRE Phonemic Decoding (PD) Efficiency. TOWRE PD raw data were fairly symmetrically distributed, with only 3.4% of z-scores with a value of ≥ 1.96 . Residuals also approximated a normal distribution and showed constant variance across groups and time points. Similarly, random effects were also normally distributed and centered around zero, and case-deletion diagnostics did not reveal the presence of any highly influential subjects or observations.

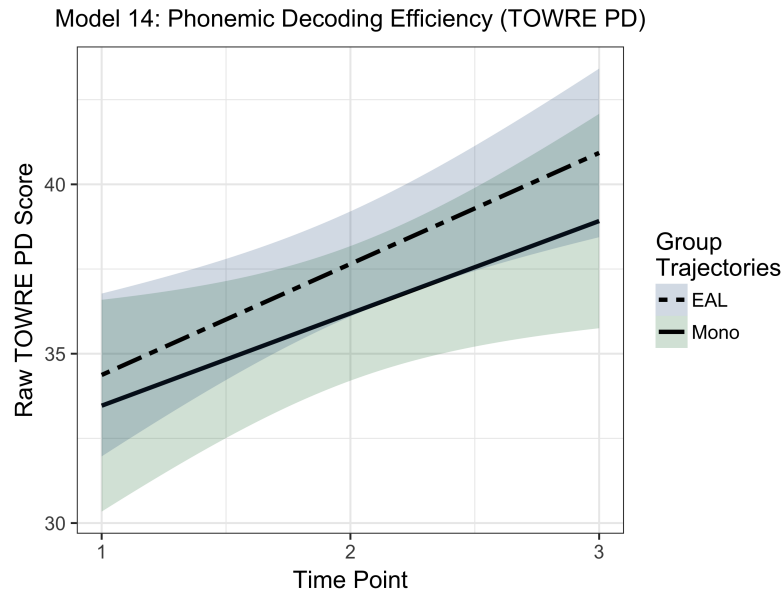


Figure 4.18: Linear Mixed Modelling of t1-t3 Phonemic Decoding Efficiency

The inclusion of a random slope term resulted in improved model fit ($\Delta\text{AIC} = -26.78$, $\chi^2(2) = 30.78$, $p < .001$) as did the addition of fixed effects of *time* and *group* ($\Delta\text{AIC} = -46.66$, $\chi^2(2) = 50.67$, $p < .001$). The inclusion of a *time* \times *group* interaction did not improve fit, however, and was removed from the model ($\Delta\text{AIC} = 1.95$, $\chi^2(1) = 0.05$, $p < .823$). In the FGM, time was a significant predictor of performance, with children reading an additional 2.97 words on average for subsequent time points ($p < .001$), but group was not a significant predictor ($p = .737$) despite a trend for more words read by the EAL group ($\beta = 0.70$). Like Sight Word reading efficiency, the final model for Phonemic Decoding efficiency performance represented a very good fit to the data (pseudo R^2 -conditional = .90), with a high degree of within-subject consistency (ICC = 0.91).

Both groups exhibited similar trajectories in performance over time (Intercepts: Mono = 30.30; EAL = 31.35; Slopes: Mono = 3.06; EAL = 2.90; Figure 4.18). Interestingly, unlike in Sight Word efficiency, it was the monolingual group that demonstrated a steeper slope in Phonemic Decoding performance, although descriptive statistics present a situation in which the dominance of the EAL group on this measure continued to grow in magnitude over time, reaching $g = 0.23$ by t3.

4.4.7.2 Passage Reading Rate, Accuracy, and Comprehension (YARC Primary)

Passage reading performance was measured with YARC Primary, yielding scores for rate, accuracy, and comprehension. As indicated in Table 4.1, passage reading variables are the only ones to be analysed using both raw and standard scores (see Section 4.1.2 for a justification of this strategy). Note that graphs in this section will display group trajectories for standard scores in the left panel (a), and trajectories for passage 3 raw scores in the right panel (b).

Overall, analysis of standard scores on passage reading variables (Table 4.16) revealed that the two groups performed on average at or above the norming population mean of 100 on passage reading rate, accuracy, and comprehension, suggesting that, as a group, children were broadly performing in line in relation to expectations for their age. Interestingly, the two groups did not

differ significantly in their standard scores relating to passage reading rate or comprehension performance, although there was a trend for the monolingual group to obtain slightly higher scores in descriptive statistics (Table 4.7). Conversely, a significant monolingual advantage emerged in reading accuracy, as well as a significant interaction over time whereby the monolingual group's reading accuracy scores decelerated over time and the EAL group's accelerated such that the initial group discrepancy at t1 was no longer present by t3.

Analysis of raw scores on passage 3 alone (Table 4.17) generally conformed to the same patterns found in analyses of standard scores; that is, no significant group differences in passage reading rate, and a significant monolingual group advantage in passage reading accuracy. However, the monolingual trend for higher reading comprehension was found to be statistically significant. No interaction term was significant in any passage 3 raw score models. Finally, linear mixed modelling of passage 3 raw scores generally resulted in slightly better model fit than that for standard scores, and all models utilising raw scores were significantly improved by the inclusion of random slope terms. Discussion of model fitting procedures will follow below, beginning with standard scores.

Table 4.16: Linear Mixed Modelling of Passage Reading Measures (Standard Scores)

	Model 15: Passage Reading Rate		Model 16: Passage Reading Accuracy		Model 17: Passage Reading Comprehension	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects						
Intercept	103.27 (2.13)	48.47 **	105.67 (2.07)	51.16 **	100.14 (1.44)	69.46 **
Linear	0.58 (0.34)	1.72	-1.31 (0.66)	-1.99 *	0.45 (0.47)	0.97
Group: EAL	-0.18 (2.63)	-0.07	-8.24 (2.69)	-3.07 **	-1.90 (1.43)	-1.33
Time × Group	-	-	2.52 (0.86)	2.92 **	-	-
Random Effects						
Intercept Variance	129.10		74.69		28.06	
Slope Variance	-		-		-	
Residual Variance	17.53		28.08		34.45	
Change in Residual (%)	-1.13		-3.87		0.00	
UGM AIC	1602.94		1645.0		1608.90	
FGM AIC	1598.23		1633.10		1606.20	
ICC	0.88		0.73		0.45	
Pseudo R ² -marginal	0.00		0.03		0.02	
Pseudo R ² -conditional	0.88		0.74		0.46	
Cases; Obs.	81; 236		81; 236		81; 236	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed); Standard scores utilised in all models

Table 4.17: Linear Mixed Modelling of Passage Reading Measures (Passage 3 only)

	Model 15.1: Passage Reading Rate		Model 16.1: Passage Reading Accuracy		Model 17.1: Passage Reading Comprehension	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects						
Intercept	121.24 (6.30)	19.26 **	5.83 (0.66)	8.84 **	4.09 (0.30)	13.80 **
Time	-15.04 (1.60)	-9.41 **	-1.21 (0.20)	-6.13 **	0.90 (0.09)	10.11 **
Group: EAL	0.88 (4.42)	0.20	1.27 (0.56)	2.25 *	-0.63 (0.25)	-2.49 *
Time x Group	-	-	-	-	-	-
Random Effects						
Intercept Variance	2246.10		18.68		2.86	
Slope Variance	126.10		1.66		0.11	
Residual Variance	140.10		2.55		0.95	
Change in Residual (%)	-69.72		-51.71		-48.54	
UGM AIC	2187.80		1150.90		854.90	
FGM AIC	2032.08		1087.16		768.74	
ICC	0.83		0.88		0.75	
Pseudo R ² -marginal	0.14		0.12		0.23	
Pseudo R ² -conditional	0.87		0.77		0.65	
Cases; Obs.	81; 230		80; 228		81; 227	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed);

Raw scores utilised in all models

4.4.7.2.1 Passage Reading Rate

Standard Scores. YARC passage reading rate standard scores approximated a normal distribution albeit with a slight negative skew due to a number of very rapid readers. One child from the EAL group was indicated by case-deletion diagnostics to have a high level of influence upon the final model (a standard score of 89 at t3). However, since diagnostic tests did not reveal a high degree of influence of this child on the fixed effect of *group*, and given that a number of children scored similar to or lower than this child, the decision was made to retain all of the data. Residuals were normally distributed with constant variance across time points and groups, and random effects also approximated a normal distribution.

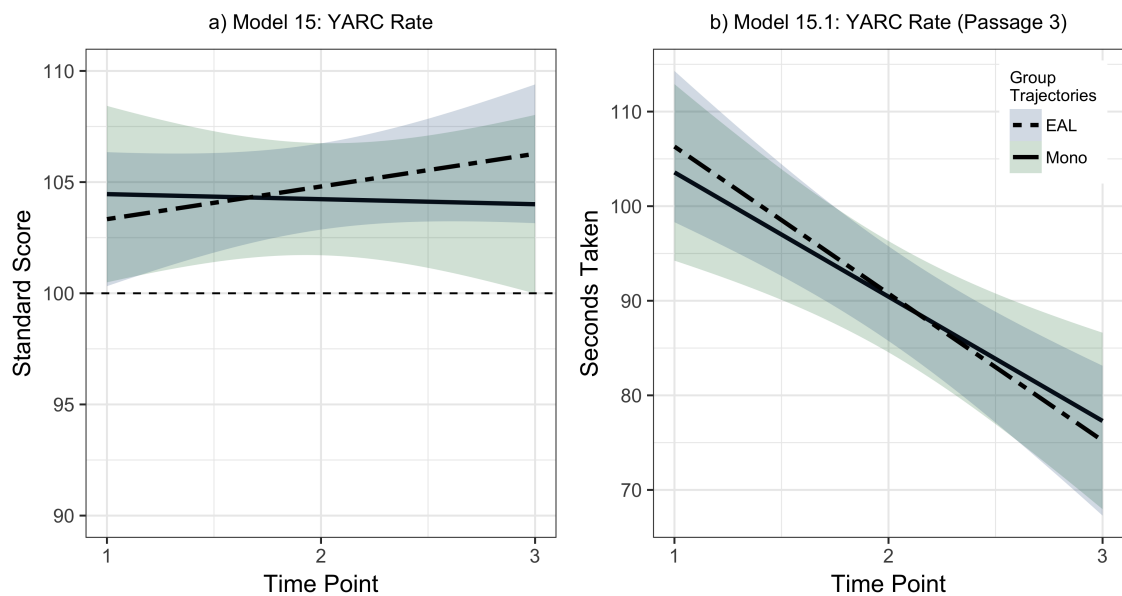


Figure 4.19: Linear Mixed Modelling of t1-t3 Passage Reading Rate. Standard scores represented in panel a) and passage 3 raw scores represented in panel b). Vertical dotted line at $y = 100$ in panel a) represents norming population mean.

The inclusion of a random slope term did not improve model fit ($\Delta\text{AIC} = 3.09$, $\chi^2(2) = 0.91$, $p = .635$); neither did the addition of fixed effects of *time* and *group* ($\Delta\text{AIC} = 1.05$, $\chi^2(2) = 2.95$, $p = .223$) nor the *time* \times *group* interaction ($\Delta\text{AIC} = 0.26$, $\chi^2(1) = 1.74$, $p = .187$). In the FGM, neither *time* ($p = .088$) nor *group* ($p = .946$) were significant predictors of performance over time. It is important to note that standard scores represent ranking relative to age-expectations based on performance of similarly aged examinees in the norming population. Therefore, these data, along with a high ICC (0.88) show that children were maintaining their position relative to expectations across the three time points. An intercept term of $\beta = 103.27$ indicates that both groups of children were performing within the average range in passage reading rate. The groups differed in the steepness of their slopes, with the EAL group improving with a faster rate relative to age-related expectations (Intercepts: Mono = 104.29; EAL = 102.38; Slopes: Mono = 0.07; EAL = 0.94; Figure 4.19). Overall, children varied mostly in terms of their initial performance at t1, and fixed effects accounted for an extremely small amount of total variance (pseudo R^2 -marginal = 0.002).

Passage 3 Raw Scores. Linear mixed modelling of raw scores for passage 3 reading rate (Model 15.1, Table 4.17) exhibited a fairly normal distribution (5.22% of z-scores ≥ 1.96). Residu-

als and random effects followed a normal distribution, and case-deletion diagnostics did not reveal the presence of any influential subjects or observations. Note that because raw scores pertain to the total amount of time taken (in seconds), both groups showed downward slopes as they took less time to read the passage at each time point (see Figure 4.19).

The addition of a random slope term to the UGM resulted in significantly improved model fit ($\Delta\text{AIC} = -87.98$, $\chi^2(2) = 91.98$, $p < .001$). Similarly, the addition of fixed factors *time* and *group* also resulted in significantly improved fit ($\Delta\text{AIC} = -56.78$, $\chi^2(2) = 60.78$, $p < .001$) but not the *time* \times *group* interaction term ($\Delta\text{AIC} = 1.99$, $\chi^2(1) = 0.01$, $p = .923$), which was removed. Closer inspection revealed that only *time* ($\beta = -15.03$, $p < .001$) and not *group* ($\beta = 0.88$, $p = .842$) accounted for improvement in model fit. Indeed, both groups exhibited very similar developmental trajectories in passage reading rate (Intercepts: Mono = 121.82; EAL = 121.90; Slopes: Mono = -15.23; EAL = -15.04; Figure 4.19). Thus, linear mixed modelling of both standard and raw scores suggested that the two groups did not differ from one another in their passage reading rate, and that both made very similar progress over time in this aspect of passage reading.

4.4.7.2.2 Passage Reading Accuracy

Standard Scores. YARC Passage Reading Accuracy data followed a normal distribution (3.39% of z-scores ≥ 1.96), as did random effects and residuals, which showed fairly constant variance across time points and groups. Case-deletion diagnostics did not reveal the presence of any highly influential subjects or observations and thus all data were retained for analysis.

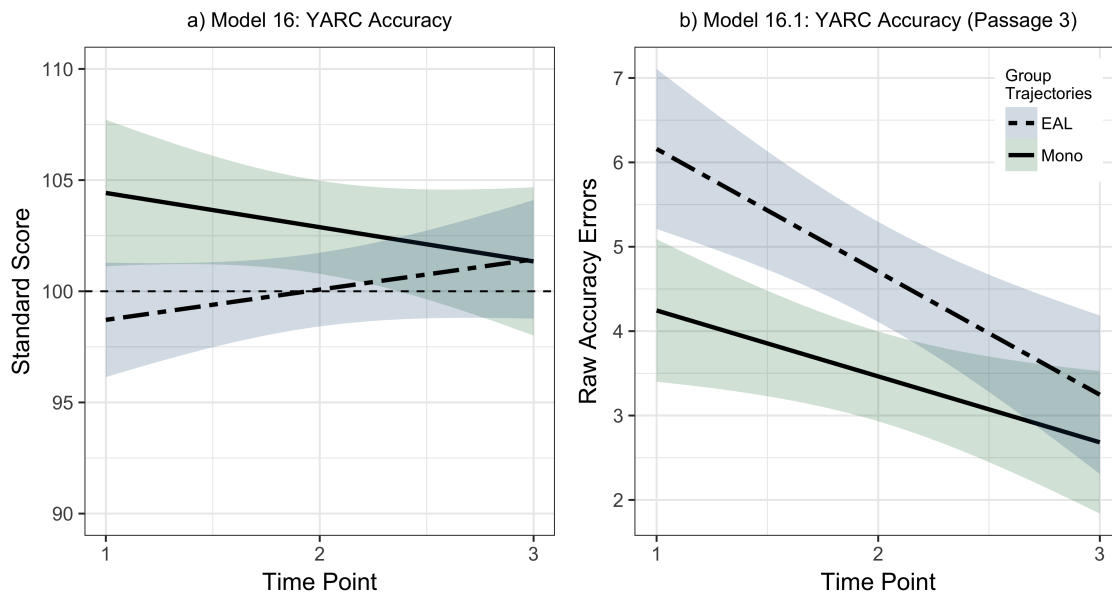


Figure 4.20: Linear Mixed Modelling of t1-t3 Passage Reading Accuracy. Standard scores represented in panel a) and passage 3 raw scores represented in panel b). Vertical dotted line at $y = 100$ in panel a) represents norming population mean.

The inclusion of a random slope term to the UGM did not result in improved model fit ($\Delta\text{AIC} = 2.88$, $\chi^2(2) = 1.12$, $p = .570$) and neither did the addition of fixed effects of *time* and *group* ($\Delta\text{AIC} = 1.39$, $\chi^2(2) = 2.61$, $p = .271$). The *time* \times *group* interaction, however, did improve model fit to a small degree ($\Delta\text{AIC} = -6.43$, $\chi^2(1) = 8.43$, $p = .004$). Indeed, the two groups differed both in terms

4.4. Linear Mixed Modelling

of performance at t1 as well as direction of slopes (Intercepts: Mono = 105.70; EAL = 97.46; Slopes: Mono = -1.34; EAL = 1.18; Figure 4.20). This crossover is reflected in the decreasing magnitude of group differences across time points (from $g = 0.63$ to $g = 0.05$; $\beta = 2.52$). In general, children showed a slightly lower level of within-subject consistency in passage reading accuracy (ICC = 0.73). The model accounted for a modest amount of total variance (pseudo R^2 -conditional = 0.74), with very little of this accounted for by fixed effects.

Passage 3 Raw Scores. Passage 3 accuracy raw scores (Model 16.1, Table 4.17) did approximate a normal distribution (5.65% z-scores ≥ 1.96), however case-deletion diagnostics revealed the presence of one influential subject who made 0, 12, and 11 accuracy errors at each time point, respectively. Removal of this subject increased the *group* coefficient *t*-value from 1.70 to 2.25, thus exerting a statistically significant impact on the model. Therefore, this subject was removed and the model was rerun. The resulting data followed a normal distribution, as did residuals and random effects. Note that raw scores for passage 3 reading accuracy indicate the raw number of pronunciation errors made; as such, a negative coefficient is interpreted as a lower raw number of errors.

The addition of a random slope term resulted in significantly improved model fit ($\Delta AIC = -33.79$, $\chi^2(2) = 37.79$, $p < .001$). The fixed effects of *time* and *group* also contributed to significantly better model fit ($\Delta AIC = -31.39$, $\chi^2(2) = 35.40$, $p < .001$), but not the *time* \times *group* interaction term ($\Delta AIC = -0.62$, $\chi^2(1) = 1.38$, $p = .241$). In the FGM, both *time* ($\beta = -1.21$, $p < .001$) and *group* ($\beta = 1.27$, $p = .027$) were predictive of children's performance. This pattern indicated that, while all children made significantly fewer accuracy errors over time, children in the EAL group made significantly more errors than their monolingual peers. Interestingly, EAL learners also made a slightly faster rate of progress over time, converging with their monolingual peers by t3 (Intercepts: Mono = 5.32; EAL = 7.56; Slopes: Mono = -0.91; EAL = -1.40; Figure 4.20). The model represented a fairly good fit to the data (pseudo R^2 -conditional = 0.77).

4.4.7.2.3 Passage Reading Comprehension

Standard Scores. YARC passage reading comprehension data approximated a normal distribution albeit with a small number of high-scoring outliers (0.85% of z-scores ≥ 3.29) which tended to be monolingual children. However, case-deletion diagnostics did not reveal the presence of any highly influential subjects or observations, and thus all data were utilised. Residuals were normally distributed with constant variance across covariates, although there was a higher degree of variability in general in the monolingual group and for all children at t1. Random effects were also approximately normally distributed and constant across covariates.

As in the two preceding models, the inclusion of a random slope did not result in improved model fit ($\Delta AIC = 3.68$, $\chi^2(2) = 0.32$, $p = .854$), and neither did the addition of fixed effects of *time* and *group* ($\Delta AIC = 1.25$, $\chi^2(2) = 2.75$, $p = .252$). The *time* \times *group* interaction similarly did not improve fit ($\Delta AIC = 1.51$, $\chi^2(1) = 0.49$, $p = .483$). Indeed, the two groups exhibited similar trajectories in reading comprehension performance over time, although the EAL group did experience a faster rate of growth (Intercepts: Mono = 100.91; EAL = 97.68; Slopes: Mono = 0.07; EAL = 0.74; Figure 4.21). In general, children showed a considerably lower level of within-subject consistency over time as compared to rate and accuracy (ICC = 0.45). Model fit for reading

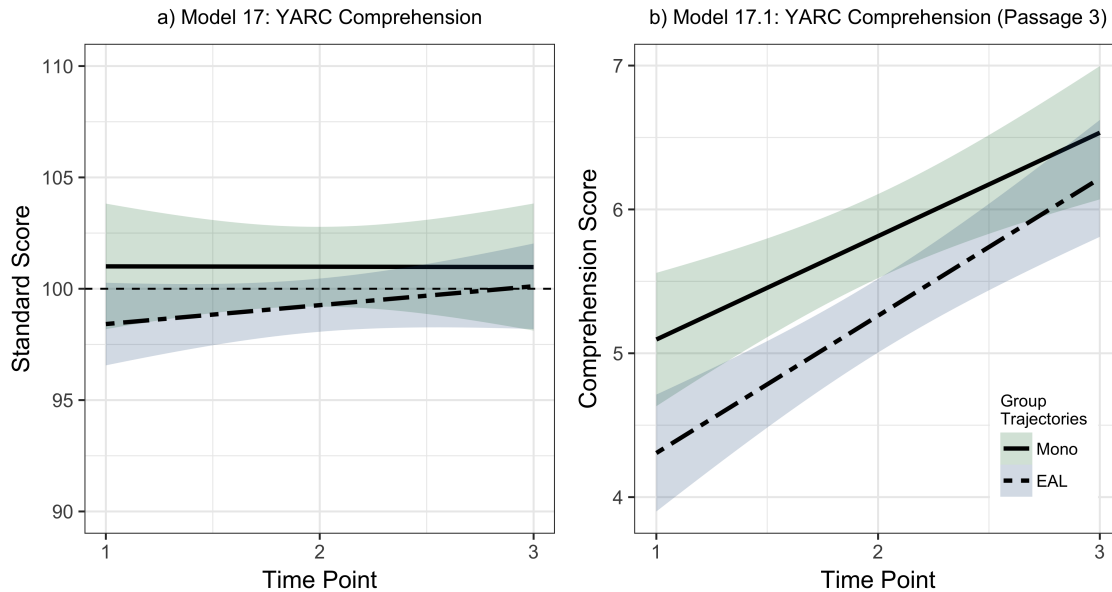


Figure 4.21: Linear Mixed Modelling of t1-t3 Passage Reading Comprehension. Standard scores represented in panel a) and passage 3 raw scores represented in panel b). Vertical dotted line at $y = 100$ in panel a) represents norming population mean.

comprehension was substantially poorer than for other passage reading variables (pseudo R^2 -conditional = 0.46).

Passage 3 Raw Scores. Raw data for passage 3 comprehension scores followed a normal distribution (3.96% of z-scores ≥ 1.96). Residuals centered around zero, and random effects were also normally distributed. Case-deletion diagnostics revealed the presence of four highly influential observations (all monolingual children at t2 and t3; 3 out of 4 low-scoring). Deletion of these observations altered the *group* coefficient from $\beta = -0.48$, ($p = .067$) to $\beta = -0.62$, ($p = .015$), and were thus removed from the dataset.

The addition of a random slope term resulted in significantly improved model fit ($\Delta AIC = -21.45$, $\chi^2(2) = 25.45$, $p < .001$), as did the fixed effects of *time* and *group* ($\Delta AIC = -71.04$, $\chi^2(2) = 75.04$, $p < .001$). The *time* \times *group* interaction term did not significantly improve model fit ($\Delta AIC = 1.46$, $\chi^2(1) = 0.54$, $p = .463$) and was therefore removed. Both fixed effects were significant predictors of children's passage reading comprehension performance (*time*: $\beta = 0.89$, $p < .001$; *group*: $\beta = -0.62$, $p = .015$), with all children answering more comprehension questions correctly over time, and monolingual children answering significantly more questions correctly. Although EAL learners started on a lower intercept at t1, they showed a trend to answer slightly more questions correctly over time (Intercepts = Mono: 4.29; EAL: 3.33; Slopes = Mono: 0.82; EAL: 0.95; Figure 4.21). Across all passage 3 models, children showed the lowest level of within-subject consistency for comprehension (Model 17.1; ICC = 0.65), which was relatively less stable than passage reading accuracy (Model 16.1; ICC = 0.77) and passage reading rate (Model 15.1; ICC = 0.77).

Overall, therefore, models for standard and raw scores were fairly similar in terms of no significant group differences in passage reading rate, as well as a significant monolingual group advantage in passage reading accuracy. The models differed with respect to comprehension,

4.4. Linear Mixed Modelling

however, with only passage 3 raw scores indicating a significant monolingual group advantage. Again, this result must be interpreted cautiously, given that it refers to performance on only one passage.

4.4.7.3 Writing

Results from the bespoke narrative writing task compared similarly to those from oral narrative retell measures; specifically, in their writing, children in the EAL group tended to produce a higher number of total T-Units, but unlike in oral narrative, these were not significantly shorter than those of their monolingual peers. Both groups showed positive growth in mean length of T-Unit (MLTw) over time. Negative relationships between MLTw and total T-Units⁸, however, suggest that caution is warranted in interpretation of these data, as in some cases a small pool of particularly long T-Units served to inflate children's MLTw scores. It is unlikely that this issue would be resolved by using an alternative metric such as the median, as each of these children's individual sentences tended to be of similar length. Instead, a larger writing sample may have produced more balanced data with more varied sentence structures.

Linear mixed modelling of writing error rates initially revealed no significant group differences; however, such differences became apparent once error type was taken into account. Particularly, a relatively higher morphosyntactic error rate served to distinguish the writing of children with EAL, while these children made slightly fewer spelling errors. Finally, it is unfortunate that no comparable measure of lexical diversity (e.g. Root TTR) could be applied to writing data due to the very low number of tokens (number of different words) used. Details of linear mixed modelling of writing variables will follow below. Models 18 to 20 are presented in Table 4.18.

⁸Pearson correlation coefficients: -.43, -.48, and -.50 at t1, t2, and t3, respectively. All coefficients $p < .01$.

Table 4.18: Linear Mixed Modelling of Written Narrative Measures

	Model 18: T-Units		Model 19: MLTw		Model 20: Error Rate	
	Estimate (SE)	t	Estimate (SE)	t	Estimate (SE)	t
Fixed Effects						
Intercept	5.63 (0.42)	13.36 **	9.30 (0.55)	16.94 **	0.94 (0.12)	7.91 **
Time	-0.15 (0.15)	-0.99	0.67 (0.25)	2.73 **	-0.14 (0.03)	-4.45 **
Group: EAL	0.88 (0.39)	2.23 *	-0.75 (0.48)	-1.58	0.06 (0.12)	0.53
Time × Group	-	-	-	-	-	-
Random Effects						
Intercept Variance	1.86		0.15		0.42	
Slope Variance	-		0.92		-	
Residual Variance	3.41		7.46		0.15	
Change in Residual (%)	0.06		-14.99		-15.84	
UGM AIC	1040.50		1223.0		385.1	
FGM AIC	1042.03		1212.37		378.70	
ICC	0.36		0.02		0.74	
Pseudo R ² -marginal	0.04		0.04		0.04	
Pseudo R ² -conditional	0.38		0.31		0.63	
Cases; Obs.	81; 235		81; 235		81; 235	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed);
Raw scores utilised in all models

4.4.7.3.1 Total T-Units

Data for the total number of T-units produced displayed a discernibly positive skew as most children tended to produce only a small amount of writing. This pattern applied equally to children in both groups, with the majority writing between 2 and 6 T-units in total. There were some outliers in the data (2.13% of z-scores ≥ 2.59): nine children produced between 11 and 15 T-units, but such potential outliers were found in both groups and are not considered particularly unusual, especially since many of these sentences were short in length. As a result, the decision was made not to remove these cases from the dataset. This decision was supported by case-deletion diagnostics, which did not reveal any of these subjects or their individual observations to be highly influential. Residuals were normally distributed and showed constant variance across groups and time points, although there was slightly higher variation in the EAL group. Random effects also followed a normal distribution.

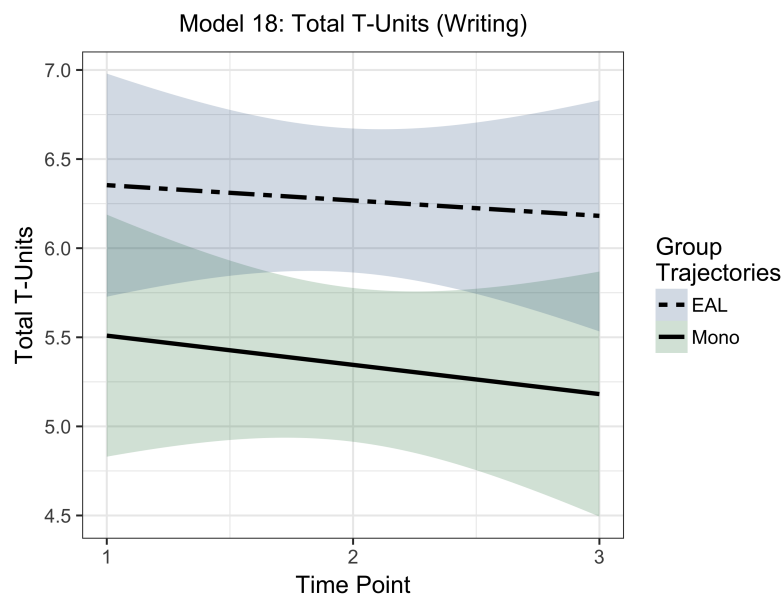


Figure 4.22: Linear Mixed Modelling of t1-t3 Total T-Units in Writing

The inclusion of a random slope term in the model did not result in improved model fit ($\Delta\text{AIC} = 2.54$, $\chi^2(2) = 1.47$, $p = .481$), and the addition of fixed effects of *group* and *time* to the intercept-only model improved fit only marginally ($\Delta\text{AIC} = -1.98$, $\chi^2(2) = 5.97$, $p = .050$). The inclusion of a *time* \times *group* interaction did not further improve fit ($\Delta\text{AIC} = 1.98$, $\chi^2(1) = 0.02$, $p = .887$) and was removed from the model. Children produced slightly fewer T-units over time ($\beta = -0.15$), however this effect was not statistically significant. On the other hand, *group* was a significant predictor of performance, as children with EAL produced on average 0.88 more T-units than their monolingual peers ($p = .029$). Therefore, akin to the oral narrative language sample data, the EAL group also appeared to show a significantly higher level of productivity in written language.

The two groups differed in terms of their initial performance at t1 as well as in the steepness of their slopes (Intercepts: Mono = 5.68; EAL = 6.46; Slopes: Mono = -0.17; EAL = -0.12; Figure 4.19). Therefore, while both groups showed a trend to produce fewer T-units over time, the magnitude of this decrease was slightly higher in the monolingual group. As indicated by descriptive

statistics in Table 4.8, by the end of the study, the monolingual group's decrease, coupled with the EAL group's increase in T-units, resulted in an effect size of $g = 0.52$. The final model represented fairly poor fit to the data (pseudo R^2 -conditional = 0.38), with a relatively low degree of within-subject consistency across time points (ICC = 0.31).

4.4.7.3.2 Mean Length of T-Unit in Words (MLTw)

Data for MLTw did present some challenges to the assumption of a univariate normal distribution (1.28% of z-scores ≥ 3.29). As in modelling of MLUw, this is likely an artefact of the generally low level of children's productivity, meaning that a very small number of T-units with, say, complex phrase structure or subordinating conjunctions, were likely to have artificially inflated mean length of production. Thus, caution should be taken when calculating measures of MLTw from a pool of between 2 to 15 T-units. As an example, Table 4.19 below provides examples from children who produced a small number of relatively lengthy T-units (spelling errors retained). These children tended to make use of subordinating conjunctions (*because, which, where*) and to elongate their sentences by adding additional clauses with elided subjects (*and build. . . , and playing. . . , etc.*), which count only as one T-unit. Additionally, it was noted during testing that some children appeared to sacrifice productivity for the sake of producing longer, more highly-crafted sentences.

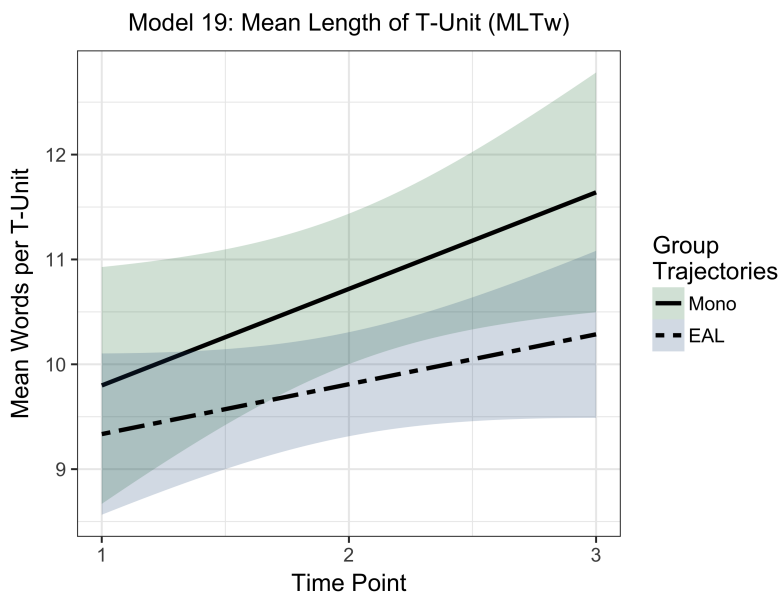


Figure 4.23: Linear Mixed Modelling of t1-t3 Mean Length of T-Unit in Words

Despite a number of potential outliers in the MLTw data, case-deletion diagnostics did not reveal any subjects or observations to be highly influential and thus the decision was made to retain all data for analysis. Residuals approximated a normal distribution with fairly constant variance across groups and time points. Additionally, random effects were also centered around zero, with a slightly higher degree of variability in slopes than intercepts.

The inclusion of a random slope term resulted in improved model fit ($\Delta AIC = -6.59$, $\chi^2(2) = 10.59$, $p = .005$), as did the addition of fixed effects of *time* and *group* ($\Delta AIC = -5.79$, $\chi^2(2) = 9.79$, $p = .007$). However, the *time* \times *group* interaction did not improve upon model fit ($\Delta AIC =$

4.4. Linear Mixed Modelling

Table 4.19: Examples of T-Units in the Bespoke Writing Task

t3; Mono	My favourite thing to do is play games such as Minecraft which is a game where you get creative and build anything you want.
t2; Mono	I also like my iPad because I love going on you tube and playing on GT racing and mincraft to mess about.
t3; EAL	and I like to play with my freinds and go on my phone to play games because it's fun and exiting.
t3; EAL	and then it was the end of the episode which was so anoying because I don't know what happens next.

1.26, $\chi^2(1) = 0.74$, $p = .391$) and was thus removed from the final model. In the FGM, *time* was a significant predictor of performance, as children's MLTw increased by an average of 0.67 words at each time point ($p = .008$). The two groups did not differ significantly in MLTw ($\beta = -0.75$; $p = .118$). As in analysis of total T-units, the final model for MLTw represented a fairly poor fit to the data (R^2 -conditional = 0.31), with a negligible amount of within-subject consistency over time (ICC = 0.02)⁹.

The two groups differed primarily in the slopes of their trajectories, with monolingual children making a faster rate of progress than their EAL peers (Intercepts: Mono = 8.89; EAL = 8.84; Slopes: Mono = 0.91; EAL = 0.49; Figure 4.23). As indicated by descriptive statistics (Table 4.8), the magnitude of the monolingual advantage remained stable between t1 and t2 but had increased by t3 ($g = 0.38$); in other words, although the two groups tended to perform similarly at t1 with regard to the mean length of their T-units, the groups diverged over time thanks to the trend for monolingual children to produce longer T-units at each subsequent time point.

4.4.7.3.3 Writing Error Analysis

Writing samples were analysed for spelling and morphosyntactic errors. In order to obtain a standardised measure of error rate, total number of errors (spelling plus morphosyntactic) was divided by total number of T-units (overall error rate; separate models for spelling and morphosyntactic errors are reported below). Analogous to oral narrative retell data, many children made no errors at all¹⁰, which resulted in positively skewed data. Despite this, residuals and random effects were normally distributed and displayed fairly constant variance across groups and time points. There were some outliers in both tails of the distribution (2.98% of z-scores ≥ 2.58), although case-deletion diagnostics did not reveal the presence of any highly influential subjects or observations.

The inclusion of a random slope in the UGM significantly improved model fit ($\Delta AIC = -2.93$, $\chi^2(2) = 6.93$, $p < .05$), as did the fixed effects of *time* and *group* ($\Delta AIC = -14.93$, $\chi^2(2) = 18.93$, p

⁹Note, however, that the intercept-only model (constructed in step 1 of the model building process; Table 4.9) derived an ICC of 0.18, justifying the use of a linear mixed model for children's performance on this variable. The ICC is also taken to represent the amount of variance in the response variable (i.e. MLTw) that is attributable to the random effect of subject (West et al., 2007). Therefore, a reduction in ICC may occur as a result of other variables within the model accounting for variance at this level within the model.

¹⁰Error rates differed by type: 17% of all children made no spelling errors at any time point, while 62% made no morphosyntactic errors.

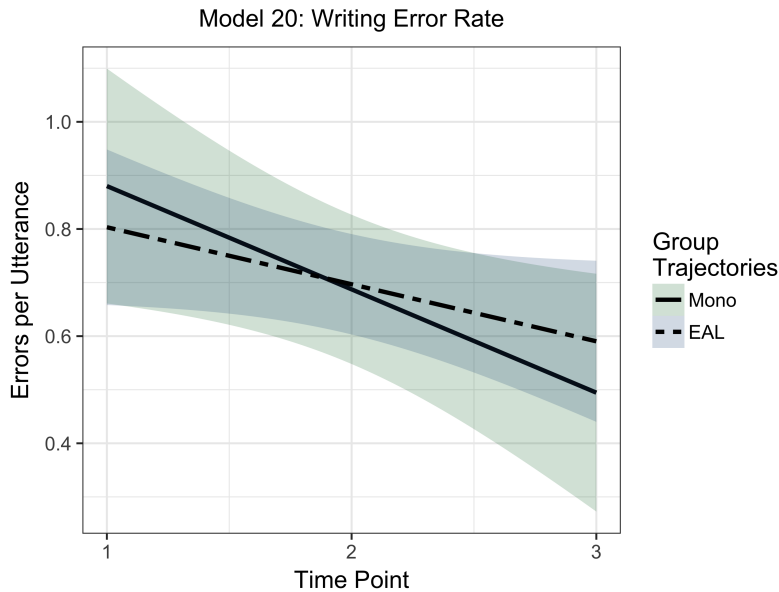


Figure 4.24: Linear Mixed Modelling of t1-t3 Errors in Writing

< .001). However, the *time* \times *group* interaction was not significant ($\Delta\text{AIC} = 0.19$, $\chi^2(1) = 2.19$, $p = .139$) and was therefore removed from the final model. In terms of writing errors, then, all children made on average 0.14 fewer errors per T-Unit over time ($p < .001$), but the two groups did not differ significantly from one another in overall error rates ($p = .601$). The final model represented a relatively better fit to the data in comparison to previous writing variable models ($R^2\text{-conditional} = 0.63$) with a fairly high degree of within-subject consistency ($\text{ICC} = 0.74$).

At t1, children in the monolingual group made a slightly higher rate of errors than their EAL peers, but also showed a steeper decline in error rate over time (Intercepts: Mono = 1.09; EAL = 0.91; Slopes: Mono = -0.20; EAL = -0.10; Figure 4.24). As shown by descriptive statistics, it was the monolingual group that made very slightly more errors per T-unit at t1 and t2; however, although this rate had decreased for both groups by t3, the relatively steeper downward slope of the monolingual group resulted in a modest effect size of $g = -0.21$, as the EAL group continued to make more errors.

As in the case of oral narrative errors, it was of interest to separate written narrative error rate into spelling and morphosyntactic types (LMMs presented in Table 4.20). Model data and fitting procedures are presented below.

Table 4.20: LMMs for Written Narrative Error Types

	Model 20.1:		Model 20.2:	
	Spelling Error Rate Estimate (SE)	t	Morphosyntactic Error Rate Estimate (SE)	t
Fixed Effects				
Intercept	0.83 (0.11)	7.85 **	0.09 (0.03)	3.57 **
Time	-0.11 (0.03)	-3.73 **	-0.02 (0.01)	2.06 *
Group: EAL	-0.03 (0.11)	-0.27	0.07 (0.02)	3.51 **
Time x Group	-	-	-	-
Random Effects				
Intercept Variance	0.29		0.00	
Slope Variance	0.01		-	
Residual Variance	0.12		0.01	
Change in Residual (%)	-17.43		-2.15	
UGM AIC	333.96		-257.59	
FGM AIC	338.10		-249.12	
ICC	0.71		0.19	
Pseudo R ² -marginal	0.03		0.08	
Pseudo R ² -conditional	0.63		0.26	

** t-statistic is significant at the 0.01 level (2-tailed); * t-statistic is significant at the 0.05 level (2-tailed)

Writing Error Analysis: Spelling Error Rate. With regards to spelling error rate, residuals followed a normal distribution and were constant across time points and groups. Additionally, random effects were also normally distributed, and case-deletion diagnostics did not reveal any highly influential observations. The inclusion of a random slope term significantly improved model fit ($\Delta\text{AIC} = -6.01$, $\chi^2(2) = 10.01$, $p = .007$), as did the addition of fixed effects ($\Delta\text{AIC} = -16.47$, $\chi^2(2) = 14.46$, $p = .007$). The addition of a *time* \times *group* interaction term also significantly improved fit ($\Delta\text{AIC} = -1.87$, $\chi^2(1) = 3.87$, $p = .049$), ostensibly due to the steeper negative slope of the monolingual group. Overall, the two groups did not differ significantly in their spelling error rates ($\beta = -0.31$, $p = .109$), and although all children made fewer spelling errors over time ($\beta = -0.19$, $p < .001$), the monolingual group showed a trend for a steeper decline in error rate relative to their EAL peers ($\beta = 0.12$, $p = .053$). This developmental picture is supported by group-level trajectories (Intercepts: Mono = 1.02; EAL = 0.72; Slopes: Mono = -0.19; EAL = -0.07; Figure 4.25). The final model for spelling error rate represented a modest fit to the data ($R^2\text{-conditional} = 0.67$).

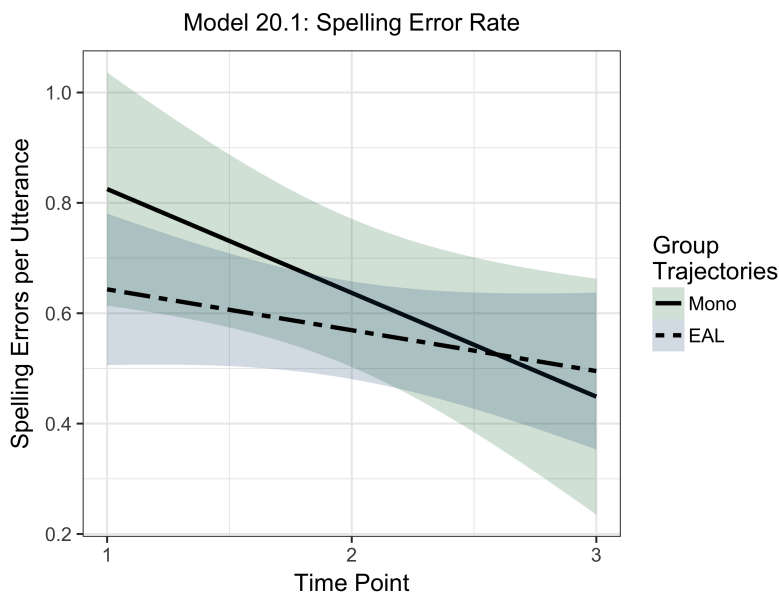


Figure 4.25: Linear Mixed Modelling of t1-t3 Spelling Error Rate in Writing

Writing Error Analysis: Morphosyntactic Error Rate. Error rate data generally followed a normal distribution (5.11% of z-scores ≥ 1.96), although again, the majority of children made no or few errors. Residuals centered around zero and showed constant variance across covariates. Random effects were normally distributed, and case-deletion diagnostics did not reveal any highly influential observations. The addition of a random slope to the UGM did not improve model fit ($\Delta\text{AIC} = 3.98$, $\chi^2(2) = 0.02$, $p = .991$), however the inclusion of fixed effects did improve fit ($\Delta\text{AIC} = -11.84$, $\chi^2(2) = 15.84$, $p < .001$). A *time* \times *group* interaction term did not improve fit ($\Delta\text{AIC} = 0.39$, $\chi^2(1) = 1.61$, $p = .204$) and was removed from the model. In the FGM, children also made significantly fewer errors over time ($\beta = -0.02$, $p = .049$), and in contrast to spelling error rate, children in the EAL group made a significantly higher rate of morphosyntactic errors than their monolingual peers ($\beta = -0.08$, $p < .001$). Again, in contrast to the previous model, it

4.4. Linear Mixed Modelling

was the EAL group that made more errors at t1 and subsequently improved at a faster rate over time (Intercepts: Mono = 0.06; EAL = 0.19; Slopes: Mono = -0.01; EAL = -0.03; Figure 4.26).

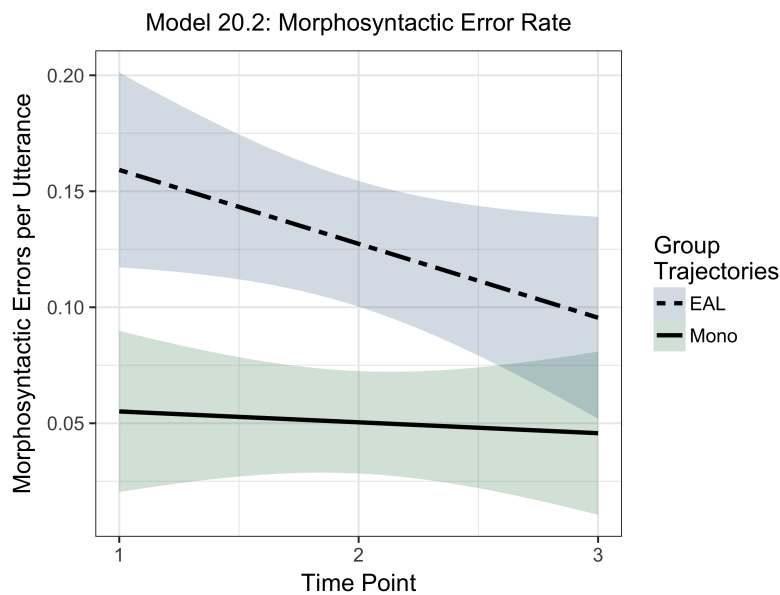


Figure 4.26: Linear Mixed Modelling of t1-t3 Morphosyntactic Errors in Writing

4.4.8 Summary of Linear Mixed Effects Modelling

Before results are discussed within the wider context of relevant literature, a summary of all intercepts and slopes for both groups for all 25 linear mixed models is provided in Table 4.21 below¹¹. In direct reference to research questions identified at the beginning of this chapter, this table provides an indication of the extent to which the monolingual and EAL group differed in performance at t1 (intercepts), whether the trajectories of the two groups were significantly different (β_{group}), and the relative steepness of these trajectories (slopes). The table reveals trends for monolingual group advantages in vocabulary and oral language measures, oral narrative, spoonerisms, all measures of passage reading, and morphosyntactic error rate in writing. In contrast, the EAL group exhibited trends for advantages in oral and written narrative productivity (MLUw and MLTw, respectively), RAN, and single-word reading efficiency. Despite consistent monolingual advantages at t1, EAL learners tended to make a slightly faster rate of progress over time than their monolingual peers, particularly in passage reading accuracy (standard scores), vocabulary depth, and expressive grammar, although this was seldom sufficient to result in convergence between the two groups by t3.

¹¹Note that *time* × *group* interaction terms are taken from Step 4 of the model building process (Table 4.9, page 92). The only FGMs to incorporate an interaction term were Model 16 (YARC passage reading accuracy) and Model 20.1 (spelling error rate); all other interaction terms are presented in this table for comparison only and did not form part of the FGM

Table 4.21: Summary of Intercepts and Slopes for Mono and EAL Group on All Variables

Variable Group	Model	Variable	Intercept		β_{group}	Slope		Time \times Group	
			Mono	EAL		Mono	EAL	Interaction	<i>p</i>
Vocabulary	1	Receptive Vocabulary (BPVS)	96.43	88.02	-2.37 *	7.47	7.75	0.25	.789
	2	Expressive Vocabulary (CELF EV)	36.07	31.43	-2.85 **	2.28	2.64	0.36	.458
	3	Vocabulary Depth (WISC VC)	25.07	22.64	-1.37	1.84	2.29	0.47	.192
Oral Language	4	Listening Comprehension (CELF USP Std.)	8.13	6.43	-1.35	-0.02	0.50	0.52	.057
	5	Expressive Grammar (CELF FS)	30.23	26.93	-2.15	5.09	5.63	0.56	.416
	5.1	Expressive Grammar (CELF FS 17)	20.44	15.99	-2.60 **	1.96	2.82	0.86	.118
	6	Oral Narrative: Utterances	22.36	23.99	3.00 **	-0.58	-0.49	-0.17	.809
	7	Oral Narrative: MLUw	7.13	6.48	-3.53 **	0.60	0.58	-0.01	.941
	8	Oral Narrative: Lexical Diversity	6.09	6.02	-0.69	0.24	0.24	0.00	.999
	9	Oral Narrative: Error Rate	0.066	0.132	3.28 **	0.002	-0.007	-0.008	.449
	9.1	Oral Narrative: Morphosyntactic Error Rate	0.03	0.04	0.019	0.003	0.007	0.004	.624
	9.2	Oral Narrative: Semantic Error Rate	0.03	0.04	0.012	-0.004	-0.004	0.000	.968
Phonological Processing	10	Rapid Letter Naming (CTOPP)	51.46	49.52	-2.05	-7.18	-7.33	-0.27	.783
	11	Rapid Digit Naming (CTOPP)	40.37	36.09	1.94	0.76	1.18	-1.07	.357
	12	Spoonerisms (PhAB)	14.80	11.93	-2.10 *	2.35	2.58	0.23	.679

Note: Std. = standardised score; P3 = YARC passage 3 raw score; * $p < .05$; ** $p < .01$. Statistics for oral narrative error rates reported to up to 3 d.p.
Table continued overleaf.

Table 4.21 Cont'd: Summary of Intercepts and Slopes for Mono and EAL Group on All Variables

Variable Group	Model	Variable	Intercept		β_{group}	Slope		Time \times Group	
			Mono	EAL		Mono	EAL	Interaction	p
Literacy	13	TOWRE Sight-Word Efficiency	59.27	61.29	1.14	3.89	3.94	0.05	.947
	14	TOWRE Phonemic Decoding	30.30	31.35	0.34	3.06	2.90	-0.16	.823
	15	Reading Rate (YARC Std.)	104.29	102.38	-0.07	0.07 *	0.94	0.90	.188
	15.1	Reading Rate (YARC P3)	121.82	121.90	0.88	-15.23	-15.04	0.31	.924
	16	Reading Accuracy (YARC Std.)	105.70	97.46	-3.07 **	-1.34	1.18	2.52	.004
	16.1	Reading Accuracy (YARC P3)	5.32	7.56	1.27 *	-0.91	-1.40	-0.46	.241
	17	Reading Comprehension (YARC Std.)	100.91	97.68	-1.33	0.07	0.74	0.67	.484
	17.1	Reading Comprehension (YARC P3)	4.29	3.33	-0.62 *	0.82	0.95	0.22	.237
	18	Writing: T-Units	5.68	6.46	2.23 *	-0.17	-0.12	0.04	.887
	19	Writing: MLTw	8.89	8.84	-1.58	0.91	0.49	-0.42	.392
	20	Writing: Error Rate	1.02	0.07	0.53	-0.19	-0.07	0.10	.139
	20.1	Writing: Spelling Error Rate	1.02	0.72	-0.31	0.19	-0.07	0.12	.053
	20.2	Writing: Morphosyntactic Error Rate	0.06	0.19	0.07	-0.01	-0.03	-0.03	.188

Note: Std. = standardised score; P3 = YARC passage 3 raw score; * $p < .05$; ** $p < .01$.

4.5 Discussion

The longitudinal study followed the language and literacy development of 33 monolingual and 48 EAL learners from schools within communities of medium to high social deprivation. The children were followed in the run up to the end of their primary school education in KS2, a period in which the National Curriculum places emphasis on oral language and vocabulary development and children are expected to have mastered the low-level skills of literacy such as letter knowledge and decoding (DfE, 2013). Key questions concerned to what extent the two groups differed in their skills at the beginning of Year 4 and how these skills developed across subsequent time points. Importantly, participants were matched not only on chronological age, nonverbal intelligence, and working memory, but also on the amount of English-medium instruction they had received, with all children having attended school in the U.K. since at least Year 1.

The present study makes an important contribution to the small amount of literature on language and literacy development in EAL learners in England by adopting a longitudinal design and robust statistical framework, namely linear mixed modelling. This approach is deemed to be more appropriate for the analysis of repeated measures data through better handling of missing data and allowing variance to be accounted for separately by fixed and random effects (West et al., 2007). Indeed, in many cases model fit was improved by the addition of both random intercepts and random slopes, reducing residual variance and better accommodating variation in children's developmental trajectories. The fitted models here provide a rich source of information about within- and between-subjects variation, thus allowing in-depth examination of developmental trajectories. A small number of group trajectories indicated some degree of nonlinear growth, however the present study was insufficiently powered to incorporate higher-order polynomial terms in linear mixed models (Singer & Willett, 2003). Future longitudinal work may assess learners over longer periods of time and incorporate more time points in order to test statistically for patterns of nonlinear growth.

In general, learners with EAL and their monolingual peers exhibited particular profiles of strengths and weaknesses in language and literacy skills. Specifically, monolingual children tended to show higher levels of vocabulary knowledge and expressive grammar ability, longer and less errorful utterances in speech and writing, and slightly higher passage reading accuracy and comprehension performance. On the other hand, the EAL group showed relative strengths in rapid naming and single-word reading tasks. In terms of developmental trajectories, although the EAL group made a faster rate of progress than the monolingual group on most measures, slopes were not *sufficiently* steep to result in convergence by t3.

The discussion will follow the order of variables discussed in Sections 4.4.3 to 4.4.7: for each variable, the discussion will begin with the extent of monolingual-EAL group differences at t1, followed by the trajectories of each group over time. Findings of the present study will be considered within the context of relevant literature; although longitudinal studies of language and literacy development in mono- and bilingual learners are less numerous (particularly in the U.K.), links will be made to relevant studies when applicable. To ease interpretation, reference will be made to group mean scores in parentheses as well as effect sizes.

4.5.1 Vocabulary Knowledge

Children in the present study were assessed on three measures of vocabulary knowledge, including two measures of breadth (receptive: BPVS-III; expressive: CELF-IV EV) and one measure of depth (expressive: WISC-IV VC).

4.5.1.1 Group Differences in Vocabulary Knowledge at t1

Receptive and Expressive Vocabulary Breadth. In terms of performance at t1, the monolingual group exhibited significant advantages on both measures of vocabulary breadth knowledge. Despite this, the absolute difference in the raw number of correctly identified words by the monolingual group was relatively small, at 7.94 for receptive and 3.77 for expressive knowledge. Effect sizes for group differences at t1 were modest, at $g = 0.51$ for receptive and $g = 0.56$ for expressive knowledge, respectively. Standardised scores revealed that both groups were performing within the average to below-average range on measures of vocabulary breadth (Receptive means - Mono: 87.80; EAL: 80.13; expressive means - Mono: 8.91; EAL: 7.35; somewhat below expectations relative to norming populations of the BPVS and CELF EV).

The monolingual group advantage in vocabulary knowledge breadth found here is supported by other studies of similarly-aged bilingual learners in England and other countries (Babayiğit, 2014a, 2015; Beech & Keys, 1997; Burgoyne et al., 2009, 2011a; Droop & Verhoeven, 2003; Hutchinson et al., 2003; Limbird et al., 2014; Mahon & Crutchley, 2006). However, these studies tend to report group differences of larger magnitude than those found here, for instance: $d = 1.5$ to 2.17 in Hutchinson et al. (2003) using the Test of Word Knowledge; $d = 1.93$ to 2.62 in Droop and Verhoeven (2003) using a Dutch-language multiple-choice breadth measure; and $d = 2.18$ to 2.54 in Lervåg and Aukrust (2010) using the PPVT. However, even the relatively more modest effect sizes found in studies of EAL learners in England, including those using the BPVS, of between $d = 0.75$ and 1.12 (Babayiğit, 2014a; Burgoyne et al., 2009, 2011a; Mahon & Crutchley, 2006) are higher than the magnitude of those found in the present study, which did not rise above $g = 0.71$ for any vocabulary measure at any time point.

One reason for this discrepancy may be the matching of the two groups in the present study on amount of English language instruction, a strategy not typically employed in other U.K.-based studies of EAL learners. However, it should be noted that 9 to 10 year-old EAL learners in Babayiğit (2014a) similarly had a minimum of four years of English-medium instruction, and yet significantly underperformed in relation to their monolingual peers on the BPVS by a magnitude of $d = 1.12$. One alternative reason for this discrepancy may be the extent to which the monolingual comparison group underperformed in relation to the norming population of the BPVS: in other words, had the monolingual group scored closer to the norming population average (i.e. a standard score of 100), the magnitude of the group difference in receptive vocabulary knowledge may have been larger, and therefore more akin to that reported by studies alluded to above. Indeed, the monolingual group did perform in the (low) average range (with a mean raw score of 103 relating to a standard score of 87.80 at t1). Once again, however, comparison with the results of Babayiğit (2014a) poses a challenge to this conclusion. To aid interpretation, Table 4.22 below contrasts group mean raw scores, standard deviations, and effect sizes for the results of Babayiğit

(2014a) and t3 of the present study, when participants in both studies were aged 9 to 10 years old.

Table 4.22: Comparison of Monolingual-EAL Group Discrepancies on the BPVS (Raw Scores)

	Monolingual	EAL	Effect Size
t3 of the present study	116.76 (13.77)	110.84 (16.76)	0.38
Babayiğit (2014a)	100.45 (14.15)	83.28 (16.36)	1.12

Note: Effect size in *g* for present study and *d* for Babayiğit (2014a)

It can be seen here that the relatively larger discrepancy in group performance in Babayiğit (2014a) appears to be due to the particularly low performance of the EAL group, which scored on average around 17 raw scores lower than the monolingual group, in contrast to a group discrepancy of around 6 raw scores in the present study. Another striking difference here is the relatively higher score of the monolingual group in the present study relative to that in Babayiğit (2014a). Therefore, lower than expected performance of the monolingual group in this study as a cause of relatively smaller monolingual-EAL group discrepancies in vocabulary performance is not supported. Rather, it would appear that the relatively more similar performance between the two groups in the present study is due to the *higher* than expected performance of the EAL group relative to the pattern reported in Babayiğit (2014a). Unfortunately, such comparisons with other measures in the battery are not permitted due to use of different measures and age cohorts.

Expressive Vocabulary Depth. Although the monolingual group did exhibit an advantage in vocabulary depth at t1, the magnitude of this trend was smaller than for vocabulary breadth measures ($g = 0.27$) and did not reach statistical significance. Reference to scaled scores revealed that both groups scored within the average range (Mono: 9.70; EAL: 8.90). Again, the absolute difference in the raw number of points scored by the monolingual group was small, at 1.86. To some extent the lack of a significant group difference may be due to the particular measure employed, as the WISC VC has a very limited scoring range of 0 to 2 and therefore may have been insensitive to finely detailed knowledge¹². Therefore, in order to test the robustness of these findings, children's responses on the WISC VC subtest at t1 were subjected to an alternative scoring method emanating from the intervention study in Chapter 6. On this measure, examinees receive a score from 0 to 8 for each stimulus word they are asked to define (for a full description of this measure, see Section 6.9.1 on page 195). Points are awarded across four categories of knowledge, including a straightforward definition, background knowledge (related concepts or personal experiences), lexical knowledge (synonyms, related words and phrases), and non-verbal (for example, gesturing a circular motion for the word *island*). Statistics for group performance and comparisons on this measure are provided in Appendix 4.3 on page 271. Despite the use of this supposedly more sensitive measure, the two groups still did not differ significantly in total score ($F(80,1) = 0.00, p = .985$), although monolingual children did make significantly more use of back-

¹²While many children indicated some level of understanding of the target words, their responses were not always acknowledged by the scoring criteria. For example, when asked 'What is a *pest*?' one child responded, 'pest is a word used by adults to describe children', which received a score of 0 according to the scoring guidelines but nevertheless did indicate some level of knowledge.

4.5. Discussion

ground knowledge in their responses ($F(80,1) = 5.93, p = .017$). While this generally points to a level of similarity between the two scoring procedures, it must be noted that in the bespoke word knowledge assessment, examinees are also required to use target vocabulary in a sentence, thus providing estimates of receptive as well as productive knowledge unlike the WISC VC which focuses exclusively on meaning rather than use (Nation, 2001). Nevertheless, results from the application of this bespoke scoring rubric bear similarity to Vermeer (2001), in which young mono- and bilingual children did not differ significantly in their vocabulary depth knowledge once credit was awarded for nonverbal and exemplar responses.

There is some research to support the existence of monolingual advantages in vocabulary depth. For example, Lervåg and Aukrust (2010) found large monolingual advantages on two separate measures of vocabulary depth knowledge in a large sample of 7 year-olds in Norway of between $d = 1.35$ and 1.59 . Similarly, in a vocabulary definition task with 8 year-olds in the Netherlands, Droop and Verhoeven (2003) found monolingual advantages of a similar magnitude, ranging from $d = 1.38$ to 1.85 . These studies present a striking contrast to the magnitude of the monolingual advantage found in vocabulary depth knowledge in the present study of just $g = 0.27$. This discrepancy is potentially due to differences not only in assessments used (although Lervåg and Aukrust did also make use of the WISC VC), but also to crucial differences in the samples of children recruited; specifically, by age 7, participants in Lervåg and Aukrust (2010) had only just begun formal instruction, and 8 year-old participants in Droop and Verhoeven (2003) were recruited specifically from very low-SES backgrounds. In contrast, the children in the present study had been in receipt of formal instruction since at least Year 1, and although recruited from schools in neighbourhoods of slightly higher than average social deprivation, this did not approximate the same level of deprivation described in the latter study.

Taken together, these findings suggest that the word knowledge of the two groups of children at t1 differed to a larger extent in a quantitative rather than a qualitative sense, as larger discrepancies were observed between measures of breadth than depth. The lack of a significant group difference in depth of vocabulary knowledge represents a novel finding within the U.K. EAL literature, and is further supported by the use of a bespoke rubric which indicated that both groups of children tended largely to produce the same types of knowledge in response to a request for verbal definitions of words. The following subsection will consider vocabulary growth over time, firstly in terms of vocabulary breadth and subsequently in terms of vocabulary depth.

4.5.1.2 Group Differences in Trajectories for Vocabulary Knowledge

Receptive and Expressive Vocabulary Breadth. In terms of both receptive and expressive vocabulary breadth knowledge, both groups of children made a significant rate of progress over time. The average absolute raw increase across both groups in correctly identified words between each time point was higher for receptive (7.66) than expressive (2.49) breadth, reflecting the different ranges and maximum scores of the measures employed (i.e. BPVS and CELF EV, respectively). Reference to the summary of intercepts and slopes in Table 4.21 indicates that despite their lower starting point at t1, EAL learners made a slightly faster rate of progress than their monolingual peers on both measures across the course of the study (receptive slopes = Mono: 7.47; EAL: 7.75; expressive slopes = Mono: 2.28; EAL: 2.64). This slightly faster rate of progress, however, was not sufficient to close the gap by t3, as indicated by the lack of any significant time \times group

interaction term (see Table 4.10). There was a degree of nonlinearity in the data: performance on receptive vocabulary breadth (BPVS-III) for both groups accelerated between t1 and t2, before decelerating by t3. For expressive breadth performance, this same pattern of acceleration and deceleration was found again for the monolingual group but not the EAL group, which showed a fairly constant rate of acceleration across all time points. Nevertheless, in both cases the magnitude of group differences at t3 for receptive and expressive breadth had decreased to below that found at t1, suggesting some convergence between the groups.

Although there is comparatively little longitudinal research on vocabulary development in EAL learners, the findings from the present study do compare closely with others in the literature. The results of Hutchinson et al. (2003) and Burgoyne et al. (2011) bear very close similarity to the trajectories described above for receptive and expressive vocabulary breadth. Although these studies recruited and followed children at earlier developmental stages (Years 2 and 4; ages 6 to 9), results showed that EAL and monolingual groups made a very similar rate of progress across time in both receptive and expressive breadth knowledge, such that monolingual advantages at t1 remained by t3. Interestingly, results of the present study also show a deceleration of the monolingual group in receptive breadth knowledge between the penultimate and final time point, but do not provide evidence of a significantly faster *rate* of growth of EAL learners in vocabulary knowledge. On the other hand, there are longitudinal studies which indicate a significantly faster rate of vocabulary development of bilingual learners: particularly, bilingual children followed between Grades 1-6 in Farnia and Geva (2011) in Canada, and between Grades 2-4 in Simos et al. (2014) in Greece. Additionally, both studies also found evidence of nonlinearity in this development due to faster rates of growth in earlier grades, followed by deceleration in later ones. This is supported to some extent by the results of the present study, although more time points would provide a more robust comparison (cf. 6 in Farnia & Geva and 5 in Simos et al).

Expressive Vocabulary Depth. Similar to trajectories for vocabulary breadth measures, both groups of children made significant progress over time in expressive vocabulary depth knowledge, scoring an additional average of 2.08 raw points at each subsequent time point. Reference to Table 4.21 shows that the EAL group made a faster rate of progress over time than the monolingual group in WISC VC performance (Slopes = Mono: 1.84; EAL: 2.29), although not significantly so, as indicated by the lack of a significant time \times group interaction term. However, the combination of a lower intercept at t1 and a faster rate of development resulted in a closing of the gap by t3, at which point the effect size of the group difference, although negligible, had reversed in the direction of an EAL group advantage ($d = -0.03$).

Studies of early language development show that both monolingual and bilingual learners make progress over time in vocabulary depth knowledge, and there is some evidence to suggest that the trajectories of the two groups are similar (Hadley et al., 2016; Karlsen et al., 2017). Karlsen et al. (2017) assessed a sample of 191 monolingual Norwegian and 66 bilingual Urdu / Punjabi-Norwegian 5 year-olds on measures of receptive vocabulary breadth and depth at the end of kindergarten and again in the first year of school. Results showed that although both groups of children made progress in their vocabulary knowledge over time, they did not differ significantly from one another in the rate of this progress, despite the bilingual group beginning on a lower intercept on both measures. Interestingly, however, and in accordance with results from the present study, the two groups did converge slightly over time in vocabulary depth knowledge

4.5. Discussion

from $d = 1.35$ to 1.06 . Finally, more evidence of the similarity between mono- and bilingual learners in vocabulary depth trajectories comes from intervention studies showing that both groups of children benefit equally from vocabulary instruction, discussed further in Chapter 5 (e.g. Carlo et al., 2004; Silverman, 2007).

Summary. Research question 1 asked to what extent the two groups differed in their performance at t1. Although the monolingual group exhibited advantages on all three measures of vocabulary knowledge, group differences were significant only for measures of breadth and not depth of knowledge. Effect sizes of group differences in all vocabulary variables were smaller than those commonly reported in the literature, especially for vocabulary depth, and it should be noted again that the two groups did not differ to a large extent in terms of absolute number of words correctly identified. The lack of a significant group difference in vocabulary depth is supported by an alternative analysis employing a bespoke scoring rubric, which indicated that children in both groups tended to provide the same kinds of answers when asked to give a definition, with the exception of the monolingual group producing significantly more answers relating to background knowledge.

Research question 2 asked to what extent the developmental trajectories of the two groups were comparable across the course of the study. Results showed that EAL learners made a consistently faster rate of progress over time than their monolingual peers on all vocabulary measures but that, where t1 group differences were significant (i.e. in receptive and expressive breadth), this faster rate of progress was not sufficient to close the gap entirely by t3. In contrast, the monolingual advantage in vocabulary *depth* at t1 was no longer present at t3. One other interesting observation was that while the monolingual group tended to decelerate between t2 and t3 on some vocabulary measures, the EAL group was more likely to maintain a similar trajectory throughout all time points. In sum, these findings suggest that even at t1, the two groups of learners were already performing more similarly to each other than what is reported in previous studies, and that convergence between the two groups in receptive and expressive vocabulary breadth did not take place to any significant degree.

4.5.2 Oral Language

Alongside vocabulary, other measures of oral language in the present study included listening comprehension (CELF USP) and expressive grammar (CELF FS). Due to children attempting different passages on the listening comprehension subtest according to their age, performance was modelled using scaled scores, while development in expressive grammar was modelled with two versions of raw scores. As in the previous subsection, t1 group differences and developmental trajectories will now be considered separately for performance on the two measures.

4.5.2.1 Group Differences in Other Oral Language Measures at t1

Listening Comprehension. At t1, the groups did not differ significantly in listening comprehension skill. Both groups were found to be performing within the average range in terms of scaled scores (Mono: 9.36; EAL: 8.29), but raw data showed a relatively low level of within-subject consistency. At t2, a number of children appeared to have difficulties with the CELF USP subtest: as discussed in Section 4.5.1, CELF USP listening comprehension passages change somewhat abruptly in dif-

difficulty level, and the relatively young age of the children attempting these more difficult passages at t2 was within the lower range of the CELF age band, which was ultimately reflected in children's scaled scores.

The lack of a significant monolingual advantage in listening comprehension here is not supported by other studies of similarly-aged EAL learners in England. For instance, in a cross-sectional study using the same CELF USP subtest, Babayiğit (2014a) found a significant monolingual advantage in performance using raw scores ($d = 0.72$), and Hutchinson et al. (2003) obtained a large and significant monolingual advantage of $d = 1.60$ on an audio-recorded version of the NARA by the end of the longitudinal study in Year 4 (age 8-9). Similarly, significant monolingual advantages were obtained by the studies of Burgoyne and colleagues (2009; 2011a) also using the NARA in this way, albeit of a smaller magnitude ($d = 0.45$ to 0.54). There is some work to suggest that open-ended listening comprehension assessments may disadvantage children with lower levels of oral language competency (McKendry & Murphy, 2011; discussed further in Chapter 8). This speaks to a wider point regarding the oral English language proficiency of EAL learners in the present study, which, as discussed in Section 4.5.1 above, appeared to converge more closely with that of their monolingual peers compared to patterns reported in previous studies. Indeed, it is important to note that EAL learners in Hutchinson et al. (2003), Burgoyne et al. (2009; 2011a), and Babayiğit (2014a) also performed significantly below their monolingual peers in other aspects of oral language including vocabulary and grammar. Therefore, the relatively better-developed English oral language skills of the EAL learners in the present study may have served to improve their performance in listening comprehension.

Expressive Grammar. Expressive grammar was measured through children's ability to produce complete, error-free sentences about stimulus pictures in the CELF FS subtest. There was a moderate though non-significant monolingual group advantage in expressive grammar at t1 ($g = 0.56$). However, as children attempted slightly different subsets of items on the FS subtest according to their age, an additional analysis using only the 17 commonly attempted items at each time point was also conducted. In this analysis, the fixed effect of *group* did reach statistical significance (Model 5.1; Table 4.11), providing more robust evidence of a monolingual advantage. In both models, however, the absolute group difference was small, at 2.15 points for FS raw score and 2.60 for the 17 commonly attempted items. Scaled scores indicate that while the monolingual group tended to score within the average range, the EAL group was performing below age-related expectations (Mono: 7.39; EAL: 5.71) and indeed, relatively more children in the EAL group did reach discontinuation criteria on this assessment.

As reflected in the CELF FS scoring manual, productive use of target words in this assessment requires knowledge of transitivity (e.g. knowing that *gave* requires three arguments and knowing the relationship between these arguments), polarity (e.g. correct use of negation with *unless* and *neither*), and clausal structure and dependency (e.g. coordination and subordination using *and*, *or* and *because*; correct placement of *although*, *instead*, *unless* and *however* in relation to other, dependent clauses). However, the fact that errors not related to the target word are also penalised in this test means that it does not necessarily provide an indication of how able examinees are to apply such rules to the specific words and constructions in question. For example, at t1 one child's response for the word *and* was: *Me and my wife was raking up some soil*. Despite the fact that this child correctly employed *and* as a coordinating conjunction between two noun phrases, the

4.5. Discussion

agreement error (*was*) reduces the score from 2 to 1 and is thus not a reflection of understanding of coordination per se. For this reason, performance on the CELF FS will be interpreted as a more general measure of expressive grammar skill in this study, i.e. not one relating solely to the target vocabulary employed.

The trend for a monolingual advantage in expressive grammar found here is supported by other studies in the literature employing different measures of morphosyntactic skill, including sentence repetition tasks (Babayiğit, 2014a; Droop & Verhoeven, 2003; Komeli & Marshall, 2013), sentence production tasks (Geva & Farnia, 2012; Silverman et al., 2015), oral cloze tasks (Chappelle & Siegel, 1999; Lesaux et al., 2006; Lesaux & Siegel, 2003) and picture judgement tasks (e.g. the TROG; Hutchinson et al., 2003). Again, however, the magnitude of monolingual-bilingual group differences found in these studies is considerably larger than that found here: for instance, $d = 1.13$ on the Recalling Sentences subtest of the CELF (Babayiğit, 2014a), up to $d = 1.70$ on the TROG (Hutchinson et al., 2003), $d = 0.93$ on the CELF FS subtest (Silverman et al., 2015), and up to $d = 1.89$ on a measure of morphological knowledge (Droop & Verhoeven, 2003). Again, it appeared that the two groups of children in the present study were more closely matched in their English oral language skills compared to studies enumerated above. Unlike for vocabulary breadth, it is difficult to determine whether this closer convergence was due to the monolingual group performing lower than expected, or the EAL group performing higher than expected, as standardised scores are typically not reported. Despite this modestly proportioned group discrepancy at t1, however, it is important to note that expressive grammar continued to be a challenging domain for EAL learners throughout the study, in both oral and written modes of expression, as discussed further in Sections 4.5.3 and 4.5.7.

4.5.2.2 Group Differences in Trajectories for Other Oral Language Measures

Listening Comprehension. Linear modelling of listening comprehension performance revealed significant progress over time, constituting an increase of 0.28 scaled scores at each subsequent time point for all children. However, disaggregation of data indicated strikingly different trajectories of each group: as shown in Table 4.21, while the monolingual group plateaued in listening comprehension performance across the study, the EAL group made an average improvement of 0.50 scaled scores at each time point. Despite these different trajectories, a time \times group interaction provided only a marginally significant improvement in model fit, and was not included in the final model. As in previous analyses, a degree of nonlinearity was present in the data, such that the magnitude of the group difference remained fairly stable between t1 and t2 but decreased substantially by t3.

This pattern departs from the findings of Hutchinson et al. (2003) in which an initial monolingual advantage in listening comprehension increased in magnitude over time from $d = 1.28$ to 1.60. Similarly, results here also departed from patterns found in the three time points of Droop and Verhoeven (2003) in which a monolingual advantage in listening comprehension first increased and then decreased over time; by the end of this latter study, too, a significant monolingual advantage remained of between $d = 0.82$ and 1.13, substantially larger than effects found in the present study.

Expressive Grammar. Both groups of children made a steady and significant rate of progress over time on the CELF FS, with the EAL group making a slightly faster rate (slopes = Mono: 5.09;

EAL: 5.63). This faster rate of progress appeared to result in convergence between the groups by t2 (from $g = 0.56$ to 0.21), but a slight deceleration of the EAL group at t3 served to widen this difference once more ($g = 0.40$). A very similar pattern was obtained using the 17 commonly attempted items only (see Tables 4.3 and 4.11).

Longitudinal studies present a rather mixed picture regarding developmental trajectories in grammatical development of mono- and bilingual comparison groups. Firstly, in agreement with results of the present study, Droop and Verhoeven (2003) found that large initial monolingual advantages in morphological knowledge decreased in magnitude from Grade 3 to 4 (e.g. from $d = 1.50$ to 0.98 for a Turkish-Dutch comparison group, and from $d = 2.34$ to 1.89 for a Moroccan-Dutch group). In contrast, however, the results of Hutchinson et al. (2003) present a more equivocal picture, with EAL learners between Years 2 and 4 first diverging and then converging with their monolingual peers in receptive grammar performance (TROG; a similar pattern applied here to CELF FS raw score). An interesting comparison can be made with Geva and Farnia (2012), in which the CELF-III FS subtest was administered as a measure of expressive grammar. In Grade 5 (age 10-11), the monolingual group in this study showed a modest advantage of $d = 0.43$, representing striking similarity with the effect found in the present study of $g = 0.40$. It is reported that 81% of participants in Geva and Farnia (2011) were born in the host country of Canada, and thus were likely being exposed to English from a very young age; in a similar fashion, amongst the 36 children from the EAL group who returned parental consent forms in the present study, 92% were born in England. Thus, the similarity between the two studies serves as further evidence that monolingual-bilingual group differences appear to be reduced in magnitude when bilingual learners have had a greater deal of exposure to the target language.

Summary. In terms of research question 1, group differences at t1 were apparent in both listening comprehension and expressive grammar skill, in which EAL learners tended to begin on a lower intercept than their monolingual peers. This monolingual advantage, however, was statistically significant only for expressive grammar when analysis was based upon the stimuli commonly attempted by all children. As with vocabulary measures, group differences in this study were generally smaller in magnitude than those found in the literature, and in which groups are not matched on educational experience.

Regarding research question 2, despite the fact that children with EAL started on a lower intercept than their monolingual peers in oral language measures, they did make a faster rate of progress over time, although not significantly so. Convergence between the groups was also found in listening comprehension, although reliance on standardised scores alone provides a less detailed picture of development, and the crossing of the CELF USP age threshold at t2 represented a considerable increase in difficulty for some children, potentially biasing analysis of development.

Interestingly, similar patterns emerged across vocabulary and other oral language measures: as unconstrained skills, children continued to improve in their vocabulary and grammar performance over time, with steady but gradual convergence between the two groups. Small group discrepancies in these skills present a picture in which EAL learners were more closely matched to their monolingual peers in terms of their oral English language proficiency relative to previous studies which tend to report group discrepancies of considerably larger magnitude.

4.5.3 Oral Narrative

Children's morphosyntactic skills were assessed through an oral narrative task in which they were asked to retell a short story with the help of a picture book (Peter and the Cat). Variables included the total number of utterances, MLU in words (MLUw), lexical diversity (Root TTR), and error rate per utterance, which was divided into morphosyntactic and semantic error types. All transcripts were transcribed and analysed using CLAN (MacWhinney, 2000).

4.5.3.1 Group Differences in Oral Narrative Measures at t1

An interesting pattern emerged at t1 in which children with EAL produced a significantly higher number of utterances ($g = -0.45$) which were significantly shorter in length ($g = 0.52$) and contained significantly more errors ($g = -0.64$) than those of their monolingual peers. While analysis of error types (morphosyntactic and semantic) posed some challenges for statistical modelling, the retellings of the two groups did not differ in lexical diversity ($g = 0.01$), interpreted as the amount of novel vocabulary employed. In other words, despite the consistent monolingual advantage in vocabulary knowledge (Section 4.5.1), the two groups of children tended to employ similar lexical knowledge during an expressive oral retell task. Children with EAL were found to make a higher proportion of overregularisation and past tense errors than their monolingual peers, with both groups producing similar proportions of agreement errors (see Section 4.3.3).

Although much of the research on oral narrative skill concerns language impairment or comparison between bilingual children's retellings in each of their languages, some studies have also compared monolingual and bilingual typically developing children (Hipfner-Boucher et al., 2014; Pearson, 2002). The Miami study (Pearson, 2002) compared oral narrative skills in monolingual English and bilingual Spanish-English groups in Grades 2 and 5, the latter split into bilingual students from immersion school contexts and two-way contexts (for purposes of comparison, reference will be made here to the immersion group only, as this best mirrors the educational experience of EAL learners in England). In Grade 2 (age 7-8), the oral narrative retellings of the two groups using the Frog Story differed significantly on a composite score of language including complex syntax, lexicon, and morphosyntactic accuracy. As found in the present study, the monolingual group did show a tendency for a higher MLUw (Mono: 7.7; Bilingual: 7.2), but also higher lexical diversity as measured through number of different words (NDW; although it should be noted that NDW does not take into account the differing length of children's retellings, Malvern et al., 2004; see Section 3.4.4.1). In contrast, the present study, which utilised a different measure of lexical diversity (Root TTR) that was able to take into account the significantly higher number of total utterances produced by the EAL group, did not find substantial group differences in the vocabulary employed during children's retellings (Mono: 6.28; EAL: 6.28).

The results of the present study also compare similarly with those of Hipfner-Boucher and colleagues (2014) in which young mono- and bilingual children (aged 3;10 to 5;9) retold narratives using the Bus Story. While the retellings of the groups did not differ in their overall structure and total number of utterances, significant monolingual advantages were found in NDW, number of grammatically acceptable utterances, and MLUw; specifically, it was children with a lower amount of English language exposure (the 'ELL minority' group) who produced significantly shorter utterances (ELL-Minority: 5.99; ELL-English: 7.72; monolingual English: 8.04). Ratings of MLUw

in the present study bore resemblance to those reported in both Pearson (2002) and Hipfner-Boucher et al. (2014), being in the region of 7 to 8 words per utterance, although this is somewhat higher than that reported in the norming study of Rice et al. (2006) of around 5 words per utterance for 8 year-old monolingual children. This difference perhaps speaks to a limitation in terms of the relatively small language samples produced¹³, as a minimum threshold of around 100 complete and intelligible utterances is often recommended and implemented in language sample analysis (Lee, 1971; MacWhinney, 2000; Rice et al., 2010). However, it should be noted that studies with younger children, for example, do often use a much lower threshold of between 5 to 15 complete and intelligible utterances (e.g. Hipfner-Boucher et al., 2014; Rojas & Iglesias, 2013).

4.5.3.2 Group Differences in Trajectories for Oral Narrative

In general, both groups of children showed very similar trajectories in oral narrative variables, which served to maintain group differences found at t1. Particularly, this developmental pattern was one of slightly fewer but longer utterances over time, with a modest increase in lexical diversity. Both groups showed only subtle changes in error rate over time, with the monolingual group making slightly more and the EAL group making relatively fewer errors per utterance. However, trajectories differed according to error type, morphosyntactic error rate increasing relatively more for the EAL group, and semantic errors decreasing to a similar rate in both groups. Thus, by the end of the study at t3, the monolingual group had maintained its advantage in terms of its higher MLUw and lower error rate.

Currently there is very little longitudinal work examining oral narrative development of mono- and bilingual populations in tandem. However, some indication of developmental trajectories is provided by studies discussed above, as well as studies of oral narrative development in bilingual samples only. In the cross-sectional Miami study (Pearson, 2002), different groups of monolingual English and bilingual Spanish-English participants were recruited in Grades 2 (age 7-8) and 5 (age 10-11). While conclusions from this study regarding developmental trajectories are therefore limited, it should be noted that both grade cohorts of children were very similar in terms of SES and home language. Results showed that while group differences in a composite measure of language (i.e. lexicon, complex syntax, and morphological accuracy) were present in both grades, the magnitude of this difference was smaller in Grade 5, suggesting some 'convergence' between the Grade 5 groups in expressive language skill, and specifically morphosyntax. Additionally, while all children in Grade 5 produced longer utterances (MLUw) than in Grade 2, this difference was significant only for the bilingual group, showing a faster 'rate of progress' over time, albeit in a cross-sectional sense. This pattern contrasts sharply with the almost equivalent MLUw trajectories of the two groups in the present study (slopes = Mono: 0.60; EAL: 0.58). The 'trajectories' of the two groups in Pearson (2002) also differed on the subcomponents of the language composite score: while the monolingual group made little or no progress between Grade 2 and 5, the bilingual group showed significantly higher scores in Grade 5 than in Grade 2, again suggesting 'convergence' among the sample of older children. The two groups showed very similar upward trajectories for morphosyntactic accuracy, unlike results in the present study which indicated an

¹³Children in the present study produced between 20.66 and 24.04 utterances on average across the three time points.

4.5. Discussion

initially higher error rate of the EAL group, followed by a slight decrease in error rate over time (in contrast to the monolingual group which exhibited a relatively flat trajectory across time points).

In a somewhat limited fashion, comparison may also be made here to studies which investigate oral narrative development in each of bilingual children's two languages separately. For instance, across the six data collection points in Rojas and Iglesias (2013) and four in Miller et al. (2006), young Spanish-English bilingual students who retold the Frog Story were found to make a continuous, linear, upward rate of development in NDW. The two studies also indicate an upward progression in MLUw in their bilingual samples, rising steadily from kindergarten until Grade 2 and 3, where means were 7.97 (Grade 2; Rojas & Iglesias, 2013) and 7.1 (Grade 3; Miller et al., 2006), and thus fairly closely matched to the mean MLUw of the EAL group at t1 in this study (7.07). However, one point of departure in relation to Roja and Iglesias (2013) is the relatively steeper slopes of children's progression in MLUw, up to a rate of 1.18 words between the final two time points in this study; in comparison, the upward rate in raw MLUw between subsequent time points in the present study was in the range of only 0.45 to 0.75. The relatively slower rate of growth in MLUw here may be accounted for to some extent by the older age of the participants (ages 8;2 to 9;4 at t1), as growth modelling studies report a relatively faster rate of development in the early stages, followed by deceleration (e.g. Simos et al., 2014; Farnia & Geva, 2011).

Summary. Firstly, with regard to research question 1, performance at t1 revealed a situation in which children with EAL produced slightly but significantly more utterances which were significantly shorter in length and contained more errors than those of their monolingual peers. In contrast, the two groups did not differ in terms of the diversity of the vocabulary they employed during retelling. EAL learners showed marginally significant trends for a higher rate of morphosyntactic as well as semantic errors in spoken language.

Secondly, regarding research question 2, both groups made significant progress over time in their general productivity (i.e. total number of utterances), MLUw, and lexical diversity. Interestingly, there was a trend for both groups of children to make *more* morphosyntactic errors and *fewer* semantic errors over time. Results in the present work contrast with studies showing a greater degree of convergence between monolingual and bilingual groups of learners, which is likely an artefact of the similar intercepts of the two groups at t1 and their similar rates of progress over time. These data represent a novel contribution to the field in providing a developmental portrait of oral narrative development in older children learning EAL alongside that of their monolingual peers. Specifically, results here point to the need for sustained focus on the expressive grammar skills of EAL pupils as justified by the finding that as a group, EAL learners continued to make a similar level of morphosyntactic errors in their spoken language across the 18 months of the study.

4.5.4 Phonological Processing

All children were assessed on their rapid automatized naming (RAN) ability of letter and digit stimuli (CTOPP), as well as their ability to simultaneously store and manipulate phonological stimuli in a spoonerism task (PhAB).

4.5.4.1 Group Differences in Phonological Processing at t1

Rapid Automatisated Naming. At t1, the two groups did not differ significantly in RAN of letters ($g = 0.41$) or of letters ($g = 0.14$)¹⁴. Reference to scaled scores revealed that both groups were performing within the average range for RAN of letters (Mono: 10.30; EAL: 11.08) as well as RAN of digits (Mono: 9.36; EAL: 9.48), somewhat higher than performance in other measures discussed thus far (see Tables 4.2 and 4.3 for comparison of standardised scores on vocabulary and other oral language measures).

Research evidence for advantages of bilingual learners in RAN tasks is somewhat mixed. The longitudinal studies of Geva and Farnia (2012) and Geva and Yaghoub Zadeh (2006) in Canada both report statistically significant bilingual advantages in RAN of letters in samples of Grade 5 ($d = 0.45$) and Grade 2 ($d = 0.66$) children, respectively, and in England, 10 to 11 year-old Sylheti-speaking EAL learners in Frederickson and Frith (1998) significantly outperformed their monolingual peers on a measure of rapid digit naming ($d = 0.42$). However, this contrasts with other findings of significant monolingual group advantages using similar cohorts and measures, for instance in Grade 2 in Lesaux and Siegel (2003; $d = 0.32$), and Grade 1 in Geva, Yaghoub Zadeh, and Schuster (2000; $d = 1.32$) as well as Chiappe and Siegel (1999; object naming speed; $d = 0.42$). However, it is important to note that children in these studies were in earlier stages of literacy acquisition relative to children in the present study. Additionally, comparison with other work generally is difficult given that studies tend to assess only RAN of letters and not digits, and often bilingual students are not matched to their monolingual peers in terms of length of residence or instructional experience (for example, participants in Jean & Geva (2009) and Geva and Yaghoub Zadeh (2006) had a minimum of four months residence in the host country; although participants in Chiappe and Siegel (1999) had all received at least one year of instruction). In contrast, the two groups of children in the present study, who were matched on amount of English language instruction, tended to perform differently only to a small or moderate and non-significant degree.

Spoonerisms. In contrast to RAN measures, the fixed effect of *group* was significant for spoonerism performance, with monolingual children outperforming their EAL peers (g at t1 = 0.42). There were a number of children in both groups who particularly struggled with the spoonerisms subtest, and who tended to perform similarly across subsequent time points, as indicated by the relatively high ICC of 0.81 (Model 12 in Table 4.13). Unfortunately, standardised scores were not available for this measure.

To the author's knowledge, there exists only one published study comparing the spoonerism performance of EAL and monolingual learners in the U.K. Frederickson and Frith (1998) similarly found statistically equivalent spoonerism performance of 9 to 10 year-old monolingual children in comparison to their Sylheti-speaking peers in London. It is also important to note that the EAL learners in this study had also been in receipt of formal education since age 5. One possible reason for the relatively higher spoonerism performance of the monolingual group in the present study may have been the availability of additional vocabulary knowledge from which to draw during the task. Indeed, t1 BPVS raw scores correlated significantly with t1 PhAB raw scores to similar degrees across the monolingual ($r = .437, p = .011$) and EAL group ($r = .462, p < .001$). Addition-

¹⁴The reader is reminded that a higher score on the CTOPP subtests is interpreted as slower naming speed and thus a positive effect size indicates faster naming speed of the EAL group.

4.5. Discussion

ally, a relatively stronger correlation was observed between t1 CELF EV raw score and t1 PhAB score in the monolingual ($r = .503, p < .001$) than the EAL group ($r = .326, p = .024$). Although this trend was suggestive of an association between expressive vocabulary and spoonerism performance, strength of correlation did not differ significantly between the two groups ($z = 0.91, p = .181$).

4.5.4.2 Group Differences in Trajectories for Phonological Processing

Rapid Automatisised Naming. Growth in children's performance in RAN of letters showed clear non-linearity, with both groups of children first decelerating between t1 and t2 and then accelerating by t3. Interestingly, the overall growth pattern (Model 10 in Table 4.14) was one of slower performance over time, as children's naming speed at t3 was on average slower than at t1. The magnitude of the (non-significant) EAL advantage remained fairly stable between t1 and t2 ($g = 0.39$ to 0.41) but increased slightly by t3 ($g = 0.52$), with the EAL group naming letters on average 4.81 seconds faster (see Table 4.5). Non-linearity was again present in RAN of digits over time, with both groups speeding up relatively more quickly between t1 and t2, and plateauing by t3. As with RAN of letters, the groups diverged in performance over time, beginning at $g = 0.14$ and finally ending at $g = 0.62$. Thus, for the most part, the two groups mirrored one another fairly closely in their trajectories, becoming slower in letter naming but faster in digit naming over the course of the study.

Some work points to the relatively faster rate of development of bilingual learners in RAN performance. For instance, in both early and later stages of development (i.e. between kindergarten and Grade 2 and between Grades 5 and 6, respectively), bilingual students have been found to make a significantly faster rate of progress than their monolingual peers in RAN of letters (Jean & Geva, 2009; Geva & Farnia, 2012; Lesaux & Siegel, 2003), although unfortunately at the time of writing there are no known studies which chart the RAN development of EAL learners in the U.K.

Spoonerisms. The relatively higher performance of the monolingual group on spoonerism performance at t1 was maintained throughout the study. The magnitude of this group difference remained fairly stable between the first two time points ($g = 0.41$ to 0.42), but decreased slightly by t3 ($g = 0.34$), suggesting some level of convergence between the groups. Again, longitudinal studies of phonological processing, and particularly spoonerism, performance in EAL learners are not available for comparison.

Summary. Regarding research question 1, at t1, children with EAL tended to show strengths relative to their monolingual peers on RAN of letters and digits, whereas monolingual children showed a significant advantage in spoonerism performance. Bilingual advantages in RAN are less robust in the literature, but have nevertheless been found for cohorts of a similar age. In contrast to this work, the results of the present study present no evidence for an EAL group advantage in RAN, perhaps again suggesting that matching groups on educational experience or the generally similar English language proficiency of the two groups served to reduce group differences. Finally, in contrast to the EAL advantage in RAN, it was the monolingual group that performed significantly better on a spoonerism task. It is possible that the higher vocabulary knowledge of the monolingual group may have contributed to this finding. Again, these data represent a novel contribution to the field, as there are currently no published studies examining RAN of both letters and digits, and only one of spoonerism performance in the U.K. EAL population.

Regarding research question 2, while the two groups showed convergence over time in spoonerism performance, trajectories in RAN were more inconsistent. In RAN of letters, for instance, all children generally became slower over time (representing a departure from other studies), first increasing and then decreasing in the total time taken to name all stimuli; in contrast, both groups made rapid improvements in RAN of digits between t1 and t2, before levelling off by t3. Despite this nonlinear development, the EAL advantage became larger in magnitude at each subsequent time point, suggesting that these children were continuing to outpace their monolingual peers over time.

4.5.5 Single-Word Reading Efficiency

Both subtests of the TOWRE (i.e. sight-word reading and pseudoword decoding) were administered as measures of single-word reading efficiency, in which children were required to read as many single words as possible within a 45-second limit.

4.5.5.1 Group Differences in Single-Word Decoding at t1

At t1, the EAL group outperformed the monolingual group on both measures of single-word reading efficiency, but not significantly so. The magnitude of this EAL advantage was relatively larger for sight-word efficiency (SW; $g = 0.25$) than for pseudoword decoding (PD; $g = 0.10$), a finding which was also reflected in the linear mixed model (Table 4.15). Once more, the absolute difference between the groups was small in practical terms, with the EAL group on average reading an additional 1.14 words correctly on the SW subtest, and an additional 0.34 words on the PD subtest. Reference to standard scores shows that both groups were performing within the average range on both subtests (group means between 104.12 and 106.08), in contrast to their slightly lower than expected performance on oral language measures discussed above.

A number of studies report equal or better performance of bilingual children in relation to their monolingual peers in single-word and pseudoword reading skill (Bowyer-Crane et al., 2017; Burgoyne et al., 2009; Chiappe & Siegel, 1999; Droop & Verhoeven 2003; Geva et al., 2000; Geva & Farnia, 2012; Jonejan et al., 2007; Lesaux et al., 2007). For instance, using experimental measures of word reading fluency, Geva and Farnia (2012) found a slight but significant bilingual advantage in Grade 5 children in their longitudinal study ($d = 0.18$), and Droop and Verhoeven (2003) found bilingual advantages in a sample of Grade 3-4 children in the Netherlands, particularly for less orthographically complex words. Despite findings of bilingual advantages, however, other studies using the same or similar measures report very small and non-significant advantages of monolingual children (Jean & Geva, 2009), or equivalence in performance across the two groups (e.g. Chiappe & Siegel, 1999; Lesaux et al., 2007, including nonword reading efficiency; Lervåg & Aukrust, 2010, using the TOWRE).

There is also work from the U.K. to suggest advantages of EAL learners in single-word reading assessments. Amongst a sample of children with oral language weaknesses in Bowyer-Crane et al. (2017), EAL learners in Reception (age 4-5) significantly outperformed their monolingual peers on the YARC Early Word Reading subtest ($d = 0.31$), and similarly performed more highly in Year 1 (age 5-6) on the Diagnostic Test of Word Reading Processes ($d = 0.52$), suggesting that strengths in single-word reading skill are present early in EAL learners' development. This relative strength

4.5. Discussion

has also been found in older EAL learners, namely, amongst samples of Year 4 children (age 8-9) in studies by Burgoyne and colleagues (2009; 2011a) who significantly outperformed their monolingual peers on the word identification subtest of the WRAT-3 ($d = 0.48$ to 0.59). The effect sizes from these studies, again, are larger than the EAL advantage found in the present study, although it should be noted that the TOWRE assesses both accuracy and speed (fluency), and thus results are not directly comparable to non-timed measures.

4.5.5.2 Group Differences in Trajectories for Single-Word Decoding

Both groups of children made a significant rate of progress over time in single-word reading efficiency. The initial EAL group advantages seen at t1 were maintained throughout the study due to the similar slopes of the two groups, in both the SW (Mono: 3.89; EAL: 3.94) and PD subtest (Mono: 3.06; EAL: 2.90). In fact, by the end of the study the t1 EAL advantage had become larger in magnitude, increasing from $g = 0.25$ to 0.40 for SW efficiency, and from $g = 0.10$ to 0.23 for PD efficiency. There was some nonlinearity in development, with group differences remaining stable between t1 and t2, and widening by t3. Interestingly, in the parental questionnaire responses, 54% of respondents indicated that their children had some literacy ability in a language other than English, and in the child language questionnaire administered during testing at t1, 16 children indicated that they had attended or were attending 'Arabic school', or a madrasa, in which children are taught to decode the Qur'an (Rosowsky, 2001). It is possible that some children may have continued to receive this instruction throughout the course of the study, hence accounting for a widening of the EAL advantage in single-word reading efficiency measures. Indeed, phonological skills have been found to transfer across languages (Durgonoğlu et al., 1993), and bilingual children who receive instruction in two scripts have been shown to exhibit advantages in word recognition skills (Leikin, Schwartz & Share, 2010).

The widening bilingual advantage in single-word reading in the present study is supported by other work. For instance, a mixed-language bilingual cohort in Lesaux et al. (2007) performed on par with a monolingual comparison group in kindergarten, but had begun to slightly outperform this group by Grade 4 ($d = 0.14$). Similarly, in the longitudinal study of Droop and Verhoeven (2003), a monolingual Dutch group initially performed very similarly in relation to two bilingual comparison groups in single word reading efficiency, only for the bilingual groups to overtake the monolingual group over time. In this latter study, too, bilingual children increased their advantage relative to their monolingual peers in their reading of orthographically simple and polysyllabic words.

Summary. Regarding research question 1, similar to RAN performance, the two groups of children did not differ significantly in single-word reading efficiency, despite a trend for slightly faster performance of the EAL group. At the same time, it was interesting to note that this appeared to be a relative strength for both groups of children, suggesting that their sight-word reading skills were generally well-developed, and somewhat ahead of their oral language and vocabulary skills. Findings of bilingual advantages in single-word reading are common in the literature, and have also been found in U.K.-based EAL populations.

In terms of research question 2, both groups of children made significant progress over time in their single-word reading efficiency, continuing to perform on average slightly above the norming population average. In accordance with other work, the small and non-significant advantage exhibited by the EAL group at t1 grew in magnitude across the study, although the EAL group

made a slightly faster rate of progress in sight-word reading, while the monolingual group did so in pseudoword reading. It is a limitation of this study that an untimed measure of single-word reading was not included in the assessment battery, as timed and untimed tasks have been shown to differentiate different subgroups of readers (e.g. Meisinger, Bloom & Hynd, 2010; Wolf & Bowers, 1993). However, the data in this study represent an important contribution to U.K.-based studies on literacy development in EAL learners, which typically do not employ measures of single-word reading efficiency.

4.5.6 Passage Reading

The YARC was employed as a measure of passage reading skill, yielding outcome scores for reading rate, accuracy, and comprehension. The overall pattern of results revealed that monolingual children and their EAL peers did not differ significantly from one another in the rate with which they read passages, nor in their ability to comprehend what they had read. However, the two groups did differ significantly in passage reading accuracy, with children in the EAL group making a higher proportion of errors. Additionally, passage reading accuracy was the only variable in any model to show a significant *time* × *group* interaction term, whereby the EAL group made a significantly faster rate of progress over time than the monolingual group, such that the initial monolingual advantage at t1 was no longer present by t3 (a pattern similar to that seen in vocabulary depth).

Given that different children attempted different sets of passages, performance was modelled using standard scores. However, analysis is also informed by reference to raw scores based on passage 3, which 75 out of 81 children read at t1 and all children read at t2 and t3 (see Section 3.4.6.2). Whether analysis was based on standard or raw scores, linear mixed models generally accounted for the highest proportion of variance in passage reading rate and accuracy, and for the lowest in passage reading comprehension, in which children were also less consistent in performance over time.

4.5.6.1 Group Differences in Passage Reading Measures at t1

Firstly in terms of passage reading rate, both groups performed within the average range at t1 (Mono: 104.09; EAL: 103.27), and although there was a slight monolingual advantage ($g = 0.07$), this was not statistically significant. A similar pattern was found with raw scores (total seconds taken to read passage 3), this time with a slightly larger but still small and non-significant monolingual advantage (Mono: 105.00; EAL: 109.16; $g = 0.11$).

Secondly in terms of passage reading accuracy, although both groups scored within the average range, there was a statistically significant and moderately sized monolingual advantage (Mono: 104.76; EAL: 98.48; $g = 0.63$), with the monolingual group making fewer errors such as omissions, mispronunciations, and so on. This was also reflected in the analysis of raw scores based on passage 3 ($g = -0.16$).

Thirdly in terms of reading comprehension, both groups again scored within the average range at t1 (Mono: 103.39; EAL: 97.81), and although there was a tendency for the monolingual group to perform more highly ($g = 0.31$), this was not statistically significant in the modelling of standard scores. On the other hand, modelling of raw scores on passage 3 alone did produce a significant

4.5. Discussion

effect of group, but showed a similarly sized monolingual advantage in reading comprehension performance (Mono: 5.00; EAL: 4.27; $g = 0.40$). A summary of group effects for standard and raw scores in models of passage reading is presented in Table 4.23.

Table 4.23: Summary of Group Coefficients on Passage Reading Measures According to Raw or Standard Score

	Group Effect (β) Standard Scores	Group Effect (β) Raw Scores
YARC Rate	-0.18	0.88
YARC Accuracy	-8.24 **	1.27
YARC Comprehension	-1.33	-0.63 *

Note: YARC = York Assessment of Reading Comprehension; ** $p \leq .01$; * $p \leq .05$.

In general, the comparison of monolingual and EAL group performance in passage reading measures in the present study does not align with that of other research, which tends to report relative strengths of bilingual children in relation to their monolingual peers in passage reading accuracy, and relative weaknesses in passage reading comprehension. Despite this, the lack of a significant group difference in passage reading rate is supported by other work, for instance among 8-10 year-old EAL learners in the studies of Babayiğit (2014a) and Burgoyne et al. (2011a) on the NARA-R, as well as Geva and Farnia (2012) on an experimental text reading fluency measure.

In contrast to results presented here, studies often report bilingual group advantages in measures of passage reading accuracy (Babayiğit, 2014a; Bowyer-Crane et al., 2017; Burgoyne et al., 2009; Hutchinson et al., 2003). In fact, the higher passage reading accuracy performance of EAL learners relative to their monolingual peers in Burgoyne et al. (2009) served to inflate their reading comprehension scores on the NARA-R, as progress on this assessment is determined by the number of reading accuracy errors made, meaning that EAL learners were able to attempt more passages before discontinuing. The opposite pattern was found in the present study, whereby EAL learners tended to make *more* accuracy errors, and thus discontinue *earlier* than their monolingual peers¹⁵. The statistically significantly higher passage reading accuracy scores of EAL learners in Burgoyne et al. (2009, $d = 0.48$) contrasts with the relatively smaller and non-significant EAL group advantages in Hutchinson et al. (2003) who also utilised the NARA-R. Again, however, both studies contrast with the findings in the present study, which revealed a significant monolingual group advantage at t1 ($g = 0.63$). To the author's knowledge, monolingual advantages in passage reading accuracy are reported only in one study (Frederickson & Frith, 1998), which also reported a medium-sized effect for this pattern based on performance on the NARA ($d = 0.49$).

A second point of departure for the findings of the present study in relation to the literature is the lack of a significant difference in the passage reading comprehension performance

¹⁵For this reason, the use of standardised scores in the present study is advantageous in accounting for the fact that different children met discontinuation criteria either due to making too many comprehension errors and/or accuracy errors.

of the monolingual and EAL group (when considering standardised scores¹⁶). Indeed, monolingual advantages in this aspect of passage reading are commonly reported from both international and U.K.-based studies, with effect sizes ranging from around half a standard deviation upwards (Babayigit 2014a, 2015; Burgoyne et al 2009, 2011a, 2011b; Droop & Verhoeven 2003; Frederickson & Frith, 1998; Geva & Farnia 2012; Hutchinson et al 2003; Lesaux et al., 2007; Lervåg & Aukrust, 2010). Additional robust evidence for the weaknesses of bilingual learners in reading comprehension is provided by a meta-analysis which examined literacy skills in mono- and bilingual children (Melby-Lervåg & Lervåg, 2014). This study found a medium-sized monolingual advantage in reading comprehension of $d = -0.62$, the magnitude of which was significantly larger for studies using passage-reading ($d = -0.43$) as opposed to sentence-reading tasks ($d = -0.78$).

Despite support for monolingual advantages in passage reading comprehension in the literature, the present findings of a non-significant group difference in reading comprehension performance do not stand entirely alone in the literature. For instance, Grade 2 bilingual learners in Lesaux and Siegel (2003) also performed on a par with their monolingual peers on the Stanford Diagnostic Reading Test (percentile score; $d = 0.05$), and 5 to 6 year-old EAL learners with oral language weaknesses in Bowyer-Crane et al. (2017) actually slightly outperformed their monolingual peers (also with oral language weaknesses) on YARC passage reading comprehension score ($d = 0.24$), although not to a statistically significant degree. While the oral language weaknesses of the children in the latter study prohibit direct comparison with similarly-focused studies of unselected samples of children, such as in the present work, it is nevertheless interesting that monolingual group advantages were not found when both groups of children were characterised as having oral language weaknesses. This observation also applies to some extent in the present study; as discussed above, reference to standard scores reveals that both the monolingual and EAL group were performing slightly lower than expected in their oral language and vocabulary skills, and did not differ from one another to a large degree in an absolute sense. Therefore, the slightly low level of oral language of both groups may have contributed to their similar reading comprehension performance.

The matching of the monolingual and EAL groups in this study on amount of educational instruction is unlikely to account entirely for the groups' equivalent reading comprehension performance, as despite also being matched in this regard, 9 to 11 year-old EAL learners in Frederickson and Frith (1998) and Babayigit (2014a) were still found to significantly underperform in relation to their monolingual peers on the NARA and YARC, respectively. Alternatively, this discrepant finding may be due to the sensitivity of the YARC: in particular, while the commonly-utilised NARA requires examinees to read all stimulus passages until a pre-specified number of errors is made, scores on the YARC are derived from only the two highest passages attempted. Therefore, the two assessments differ critically in the range of reading material that children attempt, with the YARC being a potentially less sensitive measure of reading comprehension skill (Colenbrander et al., 2017). In other words, different results may be obtained when children's reading comprehension skills are assessed over a larger range of reading material.

¹⁶A significant monolingual advantage in reading comprehension was found when analysis took account of performance on passage 3 alone; however, for the purposes of the discussion, the analysis using standard scores is preferred as a more robust measure of passage reading comprehension skill (see Section 3.4.6.2).

4.5. Discussion

The roughly equivalent performance of children in the EAL group to their monolingual peers in the present study may be better understood with reference to the predictions of the SVR (Gough & Tunmer, 1986), in which reading comprehension performance is predicted both by *decoding* and *linguistic comprehension*. In terms of decoding, EAL learners were found to have relative strengths in single-word reading efficiency (both words and nonwords), and in terms of linguistic comprehension and oral language skills, the EAL group did display some weaknesses in vocabulary knowledge and expressive grammar, but to a far lesser extent in listening comprehension performance. Therefore, taken together, the linguistic-cognitive profile of the EAL group would indeed predict adequate, if very slightly lower, reading comprehension skill in relation to that of the monolingual group. Such a prediction is borne out by the data, which indicate a slight and non-significant monolingual advantage in passage reading comprehension performance in standard scores.

4.5.6.2 Group Differences in Trajectories for Passage Reading Measures

Between t1 and t3, the overall developmental portrait in passage reading performance between the two groups was one in which EAL learners outpaced their monolingual peers in reading rate (standard scores; $g = 0.07$ to -0.21), closed the gap in passage reading accuracy ($g = 0.63$ to 0.05), and made modest gains in comprehension skill ($g = 0.31$ to 0.11). Both groups scored within the average range at all time points. Interestingly, passage reading accuracy was the only linear mixed model to contain a significant *time* \times *group* interaction term ($\beta = 2.52$, $p = .004$; Table 4.16), representing significantly different trajectories of the two groups over time such that the monolingual group's passage reading accuracy score tended to decline while the EAL group's score tended to increase, resulting in a closing of the gap by t3 (see Figure 4.20).

Although the present study departs from other work in terms of the *lack* of a significant monolingual advantage in reading comprehension and the presence of a monolingual advantage in passage reading accuracy at t1, it does accord with other work in terms of the developmental trajectories of the two groups. For example, the longitudinal studies of Burgoyne et al. (2011) and Hutchinson et al. (2003) did not find significant *time* \times *group* interaction terms in any passage reading measures, suggesting that gaps in performance were not closed to a significant degree over time. In the present study, too, the two groups made broadly similar rates of progress with the exception of passage reading accuracy, in which the EAL learners did manage to close the gap by t3. It is also noteworthy that, as found in the present study, EAL learners in Burgoyne et al. (2011) also increased their relative advantage in passage reading rate. In contrast to this work, however, the present study provides a more in-depth examination of developmental trajectories by virtue of reporting group-specific intercepts and slopes (as opposed to an overall main effect of group in a traditional ANOVA). By doing so, the study presents a picture in which all children continue to make progress in passage reading performance, although at relatively different rates according to group status.

An interesting comparison may be drawn between the results of the present study and those of Lervåg and Aukrust (2010), who followed a cohort of monolingual Norwegian and bilingual Urdu-Norwegian children at school entry (age 7-8) across four testing points also over 18 months. All children showed significant growth in passage reading comprehension performance as measured by the NARA-II and the Woodcock Reading Mastery Test-R passage reading subtest, although

the monolingual group began on a higher intercept and, contrary to the results of the present study, went on to make a significantly faster rate of progress than the bilingual group. In contrast, different results were obtained by the growth curve modelling study of Nakamoto, Lindsay and Manis (2007), in which 303 Spanish-English ELLs were followed between Grades 1-3 and 5-6 and assessed on the Woodcock-Johnson-III passage comprehension cloze procedure. When performance was compared with that of a normative English monolingual sample, it was found that the ELLs made a similar rate of progress in passage reading comprehension skill in the early grades, but began to diverge by Grade 5 when their trajectories decelerated. The explanation for this divergence, and for the discrepancy between the results of these two studies, comes from the observation that the mastery of word-level decoding has a longer developmental trajectory in less consistent orthographies such as English (Lervåg & Aukrust, 2010), and as higher demands begin to be placed on oral language and 'reading to learn', bilingual learners, who show weaknesses relative to their monolingual peers in this area, begin to experience comprehension difficulties. In terms of the present study, the EAL group appeared to be at a slightly earlier developmental stage relative to the monolingual group in passage reading accuracy, although it is interesting to note that as EAL learners closed the gap in reading accuracy, they also closed the gap in reading comprehension. However, despite the fact that the children in Nakamoto et al. (2007) and the present study had both been learning to read in English, the present study did not show divergence in performance of the EAL and monolingual group. Such a discrepancy may be due to the socio-economic and cultural make-up of cohorts involved, as well as particular reading comprehension measures employed.

Summary. After an equal amount of English-medium instruction, monolingual and EAL learners in the present study were performing within the average range in their passage reading rate, accuracy, and comprehension. Although the monolingual group did exhibit a slight advantage in reading comprehension, this did not reach statistical significance in an analysis using standard scores. More unexpected, however, was a significant monolingual advantage in passage reading accuracy which, due to the administration and scoring procedures of the YARC, did not serve to inflate comprehension scores. On the other hand, EAL learners did display a slight advantage in passage reading rate relative to their monolingual peers. As such, results depart somewhat from findings of studies with similarly-aged EAL learners in the U.K. (i.e. regarding passage reading accuracy and comprehension). Linear mixed modelling of raw scores on a single passage attempted by most children at t1 and by all children at subsequent time points generally supported this developmental picture, but also revealed a significant monolingual advantage in reading comprehension, and so it must be emphasised that the two groups were not performing entirely similarly overall.

Regarding progress over time, the composition of group differences at t1 had changed for all three aspects of passage reading by the end of the study at t3. Specifically, EAL learners increased their advantage in passage reading rate, made a significantly faster rate of progress in reading accuracy, and showed tentative signs of closing the gap in reading comprehension. Conclusions about closing the gap in passage reading performance, however, must be tempered by reference to the fact that both groups were scoring within the average range in relation to the YARC norming sample. In summary, after a minimum of four years of instruction, and despite significant weaknesses in receptive and expressive vocabulary knowledge, EAL learners generally

4.5. Discussion

performed similarly to their monolingual peers in passage reading skill, and showed evidence of closing gaps in performance over time.

4.5.7 Writing

Written language was assessed using a bespoke assessment in which children were asked to write as much as possible in five minutes about a given topic prompt such as 'My favourite thing to do'. To some extent, children's writing mirrored that of their oral narrative retelling, in that EAL learners produced relatively more T-units, which tended to be of shorter length than those of their monolingual peers. Error analysis showed little difference between the groups, but when disaggregated by type, monolingual children tended to make more spelling errors, while EAL learners made a significantly higher rate of morphosyntactic errors.

4.5.7.1 Group Differences in Writing at t1

At t1, EAL learners produced significantly more T-units than their monolingual peers ($g = 0.41$), although as in oral narrative, the absolute difference between the groups was small (Mono = 5.45; EAL = 6.48). It was noted during testing that a number of children appeared to sacrifice productivity for sentence quality; as a result, the pool of sentences from which to calculate other measures of writing was small, in contrast to the higher number of utterances produced in the oral narrative retell task. Again, similarly to oral narrative retell performance, EAL learners tended to produce shorter T-units than their monolingual peers, but this effect was small in absolute terms. Error analysis indicated no significant difference between the groups in total error rate, although a different pattern emerged once errors were analysed according to type. Specifically, the monolingual group made a slightly, though non-significantly, higher rate of spelling errors at t1 ($g = 0.26$), while EAL learners made a significantly higher rate of morphosyntactic errors, representing a large effect size ($g = 0.81$). The majority of morphosyntactic errors were in over-regularisation, agreement, and missing past tense. Again, the majority of these errors were accounted for by the writing of EAL learners, aside from agreement errors which occurred roughly equally across the two groups.

The study of writing is often considered within a 'simple view of writing' framework (Juel et al., 1986; see Section 2.2.5) in which writing ability is underpinned by transcription factors (e.g. spelling, handwriting, punctuation), as well as composition factors (e.g. quality of writing, generation and organisation of ideas). This framework has been applied to writing development in mono- and bilingual samples of learners. For instance, Silverman et al. (2015) found a significantly higher level of writing quality (i.e. composition skill) among English-speaking monolingual 8 to 11 year-olds using the Test of Written Language, but significantly better spelling ability among a Spanish-English bilingual comparison group ($d = 0.62$). Indeed, the relatively higher spelling ability of the EAL group in the present study is well-supported by research, including a meta-analysis of 18 studies which indicated a mean effect size of $g = 0.81$, showing a large bilingual advantage in real-word spelling tasks across ages 6 to 13 (Zhao, Quiroz, Dixon & Joshi, 2016).

Unfortunately, due to time and resource restrictions, the examination of writing in the present study was limited to productivity (total utterances), mean T-unit length, and spelling and morphosyntactic error rates, so as to form a comparison with oral narrative measures. Nevertheless,

some comparison is permitted between these and other studies examining the writing performance of EAL learners in the U.K. In particular, the bilingual advantage in real-word spelling has also been found in EAL learners in upper primary school (Babayiğit, 2014b), and interesting comparisons may also be drawn between EAL learners' morphosyntactic error patterns in writing and results of Cameron and Besser (2004). This latter study assessed and compared the writing of EAL learners and their monolingual peers in Key Stages 2 and 3, with all EAL learners having a minimum of five years of English-medium instruction. Analyses showed that EAL learners used fewer subordinating conjunctions, shorter verb phrases, made more agreement errors, and omitted a higher proportion of prepositions than their monolingual peers; this latter finding, particularly, is said by the authors to be highly characteristic of writing in EAL learners. Similarly, in the present study, EAL learners tended to produce shorter T-units (in many cases indicative of lower use of subordination), and produced more errors including subject-verb agreement and particularly omission or erroneous use of obligatory elements such as prepositions and determiners (e.g. *it was small ride, or on the six weeks holiday I went to Spain*). Thus, results in the present study further highlight the additional learning needs of many EAL learners in expressive grammar, both in spoken and written language.

4.5.7.2 Group Differences in Trajectories for Writing

Over the course of the study, there was a tendency for productive measures of writing to remain fairly stable, while a relatively greater degree of change was observed in incidence of writing errors. Between t1 and t3, the EAL group maintained its higher number of T-units, increasing slightly from $g = 0.41$ to 0.52. The two groups diverged somewhat in terms of MLTw due to the differences in the steepness of group trajectories (Mono = 0.91; EAL = 0.49). Thus, while both groups continued to produce longer T-units at each subsequent time point, the monolingual group appeared to do this at almost twice the rate of the EAL group. Interestingly, this pattern was not found in oral narrative measures, wherein the two groups made very similar progress over time in MLUw.

On the other hand, longitudinal modelling of writing error rates revealed different patterns depending on the variable used. In terms of total error rate, the monolingual group began on a higher intercept but showed a steeper decline in error rate over time and as a result was making fewer errors than the EAL group by t3 (see Figure 4.24). Further analysis by error type revealed this pattern to be due to a decrease in the spelling error rate of the monolingual group. In contrast, the morphosyntactic error rate of the monolingual group remained very low and stable across the study, while the EAL group made fewer errors over time (Slopes = Mono = -0.01; EAL = -0.03). Despite some convergence between the groups in morphosyntactic error rate, there remained a very clear monolingual advantage by t3 ($g = 0.41$).

Writing development of monolingual and bilingual samples has been less often considered from a longitudinal perspective; however, some work has examined progression in spelling ability. For instance, studies of children in Canada have found evidence of an early bilingual advantage in single-word spelling (Jonejan et al., 2007; Lesaux & Siegel, 2003). In these studies, it was the bilingual group that maintained its advantage on single-word spelling measures, in contrast to findings of the present study which showed an initial bilingual advantage at t1, followed by the tendency for an interaction between time and group, such that the monolingual children were

4.5. Discussion

outperforming their EAL peers by the end of the study. This pattern also contrasts with that reported in Jonejan et al. (2007) in which monolingual and bilingual children made very similar rates of progress over time. However, it is important to point out crucial differences in the analysis of spelling skills in single-word spelling tests of increasing difficulty as opposed to free writing tasks in which children choose which words to attempt. Similar to criticisms levied at the YARC (Colenbrander et al., 2017), different results might have arisen had participants been able to attempt spelling a wider range of words.

Finally, although the study into EAL writing in Cameron and Besser (2004) was not longitudinal in nature, the authors did compare writing samples of children at two educational stages; namely KS2 (age 7-11) and KS4 (age 14-16). The study found a number of similarities between the writing of children in each key stage, including similar use of adverbials and number of subordinate clauses, but also fewer agreement errors in KS2. Thus, the study provides evidence of important differences between the written language abilities of monolingual and EAL pupils up until the end of compulsory education.

Summary. At t1, children learning EAL tended to produce slightly more T-units, although these also tended to be shorter than those of their monolingual peers. Although the two groups appeared to differ little in terms of the average number of errors per T-unit, this picture changed once errors were disaggregated by type: specifically, EAL learners exhibited a lower rate of spelling errors, but a higher rate of morphosyntactic errors. The finding of a bilingual advantage in spelling is supported by the literature, and similarly, there is evidence that EAL learners make a higher proportion of morphosyntactic errors in their writing than their monolingual peers even after five years of instruction.

In terms of progress over time, the two groups made a relatively more stable rate of progress in productive measures (total T-units, MLTw) than in error rate. Effect sizes reversed direction over time for total error rate due to the monolingual group improving more quickly than the EAL group in spelling; on the other hand, while EAL learners did show a tendency to make fewer morphosyntactic errors over time, their average rate of progress was insufficient to close the gap by t3. While there was some degree of similarity between oral and written narrative measures (e.g. in total number of T/C-units, MLTw, and MLUw), some differences were observed in group slopes: specifically, the monolingual group made roughly double the rate of progress of the EAL group in MLTw, but the same rate in MLUw, suggesting some disparity between oral and written language development in the two groups. The close connection between oral and written narrative measures in general, but divergence in development of MLTw, represents a novel finding among this population of learners in the U.K.

4.5.8 Summary

In keeping with results of studies of EAL learners in the U.K. (e.g. Hutchinson et al., 2003; Burgoyne et al., 2009, 2011a; Babayiğit, 2014a), at t1 the monolingual group in the present study exhibited significant advantages in receptive and expressive measures of vocabulary, expressive grammar (CELF FS common items), and spoonerism performance, and produced significantly longer and less errorful utterances in speech than their EAL learning peers. Where EAL learners showed relative weaknesses in other areas, these effects were not statistically significant

(namely listening comprehension, lexical diversity in spoken language, or length or error rate of sentences in written language. In contrast, it was EAL learners who showed significant advantages in measures of oral and written language productivity (total C-units in speech and T-units in writing); although EAL learners showed trends for faster RAN of letters and digits as well as higher single-word reading efficiency, these effects were not statistically significant. These results also generally accord with the international and U.K.-based literature, in which bilingual learners are found to exhibit specific profiles of strengths (lexical access, single-word decoding, spelling) and weaknesses (oral language, vocabulary, and grammar; cf. 'profile effects' Oller et al., 2007; Cline & Shamsi, 2000; Geva & Farnia, 2012; Hutchinson et al. 2003; Burgoyne et al. 2009, 2011a).

However, the results of the present study depart from the literature in terms of passage reading performance, in which EAL learners performed significantly below their monolingual peers in reading accuracy, and in which the two groups did not differ significantly overall in their passage reading comprehension (although this was the case for YARC passage 3 alone). Despite these findings, however, both groups were found to be performing within the average range. Another unexpected finding was the pattern for all children to become slower over time in RAN of letters, while they became faster in RAN of digits.

A number of the group difference effect sizes in the present study are moderately to substantially smaller than those reported in other studies with younger bilingual learners who are not necessarily matched with their monolingual peers on instructional experience. However, the fact that significant group effects were found after three to four years of ordinary classroom instruction suggests that day-to-day school experience is not sufficient to fully address the 'disadvantages' or 'lags' of EAL learners relative to their monolingual peers in certain skills. The need for explicit, targeted, oral language instruction has been recommended for all learners generally (Bercow, 2008) and specifically for children with EAL (e.g. Babayiğit, 2014a; Burgoyne et al., 2011; Hutchinson et al., 2003); indeed, such a recommendation would similarly be warranted from the analyses presented here. Brief comparison with the results of Babayiğit (2014a) does appear to suggest that the EAL learners in this study were performing closer to their monolingual peers than has previously been found in U.K.-based studies. To some extent, the similarity between the two groups may have been due to higher than expected performance of the EAL group, or alternatively lower than expected performance of the monolingual group (as indicated by reference to standardised scores). While further work will be required in order to determine age- or year-level expectations for EAL learners as regarding language and literacy development, it should be emphasised that such similar performance between the two groups should not result in complacency on the part of educational practitioners. Rather, it remains paramount that all learners continue to receive access to rich and high-quality language and literacy teaching (Bercow, 2008).

Following from this conclusion, questions may be asked as to the role of explicit instruction of EAL learners in altering developmental trajectories and promoting a closing of the gap in performance between these children and their monolingual peers. The following three chapters deal explicitly with the design and implementation of a vocabulary teaching intervention for a subgroup of EAL learners from the longitudinal cohort study. Chapter 5 will review literature concerning factors affecting word learning, effective facets of vocabulary instruction, and intervention studies with bilingual learners. Chapters 6 and 7 will then present the methods, results, and discussion of the intervention.

Chapter 5

Literature Review II: Vocabulary Instruction for Mono- and Bilingual Learners

Bilingual children, including EAL learners in England, are often found to have lower levels of vocabulary knowledge than their monolingual peers (Section 2.1.2.3) and indeed, the results of the longitudinal study described in Chapter 4 confirmed this pattern. The input-dependent, unconstrained, and multidimensional nature of word knowledge, as well as its role in reading comprehension (Section 2.3), make it a strong candidate for explicit instruction, especially within samples of bilingual learners who possess adequate decoding skill but lower levels of reading comprehension than their monolingual peers (Murphy & Unthiah, 2015). Due to split exposure, the vocabulary knowledge of bilingual children tends to be distributed across their languages, often leading to a smaller stock of word knowledge in one language than that possessed by a monolingual speaker of that language (Pearson et al., 1993). While this may be an expected developmental pattern in the case of bilingual vocabulary development, it remains incumbent upon educators, especially within a 'monolingual' or 'submersion' form of education such as that employed in England (Baker, 2006; Section 1.1), to promote the vocabulary knowledge of EAL learners who are tasked with accessing the same English language curriculum and assessments as their monolingual peers.

The purpose of the following literature review is to provide some background regarding important factors in children's word learning, with the ultimate aim of informing the design and implementation of a bespoke vocabulary intervention programme described in Chapter 6. The review will begin by considering children's implicit vocabulary acquisition as a result of exposure to oral and written language (but focusing primarily on written language), and will then move on to discuss explicit teaching of vocabulary, namely definitional and contextual approaches. Following this, some key considerations in the design of vocabulary interventions will be discussed, including the selection of words to teach, contextual affordances, active engagement with target words, dosage and multiple exposures, and provision of child-friendly definitions. As much of this literature is based on word learning processes and instruction in monolingual children, the final section of the chapter will focus specifically on vocabulary intervention studies in populations of bilingual learners, including children learning EAL in England.

5.1 Incidental Learning of Word Meanings

School-age children are typically found to increase their word knowledge by a rate of 2,500 to 3,000 words a year (Beck & McKeown, 1991; Nagy & Scott, 2000), however this rate is somewhat smaller if one considers only root words (Biemiller & Slonim, 2001). Such a rate of learning can-

not be due solely to explicit classroom instruction, pointing to the conclusion that children must acquire a large amount of vocabulary through exposure to oral and written language (Cunningham, 2005; Krashen, 1989). Indeed, studies show that children and adults can successfully glean the meaning of novel vocabulary through only brief encounters in texts (Nagy, Herman & Anderson, 1985; NRP, 2000; Martin-Chang & Levesque, 2013; Webb, 2007). Such learning may be described as incidental, as opposed to intentional, in that word meanings are acquired through activities (such as free reading) which do not have the explicit aim of teaching vocabulary, and in which the meanings of unfamiliar words are merely expected to be inferred (NRP, 2000; Swanborn & de Glopper, 1999).

The rate at which incidental learning of vocabulary occurs from reading has been found to be a function of prior vocabulary knowledge and age. In a meta-analysis of 15 studies of incidental word learning of learners in Grades 4 to 11 (ages 9 to 17), Swanborn and de Glopper (1999) found a negative relationship between density of unknown words and probability of novel word learning. For instance, at a rate of 1 unknown word for every 150 words, there is a 30% probability of learning; when the density increases to 1 unknown word for every 75 words, this probability decreases to 14%. Additionally, probability of incidental word learning increased from 8% in Grade 4 to 33% in Grade 11, suggesting that children are able to acquire higher levels of vocabulary incidentally over time, potentially as a result of access to an increasingly large stock of word knowledge, serving to decrease the ratio of known to unknown words. In McKeown's (1985) study in the U.S., 30 monolingual fifth-graders were allocated to groups according to their performance on the vocabulary subtest of the Stanford Achievement Test (15 'low' and 15 'high' ability children). Nonwords were presented within a number of different sentential contexts, for example "Standing in front of it we all agreed that it seemed like a *narp* house" (p.485). When asked to define the target non-words, participants were awarded points for a range of behaviours, including correctly deciphering the word's meaning, giving justifications for their answers based on the context, and correctly discriminating good and bad exemplar sentences containing the target word. The high ability group significantly outperformed the low ability group, suggesting that verbal aptitude, as indicated by prior vocabulary knowledge and verbal reasoning ability, may contribute to the ability to correctly derive meaning from context.

In summary, the results of studies of incidental word learning suggest that novel word meanings may be acquired through exposure alone (Nagy et al., 1985; Martin-Chang & Levesque, 2013; Webb, 2007). However, the probability of this learning is low, especially in contexts in which there is a high density of unknown words (Swanborn & de Glopper, 1999) and for learners with lower levels of prior vocabulary knowledge (McKeown, 1985). As a result, a young EAL learner with a relatively lower level of English vocabulary knowledge may face difficulties in acquiring word knowledge in an incidental fashion in comparison to a monolingual peer who is likely to have a larger stock of word knowledge to draw from. Another issue, not addressed explicitly by studies of incidental word learning discussed above, is the *depth* of vocabulary knowledge that is typically acquired through incidental learning (e.g. exactly what knowledge is acquired in terms of form, meaning, and function from only a small number of exposures; Nation, 2001; see Section 5.3.4 below). Given that EAL learners in England tend to have lower levels of vocabulary knowledge than their monolingual peers and that the two groups of children show similar developmental trajectories in vocabulary acquisition as a result of engagement in the mainstream curriculum (Section

2.1.2.3), a case can be made for explicit and targeted vocabulary instruction for EAL learners in order to close the gap in knowledge with their monolingual peers. The review now turns to the role of explicit instruction in vocabulary learning, and in particular will introduce a distinction between definitional and contextual approaches to vocabulary teaching.

5.2 Explicit Learning of Word Meanings: Definitional and Contextual Instructional Methods

While children appear to learn words from incidental exposure alone (Nation, 2001; Swanborn & de Glopper, 1999), deliberate or 'explicit' instruction in word meaning is also shown to be a powerful source of novel vocabulary acquisition: indeed, meta-analyses support the responsiveness of vocabulary knowledge to explicit instruction, showing a wide range of methods (or combinations thereof) to be effective for young word learners (Elleman, Lindo, Morphy & Compton, 2009; Elleman, Steacy, Olinghouse & Compton, 2017; Hairrell, Rupley & Simmons, 2011; Marulis & Neuman, 2010; Stahl & Fairbanks, 1986; Wright & Cervetti, 2017).

Early research in vocabulary instruction sought to compare the efficacy of definitional and contextual approaches to the acquisition and retention of word knowledge. A definitional approach is one in which word meanings are conveyed explicitly in the form of traditional definitions, while a contextual approach places emphasis upon contextual cues, typically by placing novel words within sentences or passages (Stahl & Fairbanks, 1986). In a meta-analysis of 52 studies involving participants of kindergarten to college age, Stahl and Fairbanks (1986) contrasted vocabulary instruction programmes with a definitional or contextual emphasis, as well as those with a balance between the two. The review firstly looked at the influence of teaching style on vocabulary knowledge in and of itself, and found that all types of approaches were found to result in large and statistically significant improvements ($d = 0.76$ to 2.36). However, when assessing the influence of teaching style on children's passage comprehension skill, the review found that methods with a balanced approach or definitional emphasis were the only ones to result in significant improvements ($d = 1.40$ and 0.76 , respectively). Thus, results suggested that all forms of instruction were successful in improving vocabulary knowledge itself, but that a combination of definitional and contextual information may serve to best promote reading comprehension. Similarly, a more recent meta-analysis of 64 vocabulary instruction studies by Marulis and Neuman (2010) found that young monolingual learners up to age 6 benefited significantly more from definition-only or combined definition-and-context methods as compared with context-only methods. However, it should be noted that the studies in this latter review are likely to have targeted knowledge of fairly simple, concrete vocabulary for children of kindergarten age and therefore may lack direct comparison to vocabulary instruction in older children, which may involve more abstract vocabulary presented in longer and more detailed contexts (e.g. McKeown, Beck, Omanson & Perfetti, 1983; Snow, Lawrence & White, 2009).

Nash and Snowling (2006) compared definitional and contextual approaches to vocabulary instruction with a sample of 24 Year 3 pupils (aged 7 to 8 years) in England. After the screening of an entire Year 3 cohort ($n=71$), those children who scored in the bottom third in receptive vocabulary (BPVS-II, Dunn, Dunn, Whetton & Burley, 1997) and/or a composite measure of vo-

cabulary and narrative (Expression, Reception and Recall of Narrative Instrument, Bishop, 2004) were selected to take part and randomly allocated to either a definition or context condition. All children received two 30-minute teaching sessions once a week for six weeks, covering two words per session. Activities in the definition condition included work on simplified definitions, links to personal experiences, spelling, and recall, whereas children in the context condition read short bespoke passages containing target words and cues and then filled in semantic maps. Although both groups made gains in word knowledge at immediate post-test, the only statistically significant difference was the relatively higher performance of the children in the context condition on a bespoke verbal definitions outcome measure at delayed 3-month post-test. According to the authors, work with passages in the context condition may have been more enjoyable and interactive for children, and the appearance of target words alongside other important semantic, syntactic, and pragmatic information may have resulted in improved ability to give a definition.

Wilkinson and Houston-Price (2013) experimentally manipulated definitional and contextual constraints to investigate word learning outcomes in a sample of 165 monolingual children in two age groups (7 and 9 year-olds). In whole class settings, pupils learned target words which appeared in the BPVS-II using one of two approaches: in a definition condition, words were presented in context either with or without a definition, and in a context condition words were presented in context either three times in the same story, or three times in different stories. Note that all conditions included some form of context in which target words were presented within written passages. Receptive vocabulary knowledge, as measured by knowledge of target and control words from the BPVS-II, was assessed at baseline and at two time points following the end of the teaching (t1 and t2). Although all children showed significant gains in vocabulary knowledge by t1, and retained this at t2, performance on the BPVS indicates only receptive knowledge, in contrast to Nash and Snowling (2006) who also included an expressive outcome measure, and therefore it is unknown to what extent the children acquired additional aspects of word meaning, such as their function (Nation, 2001; Section 5.1). Crucially in this study, however, the addition of a definition resulted in significantly larger gains in word learning, whereas no differences were observed between the two context conditions. Thus, it would appear that the provision of explicit information relating to word meaning was more facilitative of novel word learning than varying the contexts in which words appeared. It should be noted that prior vocabulary knowledge (as measured by baseline BPVS-II raw scores) accounted for 24% of variance in word learning at t1 and a similar amount at t2, supporting the role of prior knowledge in acquisition of novel vocabulary. The results of this study support the role of definitions in word learning but are less conclusive with regard to the role of context, showing no significant differences when words were presented within the same story or across different stories.

In summary, while there is general support for the effectiveness of definitions in vocabulary instruction studies with monolingual learners, the importance of context should not be downplayed, as this is also shown to result in gains in knowledge. The complementary roles of definitional and contextual information may be expected to result in positive instructional outcomes, given that vocabulary knowledge is conceptualised as multidimensional and consisting of not only meaning, but also of form and function (Section 2.1.2). Importantly, the way in which vocabulary knowledge is measured may reveal subtle differences in children's performance according to having received definitional or contextual instruction, i.e. awarding points for contextual information about a tar-

get word or for its correct use within a sentence (this issue is returned to in Section 6.9.1). The following section will consider key considerations in the design and implementation of vocabulary intervention studies based on work with monolingual learners.

5.3 Key Considerations in Vocabulary Instruction with Monolingual Children

Vocabulary intervention studies employ a wide range of word learning strategies, including the use of explicit definitions (McKeown et al., 1983; Nash & Snowling, 2006; Wilkinson & Houston-Price, 2013), semantic maps and illustrations (Clarke, Snowling, Truelove & Hulme, 2010; Nash & Snowling, 2006; Rupley & Nichols, 2005), activities on the morphosyntactic properties of words (Baumann et al., 2003; Elleman et al., 2017; Silverman et al., 2014), use of mnemonics, metacognition and multimedia (Hairrell et al., 2011), connecting new words to previous knowledge and personal experiences (Goerrs et al., 1999), and affording opportunities to use new vocabulary (Gillanders, Castro & Franco, 2014). Furthermore, the National Reading Panel (2000) pinpointed a number of particularly effective practices for vocabulary instruction including the use of context, multiple exposures to words, and active engagement with new vocabulary. The aim of the following section is to review key considerations in vocabulary interventions with an aim to informing the design and implementation of the bespoke intervention described in Chapter 6. Although much of this work has been carried out with monolingual populations, a number of strategies employed in these studies have also been utilised in vocabulary interventions with bilingual learners (e.g. Carlo et al., 2004; Dockrell et al., 2010; Schaefer et al., *under review*; discussed in Section 5.5).

5.3.1 Word Selection

An important consideration in vocabulary instruction is the choice of words to teach. Although it is common for word selection procedures to go unreported in the literature (Marulis & Neuman, 2010), different approaches include: the selection of themes around which vocabulary is chosen (e.g. Word Generation; Snow et al., 2009); selection of vocabulary occurring in official curricula (e.g. Fricke et al., 2013; St. John & Vance, 2014), which may be a desirable strategy due to ‘vocabulary recycling’ over a longer period of instruction (Gardner, 2013); selection of vocabulary in standardised assessments for ease of measurement of progress (e.g. Dockrell, Stuart & King, 2010; Wilkinson & Houston-Price, 2013); selection of words that are already partially known (Biemiller, 2005) or unlikely to be known (Crevecoeur, Coyne & McCoach, 2014); and the use of metrics from large corpora such as age of acquisition, frequency, and imageability (as used in Nash and Snowling, 2006; see also Brysbaert, Warriner & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012).

In their influential Robust Vocabulary Instruction framework, Beck et al. (2002) delineate three tiers of word knowledge which may be considered along a continuum from easy to difficult. Tier-1 words are unlikely to require effortful learning or instruction, as they are commonly used in spoken language between interlocutors who share the same knowledge of these words, for example *chair*, *walk*, *thirsty*, and so on. Tier-3 words, at the opposing end of this continuum, are considered specific or technical in nature, not necessarily known by most people, and rarely

occurring in spoken language, for example *archipelago*, *refraction*, and *mammalian*. Words in the middle Tier-2 category are said to be ‘general but sophisticated’, occurring across a range of domains and being typical of mature language users, for example *absurd*, *pleasant*, and *slumber*. The acquisition of Tier-2 words has applicability to a broad range of literature and typically takes precedence in third or fourth grade (age 8-9) when a good deal of ‘core vocabulary’ has already been acquired (McKeown, 1993). While there are no strict criteria for the selection of Tier-2 words, it is suggested that words be chosen for their importance, utility, instructional potential, and the conceptual understanding they bestow upon learners (Beck et al., 2002). Additionally, targeting of Tier-2 vocabulary is advantageous in that children are likely to already possess vocabulary to express the concept in question, for example already knowing *lucky* before learning *fortunate* (Beck et al., 2005). Prior possession of such conceptual knowledge is therefore likely to aid vocabulary instruction, by striking a balance between the knowledge children are likely to already possess and the new, more nuanced knowledge they are expected to acquire¹.

In summary, utility appears to be a key criterion by which vocabulary is selected for instruction. A number of approaches subscribe to the view that instructional time should be spent on words that are likely to offer the largest benefit to learners, for instance in enhancing their ability to engage with curricula, in terms of the relative ease with which new words may be acquired (e.g. based on their frequency or on learners’ prior vocabulary knowledge), and in promoting comprehension across a wide range of reading materials (Beck et al., 2002; Biemiller, 2005; Bates, 2008). Similarly, while utility is also a key consideration in the selection of words for instruction with bilingual populations, additional recommendations include, where appropriate, targeting of simpler ‘core’ vocabulary which bilingual learners may be less likely to possess than their monolingual peers (Gersten et al., 2007). Indeed, there is work to suggest that a low level of target language proficiency may affect the extent to which EAL learners can benefit from oral language and vocabulary intervention (Dockrell et al., 2010; see Section 5.5).

5.3.2 Contextual Affordances

The contextual environment plays an important role in the probability of acquiring new word knowledge (Beck et al., 2002; Swanborn & de Glopper, 1999). When faced with an unfamiliar word, learners are tasked with sorting relevant from irrelevant cues and combining them to form a mental definition: such cues may invoke descriptive properties, values, and causality, and may be spatial or temporal in nature (Sternberg, 1987). Unfamiliar vocabulary is rarely presented in isolation (Bachman & Palmer, 1996), and therefore contextual information is likely to play a significant role in word learning.

The Robust Vocabulary Instruction framework of Beck and colleagues (Beck, McKeown & McCaslin, 1983; Beck, McKeown & Kucan, 2002) describes a taxonomy of contextual environments as they relate to word learning. On a spectrum from least to most facilitative, contextual environments may be described as: (i) *misdirective*, leading to an incorrect interpretation of the word’s

¹Consider, for example, the Tier-1 word *scared* and the Tier-2 word *petrified*. Prior knowledge of *scared* is likely to aid in teaching the word *petrified* by virtue of the fact that both words represent the same underlying concept. Additionally, comparison between the two words may encourage a subtle distinction in semantics; in this case, *petrified* being used to mean more intense fear (i.e. being derived from the verb ‘petrify’ - to convert into stone; to paralyse with strong emotion).

5.3. Key Considerations in Vocabulary Instruction with Monolingual Children

meaning; (ii) *nondirective*, offering no meaning at all; (iii) *general*, providing only a general or basic sense of meaning and; (iv) *directive*, affording a more highly specified definition. Beck et al. (2002) present example passages to illustrate the role of context in deriving word meaning. For instance, passage (1) below illustrates a nondirective context, in which the meaning of the target word 'lumbering' (moving heavily or clumsily) is not directly arrived at and therefore does not aid word knowledge acquisition to a high degree. In contrast, consider passage (2), which provides a directive context for the target word 'commotion' (a state of disturbance or agitation). Unlike the nondirective context, this example provides cues ('yelping', 'tripping') as well as a definitional phrase ('noise and confusion'):

1. 'Dan heard the door open and wondered who had arrived. He couldn't make out the voices. Then he recognised the *lumbering* footsteps on the stairs and knew it was Aunt Grace.'
2. 'When the cat pounced on the dog, he leapt up, yelping, and knocked down a shelf of books. The animals ran past Wendy, tripping her. She cried out and fell to the floor. As the noise and confusion mounted, Mother hollered upstairs, "What's all that *commotion*?"' (Beck et al. 2002, p.5)

While an exhaustive list of possible context cues is not provided by the work of Beck and colleagues, some recent work has examined a range of contextual devices which may be used for encouraging understanding of unfamiliar vocabulary. Based on a sample of 296 target words found in 13 narrative and expository children's books in the U.S., Dowds, Haverback and Parkinson (2016) devised a rating system for the use of contextual cues used with the aim of promoting novel word knowledge. High degrees of inter-rater reliability were obtained for 16 possible strategies, including: cause and effect (i.e. A causes B); comparison or contrast; features such as concepts, properties, and locations (particularly for nouns); grammatical use; prior knowledge or schema activation; and provision of synonyms or antonyms. The results of this study suggest the availability and use of a repertoire of contextual cues which may be used by authors to increase the probability of word learning as a result of reading, although unfortunately the study did not provide information regarding the typical incidence or popularity of the strategies identified, and it is unknown which strategies are most successful in promoting word learning.

In summary, a discussion of contextual constraints in word learning is facilitative of the design and delivery of a vocabulary intervention; specifically, the type of contextual information in intervention materials may be explicitly manipulated in order to promote likelihood of word learning, a strategy carried out in the design of the vocabulary intervention study (Chapter 6). As intervention work with EAL learners in England has tended to focus on very young children (e.g. nursery and reception, age 4-5) with limited literacy skills, intervention materials in these studies tend to constitute storybooks in which contextual information is not explicitly manipulated in accordance with the taxonomy of Beck et al. (2002) discussed above (e.g. Crevecoeur et al., 2014; Dockrell et al., 2010; Silverman, 2007).

5.3.3 Depth of Processing and Active Engagement

As alluded to in Section 5.1, children acquire some knowledge of unfamiliar words incidentally through reading or listening, but superficial exposure is unlikely to result in depth of word knowl-

edge, especially to an extent to which learners are able to use new vocabulary productively (also see discussion of Webb, 2007 in the following section). Therefore, a further consideration in vocabulary instruction is the role of depth of processing and active engagement in the use of unfamiliar words. In this section, the involvement load hypothesis (Laufer & Hulstijn, 2001) will also be introduced as a theoretical foundation from which to consider acquisition of vocabulary depth and the design of vocabulary learning activities.

Stahl (1985) discusses the concept of depth of processing in vocabulary acquisition, with three stages that move increasingly from receptive to expressive knowledge. At the shallowest level, novel word learning may involve merely *associations* between words, while at an intermediate level of depth, a learner may be able to apply word knowledge in order to indicate *comprehension*, and at the deepest level, learners may be able to *generate* new contexts and definitions for a word. There is evidence for the role of active or deep processing in word learning, typically evidenced through conditions in which learners make use of sentential context, provide novel examples, and engage in discussion around word meanings (Bowyer-Crane et al., 2008; Carnine, Kameenui & Coyle, 1984; Craik & Tulving, 1975; Coyne, Simmons, Kameenui & Stoolmiller, 2004; Dockrell et al., 2010; Gipe, 1979; McKeown et al., 1983; Nation, 2001; Stahl & Fairbanks, 1986). In their meta-analysis, Stahl and Fairbanks (1986) found that all depths of vocabulary instruction were effective to some degree for vocabulary acquisition in and of itself, but that deeper or 'generative' methods were particularly effective for reading comprehension. This finding is supported by the more recent meta-analysis of Elleman et al. (2009), which found that instructional methods with a 'high level' of discussion (i.e. representing a deeper level of processing by utilising background knowledge and presenting words in multiple contexts) were positively associated with gains in vocabulary knowledge, in contrast to methods which did not use such strategies.

One commonly used method in vocabulary intervention studies is to prompt active engagement by encouraging learners to reflect on their personal experiences as they relate to target words, keep records of encounters with new words, and explore shades of meaning and relationships among words (Beck et al., 2002; Clarke et al., 2010; Nash & Snowling, 2006; Stahl & Fairbanks, 1986; Wright & Cervetti, 2017). Although the efficacy of such strategies has not been systematically compared alongside opposing methods, instructional programmes that do incorporate active engagement have been found to produce significant gains in word knowledge for both mono- and bilingual learners (e.g. Clarke et al., 2010; Dockrell et al., 2010; Silverman, 2007) and indeed, active engagement and frequent opportunities to use novel vocabulary are strategies particularly recommended for bilingual learners by researchers and practitioners alike (Gersten & Baker, 2000).

The involvement load hypothesis (Laufer and Hulstijn, 2001) represents an attempt to operationalise depth of processing in the context of vocabulary acquisition. Laufer and Hulstijn (2001) present a motivational-cognitive model consisting of three major components: *need*, representing recognition on behalf of the learner of the need to acquire new knowledge (i.e. when a word is recognised as unfamiliar) or to meet task requirements; *search*, representing an attempt to discover or infer the meaning of unknown words (e.g. looking in a dictionary); and *evaluation*, representing comparison of target word meanings with other words (e.g. through the combination of words in sentence- or passage-writing tasks). The three components vary in their involvement load, ranging between zero, moderate, and strong. For example, a cloze exercise in which learn-

5.3. Key Considerations in Vocabulary Instruction with Monolingual Children

ers are required to select the correct option from among a list invokes moderate *need*, no *search*, and no *evaluation* (as the correct answer is provided). On the other hand, a written composition using target words invokes moderate *need*, no *search*, and strong *evaluation* (as this task involves combination of word meanings with others in context). Importantly, Laufer and Hulstijn (2001) note the applicability of involvement load to any vocabulary learning task, regardless of modality (i.e. oral or written). The hypothesis predicts that tasks inducing a higher involvement load will lead to a higher rate of learning and retention of novel vocabulary than tasks with a low involvement load, as calculated by a total 'involvement index' (e.g. moderate need equals a score of 1, no search equals a score of zero, and strong evaluation equals a score of 2, resulting in an involvement index of 3).

There is empirical support for the predictions of the involvement load hypothesis. For instance, in an early study, Hulstijn and Laufer (2001) assigned 188 adult English as a foreign language students to one of three vocabulary learning tasks of increasing involvement load. Students were tasked with learning 10 unfamiliar words through either (i) a reading comprehension task with marginal glosses of target words; (ii) a reading comprehension task requiring students to choose and fill in target words; or (iii) a writing composition task requiring students to construct short essays using target words, for which explanations and examples of usage had been supplied. At both immediate and delayed (two-week) posttest, the writing composition group exhibited significantly higher retention of target vocabulary than either of the other two groups, providing support for the notion that higher involvement load leads to higher retention of word knowledge; specifically, it is likely that the strong *evaluation* element of composition writing accounted for relatively higher gains, as this represents operationalisation of a deeper level of processing (Hulstijn and Laufer, 2001; Kim, 2008).

Since its conception, the predictions of the involvement load hypothesis have been largely supported by replication studies (Kim, 2008; Nassaji & Hu, 2012; Zou, 2017), although the hypothesis has also been criticised for its arbitrary divisions in determining involvement load (i.e. between 'moderate' and 'strong' involvement; Zou, 2017). For example, there is conflicting evidence on the effect of gradation of involvement load on vocabulary acquisition: while Kim (2008) found no statistically significant differences in vocabulary retention between a group of adult learners who wrote sentences compared to those who wrote compositions, Zou (2017) did find significantly higher retention for a group of students who wrote compositions, despite the two tasks being said to have the same involvement load. Zou (2017) argues for the separation of 'strong' and 'very strong' *evaluation* due to differing levels of structural organisation and planning required by sentence- and composition-writing. Particularly, while sentences in sentence-writing tasks are free-standing, sentences in composition-writing tasks require attention to overall coherence and linkage, thus resulting in a deeper level of processing.

The involvement load hypothesis is deemed to represent an appropriate theoretical foundation for aspects of the present vocabulary intervention study due to its specific focus on acquisition of vocabulary in populations of bilingual learners. Although the hypothesis is centred on adult learners, studies do support the efficacy of deeper processing in children's vocabulary acquisition, and the concept of involvement load does offer direct pedagogical implications. Similar to the Robust Vocabulary Instruction framework of Beck et al. (2002), the involvement load hypothesis

offers opportunities for the explicit manipulation of vocabulary learning activities in order to yield gains in learning and retention.

5.3.4 Dosage and Multiple Exposures

Dosage refers to the total duration, frequency, and intensity of instruction received (Marulis & Neuman, 2010). Repeated exposure is shown to result in better retention of novel vocabulary as compared to a single exposure alone (Elleman et al., 2009; McKeown et al., 1983; Nation, 2001; Webb, 2007). One hundred and thirty five fifth-grade (age 10-11) U.S. monolingual students in Jenkins et al. (1989) were randomly assigned to receive instruction in individual word meanings or instruction in deriving word meaning from context: these groups were then split according to the amount of exposure they received per each of 45 target words (*low* – 1 exposure; *medium* – 3 exposures; *high* – 6 exposures). Results showed a relationship between rate of novel word retention and number of exposures to target words, with significantly higher rates of learning for medium and high exposure groups, resulting in 74% and 89% retention, respectively. Similarly, monolingual fourth-grade students in McKeown et al. (1983) were randomly assigned to one of two frequency conditions in a 5-month vocabulary intervention: those in the ‘some’ condition received between 10-18 exposures to target words, while those in the ‘many’ condition received between 26-40 exposures. Word learning performance was compared to a control group that received no exposure to the target words at all. Both experimental groups showed large gains in word knowledge relative to the control group, but did not differ significantly from one another, suggesting that gains may level off after a certain level of exposure (i.e. around 18 to 26 exposures).

There is some evidence that repeated exposure to novel vocabulary has differential effects on different aspects of word knowledge. Webb (2007) measured the effect of varying exposure to novel vocabulary on receptive and productive forms of knowledge, including form and meaning, spelling, associations, syntax, and grammatical function. A sample of 98 young adult Japanese learners of English as a foreign language were randomly allocated to one of four experimental groups based on the number of exposures they received, in context, to ten nonsense words (1, 3, 7, or 10 exposures). While results showed a general trend for higher scores on all aspects of word knowledge as a function of number of repetitions, the study also found that performance on productive measures only increased markedly after 7 exposures. After one to three exposures, the largest gains were made in receptive knowledge of orthography, syntax, association, meaning, and finally grammatical function. Importantly, these findings serve to underline the importance of taking into consideration multiple aspects of word meaning in studies of vocabulary acquisition, and show that certain forms of knowledge are more likely to develop before others. Unfortunately, it is not known to what extent these findings apply to younger children and specifically EAL learners.

In their meta-analysis of 64 vocabulary intervention studies with young monolingual children, Marulis and Neuman (2010) explicitly examined the effect of dosage as a moderator of vocabulary learning performance. The authors found evidence for the particular effectiveness of vocabulary interventions with short duration (7 days or fewer), low frequency of teaching sessions (18 or fewer), and low intensity (20 minutes or fewer for each session). However, it should be noted that this is likely to be a result of the fact that participants in all studies were in preschool or

5.3. Key Considerations in Vocabulary Instruction with Monolingual Children

kindergarten, and a large proportion of studies involved storybook reading which, by definition, is short and goal-focused in nature. Nonetheless, the study did find that vocabulary programmes of longer duration, higher frequency, and higher intensity were also effective, albeit to a lesser extent, in promoting novel word knowledge in young children. Among the 37 vocabulary instruction studies analysed by Elleman et al. (2009), interventions of short duration (1 to 5 hours) were found to result in significantly higher rates of vocabulary learning than those of longer duration (10 to 20 hours and above). While challenges may be raised to such a conclusion given a bias in the study sample pool of short interventions (52.1%), it may be the case that successful learning was attributable to the highly-focused nature of such programmes on a particular group of to-be-taught words.

In summary, successful acquisition of novel vocabulary requires multiple exposures which are likely to have a differential impact on multiple aspects of word knowledge (McKeown et al., 1983; Jenkins et al., 1989; Webb, 2007). Furthermore, there is some evidence that short, explicitly targeted interventions can produce significant gains in target word knowledge. Such interventions, however, should not be viewed as an alternative to long-term vocabulary classroom instruction; as suggested by Biemiller (2005), short interventions are unlikely to have an impact upon general vocabulary knowledge (e.g. as measured by standardised assessments such as the PPVT or BPVS), as such an outcome would likely require a sustained strategy of the teaching of between 10 and 15 word meanings per week. Indeed, as argued by Nagy (2005), vocabulary knowledge that enables comprehension of increasingly difficult and decontextualised spoken or written language is typically acquired through ongoing exposure to 'rich' language and engagement in instructional methods which, as discussed above, require multiple exposures in multiple contexts. Nevertheless, it is important to note that short, fairly low-intensity vocabulary interventions with mono- and bilingual learners *have* been found to result in significant improvements in subsets of target vocabulary, and thus may represent a feasible strategy for EAL learners in England. While less work has explicitly examined the efficacy of dosage and multiple exposures on the vocabulary acquisition of older bilingual children (i.e. between the ages of 8 and 11), one recommendation for vocabulary teaching among this population of learners does include the targeting of specific vocabulary (August et al., 2005; Gersten & Baker, 2000; Gersten et al., 2007), which could feasibly be achieved in intervention of short duration.

A number of short-term vocabulary intervention studies involving both mono- and bilingual learners do not report the extent to which knowledge has been retained in the period of time succeeding the end of teaching (e.g. Carlo et al., 2004; Dockrell et al., 2010; McKeown et al., 1983), and where such assessments are made, these are often taken in a range between only 2 weeks to 3 months after the end of teaching (e.g. Nash & Snowling, 2006; Silverman, 2007; Wilkinson & Houston-Price, 2013). Follow-up assessments of children's word learning are important for establishing the efficacy of particular vocabulary instructional methods. Finally, while short-term vocabulary interventions with short-term follow-up periods do allow for the comparison of opposing instructional methods (e.g. contextual vs. definitional or single vs. repeated exposures), caution must be taken not to over-extrapolate the apparent efficacy of such studies to wider, long-term vocabulary teaching in schools.

5.3.5 Child-friendly Definitions

Traditional dictionary-style definitions present a number of challenges for young word learners. Particularly, space limitations constrain the ability of dictionaries to differentiate subtle meaning differences and illustrate appropriate usage (McKeown, 1993; Scott & Nagy, 1997). McKeown (1993) argues for the adoption of learner-friendly definitions by focusing on key meaning aspects and putting words into contexts that relate to real life. Indeed, there is evidence for better rates of novel word comprehension of children and adults in studies that contrast traditional with 'revised' dictionary definitions (McKeown, 1993; Gardner, 2007; Nist & Okejnik, 1995). In McKeown (1993), two groups of fifth-grade students were asked to use novel words within sentences and then asked questions about their meanings. In one group, students read traditional dictionary definitions, e.g. 'devious: *straying from the right course; not straightforward*', and in another, students read revised definitions, e.g. 'devious: *using tricky and secretive ways to do something dishonest*'. Students in the revised definitions group were judged to perform better in both sentence production and word comprehension in terms of producing more distinct examples and fewer unacceptable responses. Thus, the style of language used in definitions of target words represents an additional variable in vocabulary instruction which, similar to contextual affordances, may be explicitly manipulated with the aim of promoting novel word learning.

Simplified or child-friendly definitions have been used in intervention studies involving both monolingual learners (e.g. Clarke et al., 2010; Wilkinson & Houston-Price, 2013) and bilingual learners (e.g. Crevecoeur et al., 2014; Silverman, 2007; Schaefer et al., *under review*). Currently there are no published studies comparing the efficacy of child-friendly and traditional definitions in vocabulary interventions for EAL learners in England. However, the use of such a strategy would likely be appropriate for use in this population of learners, given that child-friendly definitions reduce demands on decontextualised vocabulary knowledge.

5.4 Summary

Children's vocabularies increase markedly in size throughout their educational careers, and the source of such learning is unlikely to be due entirely to classroom instruction (Beck & McKeown, 1991; Cunningham, 2005). Despite this, a case may be made for explicit vocabulary instruction for three reasons: firstly, the probability of learning the meaning of an unknown word through superficial exposure to oral or written language alone is low, especially for younger children, and this probability drops in contexts with a higher density of unknown words (Swanborn & de Glopper, 1999); secondly and relatedly, individual differences in prior word knowledge affect the ability to learn word meanings from context (McKeown, 1985), such that not all learners stand to benefit equally from incidental learning – it follows, therefore, that children with lower word knowledge such as EAL learners will be less likely to acquire vocabulary incidentally; and thirdly, mere superficial exposure to unfamiliar vocabulary is unlikely to result in great depth of knowledge, for example that which enables productive use (Webb, 2007).

Vocabulary knowledge is shown to be responsive to instruction and a number of meta-analyses seek to compare the most effective strategies for this pursuit (e.g. Elleman et al., 2009; Marulis & Neuman, 2010; Stahl & Fairbanks, 1986). Debates around the relative efficacy of definitional and

5.5. Oral Language and Vocabulary Instruction of Bilingual Learners

contextual approaches to vocabulary instruction are informed by the purposes of this instruction: purely definitional instruction may be of questionable utility (i.e. teaching vocabulary for the sake of learning vocabulary; Stahl & Fairbanks, 1986), and purely contextual instruction is subject to the vagaries of the amount and type of contextual information provided by unfamiliar target words. Rather, if vocabulary knowledge is conceptualised as a multidimensional construct, with words containing information pertaining to form, function, and meaning (Section 2.1.2), then it follows that a combination of definitional and contextual information is likely to promote depth of vocabulary knowledge, and indeed such balanced approaches are found to be effective for promoting reading comprehension (Stahl & Fairbanks, 1986). Furthermore, a review of key considerations in the design and implementation of vocabulary intervention reveals a number of factors which may be explicitly manipulated in order to promote the efficacy of children's word learning. The following section will explicitly consider oral language and vocabulary intervention studies with bilingual learners, many of which incorporate pedagogical strategies discussed above.

5.5 Oral Language and Vocabulary Instruction of Bilingual Learners

Intervention studies with bilingual children tend to target oral language skills such as vocabulary knowledge, morphosyntax, and narrative ability, given that this is a domain in which bilingual children are often found to underperform in relation to their monolingual peers (Section 2.1). Specifically, recommendations for the target language development of bilingual learners include increased focus on opportunities for oral language expression, targeting of specific vocabulary, and encouraging depth of understanding (August et al 2005; Babayiğit, 2014a; Burgoyne et al., 2011a; Gersten & Baker, 2000; Gersten et al., 2007). Additionally, bilingual learners may make use of first language support, for example in the form of cognates shared between languages and pre-reading materials in the first language (Carlo et al., 2004; Proctor et al., 2011; Mancilla-Martinez, 2010). Such strategies are typically employed in intervention studies conducted in the U.S. with the availability of large groups of children from the same ethnolinguistic community who share the same first language, for example in the case of Spanish-speaking children. Evidently, such a strategy would be more difficult to implement in a U.K. context due to a high degree of ethnolinguistic diversity (Section 1.3.1).

This section will discuss specific examples of oral language and vocabulary intervention studies with bilingual learners generally, and EAL learners in England specifically, in an effort to identify facets of effective instruction with an aim to informing the design and implementation of the vocabulary intervention described in Chapter 6. Much of the intervention work with bilingual learners has been carried out with young children prior to or around the onset of formal education. The review will begin with a discussion of this work before moving on to examine interventions with older bilingual learners in later educational stages.

5.5.1 Intervention Studies with Younger Bilingual Learners

The efficacy of targeted vocabulary instruction is supported in intervention studies with young bilingual learners around or prior to the beginning of formal education. Some specific examples

of multi-component Tier-2 vocabulary interventions will now be discussed, including relevant published studies with EAL learners in England. Silverman (2007) evaluated a book-reading vocabulary intervention amongst a sample of 72 kindergarten children in the U.S., with the explicit aim of comparing the word learning progress of English language learners (ELLs) and their monolingual English-only (EO) peers. Across the 14-week programme, pupils engaged with new vocabulary through literature, and teaching made use of child-friendly definitions, questions and prompts, visual aids, word comparison, repetition and reinforcement, and crucially, opportunities for children to use new vocabulary orally. Word learning was assessed through a multiple-choice picture selection task as well as a verbal definitions task. Analyses revealed a significant increase in target word knowledge between pretest and posttest, with the EO group learning on average 14 words, and the ELL group learning 19. Additionally, despite the fact that the ELL group had a significantly lower intercept on the picture selection task (i.e. a lower level of knowledge of the target words prior to teaching), children in this group made two times the rate of progress in comparison to their monolingual peers over time, and thus were able to 'close the gap' in target word knowledge by posttest. However, this faster rate of progress did not apply to performance on the verbal definitions task in which both groups made similar improvements over time. Indeed, as noted in Section 2.1.4, language assessments of an open-ended nature may be more challenging for children who are continuing to master the language orally (McKendry & Murphy, 2011).

Emerging evidence from the U.K. also points towards the efficacy of explicit oral language and vocabulary instruction of children with EAL. The Talking Time intervention of Dockrell et al. (2010) placed explicit focus on the use and modelling of oral language amongst a diverse sample of 96 nursery-aged EAL learners. Working in short group sessions over 15 weeks, Talking Time was delivered by teachers, nursery nurses, and classroom assistants, with the aim of improving children's knowledge of target vocabulary, understanding and inferences, and narrative skills. Practitioners were trained in the use of specific strategies to promote oral language development, including modelling and recasting, use of open-ended questions, and encouraging children to extend their utterances. Teaching also included the use of books, visual aids, acting out, and discussion of personal experiences. The Talking Time programme was contrasted with a similar but less oral language-focused programme 'Story Reading' as well as a non-intervention control group. At posttest, the Talking Time group showed significantly higher performance in relation to both comparison groups in naming vocabulary of taught words, verbal comprehension, and sentence repetition. However, it is noted that despite the improvement in performance of the Talking Time group, children's oral language was still below the level of their monolingual peers; perhaps it is for this reason that no improvements were seen in narrative skills, which are argued by the authors to place higher demands upon oral language. Nevertheless, in line with recommendations for language instruction of bilingual learners (Gersten & Baker, 2000), this study supports the role of opportunities for active engagement and direct targeting of vocabulary in promoting the oral language skills of young EAL learners. The explicit comparison of Talking Time alongside a similar but less oral language-focused programme indeed serves to delimit the 'active ingredients' accounting for improvements in children's performance.

Similar results are reported by the GetReady4Learning project (Schaefer et al., *under review*), an intervention study aimed at improving the listening, vocabulary, and narrative skills of 80 monolingual and 80 EAL learners in Reception year (age 4-5) in England. The study re-

5.5. Oral Language and Vocabulary Instruction of Bilingual Learners

cruited children deemed as having language weaknesses (based on performance on a composite measure of expressive vocabulary, receptive grammar, and phonological processing), who were randomly allocated to an intervention or waiting control group. Children in the intervention group received three 30-minute group sessions a week and two 15-minute individual sessions a week for 18 weeks, with all teaching being delivered by trained teaching assistants. Teaching was delivered in the form of group sessions, covering vocabulary teaching and reinforcement, training in phonological awareness, and narrative, and individual sessions which allowed the opportunity for teaching assistants (TAs) to consolidate children's learning by focusing on material that children found difficult. Results at immediate posttest and delayed 6-month posttest revealed that the only directly taught skill to improve significantly as a result of the intervention teaching was taught vocabulary; however, this applied only to expressive naming of vocabulary, and not to a verbal definitions task. The authors point to a number of potential causes for the lack of improvement of the intervention group. In particular, this intervention was shorter than other, similar studies, at 18 weeks duration (cf. Fricke et al., 2013 at 30 weeks). Additionally, challenges to implementation fidelity may have impacted results, as children's absence rate and timing constraints meant that a number of individual sessions were not delivered. Researcher observations also revealed that a number of sessions were poorly administered by TAs, despite in-depth and regular training. Finally, participants in this intervention may not have had sufficient language skills to participate in the programme, similar to conclusions reached by Dockrell et al. (2010).

Kotler, Wegerif and LeVoi (2001) assessed the efficacy of Talking Partners, an oral language teaching programme delivered to bilingual children across seven schools in England. Participants were 127 children with EAL between the ages of 5 and 8 years, 64 of whom received three 20-minute small-group sessions per week for ten weeks delivered by trained parent volunteers. Children took part in problem-solving activities which gave opportunities for extended talk and collaboration with peers, although specific vocabulary was not targeted. All children were assessed pre- and post-intervention on receptive vocabulary knowledge (BPVS-II), oral language (Record of Oral Language; Clay, Gill, Glynn, McNaughton & Salmon, 1976), and grammar (Renfrew Action Picture Test; Renfrew, 1988). In general, children in both groups improved in their performance on all measures, although the only variable to show a statistically significant improvement for the intervention group was Renfrew information score (i.e. amount of information provided about stimulus pictures). The authors also report positive changes in the children's communicative competence and confidence as well as more frequent attempts at extended talk. This study is important in considering qualitative changes to children's behaviour and communication as a result of participation in intervention; however, quantitative gains on outcome measures were small and mostly non-significant. This may have been a result of delivery of teaching by parent volunteers, as opposed to school staff (e.g. Schaefer et al., *under review*, Dockrell et al., 2010) or the decision not to target specific vocabulary.

Finally, St. John and Vance (2014) evaluated a short-term Tier-2 vocabulary intervention programme across three schools with a sample of 18 EAL learners in Year 1 (age 5-6) with and without a statement of special educational needs. Across a five-day instructional cycle lasting up to four weeks, class teachers worked with small groups of pupils to promote semantic, phonological, and visual/kinaesthetic knowledge of target words, with activities covering definitions, mind maps, word association, true/false judgement, and examples of use in context. Target words were

selected by classroom teachers in relation to two different curriculum topics (e.g. *journey, road, different, shape, and safe* for the topic of 'local area'). Teachers then taught five words from one topic, while the remaining five served as untaught control words. It is a considerable limitation of this study that target words were not held constant across the 18 children, resulting in possible differences between the three groups of children in terms of the facility and characteristics of the target words on which they happened to be assessed. Pupils' progress was assessed using a bespoke measure administered pre- and posttest to capture growth in a number of areas including definitions, phonological knowledge (e.g. number of syllables), and contextual information such as function or location. Despite substantial variability in implementation fidelity across the three schools (number of sessions ranging from 5 to 20), there was a statistically significant increase in word knowledge of both taught and untaught words between pretest and posttest. A significant increase in knowledge of untaught words may have been as a result of their links with the curriculum, but also the fact that teaching included training in the use of word learning strategies.

In summary, the conclusion reached by these studies is that explicit oral language instruction can be an effective pedagogical strategy for young bilingual learners. Unlike other language skills such as narrative retelling, vocabulary knowledge appears to be more easily and more effectively targeted, perhaps as a result of the goal-driven nature and short duration of such studies. Another key theme emerging from this work is the suggestion that the relatively low target language skills of these young bilingual learners may place limitations on the extent to which they may benefit from instruction; in other words, children may benefit most when they have a sufficient foundation of oral language and vocabulary to draw upon. Indeed, as discussed in Section 1.3.4, for many EAL learners, exposure to English begins in earnest only at school entry and thus it follows that children in the intervention studies discussed above may not have had sufficient English language proficiency in order to benefit from intervention teaching (particularly when asked to retell stories in narrative tasks, for example). Such an observation would lend credence to the specific targeting of to-be-taught vocabulary in line with individual children's prior knowledge. However, such a practice may be difficult in practical terms, potentially requiring extensive vocabulary assessment and analysis of pre-existing knowledge prior to intervention, as well as subsequent targeting of particular vocabulary to be taught. An alternative approach, and the one adopted in the present study, is to target broader categories of words (i.e. Tier-2) which are likely to be beneficial for EAL learners. Discussion will now turn to oral language and vocabulary intervention studies in older populations of bilingual learners.

5.5.2 Intervention Studies with Older Bilingual Learners

There is some research to suggest that bilingual learners in later educational stages may also benefit from explicit instruction using a range of pedagogical strategies. That research has focused on older bilingual learners indicates the ongoing language learning needs of these children; indeed, as reviewed in Section 2.1, bilingual learners, including those children learning EAL in England, are shown to lag consistently behind their monolingual peers throughout primary school education (age 5 to 11) and even in some cases into secondary schooling and beyond (Bialystok & Luk, 2012; Cameron, 2002).

5.5. Oral Language and Vocabulary Instruction of Bilingual Learners

A number of studies with older bilingual learners report successful interventions which target academic vocabulary (particularly the Word Generation programme; Lesaux, Kieffer, Faller & Kelley, 2010; Mancilla-Martinez, 2010; Snow et al., 2009; Townsend & Collins, 2009). With the exception of a study by Kotler et al. (2001) who recruited 5 to 8 year-old bilingual learners (see Section 5.5.1 above), there are no peer-reviewed studies of vocabulary instruction programmes for KS2 (age 7-11) EAL learners in the U.K. Some examples of vocabulary intervention studies from the international literature will be discussed below, including description of the activities employed in order to encourage word learning.

Carlo et al. (2004) assessed a programme aimed to improve the vocabulary knowledge of monolingual English and bilingual Spanish-English fifth-graders (age 10-11) in the U.S. In the treatment condition, small groups of students were exposed to between 10 and 12 words each week for fifteen weeks in sessions lasting up to 45 minutes. Words were selected from texts dealing with issues chosen to encourage debate, and included *allegiance*, *saga*, and *pledge*. As all bilingual pupils in this study were Spanish speakers, they were given the opportunity to preview each week's material in Spanish, and also received instruction involving Spanish-English cognates. Activities included passage reading, cloze tasks, semantic feature analysis, synonym and antonym identification, and morphological derivation. Students in the treatment condition improved significantly more than their comparison group counterparts on all researcher-developed measures including a multiple-choice measure of target word knowledge and a word association (depth) task. Importantly, gains in knowledge were very similar for monolingual and bilingual students.

With a similar sample of fifth-grade students, Proctor et al. (2011) assessed the effect of Improving Comprehension Online (ICON) on depth of students' vocabulary knowledge. Over 16 weeks, students read eight texts containing a total of 40 'power words', chosen as belonging to the Tier-2 category (Beck et al., 2002), as well as having cognate status with Spanish, for example *anxiously* / *ansiosamente*. As well as encountering target words in context, students were provided with definitions, related words, images, and translations, and were required to record personal connections in a wordlog. The Gates-MacGinitie Reading Test (MacGinitie, MacGinitie, Maria & Dreyer, 2002) served as a standardised outcome measure of reading comprehension and vocabulary, while researcher-developed measures (Vocabulary Breadth and Depth Tests; VBT, VDT) explicitly assessed target word knowledge. The VDT yielded separate scores for ability to provide a definition and to draw a picture with captions, which were summed into a VDT total score. Results of hierarchical linear modelling controlling for school-level variance revealed no effect of ICON on reading comprehension, but a significant improvement of students in the treatment condition on vocabulary depth (VDT total effect size $d = 1.34$). Language status was not predictive of outcomes assessed by bespoke measures, again suggesting similar improvements for bilingual and monolingual students, although a language group effect on standardised vocabulary was evident when initial vocabulary level was removed as a covariate.

A recent systematic review (Murphy & Unthiah, 2015) found a general dearth of high-quality language and literacy intervention studies with bilingual learners and a specific paucity of such research conducted outside of Canada and the U.S. Indeed, at the time of writing, there are no known intervention studies in the U.K. context specifically targeting the vocabulary knowledge of older primary school-aged EAL learners in England. Murphy and Unthiah (2015) thus make a

case for the 'urgent need' of such studies with EAL learners in the U.K., warning against the wholesale adoption of findings from North America due to variations in social, linguistic, and educational factors (for example difficulty in the use of cognates and first language support). Encouragingly, the review identified vocabulary instruction as a promising area of research, particularly for children learning EAL who are often found to show poorer reading comprehension ability despite good decoding skills (e.g. Babayigit, 2014a; Burgoyne et al., 2009).

Therefore, it would appear that there is a pressing need for further vocabulary intervention research with EAL learners in England, who are at risk of lagging behind their monolingual peers. In the one published study to do this with a sample of 5 to 8 year-old EAL learners (Kotler et al., 2001) vocabulary was not specifically targeted, and intervention teaching was carried out by parent volunteers. In contrast, more robust effects of specifically targeted vocabulary instruction delivered by school staff have been found in studies of younger EAL learners (Dockrell et al., 2010; Schaefer et al., *under review*; St. John & Vance, 2014), and given that studies often find similar rates of progress among mono- and bilingual children in oral language and vocabulary intervention studies (Crevecoeur et al., 2014; Proctor et al., 2011; Silverman, 2007), it is likely that such strategies could be effectively deployed with samples of older EAL learners.

5.6 Summary of Literature Review II and Aims of Vocabulary Intervention Study

Research shows that vocabulary knowledge is a key developmental need of many bilingual learners – including those in England – who are continuing to master the target language orally alongside monolingual peers who enter formal education with a larger stock of word knowledge (Section 2.1.2.3). Vocabulary is shown to be responsive to explicit instruction (Murphy & Unthiah, 2015) and a number of different considerations and strategies have been found to be effective for both mono- and bilingual learners, including opportunities for active engagement and extended talk, multiple exposures to target words, use of contextual information, and use of child-friendly definitions (Carlo et al., 2004; Dockrell et al., 2010; Silverman, 2007; Proctor et al., 2011; Schaefer et al., *under review*; St. John & Vance, 2014). Much of the vocabulary intervention work with bilingual learners has been carried out in North America, often with fairly homogeneous learner populations: although there is some work from the U.K. context with EAL learners in the KS2 age range (e.g. Kotler et al., 2001), this has been somewhat limited due to not targeting specific vocabulary, and relying on parent volunteers to carry out teaching as opposed to more highly experienced or trained practitioners.

In agreement with other findings from studies conducted in England (e.g. Babayigit, 2014a; Burgoyne et al., 2009), the results of the longitudinal cohort study (Chapter 4) confirmed a monolingual advantage in receptive and expressive vocabulary knowledge which was maintained over time due to the similar developmental trajectories of both groups of children. This pattern suggests that ordinary classroom instruction may be insufficient in order to close gaps in such knowledge, and calls for further research in explicit vocabulary instruction within this population of learners. Thus, the second aim of this thesis was to design and implement a targeted vocabulary interven-

5.6. Summary of Literature Review II and Aims of Vocabulary Intervention Study

tion with a subgroup of EAL learners from the longitudinal cohort study described in Chapters 3-4. Methods and results of the intervention are detailed in Chapters 6-7.

Research questions of the vocabulary intervention study included:

1. To what extent does a short, low-intensity, one-to-one explicit vocabulary training programme improve target vocabulary knowledge in EAL learners who are identified as having vocabulary weaknesses?
2. Specifically, what effect does this teaching have upon (a) children's receptive understanding of taught vocabulary and (b) their ability to use it productively?
3. Taking the approach of a multiple case series design (introduced in Chapter 6), how does an analysis of children's individual growth trajectories inform conclusions about the efficacy of the teaching programme?
4. Does short, low-intensity, one-to-one vocabulary teaching result in transfer onto non-explicitly taught skills such as expressive grammar and depth of vocabulary knowledge?

Chapter 6

Methods II: Vocabulary Intervention Study

This chapter will detail the methods of a short term, low-intensity, one-to-one vocabulary intervention developed for and carried out with a subgroup of EAL learners from the longitudinal cohort study. It will provide information concerning recruitment and selection of participants, the timeline and delivery of intervention teaching, target word selection, session activities, timing and duration, measures used to assess learning, and considerations of implementation fidelity.

6.1 Design and Recruitment

For ease of interpretation, recruitment of participants for the intervention study is depicted in Figure 6.1 overleaf. Children who took part in the intervention were recruited from the EAL group in the longitudinal cohort study (Chapter 3). Consent forms sent home to parents of EAL learners at the recruitment phase for the longitudinal study also asked for consent for parents to be contacted about the possibility of their children taking part in an additional sub-project, i.e. the intervention. Ten out of 48 parents did not give consent to be contacted about this possibility, resulting in a subgroup of 38 children as potential participants in the intervention. A score of -1 SD has been used as a criterion in selection of participants for additional or remedial instruction (Nash & Donaldson, 2005; Semel, Wiig & Secord, 2006; Bishop, 1997) and in the present study this was used to define 'vocabulary weakness', which has been identified as a learning need of EAL learners (Section 2.1.2.3). After t1 of the longitudinal study when all children had completed all assessments in the test battery (Section 3.4), children from the EAL group who obtained standardised scores of -1 SD or below on at least two out of three of the vocabulary measures used, i.e. BPVS, WISC VC, or CELF EV, were eligible to take part in the intervention. Subsequently, information sheets about the intervention and consent forms for participation were sent home to the parents of the 23 children who met this criterion, who were then asked to sign and return them if they wished their child to take part in the intervention teaching. Informed parental consent was received for 12 children in February 2016 (autumn term of Year 4), who were then officially recruited onto the intervention phase of the project. Note that these children continued to take part in the longitudinal project, and were assessed on the battery of language and literacy skills along with other children in the EAL and monolingual group at t2 and t3. The decision not to disqualify the intervention participants from the longitudinal study was justified by lack of overlap between intervention teaching materials (i.e. target words taught) and stimuli in standardised assessments of the longitudinal study test battery. Also note that potential transfer of intervention instruction on children's general language skills will be explicitly assessed in Section 7.2.4.

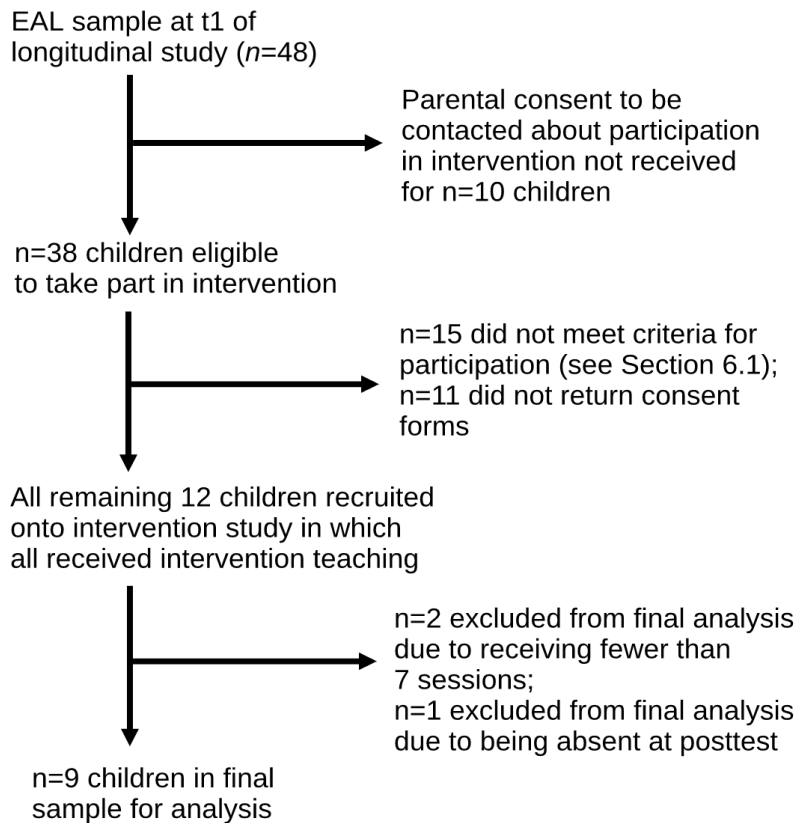


Figure 6.1: Participant Flow throughout the Intervention Study

6.2 Ethical Considerations

Ethical approval for the intervention study was obtained from the Ethics Committee of the Department of Human Communication Sciences in line with University of Sheffield ethics procedures (Appendix 6.1). As discussed above, written consent was received from the parents of children who met participation criteria. Similar to procedures in the longitudinal study, intervention coordinators received verbal assent from all 12 participating children before intervention teaching commenced.

One particular ethical concern in intervention research is the withdrawal of children from mainstream teaching activities. This concern was mitigated firstly due to the short-term and low-intensity nature of the intervention, meaning that children were not removed from their classrooms for more than 30 minutes each week for 10 weeks. Secondly, intervention sessions were timetabled according to school preferences, as such avoiding any essential teaching, in-class assessments, or other events such as sports days.

6.3 Participants

One child who took part in the intervention was not included in the final analysis due to being absent at posttest, and two further children were not included due to receiving only 6 out of 10 sessions (more information on dosage and implementation fidelity is presented in Chapter 7). A

6.4. Timeline and Delivery of Intervention

threshold of 7 out of 10 sessions was chosen to ensure that participants had the opportunity to be exposed to the majority of the intervention teaching material and to avoid assessing children on target words which they had not learned. Thus, the final analysis is based on the nine remaining children (6 female; mean age 8;7). A total of five different languages were spoken between the nine children, including: Punjabi (n=4), Arabic (n=2), Hungarian (n=1), Urdu (n=1) and Thai (n=1). Parental questionnaire data were available for 8 out of the 9 children. The majority of children had been in receipt of English-medium instruction from age 4 or under (n=7), and one child had begun to receive this by age 6.

6.4 Timeline and Delivery of Intervention

The intervention ran in the summer term of the 2015-16 academic year from April to June. Each intervention teaching session was designed to run for 25-30 minutes, once a week for ten weeks. This duration was chosen due to two reasons: firstly, as reviewed in Section 5.3.4, there is evidence to show that vocabulary interventions of short duration (i.e. both in terms of minutes per session and total hours of teaching) are found to yield significant improvements in word knowledge; and secondly, it was considered by the ESCAL team of Sheffield City Council that a relatively short session duration would be more easily accommodated by timetabling constraints of participating schools.

Intervention teaching sessions (Section 6.6) were delivered within school premises, outside of the child's main classroom, on a one-to-one basis by trained coordinators (discussed in Section 6.9). A one-to-one teaching format was chosen for three reasons: firstly, the small sample size of the intervention group allowed pairing between coordinators and children; secondly, there is research indicating that one-to-one working is effective in promoting engagement in learning, with some studies showing particular effectiveness for bilingual learners (Brooks & Thurston, 2010; Ross & Begeny, 2011); and thirdly, one-to-one pairing was considered a feasible strategy since the intervention study represented an efficacy trial, with the aim of assessing children's progress in acquisition of word knowledge in 'extremely favourable conditions' as opposed to such teaching 'in actual use' (O'Donnell, 2008, p.41). More information concerning the pairing of coordinators and intervention participants is available in Section 6.8.

The intervention ran concurrently with the longitudinal study. Due to the lack of a control group, the decision was made to include a 'double pretest' or baseline period with repeated baseline assessments before teaching took place: such a design is advantageous in reducing threats to validity such as maturation and regression to the mean (Shadish, Cook & Campbell, 2002). Given the fairly short time-frame of the intervention study, it was not possible to provide more than two pretests; additionally, two or more pretests are recommended in order to monitor non-linear change prior to intervention (Shadish, Cook & Campbell, 2002), which was deemed to be unlikely for children's knowledge of taught and untaught words. The description and administration procedures of outcome measures, including those administered during the baseline period, is provided in Section 6.9.

6.5 Target Word Selection

As discussed in Chapter 5, a number of word selection strategies have been employed in vocabulary intervention research. The present study sought to select target words considered to be within the Tier-2 category. However, as noted in Chapter 5, other than Tier-2 words being considered “general but sophisticated”, likely to occur in a range of contexts, and representing concepts which children likely already know (e.g. *lucky* and *fortunate*), no objective criteria are available for the selection of Tier-2 vocabulary. For instance, while the word *abstruse* would be considered a member of the Tier-2 category, it is unlikely such a word would be chosen for explicit instruction given its level of familiarity¹. Therefore, in order to make selection of target words more objective, metrics of word characteristics were gathered and compared in order to inform decisions; namely, age of acquisition (AoA) ratings, frequency per million words of text, and part of speech. AoA ratings and frequency metrics were extracted from the database of Kuperman et al. (2012) – see Appendix 6.2 on page 273 for ratings of all 20 target words and 10 untaught words. In addition to Tier-2 criteria (Beck et al. 2002), the following constraints were applied to potential target words: an AoA of between 6 and 10 years; and a frequency of between 5 and 50 occurrences per million words of text. These windows allowed for vocabulary to range in difficulty, and were also intended to prevent the selection of inappropriately challenging words. Words were also selected to vary in part of speech, with an even as possible mixture of verbs, nouns, and adjectives.

Alongside the 20 taught words, a list of 10 words was selected as a control group of untaught words in order to assess the specific effect of the intervention teaching. Thus, in the bespoke word knowledge assessment, children were assessed on their knowledge of a total of 30 words. The short duration of the intervention (10 weeks) and the small pool of participants (n=9) allowed for children’s knowledge of all 20 taught words to be assessed at each of the four testing points (see Figure 6.2 on page 199), in contrast to other vocabulary intervention studies of larger scope and/or longer duration which typically assess a random sample of words covered by teaching (e.g. Jenkins et al. 1989; Proctor et al. 2011; Silverman, 2007). Despite this, it was felt that an equal number of untaught words would have been burdensome for children and may have introduced fatigue effects². For this reason, the decision was made to select only 10 untaught words.

AoA and frequency ratings were also used to provide objective metrics with which to match taught and untaught words. Untaught words were selected using the following method. For each taught word, a thematically unrelated word of the same part of speech and similar ratings for frequency per million words and AoA was chosen as a potential match. This resulted in one untaught word candidate for each of the 20 taught words. A final list of 10 untaught words was chosen by selecting only those untaught word candidates with the closest match to the taught word with which they had been paired. Each taught-untaught word pair did not differ by more than 0.34 years in AoA or by 0.55 in frequency per million words, suggesting a close match between the two words based on objective metrics. However, overall group means indicated larger discrepancies: specifically, the two groups of words appeared similar in mean AoA (taught = 8.78; untaught = 8.73) but different in frequency (taught = 14.18; untaught = 8.54). The difference

¹ *Abstruse* has a frequency rating of 0.06 occurrences per million words of text (Kuperman et al. 2012).

² Based on the observation from the longitudinal study that a number of children found the WISC VC definitions task challenging, and that children in the intervention study were selected precisely because of their low English vocabulary knowledge.

6.6. Structure of Intervention Sessions

in frequency is partly accounted for by the relatively more extreme range of frequency ratings, the relatively smaller number of untaught words, and the presence of two outliers in the taught group, (*responsible* and *afford*, with frequencies of 45.06 and 44.43, respectively). Despite this trend, the two groups differed only marginally significantly in AoA ($t(28) = 1.32, p = .053$) and non-significantly in frequency per million words ($t(28) = -0.11, p = .912$), again suggesting a fair comparison between the two.

6.6 Structure of Intervention Sessions

The aim of the intervention teaching was to improve children's receptive and productive use of target vocabulary using a multimethod approach as advocated by Beck et al. (2002). Given the multidimensionality of vocabulary knowledge, many successful interventions make use of multiple strategies for teaching word meaning (Carlo et al., 2004; Dockrell et al., 2010; Silverman, 2007; Schaefer et al. *under review*) and thus a similar method was applied in the present study. Each session covered only two target words, so as to allow adequate time for completion of activities and to encourage depth of knowledge. Session activities are described in chronological order in Table 6.1 below.

Table 6.1: Intervention Session Structure Outline

Activity	Details	Minutes
Introduction	State aims for session; play word game for warm-up; consolidation from last session	3
Vocabulary Teaching (first word)	Passage reading; comprehension questions; sentence-level activity (completion / judgement); mind-map; sentence writing	10
Vocabulary Teaching (second word)	Passage reading; comprehension questions; sentence-level activity (completion / judgement); mind-map; sentence writing	10
Plenary	Summarise activities and progress made in session; go over both words one more time, giving simplified definition	2

Word games. Learners' motivation as well as their attention to and awareness of words are important factors in vocabulary teaching (Beck et al., 2002; McKeown et al., 1983; Ohanian, 2002). Therefore, five different word games (hangman, semantic category sorting, word association, and odd-one-out tasks) were employed throughout the ten weeks in order to promote children's engagement and awareness of words. Coordinators were instructed to use word games at the beginning of sessions to encourage children's engagement and attention if necessary. Games generally encouraged the child to be aware of words and to think about relationships between them.

Passages. Intervention teaching sought to strike a balance between the definitional and contextual methods described in Section 5.2, given that such a balance has been shown to result not only in gains in vocabulary knowledge in and of itself, but also to transfer to wider skills such

as reading comprehension (Elleman et al., 2009; Stahl & Fairbanks, 1986). Thus, in line with other vocabulary intervention studies with similarly-aged children (e.g. Clarke et al., 2010; Nash & Snowling, 2006; Wilkinson & Houston-Price, 2013), and drawing particularly on the methodology of Clarke et al. (2014), target words in the present study were presented within short written passages depicting stories of characters and events to illustrate use of the target word within context, for example a narrative of two friends who make a trip to a museum but are required to take public transport because it is located in a *distant* town (see Appendix 6.3 on page 275).

Additionally, as discussed in Chapter 5, contextual environments vary in the amount of information they afford to learners concerning unfamiliar vocabulary (e.g. Beck et al.'s 2002 taxonomy ranging from *misdirective* to *directive*; Section 5.3.2). Thus, passages were written in order to provide cues as to words' meanings and to encourage children to arrive at the correct interpretation, for example: "Thorpetown is quite distant, so it's a long way to walk". Intervention coordinators were instructed firstly to read through the passage with the child and ask if he/she had any ideas as to the word's meaning. If children were not able to infer the correct meaning of the target word, coordinators were instructed to provide a basic explanation by making recourse to the text to identify relevant cues. The child was then asked two comprehension questions regarding the target word which encouraged explanation or justification of reasoning, for instance, *How do you know that Thorpetown was distant?* (e.g. because Jake says 'It's a long way to walk'), and *How do you know that the dinosaur skeleton was from the distant past?* (e.g. reference to '65 million years old' and the synonym *ancient*). Written passages ranged in length from 83 to 193 words, with an average of 13.4 words per sentence. Difficulty ratings of bespoke written passages were obtained, including Flesch Reading Ease (where a score of 100 corresponds to the easiest and 0 to the most difficult), and Flesch-Kincaid grade level. Flesch Reading Ease ranged from 58.8 to 95.5 (mean=77), while Flesch-Kincaid grade level ranged from 2.7 to 9.8 (mean=5.8, corresponding to Year 6, or ages 10-11).

Sentence-level tasks. Sentence-level activities have been used in intervention studies to encourage children to consolidate their understanding of word meaning and actively engage in learning materials (e.g. Jenkins et al., 1989). Focus on productive aspects of word knowledge also accords with Nation's (2001) multidimensional conception of vocabulary in which knowledge of individual words is distinguished by form, meaning, and use, with each aspect possessing receptive and productive components. Under such a view, it follows that instruction in the *use* of novel vocabulary will likely contribute to more highly specified word knowledge. In the present study, children completed two types of sentence-level task in order to provide variation in learning activities. Firstly in sentence *judgement* tasks, the child was presented with two short sentences containing the target word and was asked to decide if the sentence made correct or incorrect use of the word. For example, for the target word *distant*: 'London and Liverpool are *distant* from each other, so it's quick to travel between them' (Incorrect; it takes a long time because they are far/distant from one another). Secondly in sentence *completion* tasks, the child was presented with a cloze-like activity consisting of a partially completed sentence, and asked to complete it according to their understanding of the word's meaning. For example: 'I'm going to the capital to ...' (e.g. see a landmark typical of a capital such as Big Ben; take a plane to go on holiday, etc.). The use of each type of sentence-level task alternated between target words such that a different task was used for each word (e.g. in Week 1, sentence judgement was used with target word

6.7. Pilot Study

distant, while sentence completion was used with target word *capital*). Coordinators encouraged children to justify their reasoning and provide suggestions if children were unsure.³

Mind Maps. Again, similar to other vocabulary intervention studies which have been shown to improve word knowledge in monolingual learners (McKeown et al., 1983; Nash & Snowling, 2006; Wilkinson & Houston-Price, 2013), children were prompted to consider word associations, idioms, synonyms, antonyms, related concepts, and personal experiences related to target words in a mind-map activity. The target word was presented on a page with blank spaces, allowing for a high degree of freedom in creativity and expression (one aspect which changed after the pilot study; see the following section). After completion of the mind map activity, the child was asked to write a sentence using the target word. Coordinators were trained to prompt children to extend their sentences where possible, for example by adding more information with subordinate clauses.

Flashcards and definitions. At the end of each session, coordinators provided short 'child-friendly' definitions (Section 5.3.5) of the target words in an attempt to consolidate children's learning. Such definitions have been shown to result in greater improvements in children's word learning (e.g. McKeown, 1993) and were considered appropriate for use with EAL learners in the present study. Additionally, visual aids are reported in some vocabulary intervention studies (Clarke et al., 2014; Hairrell et al., 2011) and are recommended for active processing of word meaning (Rupley & Nichols, 2005). Thus, at the end of each session children were shown flashcards containing the target word next to a colour photograph or illustration conveying the appropriate concept. For instance, the flashcard for *distant* contained a point-of-view photograph of a road that stretches out into the distance.

A cyclical structure to the intervention was implemented with the introduction of a recap activity at the beginning of each session, which provided children an opportunity to revisit and refresh their knowledge of words covered in the preceding week.

6.7 Pilot Study

Prior to the intervention teaching, a pilot study was conducted in order to assess the feasibility of the proposed timings and amount of material to be covered in teaching sessions. Due to a limited sample size and difficulty in recruiting participants for the pilot study, criteria were slightly less restrictive relative to those in the main intervention; namely, a score of ≤ -1 SD on any of BPVS, CELF EV, or WISC VC at t1 of the longitudinal study. It is a limitation of the study that only one child was successfully recruited to take part in the pilot. This child (age 8;9; male) obtained a standard score of 83 on the BPVS, but scores of 10 and 12 on WISC VC and CELF EV, respectively.

The pilot study took place in March 2016. This child participated in two 30-minute sessions delivered by the researcher using bespoke intervention materials and activities described above.

³In terms of the task-induced involvement load hypothesis (Laufer & Hulstijn, 2001; Section 5.3.3), sentence-level tasks represented moderate *need* and no *search* (as target words and concepts were provided); sentence judgement required comparison between sets of pre-existing sentences and thus induced moderate *evaluation*, while sentence completion involved the combination of sentence stems with target vocabulary, thus inducing strong *evaluation*. Sentence-writing in the mind-map activity also induced strong *evaluation*.

As a result of the pilot study, a decision was made to incorporate a greater variety of tasks, to introduce a warm-up activity at the beginning (word game), and to provide more opportunities for active engagement with new vocabulary, as reflected in the session structure and activities described above in Section 6.6. This was due to a low level of engagement and interest on behalf of the child who took part in the intervention, and also due to the observation that this child had few opportunities to contribute actively, especially when correct answers were given. Specifically, the mind-map activity was made more open-ended in nature by providing a blank map to encourage the child to fill in synonyms, related concepts, personal experiences, and so on, as opposed to presenting a pre-designed, filled-in mind map as was originally intended for use during intervention teaching.

6.8 Recruitment and Training of Student Coordinators

Students in the Human Communication Sciences department at the University of Sheffield were approached as potential intervention coordinators. The decision not to recruit school staff was due to the nature of the intervention study as an efficacy trial (i.e. assessing the outcome of the intervention in as 'ideal' circumstances as possible; O'Donnell, 2008), as well as potential difficulties in the logistics of recruiting and training school staff.

Students enrolled on the following courses were specifically approached: Speech and Language Sciences (BSc), Speech and Language Therapy (BMedSci and MMedSci, both leading to accreditation by the Royal College of Speech and Language Therapists), and Language and Literacy (MSc). Applicants were asked to fill out an online form to provide details of relevant teaching or clinical experience as well as to confirm availability for training. Thus, it was considered that students enrolled in higher education courses containing elements of theory and practice related to language therapy and assessment, often in a clinical setting, would represent good candidates to deliver the intervention teaching.

A recruitment email was distributed to students enrolled on the courses listed above (only 2nd, 3rd, and final-year students on BSc and BMedSci courses were contacted in order to ensure that students would have had at least one year of training and experience). The recruitment email provided information about the project and a person specification detailing the skills and experience required. Specifically, essential criteria included enrolment on one of the aforementioned university degree courses, experience of working with children, excellent communication skills, and awareness of child protection/safeguarding issues. Desirable criteria further specified experience of working in schools, working with bilingual children, working with children in the 8 to 9 year-old range, and demonstrable teaching experience.

Applications were received from a total of 20 students. Three students were unable to demonstrate any interest or experience in working with bilingual children; four students subsequently declined to take part due to other commitments, with a further four students dropping out after having been selected to take part, and one student applying after the deadline and being put on a reserve list. This resulted in a final pool of 9 students who carried out the intervention teaching. In line with criteria listed above, coordinators were studying towards degrees in speech and language therapy or speech science (BMedSci, n=4; MMedSci, n=4; and BSc, n=1), were able to demonstrate experience working with children in either a U.K. primary school context or as a

6.9. Measures and Analytical Strategy of the Intervention Study

teacher of English as a foreign language, and had an interest in bilingualism by virtue of having studied the subject, being bilingual, or having worked in settings with bilingual pupils. It should be noted that the researcher also carried out the full schedule of teaching with one of the children taking part in the intervention in order to make up a short-fall in coordinators due to drop-outs. Another reason for this decision was for the researcher to be able to compare observations with other coordinators and to gain perspective of the full delivery of the 10-week bespoke teaching programme.

Coordinators attended mandatory training to become familiar with the details and aims of the project and to gain hands-on experience with the intervention materials. Coordinators received 1 hour 45 minutes of training, the first part of which covered the background to the project, research involving children with EAL in England, effective vocabulary teaching practices, teaching materials, and logistics (i.e. organising visits to schools). Additionally, coordinators were invited to attend a feedback session in order to share experiences and raise any queries or concerns. This session was intended to be run half way through the intervention, although constraints on availability meant that it was postponed to the half term holiday at the end of May after the eighth week of teaching.

Due to the small number of children participating in the intervention ($n=12$), coordinators were paired with individual children with whom they would work throughout the ten weeks of the teaching (see Section 6.4 for a justification of one-to-one working). Seven coordinators were paired with one child each, while two coordinators were paired with two children each (four children in the same school). This decision was made due to logistical reasons (i.e. coordinators' availability of personal transport and the relatively distant location of the school in relation to the university which would have made repeat visits by public transport difficult). The researcher was paired with the final, twelfth child.

6.9 Measures and Analytical Strategy of the Intervention Study

This section will introduce the measures used in the intervention study, including the bespoke word knowledge assessment and the two standardised assessments from the main test battery of the longitudinal study. Timing and administration of measures will be outlined in Section 6.9.3. Finally, the study's multiple case series design will be introduced as an analytical strategy in Section 6.9.4.

6.9.1 Primary Outcome Measure of Word Learning

As discussed in Section 2.1.2, measures of vocabulary depth often invite examinees to provide a verbal definition of a word, which is then scored against some set of criteria (e.g. the WISC-IV Vocabulary subtest utilised in the longitudinal study). Performance on such definition tasks is said to require certain metalinguistic skills and is shown to be challenging for young children (Benelli et al., 2006; Snow et al., 1991). This has led other studies to adopt a different approach by acknowledging and awarding points for various aspects of word knowledge that young children provide, such as background knowledge, context, and gesture (e.g. Hadley et al., 2016; Vermeer,

2001). In addition, studies utilising such bespoke word knowledge assessments are found to report larger gains in word knowledge (Biemiller, 2005; Elleman et al., 2009; Proctor et al., 2011).

Word knowledge may also be evidenced through understanding of grammatical function, collocation, and other constraints of use (Nation, 2001; Treffers-Daller & Rogers, 2014). For this reason, the bespoke word knowledge task in the present study contained not only a verbal definitions task, but also a sentence production task in order to allow examination of children's productive knowledge and syntactic accuracy of target words (both tasks described below). A similar approach was utilised in a kindergarten vocabulary intervention study by Coyne et al. (2010) in which children were asked about word meanings in context, for example 'What would you be doing if you were *halting*?' Although such a context is relatively neutral, it would appear that such a question stem provides a cue to the fact that this word is a verb, for example through use of present progressive morphology *-ing*. Instead, children in the present study were simply asked to use each word within a sentence. Additionally, one other limitation of sentence production tasks is that a child may produce a sentence which, albeit not incorrect, does not reveal a word's characteristic features. One approach is to distinguish between sentence types by awarding more points to sentences that do reveal such characteristic features, while not entirely penalising more 'generic' sentences⁴ (McKeown, 1993). This approach was adopted in the present study; see below for details of the scoring procedure employed.

A bespoke word knowledge assessment was designed to capture children's baseline knowledge and growth in lexical and conceptual knowledge of the 20 taught and 10 untaught words. During administration of the assessment, all children were presented with a list of the 30 words and asked firstly to provide a verbal definition (deriving a Word Score; a maximum of 8.5 points for each word) and secondly to use each word in a sentence (deriving a Sentence Score; a maximum of 5 points per word). A summary of scoring criteria for definitions and sentences is presented below. Verbal definitions were awarded points across 4 categories (target words underlined):

- Definitional information: 1 point awarded for partial or underspecified definition (e.g. *to rescue someone is to help them*), and 2 points awarded for a full or more highly abstract definition (e.g. *to rescue someone is to save them from a dangerous situation*);
- Background knowledge, up to 4 points per word, split into three categories:
 - Situational information: up to 1 point (e.g. a hypothetical or real situation demonstrating understanding of the target, e.g. *if your friend said bad things about you, you would feel miserable*);
 - Contextually-related concepts or referents: up to 1 point (e.g. *sky* or *cockpit* for *pilot*);
 - Attributes or functions: e.g. *you have a red face when you're furious*; 1 point for one attribute or function, and 2 points for two or more attributes or functions.
- Lexical knowledge: up to 1 point in each of two subcategories, including:
 - Relevant synonyms such as *excited* for *thrilled*, or antonyms such as *agree* is the opposite of *disagree*;

⁴For example, a sentence such as *I went to the coast on Saturday* does not reveal any understanding about the target word *coast* other than the fact that it functions as a noun and that it refers to a location. In contrast, *I went to the coast and made a sandcastle* shows a deeper understanding of *coast* by virtue of associating it in a particular context.

6.9. Measures and Analytical Strategy of the Intervention Study

- Related words and phrases, e.g. morphologically related words such as *distance* for *distant*, or collocational knowledge such as *rescue attempt* or *the coast is clear*);
- Non-verbal responses: up to 0.5 points (e.g. pointing to a far object to illustrate *distant*, or showing an angry face for *furiously*).

Children's sentences were scored according to 3 categories, including:

- Syntax: up to 1 point for using target word as correct part of speech (e.g. *contagious* used appropriately as an adjective). Any syntactic errors not related to the target word were not taken into consideration (as opposed to scoring criteria for CELF Formulated Sentences (FS); see discussion in Section 4.5.2.1), so as to provide a purer indication of changes in syntactic knowledge as a result of intervention teaching;
- Morphology: up to 1 point for lack of any morphological error on target word. Note also that any sentence scoring 0 for syntax automatically received a score of 0 for morphology (a child who used a target word as the incorrect part of speech (e.g. *I will maximum the washing up liquid*) could not be awarded points for correct morphology on that word;
- Semantics: up to 3 points based on the extent to which the sentence indicates understanding of word meaning (also see McKeown, 1993): 1 point for a very simple sentence such as *I was miserable* in which the target word could be easily substituted; 2 points for a more explicit use of the target word for example with additional information, such as *I was miserable because I was cold*; 3 points for a well-specified sentence giving reason, context, an example, or synonyms, e.g. *I was miserable and felt like crying because my friend moved away*.

Test administration included two practice items, *library* and *remember* which children were asked to define and use in a sentence, using the prompts *What does ... mean?* and *Can you use ... in a sentence?* The purpose of these items was to familiarise children with the test format and to indicate to them what kinds of responses were permissible. On the two practice items only, children were prompted for the various types of knowledge they may possess of words, including not only dictionary-style definitions, but background knowledge about context and situation (e.g. *What is it like in a library?*), and lexical knowledge (e.g. *What is the opposite of remember?*).

6.9.1.1 Inter-rater Reliability

All baseline responses from the bespoke word knowledge assessment were independently scored by another doctoral student in the department of Human Communication Sciences. Training was given for the scoring of word and sentence scores, summarised in two training documents (Appendix 6.4). Cohen's kappa (κ) was calculated separately for all categories of word and sentence scores in SPSS 22, presented in Table 6.2 below. All κ -values were statistically significant at the 0.01 level, and the majority of values fell within the 'substantial' to 'excellent' range of agreement, with the exception of lexical scores for words which represented 'moderate' or 'fair' agreement (Cicchetti, 1994; Landis & Koch, 1977). Although there were disagreements among the various subcategories of the scoring rubric, disagreements in the total score (i.e. the final score calculated as the sum of scores from each subcategory) were much lower, at 13.3% for words and 19.2%

for sentences. Disagreements were discussed and changes were made where appropriate. The relatively higher rate of disagreements in the lexical category appeared to be accounted for by a small group of words including *donation*, *miserable* and *thrilled*: definitions of these words often contain synonyms (e.g. ‘a little bit sad’ for *miserable*). During formulation of the scoring rubric, such synonyms or related words were considered to be core to the definition of these words, and therefore were intended to be scored within the ‘definition’ category. This accounted for 14 out of 18 disagreements in the lexical category, and original scoring was maintained for cases in which synonyms were felt to be central to the definition of target words (i.e. points for mention of such synonyms were scored under the definition category). Despite this, children’s word and sentences scores on the bespoke word knowledge assessment indicated an overall high degree of inter-rater agreement.

Table 6.2: Inter-Rater Reliability (Cohen’s κ) for Bespoke Word Knowledge Assessment

	Definition	Background	Lexical	Total
Words	0.78	0.83	0.45	0.79
	Syntax	Morphology	Semantics	Total
Sentences	0.90	0.90	0.72	0.71

Note: all figures represent Cohen’s κ . All $p \leq .01$. Extent of inter-rater agreement according to Landis & Koch (1977): .81 - 1.00 = *almost perfect*; .61 - .80 = *substantial*; .41 - .60 = *moderate*.

6.9.2 Transfer Measures

The meta-analytical study by Elleman et al. (2009) discussed in Chapter 5 found robust evidence for transfer of vocabulary teaching onto both experimental ($d = 0.79$) and standardised ($d = 0.29$) measures of vocabulary. Therefore, in order to assess the possibility of transfer of intervention vocabulary teaching onto standardised measures in the present study, the CELF FS and WISC VC subtests were administered additionally at baseline. This allowed monitoring of intervention participants’ progress on these measures between t1, baseline, posttest (t2), and t3 (see Figure 6.2 overleaf). These measures were chosen due to their similarity with the bespoke word knowledge assessment (i.e. use of target words within a sentence in CELF FS and provision of verbal definitions of target words in WISC VC). Importantly, stimuli in these measures did not overlap with taught or untaught words of the intervention. Extent of transfer from vocabulary teaching to standardised measures is addressed by research question four in Section 7.2.4.

6.9.3 Timing and Administration of Measures

The bespoke word knowledge assessment was administered at four points across the study as depicted in Figure 6.2 overleaf: twice prior to the start of teaching, at baseline in mid-March 2016 and pretest 5.6 weeks later in late April 2016, and twice following the end of the 10-week teaching period at posttest in early July 2016 and maintenance test six months later in early to mid-January 2017. As described above, standardised measures were also administered with the bespoke word knowledge assessment at baseline and among the main test battery at posttest.

6.9. Measures and Analytical Strategy of the Intervention Study

At each of the four testing points, all responses were audio-recorded and transcribed for scoring. The bespoke word knowledge assessment was carried out by the researcher at all time points across the intervention with the exception of pretest, when intervention coordinators administered it as part of the first session. This decision was taken due to logistical constraints of visiting all participants across the five participating schools within a short time-frame and in the interests of measuring children's word knowledge as close as possible prior to the beginning of teaching.

Year	Month	Longitudinal Study	Intervention Study			
2015	Sept	<p>t1 (main test battery inc. WISC VC & CELF FS)</p>				
	Oct					
	Nov					
2016	Dec		<p>t2 (main test battery inc. WISC VC & CELF FS)</p>			
	Jan					
	Feb					
2017	Mar			<p>t3 (main test battery inc. WISC VC & CELF FS)</p>	<p>Baseline (BWK, WISC VC, CELF FS)</p>	
	Apr				<p>Pretest (BWK)</p>	
	May				<p>Intervention (10 weeks)</p>	
2017	Jun				<p>t3 (main test battery inc. WISC VC & CELF FS)</p>	<p>Posttest (BWK)</p>
	Jul					
	Aug					
Sept						
2017		Oct	<p>t3 (main test battery inc. WISC VC & CELF FS)</p>			
Nov						
Dec						
2017	Jan	<p>t3 (main test battery inc. WISC VC & CELF FS)</p>		<p>Maintenance (BWK)</p>		
	Feb					
	2017					Feb

Figure 6.2: Intervention Study Timeline in Relation to Longitudinal Study. Note: BWK = Bespoke Word Knowledge Assessment; WISC VC = Wechsler Intelligence Scale for Children IV Vocabulary subtest; CELF FS = Clinical Evaluation of Language Fundamental IV Formulated Sentences subtest. See Chapter 3 for description of main test battery.

6.9.4 Multiple Case Series Design Analytical Strategy

The small sample size of the intervention ($n=9$) placed statistical limitations on the analysis of group trajectories between the four time points of the study. Despite the availability of nonparametric procedures for this purpose (which were also utilised; see Chapter 7), the examination of individual children's trajectories allowed for a more fine-grained analysis of progress over time, and indeed this strategy was made all the more feasible by the small sample size itself. Therefore, a multiple case series design was employed for the analysis of children's individual growth

trajectories on the bespoke word knowledge assessment, both in terms of receptive and productive knowledge (word and sentence score, respectively). Multiple case series designs are advantageous in allowing in-depth examination of variation between and within individuals, where examination of group-level effects (i.e. averaging across individuals) may be misleading or may hide interesting patterns (Towgood, Meuwese, Gilbert, Turner & Burgess, 2009; Chmiliar, 2012). However, the strategy employed here does not completely forego reference to an overall or averaged group trajectory, which may provide an interesting comparison with individual trajectories. Both group and individual trajectories will be discussed in the results section of Chapter 7.

6.10 Implementation Fidelity

In order to obtain an accurate and fair estimate of the effect of an intervention, it is necessary to monitor the fidelity with which it is delivered and received. Models of implementation fidelity typically account for treatment delivery, receipt, and adherence (Dane & Schneider, 1998; Shadish et al., 2002). Fidelity of implementation in the present study was measured in three ways. Firstly, all intervention coordinators received mandatory training prior to the start of the programme in order to ensure clarity of aims and teaching methods. Secondly, coordinators were required to keep written records for each of the ten weeks (Appendix 7.1). Specifically, they were asked to record dates, start and end times of sessions, the extent to which each activity was completed (zero; partial; full), children's levels of engagement and attention, and also had the opportunity to write comments (reported in Chapter 7). Thirdly, observations of coordinators during teaching sessions were used as an indication of implementation fidelity. A brief observation checklist was devised which sought to assess not only quality and fidelity of teaching according to training, but also characteristics of the working environment and rapport between coordinators and participants (see Appendix 7.2 on page 282).

Chapter 7

Results and Discussion II: Vocabulary Intervention Study

The results of the longitudinal study indicated a trend for children learning EAL to possess lower levels of receptive and expressive English vocabulary knowledge than their monolingual peers. This pattern of performance was maintained throughout the study due to the similar trajectories of the two groups. The vocabulary intervention study, therefore, was an attempt to augment the vocabulary knowledge trajectories of children learning EAL on a specific subset of target words that did not appear in standardised assessments used in the longitudinal study. There is promising research suggesting that explicit, targeted vocabulary instruction can be effective for bilingual learners, yet this work has typically been carried out with much younger children and often not in U.K.-based contexts (Murphy & Unthiah, 2015).

This chapter will begin by restating the research questions introduced in Chapter 5, before detailing fidelity of implementation, results of children's progress in receptive and productive knowledge of taught and untaught words, extent of transfer to standardised assessments, and finally considering results with reference to relevant literature. The research questions of the intervention study were as follows:

1. To what extent does a short, low-intensity, one-to-one explicit vocabulary training programme improve the vocabulary knowledge of EAL learners who are identified as having English vocabulary weaknesses?
2. What effect does this teaching have upon (a) children's receptive understanding of taught vocabulary and (b) their ability to use it productively?
3. How do children's individual growth trajectories inform conclusions about the efficacy and adequacy of the teaching programme?
4. Does short, low-intensity, one-to-one vocabulary teaching result in transfer onto non-explicitly taught skills such as expressive grammar and depth of vocabulary knowledge?

7.1 Fidelity of Implementation

This section will consider fidelity in terms of dosage, attention and engagement, and analysis of completed mind-map activities. Both quantitative and qualitative sources of information were utilised in order to evaluate implementation fidelity and to gain a richer sense of children's experiences during the intervention teaching (Yoshikawa, Weisner, Kalil & Way, 2008). As such, reference will be made not only to descriptive statistics concerning implementation fidelity but also

to written comments by coordinators, examples of completed activities, and coordinator observations. Primary implementation fidelity statistics taken from coordinators' written records (Section 6.10) are presented in Table 7.1 below.

Table 7.1: Fidelity of Implementation of Intervention Teaching

Dosage	Mean (SD)	Min-Max
Total sessions completed (max=10)	9.44 (0.73)	8-10
Total teaching received (minutes)*	225.11 (30.89)	165-259
Session duration (minutes)	26.65 (6.06)	10-45
Mean activity completion rate (max=2)	1.89 (0.12)	1.63-2
Passage reading	1.99 (0.08)	1-2
Comprehension questions	1.99 (0.08)	1-2
Mind map	1.86 (0.42)	0-2
Sentence writing	1.79 (0.52)	0-2
Sentence judgement/completion	1.93 (0.35)	0-2
Attention and Engagement		
Mean level of engagement (max=5)	3.89 (0.85)	1-5
Mean level of attention (max=5)	3.82 (1.02)	1-5

* maximum possible amount of intervention teaching was 250-300 minutes

7.1.1 Dosage

Number and timing of sessions. Over the ten weeks, children received an average 225.11 minutes of one-to-one instruction, equating to an average of 3.75 hours per child. The mean number of sessions completed was high, at 9.44 out of 10. A small number of sessions did not run as planned due to either child absence or university examination schedules of coordinators. Attempts were made to reschedule any sessions that were missed¹. The average session duration across the 10 weeks was within the expected range of 25 to 30 minutes, although there was a considerable amount of variation (see Table 7.1). For instance, some sessions had to be stopped prematurely due to events on schools' schedules such as assemblies and sporting activities, and the average duration of week 1 sessions was considerably longer than subsequent weeks (mean = 37.4 minutes) due to the administration of the pretest during this session. In the following nine weeks of the programme, session duration did not drop below an average of 23 minutes, suggesting that intervention delivery closely approximated the minimum intended session length of 25 minutes.

Activity completion rates. Coordinators were asked to rate the completion of each activity on a scale of 0 (not completed), 1 (partially completed) to 2 (fully completed; see Appendix 7.1 on page 280). The overall average completion rate across the ten weeks was high, at 1.89 out of a possible score of 2. There was some variation according to activity type, with passage read-

¹As noted in Chapter 6, two children were excluded from analyses due to receiving fewer than 7 sessions. Of the remaining nine children, the breakdown of total number of sessions was as follows: 8 sessions (1 child); 9 sessions (3 children); 10 sessions (5 children).

7.1. Fidelity of Implementation

ing and comprehension questions being the most consistently highly completed activities, and sentence writing being the least highly completed activity. Indeed, during training, coordinators were instructed to focus on passage reading and comprehension questions in the event that they would not have sufficient time to complete all activities. The reason for this decision was to allow children an opportunity to engage with word meanings in context and engage actively with novel word knowledge by discussing and explaining their answers to the comprehension questions with coordinators.

7.1.2 Attention and Engagement

Coordinators were asked to give indications of children's level of attention (defined as making eye contact with the coordinator and the learning materials and not being distracted), and engagement, (defined as willingness to provide answers to questions and make discussion around the topic). A likert scale was provided for both ratings, ranging from 1 (very little attention or engagement) to 5 (very high attention or engagement). Ratings of attention and engagement remained moderately high on average across the ten-week programme, and were relatively consistent on a week-to-week basis (see Figure 7.1 below). Although there was a subtle trend for ratings of attention and engagement to decrease across the ten weeks, average ratings did not drop below 3.5, suggesting that children were continuing to engage with learning materials and activities.

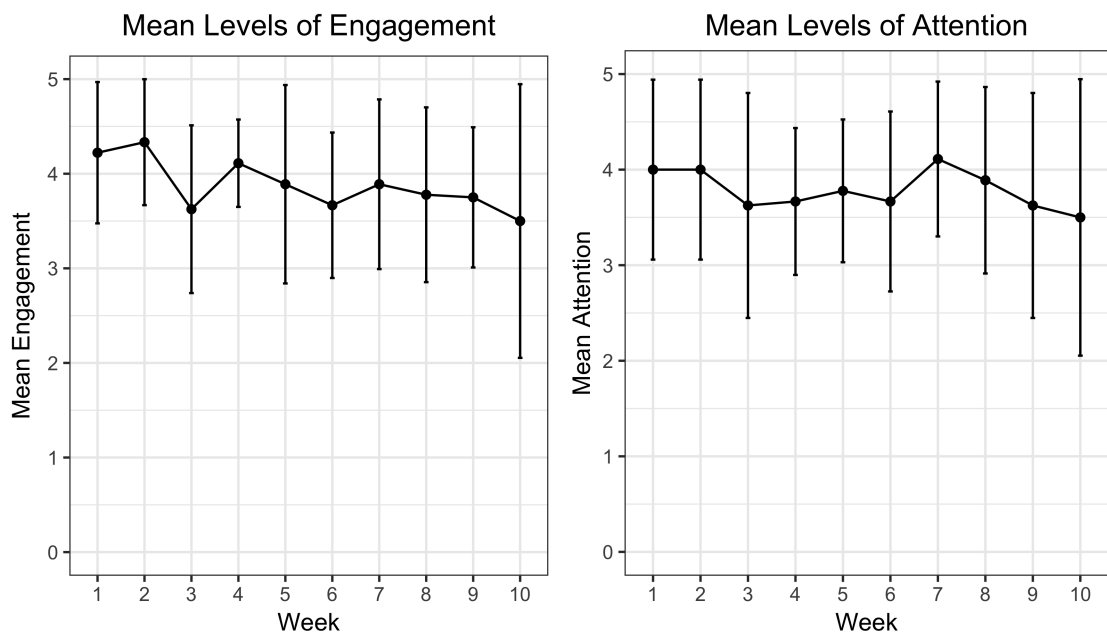


Figure 7.1: Coordinators' Mean Ratings of Children's Engagement and Attention Throughout the Intervention. Vertical bars represent 95% confidence intervals.

As well as likert ratings, coordinators also had the opportunity to provide qualitative information about attention and engagement through written comments. Firstly, it was noted that working environments were often noisy and some sessions were interrupted by other activities going on within schools (see also Schaefer et al., *under review* for a similar observation). Some children were particularly reticent to engage in the content of the session, or were highly distracted, in

the presence of other people. However, some dyads were able to relocate to less disruptive locations when space became available. Secondly, some children required substantial input and prompting from coordinators and many had difficulty in justifying or explaining the answers they provided to comprehension questions (often reverting to responses such as “because it says in the text”). The written passages appeared to be problematic for some children due to their length, and in some cases, the vocabulary they contained was perceived to be difficult². Some coordinators noted a fall in engagement during the second half of the session due to the repetition of format. Thirdly, some concerns were raised regarding sentence-level activities: one coordinator suggested that the sentence completion cloze task was too ambiguous (although it should be noted that this activity was designed to encourage expression and was therefore not designed to be overly restrictive; i.e. coordinators were not provided with lists of ‘acceptable’ answers); and during sentence judgement activities one child misunderstood instructions to identify the ‘good’ or ‘bad’ sentence (e.g. ‘when I’m ill, I feel *miserable*’ was judged to be ‘bad’ because being ill is not a good or desirable state, rather than as being correctly identified as ‘good’ as this would represent *correct usage* of the target word *miserable*).

Despite these issues, coordinators noted that children enjoyed discussing their personal experiences when they were able to relate to the words or stories, and there were some clear examples of deep engagement with word meaning: one child, in particular, commented during discussion of the word *coast* that rivers and ponds were also coasts, prompting discussion of their differences; this child also argued that to *disagree* is not necessarily to dislike. Finally, it was noted that children liked the opportunity to write and draw in the mind-map activity, and coordinators suggested that pictures and flashcards helped to disambiguate word meanings.

7.1.3 Mind Map Activities

In the mind-map activity, coordinators were instructed to discuss lexical knowledge, background knowledge, personal experiences, and associated emotions (Section 6.6). Analysis of returned mind-maps following the end of the programme indicated that coordinator-child pairs consistently discussed multiple aspects of words’ meanings, examples of which are provided in Table 7.2 and Appendix 7.3 on page 283.

7.1.4 Coordinator Observations and Feedback Session

Using the checklist in Appendix 7.2, coordinators were observed during teaching sessions in order to ensure adherence to teaching methods, timing, and utilisation of materials. Unfortunately, due to timing constraints and child absences, observations were carried out for only two coordinators. However, these observations were encouraging in that both ran the intended length, and coordinators made consistent effort to promote depth of word learning through strategies such as encouraging, recasting, and asking children to explain their answers (e.g. ‘Yes, but why was she *thrilled*? Does it say here?’). For example, in the first coordinator-child dyad, in the mind map for *cautious*, one child gave an example of having hurt her foot – the coordinator added to his and

²It should be noted that passages were written to appeal to children with differing levels of ability, and ultimately, passages were used merely as vehicles through which to provide context around unfamiliar words aligning with the recommendations in Beck et al. (2002).

7.1. Fidelity of Implementation

Table 7.2: Examples from Mind-Map Activities among the 9 Intervention Participants

Category	Examples from Mind-Maps
Background	'don't agree with something like bullying' (for <i>disagree</i>); 'someone spoils a surprise' (for <i>disaster</i>); 'lonely', 'crying', 'not allowed chocolate' (for <i>miserable</i>)
Lexical	Synonyms: 'seaside', 'beach' (for <i>coast</i>); 'angry', 'mad', 'livid' (for <i>furios</i>) Antonyms: 'village', 'small town' (for <i>capital</i>); 'not genuine', 'not real' (for <i>fraud</i>) Morphologically related: 'coastal' (for <i>coast</i>); 'fury', 'furiously' (for <i>furios</i>)
Personal	'In Pakistan we got lost - we had to navigate to nan's' (for <i>navigate</i>); 'I went to the seaside with my little brother' (for <i>coast</i>)

said 'you might have *limped* to be cautious', and 'you were cautious to not do it again'. Another example scenario included crossing the road, in which the coordinator prompted the child to justify why this might warrant caution, for example 'Why? What do you have to do to be cautious?' Additionally, the verb *to caution* was also discussed, e.g. 'I might caution you; I might warn you'.

During discussion of the word *thrilled* in the second coordinator-child dyad, the coordinator encouraged nuance of meaning when the child offered the synonym *happy*, agreeing with this, but also stating that 'it means much more than happy'. In the mind-map activity, the coordinator introduced the phrase *over the moon* as a related phrase, which was unfamiliar to the child. As this child was particularly reticent to offer examples, the coordinator asked him to think about times when he might feel *thrilled*, suggesting 'What about your birthday? Would you feel thrilled then?' Thus, interactions in the two observed sessions indicated a high level of fidelity according to the training coordinators received.

A feedback session was held as an opportunity for coordinators to discuss their observations and any issues they were experiencing. Three coordinators attended. One general theme from coordinator feedback was that children generally enjoyed participating in teaching sessions, but that engagement tended to drop slightly during the second half of the session in which the second word was introduced. Some specific issues included one child being distracted by other words in passages not relevant to the target word, and one child experiencing difficulty in providing synonyms in the mind-map activity due to a low level of vocabulary knowledge. Despite these issues, however, coordinators reported that passage comprehension questions and sentence-level activities helped to tease out children's misunderstandings (e.g. one coordinator reported that sentence-completion was an effective method of determining whether the child had understood the target word, with this child sometimes using the word as the wrong part of speech). Child-friendly definitions and flashcards were also suggested to be effective in steering children towards the correct meaning of words.

7.1.5 Summary of Implementation Fidelity

Written records of completed activities indicated that the core of the intervention teaching was delivered to a high degree of completion despite disruptive working environments and changes to scheduling. For the most part, the children were engaged and attentive, and mind-maps and

observations show clear evidence of coordinators encouraging growth in children's depth of word knowledge. While some children's difficulty with written passages and repetition in session structure may have contributed to a trend for lower levels of attention and engagement over time, coordinators reported that children generally enjoyed taking part, and discussing their personal experiences in particular.

7.2 Results of the Vocabulary Intervention Study

The results section will follow the structure of the research questions presented above. Firstly, with reference to research question 1, results will examine the extent to which the intervention teaching was effective in promoting acquisition of word knowledge, determined through analysis of children's growth in terms of *total* score (i.e. the sum of *word* and *sentence* scores) in taught and untaught words. Secondly, with reference to research question 2, results will examine children's relative growth in receptive and productive knowledge of taught words, as assessed by the separate *word* and *sentence* score components of the bespoke word knowledge test³. Thirdly, with reference to research question 3, a multiple case series design will be used to examine the individual trajectories of the nine children in order to explore patterns of progress across the study. In particular, the cases of two children who failed to benefit from the intervention teaching (BA and JG) will be discussed in further detail. Trajectories for groups and individual children are presented graphically alongside one another in reference to each research question. Finally, with reference to research question 4, analyses will be presented relating to the extent of transfer of learning on to standardised assessments between baseline and posttest.

As a general analytical strategy, research questions 1 to 3 will examine children's performance across all four time points in an effort to determine (a) progress in acquisition of word knowledge in the baseline period (baseline to pretest) before the onset of explicit intervention teaching; (b) the direction and magnitude of changes in word knowledge as a result of the intervention teaching (pretest to posttest) and; (c) the extent to which children retained this knowledge six months later (posttest to maintenance).

Given the small sample size ($n=9$) and therefore lack of power to determine normality, the decision was made to use non-parametric statistics, including the use of medians and interquartile range as measures of central tendency and dispersion, respectively (Field, 2012). Non-parametric repeated measures Wilcoxon signed-rank tests were conducted to analyse the statistical significance and magnitude of changes in children's scores between each pair of time points. Interpretation of effect sizes between each time point follows that of Cohen (1988), whereby 0.2 is considered small, 0.5 is medium, and 0.8 is large⁴. As a result of the small sample size and use of non-parametric procedure, caution should be taken in the interpretation of results presented here. For reference, parametric descriptive (means and standard deviations) and inferential statistics (repeated measures t-tests) are presented in Appendix 7.4 on page 284. In this chapter, results are presented for taught words in Table 7.3 and untaught words in Table 7.4. Additionally, both

³Due to space limitations and specific interest in taught words, separate receptive and productive trajectories for untaught words will not be discussed.

⁴The effect size r was calculated from output of Wilcoxon signed-rank tests: $r = \frac{Z}{\sqrt{N}}$ where N is the total number of observations across both points (i.e. 9 observations \times 2 = 18).

7.2. Results of the Vocabulary Intervention Study

group and individual trajectories for taught and untaught words are presented graphically in Figure 7.2. In order to give an indication of overall progress, results are presented in terms of total scores which represent the sum of *word* and *sentence* scores from the bespoke word knowledge assessment. Group and individual trajectories will be discussed in text, and are presented side by side in Figure 7.2 in order to aid interpretation (trajectories for taught words in upper panels, and for untaught words in lower panels). Note that for all figures, points represent medians, error bars represent the interquartile range, and asterisks represent statistically significant differences between adjacent time points. For reference, group and individual trajectories for all subcategories of the bespoke word knowledge assessment are presented graphically in Appendix 7.5 (page 286). Additionally, metrics relating to average scores for each of the 20 taught words at each time point, as well as standardised mean differences (Cohen's *d*) between pre- and posttest scores are presented in tabular format in Appendix 7.6 (page 288).

Table 7.3: Progress of Intervention Participants (n=9) in Total, Word, and Sentence Score for Taught Words

	Total score	Words and Sentences	Definition	Background	Lexical	Non-Verbal	Total	Sentence score	Syntax	Morphology	Semantics	Total	Baseline to Pretest			Pretest to Posttest			Posttest to Main.			
													Z	p	r	Z	p	r	Z	p	r	Z
		55.00 (28.50)	9.00 (7.00)	8.00 (6.00)	1.00 (1.50)	0.00 (0.30)	18.00 (13.80)	12.00 (5.00)	12.00 (4.50)	12.00 (4.50)	17.00 (5.00)	40.00 (17.00)	94.00 (44.75)	0.30	.767	.07	2.67	.008	.63	-1.01	.314	-.24
		9.00 (7.00)	9.00 (7.50)	5.00 (3.50)	1.00 (2.00)	0.00 (0.30)	18.00 (7.50)	12.00 (4.50)	9.00 (7.50)	17.00 (11.00)	17.00 (6.00)	27.00 (15.00)	18.00 (7.50)	0.14	.888	.03	2.67	.008	.63	-0.60	.549	.14
		8.00 (6.00)	8.00 (6.00)	5.00 (3.50)	1.00 (2.00)	0.00 (0.30)	18.00 (13.80)	12.00 (4.50)	5.00 (3.50)	10.00 (6.00)	10.00 (6.00)	35.00 (20.08)	9.00 (7.50)	-1.26	.208	-.30	1.83	.068	.43	1.55	.121	-.37
		1.00 (1.50)	1.00 (2.00)	1.00 (2.00)	1.00 (2.00)	0.00 (0.30)	18.00 (13.80)	12.00 (4.50)	1.00 (2.00)	3.00 (3.50)	3.00 (3.50)	2.00 (1.50)	2.00 (1.50)	0	1	.00	2.2	.028	.52	-1.80	.072	-.42
		0.00 (0.30)	0.00 (0.30)	0.00 (0.30)	0.00 (0.30)	0.00 (0.30)	18.00 (13.80)	12.00 (4.50)	0.00 (0.30)	0.00 (0.50)	0.00 (0.50)	0.00 (0.30)	0.00 (0.30)	0	1	.00	1	.317	.24	-1.14	.157	-.33
		18.00 (13.80)	18.00 (7.50)	18.00 (7.50)	18.00 (7.50)	18.00 (7.50)	18.00 (13.80)	18.00 (13.80)	14.50 (10.30)	29.50 (17.50)	29.50 (17.50)	35.00 (20.08)	35.00 (20.08)	-1.08	.282	-.25	2.67	.008	.63	-1.19	.223	.28
		12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (4.50)	17.00 (6.00)	17.00 (6.00)	16.00 (4.50)	16.00 (4.50)	0.71	.476	.17	2.25	.024	.53	0.81	.417	.19
		12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	12.00 (5.00)	13.00 (4.50)	17.00 (6.00)	17.00 (6.00)	16.00 (4.50)	16.00 (4.50)	0.63	.527	.15	2.26	.024	.53	0.58	.564	.14
		17.00 (9.00)	17.00 (9.00)	17.00 (9.00)	17.00 (9.00)	17.00 (9.00)	17.00 (9.00)	17.00 (9.00)	15.00 (11.00)	35.00 (23.50)	35.00 (23.50)	27.00 (15.00)	27.00 (15.00)	0.77	.440	.18	2.68	.007	.63	-1.26	.206	-.30
		40.00 (17.00)	40.00 (17.00)	40.00 (17.00)	40.00 (17.00)	40.00 (17.00)	40.00 (17.00)	40.00 (17.00)	42.00 (16.50)	70.00 (32.50)	70.00 (32.50)	62.00 (22.00)	62.00 (22.00)	0.65	.514	.15	2.67	.008	.63	-0.89	.374	-.21

Note: median and (interquartile range); r = effect size for Wilcoxon signed-rank test; Main. = Maintenance Test

Table 7.4: Progress of Intervention Participants (n=9) in Total, Word, and Sentence Score for Untaught Words

	Baseline	Pretest	Posttest	Main.	Baseline to Pretest			Pretest to Posttest			Posttest to Main.		
					Z	p	r	Z	p	r	Z	p	r
Total score	19.00 (13.00)	17.00 (20.50)	26.00 (20.00)	34.00 (17.00)	1.07	.284	.25	2.49	.013	.59	1.30	.192	.31
Word score													
Definition	5.00 (5.50)	4.00 (5.50)	4.00 (5.50)	7.00 (7.50)	-0.21	.831	-.05	1.52	.129	.36	1.62	.106	.38
Background	2.00 (1.50)	3.00 (3.00)	3.00 (2.50)	3.00 (4.00)	0.86	.389	.20	0.38	.705	.09	-0.86	.931	-.20
Lexical	0.00 (0.50)	0.00 (0.00)	0.00 (0.50)	0.00 (1.00)	-0.58	.564	-.14	0.58	.564	.14	1	.317	.24
Non-Verbal	0.00 (0.00)	0.00 (0.00)	n/a (n/a)	n/a (n/a)	n/a	n/a	n/a	1	.317	.24	n/a	n/a	n/a
Total	7.00 (7.50)	7.00 (8.00)	9.00 (7.50)	11.00 (9.00)	0.17	.863	.04	1.71	.088	.40	1.56	.119	.37
Sentence score													
Syntax	4.00 (1.50)	4.00 (2.50)	4.00 (3.00)	7.00 (3.00)	0.69	.493	.16	1.19	.234	.28	2.67	.008	.63
Morphology	4.00 (1.50)	4.00 (2.50)	4.00 (3.50)	7.00 (3.00)	0.69	.493	.16	1.19	.236	.28	1.79	.073	.42
Semantics	5.00 (4.00)	5.00 (6.00)	9.00 (10.50)	9.00 (6.00)	1.13	.258	.27	2.38	.017	.56	0.42	.676	.10
Total	12.00 (6.00)	12.00 (11.00)	17.00 (16.50)	23.00 (9.00)	0.95	.342	.22	2.54	.011	.60	1.14	.256	.27

Note: median and (interquartile range); r = effect size for Wilcoxon signed-rank test; Main. = Maintenance Test

Table 7.5: Raw Scores (top) and Percentage Change (bottom) of the 9 Children on Total, Word, and Sentence Score

Case	Baseline			Pretest			Posttest			Maintenance		
	Total	Word	Sentence	Total	Word	Sentence	Total	Word	Sentence	Total	Word	Sentence
AA	25.5	8.5	17	16	6	10	51	20	31	50	18	32
BA	49	14	35	53	11	42	61	18	43	65	15	50
FA	29	9	20	58	14	44	103	29	74	88	26	62
JG	55	14	41	54.5	14.5	40	61.5	19.5	42	59	15	44
ME	55	18	37	54	12	42	90	37	53	94	38	56
MT	64	19	45	53	16	37	100.5	29.5	71	112.5	36.5	76
NA	65	25	40	77	22	55	105	35	70	97	35	62
RA	73.5	25.5	48	50.5	21.5	29	113.5	37.5	76	101	35	66
RS	70	26	44	92	29	63	128	48	80	119	47	72
Median	55	18	40	54	14.5	42	100.5	29.5	70	94	35	62
Mean	54	17.67	36.33	56.44	16.22	40.22	90.39	30.39	60	87.28	29.5	57.78

Case	Baseline to Pretest Change %			Pretest to Posttest Change %			Posttest to Maintenance Change %		
	Total	Word	Sentence	Total	Word	Sentence	Total	Word	Sentence
AA	-37.3	-29.4	-41.2	218.8	233.3	210.0	-2.0	-10.0	3.2
BA	8.2	-21.4	20.0	15.1	63.6	2.4	6.6	-16.7	16.3
FA	100.0	55.6	120.0	77.6	107.1	68.2	-14.6	-10.3	-16.2
JG	-0.9	3.6	-2.4	12.8	34.5	5.0	-4.1	-23.1	4.8
ME	-1.8	-33.3	13.5	66.7	208.3	26.2	4.4	2.7	5.7
MT	-17.2	-15.8	-17.8	89.6	84.4	91.9	11.9	23.7	7.0
NA	18.5	-12.0	37.5	36.4	59.1	27.3	-7.6	0	-11.4
RA	-31.3	-15.7	-39.6	124.8	74.4	162.1	-11.0	-6.7	-13.2
RS	31.4	11.5	43.2	39.1	65.5	27.0	-7.0	-2.1	-10.0
Median	-0.9	-15.7	13.5	66.7	74.4	27.3	-4.1	-6.7	3.2
Mean	6.9	-6.3	14.8	75.6	103.4	68.9	-2.6	-4.7	-1.5

7.2.1 Group and Individual Trajectories for Taught and Untaught Words

Baseline Period. During the baseline period (baseline to pretest), group trajectories showed only a low rate of progress on both sets of words between baseline and pretest. Indeed, improvement between these time points was not statistically significant for taught words ($Z = 0.30$, $p = .767$, $r = .07$) or untaught words ($Z = 1.07$, $p = .284$, $r = .25$) and thus, the baseline period provided no evidence that children were making considerable progress in their knowledge of the 20 to-be-taught words or the 10 untaught words prior to explicit instruction. Individual trajectories for taught words showed some variation, with scores tending to remain stable or increase for most children; however, exceptions to this pattern were cases RA, MT, and AA, whose scores had decreased by pretest. As discussed in Section 6.9.3, the pretest was administered by coordinators, and decreases in scores at pretest may have been a result of children working with an unfamiliar adult. Individual trajectories for untaught words in the baseline period again tended to remain stable or increase slightly, and it is interesting to note the correspondence between children's growth in taught and untaught words. For example, it is telling that MT and AA, who experienced negative growth for taught words, also did so for untaught words, suggesting that increases or decreases in scores were similar across both groups of words assessed.

Pretest to Posttest Change. Group trajectories between pretest and posttest showed medium and statistically significant increases in knowledge of both taught words ($Z = 2.67$, $p = .008$, $r = .63$) and untaught words ($Z = 2.49$, $p = .013$, $r = .59$). Although children made significant progress by posttest, this knowledge was by no means at ceiling level, with the group scoring on average 100.50 points out of a possible of 170 (range: 51-128).

Analysis of individual trajectories showed some interesting patterns (individual scores and percentage change between time points are presented in Table 7.5). The average increase in taught word knowledge between pretest and posttest of 33.94 points did not characterise the trajectories of all children equally well. In particular, cases JG and BA made very little progress by posttest in terms of total score (7 and 8 raw score points, respectively). This is in contrast to case MT, who scored very similarly with respect to BA and JG at pretest but who made a substantially greater gain of 47.5 points by posttest. Yet another contrast is the case of AA, who had the lowest pretest total score of all children (16), but who again made a substantially faster rate of progress than BA or JG by posttest (35 points). The specific cases of JG and BA will be returned to in Section 7.2.3, although it is interesting to note at this point how significant improvements in word knowledge by posttest were not exhibited by all nine children. Lastly, individual trajectories of progress on untaught words suggested that the statistically significant improvement found in the group trajectory between pretest and posttest may have been accounted for by two particularly high-performing cases, RS and NA who, up to and including this point, had shown a roughly linear rate of progress in untaught word knowledge since baseline. These cases contrasted with the majority of children who continued to show either a stable or only slightly elevated rate of growth in untaught word knowledge since pretest.

Posttest to Maintenance. Although taught word knowledge did decrease slightly between posttest and maintenance, this trend did not reach statistical significance ($Z = -1.01$, $p = .314$, $r = -.24$), and represented a reduction of only 6.5 raw scores from the posttest median of 100.50. Thus, the group trajectory suggested that word knowledge gained during the intervention had largely

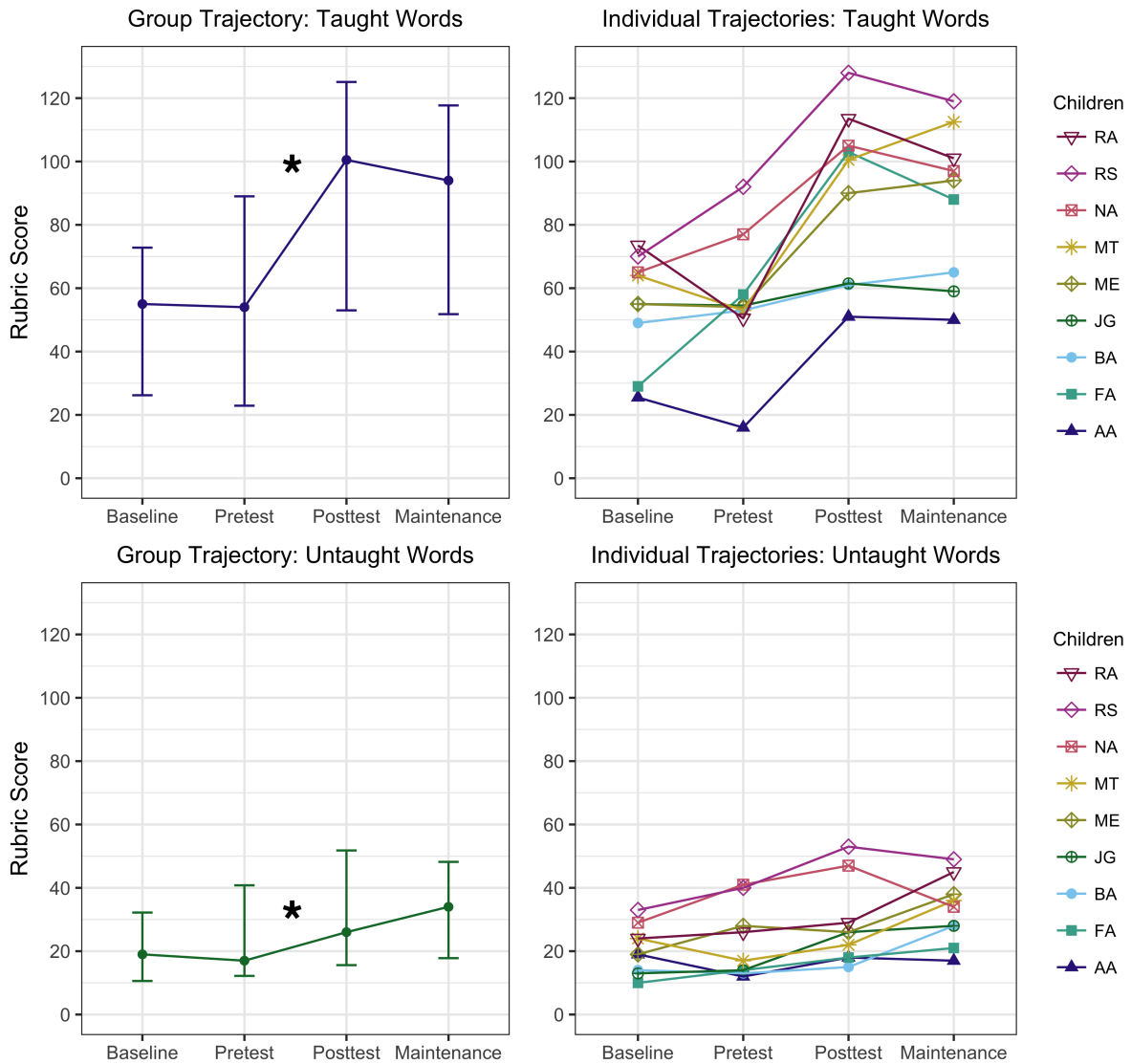


Figure 7.2: Group and Individual Trajectories in Total Score (Word + Sentence Score) for Taught and Untaught Words

been retained six months later. In contrast, untaught word total scores continued to increase between posttest and maintenance, although again this trend did not reach statistical significance ($Z = 1.30, p = .192, r = .31$). Group trajectories in taught word knowledge were found to be an accurate reflection of individual growth trajectories, with most children showing slight declines in their scores by maintenance (although one exception was MT, whose score continued to increase, albeit at a slower rate than between pretest and posttest). Similarly for untaught word knowledge, the majority of cases showed a slight increase by maintenance, and it is interesting to note that cases RS and NA, the two highest-scoring children at posttest, both experienced declines in their scores by maintenance, perhaps due to regression to the mean.

7.2.1.1 Incidence of Non-Responses

In order to further elucidate children's response patterns across the intervention, incidence of non-responses was computed for both taught and non-taught words (both in word and sentence score), where non-response is defined as a score of zero due to not offering any answer, as opposed to an incorrect answer. Counts of non-responses are presented in Table 7.6 below. As can be seen from column totals, incidence of non-responses decreased sharply and statistically significantly between pre- and posttest for taught words ($t(8) = 2.96, p = .018$), but not for untaught words ($t(8) = -0.57, p = .584$) which showed greater stability over time. Therefore, despite statistically significant increases in both taught and untaught words between pretest and posttest, response patterns in the two categories differed considerably. This serves to support the finding that increases in untaught words were due to improvements in children's expressive grammar skills, as noted above, rather than to acquisition of new knowledge of untaught words in which case a decrease in incidence of non-responses between pretest and posttest may have been expected. Additionally, it is interesting to note here that two of the lowest-scoring children, BA and JG, tended to make fewer non-responses than their higher-scoring peers, suggesting that their low scores were not merely due to a low level of verbosity.

Table 7.6: Incidence of Non-Responses Across Time Points (Taught and Untaught Words)

Child	Taught Words (n=20)				Untaught Words (n=10)			
	Baseline	Pretest	Posttest	Main.	Baseline	Pretest	Posttest	Main.
AA	22	29	16	18	12	14	12	12
BA	2	0	1	1	1	0	0	2
FA	25	7	4	5	13	6	8	10
JG	4	6	3	7	6	6	0	4
ME	14	18	0	2	6	9	10	3
MT	12	12	4	1	8	10	12	4
NA	15	0	0	4	10	0	3	9
RA	7	9	2	3	5	8	11	4
RS	11	4	0	0	6	1	3	3
Total	112	85	30	41	67	54	59	51

Note: *Main.* = Maintenance test. Individual counts represent sum of non-responses for each child across word and sentence subcategories of the bespoke word knowledge assessment.

Summary. These results do present some evidence for the efficacy of the intervention teaching on children's vocabulary knowledge of the 20 taught words. Children's progress between time points outside of the instruction period (i.e. baseline to pretest and posttest to maintenance) showed minimal or negative growth, which contrasted strikingly with rate of progress between pretest and posttest, a period of 10 weeks in which they received targeted, one-to-one vocabulary instruction. Again, it is interesting that significant growth was also observed in untaught words between pretest and posttest. Individual trajectories were generally accurate reflections of group trajectories, and showed variation in the baseline period, where roughly equal numbers of children showed increases, decreases, and little or no growth. Variation was also seen between pretest and posttest, where not all children were found to benefit equally from the intervention teaching,

in particular, the cases of BA and JG, which will be returned to in Section 7.2.3. Having examined children's overall progress in taught and untaught word knowledge across the four time points of the study (research question 1), the following section will examine trajectories in receptive and productive knowledge specifically (research question 2).

7.2.2 Group and Individual Trajectories in Receptive and Productive Knowledge

As described in Section 6.9.1, for each of the 20 taught words, children received scores on the bespoke word knowledge assessment for a verbal definition (word score; receptive knowledge) and for using the target word within a sentence (sentence score; productive knowledge). As indicated in Table 7.3, these scores were further divided into their constituent subcategories. This section will examine group and individual trajectories in receptive and productive knowledge between each time point, firstly in terms of total word score and total sentence score, and then followed by progress across the various subcategories. Due to space limitations and interest in the effect of explicit instruction on word knowledge, results will be presented here for children's progress on taught words only. Group and individual trajectories for total word score and total sentence score are presented in Figure 7.3 overleaf, and graphs depicting trajectories for word and sentence subcategories are provided in Appendix 7.5 on page 286.

Baseline Period. Between baseline and pretest, group trajectories did not change significantly for word score ($Z = -1.08$, $p = .282$, $r = -.25$) or sentence score ($Z = 0.65$, $p = .514$, $r = .15$), suggesting that children were not already making consistent gains in their receptive or productive knowledge of the 20 taught words prior to the onset of teaching. As indicated in Table 7.3, the highest scores obtained within the word score category for taught words in descending order were in *definition*, *background knowledge*, and *lexical knowledge*. Non-verbal responses were generally very rare. This pattern was retained across all subsequent time points, and shows that children tended to give not only core definitional information of words when prompted, but also relied (albeit to a lesser extent) on background and lexical knowledge in their answers.

In contrast to the slight (though non-significant) decrease observed in receptive knowledge, productive knowledge showed a slight and non-significant increase between baseline and pretest. As indicated in Table 7.3, none of the individual sentence subcategories increased significantly, with fairly even growth between *syntax*, *morphology*, and *semantics*. Children scored highest in *semantic* scores, as scores in this category ranged from 0 to 3 for each word, in contrast to scores on *syntax* and *morphology* which ranged from 0 to 1 (see Section 6.9.1 for scoring details). Thus, it appeared that, in agreement with patterns shown by total *word* and *sentence* scores, children were not making a significant rate of progress in any of the measured subcategories for taught or untaught words in receptive or productive knowledge prior to onset of explicit instruction.

Individual trajectories for receptive and productive knowledge bore close resemblance to those for *total* score in Figure 7.3; that is, children's positions relative to their peers in terms of *total* score were reflected in their trajectories for receptive and productive knowledge separately. For example, case AA, the lowest scoring child in terms of total score, was also the lowest scoring for receptive and productive knowledge separately, while a parallel conclusion can be drawn about one of the highest-scoring children, RA. As reflected in the group trajectory for progress between

7.2. Results of the Vocabulary Intervention Study

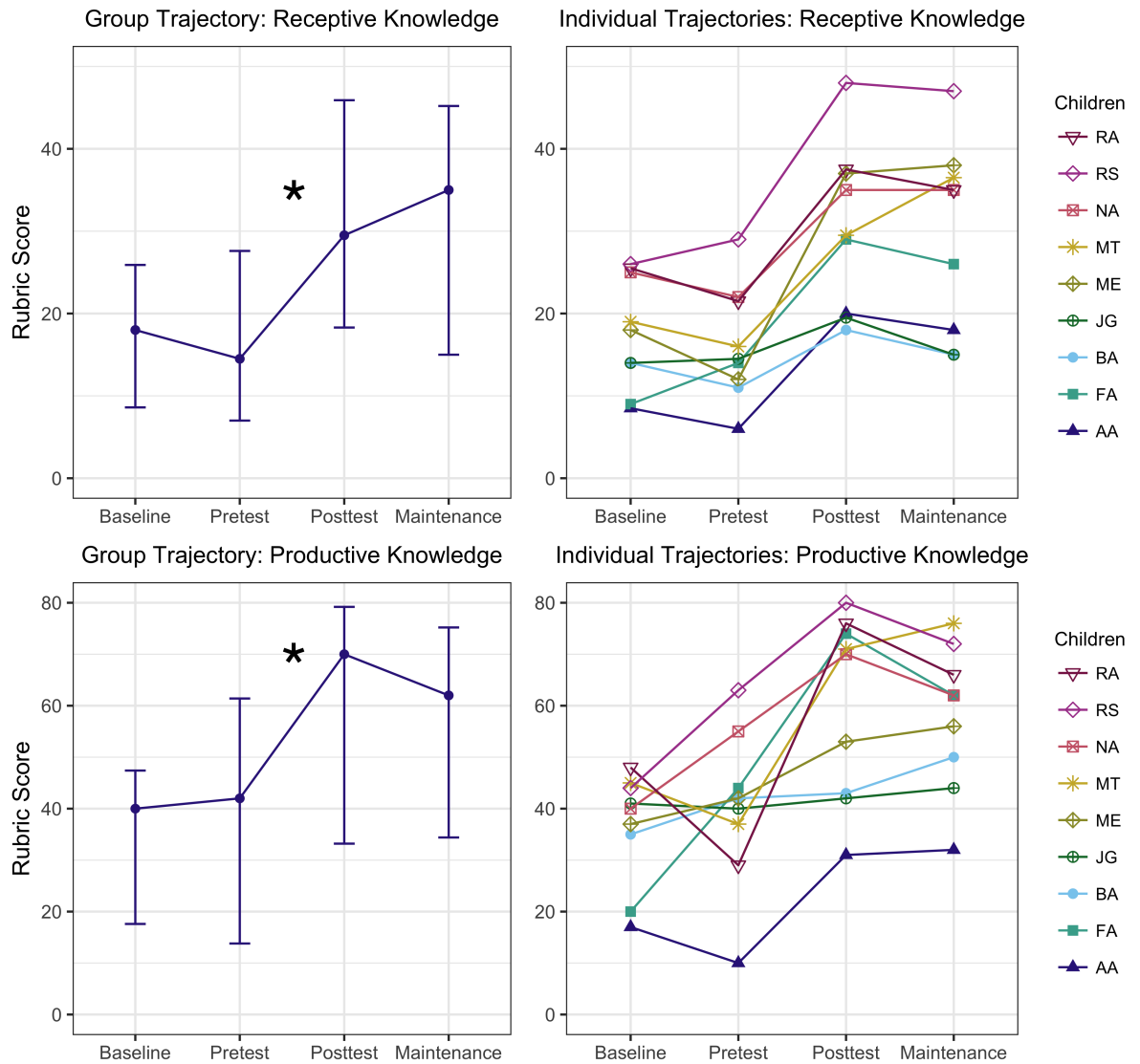


Figure 7.3: Group and Individual Trajectories in Receptive Productive Knowledge of Taught Words

baseline and pretest, children tended to make a faster rate of improvement in their receptive than productive knowledge, with extreme examples of this pattern being cases FA and RS.

Pretest to Posttest Change. By posttest, the group trajectory indicated a modest and statistically significant gain in receptive knowledge ($Z = 2.67$, $p = .008$, $r = .63$). This represented a 103% increase in median scores from 14.50 to 29.50. Word subcategories showing a significant rate of growth in this time included definition ($Z = 2.67$, $p = .008$, $r = .63$) and lexical knowledge ($Z = 2.2$, $p = .028$, $r = .52$). Improvements in lexical scores were due to the mention of synonyms (e.g. *excited* for ‘thrilled’, *fake* for ‘fraud’, *heartbroken* for ‘miserable’), and also phrasal vocabulary (e.g. *in charge of* for ‘responsible’), and antonyms (e.g. ‘*disagree* is the opposite of *agree*’). Although improvement in background knowledge did not reach statistical significance, an increase in medians from 5.00 to 10.00 does suggest that children were using more background information in their responses.

Group trajectories for productive knowledge also increased to a modest and significant degree between pretest and posttest ($Z = 2.67, p = .008, r = .63$). Of the three sentence subcategories it was *semantic* score that showed the largest improvement by posttest ($Z = 2.68, p = .007, r = .63$), followed by *morphology* ($Z = 2.26, p = .024, r = .53$) and then *syntax* ($Z = 2.25, p = .024, r = .53$). Growth in *semantic* scores was seen as a result of children providing more information in their sentences. For example, at pretest for the target word *furious*, one child said “my brother was furious” (a *semantic* score of 1), while at posttest this child’s response was “my friend was furious because someone stole something of hers” (a *semantic* score of 3). Growth was also seen in *syntax* scores as a result of using a target word as the correct part of speech; for example, for the target word *distant* at pretest one child said “I saw the distant above” (syntax score of 0; target word used incorrectly as a noun) but at posttest this child offered: “I was going on a distant ride in an airplane” (syntax score of 1; used correctly as an adjective). This example illustrates the utility of the inclusion of a productive measure of word knowledge, as this kind of qualitative change in word knowledge may not have been as readily observable in a definition-only test. Finally, no children made morphological errors related to target words; rather, where children scored 0 for morphology, this was because the target word had been used as the incorrect part of speech, and thus a morphology score would have been meaningless (see Appendix 6.3 for further scoring details).

Individual children’s pretest, posttest, and pre-to-post change scores for taught words are presented in Table 7.5 on page 210. Change scores are presented in both raw score change (top) and percentage change (bottom) in order to give an indication as to children’s improvement over time relative to their pretest scores. As with analysis of individual trajectories for *total* score, cases BA, JG, and AA were once more the lowest scorers in both receptive and productive score by posttest. Interestingly, while AA made a high rate of progress in both categories by posttest (an increase of 200%), BA and JG made very little progress, especially in productive knowledge, with gains of 2.4% and 5% relative to their pretest scores, respectively. Indeed, by posttest, the nine children had generally made a higher rate of progress in receptive (mean = 103.4%) than productive knowledge (mean = 69%) relative to their pretest scores, with extreme examples including case ME (208.3% vs. 26.2%) and RS (65.6% vs. 27%). However, some children bucked this trend by showing larger relative gains in their productive than receptive knowledge, for example cases RA (162.1% vs. 74.4%) and MT (91.9% vs. 84.4%).

Posttest to Maintenance Change. Changes in children’s scores between posttest and maintenance did not reach significance for receptive knowledge ($Z = 1.19, p = .223, r = .28$) or productive knowledge ($Z = -0.89, p = .374, r = -.21$), suggesting that children had generally retained both types of knowledge six months after the intervention. Group trajectories in Table 7.3 and Figure 7.3 utilise the median as a measure of central tendency, and appear to suggest differing directions for growth in the two categories of knowledge; namely, positive for receptive ($r = .28$) and negative for productive ($r = -.21$). However, comparison with the mean in Appendix 7.4 indicates that both scores decreased slightly by maintenance (receptive: $r = -.27$; productive: $r = -.31$). Indeed, this slight decrease is borne out by analysis of individual trajectories discussed below and depicted in the right panels of Figure 7.3. In other words, while data suggest that children’s scores decreased very slightly and to a non-significant degree between posttest and maintenance, this was of insufficient magnitude to return any of the nine children to their levels of receptive or productive

7.2. Results of the Vocabulary Intervention Study

knowledge as measured at baseline or pretest, further supporting the finding that they retained the knowledge they had gained during the teaching.

Individual trajectories between posttest and maintenance showed a degree of variation across both categories of knowledge. As discussed above, most children's scores had decreased slightly six months following the intervention, however one exception was case MT who continued to show gains in both receptive and productive knowledge. Interestingly, children who scored highest at posttest tended also to exhibit steeper decreases in productive knowledge by maintenance (particularly cases RS, RA, and FA), while the lowest scorers at posttest – particularly cases AA and JG – showed almost zero growth by this point.

Summary. The division of total scores into receptive and productive subcomponents allowed a more fine-grained investigation into the types of knowledge acquired by children in the intervention. Results showed that both receptive and productive forms of knowledge were closely tied to children's overall performance in terms of total scores described in Section 7.2.1 and that neither receptive nor productive knowledge improved significantly prior to the onset of instruction. However, some subtly different patterns did emerge across the two categories; specifically, children generally made a faster rate of progress in receptive than productive knowledge. Between pretest and posttest, both knowledge categories showed modest and significant improvements, which were retained by maintenance. For receptive knowledge, the largest gains in this time were seen in *definition* and *lexical* scores, and for productive knowledge, similarly-proportioned gains were found in *syntax*, *morphology*, and *semantics*. While receptive knowledge tended to decrease slightly between posttest and maintenance, productive knowledge continued to increase, potentially as a result of the children's continually improving grammatical skills and the addition of more information to their sentences. Results showed that some gains in productive knowledge were attributable to children learning to employ target words in sentences as the correct part of speech.

7.2.3 Further Analysis of Cases BA and JG

Analysis of individual trajectories revealed that all children made some level of progress in their knowledge of taught words between pretest and posttest, and all retained what they had learned by maintenance. However, two children (cases BA and JG) represented exceptions to this trend due to the very small magnitude of progress they had made by posttest. This section will take advantage of the multiple case series design introduced in Chapter 6 in an effort to explore possible reasons for these children's lack of progress in the intervention.

Previous studies have suggested that intervention programmes may not be optimally effective if EAL learners do not possess sufficient levels of English language proficiency to benefit from teaching activities (e.g. Dockrell et al., 2010; Schaefer et al., *under review*). Therefore, it was of interest to what extent BA and JG may also have conformed to this pattern. As shown in Table 7.5, at pretest, BA and JG possessed levels of knowledge of the to-be-taught words that were similar to those of other children (e.g. cases MT, RA, and FA) and, in one case, much higher (case AA). Since these other children all made considerably high rates of progress, BA and JG's levels of pretest knowledge were unlikely to account for their very low gains in word knowledge. As a result, the decision was made to examine BA and JG's scores on other measures in the

assessment battery of the longitudinal study. The standardised (scaled and standard) scores of all nine intervention participants for BPVS, CELF EV, WISC VC, and CELF FS assessments are presented in Table 7.7 below. As shown in this table, BA and JG were not the lowest-scoring children across all four measures, for instance, with case ME also failing to obtain a standard score on the BPVS (< 69), case AA scoring below BA and JG on both the CELF EV and WISC VC, and case FA scoring below BA and JG on the CELF FS.

Given that cases ME, AA, and FA scored below BA and JG on certain measures, it may be queried why these children too did not fail to make progress as a result of the intervention teaching. Further analysis of scores in Table 7.7 reveals that unlike BA and JG, other children appeared to compensate for their low scores on certain measures by scoring relatively higher on other standardised measures. For instance, while ME failed to obtain a standard score on the BPVS, this child also happened to score within the average range on all other measures presented here; similarly, while AA scored low on a number of measures, this child scored relatively more highly on the BPVS (with a similar situation applying to FA). Indeed, with the exception of cases BA, JG, and AA, all other children scored within the average range on at least one measure. Thus, the lack of progress experienced by BA and JG across the ten weeks of the intervention teaching may have been a result of their generally low level of English language proficiency. While other children exhibited low proficiency in certain areas such as expressive grammar or receptive vocabulary knowledge and were perhaps able to rely on their relatively higher proficiency in other domains, cases BA and JG appeared to be at a disadvantage with their low level of proficiency across different domains.

Table 7.7: Individual Children's BPVS, CELF EV, WISC VC, and CELF FS Standardised (Scaled and Standard) Scores from t1 of the Longitudinal Study (cases BA and JG highlighted in green)

Child	BPVS*	CELF EV*	WISC VC*	CELF FS
JG	<69	5	6	3
BA	71	2	4	3
AA	80	4	5	3
FA	80	10	6	1
ME	<69	7	8	9
MT	76	6	7	4
NA	81	6	8	6
RA	83	6	9	6
RS	80	6	10	10
Group Mean	76.5	5.78	7	5

Note: BPVS = British Picture Vocabulary Scale-III; CELF EV = Clinical Evaluation of Language Fundamentals-IV Expressive Vocabulary; WISC VC = Wechsler Intelligence Scale for Children-IV Vocabulary subtest; CELF FS = Formulated Sentences subtest; * measure used to determine eligibility for participation in intervention.

The relatively low rate of progress for cases BA and JG is also reflected in both receptive and productive subscores (see Appendix 7.5), especially in the latter: as seen in Table 7.5, these children made only 1 and 2 raw points of progress, respectively, in sentence total scores be-

7.2. Results of the Vocabulary Intervention Study

tween pretest and posttest. Interestingly, where BA did show improvement in sentence score, this was outside of the teaching phase, between posttest and maintenance (an improvement of 7 raw score points). Reference to BA and JG's sentence responses at posttest reveals that both children tended to use words as the wrong part of speech (target words underlined; e.g. I annual around the sky; yesterday I fraud), and to offer sentences that were nonsensical (e.g. my friend thrilled at his brother; the agony of the Great Britain is Wales, Scotland, Ireland). Both children's *syntax* scores declined between pretest and posttest, suggesting that this aspect of taught word knowledge could have been emphasised more during the teaching.

Further analysis of progress by BA and JG in each subcategory also revealed some interesting patterns. In terms of receptive subscores, while JG was more likely than BA to score highly within the *definition* category, BA showed stronger performance in the *background knowledge* category, often relying heavily on this kind of knowledge in answers given during the verbal definitions task (e.g. *you can't afford to buy a bike; you're in a shop for target word bargain*). Both BA and JG were amongst the lowest scorers in the *lexical knowledge* category. To some extent, this provides further support to the suggestion above that these children's rate of progress on target words may have been constrained by their generally low English language proficiency, including in their vocabulary knowledge; i.e. with both children scoring very low on lexical knowledge and one child relying heavily on background knowledge.

Class teachers of intervention participants were also asked to complete the CELF Observational Rating Scale (ORS; Semel, Wiig, & Secord, 2006) questionnaires for children participating in the intervention study. The purpose of the ORS is to ascertain children's difficulties in the domains of speaking, listening, reading and writing, assessing frequency of occurrence along a 4-point scale from 'never' occurs to 'always' occurs. Unfortunately, completed ORS forms were received for only 6 out of the 9 intervention participants (scores in Table 7.8 below), including BA but not JG. As a guide to interpretation, a high score on the CELF ORS indicates presence of more language difficulties.

Table 7.8: CELF Observational Rating Scale Scores for 6 Intervention Participants

Child	Speaking and Listening	Reading and Writing	Total
BA	25	14	39
FA	9	9	18
ME	9	6	15
MT	16	10	26
NA	0	0	0
RS	13	8	21

Interestingly, BA scored the highest across all four categories, further supporting the notion that this child had generally low English language proficiency which may have inhibited progress during the intervention. Two additional points concerning BA merit mention: firstly, BA's class teacher and intervention coordinator reached consensus regarding poor attention, with both noting this as a weakness for this child; secondly, during testing it was disclosed to the researcher that since joining the school, BA had made a number of prolonged trips back to his family's home

country, which likely resulted in no or very little exposure to English. ORS information was not obtained for JG; however, unlike BA, parental questionnaire data was available for this child and served to provide further contextual information. For instance, JG was the only child taking part in the intervention not to have been born in the U.K., beginning school aged 6. Thus, additional data suggest a situation in which both children who failed to make as considerable a rate of progress as their peers in their target word knowledge were those who exhibited generally low English language proficiency and, for different reasons, had likely experienced lower levels of exposure to English.

Summary. The employment of a multiple case series design allowed further investigation into the factors which may have placed limitations on the progress made by two particularly low-scoring children. Particularly, cases BA and JG appeared to possess relatively lower levels of general English language proficiency than their peers, potentially accounting for their lack of progress during the intervention.

7.2.4 Transfer to Standardised Assessments

Alongside the bespoke word knowledge assessment, children in the intervention were also administered two standardised assessments from the longitudinal study (CELF FS and WISC VC) in order to examine the possibility of any transfer effects of the intervention teaching on general language skills. Descriptive statistics for children's performance on these measures (raw and scaled scores) at t1, baseline, posttest, and maintenance, as well as inferential statistics comparing progress between time points are presented in Table 7.9. Children's individual trajectories on standardised measures (in raw and scaled scores) are presented in Table 7.10.

Unfortunately, due to timing constraints, standardised measures could not be administered additionally at pretest, and therefore these measures do not conform to the 'double pretest' design of the intervention study⁵. As a result, caution should be exercised in the interpretation of results pertaining to transfer, as it is possible that any gains in performance between baseline and posttest may have occurred in the interval preceding the actual intervention teaching (between baseline and pretest), or alternatively during both this period and the intervention teaching. Average intervals between the four time points are as follows: t1 to baseline = 5.5 months; baseline to pretest = 1.3 months; baseline to posttest = 3.1 months; posttest to t3 = 5 months⁶ (see Figure 6.2 on page 199). Group and individual trajectories in raw and scaled scores are presented graphically in Figures 7.4 and 7.5. Tables with parametric statistics (mean, SD, t-tests) are available in Appendix 7.4.

⁵This would have required additional training of intervention coordinators in the administration of CELF FS and WISC VC subtests. Additionally, given that the bespoke word knowledge assessment was already administered during children's first teaching sessions, the addition of a further two assessments would likely have been overly demanding.

⁶Note that because the maintenance test was not necessarily carried out at the exact same time as t3 of the longitudinal study, the interval between posttest and maintenance was slightly longer, at six months.

Table 7.9: Progress on Standardised Measures Across the Study for all Intervention Participants (Raw and Scaled Scores)

		t1		Posttest	t3	t1 to Baseline			Baseline to Posttest			Posttest to t3		
		Baseline	t1			Z	r	p	Z	r	p	Z	r	p
Raw Scores	WISC VC	19.00	24.00	24.00	26.00	2.41	.57	.016	1.88	.44	.061	2.39	.56	.017
		(7.00)	(9.00)	(6.50)	(11.00)									
Scaled Scores	CELF FS	27.00	29.00	38.00	40.00	1.01	.24	.312	2.31	.55	.021	1.82	.43	.068
		(11.00)	(10.00)	(6.50)	(9.00)									
Raw Scores	WISC VC	7.00	8.00	8.00	8.00	1.34	.32	.180	1.41	.33	.157	0.33	.08	.739
		(3.00)	(3.00)	(3.00)	(4.50)									
Scaled Scores	CELF FS	4.00	4.00	7.00	7.00	-0.14	.03	.887	2.33	.55	.020	1.79	.42	.073
		(4.50)	(4.00)	(2.50)	(2.50)									

Note: Descriptive statistics represent medians and (interquartile range); *r* and *p* calculated from Wilcoxon Signed-Rank Tests. Interpretation of *r* as follows: 0.2 = small, 0.5 = medium, 0.8 = large.

Table 7.10: Individual Trajectories in Standardised Measures between t1, Baseline, Posttest, and t3

Case	t1			Baseline			Posttest			t3		
	WISC VC	CELF FS		WISC VC	CELF FS		WISC VC	CELF FS		WISC VC	CELF FS	
AA	16 (5)	24 (3)		19 (6)	25 (2)		21 (6)	38 (7)		22 (6)	34 (5)	
BA	13 (4)	24 (3)		18 (6)	22 (2)		19 (6)	35 (6)		17 (5)	39 (7)	
FA	18 (6)	18 (1)		18 (6)	30 (5)		23 (7)	43 (9)		25 (8)	40 (7)	
JG	18 (6)	24 (3)		21 (7)	30 (5)		20 (6)	36 (6)		21 (6)	38 (7)	
ME	23 (8)	39 (9)		26 (9)	43 (9)		24 (7)	37 (6)		33 (10)	47 (9)	
MT	19 (7)	28 (4)		24 (8)	31 (6)		28 (10)	48 (11)		27 (8)	43 (9)	
NA	24 (8)	31 (6)		24 (8)	37 (6)		24 (7)	40 (7)		26 (7)	48 (11)	
RA	24 (9)	33 (6)		28 (10)	33 (6)		29 (10)	42 (8)		32 (11)	40 (7)	
RS	28 (10)	46 (10)		32 (11)	39 (7)		37 (12)	44 (9)		40 (13)	48 (10)	
Mdh (Mean)	19 (20.67)	28 (29.67)		24 (23.33)	31 (32.22)		24 (25.00)	40 (40.33)		26 (27.00)	40 (41.89)	
	t1-Baseline Change %			Baseline-Posttest Change %			Posttest-t3 Change %					
Case	WISC VC	CELF FS		WISC VC	CELF FS		WISC VC	CELF FS		WISC VC	CELF FS	
AA	18.8	4.2		10.5	52.0		4.8	-10.5				
BA	38.5	-8.3		5.6	59.1		-10.5	11.4				
FA	0	66.7		27.8	43.3		8.7	-7.0				
JG	16.7	25.0		-4.8	20.0		5.0	5.6				
ME	13	10.3		-7.7	-14.0		37.5	27.0				
MT	26.3	10.7		16.7	54.8		-3.6	-10.4				
NA	0	19.4		0	8.1		8.3	20.0				
RA	16.7	0		3.6	27.3		10.3	-4.8				
RS	14.3	-15.2		15.6	12.8		8.1	9.1				
Mean Change	16	12.5		7.5	29.3		7.6	4.5				

Note: Statistics in upper table represent raw scores and (scaled scores); per cent change represented in lower table

7.2. Results of the Vocabulary Intervention Study

Standardised Measures: Raw Scores. In terms of group trajectories in raw scores (depicted in Figure 7.4 below), Wilcoxon Signed-Rank tests revealed statistically significant progress in WISC VC performance preceding and following the intervention, (between t1 and baseline; $Z = 2.41$, $p = .016$, $r = .57$; and between posttest and t3; $Z = 2.39$, $p = .017$, $r = .56$), but no significant change in the relatively shorter period of time between baseline and posttest ($Z = 1.88$, $r = .44$, $p = .061$). For CELF FS, the inverse pattern was detected, whereby children did not make significant progress preceding the intervention (t1 to baseline: $Z = 1.01$, $p = .312$, $r = .24$), or following the end of the intervention (posttest to t3: $Z = 1.82$, $p = .068$, $r = .43$). In contrast, children did make a significant rate of progress on the CELF FS in the 3-month period between baseline and posttest ($Z = 2.31$, $p = .021$, $r = .54$).

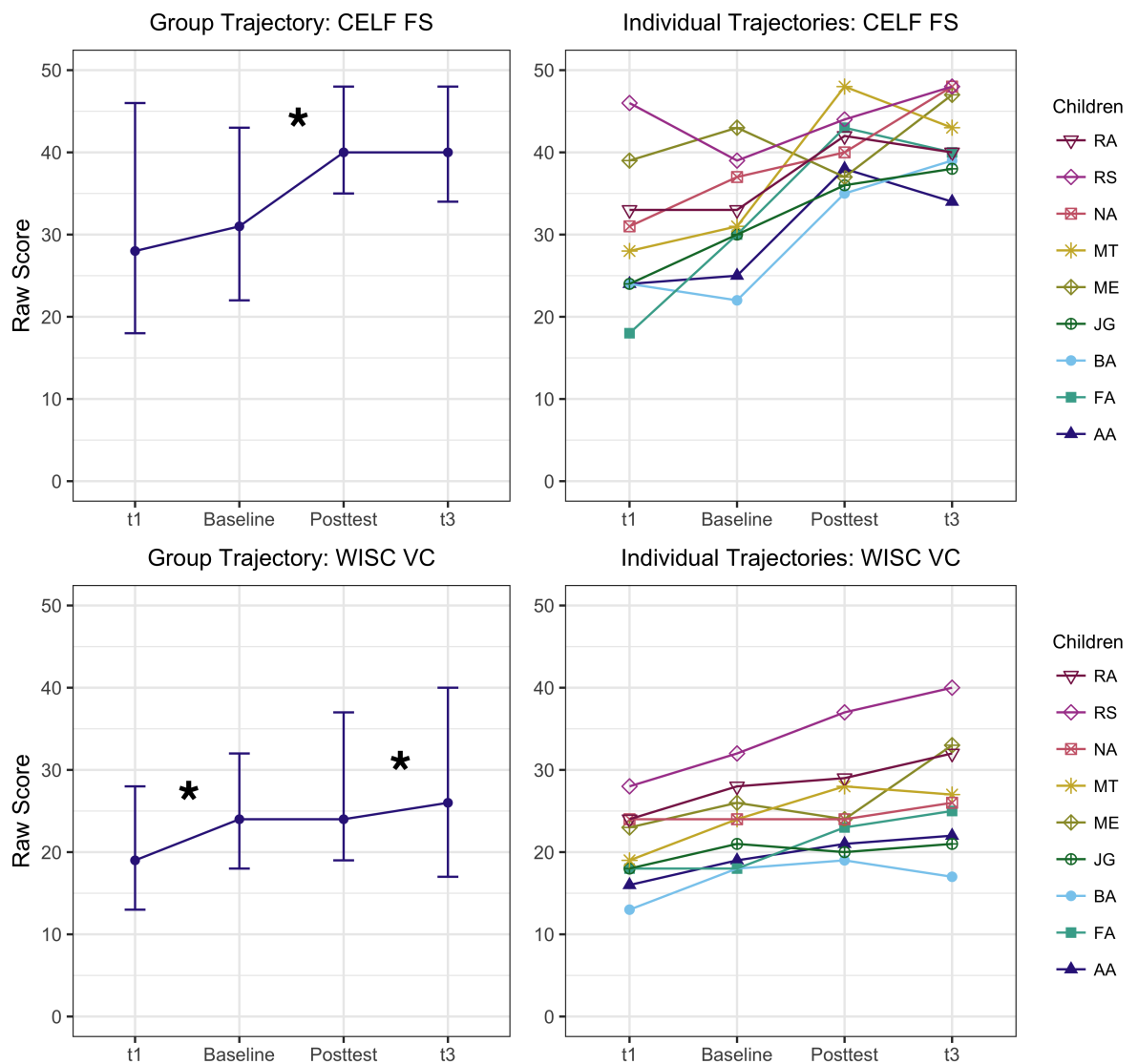


Figure 7.4: Intervention Participants' Progress on CELF FS and WISC VC (Raw Scores). Note: CELF FS = Clinical Evaluation of Language Fundamentals-IV Formulated Sentences; WISC VC = Wechsler Intelligence Scale for Children-IV Vocabulary

Individual trajectories in raw scores generally resembled group trajectories, with the rate of baseline to posttest progress being relatively larger in magnitude for CELF FS than WISC VC

performance. Table 7.10 presents individual raw and scaled scores for CELF FS and WISC VC, as well as percentage change between time points in raw scores. Data indicate considerable variation in rate of progress over time, although in line with group trajectories, children tended to make a higher rate of progress on CELF FS than WISC VC between baseline and posttest (average increases of 29.3% and 7.5%, respectively). Interestingly, case BA made the highest relative rate of progress in CELF FS by posttest (59.1%), which was not matched by this child's progress from t1 to baseline (-8.3%) or posttest to t3 (11.4%). A similar pattern applied also to cases AA and MT. One exception was that of case ME, whose CELF FS score decreased by posttest. However, given that this child's t3 CELF FS score had risen to an above-baseline level, it is possible that this score depression at baseline was due to measurement error.

Standardised Measures: Scaled Scores. In terms of scaled scores, the only statistically significant increase over the course of the study was in CELF FS between baseline and posttest ($Z = 2.33$, $p = .020$, $r = .55$), with no significant changes in WISC VC performance between any pair of time points (depicted in Figure 7.5 overleaf). This suggests that children improved their ranking in expressive grammar skill relative to age-related expectations between baseline and posttest.

Few trajectories in scaled scores represented exceptions to the general group trend across the four time points. Some children, including cases FA, JG, and MT, did make moderate progress on both measures in the 5.5 months between t1 and baseline, suggesting that their sentence construction skills were continuing to improve prior to the onset of the intervention teaching. By posttest, 8 out of 9 children had made gains according to age-related expectations on the CELF FS, particularly cases BA and AA, the two lowest-scoring children at baseline and two of the lowest-scoring children on the bespoke word knowledge assessment. Increases in scores between baseline and posttest placed five children within the average performance range of the CELF FS (i.e. scoring between 7 and 13)⁷, although fewer children crossed this threshold on the WISC VC, where performance was generally higher than on CELF FS at baseline. By the end of the longitudinal study at t3, the group diverged in direction, with roughly half of children's scores continuing to rise, and half showing decreases; nevertheless, for all children, t3 CELF FS scaled scores were either the same or higher than those measured at t1. Individual trajectories for progress on WISC VC generally matched those of the group trajectory, with children's scores tending to change only to a small degree between each time point. One exception to the group trajectory was case ME; as discussed above, this child's CELF FS raw score decreased between baseline and posttest. It should be noted, however, that this represented a drop of only 6 raw scores, but resulted in a decrease from 9 to 6 in terms of scaled score due to the scoring threshold of this assessment.

Summary. While significant gains were made in vocabulary depth (WISC VC) by the group of nine children before and after the intervention, it was expressive grammar skill (CELF FS) which showed a significant gain between baseline and posttest. In contrast, WISC VC scores did not show statistically significant gains. While this pattern may be suggestive of transfer of intervention teaching to children's expressive grammar skills, it must be borne in mind that lack of administration of standardised assessments at pretest means that it cannot be ascertained whether children made further gains in the run up to the intervention (baseline to pretest). Individual trajectories

⁷However, cases BA and JG were still not scoring within the average range by posttest, both obtaining scaled scores of 6 on CELF FS.

7.3. Discussion of Results

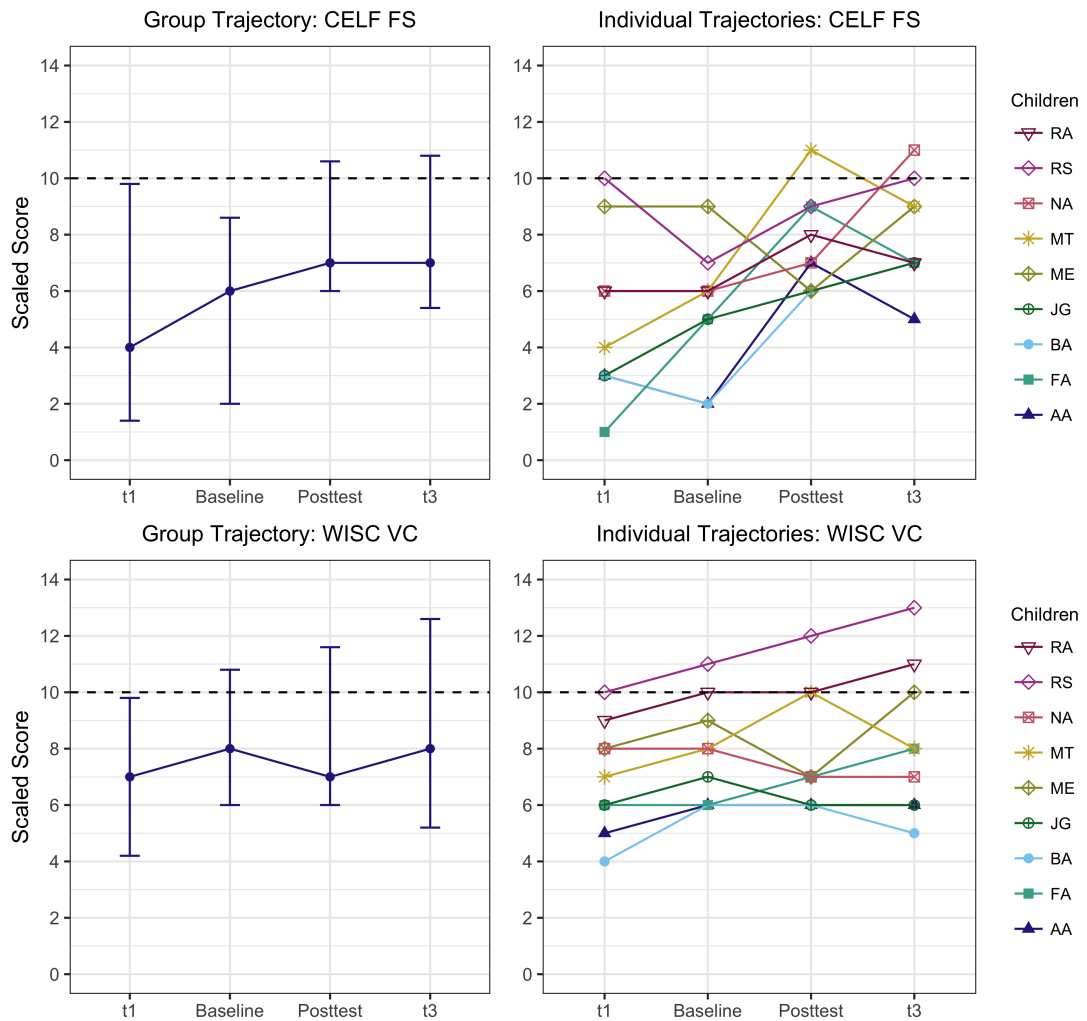


Figure 7.5: Intervention Participants' Progress on CELF FS and WISC VC (Scaled Scores). Note: CELF FS = Clinical Evaluation of Language Fundamentals-IV Formulated Sentences; WISC VC = Wechsler Intelligence Scale for Children-IV Vocabulary. Dashed line represents norming population mean of measures

confirm the general pattern of relatively larger increases in expressive grammar skill, and analysis of scaled scores indicated that a number of children were performing within the average range on the CELF FS by posttest.

7.3 Discussion of Results

The results of the intervention will now be discussed with reference to the four research questions of the study, beginning with a brief summary and then moving on to discuss the efficacy of intervention teaching, relative gains in receptive and productive knowledge, analysis of individual trajectories, and lastly transfer of vocabulary teaching to standardised assessments.

The intervention study investigated the effect of explicit instruction upon the word knowledge of a small group of EAL learners with English vocabulary weaknesses. The study represents an important contribution to the field, as currently there is very little research on oral language and

vocabulary intervention with EAL learners in England generally, and with EAL learners in Key Stage 2 (age 7-11) specifically. In accordance with research on effective vocabulary instruction methods (Chapter 5), children took part in a range of activities designed to encourage acquisition of receptive and productive word knowledge; particularly, words were accompanied by both definitional and contextual information, and children were encouraged to engage with words through discussion and sentence-level activities, including using target words within sentential contexts. The design of a bespoke word knowledge assessment with separate categories for receptive and productive knowledge, coupled with a small sample size of nine children, allowed for in-depth analysis of the effects of the intervention teaching on children's developmental trajectories.

Data provided evidence for a moderate to high degree of fidelity of implementation by intervention coordinators. Ratings of children's attention and engagement were moderately high and remained fairly stable across the 10 weeks, although coordinators reported issues working in disruptive working environments. Comments from coordinators revealed that while some children found the passages challenging and showed lower levels of engagement in the second half of the session, they did enjoy the opportunity to discuss personal experiences, and clearly explored some concepts and words in depth. While it is a limitation that not all coordinators could be observed, triangulation of implementation fidelity data supports an appropriate level of adherence to teaching methods introduced in coordinators' training.

The following discussion will consider all four research questions stated at the beginning of this chapter in light of the results of the intervention study, covering additionally the roles of definitional and contextual information and prior knowledge, measurement of word learning, and efficacy of delivery methods.

7.3.1 Efficacy of Intervention Teaching

The first research question asked to what extent a short, low-intensity, one-to-one explicit vocabulary training programme improves the vocabulary knowledge of EAL learners who are identified as having English vocabulary weaknesses. The first analysis carried out to answer this question contrasted children's trajectories in knowledge of taught and untaught words, defined as the sum total of their receptive knowledge (ability to define a word) and productive knowledge (ability to use a word a sentence). Children did not make a significant rate of progress prior to the onset of formal instruction (baseline to pretest) on either taught or untaught words. Between pretest and posttest, after the 10 weeks of instruction, children made a moderate but statistically significant rate of progress on both taught and untaught words. The number of non-responses (i.e. where children did not attempt to define or use target vocabulary) dropped considerably between pretest and posttest, suggesting that to some degree, children were acquiring *new* knowledge, rather than deepening their existing knowledge. Assessment at maintenance test six months after the end of teaching revealed that children had largely retained the word knowledge they had gained by posttest.

This section will firstly discuss results in light of relevant literature, before moving on to unexpected gains in untaught word knowledge, and finally considering what factors may have accounted for the efficacy of intervention teaching, including definitional and contextual information, involvement load of activities, one-to-one delivery, and the role of intervention coordinators.

7.3. Discussion of Results

In general, the results of the vocabulary intervention accord with similarly-focused studies in the literature which show that vocabulary knowledge is generally responsive to explicit instruction (e.g. Elleman et al., 2009; Marulis & Neuman, 2010; Stahl & Fairbanks, 1986). The intervention is also comparable to other studies in its utilisation of multiple strategies to encourage depth of lexical knowledge, including passage reading, sentence-level work, mind maps, and attention to morphosyntactic properties of words (Baumann et al., 2003; Clarke et al., 2010; Nash & Snowling, 2006; Rupley & Nichols, 2005; Wilkinson & Houston-Price, 2013; Silverman et al., 2014). A key element of the intervention was the opportunity for children to discuss their ideas and engage actively in activities, a strategy recommended by practitioners and one that has been found to result in gains in knowledge of both EAL learners and their monolingual peers (Dockrell et al., 2010; Elleman et al., 2009; Gersten & Baker, 2000).

At the time of writing, the small amount of published intervention research with EAL learners in the U.K. has typically been conducted with young children (particularly within the first 1 to 2 years of formal education), often using multicomponent teaching programmes targeting vocabulary among a set of other outcome variables (Dockrell et al., 2010; Schaefer et al., *under review*). In contrast, the present study targeted vocabulary knowledge exclusively and focused instead on older children with a greater deal of experience of English-medium education. That children learning EAL in primary school Year 4 exhibit the particularly low levels of English vocabulary knowledge found here (i.e. in line with recruitment criteria; Section 6.1) suggests the need for high-quality, targeted vocabulary and oral language instruction, of which the intervention method assessed here may represent one possibility. Despite generally low levels of English language vocabulary prior to teaching, however, the majority of participants in the intervention did make considerable gains in word knowledge by posttest, suggesting that they were capable of acquiring new word knowledge as it was delivered in this study, and that by extension, such teaching may aid some EAL learners to close gaps with their monolingual peers in English vocabulary knowledge.

Results of the intervention bore close comparison to those of St. John and Vance (2014), a small-scale study targeting Tier-2 vocabulary knowledge within a sample of 18 Year 1 pupils with EAL (age 5-6). After four weeks of instruction, children showed significant gains in both taught and untaught words on a bespoke measure capturing semantic, phonological, and contextual knowledge of different sets of curriculum-related vocabulary. The present study builds on some limitations of St. John and Vance (2014): firstly, by holding vocabulary constant across all intervention participants and therefore ensuring fair comparison across children in terms of word difficulty and characteristics; secondly, by selecting vocabulary independently of curriculum content and therefore reducing chances of incidental reinforcement through regular classroom instruction; thirdly, by administering standardised measures of oral language prior to and following intervention teaching in order to assess specific effects of the intervention, as well as administering measures more closely aligned to skills focused on by the intervention; and finally, by ensuring a higher level of implementation fidelity.

One surprising finding was a significant rate of progress in *untaught* words by posttest: indeed, progress on untaught words has also been found in other vocabulary intervention studies, as children continue to be exposed to oral language outside of the teaching they receive, particularly if target word selection is aligned with the curriculum (Hadley et al., 2016; St. John & Vance,

2014; Wilkinson & Houston-Price, 2013). However, gains in untaught word knowledge here must be qualified by three points. Firstly, one source of gain in untaught word knowledge between pre- and posttest appeared to be due to two particularly high-performing children (cases RS and NA), calling into question the generalisability of this trend for the whole group (see Figure 7.3). Secondly, disaggregated data reveal that statistically significant gains occurred only in sentence-level variables (sentence *total* score and *semantics* between pretest and posttest, and in *syntax* between posttest and maintenance; Table 7.4) – such a finding should be interpreted within the more general trend indicated in the longitudinal study for children’s expressive grammar skills to improve through regular classroom instruction and indeed, this trend also applied to taught words, albeit to a larger and more robust degree. Thirdly, it was found that while the incidence of non-responses for taught words decreased significantly between pretest and posttest, the same pattern did not apply to untaught words (Table 7.6). This supports the notion that increases in untaught word knowledge were not comparable to those in taught words, and ultimately it is arguable that gains in untaught word knowledge do not threaten the validity or utility of gains in taught word knowledge immediately after participation in the intervention.

The discussion will now consider which factors may have accounted for significant gains in children’s target word knowledge over the course of the intervention. Definitional and contextual approaches, as discussed in Section 5.3, have both been found to be effective in vocabulary instruction (Marulis & Neuman, 2010; Stahl & Fairbanks, 1986). While explicit definitions provide core semantic information, additional context can be beneficial for a number of reasons: in the present study specifically, the positioning of target words within passages allowed learners to glean other information about target words (e.g. regarding constraints of use in terms of morphosyntactic, semantic, or pragmatic knowledge), encouraged use of strategies for discerning the meaning of unknown words (e.g. searching for cues; see below), and resulted in engaging activities for children, as supported in other work (Fukink & de Glopper, 1998; Nash & Snowling, 2006). Indeed, some children in the present study were reported by coordinators to enjoy reading the bespoke passages (although others found this task difficult as a result of other vocabulary employed). It is noteworthy that some children used examples from the written passages to support their answers during administration of the word knowledge assessment. For example, children tended to draw from passages for: *rescue* (a story in which two sailors experience a storm at sea and are rescued by the coastguard); *bargain* (a story in which a father buys ice cream for his children but must bargain with the seller as he does not have enough money); and *fraud* (a story about a professor who discovers what appears to be a rare fossil, only to find out that it is fake). In some cases, contextual information from passages was presented in lieu of abstract or definitional information, e.g. for *bargain*: “if you have some money and you’re trying to buy three ice creams and it’s £5 but you only have £4. . .” In other cases, both contextual and definitional information was included in children’s responses, e.g. for *rescue*: “when you save someone; like if someone falls into the sea there’s these special people and they save them from the sea before they drown”. Thus, the written passages appeared to provide an aide memoire for some children and to represent an extra resource they could rely upon to demonstrate their knowledge. Similar findings are reported from studies that present learners with target words within passages and encourage active discussion (see, for example, Dole, Sloan and Trathan, 1995, in which Grade

7.3. Discussion of Results

4 students discussed novel word meanings with relation to the situations or characters that they described in stories).

Recent studies provide evidence to support the inclusion of contextual information in vocabulary instruction programmes, and particularly the inclusion of definitional information within written passages, resulting in word learning gains for monolingual (Nash & Snowling, 2006; Wilkinson & Houston-Price, 2013) and bilingual learners (Proctor et al., 2011; Snow et al., 2009; Mancilla-Martinez, 2010; St. John & Vance, 2014). In an experimental study with 114 monolingual fourth to sixth graders (ages 9 to 12), Carnine et al. (1984) investigated the effect of the presence of contextual information on children's ability to correctly decipher the meaning of 'difficult' vocabulary. In one condition, participants were required to match definitions to target words in isolation, and in another, target words were presented alongside synonyms, antonyms, or inferential information. Results showed a clear trend for a higher proportion of correct deductions when target words were presented alongside cues, particularly synonyms. Passages in the present study similarly made use of contextual information to aid understanding of words, including synonyms or related concepts (e.g. 'long way' for *distant*), associated actions (e.g. 'shouted and stamped his feet' for *furious*), and example situations (e.g. asking a shop assistant for help for *purchase*). Thus, the combination of definitional and contextual information may be particularly effective for children learning EAL who may have fewer linguistic resources from which to draw when encountering unfamiliar vocabulary.

In the present study, vocabulary instruction was delivered in a one-to-one fashion by student coordinators. A large literature has examined the relative benefits of different within-class grouping arrangements, with systematic reviews and meta-analyses supporting the role of one-to-one and small-group instruction in improving children's learning outcomes (e.g. Bloom, 1984; Cohen, Kulik & Kulik, 1982; Lou et al., 1996; Slavin, 1987). Small group working has also been found to be beneficial over whole-class instruction for children's engagement, including for bilingual learners (Brooks & Thurston, 2010; Ross & Begeny, 2011). Particularly, small-group arrangements allow instruction to be adapted to the abilities and needs of pupils, potentially resulting in a higher mastery of learning (Chi, Siler, Jeong, Yamauchi & Hausmann, 2001; Lou et al., 1996; Schaefer et al., *under review*; Slavin, 1987). In the case of the present study, one-to-one instruction likely allowed for greater engagement in learning materials, for example in discussion of vocabulary and relating this to personal experiences, as well as in the construction of sentences.

The student coordinators who participated in the intervention were studying towards degrees in either speech science or speech and language therapy as accredited by the Royal College of Speech and Language Therapists. Additionally, all coordinators were able to demonstrate experience of working with children on a one-to-one basis for the purpose of supporting their language and/or literacy skills in either educational or clinical settings. Thus, the training and relevant experience of coordinators is likely to have benefitted children's vocabulary development. A meta-analysis of 31 studies by Elbaum, Vaughn, Hughes and Moody (2000) found that, within small-group arrangements, intervention delivery by trained college students resulted in the largest gains in learning outcomes (average weighted effect size of $d = 1.65$), followed by delivery by para-professionals ($d = 0.65$), teachers ($d = 0.36$), and volunteers ($d = 0.26$). Other research indicates that interventions can be successfully delivered by trained paraprofessionals (O'Keefe, Slocum & Magnusson, 2013), including in the areas of reading (e.g. Miller, 2003) and vocabulary (e.g. Fien

et al., 2011). This may represent a possible mode of delivery for such an intervention in future studies, and previous language and literacy interventions conducted in the U.K. with both monolingual and EAL learners report gains in vocabulary knowledge of children who receive instruction from trained teaching assistants (Bowyer-Crane et al., 2008; Clarke et al., 2010; Dockrell et al., 2010; Fricke et al., 2013; Schaefer et al., *under review*). Application of the vocabulary study to wider school contexts will be considered in the general discussion (Chapter 8).

7.3.2 Acquisition of Receptive and Expressive Knowledge

The second research question concerned the effect of explicit instruction on children's receptive understanding of new vocabulary and their ability to use it productively. As discussed in Section 2.1.2, individuals possess a rich array of knowledge about each word in their mental lexicons. Word knowledge can be considered in terms of form, meaning, and use, including not just the meaning of a word, but also its grammatical and stylistic constraints, related concepts and associations, and spoken and written forms (Nation, 2001). Extending this multidimensional view of vocabulary knowledge to a small group of EAL learners, the present study employed a bespoke word knowledge assessment in order to measure acquisition of receptive and productive knowledge as a result of intervention teaching. Meta-analyses indicate that researcher-developed scoring rubrics such as those found in studies by Hadley et al. (2016) and Proctor et al. (2011) demonstrate significantly larger gains in word knowledge than standardised measures of vocabulary knowledge (Elleman et al., 2009; Marulis & Neuman, 2010), potentially due to a higher level of sensitivity and assessment of words that have been explicitly targeted.

Regarding receptive knowledge, children at all time points tended to receive the highest scores for *definition*, followed by *background knowledge*, and *lexical knowledge*. This suggests that they did possess some core semantic or conceptual understanding and made reference to background and lexical knowledge in their verbal definitions. By posttest, a statistically significant rate of progress was found in *definition* and *lexical knowledge*, amounting to a moderate effect size in *total* receptive score. Performance in all word subcategories apart from *definition* decreased very slightly by maintenance but not significantly so, suggesting again that children had been able to retain the knowledge from the instruction they received.

Regarding productive knowledge, statistically significant gains between pre- and posttest were found in all sentence subcategories, the largest of which was in *semantics*, as children began to include more information in their sentences. Improvements in *syntax* and *morphology* also suggest that children were acquiring knowledge about words' part of speech, and were beginning to employ these words within sentences correctly or in a more appropriate manner. In line with a general trend of improvements in children's expressive grammar in the longitudinal study, scores in *syntax* continued to grow significantly after posttest, but not in *morphology* or *semantics*.

The improvements seen in receptive and productive knowledge for taught words were both statistically significant and moderate in magnitude, and thus it can be argued that the intervention teaching resulted in improvements both in children's understanding of target words and their ability to use them within sentences. One interesting finding is that while children's *lexical knowledge* scores increased by posttest, these scores dropped by almost the same rate at which they had increased by follow-up, suggesting that sustained engagement with synonyms, antonyms,

7.3. Discussion of Results

and related terms may be required in order for knowledge to be retained over a longer period of time. Indeed, reinforcement is recognised as an important facet of effective long-term vocabulary instruction (Nation, 2001), particularly so for English language learners when combined with opportunities for active engagement (Gersten & Baker, 2000). While the intervention did provide a recap activity at the beginning of each session, this included only vocabulary discussed in the previous session, as opposed to all vocabulary covered until that point in time. On the other hand, constant reinforcement of vocabulary may have resulted in more robust gains in knowledge, and future intervention studies may incorporate such opportunities.

Results may be interpreted in light of the task-induced involvement load hypothesis (Laufer & Hulstijn, 2001), which proposes that tasks with a higher involvement load (in terms of *need*, *search*, and *evaluation*) result in a higher rate of vocabulary acquisition and retention than tasks with a low involvement load. The sentences that children constructed with the help of coordinators involved the combination of target vocabulary with other words, resulting in a deeper level of processing than merely encountering words in pre-written contexts alone (i.e. written passages). Again, the inclusion of a productive component in the bespoke word knowledge assessment was advantageous by way of indicating gains in syntactic knowledge of words as well as their meaning.

Research indicates that opportunities for learners to generate their own contexts (e.g. sentence- or composition-writing) are particularly effective for vocabulary learning in bilingual adults (Hulstijn & Laufer, 2001; Kim, 2008; Nassaji & Hu, 2012; Zou, 2017). Although the present study was not an explicit investigation of the involvement load hypothesis, results do extend support for the effectiveness of high involvement load in the vocabulary acquisition of primary school children learning EAL. Not only did sentence-writing require children to combine new word knowledge with their existing knowledge, but in many cases children made reference to the short stories in which target words were introduced, suggesting that some higher-level organisation (i.e. narrative structure) may also have encouraged retention of knowledge (Dole et al., 1995). Similarly, participants in Zou (2017) who were asked to utilise novel vocabulary in written compositions were found to make use of higher-level organisation and planning, elements proposed by Zou (2017) to be crucial to the *evaluation* component of the involvement load hypothesis, and which are less likely to be utilised in simple reading or cloze activities alone. This hypothesis proposes that, had children merely read passages – a relatively more passive activity with no *evaluation* – they may not have experienced the same rate of vocabulary acquisition. While the present study was not designed to assess such a hypothesis explicitly, it does offer support for the notion that higher involvement load may encourage learning and retention of novel vocabulary among children learning EAL.

In some cases, composition-writing has been found to be more effective than sentence-writing for vocabulary learning (Hulstijn & Laufer, 2001; Zou, 2017). Despite this, composition writing would not have been a feasible strategy in the present study due to timing constraints, but it was also noted during the writing activity in the longitudinal study that both groups of children produced very little writing in the time given, and therefore composition-writing may be a more appropriate strategy once children's writing skills have obtained a certain level of speed or fluency. Nevertheless, as noted by Laufer and Hulstijn (2001), the construct of involvement load applies regardless of modality. As a result, an alternative strategy may have been to carry out a composition-writing activity orally or to ask coordinators to dictate children's utterances. Indeed, a key principle supported by both previous research with adult learners, and by results of the present study with pri-

mary school children, is that opportunities for active engagement and learner-generated contexts are likely key factors in successful vocabulary acquisition in intentional learning environments for bilingual learners.

As discussed in Section 6.9.1, the ability to give a definition requires decontextualised and abstract language which may be challenging particularly for children with lower levels of vocabulary knowledge. Even though some children could not provide such abstract definitions, they were, nevertheless, able to use target words to a well-specified degree in context. For example, for *thrilled*, one child offered the definition ‘like if someone had a surprise birthday party [. . .] they would be really thrilled’ and the sentence ‘my sister was thrilled when she found out that she had a surprise birthday party’. Despite not providing an abstract definition (e.g. ‘extremely happy about something; a sudden sensation of pleasure or delight’), this response indicated an appropriate understanding of a real-world context in which someone may be *thrilled*. Furthermore, some children relied more heavily on background knowledge than others, particularly cases BA, FA, and MT. Importantly, this allowed them to gain credit for their receptive word knowledge where otherwise they may not have had the opportunity to do so (for example, in a more restricted scoring rubric such as that of the WISC VC). Such scoring criteria may be particularly relevant to populations of children learning EAL who may be disadvantaged in traditional assessments of vocabulary depth due to possessing a lower stock of word knowledge than their monolingual peers.

Allowing credit for appropriate use of vocabulary within context is contingent with the view that assessments of linguistic knowledge should “treat performance on a language test as a particular instance of language use” (Bachman & Palmer 1996, p.10); in other words, the view that language assessment should be as closely related to actual language use as possible. Indeed, sentence- and composition-writing are very commonly occurring classroom activities, and thus it is important for EAL learners to be able to utilise their newly acquired word knowledge in context. Future research may assess the extent to which participation in sentence-writing activities for the purposes of word learning promotes children’s confidence in knowledge of learnt words specifically, as well as their continuing use of newly acquired vocabulary in context more generally, which may in turn promote depth and consolidation of this knowledge. Further work may also explicitly assess the predictions of the involvement load hypothesis with EAL learners in the context of vocabulary intervention by manipulating the amount of load incurred by different activities across groups of learners, and comparing acquisition and retention of novel vocabulary. Given that the construct of involvement load is said not to be modality specific (Laufer & Hulstijn, 2001), activities with high load (particularly, strong *evaluation*), could be implemented in educational settings before children acquire literacy. As EAL learners have been shown to possess lower levels of oral language and vocabulary knowledge than their monolingual peers even from school entry (Basit et al., 2015; Bowyer-Crane et al., 2017; Gregory, 1996; Hirst, 1998; Parke & Drury, 2001; Strand et al., 2015), one pertinent educational strategy for promoting the English language proficiency of EAL learners may be to provide exposure to a rich linguistic environment with opportunities to engage with language structures to an extent that promotes both receptive and productive knowledge.

7.3.3 Individual Growth Trajectories

The third research question concerned what insights could be drawn about the efficacy of the intervention by examining children's individual growth trajectories. Firstly, there was a tendency for children with a low level of knowledge at baseline to remain low by posttest, relative to their peers also taking part in the intervention; that is, children tended to make similar rates of progress over time regardless of their starting positions. However, exceptions to this pattern were the trajectories of two children, JG and BA, who despite having some knowledge of the to-be-taught words in the baseline period, failed to make as high a rate of progress as other children with similar or even lower levels of knowledge.

Further investigation suggested that JG and BA may not have had sufficient levels of English oral language knowledge to benefit from the intervention instruction. This conclusion is supported by the very low scores of the two children on standardised measures of English vocabulary knowledge and expressive grammar skill (administered at t1 of the longitudinal study), and to some extent by their reliance on background knowledge and relatively low scores in the *lexical knowledge* category of the bespoke word knowledge assessment. Although BA and JG were not the consistently lowest-scoring children on these measures, they did obtain low scores *across* all measures, in contrast to other low-scoring children who tended to score within the average range on at least one measure, and who therefore were able to compensate for low levels of knowledge or skill in other domains.

This finding correlates with studies of interventions with very young EAL learners in England which suggest that the English language skills of participating children may have limited their ability to engage with the material (e.g. Dockrell et al., 2010; Schaefer et al., *under review*). Indeed, prior knowledge has also been found to be a significant correlate of children's progress in vocabulary interventions; in particular, studies find evidence of the predictive role of background knowledge (Elleman et al., 2017), general verbal aptitude (McKeown, 1985), reading ability (Cain et al., 2004; Swanborn & de Glopper, 1999), and prior vocabulary knowledge (Wilkinson & Houston-Price, 2013; Webb & Chang, 2015) on acquisition of novel vocabulary. Although extreme caution is warranted in the interpretation of current findings due to the small sample size, analysis of individual trajectories in this study provides a tentative suggestion that the nature of vocabulary intervention may need to be adjusted depending on children's prior lexical knowledge. For example, a child identified as having low general target language vocabulary knowledge may be better placed in an intervention offering coverage of simpler Tier-2 vocabulary, or alternatively, words considered to be within the Tier-1 category. This point speaks to a difficulty in the use of Beck et al.'s (2002) tiered framework, as boundaries between Tiers are not clearly demarcated; alternatively, such Tiers may be leveraged as a rough guide or starting point when selecting vocabulary to teach, but such a decision should also be combined with input from teachers and children (i.e. in terms of which words they already know).

Words within the Tier-2 category are said to require some basic understanding of the concepts they represent, for example, already knowing a simpler synonym, such as *sleep* for *slumber* (Beck et al., 2005). Accordingly, one reason why cases BA and JG may have failed to benefit from the intervention teaching is because they lacked the more general conceptual knowledge required as a foundation for acquisition of new Tier-2 words. As an illustration, most children showed some

general level of understanding at baseline and pretest that *disaster* related to a negative event, whereas BA and JG did not possess even this basic foundational concept, offering 'you are so happy' and 'powerful' as definitions. Interestingly, BA and JG also mistakenly used *disaster* as a verb when asked to use this word within a sentence, in contrast to all other children who used it correctly as a noun. At the same time, however, it must also be pointed out that a number of children, including BA and JG, appeared to possess no knowledge of certain taught words prior to intervention teaching, but nevertheless made progress in their receptive and productive knowledge, and therefore lack of conceptual knowledge may explain lack of progress only up to a point. In other words, in some cases, children may indeed have possessed foundational or conceptual knowledge, but not the particular linguistic 'label'; for example, for taught word *purchase*, the majority of children referred to the synonym *buy* in their definitions and so in this case, *purchase* did not require a large conceptual leap, but merely represented a more 'sophisticated' word.

In summary, the adoption of a multiple case-series analysis provided opportunities for in-depth examination of patterns in children's vocabulary acquisition. Future studies may take further advantage of such a design by following individual learners over longer periods of time, both before and after intervention in order to observe longer-term maturational trends in targeted and general vocabulary. Particularly, such studies may assess the extent to which ongoing opportunities for use of novel vocabulary in context lead to retention and consolidation of knowledge.

7.3.4 Transfer of Teaching onto Standardised Assessments

The possibility of transfer of intervention teaching onto generalised language skills was assessed through the additional administration of standardised assessments at baseline. There was evidence for a significantly faster rate of improvement in children's expressive grammar skills (CELF FS) between baseline and posttest. The extent to which this improvement represents transfer from the intervention teaching is uncertain, however, given that CELF FS could not be administered at pretest, meaning that progress may have occurred in the 1.3-month period between baseline and pretest. Moreover, improvements in CELF FS score during this period must be interpreted with reference to a trend found in the longitudinal study for children's expressive grammar skills to improve over time as a result of regular classroom instruction, and for EAL learners particularly to make fewer grammar errors over time in their speech and writing.

On the other hand, it is possible that the explicit training children received in sentence judgement and construction during the intervention may have accounted for their significantly faster rate of progress between baseline and posttest, relative to that between t1 and baseline. At this point, comparison with relevant literature is difficult as many intervention studies do not explicitly examine transfer to generalised language skills, including among samples of EAL learners in England. However, there is evidence for such transfer of vocabulary instruction for monolingual learners onto not only generalised vocabulary knowledge, but also wider skills such as reading comprehension, as provided by the meta-analytical study of Elleman et al. (2009) discussed in Chapter 5. Additionally, there is some evidence to suggest that training in sentence construction can lead to improvements in wider skills such as writing composition (Viel-Ruma, Houchins, Jolivet, Frederick & Gama, 2010; Datchuk, 2017), although such work has not been carried out with primary school-age EAL learners.

7.3. Discussion of Results

Where significant improvements in generalised language skills were *not* found (i.e. for vocabulary depth knowledge; WISC VC subtest), this may have been due to two factors. Firstly, stimuli in the WISC VC subtest were not directly targeted by intervention teaching, and secondly, where improvements did occur, the WISC VC subtest may not have been sensitive enough to detect changes in children's word knowledge. Arguably, however, sensitivity is unlikely to have accounted for lack of transfer, since a modified version of the WISC scoring rubric was found to yield very similar results to the original scoring rubric of the manual when applied to WISC VC performance in the longitudinal study (see Section 4.5.2.1). Additionally, the reader is reminded that, while 5.5 months passed between t1 and baseline, the period of time between baseline and posttest was shorter, at only 3.1 months; in other words, lack of significant gains in vocabulary depth may also have been due to the relatively shorter time interval.

As well as participants' continually improving expressive grammar skills, the vocabulary teaching they received may have served to promote their 'word consciousness', a concept defined as interest in and sensitivity to "nuance of meaning" (Anderson & Nagy, 1993, p.10). This, in combination with the brief training received from coordinators in sentence construction may have led children to approach the CELF FS task with heightened sensitivity towards grammatical properties of stimulus words, or to sentence construction more generally. To some extent, a similar process may have applied to children's definitions and usage of untaught words.

7.3.5 Conclusions

The results of the intervention study suggest that a short, low-intensity, vocabulary teaching intervention was effective in promoting the Tier-2 vocabulary knowledge of a small group of EAL learners with English vocabulary weaknesses. After participation in a multicomponent vocabulary teaching programme delivered by speech and language therapy students, children showed significant gains in both their receptive and productive knowledge after an average of 9.4 sessions of one-to-one instruction, and largely maintained this knowledge at 6-month maintenance. Importantly, the implementation of a bespoke word knowledge assessment allowed children to gain credit for their word knowledge without the necessity of providing definitions, and this assessment also indicated gains in other aspects of word knowledge such as grammatical function. The two children who failed to make as high a rate of gain in knowledge as their peers appeared to possess generally low English language proficiency, suggesting that after four years of English-medium instruction, such an intervention may not be equally effective for all EAL learners.

Chapter 8

General Discussion

The aims of this thesis were twofold: firstly, to follow the language and literacy development of children learning EAL and their monolingual peers in primary school, and secondly, to evaluate the efficacy of a short-term vocabulary intervention for a subgroup of EAL learners from the longitudinal study. This final chapter will discuss the results of the two studies, highlighting overarching patterns and contribution of findings to the existing literature. The final two sections of this chapter will then consider strengths, limitations and future directions, as well as educational implications of findings.

8.1 Language and Literacy Development in Children Learning EAL and their Monolingual Peers in Primary School

Primary school Year 4 is part of an important transition phase in the literacy instruction of primary school pupils, as focus of the National Curriculum shifts to the extraction and evaluation of meaning from text (DfE, 2013). Given the role of vocabulary and syntactic knowledge in linguistic comprehension, any child who does not possess sufficient oral language proficiency will therefore be placed at risk of underachievement in national high-stakes Key Stage 2 reading and writing assessments at age 10-11 and potentially beyond. Although characterised by great heterogeneity, a number of EAL learners in England are found to enter formal education with lower English language proficiency than their monolingual peers. This thesis provides evidence for the continuing need for additional, explicit support for EAL pupils in aspects of English oral language skills even after four years of English-medium instruction, but also proof of concept for one method of achieving this end.

The following sections will begin by discussing the extent to which gaps in language and literacy performance converged or diverged between EAL learners and their monolingual peers in the longitudinal study, before considering the efficacy of the vocabulary intervention study.

8.1.1 Schooling Closes Some but not All Gaps

The longitudinal cohort study compared the language and literacy skills of a group of EAL learners against those of their monolingual peers between the beginning of Year 4 and the middle of Year 5. Having attended English-medium formal education since at least Year 1, language questionnaire data revealed that the majority of EAL learners were being exposed regularly to English outside of school through family, friends, and media, and that the children themselves tended to express a preference for communicating in English and to perceive themselves as more highly

skilled in English than the home language. At t1 (the beginning of Year 4), children in the EAL group exhibited a profile of strengths and weaknesses relative to their monolingual peers which was broadly commensurate with both international and U.K.-based research (August & Shanahan, 2008; Babayiğit, 2014a; Burgoyne et al., 2009; Cline & Shamsi, 2000; Cameron & Besser, 2004; Geva & Farnia, 2012; Hutchinson et al., 2003; Oller et al., 2007). Specifically, the strengths of the EAL group were manifested in measures which required the rapid naming of letters and digits (CTOPP RLN and RDN), and accurate sight-word reading and nonword decoding (TOWRE). Additionally, EAL learners also made fewer spelling errors in their writing. On the other hand, weaknesses of the EAL group were found across both oral language and literacy domains, including in receptive (BPVS) and expressive vocabulary (CELF EV), expressive grammar (the 17 consistently attempted items on the CELF FS; morphosyntactic error rate in oral narrative task), spoonerisms (PhAB), mean length of utterance in words (Peter and the Cat oral narrative retell task), passage reading accuracy and comprehension (YARC; raw scores only), and syntactic error rate in writing (bespoke task). One interesting finding was the degree of parallelism between oral and written narrative skills, wherein the EAL group produced significantly more utterances and T-units, but of a shorter length and higher rate of syntactic errors than those of their monolingual peers. Interestingly, mean length of utterance differed significantly between the two groups only in oral and not in written narrative, suggesting that the oral narrative task may have been a more sensitive measure of children's expressive syntax.

This developmental picture was not entirely unchanged by the end of the longitudinal study 18 months later (t3; middle of Year 5). Slopes derived from linear mixed models revealed that EAL learners were able to catch up to their monolingual peers on some but not all measures. Specifically, the steeper developmental trajectories of the EAL group in receptive and expressive vocabulary breadth, spoonerisms, and passage reading rate were not *sufficiently* steep in order to close gaps in performance by t3, but the EAL group did show convergence with the monolingual group in a number of other measures. In some cases, this convergence was due to a plateauing or deceleration of monolingual learners (e.g. in listening comprehension (CELF USP) and passage reading accuracy). In other cases, however, convergence was achieved due to the steeper developmental trajectory in the EAL group alone; this pattern was particularly pronounced in the case of vocabulary depth (WISC VC), but was also evident to some extent in expressive grammar, oral narrative utterance error rate, passage reading rate and comprehension, and morphosyntactic error rate in writing. Again, this pattern of developmental trajectories was broadly in line with previous longitudinal work in England indicating that, where observed, monolingual-EAL group differences tend to remain broadly in place over time (Burgoyne et al., 2009, 2011a; Hutchinson et al., 2003), a finding also reported in the international literature (e.g. Droop & Verhoeven, 2003; Lervåg & Aukrust, 2010).

Divergence between the two groups was also observed in some domains, but to a far lesser extent than that reported in other studies of EAL learners in English primary schools (e.g. Burgoyne et al., 2009; 2011; Hutchinson et al., 2003). Particularly, where EAL learners exhibited strengths, they also tended to diverge from their monolingual peers in their trajectories; this pattern was observed for rapid naming of digits as well as sight-word reading and nonword decoding. Divergence was also observed in measures of writing: firstly, monolingual children constructed increasingly long sentences across the course of the study (at almost double the rate of their EAL

8.1. Language and Literacy Development in Children Learning EAL

learning peers); and secondly, although the monolingual group made a higher rate of spelling errors at t1, this rate dropped quickly over time, to just under that of the EAL group by t3. Writing error analysis revealed that, despite the relative strengths of EAL learners in spelling, these children were far more likely than their monolingual peers to make errors in the use of formal lexical devices, particularly concerning the appropriate use of determiners, copulas, and prepositions. This is supported by Cameron and Besser's (2004) large-scale writing analysis which similarly identified such errors as highly distinctive of the writing of EAL learners. In the present study, it is interesting to note that errors in the use of formal lexical devices were rarely occurring in the oral narrative task, but substantially pronounced in writing, again suggestive of the different sensitivity of the two tasks to different aspects of linguistic skill.

The longitudinal study also revealed a number of novel and unexpected findings in relation to the literature. For example, where bilingual learners exhibit lower levels of target language vocabulary in relation to their monolingual peers, this has been found to apply to both breadth and depth of knowledge (Lervåg & Aukrust, 2010; Vermeer, 2001); however, this pattern did not apply in the present study, in which EAL learners significantly underperformed in relation to their monolingual peers only in vocabulary breadth knowledge. This represents a novel finding in U.K.-based literature in which vocabulary depth has not typically been assessed. The role of depth and breadth in vocabulary instruction is discussed in Section 8.3.

A second unexpected finding was children's passage reading performance: particularly, children in the EAL group read passages significantly less accurately than their monolingual peers, yet exhibited very similar levels of reading comprehension skill. This contrasts sharply with previous research indicating exactly the opposite pattern (i.e. relatively higher reading accuracy but lower comprehension; Babayiğit, 2014a; Burgoyne et al., 2009, 2011; Hutchinson et al., 2003), although it should be noted that analysis of raw scores on passage 3 alone did indicate a significant monolingual group advantage in reading comprehension. While non-significant group differences for standardised scores in reading comprehension may be due in part to the relatively small amount of reading comprehension skill sampled by the YARC (discussed further in Section 8.2.2), it was also the case that the vocabulary knowledge and listening comprehension skill of the two groups did not differ greatly in an absolute sense, potentially accounting for this pattern to some extent. Similarly, the two groups did not differ to a large degree in number of passage reading accuracy errors, and it is likely that EAL learners were able to read enough of the sample passages in order to answer comprehension questions.

Two factors were particularly important in the interpretation and comparison of children's developmental trajectories in the present study. Firstly, the magnitude of group differences at t1 was rather reduced relative to other studies of EAL learners and their monolingual peers in England (e.g. Babayiğit, 2014a; Burgoyne et al., 2009, 2011; Hutchinson et al., 2003), and even where the two groups did differ significantly, this was often only to a small degree in an absolute sense. As a result, the scope for convergence between the two groups was relatively restricted, as reflected by slope terms in linear mixed models (see Table 4.21 on page 140 for a summary of slope terms across all models). Secondly, and most importantly, results must take account of the representativeness of the monolingual comparison group. Although data pertaining to socio-economic status were not available for individual participants, all eight participating schools had higher than average proportions of pupils eligible for free school meals (FSM) and were situated in neighbour-

hoods of higher-than-average levels of deprivation affecting children (Table 3.1). Indeed, this is consistent with the national picture in which ethnolinguistic communities, and by extension, many EAL learners, are situated in neighbourhoods of higher-than-average social deprivation (Casey, 2016). This is likely to have had an impact on results of the longitudinal study, as socio-economic deprivation is a well-established predictor of children's language development and educational attainment (Clegg & Ginsbourg, 2006; Hoff, 2003). Interestingly, in their analysis of National Pupil Database data, Strand et al. (2015) found FSM to have a relatively larger impact on the attainment of *non*-EAL learners; if this is the case, one may expect the performance of the monolingual group in the present study to have been relatively more depressed as a result of deprivation, resulting in smaller gaps in attainment with EAL learning peers (see also comparison with the results of Babayiğit, 2014a in Section 4.3.2). In other words, had participants been situated in neighbourhoods of average or lower-than-average levels of deprivation, discrepancies in performance between the two groups may have been larger in magnitude due to the monolingual group scoring more highly. In order to obtain a more rounded picture of development, future studies in England may seek to recruit participants from a wider range of backgrounds (e.g. Droop and Verhoeven, 2003; Oller & Eilers, 2002) in order to explicitly examine the effects of neighbourhood deprivation on the development of language and literacy skills in children learning EAL.

Reference to standardised scores, although questionable in bilingual populations, made an important contribution to the interpretation of EAL-monolingual group differences in the present study. By way of example, the EAL group mean for receptive vocabulary knowledge (BPVS) at t1 of 80.13 may have appeared particularly low in isolation. Comparison of this score with a population norming mean of 100 would suggest a large discrepancy of just under 20 standard scores; however, the monolingual group's mean of 87.80 at t1 appeared to show closer similarity between the monolingual and EAL group, perhaps because the monolingual comparison group was not representative of the national monolingual population (despite the fact that these monolingual children did score within the average range of ± 1 SD). However, this was not the case for all measures, for instance in rapid naming, single-word reading efficiency, and passage reading, in which both groups were performing relatively higher (or above norming population means). Standardised scores may therefore represent an additional tool with which to measure the English language and literacy development of EAL learners by providing information as to the relative standing of monolingual children against whom EAL learners are often compared.

In terms of theoretical applications, the results of the longitudinal study generally conform to predictions made by Paris's (2005) constraints on reading framework, in which the developmental trajectories of language and literacy skills differ according to their *scope*, *mastery*, *universality*, and *codependency* (see Section 2.2.2). This framework is particularly relevant to the study of language and literacy development in bilingualism, in which amount of target language linguistic input is necessarily less than 100 percent, having implications for scope and mastery aspects of skills particularly. Indeed, results of the present study lend support to this framework, as the most consistent discrepancies between the groups were found in the more unconstrained skills of vocabulary and syntactic knowledge, which depend to a more important degree on exposure to the target language. In contrast, EAL learners exhibited strengths in the more constrained skills of rapid naming and accurate decoding, which involve acquisition and mastery of a finite scope of knowledge. Vocabulary knowledge, particularly, represented a persistent weakness of EAL

learners, which may be interpreted by Paris's (2005) framework as owing to the fact that word knowledge is vast in its scope and continues to be acquired throughout the lifespan. As such, vocabulary is more highly dependent on continuing linguistic input, unlike letter knowledge and skills dependent on this, such as rapid naming and single-word reading efficiency, which involve finite scope and require complete mastery.

8.1.2 Developmental Trajectories can be Altered with Targeted Instruction

Due to split exposure between languages, it is common for EAL learners to begin formal schooling with lower levels of English language proficiency than their monolingual peers (see Section 1.3.4). Even in the case of regular language learning progress, then, these children are still likely to underperform in relation to their monolingual peers over time unless they make a faster *rate* of progress (NALDIC, 1999). Indeed, the results of the longitudinal study confirmed that that EAL learners were unlikely to catch up to their monolingual peers, particularly in vocabulary knowledge, through engagement in ordinary classroom instruction alone.

As a result, the intervention specifically targeted EAL learners' vocabulary knowledge in an effort to alter word-learning trajectories. It aimed to promote receptive and productive knowledge of a set of 20 words within the Tier-2 category (Beck et al., 2002; Section 5.3.1). The design and multiple testing points of the intervention study allowed comparison of vocabulary acquisition trajectories in taught words before and after receipt of explicit instruction, taking account of maturation effects prior to teaching, as well as extent of retention of learned vocabulary after teaching. Children made little progress in taught vocabulary in the baseline period preceding the teaching, but made statistically significant, moderate-to-large improvements immediately after the 10-week programme which were largely retained six months later. Thus, the vocabulary acquisition trajectories of the nine EAL learners in the intervention were explicitly targeted and altered, resulting in significant gains in receptive and productive knowledge of target words. Although significant gains were also found in untaught vocabulary, this may have been due to two particularly high-scoring children, as well as specific improvements in sentence score, interpreted within a trend for general improvements in children's expressive grammar skills over time. Children also made significant improvements on generalised expressive grammar skill (CELF FS) but not vocabulary depth knowledge (WISC VC) between baseline and posttest, however interpretation of transfer effects is limited by lack of administration of these measures at pretest.

The results of the intervention study have significance for three reasons. Firstly, there is currently a paucity of research on effective pedagogical practices for EAL learners, particularly in England. In their review of international language and literacy intervention research with bilingual learners, Murphy and Unthiah (2015) identify vocabulary as a good candidate for intervention, given that EAL learners often experience difficulties in this domain, and that vocabulary knowledge is generally responsive to intervention. At the same time, however, the review discovered very few studies that explicitly evaluated language, or specifically vocabulary, teaching programmes with bilingual learners outside of the U.S. Studies conducted in England are few, and tend to focus on very young EAL learners around the onset of formal education (Dockrell et al., 2010; Kotler et al., 2001; Schaefer et al., *under review*; St. John & Vance 2014). The present study represents an important contribution to this small literature base by focusing specifically on vocabulary knowl-

edge, and provides evidence for the efficacy of such intervention even among EAL learners in a later phase of primary education, after four years of English-medium instruction.

Secondly, and related to the above, the results of the intervention study provide proof of concept for the efficacy of explicit and targeted Tier-2 vocabulary instruction for EAL learners in KS2. While extrapolation of results is limited by the small sample size of the study ($n=9$), strengths included the creation of a bespoke word knowledge assessment and the measurement of children's progress in a parallel list of *untaught* words. The longitudinal cohort study showed that EAL learners do not catch up to their monolingual peers in the breadth of their receptive and expressive vocabulary knowledge through classroom instruction: given significant and enduring gains in children's target vocabulary knowledge after participation in the present intervention programme, the methods employed here - namely, passage reading, sentence-level work, and mind maps - may represent one effective strategy for closing this gap. Additionally, results here extend the work of EAL oral language interventions with younger children, according with recommendations for active engagement with new vocabulary (Dockrell et al., 2010; Gersten & Baker, 2000), but also making the additional recommendation of embedding novel words within meaningful contexts - for example, children often referred to the stimulus story contexts and background or related knowledge when giving definitions and constructing sentences in the bespoke word knowledge assessment. Considering that EAL learners often have smaller vocabularies in English, the provision of engaging and relatable contextual information may represent a scaffold for the acquisition of novel words.

In terms of theoretical contributions, the results of the intervention also provide some support for the task-induced involvement load hypothesis (Laufer & Hulstijn, 2001), which has not been assessed in bilingual children. In line with predictions of the hypothesis, activities encouraging depth of processing and specifically *evaluation* (the combination of novel vocabulary with prior knowledge in sentence-writing tasks) appeared to result in successful vocabulary acquisition. The present study offers some support for the application of involvement load to the design of vocabulary teaching activities for children learning EAL, and future studies may explicitly contrast, for example, sentence-writing and composition-writing in vocabulary learning. Although composition-writing was not a feasible strategy in the present study due to timing constraints and children's writing fluency, the effect of involvement load is proposed to apply independently of modality (Laufer & Hulstijn, 2001), and thus one strategy could be to encourage children to utilise novel vocabulary orally through stories. In addition, such an activity may provide opportunities to revisit and utilise previously-covered vocabulary, in turn encouraging further depth of processing through the combination of target words.

The third and final reason for the significance of these results is further insight into the role of general English language proficiency in EAL learners' ability to benefit from Tier-2 vocabulary instruction. The use of a multiple case series design revealed that two of the intervention participants, cases BA and JG, failed to benefit from the intervention teaching, despite the availability of one-to-one instruction from speech and language therapy students. Although the levels of target word knowledge that these two children possessed prior to teaching were similar to those of other participants, they failed to make appreciable progress across the ten weeks of teaching. Tier-2 vocabulary was chosen specifically for its potential utility to learners in KS2, a period in which instructional focus shifts from lower- to higher-level reading skills, and vocabulary in particular (DfE,

8.2. Strengths, Limitations and Future Directions

2013; target word selection is discussed further in Section 8.2.5). While this strategy appeared to be appropriate for the majority of participants, the generally low English language proficiency of cases BA and JG may have inhibited their progress, and thus acquisition of such Tier-2 vocabulary may not be an appropriate goal for all EAL learners, even after four years of education. The role of English language proficiency has been suggested to play a part in the success of intervention teaching in studies of young EAL learners (Dockrell et al., 2010; Schaefer et al., *under review*). Therefore, it would appear that low English language proficiency may also be a barrier to vocabulary learning in older EAL learners, as indicated in the multiple case series analysis of the present study.

8.2 Strengths, Limitations and Future Directions

This section will consider the strengths and limitations of the longitudinal and intervention study, including design and statistical framework, choice of measures, participant characteristics, and selection of target words.

8.2.1 Design and Statistical Framework

Although developmental trajectories can be inferred from cross-sectional designs, the longitudinal nature of the present study with its repeated measurements on the *same* individuals is advantageous in controlling for cohort effects and within-subject variation (Singer & Willett, 2003). While two time points are theoretically sufficient for the study of change over time, the inclusion of a third time point likely allowed more reliable estimates of trajectories and modelling of change using the linear mixed models framework discussed below (Willett, 1989). It is noteworthy that some developmental trajectories (e.g. receptive and expressive vocabulary, rapid naming of digits and letters) showed non-linear trends (Appendix 4.2, page 262). It is a limitation of the study that statistical models were insufficiently powered to detect such non-linear trajectories, i.e. through the inclusion of quadratic terms which require a minimum of four time points (Law et al., 2008; Singer & Willett, 2003). Indeed, it is noteworthy that in some cases, convergence between the two groups of children was due to deceleration of the monolingual group. Future studies sufficiently powered to detect non-linear trajectories may also seek to model the growth of the two groups over a longer period of time or to include additional testing points, as the extent and timing of acceleration or deceleration in either group may have implications for the timing and intensity of intervention. For instance, in their study of monolingual Greek and bilingual Albanian-Greek 6 to 9 year-olds, Simos et al. (2014) reported a trend for the deceleration of vocabulary acquisition over time. If a similar pattern were to be found in populations of EAL learners, such a finding could potentially influence decisions around the timing of the introduction of explicit vocabulary instruction; that is, if EAL learners are shown to acquire less vocabulary in later educational stages, intervention may be particularly effective at this stage.

Longitudinal data present challenges for the statistical modelling of growth. The linear mixed modelling (LMM) framework applied in the present study represented an advantage over traditional statistical methods such as repeated measures ANOVA by explicitly modelling dependency between data points, as opposed to applying post-hoc correction procedures (e.g. Greenhouse

& Geisser, 1959; Huynh & Feldt, 1976). Such a strategy is found to result in more efficient model parameter estimation and smaller standard errors (Burton et al., 1997; Osborne, 2008; West et al., 2007). The step-up model building strategy (described in Table 4.9) ensured parsimony and prevented model over-fitting; particularly, *group* \times *time* interaction terms did not add meaningfully to the majority of models and thus were dropped, thereby preserving degrees of freedom (note, however, that comparison of trajectories was still made possible by estimation of group intercepts and slopes). Finally, LMM lent itself well to the unequal sizes of the EAL ($n=48$) and monolingual group ($n=33$), and incorporated missing data without resorting to list-wise deletion of subjects (West et al., 2007). The decision to apply LMM in the present study represents a methodological departure from previous EAL research in England. However, with the advent of freely available software such as R and its `lme4` package, and an increasing realisation of the power and flexibility of LMM, this framework is beginning to gain traction in the psychological and second-language testing literature (Cunnings, 2012; Magezi, 2015).

8.2.2 Choice of Measures

The assessments used in the present study to measure constructs of interest in language and literacy development lend validity to results by virtue of being standardised, norm-referenced (in most cases), and by affording comparison with similar studies of EAL learners in England (e.g. Babayiğit, 2014a; Bowyer-Crane et al., 2017; Schaefer et al., *under review*). The relative focus of the test battery on language skills is justified by studies finding oral language, and particularly vocabulary, to be a domain of persistent weakness for EAL learners (see Section 2.1.2.3). The inclusion of multiple measures of vocabulary was advantageous for two reasons. Firstly, this allowed investigation of growth in vocabulary breadth as well as depth, a variable not commonly included in test batteries in EAL populations in England. Secondly, multiple vocabulary measures likely resulted in additional accuracy when selecting participants for the intervention (indeed, other intervention studies often utilise composite variables for this purpose, e.g. Bowyer-Crane et al., 2008, 2017; Fricke et al., 2013). Indeed, it was interesting to note that some intervention participants exhibited uneven profiles of vocabulary skill. For example, one child obtained a scaled score of 8 on the WISC VC, which by itself, would have disqualified her as a potential intervention participant. However, this child also failed to obtain a standard score on the BPVS (i.e. lower than the minimum possible score of 70), suggesting a considerable receptive vocabulary weakness. Therefore, use of multiple vocabulary measures was advantageous in identifying children likely to benefit from intervention teaching.

Future work may incorporate additional measures of vocabulary such as the Levels Test (Nation, 1990) in which vocabulary is grouped into frequency bands. The use of this measure may inform selection of target words for vocabulary instruction as, for instance, if EAL learners are shown to lack vocabulary above a certain frequency level, words with lower frequency of occurrence may then be selected for instruction; conversely, if EAL learners do not appear to lack vocabulary below the 2,000-word frequency band, such vocabulary may be filtered out in order to focus on less frequently occurring words. As discussed in Section 2.1.2.3.3, there has been some work in England examining multi-word phrase (MWP) vocabulary (e.g. *break the ice*) in EAL learners. In their cross-sectional study of Year 3, 4, and 5 pupils, Smith and Murphy (2014)

8.2. Strengths, Limitations and Future Directions

found significant monolingual-EAL group differences in MWP knowledge only from Year 4: future longitudinal work may provide further insight into the developmental trajectory of MWP knowledge and for example, the extent to which this aligns with general receptive and expressive vocabulary breadth.

One methodological limitation pertained to the scoring criteria of assessments in the longitudinal study, particularly the vocabulary definitions (VC) subtest of the WISC, and the Peter and the Cat oral narrative measure. Firstly, although the WISC VC measures depth of vocabulary, its relatively limited scoring range of 0-2 did not provide a great deal of information about children's word knowledge. For example, according to scoring criteria, a full definition and a synonym both merit the maximum score of 2, and credit is not awarded for expressive knowledge or other constraints on use within context (Nation, 2001; Wechsler, 2003). As discussed in Section 6.9.1, this limitation is overcome by alternative scoring rubrics in which receptive and expressive knowledge, definitions, synonyms, and real-life examples all receive credit. Indeed, this was the approach applied in the creation of the bespoke word knowledge assessment for the vocabulary intervention study, which revealed that some children possessed impartial or incorrect productive knowledge of target vocabulary which would not have been evident in a purely receptive assessment. While the inclusion of a measure of vocabulary depth in the present study represents an advantage in itself (as this aspect of vocabulary has not been assessed among samples of EAL learners in England), future studies may wish to adopt alternative measures of vocabulary depth which yield more detailed information, such as word association networks or structured interviews (Verhallen & Schoonen, 1998; Vermeer, 2001). Secondly, the Peter and the Cat oral narrative assessment was administered in order to elicit children's narrative retelling capabilities and expressive grammar. As discussed briefly in Section 2.1.3.2, children's oral narrative retells are typically examined at two levels of analysis, namely microstructure (syntax and cohesion) and macrostructure (story structure and coherence). However, due to the imprecise macrostructure scoring criteria of the Peter and the Cat manual, and concerns relating to reliability of scoring, the decision was made not to include this element of children's narrative skills in the study. One improvement may have been to administer a bespoke oral narrative measure with more clearly defined scoring criteria (see Nielsen, Dixon & Fricke, *in preparation*), or the creation of a bespoke macrostructure scoring rubric for the Peter and the Cat assessment.

As discussed above, one unexpected finding of the longitudinal study was the lack of any large and significant discrepancy in the reading comprehension performance of the two groups of children on the YARC, a result which stood in opposition to other studies of EAL learners in England which typically administer the NARA (e.g. Burgoyne et al., 2009; Hutchinson et al., 2003). This discrepancy may be understood with reference to the differing administration procedures of the two assessments. In particular, while the NARA requires examinees to read all stimulus passages until a pre-specified number of errors is made, scores on the YARC are derived from only the *two highest* passages attempted. Therefore, the two assessments differ critically in the range of reading material that children attempt, with the YARC being a potentially less sensitive measure of reading comprehension skill (Colenbrander et al., 2017). This relates to a further limitation with the YARC in that different children read different subsets of passages, limiting direct comparison across groups. However, this issue was mitigated to some extent by stipulating that

all children read the same subsets of passages from t2 onwards, and also by the use of standard scores.

The inclusion of only one measure of passage reading skill may be seen as a limitation, as different measures have been shown to place differing demands on children's decoding and comprehension skills (Cain & Oakhill, 2006; Colenbrander et al., 2017; Cutting & Scarborough, 2006). This may represent a limitation in studies of EAL learners who often exhibit strengths in decoding, which may artificially inflate comprehension ability due to attempting more passage stimuli (Burgoyne et al., 2009; Cline & Shamsi, 2000; Hutchinson et al., 2003). Although not permitted by time and resource limitations, the administration of a second passage reading measure in addition to the YARC would have been advantageous, as passage reading skills (and particularly comprehension) are known to have lower test-retest reliability than single-word reading, for instance (GL Assessment, 2011; Torgesen et al., 1999). Therefore, an additional passage reading measure may have provided the opportunity for the creation of a composite passage reading variable, resulting in a more reliable measure of this construct (a strategy recommended particularly for EAL learners who show a great deal of variability in their skills; Cline & Shamsi, 2000).

Finally, a few points will be made concerning the parent and child language questionnaires (see Section 3.5 for a summary). Firstly, availability of questionnaire data was advantageous in confirming children's EAL status, with no participants in the monolingual group and all participants in the EAL group stating that they spoke a language other than English in the home (where data were available, these patterns were also confirmed by parental questionnaires). Secondly, questionnaires provided important contextual information relating to parental education, employment, and self-rated proficiency in English language and literacy. Most importantly, however, questionnaires provided information relating to children's language exposure in the home, revealing that, as well as having received an equal amount of English-medium instruction to their monolingual peers, the majority of EAL learners in the study had been born in the U.K., and frequently spoke or were being exposed to English in the home. Unfortunately, an overall response rate of only 70% resulted in incomplete home language data for some participants. Additionally, it is possible that only parents with sufficiently high English language proficiency may have responded to the questionnaire, as language proficiency has been linked with response rates in other multilingual populations (Kappelhof, 2013). This issue may be avoided in future studies through the translation of questionnaires into multiple languages or alternatively the availability of a translator. However, given the high degree of linguistic diversity among EAL learners in the present study ($n=13$ distinct home languages), such an undertaking would have proved prohibitively costly.

8.2.3 Selection Criteria and Retention of Intervention Participants in the Longitudinal Study

The longitudinal cohort study stipulated the important criterion that all children had to have been in receipt of formal English-medium education since at least primary school Year 1. This decision was taken firstly to ensure sufficient English language skills for children to access the language and literacy assessments in the test battery, and secondly to control for the effect of differing amounts of instruction in the analysis of group differences in performance. As a result of this recruitment criterion, EAL learners in the present study are unlikely to be wholly representative

8.2. Strengths, Limitations and Future Directions

of other, unselected samples of EAL learners elsewhere (e.g. Hutchinson et al., 2003), who may vary substantially in their length of residence in England or experience of English-medium instruction.

Children in the EAL group who received intervention teaching remained in the longitudinal study, and were administered the full test battery at t2 and t3. Extent of transfer from intervention teaching onto generalised language skills (expressive grammar, CELF FS; vocabulary depth, WISC VC) was assessed explicitly in Chapter 7. Analysis indicated some evidence for transfer onto expressive grammar, although no administration of standardised assessments at pretest limits this conclusion somewhat. No transfer onto vocabulary depth knowledge was observed. Outside of transfer onto these two skills, however, it may be questioned whether participation in the intervention influenced children's performance on other measures in the test battery. Fortunately, this was mitigated by the fact that the number of intervention participants was small (12 out of 48 children in the EAL group received intervention teaching), and taught vocabulary did not overlap with stimuli in any standardised measure. Re-analysis of linear mixed models without these 12 children altered the statistical significance of *group* coefficients only in receptive vocabulary (BPVS) and expressive vocabulary (CELF EV; i.e. coefficients were no longer statistically significant) - an expected pattern, given that these variables were utilised for selection purposes. Therefore, given no changes in other variables as a result of participation in the intervention, the decision was made to retain the 12 children in the longitudinal study. Additionally, participants recruited at t1 represented an unselected sample, and therefore it was of interest to continue to follow all children's developmental trajectories until the end of the study, as opposed to removing the subgroup of those who took part in the intervention.

8.2.4 Design and Analytical Strategy of the Vocabulary Intervention

While the longitudinal study benefitted from a repeated measures design and robust statistical framework, analysis of children's progress in the vocabulary intervention was limited by the small sample size of the study and lack of a control group. However, additional measures were taken in order to improve the robustness of results, including the addition of a list of untaught words, and the establishment of a baseline period. On the other hand, the small sample size was also advantageous in allowing in-depth investigation of children's individual growth trajectories through a multiple case series design (Chmiliar, 2012). Linkage between individual growth trajectories and performance on standardised measures led to the interesting finding that two of the children who made very little progress also possessed low levels of general English vocabulary knowledge and expressive grammar skill, which may have limited their ability to benefit from the one-to-one instruction. Future work may take this into consideration when planning intervention programmes; particularly, there may be a level of language proficiency below which Tier-2 vocabulary intervention is less likely to be successful.

Statistically, the small sample size resulted in a reliance on less powerful non-parametric procedures, and it is unfortunate that the linear mixed modelling (LMM) procedure described in Chapter 4 could not also have been applied to data from the intervention (see Silverman, 2007 for an application of this strategy in an intervention design). Future studies with sufficiently large sample sizes may allow further investigation into the predictors of EAL learners' progress in vo-

cabulary acquisition. For instance, using such a procedure to analyse results from a 12-week intervention programme with 8 to 11 year-old monolingual children, Elleman et al. (2017) found that gains in knowledge of to-be-taught words were significantly associated with word frequency and prior knowledge, but interestingly not general vocabulary knowledge, counter to patterns suggested by the multiple case series analysis here. A larger-scale evaluation of such a vocabulary intervention programme would be better-placed to investigate the influence of general English language proficiency on EAL learners' vocabulary knowledge gains.

Concerning choice and administration of measures of the intervention, the creation of a bespoke word knowledge assessment (Section 6.9.1) can be considered a strength. This measure captured a high level of detail in children's word knowledge, not only in receptive understanding but also in productive use. Additionally, while the BPVS and CELF EV subtest provided measures of children's knowledge of non-overlapping sets of words, the bespoke word knowledge assessment measured both aspects of knowledge in the same set of words. This provided insight into growth patterns not only in receptive and expressive knowledge of the same words, but also across various categories such as definitions, background knowledge, and so on (Appendix 7.5, page 286). This supports the use of bespoke measures in assessing the efficacy of intervention studies, especially for vocabulary teaching programmes which result in sometimes subtle changes in knowledge, or act on particular dimensions of knowledge (e.g. productive use). Indeed, given that expressive vocabulary knowledge has been found to be a relatively stronger predictor of reading comprehension in EAL learners than their monolingual peers (e.g. Hutchinson et al., 2003), measures which capture the ability to use vocabulary productively and meaningfully are important in the study of EAL learners' developmental language trajectories¹.

8.2.5 Selection of Target Words

The intervention study attempted to take a somewhat objective approach in the selection of target words by selecting words not only meeting Tier-2 criteria as described by Beck et al. (2002), but also by reference to statistics on each word's frequency per million words and typical age of acquisition (AoA; Kuperman et al., 2012). This strategy was a success for a number of reasons: firstly, according to feedback from coordinators, children were generally engaged and able to discuss their personal experiences using the target words, suggesting that the words were accessible to them; secondly, words represented a range of difficulty, and were neither all too easy nor all too difficult; and thirdly, the use of AoA and frequency metrics allowed more objective matching between taught and untaught words. However, there were also limitations to this approach to word selection. For example, the final list of words contained some items that did not easily align with children's experiences (e.g. *fraud*), or which all or most children had appeared to have already mastered to a high degree (e.g. *responsible*). One improvement to this strategy may have been a pretesting phase in which children are asked to rate their familiarity and understanding of a larger pool of word candidates, before selecting a final list of words for which they possessed no or only

¹Although not reported in-text, additional analyses of children's progress on each target word revealed different rates of progress in receptive and productive knowledge. For example, for the words *persuade* and *rescue*, receptive knowledge increased to a relatively larger degree than productive knowledge between pre- and posttest whereas, in contrast, the opposite pattern applied to *bargain* and *wealthy*; see Appendix 7.6 on page 288. Such an example serves to illustrate the utility of measuring both receptive and productive knowledge of the same words.

8.3. Educational Implications

partial knowledge (Biemiller & Boote, 2006; Wesche & Paribakht, 1996). Similarly, selection of target words according to curriculum relevance (e.g. Fricke et al., 2013) may have provided opportunities for children to build upon a foundation of knowledge gained through classroom teaching, although this would have required additional time for analysis of curriculum materials and discussion with class teachers. Finally, although not an issue unique to the present study, it is a limitation that it could not be known to what extent children were exposed to any of the taught or untaught words outside of the intervention teaching.

8.3 Educational Implications

As children learning EAL in England are educated in mainstream classroom settings, there is an expectation that they will acquire the same English language proficiency as their monolingual peers through school attendance and engagement with the curriculum alone (Cameron & Besser, 2004; Costley, 2014). Studies suggest that bilingual learners require a period of five to seven years to acquire the same level of CALP as their monolingual peers (Hakuta et al., 2000; Thomas & Collier, 2002), and indeed this has been found to be the case for EAL learners in England (Demie, 2013). Results here showed that while the two groups of learners generally resembled one another in their language and literacy skills after three to four years of English-medium education, EAL learners continued to experience significant and enduring weaknesses in breadth of English vocabulary knowledge. Thus, the primary educational implication of the present study is a need for sustained and high-quality vocabulary instruction of EAL learners in primary school, which may in principle be achieved through the instructional methods incorporated in the intervention study carried out here.

Contrary to the finding that EAL learners differed very little from their monolingual peers in the depth of their vocabulary knowledge, an instructional focus on depth may play an important role in enlarging the word stock of EAL learners. The intervention study supported the efficacy of vocabulary acquisition as a result of participation in activities which encourage depth of processing (personal experiences; using novel vocabulary productively), and which also provide opportunities for exposure to wider vocabulary through reading passages and completing mind-maps (specifically, mind maps encouraged depth of vocabulary knowledge through synonyms, antonyms, and related phrases and concepts). In other words, a focus on vocabulary depth may be more likely to expose learners to a wider breadth of word knowledge. Additionally, given that a focus on depth may also result in improvements in expressive knowledge, this approach may be particularly well-suited to children learning EAL, for whom expressive language is often shown to be an area of developmental need (Hutchinson et al., 2003; McKendry & Murphy, 2011).

The present study provides proof of concept for the efficacy of Robust Instruction vocabulary teaching methods (Beck et al., 2002) for children learning EAL in England. As an efficacy trial, the intervention was carried out under 'ideal circumstances' (O'Donnell, 2008), with a small number of participants receiving one-to-one instruction from trained speech and language therapy students. One-to-one delivery methods may be appropriate for determining which particular aspects of an intervention teaching programme are effective: for example, as discussed in Chapter 7, one method of capitalising upon the *evaluation* element of involvement load (Laufer & Hulstijn, 2001) may be through the combination of multiple target words in narratives. Such a strategy is yet to

be trialled with children learning EAL, and an intervention with one-to-one delivery may be best placed to determine its efficacy, as this working pattern would likely serve to reduce distractions and allow practitioners to give tailored feedback, potentially providing a more accurate estimate of efficacy.

Importantly, the findings of the present study also have applicability beyond one-to-one delivery: firstly, the vocabulary knowledge of EAL learners has also been effectively targeted in previous studies utilising small-group delivery (e.g. Dockrell et al., 2010; Schaefer et al., *under review*), and secondly, the activities completed by children in the intervention are not exclusively amenable to one-to-one working, as for example, passage reading, sentence-writing, and mind-maps could feasibly be completed individually or as part of pair- or group-work in a classroom. Additionally, there is precedent in the intervention literature for successful delivery of intervention material to EAL learners by teaching assistants (Dockrell et al., 2010; Schaefer et al., *under review*), supporting the notion that such an intervention programme could possibly be delivered by school staff as opposed to externally-sourced practitioners.

8.4 Conclusion

Children learning EAL often begin formal education with lower levels of English language proficiency than their monolingual peers. However, as these children are expected to master English through classroom teaching alone, they are typically afforded little or no opportunity for dedicated, explicit English language instruction. This study showed that after an equal amount of English-medium instruction, EAL-monolingual group differences in oral language and literacy skills appeared reduced in magnitude relative to previous work, although such a conclusion is critically informed by the representativeness of the monolingual group against whom EAL learners were compared. Given the view that an EAL learner is not 'two monolinguals in one person' (Grosjean, 1998), aligning the language and literacy skills of EAL learners relative to national, monolingual norms may be considered contentious. However, it is the case in the English educational system that all children are taught *in English* and sit the same high-stakes assessments *in English*. Therefore, it is of interest to what extent EAL learners resemble their monolingual peers in their English language and literacy skills in order to establish instructional need and investigate effective instructional strategies.

Although EAL learners showed little sign of catching up to their monolingual peers over time in most oral language measures, the Tier-2 word knowledge of a small group of EAL learners was shown to be responsive to dedicated, one-to-one instruction. That significant gains in both receptive and productive knowledge occurred as a result of the intervention teaching despite children's vocabulary weaknesses is supportive of the efficacy of robust vocabulary instruction for this population of learners. While future studies may assess the extent to which such intervention is effective on a larger scale in closing gaps, the present study provides proof of concept for one method of potentially altering the developmental trajectories of EAL learners' English vocabulary knowledge, thereby ensuring their equitable access to the curriculum.

Appendices

Appendix 3.1: Longitudinal Study Ethical Approval



Downloaded: 31/03/2015

Approved: 31/03/2015

Christopher Dixon
Registration number: 140117394
Human Communication Sciences
PhD Human Communication Sciences

Dear Christopher

PROJECT TITLE: Literacy development in children learning English as an Additional Language and their monolingual peers in primary school years 4 to 6.

APPLICATION: Reference Number 002846

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 31/03/2015 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 002846 (dated 30/03/2015).
- Participant information sheet 005329 (16/02/2015)
- Participant information sheet 005328 (16/02/2015)
- Participant consent form 005342 (16/02/2015)
- Participant consent form 005350 (16/02/2015)

If during the course of the project you need to [deviate significantly from the above-approved documentation](#) please inform me since written approval will be required.

Yours sincerely

Thomas Muskett
Ethics Administrator
Human Communication Sciences

Appendix 3.2: Parental Questionnaire Results

Summary

At t1, parental questionnaires were sent home to all children participating in the study. Questionnaires were sent home with pupils to be filled in and returned by parents/carers, where they were subsequently collected from school teachers. A total of 81 questionnaires were sent home at t1: additional copies of questionnaires were sent to parents who did not return them in the first instance. By the end of t3, a total of 57 questionnaires had been returned (monolingual: n=21, 63.6%; EAL: n=36, 75%), representing an overall return rate of 70%.

1. Country of birth

Country	EAL	Mono
U.K	33 (92%)	21 (100%)
Elsewhere	3 (8%)	0

2. Age that child began formal education in English

Age	EAL	Mono
<3	13 (36%)	11 (52%)
3 to 4	19 (53%)	10 (48%)
4 to 5	3 (8%)	0
Other	1 (3%)*	0

* age 6

3.1 Maternal highest educational qualification

Level	EAL	Mono
Primary	2 (6%)	0
Secondary	11 (32%)	8 (53%)
Further	13 (38%)	6 (40%)
University	4 (12%)	1 (7%)
Postgraduate	3 (9%)	0
Other	1 (3%)*	0

* doctoral degree

3.2 Paternal highest educational qualification

Level	EAL	Mono
Primary	0	0
Secondary	12 (34%)	8 11 (73%)
Further	13 (37%)	3 (20%)
University	7 (20%)	0
Postgraduate	3 (9%)	1 (7%)
Other	0	0

4. Country respondents received education in (EAL only)

Country	Mother	Father
U.K	17 (47%)	11 (32%)
Elsewhere	19 (53%)	23 (68%)

5.1 Maternal employment status

Status	EAL	Mono
Not Employed	32 (91%)	4 (25%)
Self-Employed	0	1 (6%)
Employed	3 (9%)	11 (69%)

5.2 Paternal employment status

Status	EAL	Mono
Not Employed	8 (23%)	3 (19%)
Self-Employed	11 (31%)	3 (19%)
Employed	16 (46%)	10 (62%)

6.1 Maternal English language proficiency

6.1.1 Maternal oral language skill

Level	Proportion
None	0
Poor	6 (17%)
Average	9 (25%)
Good	8 (22%)
Excellent	13 (36%)

6.1.2 Maternal literacy skill

Level	Proportion
None	1 (3%)
Poor	6 (17%)
Average	7 (19%)
Good	9 (25%)
Excellent	13 (36%)

6.2 Paternal English language proficiency

6.2.1 Paternal oral language skill

Level	Proportion
None	0
Poor	3 (9%)
Average	8 (24%)
Good	10 (29%)
Excellent	13 (38%)

6.2.2 Paternal literacy skill

Level	Proportion
None	0
Poor	2 (6%)
Average	9 (27%)
Good	9 (27%)
Excellent	13 (39%)

Incidence of familial language or literacy difficulties

Respondents were asked: *Please indicate if any members in the child's family (parents, grandparents, siblings, cousins, etc.) currently have, or had in the past, any problems with reading, writing, speaking, or listening (in any language). This could include a diagnosis of dyslexia, language impairment, autism, as well as any undiagnosed problems with learning to read, write, spell, speak, or understand what is being said.*

In total, five respondents indicated some presence of language and/or literacy difficulties in the family (3 monolingual; 2 EAL), representing just under 9% of questionnaire respondents. Four out of the five responses involved diagnoses of dyslexia in first- and second-degree relatives, and one response indicated presence of hearing difficulties. However, there were no indications that children taking part in the study had language and/or language difficulties.

7. Whether English is spoken in the home (EAL only)

English Spoken	Proportion
English is Spoken	32 (89%)
English is not Spoken	4 (11%)

8. Where English is spoken in the home, how often?

Frequency	Proportion
Never	1 (3%)
Rarely	2 (6%)
Sometimes	9 (27%)
Most of the time	16 (49%)
Always	5 (15%)

9. How often the child speaks and hears English and the home language

Language	Hears Most	Speaks Most
English	22 (67%)	28 (80%)
Home Language	11 (33%)	7 (20%)

10. Child and parental home language literacy proficiency

Ability	Parent	Child
Some Ability	9 (26%)	19 (54%)
No Ability	26 (74%)	16 (46%)

10.1 Receipt of formal instruction in the home language

Instruction	Proportion
Some Instruction	10 (33%)
No Instruction	20 (67%)

11. How often the child speaks English with various people

Frequency	Mother	Father	Siblings	Other	Friends
Never	0	1 (3%)	0	0	0
Rarely	3 (9%)	4 (12%)	1 (3%)	3 (9%)	0
Sometimes	7 (20%)	9 (27%)	1 (3%)	5 (16%)	0
Often	5 (14%)	3 (9%)	2 (6%)	4 (12%)	1 (3%)
Mostly	8 (23%)	10 (29%)	6 (17%)	5 (16%)	8 (23%)
All the Time	12 (34%)	7 (21%)	25 (71%)	15 (47%)	26 (74%)

12. Media consumption

Media	Books	TV	Internet
In English	35 (97%)	27 (75%)	35 (97%)
In Home Language	1 (3%)	3 (8%)	0
In Both Languages	0	6 (17%)	1 (3%)

Appendix 3.3: Results of Child Language Questionnaire

At t1 a child language preference questionnaire was administered orally to all children in the EAL group. This questionnaire ascertained background information such as the language(s) spoken in the home, children's preferences for language use, and extent of literacy skill in the home language.

1. Home languages spoken

Children's home languages included: Amharic (n=1; 2%), German (n=1; 2%), Hungarian (n=1; 2%), Nepali (n=1; 2%), Pushto (n=1; 2%), Somali (n=1; 2%), Thai (n=1; 2%), Tigrinya (n=1; 2%), Farsi (n=2; 4%), Polish (n=2; 4%), Turkish (n=2; 4%), Bengali (n=5; 11%), Urdu (n=7; 15%), Arabic (10; 21%), Punjabi (n=11; 23%)

2. Number of languages other than English spoken in the home

Other Languages	Proportion
1 Language	44 (94%)
2 Languages	3 (6%)

3. Language spoken most often at home

Language	Proportion
English	29 (62%)
Home Language	17 (36%)
Both	1 (2%)

4. Which language children prefer to speak at home

Language	Proportion
English	29 (63%)
Home Language	12 (26%)
Both	5 (11%)

5.1 Literacy ability in the home language: reading

Reading Ability	Proportion
Some Ability	22 (48%)
No Ability	24 (52%)

5.2. Literacy ability in the home language: writing

Writing Ability	Proportion
Some Ability	13 (28%)
No Ability	33 (72%)

6. Where children have some ability in home language literacy, what is the extent of formal instruction in the home language?

Instruction	Proportion
Formal Instruction	12 (50%)
No Instruction	12 (50%)

7. Which languages are spoken most often with different individuals

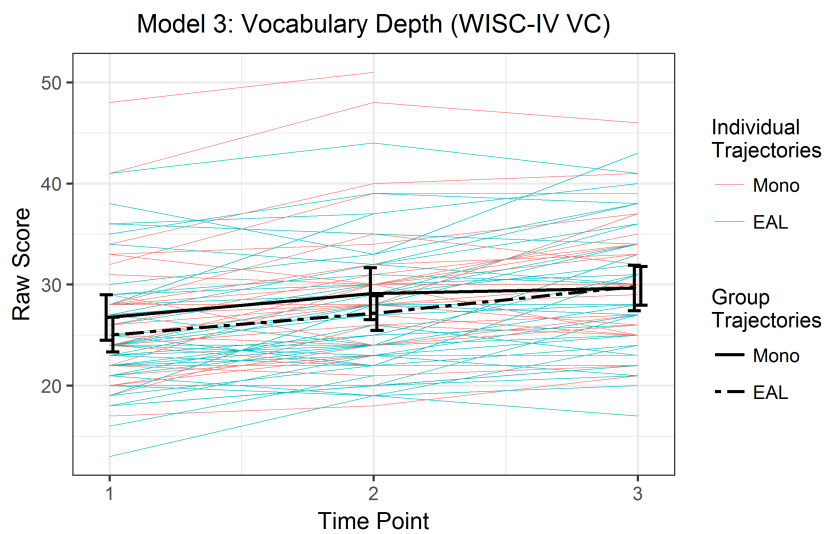
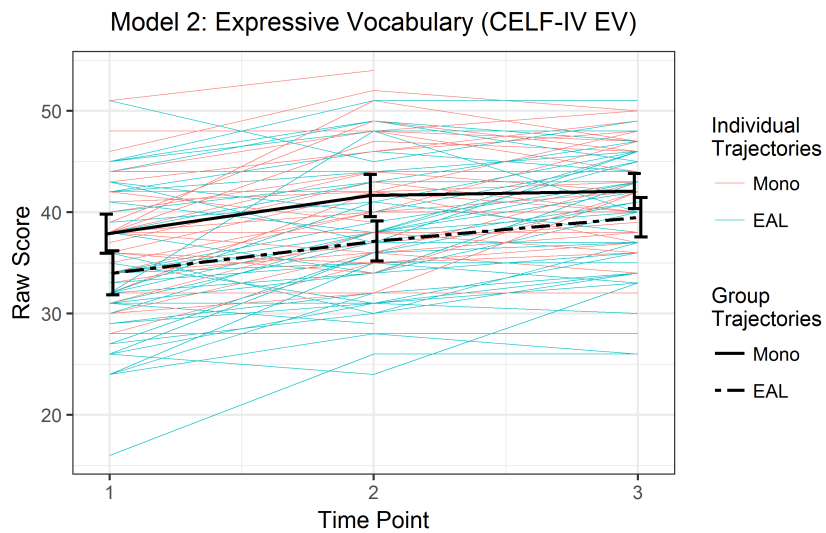
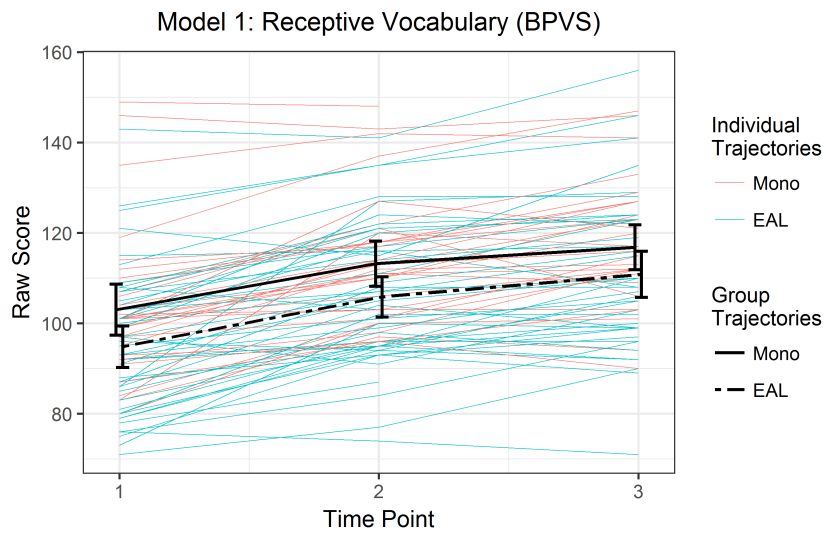
Language	Mother	Father	Siblings	Others
English	20 (44%)	23 (52%)	34 (83%)	3 (13%)
Home Language	20 (44%)	19 (43%)	1 (2%)	19 (83%)
Both	6 (12%)	2 (5%)	6 (15%)	1 (4%)

Appendix 4.1: Descriptive Statistics of Background Measures for Both Groups at all Time Points

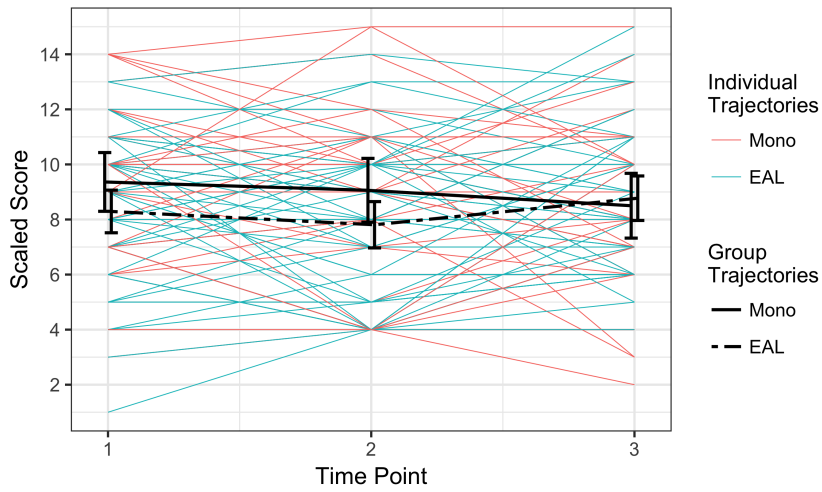
	Time	Mono			EAL			Effect Size of Group Difference		
		N	Mean (SD)	Range	N	Mean (SD)	Range	α	g	95% CI
WISC MR (Max = 35)	1	33	14.00 (4.16) <i>7.36 (2.41)</i>	6-23 3-13	48	14.56 (4.24) <i>7.63 (2.52)</i>	8-23 4-13	-	0.13	-0.59 to 0.32
	1	33	4.45 (1.30) <i>10.36 (2.16)</i>	2-8 6-16	48	3.77 (1.24) <i>9.13 (2.13)</i>	2-7 5-15	.59	0.53	0.07 to 1.00
	2	32	4.31 (1.60) <i>9.19 (2.65)</i>	2-8 5-16	44	4.25 (1.53) <i>9.05 (2.66)</i>	2-8 5-15	.64	0.04	-0.43 to 0.51
CELF NR Backwards (Max = 16)	3	32	4.38 (1.45) <i>8.72 (2.61)</i>	2-8 5-14	44	4.61 (1.54) <i>9.34 (2.56)</i>	2-8 5-14	.43	0.16	-0.63 to 0.31
	1	33	7.94 (1.60) <i>8.61 (2.50)</i>	5-12 4-14	48	7.50 (1.17) <i>7.88 (2.11)</i>	5-11 4-13	.74	0.32	-0.14 to 0.78
	2	32	7.91 (1.53) <i>8.22 (2.52)</i>	5-12 4-14	44	7.89 (1.45) <i>8.23 (2.50)</i>	5-11 4-13	.77	0.01	-0.46 to 0.48
CELF NR Forwards (Max = 16)	3	32	8.03 (1.75) <i>8.19 (2.83)</i>	5-14 4-17	44	8.07 (1.32) <i>8.43 (2.10)</i>	5-12 4-14	.79	0.02	-0.49 to 0.45
	1	33	12.39 (2.50) <i>9.55 (2.21)</i>	7-20 4-16	48	11.27 (1.69) <i>8.38 (1.77)</i>	8-14 4-11	.78	0.54	0.08 to 1.01
	2	32	12.22 (2.55) <i>8.41 (2.54)</i>	7-20 3-16	44	12.14 (2.35) <i>8.23 (2.37)</i>	7-18 3-14	.84	0.03	-0.44 to 0.50
CELF NR Total (Max = 30)	3	32	12.41 (6.83) <i>8.03 (2.60)</i>	7-22 3-17	44	12.68 (2.11) <i>8.39 (2.35)</i>	9-19 4-15	.73	0.12	-0.59 to 0.35

Note: WISC MR = Wechsler Intelligence Scale for Children IV Matrix Reasoning; CELF NR = Clinical Evaluation of Language Fundamentals IV Number Repetition; CELF NR total score derived from sum of performance on NR Forwards and NR Backwards; scaled scores and standard deviations in italics and standard deviations in italics; α = Cronbach's alpha (internal reliability).

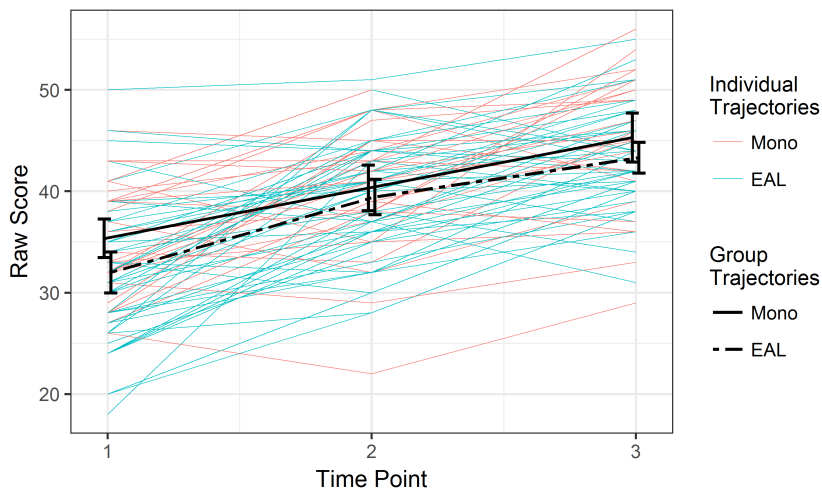
Appendix 4.2: Line Graphs Showing Group and Individual Trajectories for Language and Literacy Measures



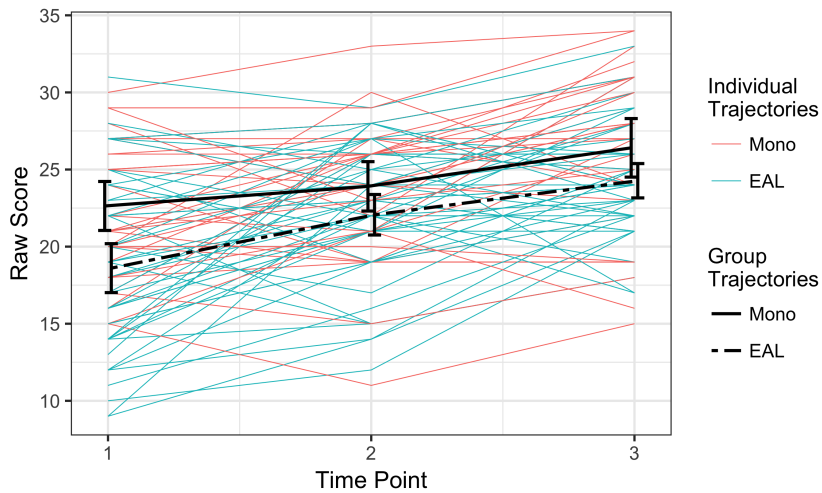
Model 4: Listening Comprehension (CELF-IV USP)



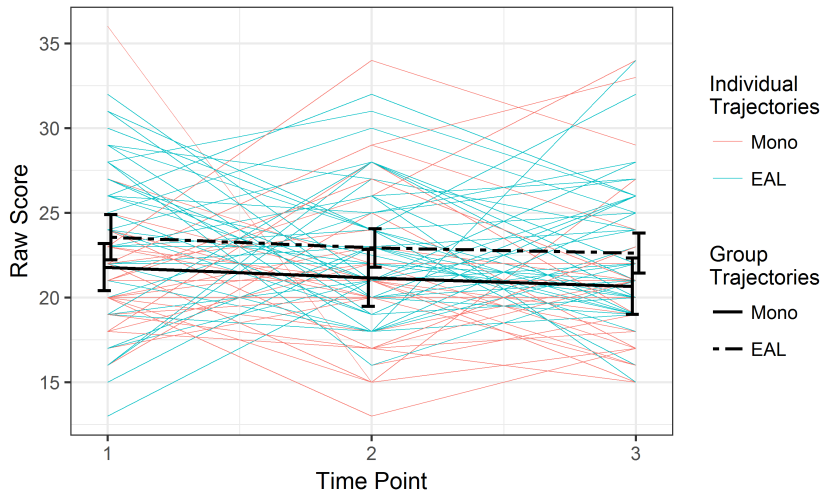
Model 5: Expressive Grammar (CELF-IV FS)



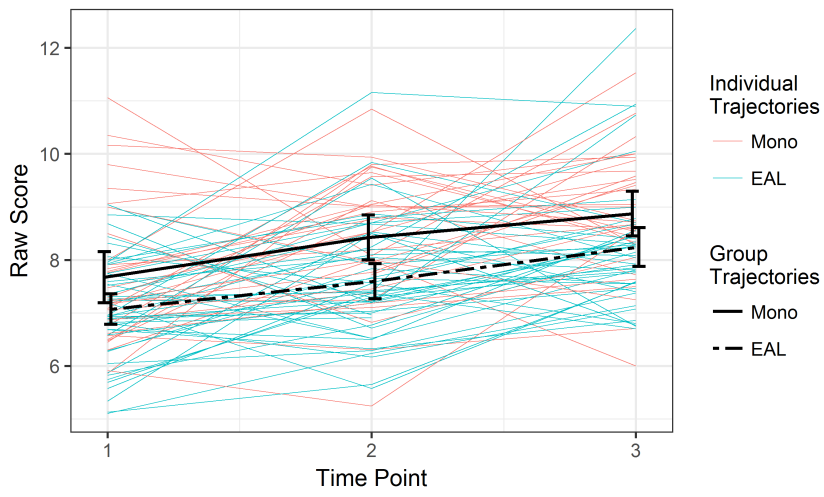
Model 5.1: Expressive Grammar (CELF-IV FS 17 Common Items)



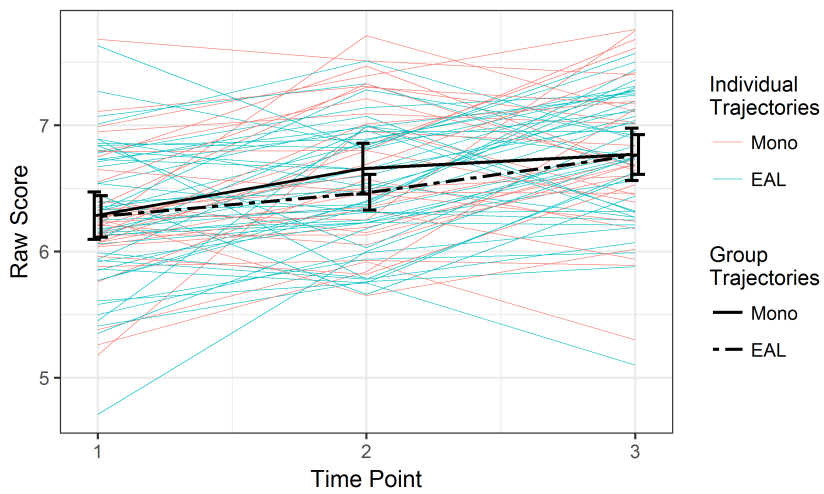
Model 6: Total Utterances



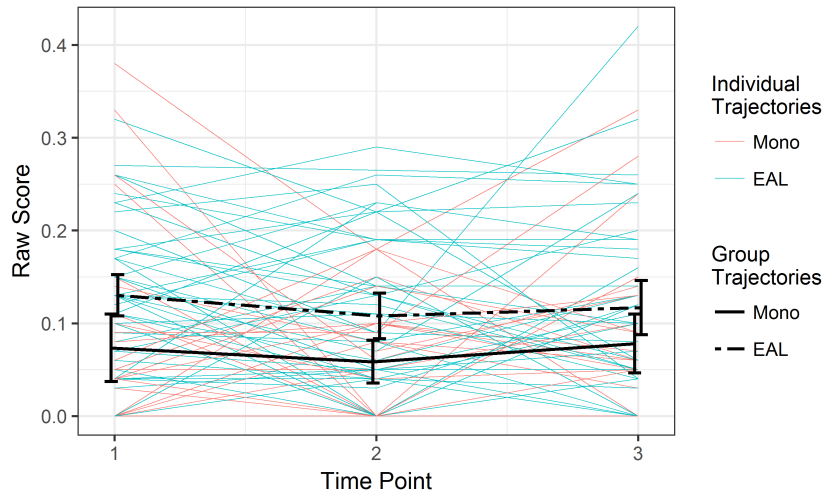
Model 7: Mean Length of Utterance in Words



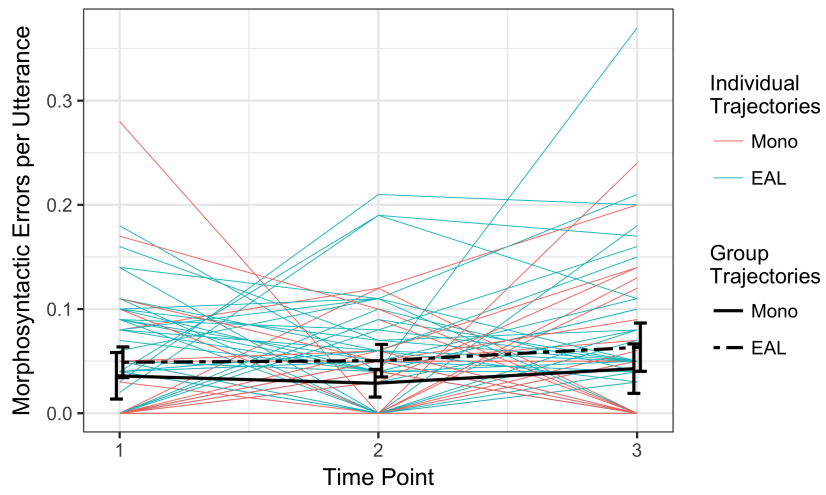
Model 8: Lexical Diversity



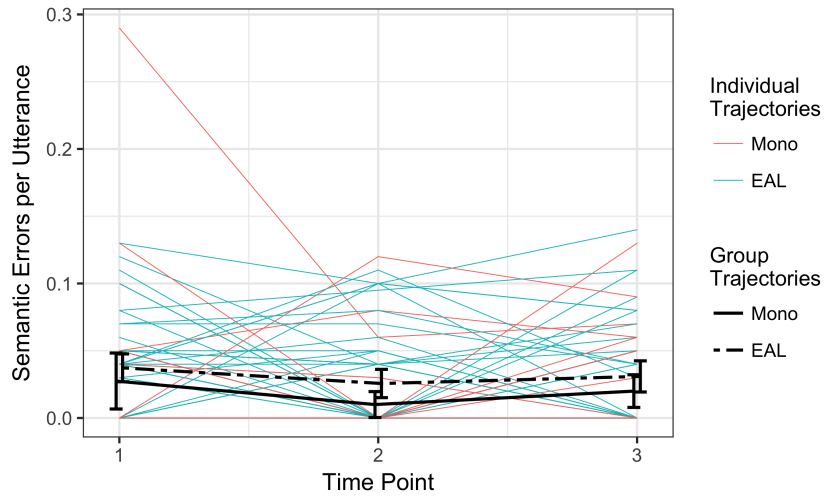
Model 9: Narrative Errors



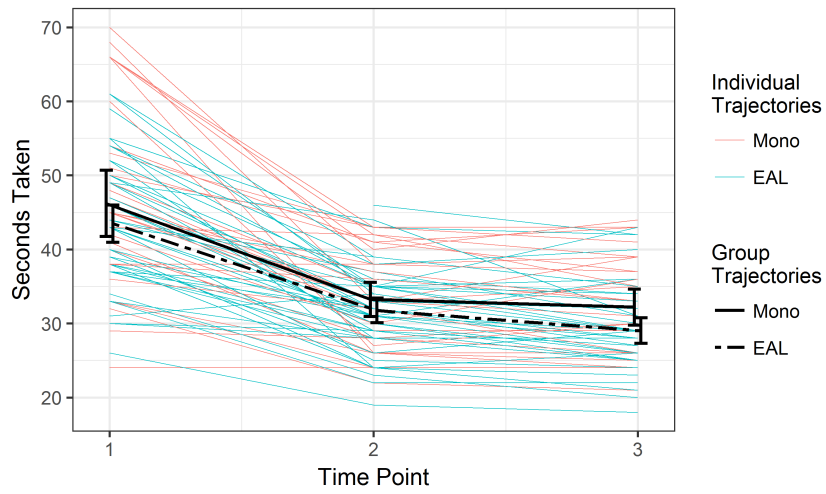
Model 9.1: Morphosyntactic Errors



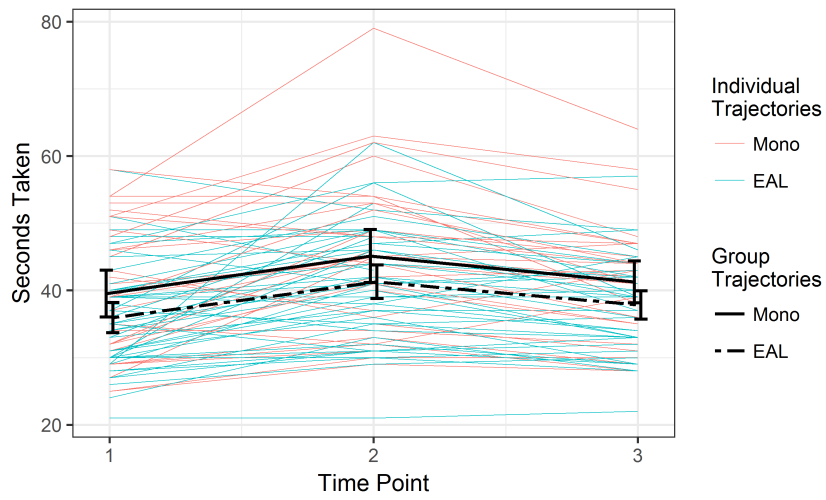
Model 9.2: Semantic Errors



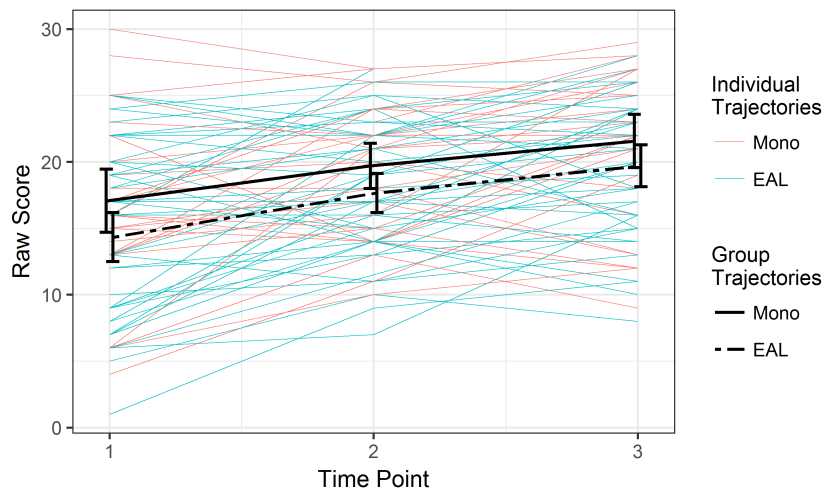
Model 10: RAN Digits



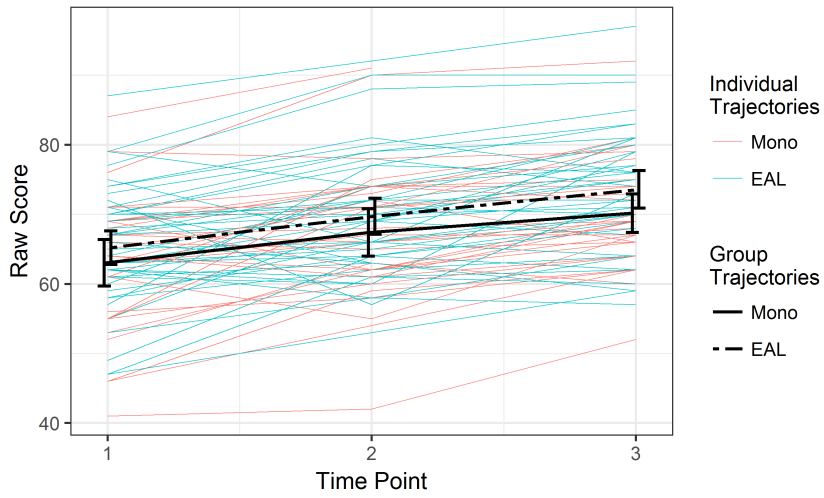
Model 11: RAN Letters



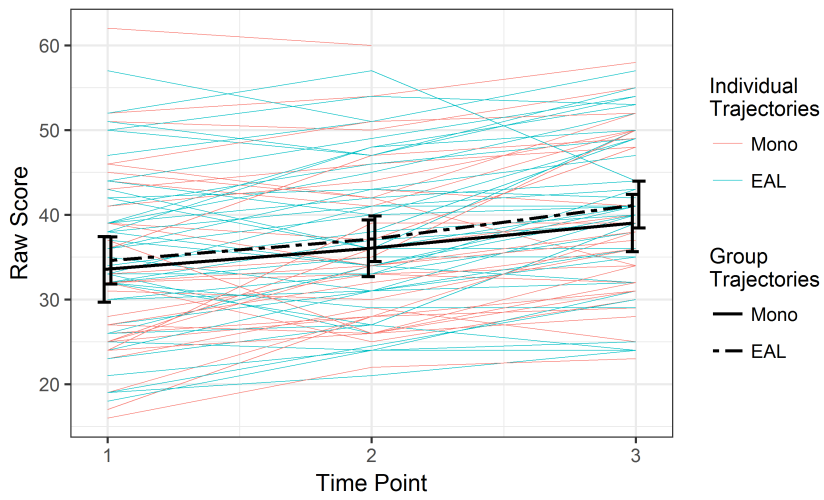
Model 12: Spoonerisms



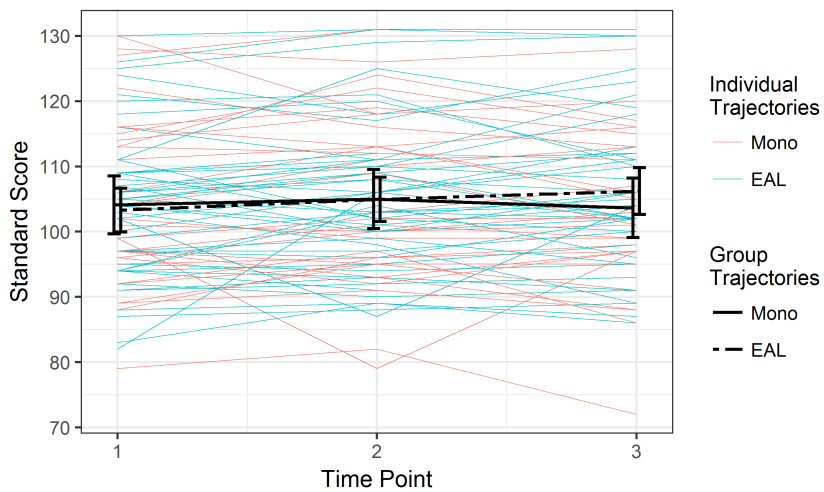
Model 13: TOWRE SW



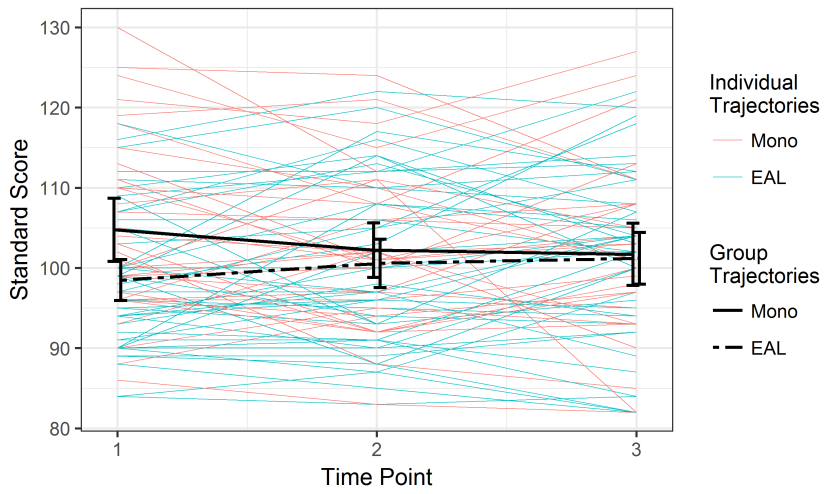
Model 14: TOWRE PD



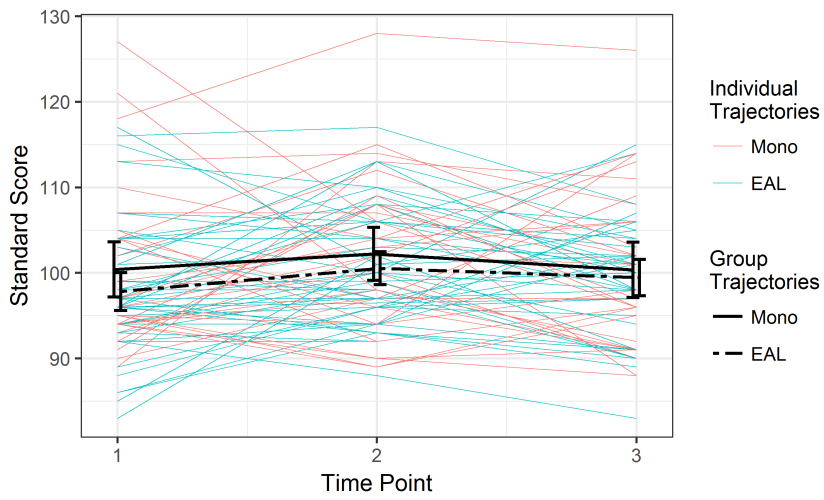
Model 15: Passage Reading Rate (YARC)



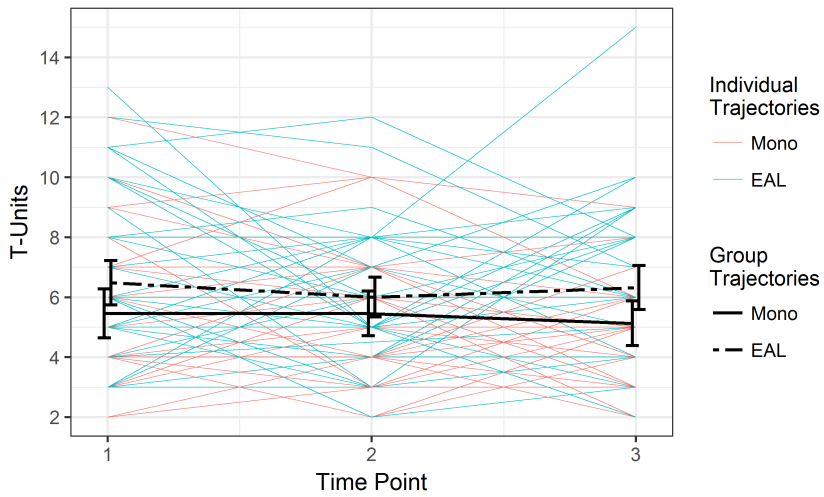
Model 16: Passage Reading Accuracy (YARC)



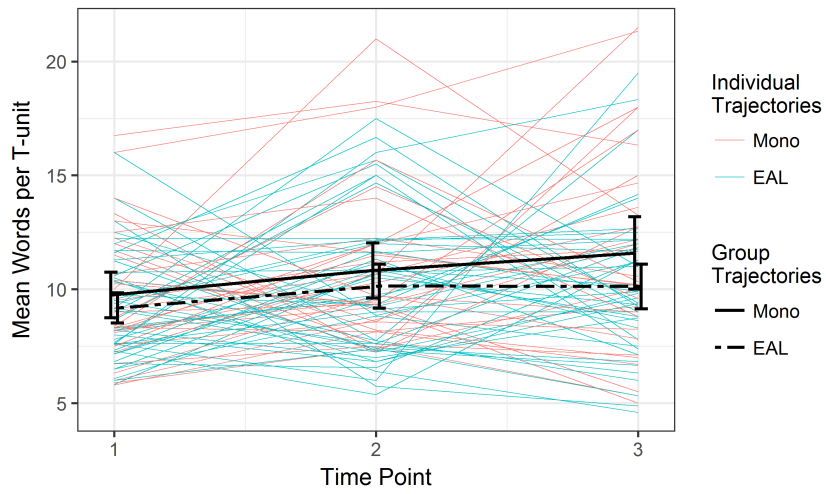
Model 17: Passage Reading Comprehension (YARC)



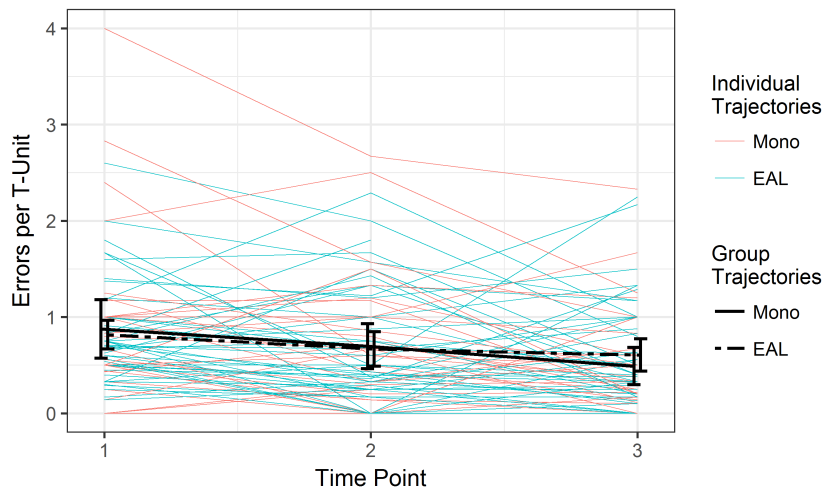
Model 18: Total T-Units



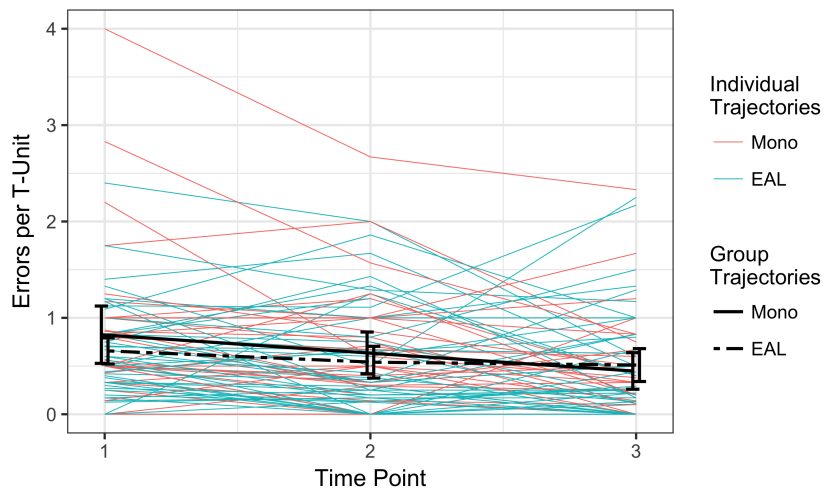
Model 19: Mean Length of T-Unit in Words



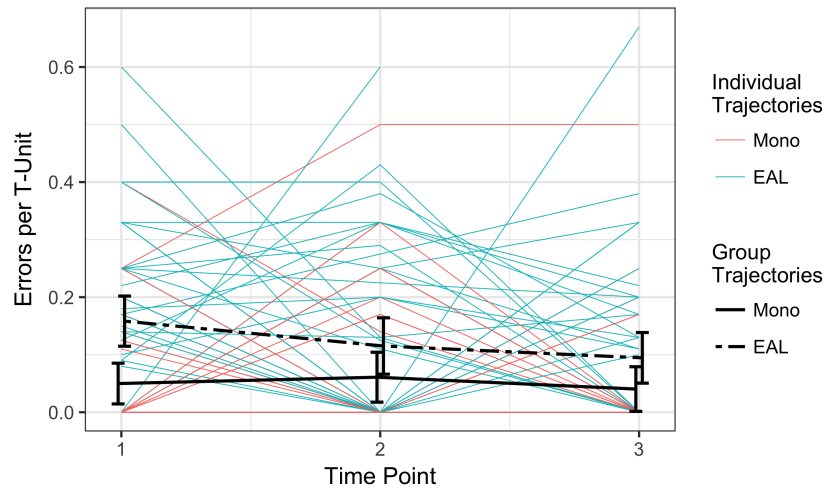
Model 20: Writing Error rate



Model 20.1: Spelling Error rate



Model 20.2: Morphosyntactic Error rate



Appendix 4.3: Descriptive Statistics from Bespoke Scoring on the WISC-IV VC Subtest at t1 of the Longitudinal Study

Category	Mono		EAL		Effect Size of Group Difference		
	Mean (SD)	Range	Mean (SD)	Range	<i>g</i>	95% CI	
Definition	14.79 (6.42)	6-37	14.81 (5.56)	4-27	0.00	-0.46 to 0.45	F(80,1) = .000
Background Knowledge	10.67 (3.45)	4-21	8.75 (3.50)	2-17	0.55	0.01 to 1.01	F(80,1) = 5.93 *
Lexical Knowledge	1.70 (1.33)	0-5	1.13 (1.25)	0-7	0.44	-0.02 to 0.90	F(80,1) = 3.88
Total Score	27.18 (9.13)	16-55	24.54 (8.53)	11-50	0.30	-0.16 to 0.76	F(80,1) = 1.77

Note: Descriptive statistics represent raw scores from Wechsler Intelligence Scale for Children IV Vocabulary subtest (bespoke scoring); * $p < .05$

Appendix 6.1: Intervention Study Ethical Approval



Downloaded: 26/01/2016
Approved: 26/01/2016

Christopher Dixon
Registration number: 140117394
Human Communication Sciences
Programme: PhD Human Communication Sciences

Dear Christopher

PROJECT TITLE: Supporting the oral language development of children learning English as an additional language

APPLICATION: Reference Number 007200

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 26/01/2016 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 007200 (dated 21/12/2015).
- Participant information sheet 1014426 version 1 (19/12/2015).
- Participant information sheet 1014427 version 1 (19/12/2015).
- Participant consent form 1014428 version 1 (19/12/2015).
- Participant consent form 1014429 version 1 (19/12/2015).
- Participant consent form 1014455 version 1 (21/12/2015).

The following optional amendments were suggested:

This is a very nicely presented and obviously carefully thought-through ethics application. There are only a few small comments suggested for the pIS and consent forms: add the full name of ESCAL to explain the acronym. would remove the sad face for children not taking part and stick to just thumbs up or thumbs down. Also - consider revising from 'without getting into trouble' - just 'you can stop at any time'. Also add somewhere that non-participation can be indicated by just refusing to come with the student - don't need to revisit the form etc? Good luck with your data collection!

If during the course of the project you need to [deviate significantly from the above-approved documentation](#) please inform me since written approval will be required.

Yours sincerely

John Mason
Ethics Administrator
Human Communication Sciences

Appendix 6.2: Characteristics for Taught Words (Frequency, AoA, and PoS)

Theme	Target word	Part of Speech	Age of Acquisition	Frequency Rating
Travelling and Distant Lands	distant	ADJ	8.95	8.61
	capital	N	8.1	12.71
	coast	N	6.43	26.69
	navigate	V	10.05	1.92
			M = 8.38	M = 12.48
Emotions and States	furious	ADJ	8.78	6
	miserable	ADJ	10.11	21.49
	thrilled	ADJ	7.62	11.06
	cautious	ADJ	8.25	3.35
			M = 8.69	M = 10.46
Wrongdoing	responsible	ADJ	8.37	45.06
	disagree	V	10.11	6.63
	persuade	V	10.15	6.39
	fraud	N	10.79	10.04
			M = 9.86	M = 17.03
Shopping and Finance	purchase	V	8.11	6.37
	bargain	N	8.72	12
	wealthy	ADJ	7.89	7.37
	afford	V	7.47	44.43
			M = 8.05	M = 17.54
Accident and Emergency	rescue	V	7.17	25.41
	disaster	N	8.22	17.27
	agony	N	9.22	3.75
	fatal	ADJ	9.05	7.1
			M = 8.42	M = 13.38
			Mean (SD) all words	Mean (SD) all words
			8.78 (1.14)	14.18 (12.56)

Note: Frequency Rating illustrates occurrence of word per one million words of printed text; Age of Acquisition (AoA) indicated in years; M = mean; characteristics for untaught words presented overleaf.

Appendix 6.2 cont'd: Characteristics for Untaught Words (Frequency, AoA, and PoS)

Theme	Word	Part of Speech	Age of Acquisition	Frequency Rating
	annual	ADJ	9.26	7.2
	budget	N	10.05	10.06
	contagious	ADJ	8.17	3.33
	donation	N	9.33	3.51
	genuine	ADJ	9.06	8.2
	identical	ADJ	8.72	5.53
	maximum	ADJ	7.6	7.76
	pilot	N	6.32	26.67
	starve	V	8.32	6.16
	tolerate	V	10.45	6.94
			Mean (SD) all words	Mean (SD) all words
			8.73 (1.20)	8.54 (6.69)

Note: Frequency Rating illustrates occurrence of word per one million words of printed text; Age of Acquisition (AoA) indicated in years. M = mean; See Section 6.5 for details on the matching between taught and untaught words.

Appendix 6.3: Example Intervention Passage

Week 1: Travelling and Distant Lands

Word 1:

Jake and Adil were best friends. They did all sorts of things together. On Saturday they had planned to visit the natural history museum in Thorpetown to see all the ancient dinosaur bones. It was the morning of the trip.

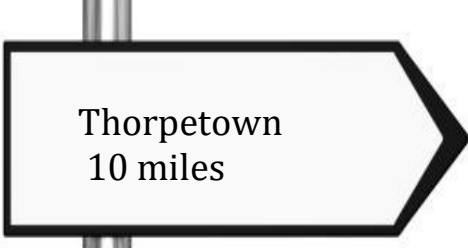
“I’m ready, let’s go!” said Adil, putting on his shoes. But Jake wasn’t so sure.

“Thorpetown is quite distant, so it’s a long way to walk. Maybe we should take the bus”, he suggested.

“Yes, you’re right” agreed Adil.

“I think my feet would get too sore if we walked all the way there!”

Once they were in the museum, they saw some wonderful and fascinating things. Adil’s attention was caught by some ancient dinosaur bones. He read the sign: “Ancestors from the distant past” But how distant exactly, Adil wondered? He was looking at a dinosaur skeleton that was 65 million years old.



Thorpetown
10 miles



Questions

1. How do you know that Thorpetown was distant?
2. How do you know that the dinosaur skeleton was from the distant past?

Appendix 6.4: Inter-Rater Guide

Word Knowledge Assessment Rubric *Word Score*

In this task, children are asked to give a verbal definition of a list of 30 words. The ability to give a dictionary-style definition is a metalinguistic task which improves with age (Benelli et al. 2006; Snow et al. 1991). Other approaches to the assessment of vocabulary depth take a different approach and give credit for non-definitional responses, for instance, in background, lexical, and gestural knowledge (e.g. Hadley et al. 2016). Indeed, this is consonant with the idea that word knowledge is comprised of many factors (Nation, 2001).

Test Administration

In this task, each stimulus word is presented to the child (verbally and in writing), who is then asked 'Can you tell me what this word means?' After two practice items to indicate examples of permissible responses, children are encouraged to give as much information as possible for the remaining 30 words. All verbal responses have been transcribed for scoring.

Scoring

Word knowledge is scored according to 4 broad categories described below.

1. Definition (0-2): an explanation of the word's meaning. This may be more or less explicit. For instance, for the word 'donate', *when you give money to somebody* would receive 1 point, whereas *when you give money to [poor people / people who need it]* would receive 2 points as this is more highly specified. Specific examples for each of the 30 words can be found in the scoring rubric.
2. Background knowledge: for any information that would not constitute a definition but nevertheless indicates understanding of the word. Background knowledge is made up of three categories:
 - Situational knowledge (0-1): any situation in which the target word would be used or would apply, for instance for the target word 'disaster': *like when you're drowning*; or for the target word 'wealthy': 'when you live in a mansion'. These represent examples and situations that apply to the target words without giving abstract, decontextualised definitions. Note: also award 1 point for an appropriate and correct example, e.g. *Madrid is the capital of Spain* for target word 'capital'.
 - Contextual knowledge (0-1): the addition of contextually related words, concepts, or phrases (similar to synonyms but related contextually rather than linguistically). For instance, mention of *sky, helmet, helicopter*, etc for 'pilot', mention of *bank card, currency, cash* for 'purchase'; mention of *kilometre, metre, miles, journey, land* for 'distant'.
 - Functions and attributes (0-2): For instance *stamp your feet and shout* for 'furious'; *cry* for 'agony'. Award 1 point for one function or attribute, and award 2 points for two or more functions or attributes.

General scoring principle: Award points for background knowledge for such items if they are presented in isolation, or if they are presented 'outside' of a definition. For example, for the word 'disagree':

when you think you're right and the other person thinks you're wrong - if you're trying to solve 5 x 5 and they say 20, you say no, it means you're disagreeing with them because it's not the right answer

Broken down:

- *when you think you're right and the other person thinks you're wrong* (definition score of 2)
- *if you're trying to solve 5 x 5 and they and they say 20, you say no, it means you're disagreeing with them because it's not the right answer* (situational knowledge score of 1)

Additionally, sometimes it may seem that a definition is 'hiding' in the form of background knowledge. For example, if a child said something like *A disaster is like an earthquake*, this would not receive any score for definition because it is merely a simple example. Perhaps the child does have an abstract sense of what the word 'disaster' means, but if this isn't explicitly stated, then a response cannot be given a definition score.

In other cases, a situational example may function as a vehicle for a definition, e.g. a definition for 'thrilled' must contain notions of happiness and/or excitedness. For example: *if you went on a rollercoaster and you were really happy and excited, you're thrilled* – this would receive a score of 2 for definition (mention of 'really happy and excited') and 1 for background knowledge (mention of rollercoaster example).

1. Lexical knowledge: This category is comprised of two subcategories.

- Synonymy / Antonymy (0-1): Award 1 point for any correct synonym (e.g. *careful* for 'cautious', *angry* for 'furious', or *deadly* for 'fatal') or any correct antonym as long as it is explicitly stated as the opposite (e.g. *agree* for 'disagree', *poor* for 'wealthy', or *happy* for 'miserable').
- Morphological / Collocational knowledge (0-1): Award 1 point for the mention of derivationally related forms of target words, such as *distance* for 'distant', *tolerant* for 'tolerate', etc. Award 1 point for any idiomatic or formulaic phrases containing the target word, e.g. 'the coast is clear' or 'rescue attempt'.

Word Knowledge Assessment Rubric Sentence Score

For this task, children are asked to use the target word within a sentence, e.g. *Can you put the word distant in a sentence?* The justification for this activity is that word knowledge also comprises a productive element which may include not only a word's form and meaning, but also its grammatical functions (Nation, 2001).

Please take note of the following scoring conventions:

Incorrect Sentences

Sentences may be incorrect (and score 0 points) for one or more of the following reasons:

1. Sentence does not contain the target word.
2. Sentence is incomplete, e.g. *The capital of England*.
3. Sentence uses a derivational form of the target word, e.g. *distance* for target word 'distant' or *donated* for target word 'donation'.
4. Sentence uses a neologised form of the target word, usually into a different part of speech e.g. *I disasterly went for a walk*; *I maximumed all the money*.
5. The target word is interpreted incorrectly or mistaken for another word. E.g. 'I bought a budgie [budget]', 'Today we went on the rollercoaster [coast]', and 'I was purring [purchase] like a cat'.
6. An unrelated sense of the target word is used, E.g. 'She forgot to put a capital letter' [wrong sense of capital – see scoring rubric].
7. (No responses are marked as 'NR' and receive a score of 0).

Correct Sentences

1. Syntax (0-1): is the word used as the correct part of speech? If the target word is a noun, it may: be used as an argument (e.g. subject or object), be the complement of a verb, have nominal morphology (e.g. plurality), be preceded by a determiner, etc., If it is a verb, it may take arguments as subjects or objects, show inflectional morphology (tense, agreement), etc.

Assign a score of 1 if and only if the word is used correctly according to its part of speech, e.g. *Last summer I went to the coast*, *They rescued the cat*.

Note: scoring here is only concerned with errors of the target word and not the sentence as a whole. For instance, an error elsewhere in a sentence will not affect the scoring of correct usage of the target word, e.g. *Yesterday I were going to the coast* would still receive a score of 1 for syntax.

2. Morphology (0-1): does the target word contain any morphological errors? Assign a score of 1 if and only if the target word does not contain any morphological errors (e.g. in agreement or number) related to the target word, for example: *The woman purchases a bag in the shop*. Any error in agreement or number receives a score of 0, e.g. *I am starve*, *The man purchase his mother*. Again, this applies only to the target word – other errors in the sentence will not affect this score. Note: if the word is used as the wrong part of speech (i.e. receives a score of 0 for syntax), it will necessarily also receive a score of 0 for morphology, e.g. *Last year I fraud my teacher* cannot be interpreted because fraud is not a verb and therefore cannot 'miss' or 'lack' verbal morphology.
3. Semantics (0-3): how well does the sentence display knowledge of the target word? A short sentence such as 'The holiday was a disaster' is admissible and appropriate, but a sentence such as 'The holiday was a disaster because the flight was cancelled' shows deeper understanding of why it was a disaster, and gives an explicit example. Unlike scores for syntax and morphology which range from 0 to 1, scores for semantics range from 0 to 3. See below for examples of sentence scoring:

- 0 points: Sentence does not make sense semantically, e.g. *I was contagious to go on the stage* (nervous); *I persuaded my sister how to speak in English* (taught); *I purchase myself*; *I was fatal*. These examples are easier to score because they are categorically incorrect. However, other sentences may present more of a challenge due to vagueness or ambiguity, e.g. *I disagree with my friend because he is nasty to me*, or *I am persuaded*. In these cases, it is up to your judgement whether to award 0 or 1 point.
- 1 point: Sentence has a very simple structure, e.g. *I was starving*, *I am going to rescue the cat*, *I was miserable*; *I went to the coast yesterday*. Syntactically and morphologically these sentences are correct, but there is little information to measure depth of knowledge. Note: any target words that are used as part of an idiom or formulaic phrase receive 1 point.
- 2 points: Sentence provides more information through use of adverbs, adjectives, subordinating conjunctions, prepositional phrases, relative clauses, etc. but the sentence is still slightly vague and inexplicit, e.g. *I was miserable to go to school*; *In literacy we had to write a letter to persuade someone*; *I was in agony when I fell off the swings*. These sentences receive 2 points because they provide more context around the target word, although they do not provide explicit information about meaning (semantics) – in *I was miserable to go to school*, the target word is used correctly, and the addition of *to go to school* tells us that this is a situation in which one might be miserable; however, we are not given any explicit indication as to the meaning of ‘miserable’ or why this situation would cause someone to become miserable.
 Note: adverbial phrases of time or place do not warrant 2 points, e.g. *I was starving*, and *I was starving on Monday* both receive 1 point, as *on Monday* does not add meaningfully to the sense of *starving*.
- 3 points: The sentence is well-formed around the target word and gives a more explicit meaning, e.g. *I were fasting and I was starved to death because I didn't have nothing to eat* (note that the other errors in this sentence do not detract from the final score); *the man couldn't afford anything because he had no money*; *I was on the coast near the sea*; *the man in the shop let me have it for half price: it was a bargain*; *I felt miserable today: I was very sad and I didn't want to go to school*. Generally, 3-point sentences may contain subordinate clauses (e.g. *x because y*), synonyms (e.g. *miserable* and *sad*), or an explicit reason for something (e.g. something that is half price is likely to make it a bargain).

3. Comments

Optional: please provide brief comments about how the session went – were there any issues with the materials, the activities, the working space/environment?



Appendix 7.2: Intervention Observation Checklist

Environment

1. Is the workspace conducive to learning? I.e. is there a lot of noise, is there enough space to work, are seating arrangements adequate?

Structure of the session

1. How long did the session last?
2. Which activities were fully or partially completed?

Recap

1. Was there a recap of last week's words at the beginning of the session?
2. Did the child remember words from the previous week?

Passage reading activity

1. Who read the passage? If the child read the passage, how did he/she manage, and did the coordinator have to assist?
2. Did the coordinator point out cues to word meanings in the text?

Sentence tasks (judgement and completion)

1. Where wrong answers were given by the child, did the coordinator give adequate explanations?
2. In the sentence completion activity did the coordinator correct any errors or explain why they were errors?

Mina Map activity

1. What kinds of prompts did the coordinator give?
2. What kinds of information were added to the map? I.e. situational, contextual, lexical? In other words, was there evidence of increasing vocabulary depth?
3. Did the coordinator add any items to the map herself?
4. In the sentence writing task did the coordinator prompt the child to make a longer/better sentence?

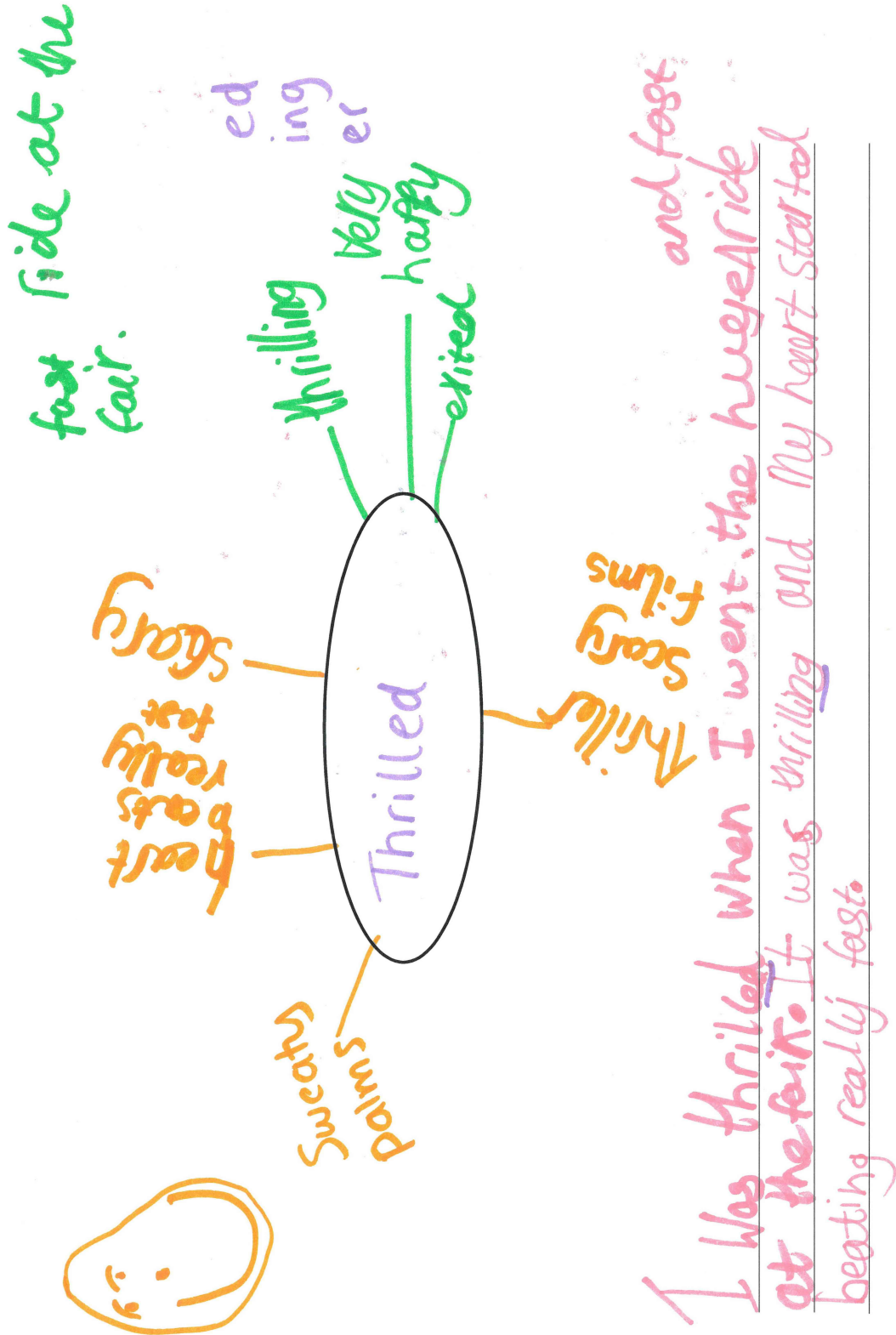
Word games

1. Was a game played at the beginning? If so, did this run smoothly and engage the child?

General

1. Was there a general level of rapport?
2. Did the child seem engaged and interested?

Appendix 7.3: Example of mind-map activity from intervention



Appendix 4.1: Appendix 7.4: Parametric Statistics for Word, Sentence, and Total Score for the 20 Taught Words

	Baseline		Baseline to Pretest		Pretest to Posttest		Posttest to Maintenance	
	Pretest	Main.	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>	<i>t</i>	<i>p</i>
Total score	54.00 (17.03)	87.28 (24.13)	-0.44	.668	-5.66	<.001	1.06	.321
Words and Sentences								
Word score	9.11 (3.66)	17.89 (8.51)	n/a	n/a	-5.48	<.001	0.72	.490
Definition								
Background	7.44 (3.24)	10.22 (4.32)	1.32	.223	-2.10	.069	-1.66	.135
Lexical	1.00 (1.00)	1.33 (0.87)	n/a	n/a	-2.77	.024	2.07	.073
Non-Verbal	0.11 (0.22)	0.06 (0.17)	n/a	n/a	-1.00	.347	1.51	.169
Total	17.67 (6.82)	29.5 (11.47)	1.22	.257	-7.16	<.001	0.78	.455
Sentence score	10.67 (2.83)	15.78 (3.19)	-0.72	.487	-3.08	.015	-0.76	.471
Morphology	10.78 (2.91)	15.67 (3.24)	-0.69	.509	-3.01	.017	-0.55	.594
Semantics	14.89 (5.62)	26.33 (8.15)	-0.90	.394	-4.18	.003	1.34	.217
Total	36.33 (10.89)	57.78 (13.94)	-0.83	.429	-3.93	.004	0.92	.382

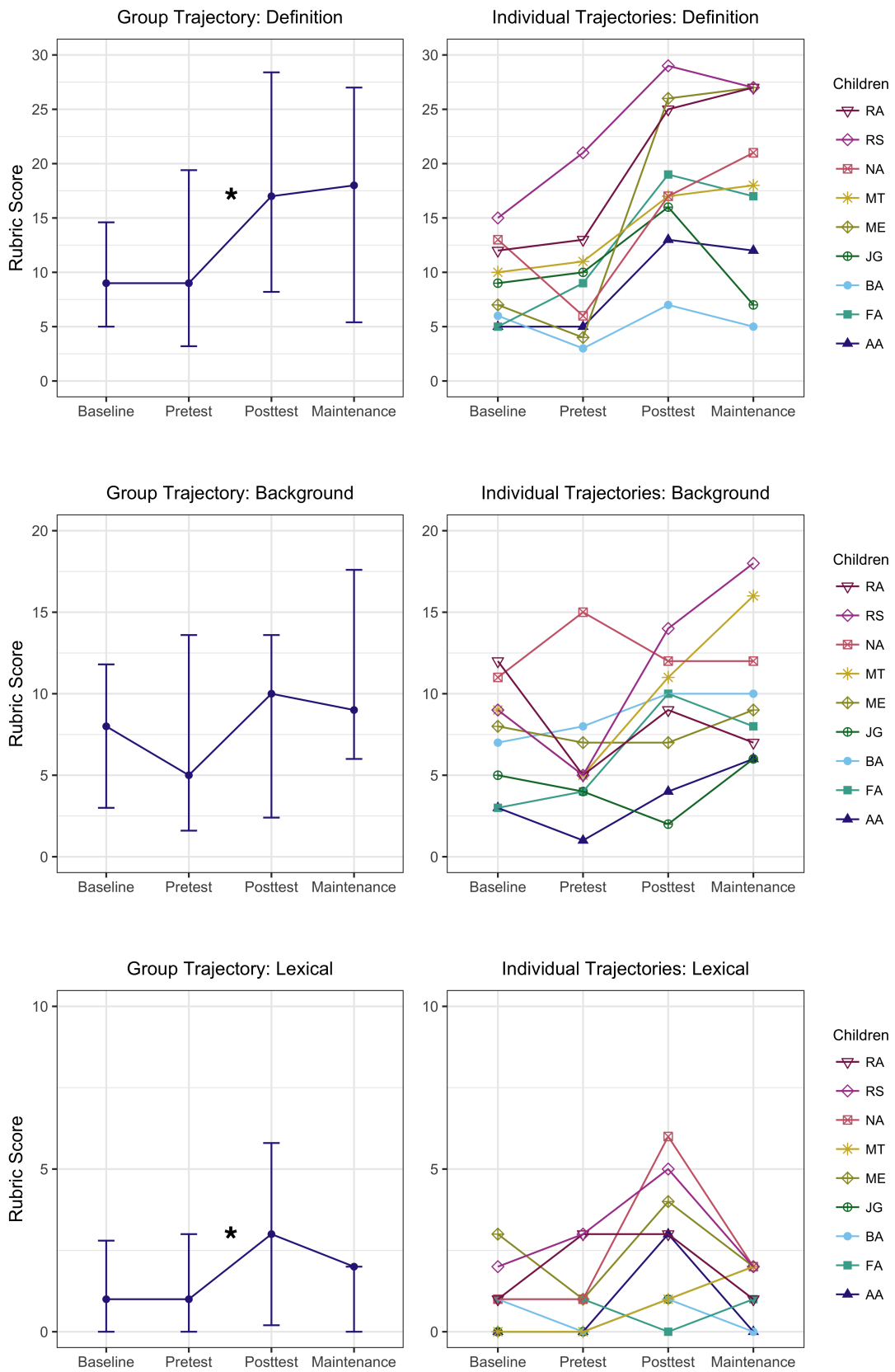
Note: statistics in columns Baseline to Main. Represent mean and (SD); *t* and *p* statistics reported from dependent t-tests; *d* = effect size in Cohen's *d*; Main = maintenance test.

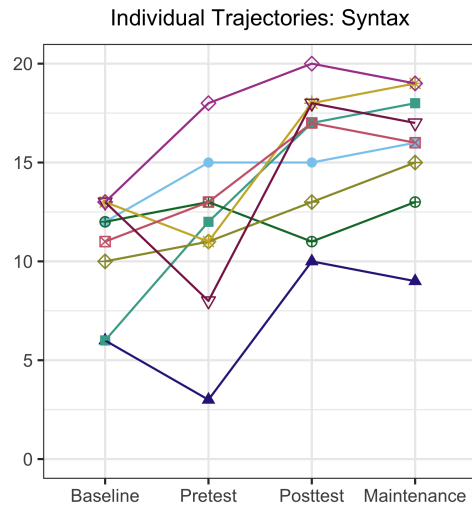
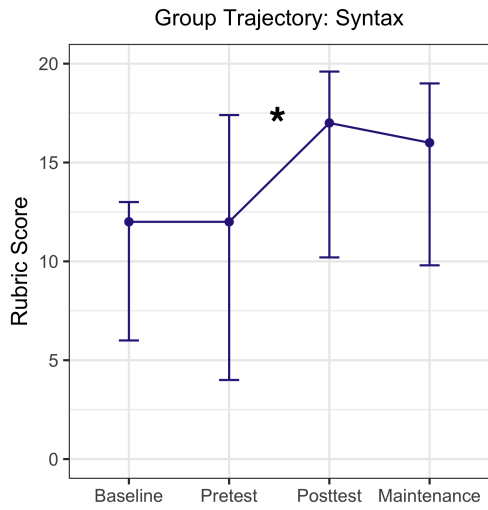
Appendix 7.4 cont'd: Parametric Statistics for Word, Sentence, and Total Score for the 10 Untaught Words

	Baseline			Baseline to Pretest			Pretest to Posttest			Posttest to Maintenance		
	Pretest	Posttest	Main.	<i>t</i>	<i>p</i>	<i>d</i>	<i>t</i>	<i>p</i>	<i>d</i>	<i>t</i>	<i>p</i>	<i>d</i>
Total score	20.56 (7.63)	22.78 (11.54)	32.89 (10.52)	-1.01	.323	0.34	-3.48	<.01	1.16	1.43	.191	0.48
Word score												
Definition	4.11 (2.93)	3.89 (2.89)	6.11 (3.44)	0.32	.760	0.11	-1.67	.133	0.56	-1.89	.096	0.63
Background	2.67 (1.22)	3.33 (2.18)	3.33 (2.18)	-1.03	.332	0.34	-0.36	.729	0.12	0.16	.874	-0.05
Lexical	0.22 (0.44)	0.11 (0.33)	0.44 (0.53)	0.55	.594	-0.18	-0.55	.594	0.18	1	.347	0.33
Non-Verbal	0.11 (0.33)	0 (0)	0 (0)	1	.347	-0.33	1	.347	0.33	1	.347	-0.33
Total	7.11 (3.79)	7.33 (4.61)	9.89 (4.37)	-0.31	.766	0.10	-1.89	.095	0.63	-1.77	.116	0.59
Sentence score												
Syntax	4.11 (1.17)	4.67 (2.12)	6.67 (1.66)	-0.89	.401	0.30	-1.18	.272	0.39	-2.19	.059	0.73
Morphology	4.11 (1.17)	4.67 (2.12)	6.67 (1.66)	-0.89	.401	0.30	-1.25	.247	0.42	-2.04	.076	0.68
Semantics	5.22 (2.64)	6.11 (3.82)	9.67 (4.36)	-1.32	.225	0.44	-3.38	<.01	1.13	-0.18	.864	0.06
Total	13.44 (4.72)	15.44 (7.73)	23.00 (6.86)	-1.10	.305	0.37	-3.28	.011	1.09	-1.27	.240	0.42

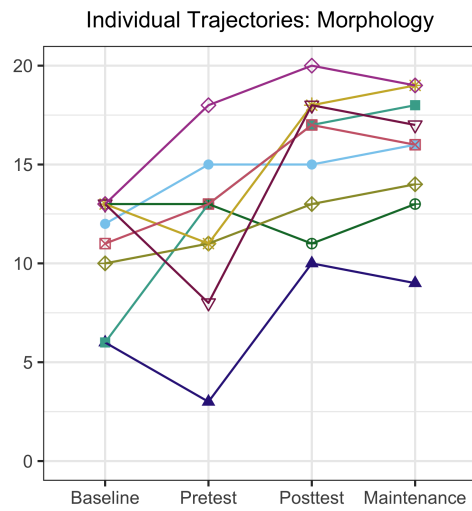
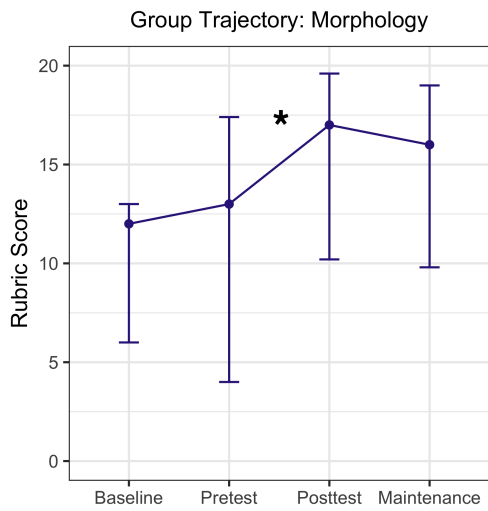
Note: statistics in columns Baseline to Main. Represent mean and (SD); *t* and *p* statistics reported from dependent t-tests; *d* = effect size in Cohen's *d*; Main = maintenance test.

Appendix 7.5: Group and Individual Trajectories for Subcategories of the Bespoke Word Knowledge Assessment

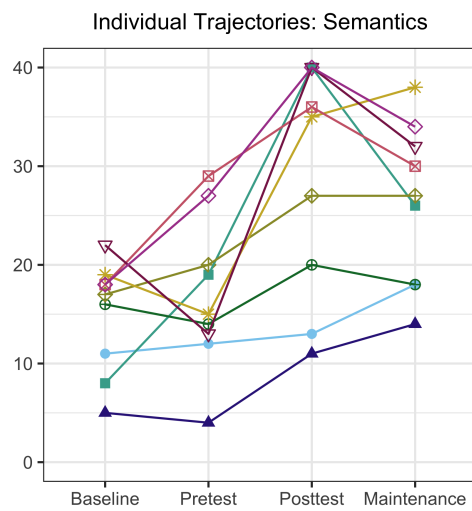
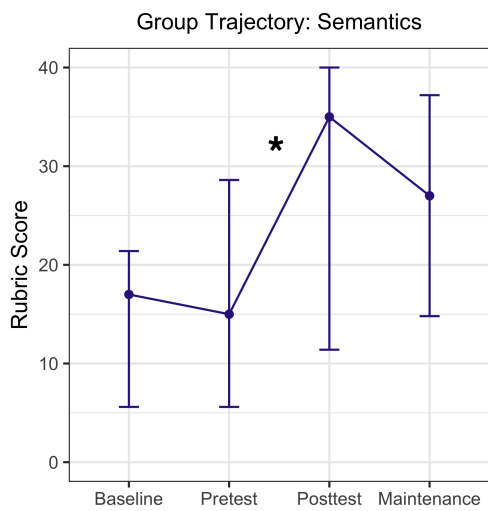




- Children
- RA
 - RS
 - NA
 - MT
 - ME
 - JG
 - BA
 - FA
 - AA



- Children
- RA
 - RS
 - NA
 - MT
 - ME
 - JG
 - BA
 - FA
 - AA



- Children
- RA
 - RS
 - NA
 - MT
 - ME
 - JG
 - BA
 - FA
 - AA

Appendix 7.6: Raw Scores at each Timepoint and Pre-to-Post Changes for the 20 Taught Words

Word	Raw Scores (Total Word Score)			Pretest to Posttest Change		Pretest to Posttest Change Gain (<i>d</i>)		
	Baseline	Pretest	Posttest	Maintenance	Raw Score	Effect Size (<i>d</i>)	Word Score	Sentence Score
afford	1.50	0.92	1.80	1.78	0.88	1.11	1.17	0.39
agony	0.00	0.08	0.70	1.00	0.62	0.62	0.65	0.78
bargain	0.42	0.33	1.50	1.56	1.17	1.07	1.12	1.72
capital	1.08	0.83	1.50	1.89	0.67	0.60	0.63	0.34
cautious	0.25	0.17	0.20	1.00	0.03	0.48	0.50	0.07
coast	0.33	0.63	1.71	1.61	1.08	0.80	0.84	0.59
disagree	1.46	1.75	2.00	2.00	0.25	0.19	0.20	-0.16
disaster	1.33	1.08	1.30	1.33	0.22	0.18	0.18	0.89
distant	1.54	1.83	2.00	2.00	0.17	0.35	0.37	0.65
fatal	0.00	0.00	0.50	0.00	0.50	0.60	0.63	0.42
fraud	0.33	0.33	0.78	0.89	0.44	0.42	0.44	0.68
furious	0.83	0.67	1.60	1.00	0.93	0.63	0.67	0.61
miserable	0.89	0.56	1.22	1.11	0.67	0.91	0.96	0.42
navigate	0.00	0.67	1.30	0.67	0.63	0.62	0.65	0.90
persuade	0.78	0.67	1.25	1.44	0.58	0.90	0.94	-0.25
purchase	0.75	0.67	1.91	2.22	1.24	0.80	0.84	0.93
rescue	1.92	1.33	2.00	2.22	0.67	0.47	0.49	-0.31
responsible	1.78	1.78	2.22	2.00	0.44	0.80	0.84	0.56
thrilled	0.58	0.67	1.50	1.44	0.83	0.42	0.44	0.47
wealthy	0.00	0.42	1.50	2.00	1.08	0.85	0.89	1.49

Note: *d* = effect size in Cohen's *d*.

Appendix 7.7: Progress on Standardised Measures Across the Study (WISC VC and CELF FS) - Parametric Statistics

		t1		Baseline		Posttest		t3		t1 to Baseline			Baseline to Posttest			Posttest to t3		
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	t	p	d	t	p	d	t	p	d
Raw Scores	WISC VC	20.33	(4.72)	24.00	(4.82)	25.00	(5.61)	27.00	(7.04)	-4.81	.001	1.60	-1.96	.087	0.65	-1.92	0.64	.091
	CELF FS	29.67	(8.67)	32.22	(6.65)	40.33	(4.27)	41.89	(4.94)	-1.41	.196	0.47	-3.48	.008	1.16	-0.87	0.29	.412
Scaled Scores	WISC VC	7.00	(1.93)	7.89	(1.83)	7.89	(2.20)	8.22	(2.64)	-3.05	.016	1.48	0	1	0	-0.71	0.24	.450
	CELF FS	5.00	(3.00)	5.33	(2.24)	7.67	(1.73)	8.00	(1.87)	-0.49	.641	0.16	-2.22	.007	0.92	-0.45	0.15	.667

Note: Descriptive statistics represent means and (standard deviation); *r* and *p* calculated from dependent t-tests (df=8 for all t-tests)
 Interpretation of *r* as follows: 0.2 = small, 0.5 = medium, 0.8 = large.

References

- Aaron, P., Joshi, R., Ayotollah, M., Ellsberry, A., Henderson, J., & Lindsey, K. (1999). Decoding and sight-word naming: Are they independent components of word recognition skill? *Reading & Writing, 11*, 89-127.
- Abbott, R., Berninger, V., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in Grades 1 to 7. *Journal of Educational Psychology, 102*, 281.
- Adams, M. J. (1990). *Beginning to read: thinking and learning about print*. Cambridge, Mass.: MIT Press.
- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*, 207-245.
- Adlof, S. M., & Catts, H. W. (2015). Morphosyntax in poor comprehenders. *Reading & Writing, 28*, 1051-1070.
- Allen, M. S., Kertoy, M. K., Sherblom, J. C., & Pettit, J. M. (1994). Children's narrative productions: A comparison of personal event and fictional stories. *Applied Psycholinguistics, 15*, 149-176.
- Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (p. 77-117). Newark, DE: International Reading Association.
- Anglin, J. M. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 58*, 1-166.
- Anwar, M. (1998). *Between cultures: Continuity and change in the lives of young Asians*. London: Routledge.
- Applebee, A. N. (1978). *The child's concept of story: Ages two to seventeen*. Chicago: University of Chicago Press.
- Araújo, S., Petersson, K., Reis, A., & Faísca, L. (2015). Rapid automatized naming and reading performance: A meta-analysis. *Journal of Educational Psychology, 107*, 868-883.
- August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice, 20*, 50-57.
- Babayiğit, S. (2014a). Contributions of word-level and verbal skills to written expression: Comparison of learners who speak English as a first (L1) and second language (L2). *Reading & Writing, 27*, 1207-1229.
- Babayiğit, S. (2014b). The role of oral language skills in reading and listening comprehension of text: A comparison of monolingual (L1) and bilingual (L2) speakers of English language. *Journal of Research in Reading, 37*, 22-47.

- Babayigit, S. (2015). The relations between word reading, oral language, and reading comprehension in children who speak English as a first (L1) and second language (L2): A multigroup structural analysis. *Reading & Writing, 28*, 527-544.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Baetens Beardsmore, H. (1982). *Bilingualism: Basic principles*. Clevedon: Tieto.
- Baker, C. (2006). *Foundations of bilingual education and bilingualism*. Clevedon: Multilingual Matters.
- Banerjee, M., & Frees, E. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association, 92*, 999-1005.
- Barac, R., & Bialystok, E. (2011). Cognitive development of bilingual children. *Language Teaching: Surveys and Studies, 44*, 36-54.
- Barnes, M. A., Dennis, M., & Haefele-Kalvaitis, J. (1996). The effects of knowledge availability and knowledge accessibility on coherence and elaborative inferencing in children from six to fifteen years of age. *Journal of Experimental Child Psychology, 61*, 216-241.
- Bartón, K. (2016). *MuMIn: Multi-model inference*. Retrieved from <https://cran.r-project.org/web/packages/MuMIn/index.html>
- Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.
- Bates, L. (2008). Responsible vocabulary word selection: Turning the tide of 50-cent terms. *The English Journal, 97*, 68-76.
- Baumann, J. F., Edwards, E. C., Boland, E. M., Olejnik, S., Kame, & enui, E. J. (2003). Vocabulary tricks: Effects of instruction in morphology and context on fifth-grade students' ability to derive and infer word meanings. *American Educational Research Journal, 40*, 447-494.
- Beck, I., McKeown, M., & Kucan, L. (2005). Choosing words to teach. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (p. 209-222). New York: Routledge.
- Beck, I. L., & McKeown, M. G. (1991). Conditions of vocabulary acquisition. In R. Barr, M. Kamil, P. Mosenthal, & D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, p. 780-814). Mahwah, NJ: Lawrence Erlbaum Associates.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford Press.
- Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal, 83*, 177-181.
- Beech, J., & Keys, A. (1997). Reading, vocabulary and language preference in 7- to 8-year-old bilingual Asian children. *British Journal of Educational Psychology, 67*, 405-414.

- Benelli, B., Belacchi, C., Gini, G., & Lucangeli, D. (2006). To define means to say what you know about things: The development of definitional skills as metalinguistic acquisition. *Journal of Child Language, 33*, 71-97.
- Bercow, J. (n.d.). *The Bercow Report: A review of services for children and young people (0-19) with speech, language and communication needs*.
- Berman, R. A. (1988). On the ability to relate events in narrative. *Discourse Processes, 11*, 469-497.
- Berninger, V. W. (2000). Development of language by hand and its connections with language by ear, mouth, and eye. *Topics in Language Disorders, 20*, 65-84.
- Bialystok, E., Craik, F. I. M., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia, 45*, 459-464.
- Bialystok, E., Peets, K. F., Yang, S., & Luk, G. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition, 13*, 525-531.
- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93*, 498-520.
- Bijeljac-babic, R., Biardeau, A., & Grainger, J. (1997). Masked orthographic priming in bilingual word recognition. *Memory and Cognition, 25*, 447-457.
- Bishop, D. (1997). *Uncommon understanding: Development and disorders of language comprehension in children*. Hove: Psychology Press.
- Bishop, D. (2003a). *Children's Communication Checklist-2*. London: Pearson.
- Bishop, D. (2003b). *Test for Reception of Grammar-2*. Harcourt Assessment.
- Bishop, D. (2004). *Expression, Reception and Recall of Narrative Instrument*. London: Harcourt Assessment.
- Blackledge, A., & Creese, A. (2010). *Multilingualism: A critical perspective*. London: Continuum.
- Bliese, P., & Ployhart, R. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods, 5*, 362-387.
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.
- Bowen, N., & Guo, S. (2012). *Structural equation modeling*. Oxford: Oxford University Press.
- Bowey, J. (2005). Predicting individual differences in learning to read. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (p. 155-172). Oxford: Blackwell.
- Bowyer-Crane, C., Fricke, S., Schaefer, B., Lervåg, A., & Hulme, C. (2017). Early literacy and comprehension skills in children learning English as an additional language and monolingual children with language weaknesses. *Reading & Writing, 30*, 771-790.

- Bowyer-Crane, C., Snowling, M. J., Duff, F. J., Fieldsend, E., Carroll, J. M., Miles, J., . . . Hulme, C. (2008). Improving early language and literacy skills: Differential effects of an oral language versus a phonology with reading intervention. *Journal of Child Psychology and Psychiatry*, *49*, 422-432.
- Brooks, K., & Thurston, L. (2010). English Language Learner academic engagement and instructional grouping configurations. *American Secondary Education*, *39*, 45-60.
- Brownell, R. (2000). *Expressive and receptive one-word picture vocabulary tests (3rd Ed.)*. Novato, CA: Academic Therapy Publications.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage.
- Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904-911.
- Burgoyne, K., Kelly, J. M., Whiteley, H. E., & Spooner, A. (2009). The comprehension skills of children learning English as an additional language. *British Journal of Educational Psychology*, *79*, 735-747.
- Burgoyne, K., Whiteley, H. E., & Hutchinson, J. M. (2011). The development of comprehension and reading-related skills in children learning English as an additional language and their monolingual, English-speaking peers. *British Journal of Educational Psychology*, *81*, 344-354.
- Burgoyne, K., Whiteley, H. E., & Kelly, J. M. (2011). The role of background knowledge in text comprehension for children learning English as an additional language. *Journal of Research in Reading*, *36*, 132-148.
- Burton, P., Gurrin, L., & Sly, P. (1998). Tutorial in biostatistics. extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modeling. *Statistics in Medicine*, *17*, 1261-1291.
- Cain, K. (2003). Text comprehension and its relation to coherence and cohesion in children's fictional narratives. *British Journal of Developmental Psychology*, *21*, 335-351.
- Cain, K. (2007). Syntactic awareness and reading ability: Is there any evidence for a special relationship? *Applied Psycholinguistics*, *28*, 679-694.
- Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*, 683-696.
- Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'année Psychologique*, *114*, 647-662.
- Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*, 31-42.

- Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading & Writing, 11*, 489-503.
- Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory and Cognition, 29*, 850-859.
- Cajkler, W., & Hall, B. (2009). 'when they first come in what do you do?' English as an additional language and Newly Qualified Teachers. *Language and Education, 23*, 153-170.
- Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research, 6*, 145-73.
- Cameron, L., & Besser, S. (2004). *Writing in English as an additional language at Key Stage 2*. Department for Education and Skills.
- Caravolas, M., Lervåg, A., Defior, S., Malkova, G., & Hulme, C. (2012). Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. *Psychological Science, 24*, 1398-1407.
- Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (p. 264-293). Cambridge, MA: MIT Press.
- Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., ... White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-Language Learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*, 188-215.
- Carnine, D., nui, E. J., & Coyle, G. (1984). Utilization of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly, 19*, 188-204.
- Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. London: Routledge.
- Carver, R. P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Literacy Research, 26*, 413-437.
- Casey, L. (2016). *The Casey review: A review into opportunity and integration*. London: Crown Copyright.
- Cattani, A., Abbot-Smith, K., Farag, R., Krott, A., Arreckx, F., Dennis, I., & Floccia, C. (2014). How much exposure to English is necessary for a bilingual toddler to perform like a monolingual peer in language tests? *International Journal of Language & Communication Disorders, 49*, 649-671.
- Catts, H., Adlof, S., & Weismer, S. (2006). Language deficits in poor comprehenders: A case for the Simple View of Reading. *Journal of Speech, Language, and Hearing Research, 49*, 278-293.
- Centre for Information on Language Teaching and Research. (2005). *Language trends 2005: Community language learning in England, Wales and Scotland* (Tech. Rep.). National Centre for Languages.

- Chall, J., Jacobs, V., & Baldwin, L. (1990). *The reading crisis: Why poor children fall behind*. Cambridge, MA: Harvard University Press.
- Chi, M., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-533.
- Chiappe, P., & Siegel, L. S. (1999). Phonological awareness and reading acquisition in English- and Punjabi-speaking Canadian children. *Journal of Educational Psychology*, *91*, 20-28.
- Chmiliar, L. (2012). Multiple-case designs. In A. Mills, G. Durepos, & E. Wiebe (Eds.), *Encyclopedia of case study research* (p. 583-584). Thousand Oaks: Sage.
- Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-90.
- Clarke, P., Truelove, E., Hulme, C., & Snowling, M. J. (2014). *Developing reading comprehension*. Chichester: Blackwell.
- Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*, 1106-16.
- Clay, M., Gill, M., Glynn, E., McNaughton, A., & Salmon, K. (1976). *Record of Oral Language*. Wellington: N.Z.E.I.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Collier, V. P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, *21*, 617-641.
- Collier, V. P. (1989). How long? a synthesis of research on academic achievement in a second language. *TESOL Quarterly*, *23*, 509-31.
- Coltheart, M. (2005). Modeling reading: The Dual-Route Approach. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (p. 6-23). Oxford: Blackwell.
- Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589-608.
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204-256.
- Costley, T. (2014). English as an additional language, policy and the teaching and learning of English in England. *Language and Education*, *28*, 276-292.
- Coyne, M., McCoach, B., Loftus, S., Zipoli, R., Ruby, M., Crecevoeur, Y., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness*, *3*, 93-120.
- Coyne, M., Simmons, D., Kame'enui, E., & Stoolmiller, M. (2004). Teaching vocabulary during shared storybook readings: An examination of differential effects. *Exceptionality*, *12*, 145-162.

- Cragg, L., & Nation, K. (2006). Exploring written narrative in children with poor reading comprehension. *Educational Psychology, 26*, 55-72.
- Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.
- Crouch, J., & Stonehouse, M. (2016). *A question of identity and equality in multicultural Britain*.
- Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism, 9*, 121-129.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*, 222-251.
- Cummins, J. (1981a). Age on arrival and immigrant second language learning in Canada. *Applied Linguistics, 2*, 132-149.
- Cummins, J. (1981b). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework* (p. 3-49). Los Angeles: National Dissemination and Assessment Center.
- Cummins, J. (1991). Interdependence of first- and second-language proficiency in bilingual children. In E. Bialystok (Ed.), *Language processing in bilingual children* (p. 70-89). Cambridge: CUP.
- Cunningham, A. (2005). Vocabulary growth through independent reading and reading aloud to children. In M. Kamil & E. Hiebert (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (p. 45-68). New York: Routledge.
- Cunningham, A., & Stanovich, K. (1997). Early reading acquisition and its relation to reading experience and ability ten years later. *Developmental Psychology, 33*, 934-45.
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research, 28*, 369-382.
- Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*, 277-99.
- Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles? *Annual Review of Applied Linguistics, 31*, 182-204.
- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*, 23-45.
- Datchuk, S. (2017). A direct instruction and precision teaching intervention to improve the sentence construction of middle school students with writing difficulties. *The Journal of Special Education, 51*, 62-71.

- De Houwer, A. (1995). Bilingual language acquisition. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (p. 219-250). Oxford: Blackwell.
- De Houwer, A. (2009). *Bilingual first language acquisition*. Bristol: Multilingual Matters.
- De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual-monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied Psycholinguistics*, *35*, 1189-1211.
- Demie, F. (2013). English as an additional language pupils: how long does it take to acquire English fluency? *Language and Education*, *27*, 59-69.
- Deno, S. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *16*, 99-104.
- Department for Education. (2013). *The National Curriculum in England: Key Stages 1 and 2 framework document*. Retrieved from <https://www.gov.uk/government/publications/national-curriculum-in-england-primary-curriculum>
- Department for Education. (2014). *Statutory framework for the Early Years Foundation Stage: Setting the standards for learning, development and care for children from birth to five*.
- Department for Education. (2015a). *Schools, pupils and their characteristics: January 2015 (no. SFR19/2015: National tables)*.
- Department for Education. (2015b). *Schools, pupils and their characteristics January 2015 (no. SFR16/2015: Local authority and regional tables)*.
- Department for Education. (2015c). *Proficiency in English*. Retrieved from http://defenddigitalme.com/wp-content/uploads/2016/07/RFC_875_-_new_data_item_for_proficiency_in_English.pdf
- Department for Education. (n.d). *English indices of deprivation 2015 (IDACI postcode search)*. Retrieved from <http://imd-by-postcode.opendatacommunities.org/>
- Department for Education and Skills. (2004). *Pupil characteristics and class sizes in maintained schools in England*. Retrieved from <https://www.naldic.org.uk/Resources/NALDIC/Research%20and%20Information/Documents/EALpupilsbyLEA20042013.xls>
- Department for Education and Skills. (2007). *Ensuring the attainment of pupils learning EAL*. Retrieved from https://www.naldic.org.uk/Resources/NALDIC/Teaching%20and%20Learning/ks3_ws_eal_mgmt_gd_sch_strat.pdf
- Dickinson, D. K., McCabe, A., Clark-Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics*, *25*, 323-347.
- Dijkstra, T., & van Heuven, W. (1998). The BIA model and bilingual word recognition. In J. Grainer & A. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (p. 189-226). Mahwah, NJ: Lawrence Erlbaum Associates.
- Dockrell, J. E., Stuart, M., & King, D. (2010). Supporting early oral language skills for English language learners in inner city preschool provision. *British Journal of Educational Psychology*, *80*, 497-515.

- Dole, J., Sloan, C., & Trathen, W. (1995). Teaching vocabulary within the context of literature. *Journal of Reading, 38*, 452-460.
- Dowds, S., Haverback, H., & Parkinson, M. (2016). Classifying the context clues in children's text. *The Journal of Experimental Education, 84*, 1-22.
- Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly, 38*, 78-103.
- Duncan, L., Seymour, P., & Hill, S. (2000). A small-to-large unit progression in metaphonological awareness and reading? *Human Experimental Psychology, 53*, 1081-1104.
- Dunn, L., Dunn, D., & NFER. (2009). *British Picture Vocabulary Scale - Third Edition*. London: GL Assessment.
- Dunn, L., & Dunn, L. (1997). *The Peabody Picture Vocabulary Test - Third Edition*. Circle Pines, MN: American Guidance Service.
- Dunn, L., & Dunn, L. (2007). *The Peabody Picture Vocabulary Test - Fourth Edition*. Circle Pines, MN: American Guidance Service.
- Duran, L. K., Roseth, C. J., & Hoffman, P. (2010). An experimental study comparing English-only and transitional bilingual education on Spanish-speaking preschoolers' early literacy development. *Early Childhood Research Quarterly, 25*, 207-217.
- Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85*, 453-465.
- Educational Endowment Foundation. (2015). *Cash for pupils without English as first language should be better targeted, as research shows big differences in results*. Retrieved 30.1.15, from <https://educationendowmentfoundation.org.uk/news/cash-for-pupils-without-english-as-first-language-should-be-better-targeted/>
- Edwards, S., Letts, C., & Sinka, I. (2011). *The New Reynell Developmental Language Scales*. London: GL Assessment.
- Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading, 18*, 116-125.
- Elbaum, B., Vaughn, S., Hughes, M. T., & Watson Moody, S. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? a meta-analysis of the intervention research. *Journal of Educational Psychology, 92*, 605-619.
- Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2*, 1-44.
- Elleman, A. M., Steacy, L. M., Olinghouse, N. G., & Compton, D. L. (2017). Examining child and word characteristics in vocabulary learning of struggling readers. *Scientific Studies of Reading, 21*, 133-145.
- Elliot, C. (1996). *Dansk evneprøve [Danish Ability Scales]*. Dansk Psykologisk Forlag.

- Elliott, C., Smith, P., & McCulloch, K. (1997). *The British Ability Scales II*. Windsor: NFER-Nelson.
- Ellis, M., Hester, H., & Barrs, T. (1990). *Patterns of learning*. London: Centre for Language in Primary Education.
- Feagans, L., & Short, E. J. (1984). Developmental differences in the comprehension and production of narratives by reading-disabled and normally achieving children. *Child Development, 55*, 1727-1736.
- Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage.
- Fien, H., Santoro, L., Baker, S., Park, Y., Chard, D., Williams, S., & Haria, P. (2011). Enhancing teacher read alouds with small-group vocabulary instruction for students with low vocabulary in first-grade classrooms. *School Psychology Review, 40*, 307-318.
- Fischer, S. (2001). *A history of writing*. London: Reaktion Books Ltd.
- Foorman, B., Herrera, S., Petscher, Y., Mitchell, A., & Trueman, A. (2015). The structure of oral language and reading and their relation to comprehension in kindergarten through Grade 2. *Reading & Writing, 28*, 655-681.
- Forum for Research in Literacy and Language. (2012). *Diagnostic Test of Word Reading Processes*. London: GL Assessment.
- Frederickson, N., & Frith, U. (1998). Identifying dyslexia in bilingual children: A phonological approach with inner London Sylheti speakers. *Dyslexia, 4*, 119-131.
- Frederickson, N., Frith, U., & Reason, R. (1997). *Phonological Assessment Battery - Standardised Edition*. London: NFER-Nelson.
- Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly, 18*, 277-294.
- Fricke, S., Bowyer-Crane, C., Haley, A. J., Hulme, C., & Snowling, M. J. (2013). Efficacy of language intervention in the early years. *Journal of Child Psychology and Psychiatry, 54*, 280-90.
- Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., . . . Hulme, C. (2017). The efficacy of early language intervention in mainstream school settings: A randomized controlled trial. *Journal of Child Psychology & Psychiatry, 58*, 1141–1151.
- Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall, & M. Coltheart (Eds.), *Surface dyslexia, neuropsychological and cognitive studies of phonological reading* (p. 301-330). London: Lawrence Erlbaum Associates.
- Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 104-115.
- Gardner, D. (2007). Children's immediate understanding of vocabulary: Contexts and dictionary definitions. *Reading Psychology, 28*, 331-373.

- Gardner, D. (2013). *Exploring vocabulary: Language in action*. London: Routledge.
- Garnham, A. (1982). Testing psychological theories about inference making. *Memory and Cognition*, 10, 341-349.
- Garton, A., & Pratt, C. (2009). Cultural and developmental predispositions to literacy. In D. Olson & N. Torrance (Eds.), *The cambridge handbook of literacy* (p. 501-519). Cambridge: CUP.
- Gathercole, V. C. M. (2007). Miami and North Wales, so far and yet so near: A constructivist account of morphosyntactic development in bilingual children. *International Journal of Bilingual Education and Bilingualism*, 10, 224-247.
- Gathercole, V. C. M., Thomas, E. M., Kennedy, I., Prys, C., Young, N., Viñas Guasch, N., . . . Jones, L. (2014). Does language dominance affect cognitive performance in bilinguals? Lifespan evidence from preschoolers through older adults on card sorting, Simon, and metalinguistic tasks. *Frontiers in psychology*, 5, 1-14.
- Gaux, C., & Gombert, J. (1999). Implicit and explicit syntactic knowledge and reading in pre-adolescents. *British Journal of Developmental Psychology*, 17, 169-188.
- Gernsbacher, M. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gernsbacher, M., & Foertsch, J. (1999). Three models of discourse comprehension. In S. Garrod & M. Pickering (Eds.), *Language processing* (p. 283-299). East Sussex: Psychology Press.
- Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, 26, 67-86.
- Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children*, 66, 454-470.
- Gersten, R., Baker, S., Shanahan, T., Linan-Thompson, S., Collins, P., & Scarcella, R. (2007). *Effective literacy and English language instruction for English learners in the elementary grades: A practice guide* (Tech. Rep.). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading & Writing*, 25, 1819-1845.
- Geva, E., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading & Writing*, 12, 1-30.
- Gibbons, R. D., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, 6, 79-107.
- Gillanders, C., Castro, D. C., & Franco, X. (2014). Learning words for life: Promoting vocabulary in Dual Language Learners. *Reading Teacher*, 68, 213-221.
- Gipe, J. (1978). Investigating techniques for teaching word meanings. *Reading Research Quarterly*, 14, 624-644.

- Goff, D. A., Pratt, C., & Ong, B. (2005). The relations between children's reading comprehension, working memory, language skills and components of reading decoding in a normal sample. *Reading & Writing Quarterly*, 18, 583-616.
- Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the Weaker Links Hypothesis. *Journal of Memory and Language*, 58, 787-814.
- Goswami, U. (2000). Phonological and lexical processes. In M. L. Kamil (Ed.), *Handbook of reading research* (Vol. 3, p. 251–267). Mahwah, NJ: Lawrence Erlbaum Associates.
- Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read*. Hove: Lawrence Erlbaum Associates.
- Gough, P., Hoover, W., & Peterson, C. (1996). Some observations on a Simple View of Reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (p. 1-14). New Jersey: Lawrence Erlbaum Associates.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6-10.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95-112.
- Gregory, E. (1996). Learning from the community – a family literacy project with Bangladeshi-origin children in London. In S. Wolfendale & K. Topping (Eds.), *Family involvement in literacy: Effective partnerships in education* (p. 89-102). London: Cassell.
- Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1, 131-149.
- Guasti, M. (2002). *Language acquisition: The growth of grammar*. Cambridge, Mass.: MIT Press.
- Guiraud, P. (1959). *Problemes et methodes de la statistique linguistique*. Dordrecht: Reidel.
- Hadley, E. B., Dickinson, D. K., Hirsh-Pasek, K., Golinkoff, R. M., & Nesbitt, K. T. (2016). Examining the acquisition of vocabulary knowledge depth among preschool students. *Reading Research Quarterly*, 51, 181-198.
- Hairrell, A., Rupley, W., & Simmons, D. (2011). The state of vocabulary research. *Literacy Research and Instruction*, 50, 253-271.
- Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take to English learners to attain proficiency?* (Tech. Rep.). Stanford University.
- Hakuta, K., & Diaz, R. (1985). The relationship between degree of bilingualism and cognitive ability: A critical discussion and some new longitudinal data. In K. Nelson (Ed.), *Children's language* (p. 319-344). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662-720.
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, Md.: P.H. Brookes.
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*, 369-388.
- Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.
- Heilmann, J., Miller, J. F., Nockerts, A., & Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology*, *19*, 154-166.
- Henson, R. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.
- Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., & Girolametto, L. (2015). Narrative abilities in subgroups of English language learners and monolingual peers. , *19*, 677-692.
- Hirst, K. (1998). Preschool literacy experiences of children in Punjabi-, Urdu-, and Gujerati-speaking families in England. *British Educational Research Journal*, *24*, 415-429.
- Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*, 1368-1378.
- Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*, 4-14.
- Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, *39*, 1-27.
- Hogan, T., Adlof, S., & Alonzo, C. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology*, *16*, 199-207.
- Hoover, W., & Gough, P. (1990). The Simple View of Reading. *Reading & Writing*, *2*, 127-160.
- Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read. *Psychological Science*, *23*, 572-577.
- Hulme, C., & Snowling, M. J. (2009). *Developmental disorders of language learning and cognition*. Chichester: Blackwell.
- Hutchinson, J. M., Whiteley, H. E., Smith, C. D., & Connors, L. (2003). The developmental progression of comprehension-related skills in children learning EAL. *Journal of Research in Reading*, *26*, 19-32.

- Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*, 236-248.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*, 69-82.
- IBM. (2013). *IBM SPSS statistics for Windows, v.22*.
- Jean, M., & Geva, E. (2009). The development of vocabulary in English as a second language children and its role in predicting word recognition ability. *Applied Psycholinguistics, 30*, 153-185.
- Jenkins, J. R., Matlock, B., & Slocum, T. A. (1989). Two approaches to vocabulary instruction: The teaching of individual word meanings and practice in deriving word meaning from context. *Reading Research Quarterly, 24*, 215-235.
- Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics, 24*, 131-61.
- Jobard, G., Crivello, F., & Tzourio-Mazoyer, N. (2003). Evaluation of the dual route theory of reading: A metaanalysis of 35 neuroimaging studies. *Neuroimage, 20*, 693-712.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*, 60-99.
- Jonejan, W., Verhoeven, L., & Siegel, L. (2007). Predictors of reading and spelling abilities in first- and second-language learners. *Journal of Educational Psychology, 99*, 835-851.
- Joshi, R. (2005). Vocabulary: A critical component of comprehension. *Reading & Writing Quarterly, 21*, 209-219.
- Joshi, R., & Aaron, P. (2000). The component model of reading: The Simple View made a little more complex. *Reading Psychology, 21*, 85-97.
- Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second Grade. *Journal of Educational Psychology, 78*, 243-255.
- Junker, D., & Stockman, I. (2002). Expressive vocabulary of German-English bilingual toddlers. *American Journal of Speech-Language Pathology, 11*, 381-394.
- Justice, L., Mashburn, A., & Petscher, Y. (2013). Very early language skills of fifth-grade poor comprehenders. *Journal of Research in Reading, 36*, 172-185.
- Kamil, M., & Hiebert, E. (2005). Teaching and learning vocabulary: Perspectives and persistent issues. In *Teaching and learning vocabulary: Bringing research to practice* (p. 1-23). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kappelhof, J. (2013). The effect of different survey designs on nonresponse in surveys among non-Western minorities in the Netherlands. *Survey Research Methods, 8*, 81-98.

- Karlsen, J., Lyster, S.-A. H., & Lervåg, A. (2017). Vocabulary development in norwegian L1 and L2 learners in the kindergarten–school transition. *Journal of Child Language, 44*, 402-426.
- Keith, M., & Nicoladis, E. (2013). The role of within-language vocabulary size in children's semantic development: Evidence from bilingual children. *Journal of Child Language, 40*, 873-884.
- Kendeou, P., Savage, R., & van den Broek, P. (2009). Revisiting the Simple View of Reading. *British Journal of Educational Psychology, 79*, 353-370.
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from Restricted Maximum Likelihood. *Biometrics, 53*, 983-997.
- Kessler, B., & Treiman, R. (2003). Is English spelling chaotic? misconceptions concerning its irregularity. *Reading Psychology, 24*, 267-289.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction—integration model. *Psychological Review, 95*, 163-182.
- Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.
- Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning, 57*, 1-44.
- Kotler, A., Wegerif, R., & LeVoi, M. (2001). Oracy and the educational achievement of pupils with English as an additional language: The impact of bringing 'Talking Partners' into Bradford schools. *International Journal of Bilingual Education and Bilingualism, 4*, 403-19.
- Krashen, S. (1989). We acquire vocabulary and spelling by reading - additional evidence for the Input Hypothesis. *Modern Language Journal, 73*, 440-464.
- Krashen, S., Long, M., & Scarcella, R. (1979). Age, rate and eventual attainment in second language acquisition. *TESOL Quarterly, 13*, 573-582.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978-990.
- Kuznetsova, A., Brockhoff, P., & Christensen, R. (2016). *lmerTest*. Retrieved from <https://cran.r-project.org/web/packages/lmerTest/index.html>
- Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.
- Language and Reading Research Consortium. (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly, 50*, 151-169.
- Laufer, B. (1989). A factor of difficulty in vocabulary learning: Deceptive transparency. In I. Nation & R. Carter (Eds.), *Vocabulary acquisition* (p. 10-20). Free University Press.
- Leitão, S., & Allan, L. (2003). *Peter and the cat narrative assessment*. Middleton: Black Sheep Press.

- Lepola, J., Lynch, J., Laakkonen, E., Silvén, M., & Niemi, P. (2012). The role of inference making and other language skills in the development of narrative listening comprehension in 4–6 year-old children. *Reading Research Quarterly, 47*, 259-282.
- Lervåg, A., & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry, 51*, 612-620.
- Lesaux, N. (2015). Reading development among English learners. In C. Connor & P. McCardle (Eds.), *Advances in reading intervention: Research to practice to research* (p. 155-166). Baltimore: Paul H. Brookes.
- Lesaux, N., & Siegel, L. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology, 39*, 1005-1019.
- Lesaux, N. K., Lipka, O., & Siegel, L. S. (2006). Investigating cognitive and linguistic abilities that influence the reading comprehension skills of children from diverse linguistic backgrounds. *Reading & Writing, 19*, 99-131.
- Leseman, P. P. M. (2000). Bilingual vocabulary development of Turkish preschoolers in the Netherlands. *Journal of Multilingual and Multicultural Development, 21*, 93-112.
- Leung, C. (2001). English as an additional language: Distinct language focus or diffused curriculum concerns? *Language and Education, 15*, 33-55.
- Liberman, I. Y., & Shankweiler, D. (1985). Phonology and the problems of learning to read and write. *Remedial and Special Education, 6*, 8-17.
- Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18*, 201-212.
- Liles, B. Z., & Purcell, S. (1987). Departures in the spoken narratives of normal and language-disordered children. *Applied Psycholinguistics, 8*, 185-202.
- Long, M. H. (1990). Maturation constraints on language development. *Studies in Second Language Acquisition, 12*, 251-285.
- Lothers, M., & Lothers, L. (2012). *Mirpuri immigrants in England: A sociolinguistic survey*.
- Lynch, J. S., van den Broek, P., Kremer, K. E., Kendeou, P., White, M. J., & Lorch, E. P. (2008). The development of narrative comprehension and its relation to other early reading skills. *Reading Psychology, 29*, 327-365.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Magezi, D. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology, 6*, 1-7.
- Mahon, M., & Crutchley, A. (2006). Performance of typically-developing school-age children with English as an additional language on the British Picture Vocabulary Scales II. *Child Language Teaching and Therapy, 22*, 333-351.

- Mancilla-Martinez, J. (2010). Word meanings matter: Cultivating English vocabulary knowledge in fifth-grade Spanish-speaking language minority learners. *TESOL Quarterly*, *44*, 669-699.
- Mandler, J. M., Scribner, S., Cole, M., & Deforest, M. (1980). Cross-cultural invariance in story recall. *Child Development*, *51*, 19-26.
- Manis, F. R., Doi, L. M., & Bhadha, B. (2000). Naming speed, phonological awareness, and orthographic knowledge in second graders. *Journal of Learning Disabilities*, *33*, 325-333.
- Martin-Chang, S., & Levesque, K. (2013). Taken out of context: Differential processing in contextual and isolated word reading. *Journal of Research in Reading*, *36*, 330-349.
- Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*, 300-335.
- Mayer, M. (1969). *Frog, Where Are You?* New York: Puffin.
- McCabe, A., & Rollins, P. (1994). Assessment of preschool narrative skills. *American Journal of Speech-Language Pathology*, *3*, 45-56.
- McCarthur Communicative Development Inventory*. (1989). San Diego: University of California, Center for Research in Language.
- McKendry, M. G., & Murphy, V. (2011). A comparative study of listening comprehension measures in English as an additional language and native English-speaking primary school children. *Evaluation & Research in Education*, *24*, 17-40.
- McKeown, M. G. (1985). The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly*, *20*, 482-496.
- McKeown, M. G. (1993). Creating effective definitions for young word learners. *Reading Research Quarterly*, *28*, 17-31.
- Meisel, J. M. (2011). *First and second language acquisition: Parallels and differences*. Cambridge: CUP.
- Meisinger, E., Bloom, J., & Hund, G. (2010). Reading fluency: Implications for the assessment of children with reading disabilities. *Annals of Dyslexia*, *60*, 1-17.
- Melby-Lervåg, M., & Lervåg, A. (2013). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*, 409-433.
- Merritt, D., & Liles, B. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *Journal of Speech and Hearing Research*, *30*, 539-552.
- Miller, J., & Iglesias, A. (2012). *Systematic Analysis of Language Transcripts (SALT)*. Middleton, WI: SALT Software, LLC.
- Murphy, V., & Unthiah, A. (2015). *A systematic review of intervention research examining English language and literacy development in children with English as an additional language (EAL)* (Tech. Rep.). University of Oxford.

- Murphy, V. A., Macaro, E., Alba, S., & Cipolla, C. (2015). The influence of learning a second language in primary school on developing first language literacy skills. *Applied Psycholinguistics*, *36*, 1133-1153.
- Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, *40*, 665-81.
- Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative models for small samples in psychological research: Applying linear mixed effects models and generalized estimating equations to repeated measures data. *Educational and Psychological Measurement*, *76*, 64-87.
- Nagy, W. (2005). Why vocabulary instruction needs to be long-term and comprehensive. In E. H. Hiebert & M. L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (p. 27-44). New York: Routledge.
- Nagy, W., & Scott, J. (2000). Vocabulary processes. In D. Pearson, R. Barr, & M. Kamil (Eds.), *Handbook of reading research* (Vol. 3, p. 269-284). Mahwah, NJ: Lawrence Erlbaum Associates.
- Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*, 233-253.
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133-142.
- NALDIC. (1999). *The distinctiveness of EAL: A cross-curriculum discipline* (Tech. Rep.).
- NALDIC. (2014). *The national audit of English as an additional language training and development provision*. Retrieved from <http://www.naldic.org.uk/Resources/NALDIC/Research%20and%20Information/Documents/NALDIC%20-%20EAL%20Audit%202014%20FINAL%20FINAL%20YF%200ct%2014.pdf>
- Nash, H., & Snowling, M. (2006). Teaching new words to children with poor existing vocabulary knowledge a controlled evaluation of the definition and context methods. *International Journal of Language & Communication Disorders*, *41*, 335-354.
- Nash, M., & Donaldson, M. L. (2005). Word learning in children with vocabulary deficits. *Journal of Speech, Language, and Hearing Research*, *48*, 439.
- NASSEA. (2001). *NASSEA EAL Assessment System*. Dukinfield: NASSEA.
- NASSEA. (2015). *The NASSEA Assessment Framework*.
- Nation, K., & Angell, P. (2006). Learning to read and reading to learn. *London Review of Education*, *4*, 77-87.
- Nation, K., Clarke, P., Marshall, C., & Durand, M. (2004). Hidden language impairments in children: Parallels between poor reading comprehension and specific language impairment? *Journal of Speech, Language, and Hearing Research*, *47*, 199-211.
- Nation, K., Cocksey, J., Taylor, J., & Bishop, D. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry*, *51*, 1031-1039.

- Nation, K., & Snowling, M. J. (2000). Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied Psycholinguistics*, *21*, 229-241.
- Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading*, *27*, 342-356.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: CUP.
- Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (p. 6-19). Cambridge: CUP.
- National Reading Panel. (2000). *Report of the national reading panel: Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (reports of the subgroups)* (Tech. Rep.). National Institute of Child Health and Human Development.
- Neale, M. (1997). *Neale Analysis of Reading Ability*. Windsor: NFER-Nelson.
- Nezlek, J. (2013). *Multilevel modeling for social and personality psychology*. Sage.
- Ng, B. C., & Wigglesworth, G. (2007). *Bilingualism: An advanced resource book*. London: Routledge.
- Nieuwenhuis, R., te Grotenhuis, M., & Pelzer, B. (2012). Influence.ME: Tools for detecting influential data in mixed effects models. *R Journal*, *4*, 38-47.
- Nist, S. L., & Olejnik, S. (1995). The role of context and dictionary definitions on varying levels of word knowledge. *Reading Research Quarterly*, *30*, 172-193.
- Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*, 443-468.
- Office for National Statistics. (2011). *2011 Census Analysis, Language in England and Wales*. Retrieved from <http://www.ons.gov.uk/ons/rel/census/2011-census-analysis/language-in-england-and-wales-2011/index.html>
- OFSTED. (2005). *Could they do even better? the writing of advanced bilingual learners of English at Key Stage 2: HMI survey of good practice* (Tech. Rep.).
- O'Grady, W. (1997). *Syntactic development*. Chicago: Chicago University Press.
- Ohanian, S. (2002). *The great word catalogue: FUNdamental activities for building vocabulary*. U.S.: Heinemann Educational Books.
- O'Keefe, B., Slocum, T., & Magnusson, R. (2013). The effects of a fluency training package on paraprofessionals' presentation of a reading intervention. *The Journal of Special Education*, *47*, 14-27.
- Osborne, J. W. (2008). *Best practices in quantitative methods*. [Electronic Resource]: Sage.
- Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading & Writing*, *23*, 189-208.

- Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98*, 554-566.
- Paap, K., & Greenberg, Z. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66*, 232-258.
- Palmer, B., Shackelford, V., Miller, S., & Leclere, J. (2006). Bridging two worlds: Reading comprehension, figurative language instruction, and the English-Language Learner. *Journal of Adolescent & Adult Literacy, 50*, 258-267.
- Paradis, J., & Genesee, F. (1996). Syntactic acquisition in bilingual children: Autonomous or interdependent? *Studies in Second Language Acquisition, 18*, 1-25.
- Paradis, J., Genesee, F., & Crago, M. B. (2010). *Dual language development and disorders: A handbook on bilingualism and second language learning*. Baltimore, MD: Paul H. Brookes.
- Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly, 38*, 36-76.
- Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*, 184-202.
- Parke, T., & Drury, R. (2001). Language development at home and school: Gains and losses in young bilinguals. *Early Years, 21*, 117-127.
- Pearson, B. (2002). Narrative competence among monolingual and bilingual school children in Miami. In K. Oller & R. Eilers (Eds.), *Language and literacy in bilingual children* (p. 135-174). Clevedon: Multilingual Matters.
- Pearson, B. Z., Fernández, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning, 43*, 93-120.
- Peng, C.-Y. J., & Chen, L.-T. (2013). Beyond Cohen's d: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education, 82*, 1-29.
- Perfetti, C. (1985). *Reading ability*. New York: OUP.
- Perfetti, C. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (p. 167-197). Oxford: OUP.
- Perfetti, C. (2003). A Universal Grammar of reading. *Scientific Studies of Reading, 7*, 3-24.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357-383.
- Perfetti, C., Marron, M., & Foltz, P. (1996). Sources of comprehension failure: Perspectives and case studies. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: processes and intervention* (p. 137-165). Mahwah, NJ: Lawrence Erlbaum Associates.
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*, 22-37.

- Peterson, C., & Dodsworth, P. (1991). A longitudinal analysis of young children' cohesion and noun specification in narratives. *Journal of Child Language, 18*, 397-415.
- Pinheiro, J. C., & Bates, D. M. (n.d.). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Plaut, D. (2005). Connectionist approaches to reading. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (p. 24-38). Oxford: Blackwell.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.
- Proctor, C., Dalton, B., Uccelli, P., Biancarosa, G., Mo, E., Snow, C., & Neugebauer, S. (2011). Improving comprehension online: Effects of deep vocabulary instruction with bilingual and monolingual fifth graders. *Reading & Writing, 24*, 517-544.
- Proctor, D. C., Carlo, M., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology, 97*, 246-256.
- Qualifications, & Authority, C. (2000). *A language in common: Assessing English as an additional language*. London: QCA.
- R Core Team. (2017). *R: A language and environment for statistical computing*.
- Rao, C., Vaid, J., Srinivasan, N., & Chen, H.-C. (2011). Orthographic characteristics speed Hindi word naming but slow Urdu naming: Evidence from Hindi/Urdu biliterates. *Reading & Writing, 24*, 679-695.
- Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice-Hall.
- Renfrew, C. (2003). *Action Picture Test*. Milton Keynes: Speechmark Publishing Ltd.
- Renfrew, C., Cowley, J., & Glasgow, C. (1994). *The renfrew bus story language screening by narrative recall (american edition)*. The Centerville School Delaware.
- Ricketts, J., Nation, K., & Bishop, D. V. M. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading, 11*, 235-257.
- Roman, A. A., Kirby, J., Parrila, R., Wade-Woolley, L., & Deacon, S. (2009). Toward a comprehensive view of the skills involved in word reading in Grades 4, 6, and 8. *Journal of Experimental Child Psychology, 102*, 96-113.
- Rosowsky, A. (2001). Decoding as a cultural practice and its effects on the reading process of bilingual pupils. *Language and Education, 15*, 56-70.
- Rosowsky, A. (2010). 'writing it in English': Script choices among young multilingual Muslims in the U.K. *Journal of Multilingual and Multicultural Development, 31*, 163-179.
- Ross, S., & Begeny, J. (2011). Improving latino, English language learners' reading fluency: The effects of small-group and one-on-one intervention. *Psychology in the Schools, 48*, 604-618.

- Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *Journal of Educational Research, 95*, 259-72.
- Rupley, W. H., & Nichols, W. D. (2005). Vocabulary instruction for the struggling reader. *Reading & Writing Quarterly, 21*, 239-260.
- Russell, D. H., & Saadeh, I. Q. (1962). Qualitative levels in children's vocabularies. *Journal of Educational Psychology, 53*, 170-174.
- Sacre, L., & Masterson, J. (2000). *Single Word Spelling Test*. London: nfer-Nelson.
- Scarborough, H. S. (1990). Very early language deficits in dyslexic children. *Child Development, 61*, 1728-1743.
- Schaefer, B., Fricke, S., Bowyer-Crane, C., Millard, G., & Hulme, C. (Under Review). Oral language intervention for children with EAL and monolingual peers with language weaknesses. *Language, Speech, and Hearing Services in Schools*.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal, 95*, 26-43.
- Schonell, F., & Goodacre, E. (1971). *The psychology and teaching of reading*. London: Oliver & Boyd.
- Schreiber, P. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior, 12*, 177-186.
- Scott, J. A., & Nagy, W. E. (1997). Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly, 32*, 184-200.
- Semel, E., Wiig, E., & Secord, W. (2006). *Clinical Evaluation of Language Fundamentals: Fourth U.K. Edition*. London: Pearson Assessment.
- Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143-174.
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.
- Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition, 55*, 151-218.
- Silverman, R. (2007). Vocabulary development of English-language and English-only learners in kindergarten. *The Elementary School Journal, 107*, 365-383.
- Silverman, R. D., Proctor, C. P., Haring, J. R., Doyle, B., Mitchell, M. A., & Meyer, A. G. (2014). Teachers' instruction and students' vocabulary and comprehension: An exploratory study with English monolingual and Spanish-English bilingual students in Grades 3-5. *Reading Research Quarterly, 49*, 31-60.
- Silverman, R. D., Proctor, C. P., Haring, J. R., Hartranft, A. M., Doyle, B., & Zelinke, S. B. (2015). Language skills and reading comprehension in English monolingual and Spanish-English bilingual children in Grades 2-5. *Reading & Writing, 28*, 1381-1405.

- Simos, P. G., Sideridis, G. D., Mouzaki, A., Chatzidaki, A., & Tzeveleki, M. (2014). Vocabulary growth in second language among immigrant school-aged children in Greece. *Applied Psycholinguistics, 35*, 621-647.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.
- Skutnabb-Kangas, T. (1981). *Bilingualism or not: The education of minorities*. Clevedon: Multilingual Matters.
- Smith, S., & Murphy, V. (2015). Measuring productive elements of multi-word phrase vocabulary knowledge among children with English as an additional or only language. *Reading & Writing, 28*, 347-369.
- Snow, C., Burns, M., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.
- Snow, C. E., Lawrence, J. F., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness, 2*, 325-344.
- Snowling, M. J., Stothard, S., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., ... Hulme, C. (2011). *York Assessment of Reading for Comprehension*. London: GL Assessment.
- Stahl, S. A. (1985). To teach a word well: A framework for vocabulary instruction. *Reading World, 24*, 16-27.
- Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*, 72-110.
- Stein, N., & Albro, E. (1997). Building complexity and coherence: Children's use of goal-structured knowledge in telling stories. In M. Bamberg (Ed.), *Narrative development: Six approaches* (p. 5-44). Mahwah, NJ: Lawrence Erlbaum Associates.
- Stein, N., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *New directions in discourse processing*. Norwood, NJ: Ablex.
- Sternberg, R. (1987). Most vocabulary is learned from context. In M. McKeown & M. Curtis (Eds.), *The nature of vocabulary acquisition* (p. 89-106). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Stothard, S., & Hulme, C. (1992). Reading comprehension difficulties in children. *Reading & Writing, 4*, 245-256.
- Stothard, S., & Hulme, C. (1996). A comparison of reading comprehension difficulties in children. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (p. 93-112). Mahwah, NJ: Lawrence Erlbaum Associates.
- Strand, S., Malmberg, L., & Hall, J. (2015). *English as an additional language (EAL) and educational achievement in England: An analysis of the National Pupil Database* (Tech. Rep.). University of Oxford.

- Sumption, M., & Somerville, W. (2010). *The U.K.'s new europeans: Progress and challenges five years after accession* (Tech. Rep.). Equality and Human Rights Commission & Migration Policy Institute.
- Swain, M., & Johnson, K. (1997). Immersion education: A category within bilingual education. In J. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (p. 1-16). Cambridge: CUP.
- Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69, 261-85.
- Swann, M. (1985). *Education for all: The report of the committee of inquiry into the education of children from ethnic minority groups* (Tech. Rep.). HMSO.
- Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, 10, 381-98.
- Taras, J., Meisinger, E., & Dickens, R. (2015). Test Review (TOWRE-2). *Canadian Journal of School Psychology*, 30, 320-326.
- Thomas, W., & Collier, V. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement* (Tech. Rep.). Centre for Research on Education, Diversity & Excellence.
- Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, 15, 426-445.
- Thordardottir, E. (2015). The relationship between bilingual exposure and morphosyntactic development. *International Journal of Speech-Language Pathology*, 17, 97-114.
- Tilstra, J., McMaster, K., Van Den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: components of the Simple View of reading across grade levels. *Journal of Research in Reading*, 32, 383-401.
- Tobia, V., & Bonifacci, P. (2015). The Simple View of reading in a transparent orthography: The stronger role of oral comprehension. *Reading & Writing*, 28, 939-957.
- Toolan, M. J. (2001). *Narrative: A critical linguistic introduction*. London: Routledge.
- Torgesen, J., Alexander, A., Wagner, R., Rashotte, C., Voeller, K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal Of Learning Disabilities*, 34, 33-78.
- Torgesen, J., Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency - Second Edition*. Austin: Pro-ED.
- Training and Development Agency. (2011). *Languages in training: English as an additional language (EAL)* (Tech. Rep.).
- Treffers-Daller, J., & Rogers, V. (2014). Knowledge of grammatical use. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (p. 106-122). Basingstoke: Palgrave Macmillan.
- Tsimpli, I. M. (2017). *Multilingual education for multilingual speakers (policy paper)* (Tech. Rep.). Multilingualism: Empowering Individuals, Transforming Soci-

- eties (MEITS). Retrieved from <http://www.meits.org/policy-papers/category/ianthi-maria-tsimpli>
- Tunmer, W. E., & Chapman, J. W. (2012). The Simple View of Reading Redux: Vocabulary knowledge and the Independent Components Hypothesis. *Journal of Learning Disabilities, 45*, 453-466.
- Ukrainetz, T. A., Justice, L. M., Kaderavek, J. N., Eisenberg, S. L., Gillam, R. B., & Harm, H. M. (2005). The development of expressive elaboration in fictional narratives. *Journal of Speech, Language, and Hearing Research, 48*, 1363-1377.
- Unsworth, S. (2013). Current issues in multilingual first language acquisition. *Annual Review of Applied Linguistics, 33*, 21-50.
- van Heuven, W., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal Of Memory And Language, 39*, 458-483.
- Vellutino, F., Scanlon, D., Small, S., & Tanzman, M. (1991). The linguistic basis of reading ability: Converting oral to written language. *Text, 11*, 99-133.
- Vellutino, F., Tunmer, W., Jaccard, J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*, 3-32.
- Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. New York: Springer.
- Verhallen, M., & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics, 14*, 344-363.
- Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*, 407-423.
- Verhoeven, L., & Van Leeuwe, J. (2012). The simple view of second language reading throughout the primary grades. *Reading & Writing, 25*, 1805-1818.
- Verhoeven, L. T. (1994). Transfer in bilingual development: The Linguistic Interdependence Hypothesis revisited. *Language Learning, 44*, 381-415.
- Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*, 65-83.
- Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*, 217-234.
- Viel-Ruma, K., Houchkins, D., Jolivet, K., Frederick, L., & Gama, R. (2010). Direct instruction in written expression: The effects on English speakers and English Language Learners with disabilities. *Learning Disabilities Research & Practice, 25*, 97-108.
- Wagner, R., & Torgesen, J. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*, 192-212.
- Wagner, R., Torgesen, J., & Rashotte, C. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology, 30*, 73-87.

- Wagner, R., Torgesen, J., & Rashotte, C. (1999). *Comprehensive Test of Phonological Processing*. Austin: Pro-ED.
- Wainwright, P. E., Leatherdale, S. T., & Dubin, J. A. (2007). Advantages of mixed effects models over traditional ANOVA models in developmental studies: A worked example in a mouse model of fetal alcohol syndrome. *Developmental Psychobiology*, *49*, 664.
- Wang, M., Park, Y., & Lee, K. R. (2006). Korean-English biliteracy acquisition: Cross-language phonological and orthographic transfer. *Journal of Educational Psychology*, *98*, 148-158.
- Wardman, C. (2013). Interactions between EAL pupils, specialist teachers and TAs during withdrawal from the mainstream in U.K. primary schools. *International Journal of Primary, Elementary and Early Years Education*, *41*, 647-663.
- Warmington, M., & Hulme, C. (2012). Phoneme awareness, visual-verbal paired-associate learning, and rapid automatized naming as predictors of individual differences in reading ability. *Scientific Studies of Reading*, *16*, 45-62.
- Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46-65.
- Webb, S., & Chang, A. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*, 651-675.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children - Fourth Edition*. San Antonio: Harcourt Assessment.
- Wechsler, D. (2005). *Wechsler Individual Achievement Test 2nd Ed*. Psychological Corporation.
- Wesche, M., Paribakht, S., & Harley, B. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *53*, 13-40.
- West, B. T., Welch, K. B., & Galecki, A. (2007). *Linear mixed models: A practical guide using statistical software*. London: Chapman & Hall/CRC.
- Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, *69*, 848-872.
- Whiteside, K. E., Gooch, D., & Norbury, C. F. (2017). English language proficiency and early school attainment among children learning English as an Additional Language. *Child Development*, *88*, 812-827.
- Wiig, E., & Secord, W. (1992). *Test of Word Knowledge*. Psychological Corporation Limited.
- Wilkinson, G. (1993). *Wide Range Achievement Test 3*. Psychological Corporation Limited.
- Wilkinson, K. S., & Houston-Price, C. (2013). Once upon a time, there was a pulchritudinous princess ...: The role of word definitions and multiple story contexts in children's learning of difficult vocabulary. *Applied Psycholinguistics*, *34*, 591-613.

- Willett, J. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*, 587-602.
- Wolf, M. (1991). Naming speed and reading: The contribution of the cognitive neurosciences. *Reading Research Quarterly*, *26*, 123-141.
- Wolf, M., Bally, H., & Morris, R. (1986). Automaticity, retrieval processes, and reading: A longitudinal study in average and impaired readers. *Child Development*, *57*, 988-1000.
- Wolf, M., Bowers, P. G., & Biddle, K. (2000). Naming-speed processes, timing, and reading: A conceptual review. *Journal of Learning Disabilities*, *33*, 387-407.
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, *5*, 211-239.
- Woodcock, R. (1987). *Woodcock Reading Mastery Test*. American Guidance Service.
- Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*, 203-226.
- Yoshikawa, H., Weisner, T. S., Kalil, A., & Way, N. (2008). Mixing qualitative and quantitative research in developmental science: Uses and methodological choices. *Developmental Psychology*, *44*, 344-354.
- Yuill, N., & Oakhill, J. (1991). *Children's problems in text comprehension: An experimental investigation*. Cambridge: Cambridge University Press.
- Zhao, J., Quiroz, B., Dixon, L., & Joshi, M. (2016). Comparing bilingual to monolingual learners on English spelling: A meta-analysis. *Dyslexia*, *22*, 193-213.
- Ziegler, J., Bertrand, D., Toth, D., Csepe, V., Reis, A., Faisca, L., . . . Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*, 551-559.