

**The effects of bilingualism, executive functioning, and numeracy on
children's semantic-pragmatic acquisition of logical quantifiers and
operators**

Haifa Eid Alatawi

Submitted in accordance with the requirements for the degree of Doctor of
Philosophy

The University of Leeds
School of Languages, Cultures and Societies

September, 2016

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

© 2016 The University of Leeds and Haifa Alatawi.

Acknowledgements

First and foremost, I thank you God for all the opportunities I had, for all the lessons I learned and for all the great people I met through this journey.

I feel very indebted to many people who helped me bring this project to life. This research would not have been possible without the wonderful people who agreed to participate. I feel extremely indebted to all the children who agreed to take part—collecting data from these little angels was the very best aspect of this project, and I am very thankful to all the parents who allowed their children to participate in my project. I am also very thankful to all the children's schools and head-teachers who allowed me to collect data: the Iraqi (Alrasoul Alazam), Libyan, and Saudi Arabic schools; Holy Trinity Church Primary School; Lily Croft Primary School; and King Abdul Aziz Model Schools Kindergarten. I am very thankful to all the adult participants for their cooperation and time. I would like also to express my sincere gratitude to Mr John Juniper; despite moving to France, he continuously supported me through my MA and PhD studies and generously tried to help me with my participants as much as he could. I was blessed to be one of your students, John; I feel I owe you a lot and I am very thankful for your support and friendship.

I am very grateful to my supervisors, Dr Catherine Davies and Prof Cecile De Cat for all their help, guidance and efforts. I really appreciate their thorough feedback on my work, and the time they spent reading through my long thesis. I have learnt a lot from them, and I am very thankful for this.

This project would not have included a reliable sample of English children without the help of Dr Napoleon Katsos. After contacting more than 170 schools, only one school agreed to cooperate with a limited sample so I thought I would have no hope of completing this research in the way I wished, and I was very fortunate in this regard to meet Dr Katsos at an academic event; he generously directed me to a school that would open its doors to me. I feel truly indebted to him, and will never forget his generous help.

I am also very thankful to all the staff in the Linguistics and Phonetics Department at Leeds, especially Dr Leendert Plug for his thorough feedback on my first-year report, from which I learned a lot. Great thanks also to all my colleagues at the university, especially those who allowed me to test their children and who directed me to schools that might agree to cooperate.

In addition, I am very thankful for the scholarship I received from the Saudi Ministry of Higher Education, and also to my family for covering the fees for my fieldwork inside and outside the UK.

Great thanks go to my dearest friends, who always stand beside me, Noor Elhuda Safwat and John and Kate Juniper. I thank you, Noor, for always being there to encourage and support me, even in the most difficult times that your country, Egypt, passed through during the time I was doing my PhD. To my dear friends John and Kate, for all their support and encouragement; you are really great people.

Finally, no words will suffice to express my gratitude to my amazing family, who all should be recognised as part of this project. I thank you Mum for doing your best to help me complete this project successfully. I thank you for your encouragement and your help with the schools and participants while I was conducting my research. You are literally the best. I am also very thankful to my best and most supportive friends: my wonderful siblings. I thank you for always being happy to listen to my theoretical ideas despite having different scientific backgrounds, and I am very thankful for your bright brains, which not only easily grasped my abstract ideas, but made compelling arguments against some of them.

I dedicate this work to the memory of my beloved father, whom I owe everything I have achieved and will achieve in future. Thank you, Dad, for all the great things that my siblings and I had learned from you. Thank you for teaching us the importance of a good education and for sharing with us your ideal vision of knowledge. Thank you for supporting all my decisions and for helping me to achieve my goals. Thank you, Dad, for everything.

Abstract

This research investigates children's semantic and pragmatic competence using the logical quantifiers 'most' and 'some', and the operators 'or' and 'and' in English and Arabic. It includes two main studies, with a sample of 30 Arabic-bidialectal, 26 English-monolingual, and 30 Arabic-and-English-bilingual pre-schoolers (mean age 5;6).

Study 1 explored the relationship between children's semantic comprehension of quantifiers and their numeracy skills, and asked two questions: a) do children comprehend quantifiers in semantically appropriate ways, and b) to what extent does acquisition of numerical system affect acquisition of quantifiers? The study applied two semantic tasks (perception v. production) and four numerical tasks (how-many, give-a-number, non-verbal ordinal, and estimating-magnitude-numerically). Most children showed very good numeracy skills; all performed better on the production than on the perception task, and Arabic children had significantly lower quantifier comprehension than the other groups. Ability on the give-a-number task (measuring ability to produce sets representing numerical values) had a significant effect on comprehension of 'some'.

Study 2 explored the relationship between pragmatic competence and bilingualism, with a focus on scalar implicature. It asked whether any superior pragmatic competence in bilinguals is due to a cognitive advantage over monolinguals. It applied two ternary-response judgment tasks to assess pragmatic ability in two conditions (enriched context v. no context), and two cognitive tasks: an inhibitory control task and a short-term memory task. A bilingual advantage was found only on pragmatic, not cognitive, tasks; however, cognitive tasks had strong effects on pragmatic performance. These results are discussed vis-à-vis theories of implicature processing.

The main contributions of this research are to a) theoretically establish how quantifiers and numbers are associated by linking theories of abstract and number word representations and then testing this relation empirically, b) show that the bilingual advantage emerges in both English and Arabic, and c) provide evidence that implicature processing is cognitively effortful.

Table of Contents

Acknowledgements	iii
Abstract	v
Table of Contents	vi
List of Tables	x
List of Figures	xiii
List of Abbreviations	xv
Notes on the Transliteration	xvi
Chapter 1	1
Introduction	1
1.1 Overview	1
1.1.1 Study 1: Children’s comprehension of quantifiers and operators and the potential effect of numeracy	1
1.1.2 Study 2: The potential effect of bilingualism on pragmatic competence	3
1.2. Context of the current research	5
1.3 Outline of the thesis	6
Chapter 2	9
Literature Review	9
2.1 Introduction	9
2.2 Numeracy and quantifier comprehension	10
2.2.1 Empirical findings on children’s acquisition of numbers and quantifiers.....	12
2.2.1.1 Acquisition of numbers.....	12
2.2.1.2 Acquisition of quantifiers	15
2.2.1.3 Which is acquired first, quantifiers or numbers, and why?.....	16
2.2.2 The relationship between numbers and quantifiers.....	17
2.2.2.1 Quantifiers as abstract concepts.....	18
2.2.2.2 Mental representation of the numerical system and its possible role in the acquisition of quantifiers.....	21
2.2.2.3 Neural basis of quantifiers and numbers.....	23
2.2.3 Factors that might affect children’s acquisition of numbers and quantifiers .	26
2.2.4 A summary of the research on development of numbers and quantifiers.....	28
2.3 Pragmatic competence	29
2.3.1 Pragmatic theories.....	29
2.3.1.1 Grice’s theory of implicature.....	29
2.3.1.2 Theories on implicature processing.....	35
2.3.2 Empirical evidence on scalar implicature processing at the surface and neural levels.....	38
2.3.2.1 Scalar implicature	39
2.3.2.2 Types of scale.....	40
2.3.2.3 Evidence on scalar implicature processing at the surface level	42
2.3.2.4 Measuring the cost of processing by rate of pragmatically enriched responses: Evidence from bilinguals	49
2.3.2.5 Measuring the cost of processing by neural activation	53
2.3.3. Acquisition of pragmatics	58
2.3.3.1 Children’s comprehension of scalar implicature.....	58
2.3.4 A summary of pragmatic competence.....	63

2.4 The impact of bilingualism on pragmatic and cognitive abilities	64
2.4.1 Bilingualism and cognitive development.....	64
2.4.2 Empirical findings of an effect of bilingualism on executive function.....	65
2.4.3 Understanding the reason for the scantiness of evidence of a bilingualism effect.....	67
2.4.4 Bilingualism and pragmatic competence	70
2.4.4.1 Is there a bilingual pragmatic advantage?.....	71
2.4.4.2 How can the superior pragmatic performance of bilinguals be explained in terms of their EF abilities?.....	76
2.4.5 Bilingualism, executive functioning (EF), and theory of mind	79
2.4.6 Bilingualism, language proficiency, and socio-economic status (SES).....	82
2.4.7 A summary of bilingualism’s impact.....	83
2.5 The current research.....	84
2.6 Chapter summary	85
Chapter 3	87
Methodology	87
3.1 Introduction	87
3.2 Philosophical approach to research.....	88
3.2.1 Justification for the study’s methodological approach.....	89
3.2.2 Empirical methods in social science research: Benefits and challenges	89
3.3 The research design.....	91
3.3.1 Pilot study	92
3.3.2 Sampling.....	93
3.3.3 Location of testing.....	93
3.3.4 Ethical issues	94
3.3.5 Data analysis.....	94
3.4 Method	95
3.4.1 Participants	95
3.4.2 Materials and procedure	98
3.4.3 Controlling for confounding variables	98
3.4.3.1 Language proficiency test.....	98
3.4.3.2 Non-verbal IQ test: Matrix Reasoning	101
3.4.3.3 Socio-economic status (SES) measure	101
3.4.3.4 Language background	104
3.4.4 Study 1: Children’s comprehension of quantifiers and operators and the potential effect of numeracy	105
3.4.4.1 Semantic performance.....	106
3.4.4.2 Number tasks	109
3.4.5 Study 2: The potential effect of bilingualism on pragmatic competence	115
3.4.5.1 Pragmatic performance	116
3.4.5.2 Cognitive performance	122
3.5 Summary.....	127
Chapter 4	129
Results and Analyses	129
4.1 Introduction	129
4.2 Background characteristics	130
4.2.1 Child participants	130
4.2.1.1 General measures.....	130
4.2.1.2 Child participants’ socio-economic status (SES)	133
4.2.1.3 Language measures.....	136
4.2.2 Adult participants.....	148
4.2.3 A summary of background measures	149

4.3 Results of Study 1: Children’s comprehension of quantifiers and operators and the potential effect of numeracy	150
4.3.1 Semantic performance	150
4.3.1.1 Give-a-quantifier task (experiment 1)	151
4.3.1.2 Estimating-magnitude-proportionally task (experiment 2)	160
4.3.1.3 Children’s performance on ‘some’ and ‘most’: Perception v. production.....	167
4.3.1.4 A summary of semantic performance	172
4.3.2 Performance on number tasks.....	174
4.3.2.1 Results for the how-many task.....	174
4.3.2.2 Results for the give-a-number task	177
4.3.2.3 Non-verbal ordinal task.....	180
4.3.2.4 Estimating-magnitude-numerically task	180
4.3.2.5 The potential effect of pre-school learning on children’s numeracy skills	182
4.3.2.6 A summary of performance on number tasks	184
4.3.3. The relationship between numbers and quantifiers.....	186
4.4 Results of Study (2): The potential effect of bilingualism on pragmatic competence.....	189
4.4.1 Pragmatic performance: Ternary-response investigation	189
4.4.1.1 Results for experiment 3 (enriched context).....	189
4.4.1.2 Results for experiment 4 (no context)	199
4.4.1.3 Enriched context v. no context	207
4.4.1.4 A summary of pragmatic performance (based on ternary responses).....	226
4.4.2 Pragmatic performance: Another way of analysing data.....	228
4.4.2.1 Experiment 3 (enriched context).....	230
4.4.2.2 Experiment 4 (no context)	234
4.4.2.3 Children’s performance on different scales: Horn, ad hoc, and encyclopaedic....	238
4.4.2.4 A summary of the new pragmatic analyses	244
4.4.3 Cognitive performance.....	245
4.4.3.1 The Simon task.....	246
4.4.3.2 The Corsi blocks task.....	257
4.4.3.3 A summary of cognitive performance	260
4.4.4 The relationship between children’s pragmatic and cognitive performance	261
4.5 A summary of results.....	265
4.5.1 A summary of participants’ basic measures	265
4.5.2 A summary of semantic results	266
4.5.3 Number task results.....	267
4.5.4 A summary of pragmatic results.....	268
4.5.4.1 Performance with ternary responses.....	269
4.5.4.2 Performance with binary responses.....	270
4.5.5 Cognitive task results.....	271
4.5.6 Chapter summary.....	272
Chapter 5	273
Discussion.....	273
5.1 Introduction	273
5.2 Study 1: Children’s comprehension of quantifiers and operators and the potential effect of numeracy	273
5.2.1 Performance on semantic tasks	274
5.2.1.1 Semantic performance on the comprehension task.....	274
5.2.1.2 Semantic performance on the production task.....	276
5.2.1.3 Perception v. production.....	277
5.2.2 Performance on number tasks.....	279
5.2.2.1 Acquisition of the exact numerical system.....	279
5.2.2.2 Acquisition of the approximate numerical system.....	282
5.2.3 The relationship between numbers and quantifiers.....	283

5.3 Study 2: The potential effect of bilingualism on pragmatic competence ...	287
5.3.1 Pragmatic performance	288
5.3.1.1 Children v. adults.....	289
5.3.1.2 Is there a bilingual pragmatic advantage?.....	299
5.3.2 Cognitive performance.....	302
5.3.3 The relationship between pragmatic and cognitive advantage	305
5.3.4 Implications for implicature processing theories	307
5.4 Summary of discussion.....	312
Chapter 6	314
Conclusion	314
6.1 Introduction	314
6.2 Summary and main findings	314
6.3 Contribution and possible implications of the current research	318
6.3.1 Contribution of Study 1: The relationship between numbers and quantifiers	318
6.3.2 Contribution of Study 2: The relationship between bilingualism, EF, and pragmatic competence	319
6.4 Limitations.....	320
6.5 Future work.....	322
References	323
Appendices.....	348
Appendix 1. Language exposure questionnaire	348
Appendix 2. Item list for Experiment 3	355
Appendix 3. Sample stimuli used for Experiment 3.....	359
Appendix 4. Item list for Experiment 4	361
Appendix 5. Predictors for pragmatic performance: Ordinal logistic regression 	363

List of Tables

Table 3.1. Information on tested child and adult participants in each group	96
Table 3.2. Sample stimuli from the no-context (pragmatic) ternary-judgement task (experiment 4)	122
Table 4.1. Background characteristics of bilingual, Arabic-speaking, and English-speaking children.....	131
Table 4.2. Parents' level of education in each child group, classified according to the UNESCO international education scale	133
Table 4.3. Child participants' SES by the international Family Affluence Scale	135
Table 4.4. Number of bilingual and Arabic children who have limited exposure to either of the two languages/dialects	140
Table 4.5. Average age at first exposure to languages/dialects in bilingual and Arabic children	144
Table 4.6. Individual results for first exposure to the second language/dialect in bilingual and Arabic groups	145
Table 4.7. Background features for the Arabic and English adults.....	148
Table 4.8. Criteria for correct (adult-like) responses for each quantifier/operator within a given set size.....	152
Table 4.9. Number, average age, and performance of children who provided wrong responses consistently (in all three trials for each quantifier/operator) within each group.....	157
Table 4.10. Arabic and English adults' responses for the quantifiers 'most' and 'some' in the give-a-quantifier task.....	159
Table 4.11. GLM regression results: Exploring the relationship between children's perception and production of the quantifiers 'most' and 'some'	171
Table 4.12. A summary of within-group differences in the perception and production of 'most' and 'some' (experiment 1 v. experiment 2).....	173
Table 4.13. Number and percentage of participants in each child group who correctly counted set sizes of {10} and {14} in the how-many task.....	175
Table 4.14. Performance and age of children who could not complete the how-many task accurately	176
Table 4.15. Number and percentage of cardinal-principle-knowers in each child group based on the results of give-a-number task	178
Table 4.16. Individual results for Arabic children who could not complete the how-many and/or give-a-number tasks successfully	179
Table 4.17. GLM regression results: Predictors for children's semantic performance (experiment 1).	188
Table 4.18. Bilingual, Arabic and English children's responses (as a percentage of each type of utterance) in experiment 3.....	193

Table 4.19. Percentages of Arabic and English adults' ternary responses in the three conditions of experiments 3.....	197
Table 4.20. Bilingual, Arabic and English children's responses (as a percentage of each type of utterance) in experiment 4.....	201
Table 4.21. Breakdown of Arabic and English adults' ternary responses over the three conditions of experiment 4	206
Table 4.22. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): Bilingual children in English	209
Table 4.23. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): Bilingual children in English.....	210
Table 4.24. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): Bilingual children in English.....	211
Table 4.25. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): Bilingual children in Arabic.....	212
Table 4.26. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): Bilingual children in Arabic	213
Table 4.27. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): Bilingual children in Arabic	214
Table 4.28. Experiment 3*experiment 4 cross-tabulation (under-informative v. Infelicitous): Arabic children (monolingual).....	215
Table 4.29. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): Arabic children (monolingual)	216
Table 4.30. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): Arabic children (monolingual)	217
Table 4.31. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): English children (monolingual)	218
Table 4.32. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): English children (monolingual).....	219
Table 4.33. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): English children (monolingual).....	220
Table 4.34. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): Arabic adults	221
Table 4.35. Experiment 3*experiment 4 cross-tabulation: Arabic adults (optimal v. felicitous).....	222
Table 4.36. Experiment 3*experiment 4 cross-tabulation: Arabic adults (false v. bizarre).....	223
Table 4.37. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): English adults.....	224
Table 4.38. Experiment 3*experiment 4 cross-tabulation: Arabic adults (optimal v. felicitous).....	225
Table 4.39. Experiment 3*experiment 4 cross-tabulation: English adults (false v. bizarre).....	226
Table 4.40. A summary of the groups' performance in the parallel conditions of experiments 3 and 4 (context v. no context).....	228

Table 4.41. Summary of children’s pragmatic performance on Horn and ad hoc scales	245
Table 4.42. Coefficients of a Cox proportional hazard model fitted to the time to correct response	253
Table 4.43. Coefficients of a GLM regression fitted to Corsi Blocks score	259
Table 4.44. GLM regression results: Predictors for children’s pragmatic performance on under-informative items in the two context conditions.	264
Table 4.45. A summary of participants’ basic measures	266
Table 4.46. A summary of between-group differences in performance in experiments 1 and 2	267
Table 4.47. A summary of participants’ performance on the number tasks.....	268
Table 4.48. A summary of between-group comparisons of pragmatic performance in the under-informative/infelicitous condition between experiments 3 and 4 (ternary responses).....	269
Table 4.49. A summary of between-group comparisons on pragmatic performance in the under-informative condition in experiments 3 and 4 (binary response).....	271
Table 4.50. A summary of participants’ performance on the cognitive tasks	272

List of Figures

Figure 3.1. Sample item from the estimating-magnitude-proportionally task	108
Figure 3.2. Stimuli used in the how-many task	111
Figure 3.3. Sample item from the non-verbal ordinal task	113
Figure 3.4. Sample item from the estimating-magnitude-numerically task	115
Figure 3.5. A sample of the (enriched context) ternary-judgement task	118
Figure 3.6. Sample trial from the Simon task (incongruent condition)	124
Figure 3.7. Sample display from the Corsi blocks task	126
Figure 4.1. Results for family wealth using the Family Affluence Scale (FAS)	135
Figure 4.2. Children's average raw scores in the receptive vocabulary test	137
Figure 4.3. Average ratio of Arabic-English use in the bilingual child group and Colloquial-Standard Arabic in the Arabic child group	140
Figure 4.4. Proportions of correct (adult-like) responses given for each quantifier/operator in the bilingual (English, Arabic), Arabic, and English children's groups	153
Figure 4.5. Breakdown of bilingual children's use of the quantifiers 'most' and 'some' to describe various proportions	161
Figure 4.6. Breakdown of English children's use of the quantifiers 'most' and 'some' to describe various proportions	161
Figure 4.7. Breakdown of Arabic children's use of the quantifiers 'most' and 'some' to describe various proportions	162
Figure 4.8. Percentages of total correct responses (adult-like choice) for 'most' and 'some' in the estimating-magnitude task by the child groups	164
Figure 4.9. Breakdown of Arabic adults' use of the quantifiers 'most' and 'some' to describe various proportions	166
Figure 4.10. Breakdown of English adults' use of the quantifiers 'most' and 'some' to describe various proportions	167
Figure 4.11. Percentages of children's correct (adult-like) responses given for 'most' in experiment 1 (perception) v. experiment 2 (production)	168
Figure 4.12. Percentages of children's correct (adult-like) responses given to 'some' in experiment 1 (perception) v. experiment 2 (production)	169
Figure 4.13. Average numeral response by set size for bilingual, Arabic, and English children in the estimating-magnitude-numerically task	181
Figure 4.14. Average numeral by set size for Arabic and English adults in the estimating-magnitude-numerically task	182
Figure 4.15. Children's age at entering pre-school and their traditional length of pre- schooling	184
Figure 4.16. Child groups' average percentages of correct penalisation of the under- informative items in experiment 3	231

Figure 4.17. Adult groups' average percentages of correct penalisation of the under-informative items in experiment 3	234
Figure 4.18. Child groups' average percentages of correct penalisation of the under-informative items in experiment 4	235
Figure 4.19. Adult groups' average percentages of correct penalisation for the under-informative items in experiment 4	237
Figure 4.20. The child groups' pragmatic performance (%) on the Horn scale v. the ad hoc scale in experiment 3	239
Figure 4.21. Child groups' performance (%) on the two scales (Horn v. encyclopaedic) in experiment 4	240
Figure 4.22. Child groups' performance (%) on Horn lexical scale in experiment 3 v. experiment 4	241
Figure 4.23. Child groups' performance (%) on the ad hoc v. the encyclopaedic scale in experiment 3 versus experiment 4	243
Figure 4.24. Child participants' average accuracy in the Simon task (by condition)	247
Figure 4.25. Average RT of correct responses in the congruent and incongruent trials of the Simon task for the bilingual, Arabic, and English children	248
Figure 4.26. Average global RT of correct responses in the Simon task for the bilingual, Arabic, and English children	250
Figure 4.27. Average RT of the Simon effect for the bilingual, Arabic, and English children	251
Figure 4.28. Adult participants' average accuracy on the Simon task	255
Figure 4.29. Average RT of correct responses in congruent and incongruent Simon task trials for Arabic and English adults	255
Figure 4.30. Average global RT of correct responses in the Simon task trials for Arabic and English adults	256
Figure 4.31. Average RT of the Simon effect for the Arabic and English adults	257
Figure 4.32. Average scores of bilingual, Arabic, and English children on the Corsi blocks task as a measure of STM span	258
Figure 4.33. Average scores of Arabic and English adults on the Corsi block task as a measure of short-term memory span	259

List of Abbreviations

ANT	Attentional Network Task
AQ	Autism-Spectrum Quotient
BPVS	British Picture Vocabulary Scale
CVT	Conversational Violations Test
D1	First dialect
D2	Second dialect
DL	Dominant Language
EF	Executive Functioning
ERP	Event-related potential
FAS	Family Affluence Scale
fMRI	Functional magnetic resonance imaging
GCI	Generalised Conversational Implicature
GLM	Generalised linear model
ISCED	International Standard Classification of Education
L1	First language
L2	Second language
LN/ AM	Letter Number/ Animal Music
LTM	Long-term memory
MMN	Mismatch negativity
NVIQ	Non-verbal IQ
PCI	Particularised Conversational Implicature
PH	Proportional Hazard
PPVT	Peabody Picture Vocabulary Test
RT	Reaction time
SES	Socioeconomic Status
STM	Short-term memory
ToM	Theory of Mind
Under-info	Under-informative
WM	Working memory

Notes on the Transliteration

The transliterations of the Arabic sounds used in the study are as follows:

Emphatics: /d̥/ (= ض), /d̥ˤ/ (= ظ)

Pharyngeal: /ħ/ (= ح), /ʕ/ (= ع)

Post alveolar affricate: (= ج) is transcribed as /j/ in the study; however, in literature this sound is either transcribed as /ǧ/ or /j/

Glottal: /ʔ/ (= ء)

The glide (ي) is transcribed as /y/, following the Arabic dialectal tradition.

Voiced labial-velar approximant, /w/, is transcribed as /w/.

Long vowels are phonologically transcribed by doubling the vowel itself (e.g. /aa/).

Chapter 1

Introduction

1.1 Overview

The current research was first designed to test children's pragmatic ability to derive implicatures when certain scalar expressions (quantifiers) are used in a context (e.g. <*all, most, some*>). Since ability to derive implicatures from quantifiers used in a context depends, at least partially, on semantic comprehension of these quantifiers, the research also explored how children comprehend the semantic meanings of these terms. In early investigations (i.e. in the pilot study) as the project was being developed, some of the children showed very poor comprehension of 'most' and 'some'; thus, to understand possible reasons for this weak semantic performance, and given the possible relationship between numeracy skills and children's comprehension of quantifier terms (e.g. Barner, Chow, & Yang, 2009), I attempted to establish theoretically how and why quantifiers and numbers might be associated, and then to examine this potential relationship empirically by including tasks measuring numeracy skills. Thus, the current project encompasses two studies, on the same participants. The first study explored how children understand the logical quantifiers *most* and *some*, the operators *or* and *and*, and their Arabic equivalents, and also examined to what extent children's numeracy skills might affect their semantic comprehension of the quantifiers. The second study explored children's pragmatic ability to derive implicatures, and the potential effects of bilingualism and executive functioning (EF) abilities (specifically, inhibition and short-term memory [STM]) on pragmatic performance. Below, a fuller description of each study is given.

1.1.1 Study 1: Children's comprehension of quantifiers and operators and the potential effect of numeracy

This study was built on my own observations and the conclusions of existing literature about the relationship between quantifiers and numbers. I found that although some studies have referred to such a relationship, there has been no clear

explanation of how and why numbers and quantifiers are associated; thus, I attempted to establish a clear link between the two. To do so, I proposed a novel way to link the two systems, based on theories of the representation of abstract words (e.g. Paivio, Yuille, & Madigan, 1986; Louwrese & Jeuniaux, 2010) and the representation of numbers (e.g. Brannon, Wusthoff, Gallistel, & Gibbon, 2001; Dehaene, 2003; Cordes & Gelman, 2005). I also tried to establish the existence and nature of the possible role of the approximate numerical system, more precisely scalar variability (ability to show numerical variability when evaluating magnitudes without counting). It is the first study to try to support such a theoretical connection between the quantificational and numerical systems, which it does by exploring the neural basis for each system, and finding that the areas of the brain activated by tasks measuring numerical values approximately are similar to those involved in quantifier processing. This finding further supports the connection between the two systems. In addition, based on existing findings (Barner et al., 2009), I suggested that the comparison of children's performance on numerical tasks to that on quantifiers shows that numbers are acquired first—a claim that goes against all previous assumptions that quantifiers are acquired first (e.g. Carey, 2004; Barner et al., 2009; Piantadosi, Tenenbaum, & Goodman, 2012).

Thus, taking into account all the previous empirical findings and my own observations, the first study in this thesis aimed to answer two basic questions:

- a) Do bilingual and monolingual children comprehend the quantifiers 'most' and 'some', and the operators 'and' and 'or' in a semantically appropriate (adult-like) way?**
- b) Does numerical system acquisition promote or (possibly) hinder the acquisition of quantifiers, and to what extent?**

To explore children's semantic comprehension of quantifiers, two semantic tasks were employed: a perception task and a production task. To explore children's numeracy skills, the study adopted measures that assessed both the exact and approximate numerical systems.

1.1.2 Study 2: The potential effect of bilingualism on pragmatic competence

Despite the increased evidence of bilingualism's positive influence on human cognitive ability (e.g. Bialystok, 2011; Blumenfeld & Marian, 2011), evidence of its potential effect on (linguistic) pragmatic ability is still scant (Siegal, Matsuo, Pond, & Otsu, 2007; Siegal, Iozzi, & Surian, 2009), and the findings of recent work are still inconsistent in this regard (Antoniou, Katsos, Grohmann, & Kambanaros, 2014). This study aimed to explore the potential effect of bilingualism on pragmatic ability, more precisely how the indirect effect of bilingualism, as assessed by EF abilities, might be reflected in children's pragmatic competence to derive implicatures.

The term (*conversational*) *implicature* was coined by Grice (1975, 1989) to refer to what is implicated (and pragmatically inferred) by an utterance as opposed to what is literally said (linguistically coded). Grice defined certain 'maxims' that interlocutors are said to consider in conversational exchange; these identify principles said to be followed by people engaged in communicative exchange, such as being informative (Maxims of Quantity I (do not be under-informative) and II (do not be over-informative)), honest (Maxim of Quality), relevant (Maxim of Relevance), and clear (Maxim of Manner). If the speaker violates one of these maxims, according to Grice, the hearer will assume that this was done intentionally, since the speaker is presumed to be cooperating competently and in good faith in the construction of communication (the Cooperative Principle). Given these assumptions, the hearer should, in a successful communicative interaction, be able to infer the meaning of the speaker's implicature.

This study focused on a particular type of implicature, known as *scalar implicature*, which is assumed to require pragmatic sensitivity to the Quantity I maxim. The study takes up scalar implicatures under the assumption that they reflect pragmatic sensitivity to violation of the Gricean maxims. Below is an example of a scalar implicature generated using the weaker term on a lexical scale.

(1)

(a) *Some* of the students passed the exam.

→(a*) *Not all* of them passed (scalar implicature)

Although the different types of scale investigated in this study are discussed in chapter 2, it is useful here to explain the idea of the *lexical scale*, introduced by Horn (1972). Horn proposed that lexical terms are organised in scales that include a set of alternative expressions that are of the same grammatical category but vary in semantic informativeness. The main hypothesis regarding these scalar terms is that the use of a semantically weaker term implicates that the stronger one does not hold (Dieussaert, Verkerk, Gillard, & Schaeken, 2011). This proposition is compatible with Grice's (1975, 1989) first Maxim of Quantity (do not be under-informative). A few examples of items organised from strong to weak along such scales are <*all, most, some*>; <*and, or*>; and <*excellent, good, acceptable*>. The study also explored children's sensitivity to informativeness on two other types of scale: *ad hoc scales*, which generate context-dependent implicatures (e.g. <*Sara*>, <*Jane*>, <*Sara and Jane*>), and *encyclopaedic scales*, where implicatures are licensed by world knowledge (e.g. to clean your teeth, you need <*toothpaste*>, <*toothbrush*>, <*toothpaste and toothbrush*>) (Papafragou & Tantalou, 2004).

Another goal of the second study is to attempt to understand the computation process and cognitive effort that may be involved in implicature computation, by explaining children's pragmatic performance (rate of correct pragmatic responses) in light of two pragmatic theories of implicature processing: the *Default hypothesis* (Levinson, 2000) and the *Relevance (Context-Dependent) theory* (Sperber & Wilson, 1986/1995). I will briefly define these theoretical accounts here, and explain them in further detail in chapter 2. The Default hypothesis of implicature processing suggests that a rapid, automatic mechanism is used to process utterances such as *some of his family are attending the wedding*, which infers that not all of his family are attending—an inference subject to cancellation if additional contextual information is provided (e.g. adding *actually, they are all attending*). In contrast, the Relevance hypothesis suggests that only context-dependent inferences are computed and that this process is cognitively effortful. For instance, in the given example, according to the Relevance account, the hearer would not infer that *not all of them are attending*, since contextual assumptions are made from the beginning. To test which theoretical account of implicature processing is more plausible, the second study will explore the relationship between children's pragmatic performance and their EF abilities in order to understand the computation process from the perspective of cognitive effort. The

main question that the second study is aiming to answer is as follows:

c) Can any superior pragmatic competence in bilingual children be explained in terms of a cognitive advantage over monolinguals?

To answer this question, the study applied two ternary-response judgment tasks, to assess children's pragmatic ability in two conditions: an enriched-context condition v. a no-context condition. To assess children's cognitive ability, the study used two cognitive tasks: an inhibitory control task and a visuospatial STM task.

1.2. Context of the current research

Having explained briefly above the main scope of, and the motivation for, conducting the two studies presented here, I will now outline the context of the current research. I will briefly describe the sample in terms of language background (exposure to different languages) and highlight some features of the languages involved in this study. The study investigated 30 Arabic-bidialectal (i.e. exposed to two dialects: Colloquial and Standard Arabic), 26 English-monolingual, and 30 Arabic-and-English-bilingual pre-schoolers (mean age 5;6). Two adult groups (Arabic and English native speakers; mean ages 22;2 and 20;7, respectively) were employed as control groups.

As the participants' background implies, Arabic and English are the languages investigated in this research. The Arabic language is a Semitic language with a complex and systematic morphology which is based on derivations of tri-consonantal root morphemes (Boudelaa, Pulvermüller, Hauk, Shtyrov, & Marslen-Wilson, 2010). In contrast, English, a Germanic language, has a mostly linear morphology, based on the addition of prefixes or suffixes to a base morpheme to produce multi-morphemic words (Bick, Goelman, & Frost, 2011). These differences might not be vital for the current research, but there is one distinct feature that we are certainly concerned about: the diglossic nature of Arabic (as distinct from English). *Diglossia* refers to the existence of two varieties of the same language: Standard Arabic (representing the 'high' or 'prestige' variety, the formal dialect used in all written documents and media and in educational materials) and Colloquial Arabic (representing the 'low'

variety(/ies) used in everyday conversation) (Versteegh, 2001; Miller, 2007). What is important here is that this dual linguistic situation, in which Standard and various dialects of Colloquial Arabic are both in regular use across the Arab world, already shares some features with bilingualism per se, even before the acquisition of a foreign language.

For a more accurate judgment of whether bilinguals indeed have a pragmatic advantage over (bidialectal) monolingual Arabic-speaking children, it was necessary to include monolingual English-speaking children as well—not only because they are exposed to only one language, but also because I want to investigate whether the bilingual pragmatic advantage might be exclusive to one of the languages in question.

1.3 Outline of the thesis

The thesis will be organised as follows.

Chapter 2 presents the main theoretical concepts and empirical findings in three research areas relevant to the current work. The first section explores children's acquisition of quantifiers and numbers, and explains theoretically how the two might be associated. The second section introduces pragmatic theories on implicatures and empirical findings on implicature processing and children's acquisition of pragmatics. The third section explores the findings of prior work on bilingual cognitive and pragmatic performance.

Chapter 3 presents the research design and methodology, explaining each measure taken and the rationale for using it. I briefly list these methods here. First, basic measures were taken to control for language proficiency (receptive vocabulary) in English and in Arabic, general intellectual ability (a non-verbal IQ (NVIQ) test), socioeconomic status (SES), and language exposure (the latter two measured by questionnaires completed by parents). Semantic comprehension of quantifiers was assessed using two semantic tasks, respectively testing perception (the give-a-quantifier task adapted from Hanlon, 1987; Barner et al., 2009) and production (the estimating-magnitude-proportionally task, adapted from Yildirim, Degen, Tanenhaus, & Jaeger, 2016). Numeracy skills were assessed by employing four numerical tasks:

How-many (Sarnecka & Carey, 2008), give-a-number (Le Corre & Carrey, 2007; Sarnecka & Carey, 2008), non-verbal ordinal (Le Corre & Carrey, 2007) and estimating-magnitude-numerically (adapted from Le Corre & Carrey, 2007). Pragmatic competence was measured using two ternary judgment tasks to assess pragmatic ability in two conditions: an enriched-context condition (Katsos & Bishop, 2011) v. a no-context condition (Noveck, 2001). Cognitive ability was assessed using two cognitive tasks: an inhibitory control task (the Simon task (Simon, 1969)) and a visuospatial STM task (the Corsi blocks task (Corsi, 1973)).

Chapter 4 presents the results of descriptive and inferential analyses of the background measures, which revealed no significant differences between the child groups (30 Arabic-bidialectal, 26 English-monolingual, and 30 Arabic-and-English-bilingual pre-schoolers, mean age 5;6) in terms of age, SES, or NVIQ, though the bilingual children had significantly smaller vocabularies than the children in the other two groups. The language exposure questionnaire revealed that the bilinguals had received more input in Arabic than in English, while the Arabic children had received only very limited exposure (less than 20%) to Standard Arabic, which might suggest that they were functionally monolinguals. Regarding the semantic tasks, in general, all the children performed better in the production task than in the perception task. Performance on the perception task showed that all the children had good comprehension of the logical operators, but that on the quantifiers, the Arabic children had very poor comprehension, significantly worse than both the other child groups on both quantifiers. The English children, who had the highest score on quantifiers, performed significantly better on ‘most’ than the bilinguals, but there was no significant difference on ‘some’. On the pragmatic tasks, all the children performed better in the enriched context condition than in the no-context condition, and there was a clear bilingual pragmatic advantage. That is, the bilingual children penalised under-informative items in the two pragmatic tasks in both Arabic and English at a significantly higher rate than the other two groups regarding at least two of the terms (namely, quantifier ‘some’ and operator ‘or’), and numerically they had the best performance. In contrast to this pragmatic advantage, however, there was no cognitive advantage exclusive to the bilinguals—that is, the analyses did not reveal a significant difference in STM ability between the bilingual and English children, but the Arabic children had a significantly shorter STM span than either. The groups also

did not significantly differ in the inhibitory task; however, regression analysis revealed strong effects of these cognitive abilities (inhibition and STM) on pragmatic performance.

Chapter 5 discusses the results presented in chapter 4. The discussion suggests that the finding across groups, of better performance on the production task than the perception task, indicates that acquiring the meaning of quantifiers starts with acquiring their positions on the ordinal scale, before coming to know their semantic meaning. Several potential explanations are mentioned for Arabic children's poor comprehension on the perception task: limited exposure to quantifiers, lack of mathematical prerequisites (possibly due to delayed exposure to the numerical system), or limited STM. The bilingual children's superior pragmatic performance was speculatively attributed to better episodic *buffer ability* (Baddeley & Hitch, 1974; Baddeley, 2000), allowing them to more effectively connect information held in STM with information stored in long-term memory (LTM). The strong effect of context found was expected due to the demands of the no-context condition, which required children to draw heavily on LTM to check the validity of a statement by searching through their acquired world knowledge. The strong effects of inhibition and STM on pragmatic performance seem to provide some evidence that the process of deriving implicatures is cognitively effortful, a finding that favours the Relevance theory of implicature processing (Sperber & Wilson, 1986/1995) over the Default theory (Levinson, 2000).

Chapter 6 concludes the thesis by reviewing the previous chapters and discussing how the results contribute to the field. It also highlights some of the limitations of the current research that might be avoided in future work. Finally, directions for future work are briefly discussed.

Chapter 2

Literature Review

2.1 Introduction

The purpose of the review of literature presented in this chapter is twofold. First, it introduces the reader to the theoretical roots of this research and the ideas that it has been built on. Second, it reviews the empirical tools, findings, and conclusions of previous work not only to help us understand the importance and the originality of the current research but also to pave the way for a more meaningful interpretation of the outcomes of this study.

The present research can be described as a multidimensional project, since it was developed by connecting theoretical concepts and empirical findings from three domains of research: (1) the semantic comprehension of quantifiers and its relation to the acquisition of approximate and exact numerical systems, (2) pragmatic competence and (3) bilingualism and cognitive development. The first section of this chapter covers the domain of the acquisition of quantifiers and numbers. In this section, I introduce the reader to the question of the nature of the relation between quantifiers and numbers. Investigating this relation was essential, since children's pragmatic ability to derive scalar implicature (e.g. when quantifiers are used in a context, as in *some of the students passed the exam* and the hearer infers that not all the students passed the exam) depends at least partially on their semantic comprehension of quantifiers (such as *some* and *most*). Thus, I first review empirical findings on children's acquisition of numbers and quantifiers, then attempt to sketch how these two systems might affect each other, on the basis of several assumptions resting on theories explaining mental representations of number and abstract words; generally, this work suggests that acquisition of number precedes acquisition of quantifiers. Next, I briefly review empirical findings on factors that might affect children's acquisition of numbers and/or quantifiers (e.g. intensity of exposure and bilingualism).

After this, I introduce the second domain, pragmatic competence. Theories explaining different concepts related to implicature and the assumptions around its cognitive computation mechanisms are covered, followed by empirical evidence for these assumptions from behavioural and neural data. Then, since children are the target population in this research, I review prior work on children's acquisition of scalar implicature, the main focus of this research.

In section 3, I introduce the reader to the third and last background research domain which is the impact of bilingualism on cognitive and pragmatic abilities. In this section, we try to understand the nature of this impact, why it occurs, and how bilingualism—more precisely the changes in cognitive abilities that result from bilingualism—can affect pragmatic competence; I explain these points with reference to the theories of implicature processing that I introduce in section 2.3. Then, I briefly review other factors (e.g. socio-economic status (SES) and language proficiency) that might modify this impact of bilingualism and thus indirectly modify its (indirect) influence on children's pragmatic ability.

Finally, I will re-present and reconsider the main questions that the current research is investigating, with a brief discussion of the research's originality and contribution to the field in light of all the reviewed work. The last section of the chapter summarises the chapter's main findings.

2.2 Numeracy and quantifier comprehension

Although there is a broad literature on children's ability to generate implicatures from scalar quantifiers such as *some* used in a context (e.g. Feeney Scrafton, Duckworth, & Handley, 2004; Noveck, 2001; Papafragou & Musolino, 2003; Gronder, Klein, Carbary, & Tanenhaus, 2010, Katsos & Bishop, 2011), only a few studies have explored how children comprehend these words by producing sets that represent such quantifiers (Barner et al., 2009; Hanlon, 1987, 1988). Also, although there is some empirical evidence of the correlation between children's comprehension of quantifiers and number words (Barner et al., 2009), and although the theoretical assumption has been made of the association between the two systems (Piantadosi et al., 2012), there

is still no clear explanation of the nature and direction of this relationship—which comes first, numbers or quantifiers, and in either case why and how? Furthermore, regardless of which comes first, it seems we still know very little about the proportional nature of these vague quantifiers, which is highly sensitive to context—more precisely to set size.

In addition, some recent empirical findings suggest that even after the complete acquisition of proportional quantifiers, adults' knowledge of them can be easily modified by external factors, a fact that might indicate that the mental representation of such terms is unstable even at the developmental endpoint. For instance, a recent study conducted by Yildirim et al., (2016) attempted to explore adult listeners' beliefs regarding the two quantifiers *some*, *many* and the adoption of talker-specific interpretations of them. For example, using a forced-choice method, participants viewed a total of 25 candies (green and blue) presented with different proportions, and were asked to rate how likely the talker would be to use one of the given description (e.g. *some of the candies are blue*, *many of the candies are blue*, other). After this pre-exposure test, they watched a video depicting talkers describing various visual scenes with ambiguous proportions (e.g. 13/25), with utterances such as *some of the candies are blue* for the some-biased group and *many of the candies are blue* for the many-biased group. Then, both groups took a post-exposure test, which was exactly the same as the pre-exposure test except that it excluded 'other' from the possible choices. Comparing the proportional interpretations of *some*, *many* in pre- and post-exposure tests, they revealed a change in participants' beliefs as a result of their adoption of the talker's interpretations. That is, based on brief exposure, each group updated their expectations on how the talker would describe the proportion (e.g. 13/15). These results were confirmed in Heim et al.'s (2016) study exploring adults' comprehension of *few* and *many* and the potential effect of training on their beliefs regarding proportional quantifier meanings. If exposure to only a limited number of trials or a limited amount of feedback leads to change in previously acquired knowledge in adults, then the question might arise how children, who have not yet established a robust understanding of the sometimes vague or obscure meanings of quantifiers and who have limited cognitive resources compared to adults, especially in terms of numeric knowledge, will proportionally interpret the meanings of quantifiers.

All these issues led me to further explore the relationship between numbers and quantifiers, which seems to be more complicated than previous explanations of it in terms of plural and singular morphology (Carey, 2004). In this section, I first review empirical findings on children's acquisition of number and quantifier words; then I attempt to define the relationship between the two, with reference to theories on the mental representation of number words and abstract concepts. After outlining this relationship, I briefly review the neural regions activated while processing different types of quantifiers and then describe some factors that could modify the acquisition process of both numbers and quantifiers.

2.2.1 Empirical findings on children's acquisition of numbers and quantifiers

Despite the huge literature on children's acquisition of number words (e.g. Becker, 1989; Frye, Braisby, Lowe, Maroudas, & Nicholls, 1989; Wynn; 1990; Le Corre, Van de Walle, Brannon, & Carey, 2006; Le Corre & Carey, 2007; Sarnecka & Carey, 2008; Odic, Le Corre, & Halberda, 2015; among others) only a few studies have investigated how and when children comprehend quantifiers (e.g. Hanlon, 1987, 1988; Barner et al., 2009). This section summarises the main findings on children's acquisition of numbers and quantifiers, in order to help us understand the basic development trajectory and hopefully to pave the way to precisely explaining the relationship between the number and quantifier systems.

2.2.1.1 Acquisition of numbers

Studies that investigate children's acquisition of an exact (cardinal) numerical system have widely adapted four paradigms. The first paradigm is designed to explore children's ability to map each item in a set to its corresponding value in an abstract numerical list. The task used to test this ability is usually called the *count list* (or *sequence*) *task* (e.g. Le Corre & Carey, 2007; Sarnecka & Carey, 2008). In this task children are simply asked to count the items in a list for the experimenter. The second paradigm examines children's ability to accurately map any given set to a corresponding number word that represents the total number of set items. The task usually used to assess this ability is the *how-many task* (e.g. Becker, 1989; Le Corre et al., 2006; Sarnecka & Carey, 2008). In this task children are asked either to listen to

the experimenter counting the set items or to count the set themselves, and then to answer the experimenter's question *how many Xs are there?* The third paradigm investigates children's ability to construct a set that represents a specific number in question, using the *give-a-number task* (e.g. Le Corre et al., 2006; Le Corre & Carey, 2007; Sarnecka & Carey, 2008). In such a task, children are usually presented to different sets (e.g. 8 apples, 8 spoons, 8 carrots), and then asked to respond to a prompt such as *give the puppet 4 carrots*. The fourth paradigm tests children's acquisition of the successor function (the ability to move ahead one value in a numerical list when adding one item to a set and to move one value backward in the numerical list when removing one item from the set) and the task used to assess this principle is the *direction task*. In this task, each trial presents a child to two sets of the same size (e.g. each has 5 strawberries) given in two separate containers (e.g. plate A and plate B). The experimenter moves one strawberry from A to B, and then asks the child to guess which plate contains 4 and which contains 6 strawberries without counting (Sarnecka & Carey, 2008). Some studies have also proposed a new paradigm aiming to test children's approximate numerical system, that is, to assess their ability to map different magnitudes (without counting) approximately to values in a numerical list; the *fast card task* is the common test for this aspect (e.g. Le Corre & Carey, 2007; Odic et al., 2015). In this task, children view sets that vary in size, and in each trial they are asked to tell the number of the objects in a given set (e.g. 8 circles) as fast as they can without counting.

Starting with the acquisition of an exact numerical system, there is empirical evidence that although 2-to-3-year-old children can successfully produce small sets only up to four, their numerical ability in relation to exact systems develops dramatically when they reach the age of 4 (Wynn, 1990; Sarnecka & Carey, 2008). For instance, in a study exploring 2-to-4-year-old children's acquisition of an exact numerical system, the results revealed that the children were able to produce sets that accurately represented a given number only after first coming to understand how counting works (Sarnecka & Carey, 2008). The counting process is based on three main principles; the *one-to-one principle* (when counting, only one numeral must be given to each item in the set), the *stable-order principle* (numerals must be used in the same order in any context), and the *cardinal principle* (the last item counted in a set represents the number of items in that set) (Gelman & Gallistel, 1978). In Sarnecka and Carey's

(2008) study, only children who had mastered the cardinal principle succeeded in an arithmetic task measuring a successor function. That was expected, since the task requires a high level of numerical comprehension, and Sarnecka and Carey suggested that children who could complete the successor function task (like the direction task mentioned above) successfully might have some implicit knowledge of how counting entails a successor function. These results replicated the findings of Le Corre et al. (2006) for 2-to-4-year-old children's performance on counting tasks, which showed that variation in task demands impacted the children's performance.

Studies exploring the simultaneous acquisition of exact and approximate numerical system (also referred to as quantity) vary in their results. For instance, Le Corre and Carey (2007) explored the acquisition of exact and approximate numerical systems by 3-, 4-, and 5-year-old children and found that even children who were able to accurately count sets consisting of 10 items and successfully generate sets consisting of up to 6 items were not able to map magnitudes beyond 4 to their numerical values and failed to show scalar variability when estimating magnitudes. These results led Le Corre and Carey to suggest that mapping between large numerals and analogue magnitudes might not part of the numerical acquisition process and that children might develop this ability at about the age of four-and-a-half years old but not earlier. Although this claim is in line with Huntley-Fenner's (2001) findings among 5-to-7-year-olds, who showed adult-like ability to estimate various magnitudes, there is conflicting evidence on the development trajectory of younger children and it is still unclear. For instance, Negen and Sarnecka (2010) replicated Le Corre and Carey's fast card task and found that young children, even those who had not yet mastered the counting principle, showed scalar variability when estimating different magnitudes, while in contrast, Odic et al. (2015) claimed that their results confirm the finding of Le Corre and Carey (2007) on the absence of scalar variability in young children.

However, the results of the above-discussed studies on children's acquisition of approximate numerical system should be taken with caution, for several reasons. First, the idea of scalar viability is intrinsically approximate and not exact in nature, and the absence of scalar variability in Le Corre and Carey's study could have resulted from their use of very similar magnitudes {1, 2, 3, 4, 5, 7, 8} for sizes that exceed {4}. Furthermore, the children in that study did give larger numerical values

when evaluating sets above {4}, possibly indicating the availability of scalar variability to those children. If the assumption of the closeness of the magnitudes is valid, then, should one expect to find evidence for scalar variability with magnitudes that are more distinct? Odic et al. (2015) used slightly different magnitudes {1, 2, 3, 4, 6, 10} than those used by Le Corre and Carey, and the children who acquired the counting principle clearly showed scalar variability; however, Odic and colleagues, like Le Corre and Carey, did not consider this finding as evidence for scalar variability. Second, although Negen and Sarnecka (2010) used exactly the same magnitudes of Le Corre and Carey (2007), their results were based on a longitudinal study where data were collected by testing the same children week to week; thus, the results might be attributed to a training effect or a natural cognitive development effect rather than being taken as evidence for the availability of scalar variability.

To sum up, although the findings on children's acquisition of exact numerical systems have been replicated in various studies and are quite solid (e.g. Becker, 1989; Wynn, 1990; Le Corre et al., 2006; Le Corre & Carey, 2007; Sarnecka & Carey, 2008), acquisition of approximate numerical systems seems to need further exploration for a better understanding of its developmental process.

2.2.1.2 Acquisition of quantifiers

Some previous studies have examined how pre-schoolers comprehend the semantics of numerical and quantificational scales, but unfortunately, the proportional meanings of logical quantifiers have not received much attention in previous research (Yildirim et al., 2016). For example, using an explorative paradigm (*the give-a-quantifier task*), Hanlon (1987, 1988) and Barner et al. (2009) investigated how children understood various scalar quantifiers by asking them to act upon given statements, for instance *give the puppet some of the balls*. Hanlon's (1987) study involved children aged between 4 and 7 years old; she tested their semantic comprehension of different quantifiers (e.g. *all, some, none, any, both, either, neither*) and found that, generally, they were more competent with quantifiers whose meanings were cognitively less complicated. For example, the children exhibited a ceiling effect (100%) on *all* and *both* but were less competent with *some* (88%) and *either* (46%). A compatible finding was gained in Barner et al.'s (2009) study with younger children (2, 3, and 5

years old)—although their participants showed an appropriate ceiling effect on *all* as in Hanlon’s study, only around half of Barner et al.’s participants could produce correct sets consistently for *some*, and only around 10% for *most*. Such results might be attributed to the younger age of Barner et al.’s sample compared to Hanlon’s. Another interesting finding in Barner’s study was the correlation between children’s performance on the give-a-number task and a give-a-quantifier task; the implication of this finding is discussed further in the next section.

2.2.1.3 Which is acquired first, quantifiers or numbers, and why?

A few studies examine developments in children’s comprehension of quantifiers in tandem with their knowledge of numerals. For example, Barner et al. (2009), discussed above, used the give-a-number task to explore the relation between children’s numerical and quantificational abilities and found a significant age-independent correlation between the two domains. They considered this result to point to the possibility that quantifier acquisition may support the development of numeral acquisition. However, if one recalls the correlation between the children’s performance in the give-a-number task and that in the give-a-quantifier task, mentioned above, these results might in fact indicate that Barner’s findings are compatible with the reverse interpretation; that is, the children’s advanced performance on the numeral task would indicate that they were more competent with this system, and in the earlier stages of acquiring the quantificational system compared to the number system.

Is it possible that the correlation found by Barner et al. (2009) between children’s numeral and quantifier acquisition is accompanied by a correlation between counting ability and pragmatic comprehension of numerals and quantifiers? Hurewitz Papafragou, Gleitman, & Gelman (2006) explored (3-to-4-year-old) pre-schoolers’ pragmatic interpretations of numbers and quantifiers within contexts, and did not find any correlation between a numerical how-many task and the pragmatic task. Although the how-many task results revealed that less than half of the participants were able to correctly count a set of five objects, most of those children were able to provide correct pragmatic interpretations for sets of *two* and *four* objects but failed when the same sets were described using the quantifiers *all*, *some*. Hurewitz and colleagues

argued that this finding showed that children might employ different mechanisms when acquiring numbers and quantifiers, but they did not provide a deeper level of interpretation or suggest what these mechanisms might be.

Piantadosi et al. (2012, p. 201) indicate that ‘it is difficult to see from what basis they [quantifiers] could be learned, or why—if they are learned—they should not be learned relatively early. We therefore assume that they are available for learners by the time they start acquiring number word meanings’. However, it seems that the existing empirical findings are not able to give a clear understanding of the role the numerical system might play in the acquisition of quantifiers, and more importantly, that previous work has not provided a clear or sufficient explanation of why and how numbers can be related to quantifiers. In the next section, I will attempt to theoretically clarify the potential relationship between the numerical and quantificational systems.

2.2.2 The relationship between numbers and quantifiers

To explain the relationship between numbers and quantifiers, I will rely on theories explaining how humans come to shape mental representations of numerals and numeral systems and also of abstract terms. Before going on to explain how such theories might help bridge the two systems, let me justify briefly the two basic motivations for mapping representations of abstract and number words with quantifiers. First, the linguistic nature of quantifiers is similar to that of abstract words (Yildirim et al., 2016), and therefore number words (which have a kind of concrete nature—for example, they have exact interpretations: *three* always refers to three objects which one can visualise in the real (physical) world) can be considered to constitute lexical entries not only defining these quantifiers but also shaping mental representations for the abstract quantificational terms. This concretising support role might be especially critical at the early stages of quantifier meaning acquisition, and since the nature of a numerical system and the level of exposure to it could differ across languages and perhaps also across learners, it might be expected that these factors would be reflected in (any effect on) children’s comprehension of quantifiers (Barner et al., 2009). Second, the proportional nature of quantifiers requires the application of advanced cognitive resources, beyond just the ability to count set items

accurately. That is, the comprehension of quantifiers requires employing both approximate and exact numerical systems, for which WM as well as mathematical operations are required to evaluate/estimate various magnitudes (Heim et al., 2016). WM and basic mathematical operations (related to the successor function) are essential for the acquisition of an exact numerical system (Le Corre & Carey, 2007; Sarnecka & Carey, 2008); in the same way, scalar variability seems to play an essential role in the acquisition of an approximate numerical system (Le Corre & Carey, 2007), and perhaps thus also for the acquisition of quantifiers, at least in the early stages of shaping the mental representation. That is, it might be assumed that someone who shows scalar variability when estimating various magnitudes numerically will be able to proportionally estimate various amounts/magnitudes using quantificational scales, while someone who does not show this scalar variability when using numbers will not be able to estimate magnitudes using scalar quantifiers either.

2.2.2.1 Quantifiers as abstract concepts

To show the nature of the similarity between quantifiers and abstract terms, a number of hypotheses on the representation of concrete and abstract words will be reviewed in this section. Several researchers attempt to clarify how representational systems for abstract terms differ from those for concrete ones. According to Fodor (1998), the representation of concrete and abstract words is abstract and symbolic (i.e. independent from sensory experiences). In contrast to this view, the *embodied account* suggests that both concrete and abstract concepts are grounded in perception and sensorimotor experiences (Barsalou, 1999). Other researchers hypothesise that multiple representational systems are activated during conceptual processing of both concrete and abstract terms (Louwerse & Jeuniaux, 2010). This view is consistent with the *dual-coding theory* proposed by Paivio et al. (1986), according to which two different codes of representation, respectively linguistic and sensorimotor, are activated when processing abstract and concrete terms; concrete words require the activation of both codes, while abstract words would involve only the activation of linguistic information. The *language and situated simulation* theory suggests that during a word processing (whether it is concrete or abstract) both the linguistic and the sensorimotor systems are activated (Barsalou, Santos, Simmons, & Wilson, 2008).

Having briefly highlighted theories addressing the representational systems of concrete and abstract words, it can be now more easily explained precisely why and how quantifiers can be considered abstract terms. Although this idea is not novel and has been employed in Yildirim et al.'s recent paper (2016), in their discussion of the absence of sensorimotor representations of these terms, they unfortunately do not make it clear how humans might then shape mental representations of such terms. Thus, relying on some of the above-discussed theories and also on some empirical findings, I will try to explain further why quantifiers seem to have an abstract nature, and how humans may perhaps form mental representations of such terms.

Taking into consideration the absence of (exact) sensorimotor objects that can represent quantifiers in the real world, the dual-coding theory suggests that when acquiring meanings of abstract concepts it is likely that linguistic information is activated (Paivio et al., 1986), while the language and situated simulation theory suggests abstract concepts require activation of both linguistic and sensory-action information (Barsalou et al., 2008). Since it is not my intention to discuss the validity of these theories in explaining how we shape mental representation of concrete objects and abstract nouns, then, regardless of whether abstract terms (and in this context quantifiers) only activate linguistic information or both sensory-action and linguistic information, it still should be asked what type of information might be involved in each kind of representation (either linguistic or sensory-action). Regarding linguistic information, if we take the quantifier *some* as an example, then its literal meaning might be difficult to be defined without explicitly referring to numerals, (e.g. 'more than one and less than the whole set'). However, explicit use of numerical words might not be applied to other quantifiers such as *most*, which can be defined as (more than half), though still its definition requires applying advanced logical/mathematical operations which might be implicitly related to numerical system (Heim et al., 2016). Although this might seem a trivial assumption that lacks empirical support, it attempts to understand which kind of linguistic information seem to be essential for formalising representations for these quantifiers by hypothesising that numerical systems in general might play a role as sets of linguistic entries for acquiring representations of quantifiers.

Regarding the sensorimotor information that might be involved in shaping

representations of quantifiers, it can be hypothesised that mental representations of numbers (which have a concrete nature, as we can see in the real world what, for example, *two* means), and which are activated by sensorimotor action as well as linguistic information, can contribute to the establishment of mental representations of abstract quantifiers. In support of this claim, Casasanto (2010, p. 453) indicates that ‘[p]erhaps sensory and motor representations that result from physical interactions with the world are recycled to support abstract thought’. He gave an example of how we employ our representation of physical space to describe the abstract concepts of time (e.g. a *long* vacation; a *short* meeting) (Casasanto, 2010, p. 453). If we apply this idea to the numerical system and quantifiers, we might be unable to define the latter without relying partly or completely on the former (since *some* means more than one, while *most* refers to more than half). To explain how numerals might play a role in the acquisition of quantifiers, assume that on one occasion, a child hears his mother refer to three objects using the numeral 3, and on a different occasion hears her refer to the same quantity using *some*; the child might then *relate* the concrete quantity (3 Xs) to the abstract quantifier *some*. At a later stage, the child will generalise such knowledge and position *some* in relation to other quantities after hearing adults around her using ‘some’ to refer to unfixed quantities.

Another possibility regarding the role number words play in shaping mental representations of quantifiers appears in association with analogical reasoning. Building on Kaminski, Sloutsky and Heckler’s (2006) idea of the role of analogical reasoning in the transfer of conceptual knowledge to novel isomorphic situations, it can be suggested that children might transfer their conceptual knowledge of number words and apply it to quantifiers; and as Kaminski et al. state in their study, mapping structure is an essential part of the learning process, via analogical reasoning. This should lead to ask in which way the structures (or learning/acquisition mechanisms) of numbers and quantifiers might be similar? Is it this similarity that might lead to the acquisition of quantifier meanings, or is it the employing of the cognitive primitives developed in the process of formalising a numerical system?

2.2.2.2 Mental representation of the numerical system and its possible role in the acquisition of quantifiers

To answer the above questions on possible common mechanisms underlying the acquisition of number and quantifier words, I will first explain briefly three main theories on the mechanisms of acquisition of number representations; then, I will discuss some possible similarities and distinctions between the quantifiers and number learning mechanisms. The first theory to be addressed is *analogue magnitude theory* (Brannon et al., 2001; Dehaene, 2003). This view suggests that learners encode analogue magnitudes with cardinal values; it is characterised by two psychological aspects, namely Weber's law and scalar variability. Weber's law indicates that distinction of two quantities is a function of their proportion. For instance, 5 and 10 are easier to distinguish from one another than 45 and 50 are from one another. Scalar variability means that the standard deviation of the estimate of some quantity is a linear function of its absolute value. For example, when prevented from counting, adults estimate numerical size relying on existing cognitive mapping between numbers and analogue magnitudes. Under these circumstances, both the average and the variability of the estimates should increase at the same rate, as the sets grow larger (Cordes & Gelman, 2005).

The second theory dealing with the system of number representations is the *enriched parallel individuation* view (Le Corre & Carey, 2007). In this view, the representations of numbers rely on children's capacity to create a working memory (WM) model in which each individual in a set is represented by a unique mental symbol. For instance, for a set of 3 dogs, children would have a mental model of {X, X, X}. In other words, this view assumes that children individuate objects, manipulate sets in their WM, and compare sets using one-to-one correspondence with their mental model existed in their long-term memory (LTM).

The last theory on the representation of numbers to be covered here is the *language of thought* theory (Piantadosi et al., 2012). This is an extension of the enriched parallel theory based on the assumption that meanings are formalised using a 'language of thought', as proposed by Fodor (1975), which defines a set of primitive cognitive operations and composition principles. According to this view, humans shape representations of numerical meanings using the *lambda calculus*, described as 'a

formalism which allows complex functions to be defined as compositions of simpler primitive functions' (Piantadosi et al., 2012, p. 201). Piantadosi et al. explain that these primitive elements are the basic cognitive mechanisms that learners must comprehend how to combine in order to arrive the correct system of numerical meanings. The primitives include various operations, such as mapping sets to truth values, manipulating sets, performing logical functions on sets (e.g. P and Q, P or Q, Not P, If P x Y, functions performed on the counting routine (next, previous, equal-word), and finally *recursion*, an operation that allows learners to return the result of evaluation of a given numerical/mathematical (λ) expression on set N. These primitives are not necessarily innate; for the theory to work they must merely be accessible to children by the time they start leaning numbers (Piantadosi et al., 2012).

Keeping the review of theories on the representation of numbers in mind, we now move on to discuss possible similarities and differences between numbers and quantifiers. Regarding similarities, it is obvious that both numbers and quantifiers exhibit scalar variability, meaning that if it is assumed that the capacity to estimate various magnitudes using approximate numerical scales is part of formalising numerical representations, then children will transfer such knowledge when acquiring the quantifiers. In addition, the acquisition of counting ability is associated with the development of certain cognitive abilities that allow children to understand the cardinal principle and the successor function, in addition to WM (Le Corre & Carrey, 2007; Sarnecka & Carrey, 2008). These abilities are likely essential for the comprehension of quantifiers (McMillan, Clark, Moore, Devita, & Grossman, 2005; Heim et al., 2016). As for differences, it might seem intuitive that number words are characterised by the availability of exact (fixed) interpretations, while quantifiers are not; but this distinction might lead to concern regarding whether children apply the same primitives in this context that they developed when acquiring numbers, and if they do, when and how they figure out that quantifiers, unlike numbers, can be mapped to various set sizes depending on the context. It is clear that more research needs to be conducted to figure out how children acquire the proportional meanings of quantifiers.

To sum up, it has been indicated that number learning is a complex learning process, influenced by many factors, including pedagogical and social cues (Piantadosi et al.,

2012). This might lead us to ask whether early exposure to the numerical system might facilitate the acquisition of the representations of numerical meanings and the primitives that are claimed to be the basic cognitive components for learning numbers.

2.2.2.3 Neural basis of quantifiers and numbers

So far, I have made the claim that quantifiers are abstract concepts whose mental representations are possibly based on our numerical concepts, or more generally on quantity knowledge. In this section, my purpose is to review some of the findings of studies that have explored the neural basis of processing of numbers and different types of quantifiers to find out to what extent brain areas associated with number processing are similar (or distinct) to those involved in quantifier processing. All the studies reported in this section employed functional magnetic resonance imaging (fMRI) or event-related potential (ERP) techniques.

We begin with the neural basis of numbers. Studies that explore the neural regions activated while completing tasks requiring the application of arithmetical operations to small or large numbers show that different neural regions are associated with exact and approximate (quantity) numerical systems. For instance, Dehaene, Spelke, Pinel, Stanescu and Tsivkin (1999) used two functional brain-imaging techniques (ERP and fMRI) to explore neural activation in adults completing a mathematical task. In this task, the participants were presented with an addition problem (e.g. $4 + 5 = \dots$) and given two possible answers. In the exact condition, one of the answers was exactly right (e.g. '9' or '7'), while in the approximate condition the participants had to select the most appropriate answer from two that were not exactly right ('8' or '3'). The results revealed that the bilateral intraparietal lobes (associated with visuo-spatial and analogical mental processing) were more active in the approximate condition than in the exact condition, whereas in the exact condition the left inferior frontal areas (associated with language production and comprehension) were highly activated. Another interesting finding was that exact stimuli activated language-based regions in the brain (in the left hemisphere) while the areas activated in the approximate condition were language-independent (were in the right hemisphere). These results have been confirmed in studies including participants with damage to left or right hemispheres (for a review see Butterworth & Walsh, 2011). For example, Lemer

Dehaene, Spelke and Cohen (2003) investigated approximate and exact numerical skills using tasks that required patients with damage either in the left hemisphere (aphasia) or the right hemisphere (Gerstmann's syndrome) to solve subtraction and multiplication problems. The results revealed that participants with severe left-hemisphere area damage were less competent on the exact numerical assessment, while participants with right-hemisphere damage showed more impairment on the approximate calculation than the exact task.

With respect to the neural basis of activation while processing different quantifiers, McMillan et al. (2005) compared differences in neural activation among healthy adult participants processing first-order quantifiers *at least*, *all*, and *some* and higher-order quantifiers *most*, *more*, *even*, and *odd*. In a truth value judgement task, participants viewed different proportions of items (e.g. 4 balls: 3 blue and 1 yellow), and were asked to decide if a statement appearing on-screen accurately described the scene (e.g. *at least three of the balls are blue*). Their results showed that although all quantifiers activate the inferior parietal cortex, associated with numeracy, only higher-order quantifiers activate the prefrontal cortex, associated with executive resources like WM. The authors attributed this finding to the higher cognitive cost required to process higher-order quantifiers in two steps. That is, as McMillan and colleagues clarified, while processing a statement like *at least half of the stars are red*, the hearer has first to assess what 'half' of the given number of items is, and this quantity must be held in mind so that a comparison between the actual number of items and this value can be made.

In another study, Troiani, Peelle, Clark and Grossman (2009), used a similar paradigm to McMillan et al. (2005) but with a binary response ('Yes'/'No') to explore neural differences also between numerical and (Aristotelian) logical quantifiers (e.g. *at least* v. *some*). The outcomes revealed that the quantifier comprehension process activates two dissociable neural networks: the numerical quantifiers were processed in the lateral parietal-dorsolateral prefrontal network (associated with quantity-based or numerical processing), whereas logical quantifiers were processed in the rostral medial prefrontal-posterior cingulate network (associated with elementary logic), supported by the posterior cingulate cortex (associated with the WM network)

(Troiani et al., 2009). Troiani et al.'s (2009) results highlight the significant involvement of abstract number knowledge in the meaning of numerical quantifiers in semantic memory and the possible contribution of logic-based evaluation in the service of logical quantifiers. Although Troiani et al.'s outcomes did not associate logical quantifier processing with number knowledge, they do give evidence for the role of WM in the comprehension process of such quantifiers.

With respect to the proportional nature of quantifiers and whether it is stable in adults, Heim and colleagues (2016) investigated the neural basis of flexible adaptations in the semantics of quantifiers such as *few* v. *many*. Heim et al. (2016) also used a binary-judgement task in which adult participants viewed pictures each including a total of 50 blue and yellow circles presented in different proportions in each stimulus (i.e. 20%/30%/40%/50%/60%/70%). Each picture was presented alongside a written sentence, either *Many of the circles are yellow* or *Few of the circles are yellow*, and the participants were asked to evaluate it by pressing the 'Yes' or 'No' key. In the first block, the participants were only asked to evaluate the stimulus; in the second block, they were given feedback only on their evaluation of *many*; and the third block aimed to test the effect of this training through feedback on both their neural and behavioural performance. The feedback given in the second block aimed to change participants' beliefs about *many* by indirectly training them to apply new criteria to it. That is, if the participants disagreed that a statement like *many of the circles are blue* accurately describes a 40% proportion of blue circles, they were given negative feedback. The behavioural data revealed a change in beliefs after this indirect training, consistent with Yildirim et al.'s (2016) results; and the neural data revealed two interesting findings: First, processing of quantifiers seems to involve activation of quantity knowledge (in the parietal lobe) as well as activation of areas relevant to decision-making (in dorsolateral prefrontal regions). Second, there was increased neural activation when the participants applied the new criterion (40%) to *many*, compared to other high proportions. It is important to refer here that such activation did not generate from ambiguity, that is, proportions of 50% or higher were unambiguously *many*, whereas 40% may be ambiguous, but rather from applying new criteria for the meaning of *many* to include 40%, and this might contradict previously acquired knowledge.

Three important conclusions might be derived from the above-mentioned studies. First, different quantifiers require different cognitive costs for processing depending on their cognitive complexity, which might explain the variation in children's performance on different quantifiers found in Hanlon (1987) and Barner et al. (2009). Second, processing of quantifiers activates the neural regions responsible for numerical representations, especially areas involved in the approximate numerical system, which might be taken as evidence for a strong relation between the two systems; however, we should be tentative with this claim, since Troiani et al.'s (2009) results only associated numerical quantifiers with a numerical neural network. Last, cognitive resources, especially WM, play a significant role in processing quantifier terms. Since the current research explores children's acquisition of the logical quantifiers 'some' and 'most' as well as their STM, then, it would be interesting to find out if there is any impact of STM on the acquisition process.

2.2.3 Factors that might affect children's acquisition of numbers and quantifiers

Regardless of whether numerical knowledge represents the lexical entries for the acquisition of abstract quantifiers, children's acquisition of numerical and quantificational systems seems to be influenced by several complicated pedagogical and cultural factors as well as by level of experience/exposure (Piantadosi et al., 2012; Barner et al., 2009; Hanlon, 1987).

With respect to factors that might influence the acquisition of numerals, Anders and colleagues (2012), in a longitudinal study, explored the effects of family background, home learning environment, and pre-school learning on children's numeracy skills. To assess all these factors, they employed various standardised measures: Family background was determined based on parental language status, educational level, and occupations (as a measure of SES), while home learning environment was composed of three measures: questionnaire, interviews, and the outcomes of a reading task completed jointly by a primary caregiver and the child participant in the experimenter's presence. Pre-school learning quality was assessed on the basis of two criteria: global quality (class size, number of children in the class, health, safety, schedule, indoor and outdoor play, spaces, teacher qualifications, play materials, administration, and whether staff needs were met) and educational quality (learning

activities for verbal knowledge, mathematics, and science knowledge; accommodation for diversity and individual learning needs). Children's numeracy knowledge was measured using a standardised test to evaluate their skills in counting, recognising numbers, knowledge of shapes, and comprehension of some basic mathematical concepts (e.g. addition, subtraction). The results revealed that children's numeracy was strongly predicted by family background (especially SES and mother's educational level), and there was a strong correlation between children's numeracy and home learning environment. While pre-school quality was not directly correlated with these skills, it had an effect detected in the long term. That is, the advantage of high-quality pre-schooling was present and maintained at later ages.

To explore the possible effect of additional languages on the process of numerical acquisition, another recent study examined whether being exposed to two languages might affect children's numeracy skills, finding that the learning process was independent in each language (Wagner, Kimura, Cheung, & Barner, 2015). Wagner and colleagues explored whether children's numeracy in their first acquired language (L1) could predict numeracy in their second language (L2). They tested 2-to-5-year-old English–French and English–Spanish bilingual children's performance on three numerical tasks (the give-a-number task, *highest count* task (in which a child is asked to count as high as she can), and direction task), and found that children's results in their L1 and L2 were independent from one another. Their results revealed that the children's ability to generate sets when tested in their L2 was predicted only by either age or the highest number they could reach in their L2 but not in their L1. The results also revealed that the children's ability to produce sets accurately (in the give-a-number task) in the L2 was predicted by their performance on the same task in the L1. Wagner et al. (2015) suggest that such results might indicate that the delay in children's ability to produce accurate sets in the L2 compared to the L1 can be attributed to their extensive exposure to number words in the L1, making the delay a result of difficulties identifying which concepts correspond to which words rather than of conceptual problems with how counting works. Although the study sheds some light on whether children can transfer their numeracy skills in one language to another, given the absence of a monolingual group it remains unclear whether bilingualism hinders or facilitates the acquisition of a numerical system, and further research should be conducted to explore this domain.

With respect to the factors that might impact the acquisition of quantifiers, with the limited amount of research exploring the semantic comprehension of quantifiers out of context, we might have not much yet to say. Hanlon (1987) proposed a possible impact of intensity of exposure on the quantifier acquisition process; to test this hypothesis, Hanlon explored whether variation in parents' frequency of usage of the quantifiers might correlate with children's performance on the give-a-quantifier task. She relied on the total frequency of quantifier terms in parental speech (taken from a longitudinal study) for three children as a predictor for their performance on different quantifiers. Hanlon's results showed a strong correlation between usage frequencies among parents and the variation in children's performance on different quantifiers in the comprehension task. However, this result cannot be considered reliable and should not be generalised to all children, as Hanlon (1987) stated; she suggested that a more plausible explanation why, for example, the children exhibited a ceiling effect on *all* but not other quantifiers is that the level of cognitive complexity for *all* is lower than that for other quantifiers, rather than its frequency of usage. Barner et al.'s (2009) results give rise to another potential impact on children's comprehension of quantifiers; they found a significant correlation between children's comprehension of number words and of quantifiers. However, we still know very little about this relationship: Does it start with the approximate numerical system, analogical reasoning, cognitive resources such as WM, or even mathematical abilities? Clearly, more research should be conducted to understand the developmental trajectory here.

2.2.4 A summary of the research on development of numbers and quantifiers

This section first reviewed the empirical findings on children's acquisition of number and quantifier terms and then attempts to explain how the two systems might be associated. The reason for including this section is twofold: First, to justify the importance for the current study of exploring children's numeracy skills when examining their comprehension of quantifiers, by explaining the potential relationship between the two on the theoretical level, and second, to help me better understand and interpret my own results. The empirical findings showed that children at age 2 were able to produce sets representing small cardinals, while children even at age 3 and older were not able to generate sets representing natural language quantifiers such as

some and *most*. From the theoretical debate emerged a preponderance of evidence that numbers seem to be acquired earlier than quantifiers due to their status as exact and concrete representations, which facilitates their acquisition process, at least in the early stages of acquisition.

2.3 Pragmatic competence

This section starts by explaining Grice's theories on *implicature* (1975, 1989) and then introduces two theories explaining different possible mechanisms of implicature computation, namely the *Default theory* (Levinson, 2000) and the *Relevance theory* (Sperber & Wilson, 1986/1995). After that, I explain the specific type of implicature explored in this research—*scalar implicature*, and briefly describe the different mechanism of its computation under each of the two processing theories. To understand these processing mechanisms more deeply, I review the empirical findings on implicature processing. After this, I review the empirical findings on children's acquisition of scalar implicature, followed by a brief summary of the main findings of this section.

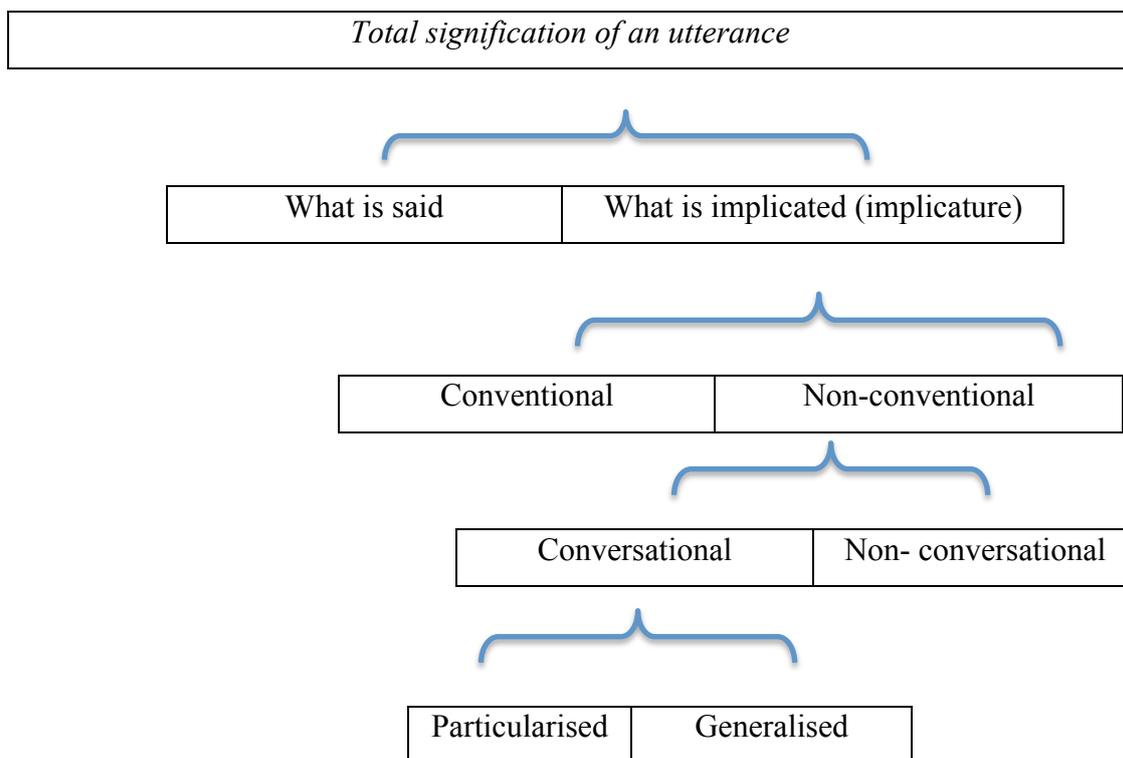
2.3.1 Pragmatic theories

The purpose of the fairly detailed explanation of the pragmatic theories of implicature in this subsection is first to establish the main concepts related the meaning and different types of implicature (drawn from Gricean theory of implicature) and then to give a clear description of the possible psychological mechanisms of implicature processing (under the Default theory and the Relevance theory, respectively). Understanding these concepts represents crucial basic knowledge for the current research and for the reader's comprehension of the empirical findings that underlie the theories.

2.3.1.1 Grice's theory of implicature

In his essay 'Logic and Conversation', Grice (1975, 1989) presented a theoretical account of the difference between what is (literally) said and what is indicated or hinted by a given utterance; he proposed the term *implicature* to refer to what is

implicated (pragmatically inferred) as opposite to what is said (linguistically coded). Grice (1978, 1989) called the sum of what is said and what is implicated the *total signification* of an utterance. He also identified several types of implicatures; his overall pragmatic theory of implicature can be schematically represented as below.



To interpret the figure and for the subsequent discussion, it might be helpful to recall briefly how Grice defined each type of implicature. First, *conventional implicature* refers to the case where ‘the conventional meaning of the words used will determine what is implicated, besides helping to determine what is said’ (Grice, 1975, p. 25); Grice gave this example to clarify his definition:

(1) He is an Englishman; he is, therefore, brave. (Grice, 1975, p. 25)

The conventional implicature here generated from the use of the lexical item *therefore*, that is to say, the relationship between ‘being brave’, and the antecedent, ‘being an Englishman’, is explicitly implicated in the utterance (Grice, 1975, pp. 25–26). Conventional implicature has two main features distinguishing it from conversational implicature (see below), namely, detachability and non-cancellability.

Detachability is the feature where the conventional implicature generated from an utterance can be re-generated using synonyms of the explicit lexical expression that produces this implicature. For instance, in example (1), one would make the same inference if the expression *therefore* were replaced with *hence*, *thus*, or *consequently*. *Non-cancellability* is the phenomenon where conventional implicature cannot be withdrawn in certain contexts without leading to a contradictory statement. To apply this to example (1), it can be seen that cancellability of the implicature would lead to an obvious contradiction, as in (1*) below.

(1*) He is an Englishman; he is, therefore, brave. Yet, his being brave does not consequently result from his being an Englishman.

Let me now move to introduce the non-conventional types of implicature, starting with *conversational implicature*, which Grice (1975, 1989) defined as an implicature giving rise to an inference generated from ‘certain general features of discourse’ (Grice, 1975, p. 26). These features are as follows. First, effective linguistic exchanges are posited to be ruled by a general principle, the *Cooperative Principle*.

Grice (1975, p. 26) defined the Cooperative Principle as ‘make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged’; it can be further articulated into four *maxims* and *submaxims* (see below). Second, when the speaker apparently violates the Cooperative Principle during the linguistic exchange, the hearer will assume that the speaker is still cooperating and made the violation purposely, and thus that the principle is functioning and can be detected at some deeper level; this is where inferences are generated.

Grice (1975, p. 26) identified the following specific maxims and submaxims under the Cooperative Principle:

The Maxims of Conversation (Grice, 1975, pp. 26–27):

Quantity:

(Submaxim) 1. Make your contribution as informative as is required (for the current purposes of the exchange).

2. Do not make your contribution more informative than is required.

Quality: Try to make your contribution one that is true.

1. Do not say what you believe to be false.

2. Do not say that for which you lack adequate evidence.

Relation:

1. Be relevant.

Manner: Be perspicuous.

1. Avoid obscurity of expression.

2. Avoid ambiguity.

3. Be brief (avoid unnecessary prolixity).

4. Be orderly.

Grice (1975, 1989) explained the connection between the Cooperative Principle (and its maxims) and conversational implicature by explaining the various ways in which an interlocutor in a talk exchange may fail to obey a maxim. That is, Grice (1975, p. 30) explained, an interlocutor

1. May quietly and unostentatiously *violate* a maxim; if so, in some cases, he will be liable to mislead.

2. May *opt out* from the operation both of the maxim and of the Cooperative Principle... He may say, for example, *I cannot say more*

3. May be faced by a *clash*. He may be unable to fulfil the first maxim of Quantity without violating the second maxim of Quality.

4. May *flout* a maxim, that is, he may blatantly fail to fulfil it.

Grice (1975, 1989) explained that conversational implicature is generated in a situation where a speaker *flouts* or is taken to flout a maxim. That is, under the assumption that the speaker is able to fulfil the maxim without violating another maxim and without opting out, and is not attempting to mislead, the hearer will assume that the speaker is committed overall to the Cooperative Principle, and that in this situation a maxim is being *exploited*; this is what gives rise to conversational implicature. Let us clarify this with one of Grice's examples where the Maxim of Quantity is being flouted. Grice (1975) gave an example of A writing a reference letter about a student applying for a philosophy job, *Dear Sir, Mr. X's command of*

English is excellent and his attendance at tutorials has been regular. Yours, etc. (p. 31). In this example, as Grice explained, A is not opting out, otherwise he would not write the letter at all; he is not unable to say more, since X is his student; and A is aware that in this letter more information is desired. Therefore, A must be wishing to convey information that he is unwilling to write down. This assumption is only plausible if A thinks that X is not good in philosophy or otherwise unsuitable for the job, and that what A is implicating.

Another situation in which a conversational implicature might be generated is when the speaker *violates* a maxim. Consider Grice's example on the violation of Relevance.

(2) A: I am out of petrol.

B: There is a garage around the corner. (Grice, 1975, p. 32)

In (2), the speaker B is implicating that the garage is open and further that it is a place where A can get petrol; if this is not what she means, the comment violates the Maxim of Relevance (barring any extraordinary or absurd interpretations of the situation). Grice (1975, 1989) called such implicature *particularised conversational implicature* (PCI) and briefly distinguished it from another type of implicature: *generalised conversational implicature* (GCI). The former type depends on contextual assumptions, as in example (2) above, whereas the latter is generated using certain words in the actually spoken utterance. Consider Grice's example on GCI:

(3) X is meeting a woman this evening. (Grice, 1975, p. 37)

In the GCI (3) above, the speaker's use of the indefinite noun phrase *a woman* indicates that the woman X is meeting is unknown to (some combination of) X, the speaker, and the addressee—not, for instance, X's mother, wife, or a platonic friend.

The last notion from Grice's theory of implicature to be clarified here is the second type of non-conventional implicature, namely *non-conversational implicature*. Levinson (1983, p. 131) defined this as a nonconventional inference 'produced by different maxims or language usage' other than the Cooperative Principle and its

maxims, for example an implicature derived from the *informativeness principle*, which might sometimes conflict with the Maxim of Quantity. According to Levinson (1983), the informativeness principle allows the hearer ‘to read into an utterance more information than it actually contains, in contrast to Quantity, which only allows the additional inference that (as far as the speaker knows) no stronger statement could be made’ (1983, p. 146). He gave the following example to clarify:

(4)

He turned on the switch and the motor started. (Levinson, 1983, p. 146)

(i) He turned on the switch and *then* the motor started.

(ii) He turned on the switch and *therefore* the motor started.

(iii) He turned on the switch and *this caused* the motor to start.

Levinson explained that upon applying the Maxim of Quantity, only (4i) and (4ii) are licensed, while the stronger implicature (4iii) should be banned. In other words, the informativeness principle allows the hearer to apply his or her knowledge of the world to derive an implicature that is informationally stronger (based on a cause-effect relation licensed by world knowledge) than the actual utterance (‘and’ might only convey an order relation). The inability of Grice’s Maxim of Quantity to generate the stronger implicature in (4iii) might be indirectly connected with a limitation pointed out by Sperber and Wilson (2012, p. 266): ‘Grice was rather non-committal on the source of pragmatic abilities and their place in the overall architecture of the mind.... He was equally non-committal on the form of the comprehension process’. That is to say, inferring the meaning in (4iii) requires the availability of a cognitive capacity (e.g. world knowledge) to understand the cause-effect relationship, and Grice’s theoretical account of communication seems to be limited in its explanatory power for this kind of inference.

To sum up, this section has aimed to cover the basic concepts of Grice’s theory of implicature, define the different types of implicature, and briefly highlight some limitations of the theory. It should also be mentioned, however, that Grice acknowledged in passing the existence of types of implicature other than those discussed above: ‘There are, of course, all sorts of other maxims (aesthetic, social, or moral in character), such as ‘Be polite’, that are also normally observed by

participants in talk exchanges, and these may also generate nonconventional implicatures' (Grice, 1975, p. 28). Such maxims were not discussed in Grice's published lectures, however, or any indication of their nature given, possibly because his proposal was 'relatively brief and only suggestive of how future work might proceed' (Levinson, 1983, p. 100). In the next section, I discuss later theories that attempt to clarify the possible mechanisms of implicature processing.

2.3.1.2 Theories on implicature processing

The debate over implicature processing centres on two main theories, the *Default theory* developed by Levinson (2000) and the *Relevance theory* proposed by Sperber and Wilson (1986/1995). In this section, I first explain the two theories in some detail; then, the possible computation mechanisms of *scalar implicature* are discussed in relation to each theory.

Default theory

The Default theory, proposed by Levinson (2000), is also known as the 'theory of Generalised Conversational Implicature'. It expands upon Grice's theory of implicatures, concerning itself basically with GCIs rather than PCIs. According to Levinson, GCIs are defeasible inferences generated from the speaker's choice of utterance (syntactic form and lexical items/lexicosemantics) via three mutual heuristics presumed to function between the speaker and hearer. The heuristics, which derive from Grice's maxims, are the *Q-heuristic* (related to quantity), the *I-heuristic* (related to informativeness), and the *M-heuristic* (related to manner). Levinson defines his Q-heuristic as 'What is not said, is not', and gives an example: *There is a blue pyramid on the red cube*, which generates the inference that 'there is not a cone on the red cube; there is not a red pyramid on the red cube' (Levinson, 2000, p. 31). The I-heuristic posits that 'What is simply described is stereotypically exemplified' (Levinson, 2000, p. 32). He explains that in an utterance such as *The blue pyramid is on the red cube*, the I-heuristic generates the inference that there is direct contact between the pyramid and the cube, since if the contact was indirect, the speaker would have specified this, because the stereotypical situation described by *on* indicates direct contact. The M-heuristic suggests that 'What is said in an abnormal

way is not normal' (p. 33). For example, in *The blue cuboid block is supported by the red cube*, the hearer can infer that the block in question is not a prototypical one due to the marked form *cuboid*.

The central claim of Default theory is that the inference computation process occurs at two levels. First, the semantic representation derived from the syntactic structure and lexical items of a sentence may be underspecified. On the basis of this semantic process, default, defeasible pragmatic inferences (GCIs) are generated to determine an utterance's meaning. After the utterance's meaning is decided by this semantic process, another, pragmatic, process starts (only under certain contextual conditions, e.g. when adding further information) to derive other inferences such as PCIs; this final process produces speaker meaning. In other words, the theory suggests that there is an intermediate level of meaning (GCIs) which comes between the literal meaning (semantics) of an utterance and its pragmatic interpretation (inferences); 'they sit midway, systematically influencing grammar and semantics on the one hand and speaker-meaning on the other'. (Levinson 2000, p. 25).

Relevance theory

In contrast to Grice's (1975, 1989) and Levinson's (2000) proposals making use of GCIs and PCIs, Sperber and Wilson (1986/1995) argue that in communication the hearer only derives context-dependent inferences (PCIs). Thus, unlike Grice and Levinson's code model of communication, which suggests that the comprehension process of an utterance is achieved by the listener's decoding the content of the message that the speaker intended to convey ('what is implicated') based on the content s/he linguistically encoded ('what is said'), for Sperber and Wilson (1986/1995) comprehension is achieved by the listener through inference of what the speaker intended to convey—the speaker's *communicative intention*—a process that goes beyond just decoding linguistic messages, as it requires the hearer to apply all information available, of various kinds (e.g. contextual assumptions, the speaker's intentions, world knowledge), to get at what the speaker intended to convey.

A second central idea in Sperber and Wilson's (1986/1995) inferential model of communication is that the pragmatic process of deriving inferences depends on the

principles of *relevance*, rather than the conversational maxims proposed by Grice (1975, 1989). Pragmatic relevance in Sperber and Wilson's (1986/1995) theory differs from Grice's maxim 'Be relevant'; in their definition, 'an input is relevant to an individual when it connects with available contextual assumptions to yield positive cognitive effects' (Sperber & Wilson, 1986/1995, p. 251). These positive cognitive effects are measured in relation to the degree of cognitive effort required to derive an inference: The smaller the mental effort required to derive an inference, the greater the relevance of the input for the receiver.

Relevance theory is based on two general assumptions about the role of relevance in cognition and communication:

First (cognitive) principle of relevance: Human cognition tends to be geared to the maximisation of relevance.

Second (communicative) principle of relevance: Every act of ostensive communication communicates a presumption of its own optimal relevance.

(Sperber & Wilson, 1986/1995, p. 260).

The *maximisation of relevance* under the cognitive principle, as Sperber and Wilson (1986/1995) explained, means that inputs are processed by making the most efficient use of the available processing (cognitive and contextual) resources, that is, the use that needs as little processing effort as possible. The communicative principle suggests that 'ostensive-inferential communication involves the use of an ostensive stimulus, designed to attract an audience's attention and focus it on the communicator's meaning' (Sperber & Wilson, 1986/1995, p. 255). In Relevance theory, the hearer can interpret the meaning conveyed by an ostensive stimulus (or an utterance) under the presumption of optimal relevance only if:

- a. The ostensive stimulus is relevant enough to be worth the addressee's effort to process; and
- b. The ostensive stimulus is the most relevant one compatible with communicator's abilities and preferences (Sperber & Wilson, 1986/1995, p. 270).

An example of *ostensive stimulus* might be, in contrast to a situation where a host notices that his guest has an empty glass and infers that she might like a drink, a situation where the guest ostentatiously waves her glass in front of the host, who will then derive the stronger conclusion that she *would* like a drink (Wilson & Sperber, 2002).

Overall, Relevance theory presupposes a psychological conception of pragmatics built on the representational theory of mind. That is, it sees that for successful communication, the hearer should go beyond what is linguistically conveyed to infer the speaker's intended meaning, which might require applying *mind-reading ability* (Sperber & Wilson, 2002).

2.3.2 Empirical evidence on scalar implicature processing at the surface and neural levels

In the past two decades, experimental studies on implicature processing have focused predominantly on scalar implicature, a process that requires sensitivity to Grice's (1975, 1989) first (sub)-maxim of Quantity: 'be as informative as required'. Most of these studies have relied for a measure on the mean reaction time (RT) associated with pragmatically enriched context, taken as evidence of recognition and processing of the enrichment by the listener (e.g. Noveck & Posada, 2003; Bott & Noveck, 2004; Katsos, Breheny, & William, 2005; Breheny, Katsos & William, 2006; Slabakova, 2010; Breheny, Ferguson & Katsos, 2012; Breheny, Ferguson & Katsos, 2013; Zajenkowski & Szymanik, 2013); some have also used rate of pragmatically enriched responses (Slabakova, 2010; Antoniou, Katsos, Grohmann, & Kambanaros, 2014). Recently, a number of studies have gone to a neural level to explore the neural activation correlates of implicature processing (e.g. Zhao, Liu, Chen, & Chen, 2015; Politzer-Ahles & Gwilliams, 2015). The purpose of this section is to map pragmatic theories about implicature processing over behavioural and neural data onto scalar implicatures for a better understanding of the possible computation process and the cognitive effort involved. Before considering the findings of prior empirical work at both surface and deeper (i.e. neural) levels, the tools these studies have used to measure processing effort, and the paradoxical evidence that they have reported, we need first to understand the type of implicature considered in all these previous

works—scalar implicature—and to briefly consider what makes it an ideal choice to explore these processing mechanisms as compared to other types of implicature.

2.3.2.1 Scalar implicature

Scalar implicatures involve sensitivity to Grice’s first (sub-)Maxim of Quantity, ‘be as informative as required’, and are generated using a given scale in a given context, usually the lexical scales proposed by Horn (1972) (e.g. (*all*>*most*>*some*), (*and*>*or*), (*must*>*might*)). In contrast to other types of implicatures that only yield PCIs (e.g. those generated from Quality and Relevance maxims), scalar implicatures are considered prototypical examples of GCIs (Breheny et al., 2006), generated by the use of the weaker term in the scale; at the same time, they allow calculation of PCI if the context requires this. The following example represents the computation mechanism for scalar implicatures (Breheny et al., 2006):

- (5) A: Did the students pass the exam?
B: *Some* of the students passed.
GCI: *Not all* the students passed.
- (6) A: Was the exam easy?
B: *Some* of the students passed.
GCI: *Not all* the students passed.
PCI: The exam was difficult.

In example (5), in the Default view only a GCI is generated—automatically from the use of *some* in context. This inference can be cancelled without conflict if the speaker adds a phrase such as *well, actually all of them passed*. In example (6), in contrast, two types of implicature arise: the context-independent GCI is generated (automatically by a default mechanism) from the use of *some*, but due to contextual requirements, the PCI ‘the exam was difficult’ is calculated, as it appropriately conveys the implicated meaning. On the other hand, under Relevance theory, the hearer would only derive a context-dependent inference, and such a process would be expected to require some cognitive effort.

Given these alternative approaches, how can one determine which is more plausible to explain implicature processing? In the attempt to answer this question, various scholars have explored the effort involved in processing scalar implicature—either on the *surface level*, through ‘e.g. the rate of responses indicating a pragmatic enrichment, or the mean reaction time associated with such an enrichment’ (Noveck & Sperber, 2012, p. 317), or on the *deeper, neural level*, by measuring research participants’ neural activation while processing scalar implicatures. Before reviewing the empirical findings on implicature processing, let us discuss the different types of scale that might arise in such a context.

2.3.2.2 Types of scale

Previous studies have investigated quantity implicatures generated by three types of scale. The first is the *lexical scale* proposed by Horn (1972), which arranges lexical terms from the same grammatical category but varying in their semantic informativeness, for example (*all>most>some*), (*and>or*), (*adore>love>like*). The idea is that the use of a semantically weaker term in an utterance implicates that the stronger scalar term does not hold (Dieussaert et al., 2011). That is, if a speaker says that *some of the students passed the exam*, then, the hearer will infer that ‘not all of the students passed the exam’. Assuming that the speaker was cooperating by being appropriately informative, she would have used *all* if that was the case, but she used the weaker term *some*, and so *all* must not apply.

The second type of scale is the *ad hoc scale*, which is context dependent, that is, generates only in a specific context. For example, in (7), interlocutor A, assuming that B is cooperating by being informative, will infer that B did not meet David, otherwise B would have said so. This type of implicature is a PCI, as it emerges only in specific contexts.

(7) A: Did you meet John and David yesterday?

B: I met John.

The last type of scale is the *encyclopaedic scale*, where implicatures are licenced by world-knowledge (Papafragou & Tantalou, 2004). Although Papafragou and Tantalou (2004) attempted to explore children’s ability to derive inferences generated from

such a scale, it might be claimed that the stimuli in their encyclopaedic condition did not actually generate inferences dependent on world-knowledge, but rather on visual context. I will explain. Supported with visual context, child participants in their experiment were presented with an animal (a bear) that had to eat a sandwich made of bread, cheese, and ham; the bear went to a dollhouse so as not to litter the room with crumbs, and when he came out, the experimenter asked him *Did you eat the sandwich?* and the bear responded *I ate the cheese*. It might be argued that in order to derive the correct inference here, a child would not need to rely on world-knowledge, since it would be enough to compare the verbal utterance with the visual context to detect the violation of informativeness.

Compare the example of Papafragou and Tantalou (2004) with the examples given below in two conditions: context and no-context.

(8) *Without context*

(i) Chemically, water consists of two ions of hydrogen.

(ii) In a typical human being, each cell consists of 23 pairs of chromosomes; 22 pairs are autosomes and the remaining pair is the sex chromosomes, which consist of XX.

(9) *With context*

Adam: Any plans for the weekend?

Sara: I am going shopping on Saturday.

Adam: OK, would you join us for lunch on Sunday?

Here, one can only infer that (8i) and (8ii) are under-informative based on one's world-knowledge. That is, water also consists of one ion of oxygen, and the XX pair of chromosomes in (ii) only represents females, while the male pair is XY. Similarly, in the context condition (9), the hearer (Adam) builds on his knowledge that the weekend has two days (Saturday and Sunday), infers that the speaker has no plans for Sunday, and thus invites her for lunch.

Intuitively, maintaining the world-knowledge relevant to utterances that use an encyclopaedic scale seems to be an essential requisite to derive implicatures from those utterances. For instance, a young child would not be able to detect the violation

in (8), which requires a level of world-knowledge we would not expect the child to acquire till a later age. Thus, if we want to test children's ability to derive inferences that depend on encyclopaedic knowledge, we should construct examples that do not rely on advanced world-knowledge. One such example employed in the present study is as follows: *To clap, you need to use your right hand*; the scale here is <both right and left hands, right hand, left hand>, and a child needs only to build on world-knowledge to infer that the action (clapping) could not be completed without the interaction of both hands.

The next section sheds light on empirical findings on implicatures generated basically from Horn's lexical scale.

2.3.2.3 Evidence on scalar implicature processing at the surface level

The approach of relying on RT as an indicator of the mechanisms from which inferences are generated is based on the assumption that if participants take a longer time to make a pragmatically enriched interpretation than a literal interpretation of a scalar implicature, this implies that the computation process does not occur by means of default mechanisms but rather that it is cognitively effortful (because more time consuming). For example, Noveck and Posada (2003) and Bott and Noveck (2004) explored response RT to statements of the form *some elephants have trunks*; participants were asked to evaluate similar statements and respond with either 'false/disagree' or 'true/agree'. The former response was taken to imply that the participants had made an enriched pragmatic interpretation ('some but not all'), while the latter would indicate a literal interpretation ('some and possibly all'). By comparing response times between cases where scalar inferences were and were not computed, both Noveck and Posada (2003) and Bott and Noveck (2004) found that participants took a longer time to make a pragmatic judgement (responding with 'false/disagree'), and concluded that such results support the Relevance view rather than the Default view.

Using similar stimuli to the studies just cited, Tomlinson, Bailey and Bott (2013) tracked computer mouse movements to explore the scalar implicature computation process. Their results revealed that when participants were asked to evaluate a

statement like *some elephants are mammals* by clicking true/false on the computer screen, the trajectory of mouse movement for participants who made pragmatic responses revealed that they first moved toward ‘true’ before ultimately selecting ‘false’, a decision which requires pragmatic enrichment. Tomlinson and colleagues suggest that these results indicate that pragmatic processing occurred in two steps, the first involving logical (literal, semantic) reading of the statement before the pragmatically enriched reading was made in the second. They suggest that the two-step model is in line with Relevance theory, in which, pragmatic processing involves two stages, the first requiring decoding/retrieving context-independent meaning, after which enrichment is implemented (or the utterance is re-interpreted with enrichment added).

However, the stimuli employed in these studies have been criticised for being completely artificial laboratory stimuli disparate from those that arise in everyday conversation (Geurts, 2010). If the claim of Noveck and Posada (2003) and Bott and Noveck (2004) that enriched inferences will take a longer time taken to derive is true, that is, such results should be replicable in a more natural and thus more valid design. The first study in this vein, by Bezuidenhout and Cutting (2002), measured the time-course of a text comprehension task with naturalistic stimuli. The findings of this study, however, contradicted the results of Noveck and Posada (2003) and Bott and Noveck (2004), actually finding that texts which required pragmatic enrichment for proper interpretation were read faster than those which did not, and thus supporting the Default account. However, Katsos et al. (2005) and Breheny et al. (2006) highlighted some experimental issues related to the stimuli used by Bezuidenhout and Cutting (2002), explaining that most of the items were borrowed from off-line studies and were aimed neither at testing nor at generating scalar implicatures. Therefore, Breheny et al. (2006) replicated Bezuidenhout and Cutting’s study to ensure that the experimental items were genuinely generating scalar implicatures. Breheny et al. (2006) measured the reading times that adult participants required to read short texts with one of two types of context: *Lower-bound contexts*, where the literal (semantic) reading of a scalar term is more suitable, as in (10), and *Upper-bound contexts*, where the enriched (pragmatic) reading of the scalar term is more suitable, as in (11).

(10) *Lower-bound context*

John heard that/the textbook for Geophysics/was very advanced./Nobody understood it properly./He heard that/if he wanted to pass the course/he should read/*the class notes or the summary.*/ (Breheny et al., 2006, p. 443; the critical phrase is in italics, as in the original)

(11) *Upper-bound context*

John was taking a university course/and working at the same time./For the exams/he had to study/from short and comprehensive sources./Depending on the course,/he decided to read/*the class notes or the summary.*/ (Breheny et al., 2006, p. 443; the critical phrase is in italics, as in the original)

Breheny et al. (2006) suggested that if participants proved to take a longer time to read lower-bound contexts (as in Bezuidenhout and Cutting's study) then this would support the Default theory and confirm the outcome of Bezuidenhout and Cutting (2002). In contrast, if participants needed longer to read the upper-bound contexts, which require pragmatic enrichment, then this gives support to the Relevance theory, and is in line with Noveck and colleagues' above-discussed studies. In the actual event, Breheny and colleagues found that reading texts which included pragmatic enrichment took longer, and thus concluded that scalar inference computation requires contextual support, as suggested by Relevance theory.

Other recent studies employed a *visual world paradigm* to explore the time-course of implicature processing. For instance, Breheny et al. (2012) adapted the *look-and-listen method*, using eye-tracking to examine the time-course of an experimental task in order to access scalar implicatures. In their study, participants watched short videos depicting an agent transferring quantities of different items to one of two locations. At the end of each video, participants viewed the last still frame and heard a pre-recorded auditory description of the events that had just happened; at the same time, their eye movements around the visual display were recorded. It was predicted that the eye movements would reveal participants' expectation of upcoming objects in the discourse before these objects were uttered; by measuring the amount of time participants needed to fixate their gaze on the target object in contexts that involved the derivation of inferences compared to contexts that did not, Breheny et al. (2012) were able to compare the time-course of the three conditions (as outlined below in

(12)) to determine whether inference derivation might cause any delay in directing the eye toward the object. Examples of the stimuli used in their study are as follows:

(12) [1] *All*

The man has poured all of the water with oranges in it into the bowl on tray B and some of the water with limes in it into the bowl on tray A.

[2] *Some early*

The man has poured some of the water with limes in it into the bowl on tray A and all of the water with oranges in it into the bowl on tray B.

[3] *Some late*

The man has poured some of the water with limes in it into the bowl on tray A and some of the water with oranges in it into the bowl on tray B. (Breheny et al., 2012, p. 449; conditions in italics as in the original)

Breheny and colleagues found that the time the participants took to fixate their gaze on the target object was very close in [1] and [2] but significantly differed in [3], and concluded on this basis that the process of deriving a scalar inference ('some but not all') is relatively rapid. The time-course in [1] and [2] significantly differed from the baseline condition [3], which could be expected due to an ambiguity effect (resulting from the agent's pouring the same quantities of water into A and B so the participants would not be able to predict the upcoming location when hearing 'some' since it would apply to both locations). These results replicated the findings of Grodner et al., (2010), who also found very rapid access to scalar implicature in such a case.

Breheny et al. (2013) conducted further research to find out whether this rapid access would also emerge with PCIs (context-dependent implicatures), and determined that it did. That is, with a stimulus such as *the woman put a spoon into box B and a spoon and a fork into box A*, Breheny et al. (2013) found that participants fixated their gaze on the relevant object (box B) once they began to hear the preposition (i.e. at the onset of *into*); the authors claimed on this basis that this implied that the participants derived the PCI ('a spoon and nothing else') very rapidly, a finding in line with Breheny et al.'s (2012) findings on scalar inferences.

Although Breheny et al. (2012, 2013) considered such results to constitute good evidence for rapid access to GCIs (scalar implicatures) and PCIs, we should be

cautious about this account (and in particular the idea that interpretation is automatic), for several reasons. First, this rapid access to inferences, whether they are considered GCIs or PCIs, might result from the nature of the task, which, unlike in previous work on implicature processing, required participants to employ both visual and verbal short-term memory (STM). That is to say, the auditory–visual interaction might facilitate the derivation process; and indeed neuroimaging research has shown that participants were able to identify audiovisual stimuli more quickly and more accurately than stimuli which were only either auditory or visual (Giard & Peronnet, 1999). Furthermore, neuroimaging evidence (Prabhakaran, Narayanan, Zhao, & Gabrieli, 2000) has shown that the integration of visual and verbal STM enhances the efficiency of working memory (WM) (the ability to manipulate information in the brain and recall it back voluntarily). Since it has also been shown that WM is involved in adults’ processing of scalar implicature (e.g. Feeney et al., 2004; De Neys & Schaeken, 2007; Dieussaert et al., 2011; Zajenkowski & Szymanik, 2013), then one might suggest that the fast access to implicature seen in Breheny’s studies might be, at least in part, a result of the high efficiency of WM.

Second, the results of Breheny et al. (2012, 2013) might be attributed to the high artificiality and the demands of the task, which makes it very unlike everyday use of language and might lead the brain to adopt different systems of reasoning than those used in everyday language comprehension (see Evans, 2003; Schroyens, Schaeken, & Handley, 2003; De Neys, 2006, for the effect of task demands on reasoning). In the real world, we do not watch events and then listen to how a communicator might describe them, as we map their utterances to what we have just viewed—we might view this as a *describing* rather than an *indicating* task, and this kind of task might lead a participant’s brain (i.e. processing system) to adjust itself to employ heuristic (conditional) reasoning, which works rapidly and automatically, rather than analytic reasoning requiring a deeper level of processing (see Barrouillet (2011) for a detailed review of the two systems of reasoning). That is to say, the hearer might (consciously or unconsciously) solve the task on the surface level (‘If *some* then B, if *all* then A’), without going to the deeper level required for implicature (pragmatic) computation (‘some but not all’). In addition, the late *some*-condition might not serve as an accurate indicator or baseline; that is, *some* and *all* might be interpreted by participants as referents to objects rather than scalar terms that could themselves

generate inferences, and it seems unclear how the ambiguity in the late *some*-condition would work as a baseline. It might be more useful to understand this phenomenon to include (unambiguous) control items and compare them with the critical ones to find out whether the hearer's behaviour will differ when, for instance, quantificational terms are not used (e.g. *the man poured the green water into X and the blue water into Y*). Similarly with the PCI, it might be expected that the brain would search for the most relevant trigger in terms of solving the task rapidly, in this case 'If *into* then B, if *and* then A'. Furthermore, if the stimuli included contexts such as *the man poured only some of the water into X and all of the water into Y* or *the man put only a spoon into X*, fast fixation cannot be taken as evidence of rapid access to the implicature, and is better understood as a brain adaptation to find the most relevant information to solve the task. This assumption of relevance is in line with the Heuristic-Analytic theory of reasoning (Evans, 1996) which has received empirical support from studies using an eye-movement paradigm (e.g. Ball Lucas, Miles, & Gale, 2003; Ball, Lucas, & Phillips, 2005).

Third, the cost of processing an implicature might not be captured on the surface (i.e. behavioural) level (whether through eye movement, mouse movement, RT, or response rate), but rather at the neural level, represented by activation in different neural regions (e.g. Zhao et al., 2015; Politzer-Ahles and Gwilliams, 2015). Of course, this goes beyond the scope of Breheny and colleagues' studies; but the results of the neural imaging research that will be discussed in section 2.3.2.5 might make us very cautious in the way we treat behavioural data. Furthermore, the increased evidence for the role of WM (e.g. De Neys & Schaeken, 2007; Dieussaert et al., 2011) in implicature might suggest that multiple tasks that measure different aspects of human cognitive and pragmatic abilities should be employed when measuring the cost of processing. We are not merely interested to find out whether implicatures are processed rapidly or slowly, but are keen to understand *why* they are processed in whichever way, and it might be that only by combining measures assessing various abilities can we understand the real reasons for the rapidity (or slowness) of inference processing.

Another study that explored the RT of scalar implicature processing, where utterances were supported by visual context, was conducted by Zajenkowski and Szymanik

(2013). They investigated the relationship between intelligence, WM, STM, and ability to direct attention, on the one hand, and RT for judging utterances with different types of quantifiers (e.g. *some of the Xs*, *more than half of the Xs*, *an even number of the Xs*), but without specifically exploring scalar implicatures. The participants' intelligence was assessed with a test of fluid intelligence (Raven's Advanced Progressive Matrices Test (Raven, Court, & Raven, 1983)), while attention was assessed with the Attentional Network Task (ANT) designed by Fan, McCandliss, Sommer, Raz, and Posner (2002). The ANT is a computerised task in which the participant is asked to press a relevant key, 'left' or 'right', depending which direction a central arrow stimulus (the target) points, while being flanked by distractor stimuli, and appearing above or below a central fixation point; RTs are recorded. WM was assessed using a reading span task, in which participants were asked whether sentences they read were true, and were asked to memorise the last word of each sentence to be recalled later in the task. STM was measured with a digit task, in which participants were presented with a series of several numerical digits on a computer screen, appearing simultaneously, for only 300 ms; after this, they viewed a test digit and were asked to decide whether it had appeared in the previously presented string. Finally, semantic processing of quantifiers was assessed with a computerised sentence–picture verification task, in which the participants viewed different pictures consisting of combinations of 15 objects in different proportions; each image had the same type of objects but differing in colour (e.g. 7 black cars and 8 white cars) accompanied with statements describing the picture (e.g. *some of the cars are white*). The participants had to decide whether the statement described the picture accurately; again, RT was measured.

The results of Zajenkowski and Szymanik (2013) reveal that participants with higher intelligence scores responded faster than those with lower intelligence scores. They also found that while WM and STM were strong predictors for RT in the quantifier processing task, attention seemed to play no significant role; especially when intelligence scores were added to the model, explaining most (more than half) of the variation in the proportional quantifiers.

2.3.2.4 Measuring the cost of processing by rate of pragmatically enriched responses: Evidence from bilinguals

A few studies attempt to interpret variation in pragmatic performance between bilingual and monolingual adults and children in terms of the Default and Relevance theories of the processing of implicature, asking whether the results better support the functioning of automatic processing mechanisms, or cognitively effortful processing. Previous studies that have explored the validity of these mechanisms by explaining the variation in their results have often relied on the rate of pragmatically enriched responses as an indicator of pragmatic competence. The idea is that, since there is established evidence of bilingualism's impact on the various core components of executive functioning (EF), proposed by Miyake et al. (2000), (for evidence of bilingualism's general cognitive advantage, see Bialystok, 2011; Bialystok, Craik, Green, & Gollan, 2009; Blumenfeld & Marian, 2011; Morales, Calvo, & Bialystok, 2013), then, a higher rate of pragmatic interpretation might suggest much easier access to implicature, a condition that can be associated with EF advantage and thus taken as indirect evidence for the Relevance perspective. That is to say, when outperformance in pragmatic tasks clearly emerges in bilingual individuals compared to their monolingual counterparts despite the former's more limited resources in a given language (Siegal, Matsuo, Pond, & Otsu, 2007; Siegag, Iozzi, & Surian, 2009), this can be taken to indicate that the process did not occur by default mechanisms but required additional cognitive effort which was served by bilingual cognitive advantage (higher active EF abilities).

For instance, Slabakova (2010) relied on rate of acceptance of under-informative items in two context conditions: enriched context v. no context. For an enriched context task, she adapted a task from Papafragou and Musolino (2003) and presented Korean and English native speakers and Korean learners of L2 English with a series of pictures depicting a fictional character ('Charlotte') who, in all the given stories, commits some action on either two out of three or three out of three items. For each story, the experimenter read the sentences describing the events to each participant, individually. At the end of each story, Charlotte's mother asks her a question like *Charlotte, what have you been doing with the candies?* and the girl responds with an utterance in either the under-informative form *I ate some of them*, or the optimal form *I ate all of them*. The participants were asked if they agreed or disagreed with

Charlotte's response. In the no-context condition, Slabakova replicated Noveck's (2001) experiment on universal and logical quantifiers, for example, *some giraffes have long necks*. Her results revealed that the Korean learners of English significantly outperformed both native groups by making more pragmatic interpretations in both conditions; within their own results, they derived more pragmatic inferences in the enriched-context condition rather than the no-context (90% v. 60%).

Slabakova suggested that such results give support to the Default theory: 'If implicatures were more effortful in terms of processing resources, the fact that L2 learners make them more often than native speakers would go against everything we know about processing in a second language'. She claimed that 'the L2 learners lack the processing resources to undo automatic pragmatic interpretations', and that since these resources were available to the natives in her experiment, they made more logical interpretations. To clarify this point further, Slabakova claimed that while evaluating the items, the natives' logical interpretations resulted from their ability to think of alternative contexts and that therefore they accepted the under-informative items, while such a privilege was not available for the L2 learners. Although Slabakova referred briefly to the possible role of EF, she disregarded it when interpreting the variation in her results.

However, Slabakova's reason for concluding that bilingual outperformance is due to default mechanisms seems implausible, for several reasons. First, she claimed that the natives in the no-context conditions were able to think of alternative logical interpretation to *some cats have tails*, such as 'some tails might be accidentally removed or cut', and therefore accepted the statement, while L2 learners could not think of such alternatives due to limited resources. If this is true, it means that both the natives and the L2 learners had enriched the meaning of *some* to 'not all', and that the subsequent rejection or acceptance of the alternative context depends on searching through world-knowledge, which cannot be measured and seems independent from linguistic resources as such. Since we would, in principle, expect adult groups to have similar levels of world knowledge, this leads us to ask: why did the L2 group not make logical interpretations as the other groups did? It is clear that language competence is not a sufficient answer, since the L2 learners outperformed monolinguals in their native language as well as in the L2. Thus, one might

reasonably disagree with Slabakova's interpretation that only automatic mechanisms for pragmatic enrichment are at play.

Antoniou and Katsos (2016) revisited their previous work (Antoniou et al., 2014) with a larger sample to examine the cognitive factors that support pragmatic development in children and to reconsider their results in terms of the Default and Relevance accounts. They tested Greek 4-to-12-year-old multilingual, bidialectal, and (monodialectal) monolingual children's pragmatic ability to derive different types of conversational implicatures (scalar implicature (Quantity maxim), Metaphor (Quality), Manner, Relevance). They also assessed children's EF abilities using a battery of cognitive tests that measured WM (the Corsi Blocks task and Backward Digit Span task (Wechsler, 1949)), inhibition (the Simon task and the Stop Signal task (Logan, 1994)), and *switching* (the ability to switch between tasks and rules flexibly, assessed by the Color-Shape task (Ellefsen, Shapiro, & Chater, 2006)). For pragmatic assessment, they used different task to assess each type of implicature.

Antoniou et al. (2014) and Antoniou and Katsos (2016) assessed children's ability to derive scalar implicatures using two tasks: an *act-out task* (adapted from Pouscoulous, Noveck, Politzer, & Bastide (2007)) and a *binary judgement* task. In the former, the children were presented with scenarios depicting five boxes and a selection of items, and were asked to construct a display matching the description they heard from a fictional character. For instance, in one stimulus, there were five boxes, each including a turtle, and the children heard an utterance such as *There are turtles in some of the boxes*; if they did not act to change the display by removing one or more of the turtles this would be taken to indicate that they failed to derive the scalar implicature. In the second scalar implicature task, the children viewed three or five cards, face down, and were asked to judge if concurrent verbal descriptions correctly represented the scene. For example, in one trial children listened to a pre-recorded utterance *there are X on Q of the cards*, where X represents item type (e.g. *rings, hearts*) and Q the quantifier type (*all, some, not all, none*); once the pre-recorded stimulus ended, the card was immediately, automatically turned over to reveal the items, and children were asked to decide if the utterance describing the scene was true or false by pressing the appropriate button. For other types of conversational implicatures, children were assessed using comprehension tasks in which they were

introduced to a fictional character ('George') and presented with several stories; at the end of each story the children were presented with two visual contexts (pictures) and were asked to select which one appropriately matched the end of the story or the event that George was describing. For example, to assess children's ability to understand metaphor, children heard stories about George and his father and at the end of each story they were asked to point out a picture that depicted how George's father felt at the end of the story. The stories ended in metaphors describing emotions of sadness (e.g. *When he returned back home George's father was a melting snowman*) or anger (e.g. *When he returned back home George's father was a thundering cannon*).

Antoniou and Katsos (2016) used a composite score for children's performance on the different types of conversational implicature and tried to explain the variation in children's performance in terms of their EF abilities, more precisely, in terms of their WM ability, since they initially found a partial correlation between WM and pragmatic performance. However, the regression analysis did not support such a relationship; instead, children's pragmatic ability was best predicted by age, language proficiency, and task version (it should be mentioned that children's language background (e.g. bilingual, bidialectal, monolingual) was not among the variables included in this model). They also conducted principal component analyses on children's performance on the different types of conversational implicatures that only showed a single factor in implicature performance. On the basis of these results, they claimed that their pragmatic results give support to pragmatic theories such as the Relevance one (e.g. Grice 1975, 1989; Sperber & Wilson, 1995) 'that treat all types of pragmatically inferred meanings as the outcome of a single pragmatic interpretation process that involves uncovering the speaker's intentions behind an utterance' (Antoniou & Katsos, 2016, p. 15).

Although Antoniou and Katsos's study gives some evidence that children might treat inferred meanings as a result of a single computation process, there are also some issues to be highlighted here. First, apart from the fact that their results were generated from different tasks, it was not clear how their pragmatic results would correspond to any of the pragmatic processing theories or inform us about the possible mechanisms of implicature processing, since some of the implicatures were completely context dependent (e.g. metaphor) while others were typical examples of

GCI (e.g. scalar implicatures), and relying on a single component of implicature performance as emerged from the principal component analysis to indicate a uniform process across the different types of implicature might not reflect a clear picture of how children treat each of PCIs and GCIs. Their reliance on a total composite score based on results from different kinds of implicature is another issue which may have affected the validity of their results. That is to say, the precise kinds of cognitive resources involved in processing scalar implicatures might be distinct from those employed in processing (e.g.) metaphor—for instance, scalar implicature computation might require the ability to apply basic mathematical operations to approximate magnitudes in a context which also requires participants to draw heavily on their WM (e.g. Feeney et al., 2004; McMillan et al., 2005; De Neys & Schaeken, 2007; Zajenkowski & Szymanik, 2013; Heim et al., 2016), while metaphor in contrast might require advanced mind-reading ability. Thus, combining the scores for these disparate implicatures would not give us the articulated understanding of the computation process that we might gain by treating them separately: we would not be able to know which EFs might affect which type of implicature and why.

To sum up, although most of the studies discussed above have yielded evidence on the possible psychological mechanisms of scalar implicature processing that indirectly supports the Relevance account, there is also a paradox in the results, or at least, the evidence presented by some studies lacks adequate empirical support when EF measures are considered. Thus, for a complementary view, I review in the next section the findings of studies that have gone to a neural level in exploring the cognitive processing cost of scalar implicature.

2.3.2.5 Measuring the cost of processing by neural activation

A recent study used similar stimuli to those used in Noveck and Posada's (2003) work to explore how Chinese-speaking adults process scalar implicature (Zhao et al., 2015). The study adopted the *mismatch negativity* (MMN) paradigm to detect neurophysiological indicators of automatic processing of scalar implicatures, using the ERP technique. Participants processed informative statements (e.g. *some animals have tails*) and under-informative statements (e.g. *some tigers have tails*) as neural activity was measured. The idea of the MMN paradigm is that when the brain receives

deviant stimuli (such as under-informative statements) that are mismatched with long-term memory (LTM) tracing, such deviant stimuli clash with stored LTM and thus induce an MMN of larger amplitude than those of standard stimuli (informative statements) that match LTM. Based on participants' results on the Autism-Spectrum Quotient (AQ) questionnaire (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), which they completed as a proxy measure of pragmatic competence, the study divided its sample into two groups (high and low pragmatic ability). By comparing ERPs while the participants heard and processed pre-recorded, highly controlled under-informative and informative stimuli (all consisting of four words uttered by a trained female speaker, with natural prosody), the study found significant ERPs only with the deviant stimuli in the high-pragmatic group, and no effect in the other group. This suggests that scalar implicature is generated by default mechanisms.

Again, however, one might disagree with the conclusions of Zhao et al. (2015), for several reasons. First, they found that the high-pragmatic group's performance differed significantly across the two informativeness conditions, in each of the left and right anterior and posterior regions (all $ps=0.001$). These high levels of activation while processing under-informative statements might indicate not default mechanisms but the opposite—evidence that the process is cognitively effortful. Second, it seems that their pragmatic results were consistent with the AQ questionnaire outcomes, and this result in itself might refute their support of Default; or put another way, the absence of any MMN effect in the low-pragmatic group, who performed equally (showed similar neural activation) in the two conditions, should be taken as further evidence to support the Relevance perspective. If the process occurs by default, and if we take into account the questionnaire results as indicator of pragmatic ability, then, when there is no context, (socio-)pragmatic ability to understand the speaker's intention seems to have no role, and the task is clearly measuring pragmatic-cognitive interaction rather than mind-reading ability, further circumscribing Zhao et al.'s conclusions. In a study that used similar methods to those employed by Zhao et al., Nieuwland, Ditman, and Kuperberg (2010) found some evidence that pragmatic ability (as measured by the communication subscale of the AQ questionnaire) affected scalar implicature processing; such results are in line with the Relevance account, since, as they stated, 'scalar inferences are not obligatory but depend on constraints from the context and possibly from neuropsychological factors' (p. 335).

Can this neural effect be detected only in terms of negativity with LTM or of effortful pragmatic processing? A recent study (Shetreet, Chierchia, & Gaab, 2014) using fMRI, measured the neural processing cost when the matching was based on visual context rather than LTM (picture-sentence matching). It also found neural activation in the anterior part of the medial frontal gyrus only in the mismatch condition, which required pragmatic enrichment; thus, it showed that the neural effect is not limited to negativity with LTM.

Further evidence for cognitive cost on the neural level comes from research using fMRI to investigate brain regions activated while processing statements including different types of quantifiers (e.g. *some*, *at least three*, *many*), but without specifically zeroing in on scalar implicatures, which has found that areas associated with EF, and especially with WM, were highly activated. For instance, McMillan et al. (2005) compared differences in neural activation among healthy adult participants processing two different types of quantifiers, either *first-order quantifiers* such as *at least*, *all*, *some* or *higher-order quantifiers* such as *most*, *more*, *even*, *odd*, during a truth value (binary) judgement task. (Participants viewed different proportions of items and were asked to decide if a statement appearing on-screen accurately described the scene). McMillan et al. found that although all quantifiers activated the inferior parietal cortex, associated with numeracy, only higher-order quantifiers activated the prefrontal cortex, associated with executive resources like WM. They attributed this finding to the higher cognitive cost required to process the higher-order quantifiers, since this processing would need to be carried out in two steps—for instance, while processing a statement like *at least half of the stars are red*, the hearer has first to assess what ‘half’ of the given number of items is and to hold this quantity in mind so that it can be compared to the actual number of items.

Other fMRI studies have involved classifying quantifiers in different ways. For instance, Troiani et al. (2009) hypothesised that there might be neural processing differences between numerical and (Aristotelian) logical quantifiers (e.g. *at least three of the Xs* v. *some of the Xs*). Using a similar paradigm to McMillan et al. (2005), they examined how adults process statements including either of these two types of quantifiers. Their findings revealed that quantifier comprehension activates two

dissociable neural networks: numerical quantifiers were processed (at least partially) in the lateral parietal-dorsolateral prefrontal network, involved with quantity-based and numerical processing, whereas logical quantifiers were processed in the rostral medial prefrontal-posterior cingulate network, which areas plays a crucial role in elementary logic, supported by the posterior cingulate cortex, which is related to the WM network. They suggest that such results highlight the substantial involvement of abstract number knowledge in interpretation of the meaning of numerical quantifiers in semantic memory and the possible contribution of logic-based evaluation in the interpretation of logical quantifiers. Troiani et al. thus give us evidence for the role of WM in the comprehension process of statements including logical quantifiers such as *some*.

Other research has explored quantifiers from the perspective of their proportional nature. For example, a recent paper conducted by Heim and colleagues (2016) investigated the neural basis of flexible adaptations conducted in quantifier semantics, for example in the interpretation of *few* v. *many*. Heim et al. (2016) used a binary judgement task in which adult participants viewed pictures each including a total of 50 items—blue and yellow circles presented in different proportions in each stimulus (in deciles from 20% to 70%). Each picture was presented alongside a written sentence, either *Many of the circles are yellow* or *Few of the circles are yellow*, and the participants were asked to evaluate the sentence by pressing a key for ‘Yes’ or ‘No’. The neural data revealed that processing of quantifiers seems to involve activation of quantity knowledge (in the parietal lobe) as well as activation of areas relevant to decision-making (in the dorsolateral prefrontal regions).

Politzer-Ahles and Gwilliams (2015) highlight some limitations of fMRI relevant to poor temporal resolution that could possibly affect accuracy in determining at which point during processing of a sentence with a scalar term the neural effect was provoked. To address these limitations, they employed techniques with high temporal resolution, such as magnetoencephalography (MEG), and adopted Breheny et al.’s (2006) paradigm to explore neural activation in participants processing scalar inferences within a context. The results revealed activation in the prefrontal cortex, which has been consistently linked to a variety of executive functions, such as WM, attention, and cognitive control (Shetreet et al., 2014).

Three important conclusions can be derived from the above-mentioned studies considered all together. First, different quantifiers have different cognitive processing costs depending on their cognitive complexity, a fact which might lead us to suspect that their associated neural activation level(s), cognitive effort, and nature of derivation of different types of implicatures might also differ. Secondly, cognitive resources, especially WM, evidently play a significant role in processing utterances including quantifiers. Thirdly, the neural effect occurred not only when the utterance contradicted information in LTM information (e.g. Nieuwland et al, 2010; Zhao et al., 2015), but also when processing did not require heavy activation of LTM, that is, when computation relied heavily on WM to compare a given statement with its visual context (e.g. Shetreet et al., 2014).

To conclude the review on scalar implicature processing, although several researchers have attempted to explore the nature of implicature processing and interpret their results in light of the two psychological hypotheses on the computation process (Default and Relevance/Context-Dependent), neither the empirical work nor the theories themselves have provided a clear description of the specific nature of the possible cognitive effort involved. The review has attempted to come to grips with this situation and to understand the nature of this processing by drawing on empirical findings from several domains. Review of these findings has shown that most of the empirical findings on implicature processing suggest that the process requires some cognitive effort, which gives (non-conclusive) support to the Relevance account over the Default. More specifically, there is fairly solid evidence of the relationship between WM (and possibly intelligence) and the ability to make pragmatic interpretations, a fact which might, further, strongly suggest that inference computation also involves EF abilities, not only contextual support. Furthermore, the findings showed that such cognitive effort is not necessarily detected based solely on findings from the surface level (through e.g. RT or rate of pragmatic responses) but that there is also potential to detect and to some degree measure it on the neural level; the existence of such results should of course make researchers very tentative about drawing evidence from results obtained on the surface level alone.

Although neuroimaging techniques might not be available in all laboratories due to their high cost, for better and more accurate understanding of the computation process researchers should rely neither only on one single task, even if it is highly controlled, nor only on pragmatic-linguistic measures, since we are not merely interested in finding out whether inferences are computed quickly or not, but more importantly, how and why they are processed quickly or slowly, and in particular how much effort is involved and what factors affect it. It seems that the most appropriate way to proceed is to combine pragmatic and cognitive measures to give a clearer image of implicature processing effort.

2.3.3. Acquisition of pragmatics

Several empirical studies have explored children's ability to comprehend implicated meaning by exploring their comprehension of, or their sensitivity to the violation of, Gricean conversational maxims. This section reviews the findings of some studies that have explored scalar implicature, a specific type of quantity implicature.

2.3.3.1 Children's comprehension of scalar implicature

This section focuses on the findings of empirical research on typically developing children's ability to derive implicatures, comparing the children's performance with that of adults, who acted as a baseline to illuminate the children's developmental endpoint.

Most previous work in this area has focused on implicatures generated from Horn's lexical scale and has used a binary-judgement paradigm, in which participants were asked to judge different stimuli and give responses such as 'True'/'False' or 'Agree'/'Disagree'. For example, Noveck (2001) explored 5-, 7-, 8-, and 11-year-old French children's ability to derive scalar implicatures, specifically, to reject stimuli containing the weaker term in a scale when the stronger is more appropriate. Within enriched contexts (visual scenarios), Noveck tested children's ability to reject an utterance such as *X might be in the box* where the more informative utterance *X must be in the box* was true. He also tested children's ability to reject under-informative statements given without context, such as *some elephants have trunks*. Noveck's results showed that children tend to make more logical interpretations than adults,

accepting under-informative statements in both conditions significantly more than the adults did (the adults rejected the under-informative items at high rates, approximately 70% and above).

Further investigation has tried to find out if children's pragmatic insensitivity to under-informative stimuli, as found in Noveck and other developmental studies, could be attributed to limited awareness on the children's part of the goal of the task rather than pragmatic delay. Papafragou and Musolino (2003) explored the ability of 5-year-old Greek children to make pragmatic interpretations of sentences in which the weaker term in a scale was used in a context. For example, they asked children to watch a scenario (acted out with toys) in which three out of three horses jumped over the fence, one by one. Then, they asked a puppet to describe the scenario; the puppet said that *some of the horses jumped over the fence*, and the children were asked if the puppet had answered well or not. Their results revealed that the children were insensitive to the violation of informativeness, and only rejected 12.5% of the under-informative items, while the adults rejected 92.5% of these items. After giving children explicit training to make them more aware that they had to judge felicity rather than truth, the children's rate of rejection of the puppet's infelicitous utterances increased to reach 52.5%, and on this basis Papafragou and Musolino (2003) suggested that children's pragmatic insensitivity to informativeness could be due to unawareness of the task goal.

However, although training effect has been found to have a positive effect on children's pragmatic competence in other studies as well, such effects have been seen to be only temporarily. For instance, Guasti et al. (2005), testing 7-year-old children in a task adopted from Noveck's (2001) '*some cats have tails*' experiment, found dramatic improvements when the children completed the task after a training session (although they did not reach adult level); however, when they repeated the same task a week afterward (with different stimuli), the improvement in pragmatic ability to make inferences had entirely disappeared. In addition, Guasti and colleagues found a strong effect of context: when children were given sufficient context (e.g., watching short videos and then hearing an under-informative utterance describing what had happened) they showed semi-adult-like pragmatic performance (rate of rejection with enriched context: children 75%, adults 83%). When they compared the children's

performance between the two context conditions (enriched context v. no context), the authors found a huge difference (rate of rejection with no context: children 11%, adults 53%). Guasti et al. (2005) argue that children at the age of 7 years are able to make consistently adult-like pragmatic interpretations when the context is sufficient, and conclude that their participants' failure in the no-context condition should be attributed to pragmatic limitations (lack of conversational background) rather than a pragmatic delay. These results are compatible with Feeney and Scafton (2004), who found strong evidence that 7-and-8-year-old English children also performed significantly better in the enriched context than in the no-context.

To learn when exactly children start to acquire the ability to derive implicatures, a number of studies have explored this question among children of different ages. For instance, Pouscoulous et al. (2007) explored sensitivity to the violation of informativeness among 4-, 5-, and 7-year-old French children and French adults, using the act-out paradigm. They presented their participants with four boxes containing small items (e.g. plastic animals). In one of the scenario, where all boxes contained a token (e.g. a turtle), the experimenter said: *I would like some boxes to contain a token*. If the participants interpreted the utterance logically, they would be expected not to change anything, but if they interpreted it pragmatically (as 'some but not all') they would be expected to remove one or more tokens. Pouscoulous and colleagues found that only 32% of 4-year-old children made logical interpretations, that is, did not act on the tokens in the under-informative condition, with a significant effect of age (that is, 5-year-olds, 7-year-olds, and adults were capable to produce such implicatures regularly). Similar results were found in Hendriks et al.'s (2009) large-scale study, which included Dutch child participants (ages 5–9 or 10–14) adolescents (ages 15–19), and adults (ages 20–64). They conducted an unsupervised binary-judgement task (experiment 3) to explore their participants' ability to derive scalar implicature <*some, all*> when sufficient visual contextual conditions are evident. Their results confirm the findings of Pouscoulous et al. (2007), as they found no significant differences between adults and children and also did not find any developmental trend between the children (that is, 5-year-olds did not generate more logical interpretations than 9-year-olds).

Papafragou and Tantalou (2004) explored younger (5 years old) Greek children's

ability to derive implicatures resulting from the ‘three scales’: lexical $\langle \textit{some}, \textit{all} \rangle$, ad hoc $\langle \{A \text{ and } B\}, \{A\}, \{B\} \rangle$ and encyclopaedic scales $\langle \textit{house}, \textit{roof} \rangle$; all their stimuli were given within enriched (visual) context. However, instead of asking adults and children to evaluate utterances with binary responses (‘Agree’/‘Disagree’, ‘True’/‘False’), as previous work had done, they asked them to reward the fictional character for the quality of its description (or, if the puppet provided an inappropriate (under-informative) description, the participant should withhold a reward) and to justify their response. For all the under-informative items, the children withheld the reward at a high rate (70% and above), justifying this by saying that the puppet did not complete the task. Although no significant difference was found among children’s rates of penalising under-informative items across the three scales, the children did perform better numerically with the ad hoc scale (90%) than with the lexical scale (77%) or the encyclopaedic scale (70%). The children’s justifications for withholding the reward from the puppet in the lexical scale (e.g. *he did not do all the Xs*) and encyclopaedic scales (e.g. *he only did the Y*) indicate that they indeed derived the scalar implicature in these conditions; in the ad hoc scale, however, as highlighted by Katsos and Cummins (2012), it was not clear whether the children withheld the reward because the puppet did not complete the task or because his response was under-informative.

Building on Papafragou and Tantalou’s (2004) study, Katsos and Bishop (2011) explored children’s ability to derive inferences generated from lexical and ad hoc scales. They asked 5-to-6-year-old English-speaking children and adults to watch short videos on Microsoft PowerPoint (using animation and pre-recorded utterances); after each video, a fictional character (‘Mr Caveman’), who was introduced as a non-native but fluent speaker of English, was asked to describe what happened. This task was first performed using a binary-judgement paradigm; that is, the participants were asked to evaluate the fictional character’s response as ‘Wrong’ or ‘Right’ and to justify their responses. Then, the same task was conducted with new child and adult participants but employing a ternary-judgement paradigm; that is, the participants were asked to reward Mr Caveman using a three-point-scale (huge, big, and small strawberries).

The results of Katsos and Bishop’s (2011) binary-judgement task revealed that the

adults rejected the under-informative items explicitly (giving a response of ‘Wrong’) and implicitly (e.g., ‘that is correct but’, ‘I do not know’, ‘half right half wrong’) at a high rate for the lexical and ad hoc scales (88% and 64% respectively) and significantly more than the children (26% and 31% for scalar (i.e. lexical scale) and non-scalar (i.e. ad hoc scale) expressions respectively). On the other hand, the results of the ternary-judgement task showed that the significant difference between the adults and children found in the binary task disappeared: the children were able to penalise the under-informative items in the two types of scale (with the big strawberry) at high rates (89% and 85% for scalar and non-scalar expressions respectively). The main contribution of Katsos and Bishop’s work was that by employing the ternary-judgement paradigm they were able to prove that children even at the age of 5 were able to detect violation of informativeness. Nevertheless, as Katsos and Bishop suggested, children, due to their limited language experience and cognitive flexibility, might not be able to correct their interlocutor when asked to do so as effectively as adults (as in the binary-judgement task) despite being sensitive to the pragmatic violation.

Other studies have explored how pre-schoolers might interpret the disjunctive operator *or*, that is, whether in a pragmatic task they apply inclusive reading (*A or B and possibly both*) or exclusive interpretation (*A or B but not both*). For instance, Singh, Wexler, Astle, Kamawar and Fox (2013) used a binary truth value judgement task in which 3-to-6-year-old children judged under-informative statements (with the disjunctive *or*) within enriched contexts. For each of the task stimuli, the children were presented with a picture depicting a boy holding some items (e.g. a banana and an apple), and were told that a puppet was going to describe the picture and that their own task would be to say if the puppet was right or wrong. For the stimuli assessing scalar implicature, there were two conditions: in condition 1, ‘One’, the boy held one item (e.g. a banana), whereas in Condition 2, ‘Both’, the boy held two items (e.g. a banana and an apple); and in both conditions the puppet’s description for each picture was (e.g.) *the boy is holding a banana or an apple* (Singh et al., 2013).

The results of Singh et al. (2013) showed that children’s performance differed significantly between the two conditions: they were more likely to judge the utterance *the boy is holding a banana or an apple* as correct when the boy was holding two

items. Interestingly, when the children's performance on the 'Both' condition was compared with that of adults, there were no significant differences, and as a result the authors assumed that the adults in their sample (26 participants) also did not compute scalar implicatures. These results, however, differ somewhat from those of Chierchia, Crain, Guasti, Gualmini and Meroni (2001), who found that pre-schoolers were able to make exclusive pragmatic interpretations of *or* (i.e. they derived the scalar implicatures) at a rate of 50%, while the adults showed a ceiling on the task by deriving implicatures at a rate of 100% (Chierchia et al. included 11 adults, and their results did not include any statistical comparisons between the groups).

To sum up, the previous results suggest that children at the age of 5 years old can show adult-like ability in deriving scalar implicatures generated when a scale associated with logical quantifiers, such as *some* and the conjunction *and* (that is, an ad hoc scale) are used in a context, but there are inconsistent results on whether children and even adults have the same pragmatic ability with under-informative utterances using the disjunction *or*. It should also be noticed that all the studies reviewed in this section involved monolingual children; there is some evidence that bilingualism might enhance children's pragmatic ability to derive inferences (e.g. Siegal et al., 2007, 2009).

In section 2.4, previous findings on bilingual children's pragmatic ability are reviewed in detail.

2.3.4 A summary of pragmatic competence

The review in this section summarises the main concepts of three pragmatic theories upon which this research is partly built: Grice's theory of implicature, and the two theories underpinning the respective processing assumptions being investigated—the Default and Relevance theories. Next, the review highlights some of the most important findings of the experimental research on implicature processing on both the surface (behavioural) and the neural levels. Because the main focus of the current research is scalar implicature and how and when children acquire the ability to derive it, empirical findings on children's acquisition of scalar implicature have been reviewed with reference to adult results, since adults represent a baseline for

understanding children's developmental end-state. In the following section, I explore how bilingualism specifically might modify children's pragmatic ability, and try to explain why such a modification might occur with reference to the two theories of implicature processing.

2.4 The impact of bilingualism on pragmatic and cognitive abilities

This section reviews empirical studies on the cognitive and pragmatic performance of bilingual children and adults and attempts to define the possible relation between these independent domains. The section starts by explaining the nature of bilingualism's impact on different cognitive abilities and pragmatic ability and why and how it might emerge. Then, the empirical evidence for this impact is reviewed. After this, empirical findings employing the Default and Relevance paradigms respectively are employed to understand theoretically how the two domains are associated, and interpreted in the light of implicature processing theories (i.e. Default (Levinson, 2000) v. Relevance (Sperber & Wilson, 1986/1995) theories).

2.4.1 Bilingualism and cognitive development

Based on the assumption that bilingualism places increased demands on certain aspects of cognition, a growing amount of research has explored whether bilingual ability has any cognitive repercussions for the development of the executive functioning (EF) system, a set of processes positioned in the frontal lobes (Schroeder & Marian, 2016; Stuss, 2011). A commonly used framework in this area, proposed by Miyake et al. (2000), is composed of three core components: *inhibition* (the ability to ignore irrelevant information), *shifting* (the ability to switch between tasks), and *working memory* or WM (the ability to keep information active in the brain for a while and recall or manipulate it) (Miyake et al., 2000).

Before going on to review the empirical findings on bilingualism's impact on EF abilities, let us briefly consider theoretically how and why bilingualism would be expected to affect EF. It is claimed that the bilingual advantage essentially originates from the continuous need for bilinguals to control both of their language systems simultaneously (Costa, Hernandez, & Sebastian-Galles, 2008; Videsott, Rosa, Wiater,

Franceschini, & Abutalebi, 2012), since in order to communicate successfully in one language, bilinguals have to avoid employing certain lexical terms or syntactic structures from another, a need that might increase their inhibition ability. There is evidence that this successful continuous resolution of language conflict allow bilinguals to develop brain structures associated with EF, such as the anterior cingulate cortex (Abutalebi et al., 2011) and the left caudate nucleus (Zou, Ding, Abutalebi, Shu, & Peng, 2011). It is also suggested that bilingualism is likely to be related to verbal intelligence, as measured by vocabulary tests, and may also contribute to EF, since shifting between two different languages necessitates a shift in attention (Costa et al., 2008).

2.4.2 Empirical findings of an effect of bilingualism on executive function

Although there is increased evidence that the period of early childhood to young adulthood is one in which EF processes develop at a relatively rapid rate, as measured by several non-linguistic EF assessments (e.g. Craik & Bialystok, 2006; Anderson & Reidy, 2012; Bialystok et al., 2009), such findings should not be taken to securely establish this principle, due to replication failures in a considerable number of studies. This section reviews some recent empirical findings of studies investigating bilingual children and focusing basically on the EF components that are measured in the current study (inhibition and WM); then, it briefly reports on the scant evidence of a bilingual advantage also found in adult participants.

We begin with inhibition. The inhibitory advantage that is seen in 4- and 5-year-old bilingual v. monolingual children's results on a Simon task (Martin-Rhee & Bialystok, 2008) was also confirmed in a recent study on 8-year-old multilingual v. bidialectal children (Antoniou et al., 2014). The Simon task is a computerised task designed to measure participants' RT to inhibit interference from conflicting stimuli (in an incongruent condition, in which stimuli appeared on the opposite side of the key to be pressed). This conflict is absent in the congruent condition (stimuli appeared on the same side as the button to be pressed) leading to faster completion of the trial. The availability of inhibitory advantage is usually measured by calculating the difference in RT between incongruent and congruent trials (called the *Simon effect* in the Simon task). It is worth noting that Martin-Rhee and Bialystok (2008) only found

a bilingual advantage in global RT (the time taken to complete the whole task), with bilinguals faster in both conditions, and no significant difference in Simon effect. However, Antoniou et al. (2014) found a bilingual advantage in both global RT and Simon effect. The bilingual inhibitory advantage has also been confirmed in tasks that measure inhibition through rate of correct trials rather than RT (e.g. Bialystok & Martin, 2004; Carlson & Meltzoff, 2008; Esposito, Baker-Ward, & Mueller, 2013). For example, children viewed two items that differed in shape and colour (e.g. circles and squares in either red or blue). In the congruent condition, the stimuli matched one or the other of the buttons they were presented with in both colour and shape, while in the incongruent, the stimuli matched one button in shape but the other in colour. Participants were instructed to match the shapes, and thus needed to inhibit the distraction from colour matching in the incongruent condition. However, in other studies, such a bilingual inhibitory advantage was found neither in tasks that depend on rate of inhibition (in children aged from 3 to 6 years old; Goldman, Negen, & Sarnecka, 2014) nor in those depending on RT (the Simon task, in 6-to-7-year-old children; Morton & Harper, 2007).

The evidence of bilingualism's positive effect on children's WM ability also seems to be unstable. That is, paradoxical results have been found in studies that measure different aspects of WM using tasks employing visual or auditory stimuli (e.g. numbers, words, shapes, or sentences), which either require children to recall stimuli in the same order (relying on STM) or to manipulate the stimuli (e.g. to recall them in reverse order). For example, Morales et al., (2013) found a bilingual WM advantage in 5-to-7-year-old children, but Blom, Küntay, Messer, Verhagen and Leseman (2014) found a bilingual WM advantage only in their 6-year-old sample and not in 5-year-old children. Although the WM advantage in bilinguals was also found in young children and adolescents (6 to 18 years old) with epilepsy (a central nervous system syndrome that has negative effects on EF) (Veenstra et al., 2016), it was absent in studies with typically developing children aged 8 years (Antoniou et al., 2014) and between 8–12 years old (Soliman, 2014; Gangopadhyay, Davidson, Weismer, & Kaushanskaya, 2016), respectively.

The inconsistent evidence on inhibitory advantage is not limited to bilingual children, but also extends to bilingual adults. Not all studies involving bilingual adults have

found that bilingualism clearly affects EF on the surface level (as measured by RT or accuracy). For instance, in an exploration of EF performance on the Simon task (to assess inhibition) in four age groups—5-year old children, 20-year old undergraduates, and middle-aged (30–59) and older (60–80) adults—Bialystok, Martin, and Viswanathan (2005) found a bilingual advantage in terms of RTs in all except the 20-year old participants. Their study suggests that as young adulthood is the age at which EF is most efficient EF, bilingualism provides no further advantages at that age, while as these EF processes decline in middle-aged and older adults, their bilingual advantage (re-)emerges. Further support for the presence of cognitive advantages in bilingual adults comes from Bialystok, Craik, Klein, and Viswanathan (2004), who tested the performance of younger and older bilingual and monolingual adults on the Simon task; the bilingual adults significantly outperformed their counterparts. However, recent studies testing bilingual and monolingual adults' performance in similar inhibition and attention tasks failed to replicate findings of bilingual advantage (Paap & Greenberg, 2013; Kousaie & Phillips, 2012a, b).

Thus, it seems that it is still unclear why bilingual advantage has emerged in some studies but not others. The next section explores possible reasons for such inconsistent results, in order, it is hoped, to give us a better understanding of bilingualism's impact.

2.4.3 Understanding the reason for the scantiness of evidence of a bilingualism effect

To understand possible reasons for the widely found absence of a bilingualism effect, let us first recall Miyake et al.'s (2000) framework, upon which most research in the area has been built. They proposed that EF is composed of three core components—inhibition, shifting, and WM; and most if not all of the previous work have used tasks that measure each component separately. This approach might have some inadequacies, however. First is the difficulty of separating these core components empirically and (also) relating them to the complex performance of real-life functions; experimental tasks can hardly produce pure measures of these components (Bialystok, 2011). Second, although these components can be considered to constitute core EF processes, Miyake et al.'s approach has been criticised for excluding processes that

are usually regarded as 'executive', such as conceptual reasoning, organisation, and planning capacities (Anderson & Reidy, 2012). In the case where one is exploring bilingualism's role in the development of EF, the situation would become rather complicated (Bialystok, 2011; Bialystok, Craik, Green, & Gollan, 2009) due to the difficulty of accurate assessment of proficiency level in each language, how extensive the practice of both languages is and has been, age of first- and second-language acquisition, and perhaps other cultural and educational factors which can hardly be controlled (Paap & Greenberg, 2013).

If we assume, hypothetically, that the reason for the lack of evidence of a bilingual influence on EF is the use of tasks that try to limit the involvement of the other components, then one should expect different results using a task that permits the components to work together to perform a coordinating function? A few studies have employed tasks that required the coordination of the three components, as well as having measured single components. The empirical findings of these studies revealed that indeed, a bilingual EF advantage was clearly detected in the coordinating task, but did not appear consistently in tasks measuring any of the single components. For instance, Bialystok, Craik, & Ruocco (2006) tested bilingual and monolingual adult participants' performance in two tasks (visual and auditory, respectively) each requiring participants to employ a single EF component (switching), and then tested their performance when these single tasks were combined into one complex task requiring a variety of EF components (inhibition, switching, and WM). To explain further, during the single tasks, the participants were asked to perform two categorisation tasks in different modalities (one visual and one auditory). The categorisation tasks respectively involved deciding whether stimuli were letters or numbers (LN) and whether stimuli were animals or musical instruments (AM). In the visual task, the participants viewed a single item (AM or LN) and were asked to categorise it by pressing the button on the appropriate side of the mouse. In the auditory task, the participants heard pre-recorded letters or numbers (LN) or certain musical instruments or animal calls (AM) and had to categorise them verbally. Each task had a duration of 60 s, after which the total number of correct classifications was counted. In the dual-task condition, participants were requested to classify simultaneously presented visual and auditory stimuli into one of the two given categories; the stimuli could be either 'related' (that is, either both LN or both AM) or

‘unrelated’ (that is, LN in one modality and AM in the other). The unrelated condition required monitoring and switching between LN and AM decisions, making it likely that it would be the more difficult condition. The results showed that although both language groups provided fewer correct classifications under the unrelated condition, there was an overall bilingual advantage only in the dual-task condition.

Would the adult bilingual advantage found in this coordinating task extend to bilingual children? The empirical findings of Bialystok (2011) for 8-year-old bilingual and monolingual children give further support to the assertion that bilingual performance cannot be attributed to a single component of EF but instead requires coordination of several components. Bialystok (2011) investigated the influence of bilingualism on children's performance on a complex task demanding coordination of various components—similar to, but slightly simpler than, the dual-task condition in Bialystok et al. (2006). Children were asked to classify visual and auditory stimuli into one of two categories: animals or musical instruments. The experiment started with a single task block for either a visual or auditory modality, followed by a dual-modality task block in which the visual and auditory stimuli were presented together. In all tests, RT and accuracy were recorded. Bialystok (2011) indicates that the dual-modality task allows children to employ all three core components of the EF system: WM, by holding rules in mind for classification purposes for both modalities; inhibition, by attending to categorisation of the target response, disregarding the irrelevant modality; and shifting, which refers to when children shift attention across stimuli so both responses can be accomplished. The results revealed that within a single modality, there was no difference in accuracy or speed between monolingual and bilingual children; however, in a dual-modality task, bilinguals were significantly better in terms of accuracy, and although they completed the task faster, the difference in RT did not reach statistical significance (Bialystok, 2011).

A recent study explored two core components of EF (WM and inhibition) in monolingual, bidialectal, and multilingual children (aged 6 to 9 years old) and found a multilingual/bidialectal advantage over monolinguals across the EF system (using a composite for of EF tasks) but not for any specific EF component (Antoniou, Grohmann, Kambanaros, & Katsos, 2016). They suggest that their results give further support to Bialystok’s EF coordination account, according to which the bilingual

advantage in EF performance is attributed to an enhanced general EF system, without a single component playing a conclusive role. However, one might suggest that aspects of their evidence could be interpreted differently to the way they have done it, since unlike Bialystok (2011), Antoniou et al. (2016) tested these components separately and then used a composite score in their analysis, whereas Bialystok's coordination account is predicated basically on the idea that EF advantage can be seen only when a task requires participants to employ various EF components simultaneously.

To sum up, the results of previous work suggest that bilingualism is likely to have an effect on cognitive ability as represented by EF processes; however, investigating the impact of bilingualism on individual processes in isolation might not provide an accurate picture, since these processes are highly integrated. This is, of course, not to say that bilingualism has not been found to impact certain individual EF component; clear evidence exists that inhibition (Blumenfeld & Marian, 2011), WM (Morales et al., 2013), and attention flexibility (Videsott et al., 2012) are influenced by bilingualism.

The next section considers whether this cognitive effect might also be recognised in pragmatic performance (regardless of age) and if so, how this relationship can be explained theoretically.

2.4.4 Bilingualism and pragmatic competence

After reviewing research addressing the cognitive performance of bilinguals and demonstrating that bilingualism seems to have a considerable effect on *cognitive* abilities, as shown by the superior performance of bilinguals on EF tasks, this study now asks (a) whether there is a bilingual *pragmatic* advantage, specifically a superior pragmatic ability to derive implicatures; and if so, (b) how can the pragmatic advantage of bilinguals (either children or adults) be explained in terms of their EF abilities? This section attempts to answer these two questions by reviewing empirical studies addressing the possible bilingualism effect on children's and adults' performance on pragmatic tasks and discussing whether the findings can be interpreted in terms of bilingual cognitive advantage.

2.4.4.1 Is there a bilingual pragmatic advantage?

Despite the established evidence for a positive effect of bilingualism on EF, only a few studies attempt to investigate if this advantage might extend to the pragmatic performance of bilingual children. One example is Siegal et al. (2007), which investigates the potential relation between bilingual cognitive advantage and children's ability to derive scalar implicatures by assessing performance on several cognitive and pragmatic tasks. The study involves child participants aged 4 to 6 years old—English monolinguals, Japanese monolinguals, and English–Japanese bilinguals. To control for language proficiency as a potential confounding variable, the study applied two tests: the British Picture Vocabulary Scale (Dunn, Dunn, Whetton, & Burley, 1997) for the English participants and the Japanese Picture Vocabulary Test (Kaiga goi hattatu kensa; Ueno, Nadio, & Iinaga, 1991) for the Japanese children. Bilinguals were assessed by both tests; however, no further language assessment was involved, leading to the question of whether these receptive vocabulary tests reflect the children's real level of language proficiency, which is not only a matter of vocabulary size (or other language skills such grammatical competence, which although they might also serve as an adequate proxy for language proficiency, might still be considered incomplete measures), but also length, intensity and quality of exposure to a certain language.

To measure participants' pragmatic competence, the study adapted a form of Papafragou and Musolino's (2003) some–all scalar implicature tasks. Children were introduced to a hand puppet that uttered statements related to a specific presented scenario and containing scalar terms; some of these statements were truthful but pragmatically inappropriate while others were truthful and pragmatically appropriate, and the children had to judge them by indicating whether the puppet answered well or not.

To explore whether bilinguals would have a cognitive advantage, which in turn would influence their pragmatic competence, the study used two EF tasks to assess two core components of EF. The first was the day/night task (Gerstadt, Hong, & Diamond, 1994), used to assess the core component inhibition. This task requires participants to

inhibit the normal classification process in order to permit opposite labelling of pictures, such as a picture of the moon labelled ‘day’ or of the sun labelled ‘night’. The second was a card sort task (an adapted version of the Wisconsin card-sorting task used by Woolfe, Want, & Siegal, 2002), used to assess switching ability. This task requires inhibition of previous labelling of stimuli by either colour or shape to allow their re-labelling on the other (substitute) dimension.

In contrast to other studies (e.g. Bialystok, 2011; Morales et al., 2013; Antoniou et al., 2014) that found a bilingual cognitive advantage, Siegal et al. (2007) found no significant difference between monolingual and bilingual performance on either of these tasks. In other words, both groups showed equivalent EF (specifically, in this case, for inhibition and switching). However, bilingual children’s performance on pragmatic tasks was significantly higher than that of monolinguals, as bilinguals were significantly more able to derive scalar implicatures by rejecting under-informative stimuli. In the attempt to explain such results, Siegal et al. suggested that the superior performance of bilinguals can be seen as evidence for what they call ‘compensation influence’, referring to a process by which bilinguals balance (compensate for) their lack of vocabulary in a language by being more competent than monolinguals in deriving conversational implicatures. That is to say, the bilinguals’ meaningfully limited vocabulary might lead them to be more alert to pragmatic traits of communication, which they might rely on more heavily to infer a speaker’s implicated meaning in their weaker language than monolinguals in that language.

A similar study (Siegal et al., 2009) was conducted to investigate whether bilingual children aged 3 to 6 years old would outperform their monolingual peers in their ability to derive implicatures generated from different types of conversational maxims (i.e. Maxims of Quantity I (do not be under-informative), Quality (be honest), Relevance (be relevant), and Manner (be clear) (Grice 1975, 1989). The sample involved three groups—Italian monolingual, Slovenian monolingual and Slovenian–Italian bilingual children—all classified as working class; their receptive vocabulary was assessed using the Italian version of the Peabody Picture Vocabulary Test (Dunn & Dunn, 2000) and another version of the test translated into Slovenian. EF abilities (inhibition and shifting) were assessed with similar measurements to those applied in Siegal et al. (2007).

To investigate children's sensitivity to the violation of Gricean maxims, they were administered a 'Conversational Violations Test' (CVT) (Siegal et al., 2009). In this test, using a laptop, the children watched short conversational exchanges among three doll speakers. In each conversation, one of the speakers asked the other two a question, and they each provided a brief response. One of their responses violated a conversational maxim while the other's did not. The children were asked to signal which doll had uttered something ridiculous or impolite. Below are two examples of the stimuli for the Gricean maxims of quantity and quality (Siegal et al., 2009, pp. 116–117).

(13)

(Sub)maxim of Quantity I

Question: *What did you get for your birthday?*

Answer: *A present.* (Alternative, appropriate answer: *A bicycle.*)

Maxim of Quality

Question: *Have you seen my dog?*

Answer: *Yes, he's in the sky.* (Alternative: *Yes, he's in the garden.*)

The outcomes demonstrate that bilingual children were significantly better than their monolingual counterparts in detecting violations of the Gricean maxims except for Quantity I, where there was no significant difference between the bilingual and monolingual groups. The authors suggested that this might be because of cultural factors promoting the children's acceptance of the under-informative answer (*A present*). Consistent with Siegal et al. (2007), this study likely indicates that the EF tasks used were not fit to purpose, since no differences in performance were found between bilingual and monolingual children. When replicating the CVT and EF tasks with pre-schoolers who had different language backgrounds (English–Japanese bilinguals versus Japanese monolinguals; German–Italian bilinguals versus Italian monolinguals), Siegal et al. (2010) had the same results.

Given this replication of the same finding of pragmatic advantage in the three studies of Siegal and colleagues, should one consider the superior pragmatic ability of bilinguals to be firmly established, and if so, how one can understand its potential

causes in the absence of EF advantage? Let us start with the pragmatic part. Katsos, Roqueta, Estevan, and Cummins (2011) raised some concerns about the stimuli used in the CVT and questioned its ability to genuinely assess children's pragmatic skills. For instance, they correctly noticed that for a child to reject a response such as *Yes, he's in the sky* when talking about a dog, s/he must draw on his/her own world-knowledge, and thus the rejection does not reflect real sensitivity to the violation of the quality maxim. In the attempt to explain such bilingual pragmatic outperformance, Siegal et al. (2007, 2009) suggested that the superior performance of bilinguals can be interpreted in terms of what they called 'compensation influence', described above.

With respect to the EF tasks, Siegal et al. (2009) drew attention to the possibility that other EF assessments might be better than the measures they used to detect bilingual EF advantages. Antoniou et al. (2014) added that the near-ceiling effect revealed in children's performance on the inhibition task (the day/night task) might be attributed to the nature of the task, which requires not the ability to suppress conflict but only sufficient response inhibition. This assumption was empirically confirmed in Esposito et al. (2013), which only found a bilingual inhibitory advantage in pre-schoolers on a task that required conflict resolution (the *bivalent shape* task (Mueller, 2010, 2011)) and not on a task merely requiring response suppression (the day/night task). The bivalent shape task presents circles and squares in either red or blue above buttons in the centre of the screen (a red circle and a blue square); congruent stimuli match one of the buttons in both colour and shape, and incongruent stimuli match one button in shape but the other in colour. Participants are directed to match the shape.

Antoniou and colleagues (2014) built on Siegal et al.'s (2009, 2010) studies in order to test 6-to-12-year-old Greek bidialectal (i.e. exposed to two varieties of Greek: Cypriot Greek and Standard Modern Greek) and multilingual (i.e. exposed to additional languages other than any varieties of Greek) children's pragmatic comprehension of relevance, manner, scalar implicatures, metaphor, and irony. They also assessed the core EF components (inhibition, shifting, and WM) using a battery of tests tailored to the children's age. For pragmatic assessment, they employed different measures. For example, to assess relevance comprehension, Antoniou et al. introduced children to a fictional character (a child of their age, 'George'); the children were told that would hear stories about George (supported with pictures

depicting the actions in the story). Then, at the end of each story, each child was asked to select one of two pictures that describe what happened in the story. For example, one of the stories depicted George asking his mother a question (e.g. *Mom, can I buy an ice-cream?*), and the mother's reply was either negative or positive answer with pragmatic enrichment (*You are ill* or *I have money in my wallet respectively*). Then, the experimenter introduced two pictures; each one suggested a possible ending to the story (one where George bought an ice-cream and one where he did not). The experimenter asked the child 'What happened at the end of the story?' One of the pictures showed an action that matched the mother's positive response (e.g. George eating an ice-cream), while the other matched the negative response (e.g. George doing something else such as playing with his toys). To assess children's comprehension of scalar implicatures based on these cases, Antoniou and colleagues used a task adapted from Pouscoulous et al. (2007), which has been described in sections 2.3.2.4 and 2.3.3.1.

The results of Antoniou et al. (2014) did not reveal strong evidence for a pragmatic advantage of multilingual over bidialectal children, although they did find strong evidence for a multilingual inhibitory advantage on the Simon task. The authors suggested various possible reasons for the absence of multilingual advantage, including the children's age, language profile, and the nature of the pragmatic assessment. Specifically, regarding assessment, Antoniou and colleagues indicated that, unlike Siegal et al.'s (2009, 2010) pragmatic task, which assessed *sensitivity to the violation* of the Gricean maxims, their task assessed children's *pragmatic comprehension* of the maxims. Thus, Antoniou et al. (2014, p.23) suggested that 'a bilingual advantage in detecting pragmatic violations does not extend to the ability to understand more complex pragmatic language like implicatures or, alternatively, that the effect of bilingualism on implicature comprehension is simply smaller and thus more difficult to detect'. They also suggested that the lack of clear pragmatic multilingual advantage might result from comparing multilinguals with bidialectals rather than 'pure' monolinguals, or that it might be attributed to the fact that the children who participated in their study were older (aged 6–12 years old) than those involved in Siegal et al. (2007, 2009, 2010) (aged 4–6 years old). Although they were not sure why this age gap would make a difference, other empirical work on children's acquisition of pragmatics does reveal that 7-year-old children show adult-

like ability in deriving inferences (e.g. Pouscoulous et al., 2007; Hendriks et al., 2009). Another possibility is that a multilingual pragmatic advantage might be detected in RT, especially given that multilinguals significantly outperformed bidialectals in the Simon task; that is to say, if the multilinguals completed the inhibition task faster than the bilinguals, it might be that their pragmatic advantage also comes in the form of faster RT rather than higher accuracy. However, this possibility cannot be examined since it is beyond the scope of Antoniou et al.'s results.

If we assume hypothetically that the children's pragmatic performance in Antoniou et al. (2014) resulted from both groups' having adult-like pragmatic competence, then one should also expect the same performance between bilingual and monolingual adults. Unfortunately, only a few studies have tested bilingualism's effect on adults' pragmatic ability. For example, Slabakova (2010) examines the pragmatic performance of advanced and intermediate Korean–English bilinguals (Korean learners of English, whose proficiency level in English was assessed by their TOEFL score) in two pragmatic tasks on scalar implicature (context v. no-context conditions). The study results revealed a bilingual pragmatic advantage in both conditions in both tasks. When there was no context, a small majority of advanced and intermediate Korean–English learners rejected the statement (at rates of 57% and 50% respectively), while native English- and Korean-speakers mostly accepted it (with rejection rates of 37% and 33% respectively). Similarly, when provided with adequate context, the bilingual adults gave more pragmatic responses (90% of their answers), while monolingual Korean- and English-speakers gave fewer (75% and 63%, respectively); these differences were statistically significant. Thus, these results might suggest that the effect of exposure to additional languages on pragmatic ability can be detected in individuals irrespective of their age or even their level of linguistic proficiency.

2.4.4.2 How can the superior pragmatic performance of bilinguals be explained in terms of their EF abilities?

To explain the potential effect of EF on pragmatic ability to derive inferences, I will elaborate an account built on four major findings, connected to theories of implicature

processing. These findings are, first, the superior pragmatic ability of bilingual children (Siegal et al., 2007, 2009, 2010) and adults (Slabakova, 2010), although Antoniou et al. (2014) found only a suggestive pragmatic advantage (as seen in their almost equal pragmatic performance to the bidialectals despite their much more limited vocabulary). Second is the bilingual advantage in various EF components (Blumenfeld & Marian, 2011; Morales et al., 2013; Videsott et al., 2012). Third is the association in monolingual adults between the rate of pragmatic response and WM (Feeney et al., 2004; De Neys & Schaeken, 2007; Dieussaert et al., 2011). Fourth is the empirical evidence that the bilingual experience results in not only distinctive brain morphology (e.g. in grey matter) but also distinctive neural activation when compared to monolinguals' brains (higher activation in language and EF areas). On the brain structure level, using morphometry techniques, Burgaleta et al. (2016) and Abutalebi et al. (2013) found increased volume in bilingual brain (grey matter) areas which are typically responsible for language processing (e.g. the *basal ganglia*, which triggers the computation of language rules such as mental grammar, and the *thalamus*, which is responsible for language production and lexical decision). On the brain function level, studies comparing bilinguals and monolinguals (using fMRI) reveal that the bilinguals show more activations of cortical areas related to language processing (Parker Jones et al., 2012; Hernandez & Meschyan, 2006) and to EF processes (e.g. switching and inhibition) (Abutalebi, 2008; Ma et al., 2014).

Theoretically, how might the differences in EF abilities, and even potentially the brain's function and structure, between bilinguals and monolinguals be associated with pragmatic performance? I will attempt to establish this relationship with reference to two theories concerning the processing of pragmatic inferences: the 'Default view' (Levinson, 2000) and Relevance theory (Sperber & Wilson, 1986/1995). Let us briefly recall the assumptions of the two theories: the Default account proposes that generalised conversational implicatures (GCIs) generated by default mechanisms are subject to cancellation if contextual assumptions require it, and that at that stage particularised conversational implicatures (PCIs) are computed. In Relevance theory, in contrast, inferences are computed by a process of derivation from context, which require some cognitive effort. If we hypothesise that GCIs are generated effortlessly, as in the Default account, then it might be reasonable to expect that monolinguals will perform better in their own language, since they will have

faster, easier access to lexical items stored in their semantic memory (which includes general knowledge, meanings, and understandings, irrespective of particular experiences (Tulving, 1972). In all previous studies that have shown a bilingual pragmatic advantage, bilingual children and adults had clearly more limited vocabulary size (Siegal et al., 2007, 2009, 2010) or less advanced language proficiency level (Slabakova, 2010) than their monolingual counterparts. Furthermore, studies that have measured semantic fluency in bilinguals and monolinguals, by asking participants to either act out or repeat an utterance loudly after hearing the stimulus, found a bilingual disadvantage both in children (Kormi-Nouri et al., 2008) and in adults (Gollan, Montoya, & Werner, 2002). Regardless of whether bilingual semantic memory might make the process of deriving inferences more effortful or at least slower, then, it seems that the Default view does not provide a sufficient explanation of the superior pragmatic performance of bilinguals. Default theory also might not open any door to understanding the potential relationship between bilingual cognitive advantage and pragmatic outperformance.

Then, does Relevance theory better interpret superior bilingual pragmatic performance? As noted above, the process of pragmatic implicature derivation is considered in the Relevance account to be a (conscious) cognitive process that requires attentional processing (Sperber & Wilson, 2002). If Zhao et al.'s (2014) result showing neural activation due to the process of deriving implicatures, and if the findings of Feeney et al. (2004), De Neys and Schaeken (2007) and Dieussaert et al. (2011) on the association between WM and pragmatic performance are taken into account, then Relevance theory might present a better description of the phenomenon, since it gives a plausible explanation, based on the assumption that the derivation process is cognitively effortful. We might understand this cognitive effort in terms of (a) processing cost (either associated with brain function (e.g. EF abilities) or brain structure/morphology (e.g. expanded grey matter or activation in areas associated with EF)), (b) rate of pragmatic responses (e.g. Slabakova, 2010), (c) processing speed (e.g. Noveck & Posada, 2003; Breheny et al., 2006), or possibly (d) context-sensitivity (e.g. Hartshorne & Snedeker, 2014). Since studies on bilingual children focused on measuring processing cost (i.e. EF) and rate of pragmatic responses, then two questions arise. First, how can we explain bilingual pragmatic outperformance when the EF advantage is absent, as in Siegal's studies? And in the opposite direction,

how can we explain the absence of bilingual pragmatic advantage when there is evidence for the EF advantage, as in Antoniou et al. (2014)? Apart from the above-discussed possible causes of the absence of either bilingual pragmatic or EF advantage, researchers should be aware that even when using the right tasks and controlling for all plausible confounding variables, the effect of bilingualism on both domains might not be always detected on the surface level (i.e. rate of pragmatic responses, speed, or context-sensitivity) but could still exist and be detectable on the neural level (as neural activation). Indeed, the findings of previous work give evidence for neural activation in the process of deriving implicatures (Zhao et al., 2014) and while completing tasks that require EF abilities (Abutalebi, 2008; Ma et al., 2014).

To sum up this discussion the pragmatic differences between bilinguals and monolinguals seem not to result from different processing techniques but rather from the bilinguals' enriched EF and expanded brain resources, which facilitate the deriving of implicatures.

2.4.5 Bilingualism, executive functioning (EF), and theory of mind

Another aspect that might help to explain pragmatic outperformance in children is *Theory of Mind* (ToM). Before going on to explain this, I would like to emphasise that this section does not aim to provide a review of recent empirical findings, but only to conduct academic due diligence and bring up some other factors that could be involved in the bilingual pragmatic advantage.

ToM, often used in research investigating the development of children's social cognition, is the ability to 'interpret other people's behaviours in terms of [the others'] internal mental states such as intentions, beliefs and desires' (Goetz, 2003, p. 1). That is, the 'theory' here is not a scholarly theory, but refers to any individual's own internal conceptual understanding of the mind. The standard task used in research that empirically explores children's understanding of others' mental states is the *false-belief task*, as introduced in the work of Wimmer and Perner (1983). In this task, a child is told a short story supported by the use of toys as props, and at the end, the child is asked what happens next. If the child is able to predict a false belief held by

the character in the story (that is, a belief which it makes sense for the character to hold within the context of the story but which differs from the belief that it is reasonable to hold from the child's position as interlocutor), then he or she passes the task and is taken to have a reasonably well developed ToM. For example, in the *unexpected transfer* false-belief task, the child is told that 'X', a character in the story, was playing in the garden, and that before leaving she put her toy in box A; then, while Sara was away, her mother (surreptitiously) moved the toy to box B. The experimenter asks the child where Sara should look for her toy. If the child chooses box A, this means he or she is able to differentiate Sara's belief, which although wrong is the reasonable one to expect Sara to hold if she has her own consciousness and ability to experience the constructed reality of the story, from what the child knows to be (real) reality. A similar task, the *unexpected content* false-belief task, was designed by Hogrefe, Wimmer, and Perner (1986); it differs in that the item itself rather than its location is changed in the character's absence. Other ToM measures are the *appearance–reality* task (measuring the ability to differentiate the appearance of an object from its real nature, i.e. a pen that looks like a small fish) and the *perspective-taking* task (assessing the ability to distinguish a speaker's physical orientation and directional perspective from that of a hearer) (see Goetz, 2003; Carlson & Moses, 2001).

A number of studies have suggested that bilingualism influences children's ToM development. For instance, comparing the performance of 3-to-4-year-old Chinese–English bilingual children with their monolingual English- and Chinese-speaking counterparts, Goetz (2003) found some evidence for a bilingual advantage on ToM tasks—bilinguals performed significantly better than monolinguals, and 4-year-old children performed significantly better than 3-year-olds.

This relationship between bilingualism and ToM has been hypothesised to be both *direct* and *indirect*. The direct influence of bilingualism on ToM development has been described in terms of two potential consequences of children's exposure to more than one language. The first, as Goetz (2003) indicates, might be that bilingualism improves children's *metalinguistic ability*, so that they can understand that a given object can be represented in more than one way linguistically. Such an understanding of (linguistic) *metarepresentation* would presumably also affect children's ToM

ability, helping them understand that an object can be represented differently by different people. The second direct consequence might be that bilingual children could become aware that particular kinds of language should be used with particular people in particular contexts. For example, a child might come to understand at an early age that he has to use a different language (e.g. English) at his nursery school than the one he uses at home (e.g. Japanese). This kind of sociolinguistic awareness, as Goetz (2003) has explained, helps the child understand that people can have different mental states than their own, an understanding that should improve their performance on ToM tasks. Further, as highlighted by Antoniou et al. (2014), metarepresentative ability is crucially important to children's pragmatic competence, as the process of implicature derivation requires children to be aware of speakers' mental states and intentions (Sperber & Wilson, 1995). The utility of this ability can also be seen in terms of the *compensatory hypothesis* in Siegel et al. (2007, 2009), in which bilingual children compensate for their less advanced vocabulary compared to their monolingual counterparts by improving their ToM understanding, and in this sense employ their advanced 'mind-reading' ability to derive more implicatures, more effectively than monolingual children.

The final, indirect relation between bilingualism and ToM that has been posited is related to EF. Although Goetz (2003) suggested the possibility that the bilingual children in her study might have better EF than monolinguals—more precisely better inhibitory control—which in turn should be reflected in better performance on ToM tasks, she did not investigate this claim empirically. However, such an empirical investigation is found in Moses and Carlson (2001), who investigate potential correlations between several inhibitory control and ToM tasks. The inhibitory control measures required either temporary inhibition of an appropriate response (as measured on a 'Delay scale', e.g. the 'Pinball task', which requires children to suppress an action until they receive a signal from the experimenter) or inhibition of an inappropriate response and activation of the conflicting appropriate one (the 'Conflict scale'), on a task such as the day/night task. Several of the tasks mentioned above (e.g., the false-belief task, the appearance-reality task) were applied by Moses and Carlson to measure the ToM development of 3- and 4-year-olds. Their findings indicated that inhibitory control strongly predicts children's ToM performance; however, this prediction was only highly significant on Conflict-scale tasks. They

explained these results by suggesting that Conflict-scale tasks require children to employ more and a wider range of EF capacities, such as WM, in order to activate the conflicted appropriate response and not merely inhibit the inappropriate one, while in the Delay tasks they only need to inhibit, not also activate, a response, and only for a short while. Thus, Moses and Carlson (2001) suggest, inhibitory control might strongly influence children's ToM, and WM might be critical for ToM reasoning.

2.4.6 Bilingualism, language proficiency, and socio-economic status (SES)

The last topic related to bilingualism's impact on cognitive abilities to be considered here concerns the extent to which language proficiency level and SES might modify this impact. The outcomes of a number of studies find positive relationships between language proficiency level in general and children's performance on certain cognitive tasks, in both monolinguals and multilinguals. For instance, the findings of Iluz-Cohen and Armon-Lotem (2012) indicate a fairly strong association between both English monolingual and Hebrew-English bilingual language proficiency, as assessed by using standardised tests of Hebrew and English ability (measuring different language abilities such as vocabulary, pronunciation, comprehension) and of key cognitive abilities (inhibition, sorting, and shifting). When children's results were compared to their performance on generic EF processes, significant correlation was found. Further evidence for the relationship between cognition and language competence comes from Videsott et al. (2012), which found a significant correlation between language competence and attention, one of the EF core components, as assessed by the Attentional Network Test (ANT). The ANT is a computer-based test that enables the exploration of the three chief components of the attentional process: *alerting*, which involves attaining and maintaining an alert state; *orienting*, which involves selecting information from sensory input; and *executive control*, which involves monitoring and resolving conflict (Posner & Petersen, 1990). The study used two criteria to assess 10-year old children's proficiency levels in Italian, Ladin, and German; the first was a self-evaluation form in questionnaire format, completed by the children and used to assess their most dominant language (DL1; that is, the language in which they had the most proficiency), while the second was an external evaluation by the children's teachers (namely, their class marks) used to assess the children's second-most-dominant language (DL2) and the language with the lowest

score (DL3). English was also incorporated, and was considered the least dominant language (DL4).

Another factor that might modify the effect of bilingualism on EF is SES. However, there is increased evidence that while there may be an effect of SES on EF, its effect is isolated and not integrated with the effect of bilingualism. For instance, De Abreu, Cruz-Santos, Tourinho, Martin, and Bialystok (2012) explored the effect of SES on bilingual and monolingual 8-year-old children's EF abilities. Their study included children from low-income families, specifically, from schools in low-income areas but that did not have severely disadvantaged neighbourhoods or difficulties with educational resources, and that had highly educated teachers. De Abreu and colleagues tested children's WM, abstract reasoning, selective attention, and interference suppression, and found a bilingual advantage in both selective attention and interference suppression, which led them to conclude that the effect of bilingualism can still emerge regardless of any SES disadvantage. However, since De Abreu et al.'s study included only children with low-income backgrounds, it might be asked: If there is an effect of SES, would one then expect it to be so powerful as to cancel out a bilingual advantage? This question was clarified in a recent study (Calvo & Bialystok, 2014) conducted with 6-year-old bilingual and monolingual children, where each group included two subgroups of children, respectively working-class and middle-class; classification was based on mother's education, which was found to be in line with the child's father's occupation and income. The study assessed two EF components (inhibition and WM), and found that both bilingualism and class contribute significantly and independently to children's EF advantage, each irrespective of the other.

2.4.7 A summary of bilingualism's impact

This section explores the empirical findings regarding bilingualism's impact on EF and pragmatic ability to derive implicatures in children and adults. Although it is still unclear why the bilingual advantage emerges in some studies and is absent in others, including highly controlled studies, the coordination assumption of Bialystok et al. (2006) and Bialystok (2011) may provide a plausible explanation. After all, even if researchers try to control for all plausible confounding variables, finding accurate

measures of bilingual experience represents a real challenge due to the difficulty of assessing how extensively a bilingual speaker uses and how often he or she switches between his or her two languages. In addition, relying on parental questionnaires to assess bilingual children may place in question the accuracy of the results.

2.5 The current research

On the basis of the literature review that has been presented in this chapter, this project conducts two empirical research studies. The first is based on the findings reviewed in section 2.2; it explores children's semantic comprehension of quantifiers and their numeracy skills in the attempt to answer two research questions: (a) Do bilingual and monolingual children comprehend the quantifiers 'most' and 'some' and the operators 'or' and 'and' in a semantically appropriate way? And (b) to what extent does the acquisition of a numerical system promote or (possibly) hinder the acquisition of quantifiers? The importance and originality of this investigation can be summarised in three points. First, it explores children's perception and production of quantifiers; to the best of the author's knowledge, previous work on children's comprehension of quantifiers explored their perception of them (in particular their ability to produce sets correctly), but not how they comprehend their proportional meaning, which a production task would reveal. Second, by exploring quantifiers along with approximate and exact numerical systems, we may find that the results can inform us which of these comes first in the acquisition process. Another contribution of this study is that it attempts to theoretically establish the relationship between numbers and quantifiers by mapping theories of abstract concept representation to theories of number word representation and to test this relation empirically. The final contribution we may expect relates to the fact that that semantic comprehension of quantifiers in Arabic has not been covered in previous research, and so the current study will shed light on findings from a new language and allow comparison between Arabic children's developmental trajectory and those of children from different backgrounds, and to draw some possibilities regarding why differences might occur.

The second study investigates the potential relation between pragmatics and bilingualism, and aims to answer one main question: Can any superior pragmatic competence in bilingual children be explained in terms of a cognitive advantage over

monolinguals? To explore this question empirically, the study employs methods used to assess children's cognitive abilities, STM, and inhibition, and whether such abilities can predict children's pragmatic performance in two pragmatic conditions (context versus no-context). The contribution of this second study is threefold. First, although there is a wealth of literature on children's acquisition of scalar implicature, this study will add a meaningful contribution by exploring the effects of STM and inhibition on children's pragmatic performance. Second, although there is some research exploring the effect of bilingualism on children's pragmatic abilities (e.g. Siegal et al., 2007; 2009; Antoniou et al., 2014), the existence and precise nature of a bilingual effect on pragmatic abilities are still unclear due either to the scant evidence regarding the role played by EF abilities (e.g. Siegal et al., 2007; 2009) or the absence of pragmatic advantage (Antoniou et al., 2014); in this sense, the current study should enhance our understanding of the effect of bilingualism on implicature and pragmatic abilities generally. Third, the study will attempt to interpret pragmatic performance in individuals who have been exposed to additional languages by connecting EF with implicature processing theories and to explain precisely how EF abilities affect pragmatic abilities by interpreting its results in relation to implicature processing theories.

2.6 Chapter summary

This chapter presents the main theoretical assumptions underpinning this study and reviews the methods, outcomes, and conclusions of empirical studies in different areas that will inform it. The first part has focused on the potential relation between number and quantifier words; theories of the mental representation of number and abstract concepts have been employed to clarify the nature and the direction of the relation. The second part of the chapter has focused on pragmatic theories and the process of acquisition and computation of scalar implicature. The third part has addressed the nature of bilingualism's effect on children's and adults' EF and pragmatic abilities, in the attempt to explain how EF and pragmatic abilities are related in terms of implicature processing theories. After the review of the concepts and findings that form the base of the present study, the last part of the chapter re-introduced the main questions that this research is exploring and highlighted briefly the importance of exploring such questions.

The next chapter will introduce the empirical methods that will be used to answer the questions raised in this chapter, and will present a clear justification for each task selected.

Chapter 3

Methodology

3.1 Introduction

This chapter gives a detailed description of the empirical methods that have been employed to answer the research questions: (a) Do bilingual and monolingual children comprehend the quantifiers ‘most’ and ‘some’ and the operators ‘or’ and ‘and’ in a semantically appropriate way? (b) To what extent does the acquisition of a numerical system promote or possibly hinder the acquisition of quantifiers? And (c) can any superior pragmatic competence in bilingual children compared to monolinguals be explained in terms of a cognitive advantage over monolinguals? The project can be divided into two main studies, conducted on the same participant sample. The first study aimed to explore the relationship between children’s semantic quantification and numeracy skills; that is, how they comprehend the target quantifiers and whether their level of acquisition of number words might impact their understanding of these quantifiers. The second study aimed to investigate the relationship between children’s pragmatic and cognitive skills; more precisely, it examined children’s ability to detect the violation of Grice’s Maxim of Informativeness, ‘to be as informative as required’ (Grice, 1975, 1989), and whether this ability is associated with certain cognitive skills, namely inhibition and STM.

The chapter starts with a brief review of the two main methodological approaches used in related research, and then highlights concisely the limitations and advantages of adopting empirical approaches in social science research. After this, it introduces the research design and the method. The method section describes the sample, the measures taken to control several potential confounding variables (language proficiency, language exposure, general mental ability, and SES), and the tasks that were used in each study, with clear justifications for the selection of these tasks and how they answer the research questions. Finally, a summary of the methods is given at the end of the chapter.

3.2 Philosophical approach to research

The two major philosophical approaches widely applied in social science research are *positivist* and *interpretive* (or *non-positivist*) approaches (Lee, 1991; Aliyu, Bello, Kasim, & Martin, 2014). Although it is not the aim of this chapter to describe the two approaches in detail, it might be useful to define them briefly before describing the current methodological approach since it was derived from both positivist and interpretive approaches.

We begin with the positivist approach. It is fundamentally based on the *ontological principle* (a philosophical principle concerned with the nature of being and the fundamental categories of reality (Neuman, 2014)), and can be described as a systematic and objective approach to acquiring knowledge (Guba & Lincoln, 1994). It assumes that reality is stable and independent of the observer's viewpoint, and that knowledge can only be acquired through empirical tools, such as those widely applied in natural science (Lee, 1991; Kura & Sulaiman, 2012; Aliyu et al., 2014). Positivist researchers usually adopt a deductive process of reasoning; they start their research with certain hypotheses or theories and then test them experimentally or empirically to see if they hold in specific instances, in a *theory-testing process* (Hyde, 2000). The main methods associated with the positivist approach are *quantitative methods*.

In contrast, the interpretive approach assumes that reality or truth is constructed by the observer's own subjective and intersubjective meaning-making as he or she interacts with the world around him or her (Orlikowski & Baroudi, 1991; Aliyu et al., 2014). Thus, 'interpretive researchers attempt to understand phenomena through accessing the meanings that participants assign to them [...] interpretive studies reject the possibility of an 'objective' or 'factual' account of events and situations, seeking instead a relativistic, albeit shared, understanding of phenomena' (Orlikowski & Baroudi, 1991, p. 5). Interpretive researchers usually adopt an inductive process of reasoning; they start with observations of specific occurrences, and seek to establish generalisations about the phenomenon under examination; this approach is called a *theory-building process* (Hyde, 2000). The main methods associated with the interpretive approach are *qualitative methods*.

The current research adopted both paths of reasoning (deductive and inductive); below, I justify philosophically how and why the two were adopted.

3.2.1 Justification for the study's methodological approach

Although this is an empirical study, I did not fully adopt the deductive process of reasoning while designing it; rather I adopted the *integrated framework* proposed by Lee (1991). This framework combined the positivist and interpretive approaches not in the sense of mixing qualitative and quantitative methods, but instead in that of adopting both deductive and inductive paths, allowing me to use the observations conducted in my pilot study (see below) alongside previous empirical findings and theoretical assumptions in the fields of language and cognitive development and then to develop my research design based on my own interpretations, which influenced my decisions when selecting experimental tasks. Thus, my role as a researcher did not stop at testing the hypotheses I started with (built on previous findings) in order to understand the cause–effect relationship, as the positivist approach suggests; neither was it limited to relying on my own subjective interpretations of the observations I made to select a specific aspect to be explored empirically. Instead, my integrated procedure involved both theory-testing and theory-building.

3.2.2 Empirical methods in social science research: Benefits and challenges

Although empirical methods allow researchers to explore different behavioural phenomena in an objective way, the empirical approach has also been criticised for the artificial nature of experiments conducted under it, which may affect the validity of research findings (Neuman, 2014). Some researchers have argued, however, that this artificiality should be seen as an advantage of experimental research rather than a disadvantage, since it permits ‘observation in a situation that has been designed and created by investigators rather than one that occurs in nature’ (Webster & Sell, 2007, p. 11). Another benefit of artificiality is that it allows the experimenter to control the study situation, including variables directly relevant to the research hypotheses and controlling for variables that could have an effect despite not being part of the hypotheses (Neuman, 2014). In the context of the current research, potential effects of several plausible confounding variables (including SES and language proficiency)

have been controlled for to understand clearly whether, for example, the differences between groups resulted from bilingualism or one of the confounding variables.

One of the main challenges in using the empirical approach in social science (in general, not only in child research) is related to the matter of simulation. In general, the creation of artificial conditions for research purposes should emulate real-world settings to the degree possible (Aliyu et al., 2014). Although great verisimilitude can be achieved in natural science, it is likely virtually impossible to meet this condition in the social sciences (Aliyu et al., 2014; Kura & Sulaiman, 2012). It has been claimed, however, that social scientists can overcome this obstacle by employing various kinds of measures, samples, analyses, and designs in order to reach a valid and reliable understanding of an occurrence (Aliyu et al., 2014; Neuman, 2014). This has been achieved in the current research by including several measures that might play a role in the acquisition process. For example, acquisition of quantifiers was tested through two tasks to enable better understanding of how children comprehend such terms. Not only this, but various tasks measuring numeracy skills were included, under the assumption that numbers might affect the acquisition of quantifiers.

Another challenge arises when experimental artificiality requires the manipulation of reality to understand the effect of a certain variable. Although this might be easily achieved in natural science, it represents a real challenge for social scientists, as it can give rise to certain ethical issues around behaviour manipulation (Neuman, 2014). For instance, if natural scientists want to study the cause–effect relation between human genes and a certain disease, they can manipulate genes in human cells in their lab to test their hypotheses without ethical concerns. In contrast, social scientists, whose study is precisely human society, cannot manipulate human life for the sake of gaining knowledge, but must use their creativity to find existing situations that can help in exploring their hypotheses (Neuman, 2014). For instance, if a researcher wants to measure the effect of not having breakfast on children’s performance in school, it will be unethical to ask parents to prevent their child from eating this meal; therefore, instead, the researcher might conduct her research in an area or a country known for severe nutritional deprivation; and compare her sample from that country with a sample taken from a country where children are expected to have better nutritional care. Of course, there are some situations where it would be much easier to

manipulate reality, for example, in the case of an experiment testing a particular educational intervention's effect on students' performance using an experimental and a control group.

Another issue related to the empirical approach is the reliability of results. This is usually assessed in terms of the ability to get similar results when replicating a study (Neuman, 2014). However, replication of research results in social science is often difficult to achieve, because studying human beings and their behaviours can be an extremely complicated process, and although the experimental approach might allow the researchers to control for a good number of variables, there will be always others that will be out of their control. In addition, the same participants might provide different answers if they are asked to re-do the same task, which would again change the research results. It might be suggested that including several tasks measuring the same phenomenon might help yield more robust findings. For example, the current research includes two pragmatic tasks, two semantic tasks, and additional tasks on the acquisition of number words; it is hoped that this multi-task design will help us better understand the child participants' developmental level. In addition, the partial replication of some previous work—since most of the tasks were adopted from previous empirical studies—will allow us to compare the findings of this research to that older work.

To conclude, it is often claimed that an empirical approach in early childhood research can provide more reliable and objective results than those obtained via the interpretive approach. However, due to the complexity of human behaviours and the various social, cultural, and pedagogical factors that can affect children's development, it is also often suggested that employing both approaches together can provide a better understanding of the phenomenon under investigation (Neuman, 2014; Kura & Sulaiman, 2012; Lee, 1991).

3.3 The research design

This project adopted a factorial experimental design, a design that taken into account the impact of several independent variables simultaneously (Neuman, 2014). Two studies are conducted in this project, using the same sample. Study 1 was designed to

explore children's semantic comprehension of quantifiers and the potential effect of acquisition of numerical system on children's semantic performance. To do so, it adopted an experimental design with 3 child groups (Arabic–English bilinguals, Arabic bidialectals, English monolinguals), 2 semantic tasks (perception v. production) and 4 number tasks (how-many, give-a-number, non-verbal ordinal, estimate-magnitudes), in addition to 2 adult control groups (Arabic, English). Study 2 aimed to identify the relationship between bilingualism, pragmatic competence and certain cognitive abilities. Its experimental design consisted of 3 child groups (Arabic–English bilinguals, Arabic bidialectals, English monolinguals), 2 pragmatic tasks (enriched context v. no context), 2 cognitive tasks (inhibition and STM), and 2 adult control groups (Arabic, English).

Apart from the variables to predict variation in children's semantic performance (e.g. numeracy skills) and pragmatic performance (e.g. cognitive skills), there might be other variables that have an impact on children's performance but were not incorporated into the study's research questions or hypotheses. These are *confounding variables*, and they have the potential to affect and obscure the causal relationship between the dependent and independent variable; therefore, they should be controlled in experimental research (Neuman, 2014). The potential confounds covered in this study are children's age, general mental ability, language proficiency, and SES. Language exposure was included in the case of bilingual and (bidialectal) Arabic-speaking children due to their exposure to more than one language or variety. The confounding variables were controlled using certain measures (described in 3.4.3), enabling the research to include them as covariates in the regression analyses for children's performance.

3.3.1 Pilot study

A pilot study was conducted on a limited sample of bilingual and Arabic-speaking children. It also included two adults groups who acted as controls for baseline for comparisons (English and Arabic adults). This pilot study applied a semantic comprehension task (give-a-quantifier), (b) two ternary judgment tasks to assess pragmatic ability in two conditions: an enriched-context condition v. a no-context condition; and (c) an inhibitory control task. Other dependent measures were adopted

to control for general intellectual ability (a non-verbal IQ test), language proficiency in English and in Arabic (receptive vocabulary test), and no further measures were taken to control for SES or language exposure. The pilot data were used to verify the design of the current research; details are integrated into the discussion below.

3.3.2 Sampling

The project used a *matched-group sample* consisting of three child groups and two adult control groups, all matched as far as possible in terms of basic independent variables such as age, number of participants, and language background. The research employed a quota sampling method to produce a quasi-representative sample (Neuman, 2014), after identifying certain categories for sampling. For child participants, the criteria were age, parental education (at least one of the parents had to have a university degree) and language situation (bilingual: exposed to Arabic and English; and monolingual: only exposed to one language, either Arabic or English). With respect to sample size, each of the bilingual and Arabic child groups contained 30 participants, while the English child group had 26. These sample sizes were adequate to enable the research to conduct certain statistical tests. For the adult control groups, there were two criteria for sampling: age and language background; that is, each participant had to be either a native speaker of Arabic or of English, with a minimal amount of exposure to languages other than the mother tongue. The sample size for each adult group was relatively small (11 English and 10 Arabic adults) since the adults were serving only as control groups to show the end developmental point of the capacities under study; this was intended to allow us to understand the performance of the child groups in more depth. Nevertheless, the main focus of the current research was on children's performance, and the use of this relatively small sample of adults was only meant to serve as a baseline.

3.3.3 Location of testing

The study conducted two types of experiments: field and laboratory experiments. The field experiments were conducted in the children's schools (bilingual children were tested in the Arabic schools they attended on the weekend) to test their abilities, while the adult control groups were tested in the lab. Apart from easing the gathering of child participants, conducting field experiments was intended to test children in a

more natural environment, fostering more natural reactions (Coolican, 2004) and consequently, ideally, reducing any deleterious effect of the artificiality of the experimental settings. The study tested the adults in the lab for two reasons: First, the effect of a controlled setting on adults' performance would be more limited, as they are better at adapting themselves to different situations in comparison with children. The second reason is that the adult participants were university students, and testing them in a university lab allowed them to easily reach the testing location and made it easier for me to find adults willing to participate.

3.3.4 Ethical issues

The ethical-moral dimension is an essential one to address adequately in research, and the current study has been conducted only after giving consideration to established ethical principles for research using human participants and only after being approved by the Performance, Visual Arts and Communication (PVAC) Faculty Research Ethics Committee of the University of Leeds (Ethics reference: PVAR 12-054). All the children who participated did so on after I secured written informed parental consent, and the adult participants completed a consent form before taking part. In addition, before testing, each child was asked personally if he/she would be happy to play some games with me (i.e., verbal assent was obtained). None of the children said they were unhappy to take part, and only two (a bilingual child and an Arabic-speaking child) asked to withdraw after completing the first task; they were immediately taken back to their classes.

3.3.5 Data analysis

All the data were first analysed with descriptive statistics (e.g. mean, range, percentages of frequency), then with inferential statistics. The inferential statistics allowed the study to explore if the groups significantly differed (using ANOVAs and alternative non-parametric tests) and to understand the relationship between different variables (using correlation and regression tests). All the analyses were conducted using SPSS software, and most of them were chosen based on Larson-Hall's (2010) discussions and recommendations.

3.4 Method

3.4.1 Participants

The 103 child participants in this study were aged between 4;1 and 7;0 years old. They can be divided into three groups: 35 English–Arabic bilingual children, 32 Arabic-speaking children, and 36 English-speaking children. However, the final analyses included only 86 children: 30 bilingual, 30 Arabic and 26 English, as shown in table 3.1. All the participating children’s parents had first-stage tertiary education (i.e. undergraduate degree), and based on this and the family wealth questionnaire (Currie, Elton, Todd, & Platt, 1997) (section 3.4.3.3.2), most of the children came from what can be described as middle-class families. The Arabic-speaking children can be described as bidialectal, as they had been exposed to two varieties of Arabic, namely, Colloquial and Standard Arabic, while the English children had had substantial exposure only to one variety of English, so it is safe to describe them as purely monolingual children. Although all those children were recruited only after parental consent, some of the parents gave their consent first by word of mouth and then returned their written consent alongside the language and socio-economic questionnaire to the school. This resulted in exclusion of some of the children after testing them. With respect to the bilingual children, five children were excluded either because they did not attend the two training sessions or because they stopped attending the Arabic school because they moved or switched schools. Some other children were excluded either because neither of their parents had at least a first-stage tertiary education, the parents did not complete all questions related to education and SES, or the parent questionnaire revealed that the child had been exposed to more than one language, although the school had introduced them to me as monolingual children.

Table 3.1. Information on tested child and adult participants in each group (further information on included participants is given in chapter 4)

Group	N of tested participants	N of Included Participants	N of excluded participants	Age range
Bilingual Children	35	30	5	4;1–7;0
Arabic Children	32	30	2	4;7–6;9
English Children	36	26	10	4;3–6;2
Arabic Adults	16	10	6	18;0–24;9
English Adults	23	11	12	18;3–24

All the bilingual children were attending British public primary school during the week and Arabic school on the weekend; in the latter, classroom instruction is heavily based on Colloquial Arabic and rarely uses Standard Arabic, while curricula and exams are given in Standard Arabic. All the bilinguals had Arab parents who were born in the Arab world and were fluent Arabic-speakers. The parents were originally either from Iraq, Libya, or Saudi Arabia, except two children who had one Arab parent and one English parent. The language background questionnaire (see appendix 1) given to parents asking about the sources of children’s exposure to Arabic showed that family members (including parents), Arabic school, and summer holidays spent in the ‘old country’ were the main sources.

The Arabic children were recruited from King Abdul Aziz Model (i.e. private) Schools Kindergarten in Tabuk, a city in the northwest of Saudi Arabia. The classroom instruction was mixed, in Colloquial and Standard Arabic, while all curricula were in Standard Arabic. The English children were recruited from Holy Trinity Church Primary School in London and Lily Croft Primary School in Bradford. Although all those children were recruited after parental consent, some of the parents gave their consent first only by word of mouth and then returned their written consent alongside the language and socio-economic questionnaires to the school.

With regard to the adult participants, there were 39 participants (23 English and 16 Arabic-speakers) aged between 18 and 24 years old (see table 3.1 above). Unfortunately, I had to exclude approximately half of the English and Arabic

participants I tested because I initially thought, mistakenly, that it would not be necessary for adult participants to complete the vocabulary test, and had to test new participants (11 English and 10 Arabic) after realising that testing adults' vocabulary would be necessary to enable comparison with that of the children, which was very important in the case of the Arabic participants since that version of the test was translated by me (that is, it was not yet validated). The adults completed all the tasks completed by the children except the mental ability test, which was tailored for children, and the adult results were used as a baseline for children's performance as they represent the endpoint of the children's cognitive development. All the adult participants were either current university students or University graduates. Most of the Arabic participants were tested in the UK, either while they were there on their summer holidays or while completing a course in the UK; only a few were tested in Saudi Arabia. All the English and Arabic adult participants reported that they had been exposed to another language but that their use of the second language was limited to educational situations, without intensive daily use, and that they had not reached a native level. Most of the English-speakers had taken some courses in a language other than English, while the Arabic-speakers had completed their university studies and exams in English. It is very common in Saudi Arabia for universities to have their curricula and exams in English, especially in natural science disciplines such as medicine, engineering, and computer science, which were the participants' majors.

Separate from the above-mentioned participants, others were recruited for a pilot study (15 English adults, 8 Arabic adults, 5 bilingual children, and 5 Arabic children). The English adults in the pilot were students at the University of Leeds, and the Arabic adults had come to the UK to take language courses, either in preparation to start a degree later on or just to improve their English level. The bilingual children were selected under the same criteria as in the main study regarding exposure to two languages and parents' educational level and were tested in the UK, while the Arabic children were tested in Saudi Arabia.

3.4.2 Materials and procedure

All the tasks were administered by a single experimenter—the present author—and all the children were tested at their schools, in a moderately quiet area that was accessible to parents and staff at all times (either a dedicated room or the school library). Bilingual children were tested in two sessions on two different days, one for English and one for Arabic, while the Arabic and English children were tested in only one session. Each session started with the vocabulary test, followed by the first pragmatic task (experiment 3: enriched context), the STM task (the Corsi blocks task), the second pragmatic task (experiment 4: no context), and then the inhibition task (the Simon task). After this, children completed the first semantic task (experiment 1: give-a-quantifier), a non-verbal IQ (NVIQ) test (see below), the two counting tasks (how-many and give-a-number), the non-verbal ordinal task, the estimating-magnitude-numerically task, and finally the estimating-magnitude-proportionally task (experiment 2) (all tasks explained in sections 3.4.4 and 3.4.5 below). Children were given short breaks between tasks, during which I asked them if they liked the task they had just completed, which of the tasks they had completed were more fun for them, or just general questions such as their favourite food or colour. The purpose of these questions was just to avoid silence during the short break. The bilingual children only completed the NVIQ test, the cognitive and number tasks, and the second semantic task in English only. Each session took approximately 45 minutes, with the second session for the bilingual children slightly shorter (around 35 minutes) due to the absence of the second semantic task in Arabic. Instructions were given in Arabic when testing in Arabic, and in English when testing in English.

3.4.3 Controlling for confounding variables

In this section, I explain how potential confounding variables were assessed and controlled. These variables were language proficiency, general mental ability, SES, and for the bilingual children and the Arabic children, who had been exposed to more than one dialect, language exposure.

3.4.3.1 Language proficiency test

The third edition of the British Picture Vocabulary Scale (BPVS; Dunn & Dunn, 2009) was used to measure participants' receptive vocabulary, as a proxy indicator of

their language proficiency (as suggested by Bialystok & Luk, 2012). The test is suitable for children as young as 3 years and up to 16 years of age and older. It consists of 168 items divided into 14 subsets by appropriate age, and there are 12 items in each subset. For each item, the participant is presented with four pictures and asked to point to the one that matched the word he or she heard. The test starts with three practice items, and then the experimenter starts the test proper with the first word in the subset corresponding to the participant's age. If a child answers all the items correctly or only makes one mistake, this subset is considered the child's 'basal' subset. If more than one error is made, the experimenter finds the basal subset by testing backwards through the preceding subsets until the mistake criterion is met. After establishing the child's basal subset, the experimenter tests forward until 8 or more errors are made in a single subset, which then represents the child's 'ceiling' subset. Children are given 1 point for each correct answer and 0 for each wrong or 'do not know' response. The sum of correct answers is calculated, and then the final score is computed by subtracting the sum from the test's total score, which is 168. All the children received neutral feedback after each trial, regardless of the accuracy of their answer: 'OK, thank you!' Corrective feedback was provided only in the practice items.

To test the bilingual and Arabic children's receptive vocabulary, I translated the BPVS into Arabic. This was the most appropriate option for specific reasons. First, the available Arabic versions of possible alternatives, the Peabody Picture Vocabulary test (Dunn & Dunn, 1997) and the Versant Arabic Test (Pearson Education, 2011) are only available in Standard Arabic, making them inappropriate given the participants' limited exposure to Standard Arabic. Second, despite attempts to create a reliable measure of children's language skills for some Gulf Arabic-speaking countries, such as Qatar (Shaalán, 2010), Saudi Arabia (Al-Akeel, 1998), or Bahrain (Mannai & Everatt, 2005), large groups of children have not yet been tested and therefore these instruments have not yet been published as reliable standardised tests (Shaalán, 2009). Even if some of these attempts could reliably assess children's vocabulary level, they could not have been adopted as a valid measure of children's vocabulary in this study, because the bilingual children have a range of home Colloquial Arabic dialects—and even within individual countries there are different regional dialects that should be taken into account before adopting any such measure. All these issues have led

various past researchers to translate existing English language and vocabulary tests and adopt them in their research (e.g. Alduais, Shoeib, Al Hammadi, Al Malki, & Alenzi, 2012; Fedda & Oweini, 2012; Alkhamra & Al-Jazi, 2016; Al-Akeel, 1998). It has been stated that the limitations that might be associated with such an approach can be overcome by employing further measures, such as parents' language questionnaire (Shaalán, 2009). Although the current study included a language exposure questionnaire completed by parents, it was adopted to assess amount of input, and did not assess vocabulary.

The use of the BPVS, which is a culture-neutral test, makes the task freely translatable to Arabic. Further, the test only gives a measure of children's receptive vocabulary; this facilitated the translation process, since I only needed to control one variable, namely lexical dialectal variation. Including different measures to assess other language skills, such as grammar, would make the translation not only more difficult but possibly also less accurate due to the potential for effects of idiosyncratic syntactic variation in addition to possible dialectal variation across individuals. The translation procedure was as follows. First, I translated all the items of the BPVS; then, the translation was checked and corrected by another native Arabic-speaker, who has a bachelor's degree in Arabic language and literature. After this, all the items were piloted with Arabic adult volunteers who spoke a range of Colloquial Arabic dialects (hailing respectively from Palestine, Jordan, Oman, Libya, Iraq, and Saudi Arabia). All the items were mutually intelligible to them, but some participants reported different words for some concepts, and these differences were considered when testing each child. The terms were the word for 'shoes', which had three lexical varieties (*kondara*, *jouti*, and *jazma*), and 'jogging', which was expressed either as *ye'jri*, *yu'rkud* or *yu'haruel*. The Iraqi speaker also mentioned a different word for 'banister', *em'hajar* as distinct from *drabzeen* or *soor eldaraj* (the latter two are mutually intelligible in other dialects). This, of course, is not to say that these are the only lexical variations between these dialects, or that all dialects in any of these countries use the same lexical expressions, but of the words in the test, these were the only ones highlighted by the participants. In addition, I checked with one teacher in each of the bilingual children's Arabic schools to ensure that they used the same lexical terms in the test, and in the pilot, I checked with one of the children's parents.

3.4.3.2 Non-verbal IQ test: Matrix Reasoning

The Matrix Reasoning test (Wechsler, 1967/2012) is a subtest of the Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV), which is administered as a measure of general intellectual ability and which is suitable for children 4 years old and older. In this test, children view an incomplete matrix and select the response option that completes it. The test consists of 26 items, and starts with three practice trials. In accordance with the test protocol, for each item, I started by asking the child to look at the pictures, and then asked: ‘Which one here [while pointing to the response option] goes here [while pointing to the empty box]?’ The children were asked to clearly indicate their response either by pointing to the picture or saying the number of the selected response. In the practice items, if the child answered correctly, I would say ‘That’s right’, and proceed to the next trial. If they gave a wrong answer, then I would say ‘Good try, but that’s not quite right’ and explain that a specific item should go with similar ones (e.g. ‘This yellow umbrella, should go in the empty box with those similar yellow umbrellas’). Once the child completed the trial items, I started the main test, with age-appropriate items. After a few items, the children began to understand the task and started to point to the selected response immediately without prompting. Children were given 1 point for each correct response and 0 for wrong answers. Scores were standardised using tables in the Administration and Scoring Manual. Only the child participants completed this task, as the test was designed for children, not adults. The bilingual and English children were tested in English, while the Arabic children were given the same instructions but in Arabic. The task took approximately 2 minutes to complete.

3.4.3.3 Socio-economic status (SES) measure

Family SES has been a central factor to explain variance in students’ high school achievement (Tenaw, 2014; Azhar, Nadeem, Naz, Perveen, & Sameen, 2013; Sirin, 2005) and also in children’s cognitive development (Dickinson & Adelson, 2014; De Abreu et al., 2012; Mezzacappa, 2004). This factor is often measured based on three aspects: parents’ educational achievement, occupational status, and family income (Bradley & Corwyn, 2002). In this study, I assessed parental education and family income (measured as explained below) as proxy indicators of family SES. Although SES is not a core interest in this study and is only considered as a potential confound

that should be controlled, to justify this approach, two questions should be briefly answered: First, how could SES potentially affect children's development, and second, why did I select these two indicators and avoid parental occupation?

Regarding the effect of SES, financial resources are associated with a child's having better opportunities in life, partially related to physical health conditions resulting from better nutrition and health care services (Bradley & Corwyn, 2002; Behrman, 1996), mental health due to better emotional caregiving and less economic and social stress (De Abreu et al., 2012; Bradley & Corwyn, 2002), and better technological facilities, which might have an impact on a child's cognitive and creative abilities (Subrahmanyam, Greenfield, Kraut & Gross, 2001; Azhar et al., 2013). Parental education's impact on children can also be clearly associated with parents' attitudes, expectations, and motivations toward their child's education. For example, it has been indicated that educated parents tend to provide their child with a more relaxed home environment, engage in richer conversations with their child, read more to them and encourage their reading habits by purchasing more books (Bradley & Corwyn, 2002).

This study measures SES based only on income and parental education and excludes parental occupation for two practical reasons. First, if parental occupation is used to reflect parents' prestigious social status (Dickinson & Adelson, 2014), we must acknowledge that what might be considered a more prestigious occupation in one society might not be in another. The General Social Survey (National Center for Education Statistics, 2004) used to classify social prestige in (e.g.) Dickinson and Adelson (2014) might not be an accurate measure due to the subjectivity involved when asking people to rate different occupations, and indeed that study's results revealed that 'occupational prestige for either parent tends to contribute less to SES than [do] education and income' (p.6). Second, when occupational is taken as a measure of SES, as in the International Standard Classification of Occupations (ISCO; International Labour Office, 1990), this has been done by basically relying on education and income (Ganzeboom & Treiman, 1996) to categorise occupations, making occupation to a large degree superfluous.

I decided to assess education and income separately; below I explain the techniques I used to do so.

Parental education

With research that includes an international sample, educational attainment is often classified using the International Standard Classification of Education (ISCED) (UNESCO, 2011). ISCED enables the comparison of education qualifications within and between countries (Schneider, 2013). Therefore, this research used the revised (2011) version of ISCED to classify parents' education. The parents were asked about the highest degree of education they attained, and then I used the ISCED mappings for each country to convert each qualification to its equivalent level on the international scale (degrees were obtained in Iraq, Libya, Saudi Arabia, and the United Kingdom). I decided to add this measure only after starting testing the bilingual children, and I obtained this information either from the head teachers at their schools or from the children's parents directly, but with the other two groups, the Arabic- and English-speaking children, I was able to ask the parents about their educational level in the questionnaire that they completed.

Family wealth

The four-item Family Affluence Scale (FAS; Currie et al., 1997) questionnaire was administered to measure family wealth as a proxy indicator of children's SES, for two reasons. First, it is usually difficult to ask individuals about their income directly, as some might be sensitive to such a question (Brese & Mirazchiyski, 2010). Thus, indirect questions about things such as possessions and vacations, helped to yield some indications about financial resources (Boyce, Torsheim, Currie, & Zambon, 2006). Second, even if it were possible to ask about parents' annual income directly, with an international sample, income alone would not tell much about the family's economic status. Instead, family income should be taken with other factors to indicate SES more accurately, for example the number of family members, the amount of taxes paid, accommodation (rented or owned?), health care, and education (free or not?). Obviously, it would be quite difficult to measure all these areas across countries, or just to estimate their effect on family income and therefore SES by asking parents about financial resources or their occupations. Therefore, it might be more accurate to adopt a measure that is associated with common consumption indicators of material deprivation rather than relying on superficial estimations of

family income. The FAS was a good available option due to its simple, brief questions, which made the task easier for parents, and due its having been used over many years as a measure of family wealth by the World Health Organization's Health Behaviour in School-Aged Children (HBSC) survey.

The questionnaire includes the following items: *Does your family own a car, van, or truck?* ('No'=0; 'Yes, one'=1, 'Yes, two or more'=2); *Does your child have his own bedroom?* ('No'=0; 'Yes'=1); *During the past 12 months, how many times did you travel away on holiday with your family?* ('Not at all'=0; 'Once'=1; 'Twice'=2; 'More than twice'=3); and *How many computers does your family own?* ('None'=0; 'One'=1; 'Two'=2; 'More than two'=3). Based on the parents' answers, each child was given a score from 0 to 9. After this, a composite FAS score was calculated using a three-point ordinal scale, where 'FAS low' (score=0, 1, 2) indicates low affluence, 'FAS medium' (score=3, 4, 5) indicates average affluence, and 'FAS high' (score=6, 7, 8, 9) indicates high affluence. In most of the analysis, where SES was used as a predictor for children's performance in different tasks, I used this re-scaled score rather than the composite score.

However, although the current research adopted this questionnaire, we should be cautious with its results, which might be influenced by various factors not covered by the instrument; for example, regardless of SES, people in Bradford are much more likely to own a car and to have children with their own bedrooms than in London.

3.4.3.4 Language background

In order to assess bilingual children's exposure and use of the two languages they speak, I adopted the Utrecht Bilingual Language Exposure Calculator questionnaire (UBiLEC) (Unsworth, 2013; see appendix 1). The questionnaire helped obtain information about when and where children use their two (or possibly more) languages, when they were first exposed to them, and how fluently they speak them. The questionnaire was developed from existing questionnaires such as Gutiérrez-Clellen and Kreiter's (2003) parent questionnaire, where Unsworth added a digital version in the form of a Microsoft Excel spreadsheet, which combines a number of algorithms that allows researchers to measure three different aspects of a child's

language exposure: (a) the quantity of exposure the child has to a specific language at the present time, (b) the quality of exposure the child has in a specific language at the present time, and (c) the quantity of exposure the child has had to a given language over time (the *cumulative length of exposure*). The outcomes of the questionnaire were given as input percentages for the child's exposure to languages A and B and also the child's use of these two languages (the output percentages). I calculated the cumulative length of exposure to a specific language manually using this equation: $\text{cumulative length} = (\text{age at testing} - \text{age at onset}) \times \text{amount of input}$ (see Unsworth et al., 2014). In the current study, for all the bilingual children, the Arabic language was considered the home (first) language (L1) and the English language was the second language (L2). The bilingual children's parents were given the chance to choose the language in which they preferred to complete the questionnaire, and English and Arabic versions were prepared.

Since Arabic-speaking children also had been exposed to two varieties of Arabic, namely Colloquial Arabic and Standard Arabic, I measured these children's amount of exposure and use of these two varieties. To do so, I translated Unsworth's language background questionnaire into Arabic, with slight modifications. I changed 'home language' to 'home dialect' (D1), which was Colloquial Arabic, and 'second language' to 'second dialect' (D2), which was Standard Arabic. All the translations were done by me and checked by two other native Arabic-speakers, one of them holding a bachelor's degree in Arabic language.

When conducting the analyses, the potential confounding variables that have been described in this section (such as age, language background, language proficiency, and SES) were first tested to see if they correlated with children's pragmatic and semantic performance; if a correlation was found, the variables were included as covariates in the regression model.

3.4.4 Study 1: Children's comprehension of quantifiers and operators and the potential effect of numeracy

The first part of this project aims to measure children's semantic quantification skills in order to empirically answer two questions: Do bilingual and monolingual children

comprehend the quantifiers *most*, *some*, and the operators *or*, and *and* in a semantically appropriate way? And to what extent does the acquisition of a numerical system promote or possibly hinder the acquisition of quantifiers? Two semantic tasks were set to answer the first question and four numerical tasks to explore numeracy skills for the second. In this section, I describe these tasks in detail and justify why they were selected and how they attain the research goals.

3.4.4.1 Semantic performance

There were two semantic tasks¹ used to investigate children's comprehension of the quantifiers; a perception task (experiment 1) and a production task (experiment 2). The research started basically with one semantic task (the give-a-quantifier task); but due to the children's weak semantic performance, the estimating-magnitude-proportionally task was added to better understand children's behaviour.

Experiment 1: Give-a-quantifier task

The goal of this experiment was to explore children's comprehension of the quantifiers 'all', 'most', 'some', and the operators 'or', and 'and'. It is adapted from the give-a-quantifier task used in Hanlon (1987) and Barner et al. (2009), in which children were asked to act upon given instructions (statements) such as *give the puppet some of the apples*. The present study used 15 trials to test children's understanding of the lexical meaning of the quantifiers and operators; three for each of the English quantifiers/operators 'all', 'most', 'some', 'or', and 'and' or their Arabic equivalents (*kul* 'all', *muāḍam* 'most', *baāḍ* 'some', *ʔw* 'or', *wa* 'and'). Stimuli used in these trials were pencils, small plastic dinosaurs, carrots, apples, spoons, flowers, and balls. In addition, small boxes, plates, and a puppet (with a small basket) were provided so that the child could move the items between these carriers. Early on, before designing the task, all the items were checked by five Arabic-speaking adults with different dialects and re-checked by the children's teachers to ensure that the items were mutually intelligible across dialects. None of the teachers reported different lexical use of the items. For the quantifiers 'all', 'most', and 'some', stimuli in each set had the same, colour, shape, and size; only plastic

¹ The terminology 'semantic' here is meant to encompass 'semantics and pragmatics'.

dinosaurs and balls came in two colours (e.g. some dinosaurs were brown and some were green). For the operators ‘or’ and ‘and’, there were three different stimuli in each set (e.g. a pencil, a carrot, and a ball). The objects were put on a child-sized table in front of the child, and before starting the test, I presented the items to the participant, to ensure that the child could differentiate between them. This was done by telling the child the name of each item: ‘Here we have apples, pens, boxes, a puppet [etc.]’, pointing to each object while naming it. I was careful not to use any quantifier when introducing the objects. Next, the participants were asked either to (for example) *[p]ut some of the apples in the box* or to *[g]ive the puppet a flower or a pen*. I gave neutral feedback to children on almost every trial. As in Barner et al. (2009), care was taken to ensure that prosody was consistent across quantifier trials (by putting a slight stress on the target quantifier). There were three trials for each quantifier, and the items were given in a pseudo-randomised order. The use of multiple objects helped avoid potential issues stemming from reuse of the same objects several times and helped complete the test faster, as there was no need to return the items to their original piles after each trial. This task took approximately 2–3 minutes to complete.

Experiment 2: Estimating magnitude proportionally

After assessing children’s comprehension of the quantifiers and operators, preliminary data revealed weak semantic comprehension of the quantifier ‘most’ by all children, and of the quantifier ‘some’ by the bilingual children (in both Arabic and English) and by the Arabic children. Therefore, the estimating-magnitude-proportionally task was added as a further assessment of children’s acquisition of these two quantifiers. This task has two aims: a) testing children’s ability to map different proportions with the appropriate quantifier, and b) assessing children’s comprehension of the lexical scale, that is, their knowledge that ‘some’ and ‘most’ have different ordinal positions in the quantifier scale according to the proportions they represent.

To achieve these two goals, I designed a production task—a mix of Yildirim et al.’s (2016) pre-exposure test and Tillman and Barner’s (2015) forced-choice task. Much like in Yildirim et al. (2016), the participants were presented with a fixed overall set

size with various breakdowns of items by proportion within it. Unlike Yildirim et al., I used 15 circles instead of 20, for two reasons. Firstly, since the participants were children, using a slightly smaller set size could make the estimation process less complicated compared to that undergone by Yildirim et al.'s adult participants, who have more advanced cognitive resources to estimate larger set sizes. Secondly and more importantly, I tested children's ability to count a set size of {14} in the how-many task (section 3.4.4.2.1), so it seemed appropriate to give them a very similar set size {15} in the estimating task. The set in each trial consisted of blue and yellow circles presented on a laptop (see figure 3.1 below), and the proportional distribution of target circles varied from trial to trial. There were 11 trials in total: 4 (critical) small proportions {2/15, 3/15, 4/15, 5/15} that were expected to be mapped to 'some', 4 (critical) large proportions {10/15, 11/15, 12/15, 13/15} to be mapped to 'most', and 3 fuzzy (filler) proportions {6/15, 7/15, 8/15} representing approximately half of the set that were used as fillers. The goal was to test the ability to map the small-condition stimuli to *some* and the large-condition to *most* and not how children would describe the proportions themselves; therefore 4 items were very similar proportionally were included in each condition, although there was only one trial for each proportion.

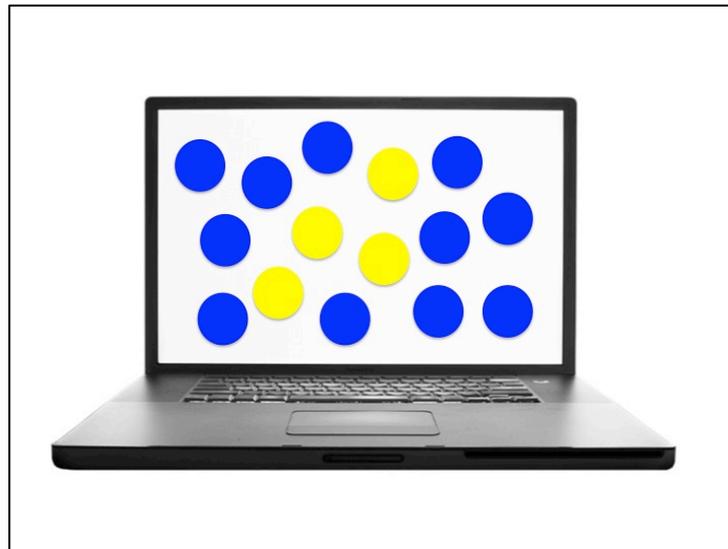


Figure 3.1. Sample item from the estimating-magnitude-proportionally task (4/15 proportion), presented on a 13-inch laptop

At the beginning of the task, I explained to the child that he/she would see blue and yellow circles appearing on the computer screen. Similar to Tillman and Barner (2015), the child was given a *forced choice*, that is, they were asked to describe *only* the yellow circles either using ‘some’ or ‘most’ (e.g. *Look at these circles; can you say if most or some of the circles are yellow?*). If a child used different quantifiers (e.g. ‘little’ or ‘many’), I explained that he/she could only use ‘some’ or ‘most’ and asked the child to try again. The trials were presented in a pseudo-randomised order, and no training was given in this task because all the children had completed this task after successfully completing the non-verbal ordinal task as a training task (see section 3.4.4.2.3 below), in which it was revealed that all the children successfully demonstrated the ability to point to a set that had more items even when the difference between the two sets in the scene was numerically very close (e.g. 2 v. 3, 6 v. 7). Neutral feedback was given to the children after each trial, and the task took approximately 2–3 minutes to complete.

3.4.4.2 Number tasks

The tasks in this section aimed to assess children’s numeracy skills. They were added to this study to explore whether some children’s (especially the Arabic children’s) poor semantic skills with quantifiers resulted from a delay in acquisition of the numerical system. Thus, the current study investigates the question: To what extent does the acquisition of the numerical system promote or possibly hinder the acquisition of quantifiers? I build my hypothesis on the potential relationship between number and quantifier words on three sources: first, Barner et al.’s (2009) findings of a correlation between children’s acquisition of number and quantifier word meaning; second, the theoretical assumption that quantifier words should be available for children by the time they start learning number word meaning (Piantadosi et al., 2012); and third, the pilot data for the pragmatic task, in which some children did not judge an utterance until they had counted the objects on the computer screen. These findings and assumptions led me to include the numeral tasks, but I went beyond exploring children’s acquisition of exact numeral systems (how-many and give-a-number tasks) to also investigate their approximate numeral systems by mapping different magnitudes to their approximate true values in the numeral list without counting (estimating-magnitude-numerically task), since such a skill might be

essential for the acquisition of quantifiers, as I explained in detail in section 2.2 (in chapter 2). Below, I describe each numeral task and explain precisely why it was included.

The how-many task

This task aimed to assess children's counting ability; more precisely, it explored children's acquisition of the three 'how-to-count' principles proposed by Gelman and Gallistel (1978): (a) the *one-to-one principle*, which implies that when counting a set, only one numeral must be given to each item in the set; (b) the *stable-order principle*, which indicates that when counting, numerals must be used in the same order in any one set as in any other set; and (c) the *cardinal principle*, which indicates that the numeral assigned to the last item in a set represents the number of items in that set. To test all three principles in one single task, I combined Sarnecka and Carey's (2008) how-many task, which explores the cardinal principle, and Le Corre and Carey's (2007) count list elicitation task, which explores the one-to-one principle. In the adapted how-many task (combining the how-many and count list elicitation tasks), I provided children with a white card (in A4 paper size; see figure 3.2 below) including a single row of 10 small apples and asked them *Can you count these apples loudly and tell me then how many apples are there?* If a child counted the set correctly and gave me the last numeral as the number of apples in the row, then I presented the second card, which featured two rows with 7 small chickens in each (set size 14, as in figure 3.2), and asked the same question. (The reason for presenting the chickens in two rows instead of one was to ensure they were large enough to count without double-counting or skipping items.) If a child counted wrong (e.g. if he/she skipped an item), I asked them to count the set one more time, more slowly, and assisted by pointing to each object as they counted. I only recorded the result of the second attempt if there were two, and regardless of the child's answer, I gave them mildly positive feedback (e.g., 'OK, thank you!'). Including this remedial counting help was also done by Le Corre and Carey (2007) in their count list elicitation task, to avoid underestimating children's counting ability, especially due to mistakes resulting from fast response, which would not reflect underlying delayed acquisition, whereas actual acquisition problems would likely also appear on recount. The task takes approximately 1–2 minutes.

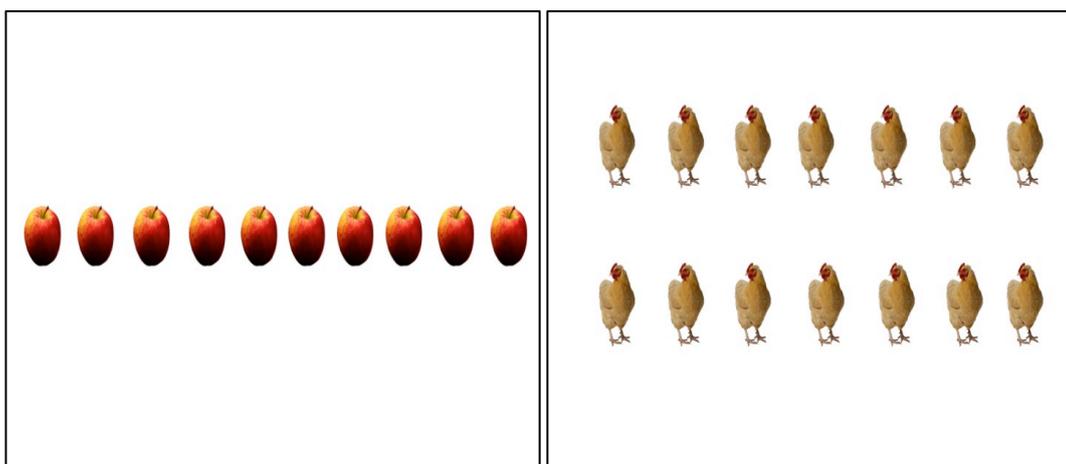


Figure 3.2. Stimuli used in the how-many task. Set size {10} (on left) and set size {14} (on right); each set presented on a white card (A4 paper size)

By asking children to count aloud, I assessed the one-to-one principle, that is, their ability to shift values on a numeral list when moving up one item in a set. Having two set sizes {10} and {14} served two goals: (a) assessing the stable-order principle by having two sets, and more importantly (b) ensuring that they could count relatively large sets as well as smaller ones. Exploring children's ability to count the {14} set accurately was important in particular for the estimating-magnitude tasks (proportionally and numerically), whereas the how-many question was important not only because it revealed whether a child had mastered the cardinal principle but also because demonstrating the ability to map sets to their true values (with the last numeral a child counted) was meaningful for the estimating-magnitude-numerically task.

Give-a-number task

Although the how-many task explores children's acquisition of three counting principles (see the how-many task above), the task itself does not demonstrate whether children understand the exact meaning of number words, but only their knowledge of counting and their ability to map sets to their true values in the numeral list. Therefore, I included a give-a-number task (adapted from Le Corre & Carey, 2007) which requires a child to build new sets from existing context.

In this task, I placed two containers on a table in front of the child, each filled with 8

small, similar plastic items, either dinosaurs or balls, and three empty boxes. The child was asked *Could you take one dinosaur out of the bowl and put it in this box?* (I used several boxes, which were cleared between trials.) Then, the child was asked to take/put another number of items, from 1 up to 6. All trials were conducted in the same order: I started with {1}, then {4}, then {2}, then {5} then {3}, and finally {6}. The reason for choosing 6 as the last number was to ensure that children could generate sets exceeding {5}, as this number could fit easily in one of a child's hands (Sarnecka & Carey, 2008). In addition, children might build their knowledge of 5 relying on analogy between their fingers on one hand and the items they put in the box, so including a {6} set would also help ascertain their actual counting ability from this perspective. Of course, it might be said that a child might generate a {6} set just by adding one item to 5 (based on the fingers of one hand), and that could be true and represent part of the learning process, but such an ability in itself would indicate that children were able to create sets above 5 accurately; especially considering that the numbers were given in a pseudo-random order. If a child produced a wrong set, I gave the child a chance to correct himself/herself by asking 'Can you count and make sure that this is an X?' where x is the required number. If the child corrected such an answer, then I would give the next number (with positive neutral feedback: 'Thank you, let's try another number'); if not, then I would record the answer and say, 'OK, good try, but this was 6, not 5; let's try another number'. There was one trial for each number, with a chance for self-correction of wrong responses, each child's score was the same as the highest number they could produce correctly. After each trial, I returned all the items to the container before asking about the next number. The give-a-number task took approximately 2 minutes to complete.

Non-verbal ordinal task

This task was adopted from Le Corre and Carey (2007); the goal was to non-verbally assess the availability and accuracy of children's analogue magnitudes. In other words, the task aimed to ensure that children had acquired the ability to spontaneously distinguish set sizes of similar magnitudes without counting. This ability was critical for estimating magnitudes proportionally and numerically (estimating magnitude proportionally and estimating magnitude numerically tasks).

On a computer screen, children viewed two sets of circles separated by a thick line simultaneously on a PowerPoint slide (see figure 3.3 below). In each of 6 trials, they were asked *Could you point to the side that has more circles?* but instructed not to count them. After the child gave an answer, I moved on to the next trial and asked the same question. After a few trials, children sometimes pointed to the side with the larger number of circles even before I asked them, at which point I ceased asking the question. None of the children counted the sets before giving an answer, so I did not have to discourage them from counting.

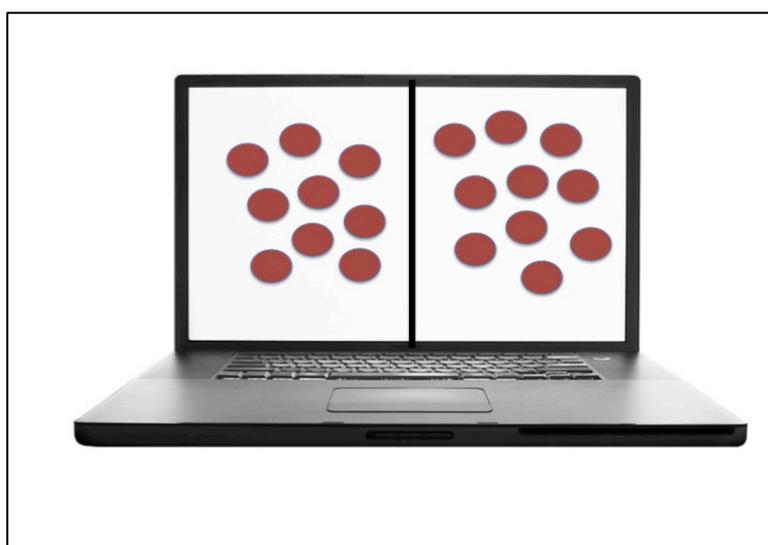


Figure 3.3. Sample item from the non-verbal ordinal task (9 v. 10), presented on a 13-inch laptop

Pairs tested were (2 v. 3, 2 v. 6, 6 v. 10, 8 v. 10, 9 v.15, 12 v. 15); each pair was presented once. Circles in all the pairs were of the same size, and for each trial, the same colour: dark red in three trials and dark blue in the other three. Using different colours was intended to make children aware that they were evaluating different pairs in each trial; the choice of colours was arbitrary. The configuration of the circles in each set (as well as the number) varied from trial to trial. The pairs were presented in a pseudo-random order. No feedback was given to the children as they went, but after completing the task, I praised the child for his or her performance. The task took approximately 1–2 minutes to complete.

Estimating-magnitude-numerically task

This task was adapted from Le Corre and Carey's (2007) fast card task. Implemented after assessing children's acquisition of verbal counting, their knowledge of exact number meanings, and their ability to clearly distinguish set sizes with larger magnitudes from smaller ones without counting, this task had two goals. The first was to investigate children's approximation numerical system, while the second and more important goal was to ensure the availability of scalar variability. This task was completed before the estimating-magnitude-proportionally task to make sure that the children were able to alter their responses as the set size changed. If children were to successfully complete the task using numbers but fail with quantifiers, one might relatively safely attribute their poor performance to the incomplete acquisition of the quantifier term rather than a lack of scalar variability; more generally, this would then suggest that a numerical system seems to be acquired earlier than quantifiers, if the opposite results were found, this would be in line with Carey (2004) and Piantadosi et al. (2012).

On a computer screen, children viewed different magnitudes (number of circles) in each trial (using PowerPoint; see figure 3.4) and were told that each circle-set would appear very quickly and they would have to tell me how many circles they saw, without counting. To encourage children to take part and prevent them from counting, I emphasised to them that they did not need to give the exact right number to win the game, but just to guess and say what number it looked like, as fast as they could. To attract children's attention and motivate them for the game, the task started with an opening scene, in which five slides flashed automatically, very quickly (3, 2, 1, Ready, Go!), as I read the countdown aloud to the child as it appeared on the screen. The first trial after this opening scene contained 2 circles, the next 1 circle, and the next 3 circles. After these three trials, set sizes varied pseudo-randomly (in a random but consistent order for all children) up to 12 circles; I pressed a button to move to the next trial once the child had estimated each magnitude presented onscreen. The reason for starting with relatively small numbers was to familiarise the children with the task, and they were encouraged with positive praise so that when reaching larger sizes their guesses would be more comfortable and confident.



Figure 3.4. Sample item from the estimating-magnitude-numerically task (set size 6), presented on a 13-inch laptop

The sets were (1, 2, 3, 4, 5, 6, 8, 12). In all trials, the circles were of the same size and colour (orange). Each number was tested in one trial, and after the first three trials, the children guessed the number immediately, that is, without my asking about the number of circles. The children received positive praise after each trial, regardless of their performance; the task took about 2 minutes to complete. If a child did not give an answer within 3–4 seconds and/or started counting, I hid the screen with an A4-sized card, explained that counting was not part of the game they were playing, and asked the child just to attempt to guess. Few children ever tried to count, and they were easily discouraged from doing so.

3.4.5 Study 2: The potential effect of bilingualism on pragmatic competence

The second part of this research aimed to empirically test the question: Can any superior pragmatic competence in bilingual children be explained in terms of a cognitive advantage over monolinguals? To do so, two pragmatic tasks and two cognitive tasks were employed to explore different child-groups' performance. In this section, I explain these tasks and my rationale for including them in detail.

3.4.5.1 Pragmatic performance

Two pragmatic experiments were employed to explore children's sensitivity to the violation of the first Gricean Maxim of Informativeness (1975, 1989). The difference between the two experiments was the presence/absence of context; in the first experiment, stimuli were presented as part of a scenario viewable on a computer screen (enriched context) and the child was asked to evaluate if a fictional character was describing what happened appropriately, while in the second experiment stimuli were statements uttered by the same fictional character as in the first experiment, and evaluating these statements depended on both the child's pragmatic sensitivity and his or her world-knowledge. The rationale for including two pragmatic tasks was twofold. First, previous empirical evidence for a bilingual pragmatic advantage (Slabakova, 2010) emerged in adults who completed a no-context task (this study's experiment 4, adopted from Noveck, 2001). However, the task itself has been criticised for using unnatural stimuli that do not arise in everyday conversation (Geurts, 2010). This led me to employ another task whose stimuli are more similar to those found in everyday language use, from Katsos and Bishop (2011; the ternary-judgement task). The second reason was to explore how children's performance would differ when context was manipulated, which I hypothesised might contribute to a better understanding of children's pragmatic ability and possibly provide some support to one of the implicature processing theories—Default (Levinson, 2000) or Relevance (Sperber & Wilson, 1986/1995)—over the other.

Experiment 3: Ternary-judgement pragmatic task (enriched context)

This experiment was adapted from Katsos and Bishop's (2011) experiment 2. It aimed to explore children's sensitivity to and rejection or tolerance of under-informative utterances. As in Katsos and Bishop (2011), a computer-based task was created by conjoining clipart pictures and animations with pre-recorded utterances on Microsoft PowerPoint slides. At the beginning of the experiment, the participants were introduced to a fictional character, called Mr Kareem (the description below takes the English trials as the base, with description of the few differences in the Arabic ones following). Mr Kareem appeared in the middle of the computer screen, introduced himself (his voice was pre-recorded by a male—non-native, but proficient—speaker of English), and asked the participants to help him learn English. I explained that Mr

Kareem spoke English very well, but would like to learn to speak English as perfectly as the participant. I then explained that they would watch some stories and that at the end of each story, I would ask a question and Mr Kareem would attempt to answer it. Participants were asked to reward his response on a three-point scale consisting of three different-sized strawberries, as used in Katsos and Bishop (2011). I told the children that strawberries were Mr Kareem's favourite food, and explained that they should reward him with a large strawberry for a very good (correct) response, a medium strawberry for a not completely correct and not completely wrong response, and a small strawberry for a wrong answer. Representations of these strawberries were placed in front of the child in a horizontal line on printed paper, as he/she watched the scenarios. After each scenario, the child was asked to reward Mr Kareem's response by grasping or pointing to the most appropriately sized strawberry; I recorded their choices. The adult participants were given a sheet on which to record their responses by ticking the most appropriate strawberry for each numbered item on a sheet of paper.

Each story started with Mr Kareem appearing at either the right top or the right bottom corner of the screen, in addition to the protagonist of the story and various items. I pointed out the protagonist and the items to the children, gesturing to them on the screen, and then asked the children to watch to see what action the protagonist would perform. For example, in one story, I told the child 'As you see, in this story, there is a girl, hearts, and stars; let's see what she is going to do'; then, the action started, with the girl moving from the right side with an eraser in her hand and erasing all 5 hearts appearing in the scenario, one by one (I used PowerPoint animation to do this; see figure 3.5, left). While the protagonist performed the action, I made remarks such as 'Look, the girl erased a heart'. After the action was completed, I asked Mr Kareem, who was ostensibly watching the scenario with us, 'OK, Mr Kareem what did the girl erase?' Mr Kareem answered with the under-informative utterance *The girl erased some of the hearts*. For *or* and *and*, only 3 items appeared on the screen: Figure 3.5 (right) gives an example of under-informative *and*, in which the protagonist (the dog) moves to pick up 2 items (the apple and the orange); in this case, Mr Kareem replied 'the dog picked up the orange'.

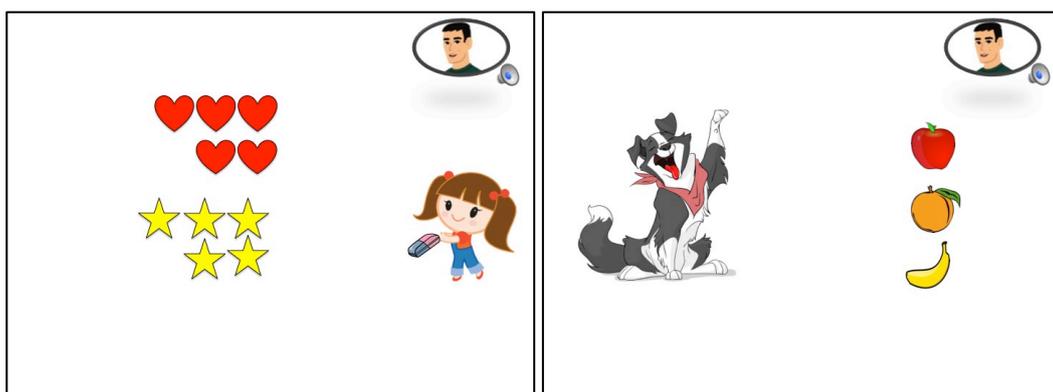


Figure 3.5. A sample of the (enriched context) ternary-judgement task (presented on a 13-inch laptop) with enriched context (for under-informative *some* on the left: the girl would erase all the hearts, and then Mr Kareem (the right top corner) would describe the action as *The girl erased some of the hearts*). For the under-informative ‘and’ on the right: the dog would pick up an apple and a banana, then Mr Kareem would describe the action as: *The dog picked up the banana*’.

Having all the items and the protagonist appear at the beginning of each story helped to keep the child’s attention and complete the task faster (since each child had to complete 32 trials in addition to several other tasks). Also, I noticed in the pilot study that the children tended to get bored with my explaining what kind of activity the protagonist in each story likes to do, especially the bilingual children who had to do the task twice (in Arabic and English). Thus, starting with all items displayed on the screen and introducing them briefly seemed like a good way to keep children’s attention and concentration till the end.

The experimental design was a 3 (condition: *under-informative, optimal, false*) x 4 (quantifier/operator: *most, some, or, and*) x 5 (groups: children—*Arabic-English bilinguals, Arabic bidialectals, English monolinguals*; adults—*Arabic natives, English natives*). There were 32 items in all—8 for each of ‘most’, ‘some’, ‘or’, and ‘and’ (see appendix 2 for a full list of items). The quantifiers ‘most’ and ‘some’ and the disjunction ‘or’ are classified as measuring pragmatic ability on the lexical scale, and ‘and’ on the ad hoc scale. Of each set of 8 items, 4 were critical items for testing the children’s ability to penalise under-informative utterances (as explained above). The other 4 items were used as control items to test the children’s ability to reject false (logically and pragmatically wrong) utterances and to accept optimal (logically and pragmatically correct) utterances. An example of a false utterance might be a scenario with 5 bananas and 5 biscuits, where a bear gives his friend, the monkey, 2

of the bananas and none of the biscuits, and when asked ‘What did the bear give to the monkey?’ Mr Kareem responds, *The bear gave the monkey some of the biscuits*. An example of an optimal utterance might be a scenario with 5 bushes and 5 fences, where a goat jumps 2 of the 5 fences and Mr Kareem, when asked ‘What did the goat jump over?’ replies, *The goat jumped over some of the fences*.

To ensure that the children understood the task and, especially, how the 3-point scale worked, I included four training items at the beginning of the task. The training items covered optimal *all* (logically and pragmatically correct) (e.g. *the elephant pushed all the trucks* when the elephant had pushed all of 5 trucks in the context), false *some* (e.g. *the giraffe ate some of the apples* when she actually ate some of the pears), optimal *or* (e.g. *I’m not sure I saw it well; the girl bought the hat or the ring*, when the girl only bought the hat) and under-informative *or* (with an image of a Sun and a moon, Mr Kareem said *In life, every day consists of a day or night*, when he should say a day *and* a night). The reason for using optimal *all* was because I found that all children showed a ceiling effect in the semantic task (experiment 1) when tested on *all*; thus, if they rewarded Mr Kareem with a small or medium strawberry when he used *all* correctly, I could be sure that they still did not comprehend how the scale worked. For those who did this, I asked them why they gave Mr Kareem a small/medium strawberry even though he said that the elephant pushed all the trucks; then, we repeated the scenario, and I said to the child *Can you see, the elephant pushed all the trucks as Mr Kareem said, and as we agreed when Mr Kareem gives us a very good answer we reward him with the large strawberry* (I pointed to the large strawberry on the printed paper). With false *some*, if a child gave a medium instead of a small strawberry, I asked him or her ‘Why did you give him a medium?’ If the child justified his or her response by saying that Mr Kareem was wrong or did not give any justification, then I just reminded the child that for completely wrong answers ‘such as this one’ we give a small strawberry, and moved to the next training item. The same procedure applied to optimal *or* (I will discuss the limitations of optimal *or* in detail in chapter 5), but with under-informative *or*, I just tried to ask a child why they chose the specific size of strawberry they had chosen, to get at the child’s world-knowledge that sensitivity to informativeness here depended on. In response to asking children for the reason they chose a specific size, either in the training or the experiment trials, I either received no justification, or, for a medium or small

strawberry, the justification was always ‘because he was wrong’, for a large strawberry, ‘because he was right’, and for a medium choice, for example, ‘that was not very good’, or ‘he was not completely right/wrong’. Fortunately, all except two of the children comprehended the logic of the 3-point scale after training. With the two non-comprehenders, I repeated the training items to ensure that they understood the scaling (using the strawberries) before starting (the acceptable indication that they understood was if they accepted the optimal ‘all’ and rejected the completely false trail). After the training, all the children received neutral feedback: ‘Thank you’; when starting the task, I did not give any feedback, but at the end of each trial, when child clearly indicated which strawberry they chose, I said ‘OK, let’s watch the next story’. The reason for not praising children for their choices is that I did not want them to think that they had received the praise for choosing the appropriate size, and then overuse that size just assuming that it was the correct one.

To test participants’ pragmatic competence in Arabic, the same scenarios were used, but the character of Mr Kareem was changed to Ms Sara, a non-native but fluent Arabic-speaker. I explained that Ms Sara knew quite a lot of Arabic, but would like to learn to speak Arabic as perfectly as the participant does. Otherwise, exactly the same procedure was used as in the English test, but with Arabic software and stimuli. To address variation among Colloquial Arabic dialects I included only the animals, objects, shapes, and verbs represented by the same lexical items across these dialects as evaluated by the adult raters (in the pilot study). The test took 16–18 minutes, in either language, for children, and roughly 13 minutes with adults.

Experiment 4: Ternary-judgement Pragmatic task (without context)

This experiment was adapted from Noveck (2001). It aimed to test children’s sensitivity to and readiness to penalise under-informative sentences given without context. Although the current study used Noveck’s experiment as a model, the procedure and materials were slightly modified. Specifically, although I asked children and adults who penalised infelicitous items why they did so (as I did in experiment 3), I did not instruct children when starting the task to justify each response, since the children in this study were younger than those in Noveck (2001) and might not be able to justify their responses articulately. Also, children were asked

to listen to statements uttered by the fictional characters that had been introduced in experiment 3, and they were asked to reward the fictional character's statement using the 3-point scale that was used in experiment 3, again with the strawberries. I merely reminded the children how the 3-point scale works and then recorded their responses. The sample in this study required very simple stimuli, as rejection or acceptance of the sentences would not rely only on children's sensitivity to/awareness of the pragmatically enriched meaning of the quantifiers but also on their world-knowledge, which would allow them to evaluate the validity of each statement.

The task design was a 3 (condition: *infelicitous*, *felicitous*, *bizarre*) x 4 (quantifier/operator: *most*, *some*, *or*, *and*) x 5 (group: children—*Arabic-English bilinguals*, *Arabic bidialectals*, *English monolinguals*; adults—*Arabic natives*, *English natives*) design consisting of 32 items: 8 for each of for 'most', 'some', 'or' and 'and' (see appendix 4 for a full list of items). Half of the items for each quantifier/operator were logically true but pragmatically inappropriate (in the *infelicitous* condition, exactly parallel to the under-informative condition in experiment 3; *infelicitous* is used here to differentiate the informativeness conditions of the two experiments, that is, informativeness in experiment 4 depends on a scale and world knowledge, while in experiment 3 it depends on scale and visual context); 2 of the items were both pragmatically felicitous and logically true (in the *felicitous* condition, which was parallel to the optimal condition in experiment 3), and the remaining 2 were logically false and pragmatically *infelicitous* (the *bizarre* condition, parallel to the false condition in experiment 3). It should be mentioned that although the *bizarre* items served the same function as the false items in experiment 3 (to ascertain whether children could reject totally wrong items and were not simply accepting all items), it differed from the false condition in that stimuli in it did not make sense in terms of the child's world-knowledge (e.g. *some flowers can talk*), while in the latter the utterance was only wrong because it referred to incorrect items in the scenario (e.g. *the giraffe ate some of the apples* when she actually ate all the pears). Thus, we expected higher rejection of *bizarre* than false items. Table 3.2 gives examples of stimuli used in each condition for each quantifier/operator. The quantifiers 'most' and 'some' and the disjunction 'or' are classified as measuring pragmatic ability on the lexical scale, and 'and' on the measuring encyclopaedic scale.

Table 3.2. Sample stimuli from the no-context (pragmatic) ternary-judgement task (experiment 4) for each quantifier/operator in the three experimental conditions

Quantifier/ Operator	Infelicitous	Felicitous	Bizarre
<i>Most</i>	Most people have a head.	Most houses have a staircase.	Most chairs can talk.
<i>Some</i>	Some elephants have trunks.	Some people wear glasses.	Some birds have telephones.
<i>Or</i>	Before going out, people wear a left shoe or right shoe.	When writing on paper, people use their left or right hands.	To survive, people can eat books or stones.
<i>And</i>	To clap, you need to use your right hand.	To make a cheese sandwich, you need bread and cheese.	To cut out a circle, you need a computer and a telephone.

All statements were translated into Arabic, and the same procedure was used with the Arabic test. As in experiment 3, all items were validated with Arabic adults. I had to change one item, *some cats have tails*, in Arabic to be *some dogs have tails*, because of the lexical variation in terms for ‘cat’ in colloquial Arabic dialects, while ‘dog’ is expressed by the same word in almost all Colloquial Arabic dialects. None of the adults had reported any other differences, and when piloting the items on children from Palestine, Iraq, Jordan, Oman, Libya, and Saudi Arabia, none of them had revealed any difficulty in grasping the items. As in experiment 3, the children never received feedback on their choice of strawberry, and the task took approximately 8–9 minutes to complete.

3.4.5.2 Cognitive performance

To explore children’s cognitive abilities, I assessed two core components of executive functioning (EF): inhibition and working memory (WM) (Miyake et al., 2000). The rationale for testing these two abilities was twofold: (a) there is established empirical evidence for a bilingual advantage in children in these two components (e.g. Bialystok & Martin 2004; Morales et al., 2013), and (b) there is empirical evidence

that WM is involved in implicature processing (e.g. De Neys & Schaeken, 2007; Dieussaert et al., 2011). It should be noted, however, that the current research only explored visuo-spatial short-term memory (STM), which is not necessarily a valid proxy for WM. Although that is a reasonable criticism, measuring WM was difficult in this study due to the young age of the participating children, some of whom found the less complicated STM task difficult to complete. I think, however, that STM can serve as a good indicator for children's WM, because STM is a core component in the WM model, wherein all resources for WM, phonological memory, and visuo-spatial STM are controlled by the same 'central executive' (Baddeley & Hitch, 1974; Baddeley, 2000). In addition, there is some empirical evidence for the effect of STM on evaluating statements including quantifier expressions (Zajenkowski & Szymanik, 2013).

Inhibitory control test: The Simon task (Simon, 1969)

This task was a computer-based version of the Simon task (programmed and run via E-Prime software) conducted as a measure of cognitive conflict inhibition, or more precisely, to assess children's ability to suppress interference from conflicting stimuli (e.g. Martin-Rhee & Bialystok, 2008; Antoniou et al., 2014). Participants were instructed to press the right arrow key on a keyboard if a green square appeared on the screen and the left arrow key if a red square appeared. To make this task easier for children, I coloured the respective keys green and red. In congruent trials, each square appeared on the same side as the appropriate button to be pressed (e.g. a red square on the left of the screen), while in incongruent trials the square appeared on the opposite side (e.g. a red square on the right of the screen); this created an inconsistency, which had to be inhibited. RT spans stimulus presentation to button press. Figure 3.6 gives a sample of an incongruent trial.



Figure 3.6. Sample trial from the Simon task (incongruent condition) as it appeared on a (13-inch) laptop screen; the relevant keys (on the keyboard) were coloured red (left) and green (right).

The task consisted of 32 trials—14 congruent and 14 incongruent, which were randomly presented, and 4 practice trials, which were excluded from the analysis. The laptop was placed on a child-sized table in front of the child, and I sat next to the children to explain the task and encourage them while they completed the training trials. After the training trials, a cartoon puppet (a dog) appeared on the screen with the word ‘Hurry!’; then, I told the children that the task would start after they pressed the button immediately, and they had to respond as fast as possible. Participants’ performance (RT and accuracy) was recorded by the software. All the children received positive praise after completing the task, which took approximately 3 minutes.

Several studies have employed the Simon task with different numbers of trials; for example, De Cat, Gusnanto and Serratrice (under-review) used a total of 48 congruent and incongruent trials, Antoniou et al. (2014) used 48 congruent trials, 48 incongruent trials, and 48 neutral condition trials (in which the coloured square was presented centrally); in both of these studies there were 8 training trials. Similar to the current research, Morton and Harper (2007) used a total of 28 congruent and incongruent trials, but with only 2 training trials. Although one might assume that increasing the number of trials might help detect a bilingual advantage on inhibition, to the best of my knowledge, there is no evidence for this; unfortunately, none of these studies justified their rationale for choosing a certain number of trials. In setting my own number, I took three issues into consideration: the children’s age, the intensity of the

tasks across the whole testing session, and the preliminary data from the pilot study. Regarding the first two issues, since the children in this study were relatively young and since they had to complete several tasks in a single long session, 28 trials (the lowest number found in previous studies) seemed more appropriate than a higher in terms of keeping their attention. In addition, a larger number of trials might be expected to have a negative effect even if children were completing the Simon task on its own, without other tasks, by causing tiredness or boredom and thus not helping to detect the child's inhibition ability. The third issue that led me to use a low number of trials (as in Morton & Harper, 2007) was that the preliminary results and feedback I received from the first child in my pilot study indicated that he was not interested in completing a relatively long task; this child's task used a child-friendly version of the Attentional Network Test (ANT) developed by Rueda et al. (2004), with 96 items (it used fish instead of arrows to make it more friendly for children). When completing the task, this pilot child participant asked consistently when the game was going to finish. This experience led me to adopt an alternative measure with a lower number of trials that assessed a specific component of attention (inhibition)—the Simon task.

Short-term memory (STM) test: The Corsi blocks task (Corsi, 1973)

This task was administered as a measure of visuo-spatial STM. I used an iPad (i.e. touchscreen) app version (called PathSpan) of a task created by Darby (2011), downloaded from the iTunes store (running it on a third generation Apple iPad 2). The iPad version was tested on pre-school children, who had no difficulties using it. The iPad screen showed 9 circular frames ('blocks') on a white background, with a sequence of circles flashing onscreen in different frames (see figure 3.7 below). Participants were told to observe the sequence and to reproduce it by touching the circles in the same order. When touched, the circles lit up to confirm that the device had detected the response. After the last circle flash occurred in each trial, the 'Done' button in the bottom-right corner immediately turned red; they were instructed to press 'Done', and then, press 'Play' in the bottom left corner, so that the next trial would start.

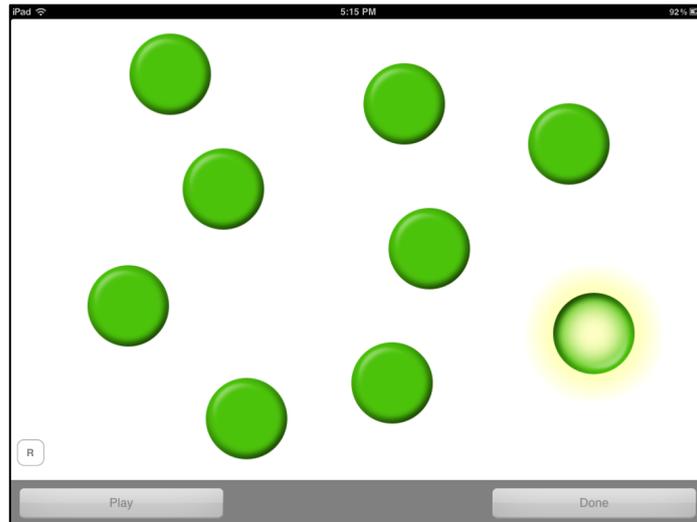


Figure 3.7. Sample display from the Corsi blocks task (on Apple iPad 2, 9.7-inch screen)

At the start of the task, I explained the process it would follow; then, we conducted 2 practice trials. Then, in the main task, a 2-flashing-circle sequence, which increased by one circle if the participant correctly remembered the sequence in one out of three trials undertaken. For each sequence, a participant had 3 trials; if the participant got any of them incorrect, the task was terminated. The participants' *length span* was defined as the longest sequence (up to 9) the participant could repeat correctly in one trial.

The task only included a forward condition (participants were only required to recall the sequence of circles flashed in the same order), not a backward condition as in the traditional Corsi blocks task, (which requires participants to repeat the sequence in reverse order as a measure of WM), because in the forward condition some children were only able to repeat a 2-block sequence, indicating that the task was difficult for them, and only a few children could reach a 6-sequence, so I avoided the backward condition. Also, I used a direct-input touchscreen instead of the traditional (manual) version or the computerised version (indirect input with a mouse) of the task, for two performance-related factors. First, the touchscreen allows the participant to interact directly with the task, and demands less hand-eye coordination than using a mouse (Shneiderman, 1991). Intuitively, this could make the task easier for the children, who might be less competent in using a computer and mouse compared to adults. Second, Robinson and Brewer (2016) also report that touchscreens offer a higher level of

engagement and may positively affect other behavioural and physical aspects, such as the participant's exhaustion and mood, which may in turn impact the cognitive resources required to complete the task. However, studies comparing adults' performance on different versions of the task have found no empirical evidence to support these claims. For example, adults tested on the traditional (manual) version versus the touchscreen version of the Corsi blocks performed approximately equivalently on length span (Brunetti, Del Gatto, & Delogu, 2014; Robinson & Brewer, 2016). These findings, however, should not be generalised to children, who have more limited cognitive resources, and for whom the benefit of using one version rather than the other could therefore be crucial to providing more accurate measures of their memory span. Hopefully, similar comparative research to that reported above will be conducted with young children in future to identify any differences in their performance on the different versions of the Corsi task.

3.5 Summary

This research consisted of two studies. The first study explored children's semantic comprehension of the English quantifiers *all*, *most*, *some*, and the operators *or*, *and* and their Arabic equivalents, with a perception task and a production task. It also explored children's acquisition of a numerical system, with four tasks (how-many, give-a-number, non-verbal ordinal, and estimating-magnitude-numerically). In this way, it investigated empirically the relationship between quantification and numeracy skills. The second study examined children's pragmatic ability to detect the violation of informativeness (with two pragmatic tasks: enriched context v. no context) and also assessed cognitive skills that might be involved in the process of implicature derivation: inhibition (using the Simon task) and STM (using the Corsi blocks task). Its goal was to find out if bilingual children would pragmatically outperform Arabic- and English-speaking children, and if so, whether their superior pragmatic performance could be interpreted in terms of a bilingual cognitive advantage. This chapter has also explained the measures used to control several potential extraneous, confounding variables that could affect children's performance undesirably: language proficiency (measured by receptive vocabulary), language exposure (measured by background language questionnaire), general mental ability (measured by NVIQ), and for SES (measured by parental education and family wealth questionnaire from the

FAS). The next chapter, chapter 4, presents the results for participants' performance in the various tasks.

Chapter 4

Results and Analyses

4.1 Introduction

This chapter provides descriptive and inferential analyses of Arabic–English bilingual, Arabic bidialectal, and English monolingual children’s performance on the semantic, pragmatic, numerical, and cognitive tasks in order to answer the research questions that were introduced in chapter 1 of this thesis. It might be useful, before starting the analyses, to repeat these questions here: (a) Do bilingual and monolingual children comprehend the quantifiers ‘most’ and ‘some’ and the operators ‘or’ and ‘and’ in a semantically appropriate way? (b) To what extent does the acquisition of a numerical system promote or possibly hinder the acquisition of quantifiers? And (c) Can any superior pragmatic competence in bilingual children be explained in terms of a cognitive advantage over monolinguals? In addition to exploring the children’s results, the analyses also consider the performance of adults, who act as controls for baseline comparison since they represent the developmental endpoint of this acquisition process (Hanlon, 1987). The adult control consisted of an Arabic group and an English group. Their performance on the various tasks is described at the end of the respective sections, after presenting the children’s results on each task.

The structure of the chapter is as follows. The first section presents information on the participants’ background characteristics; beside basic information on participants’ age and gender, the section includes the results of the measures I used to control for potential confounds such as general mental ability, language proficiency, and SES. The second section presents the results of study 1, and it is divided into two main parts. The first part aims to answer the first research question; thus, it presents results for the participants’ semantic performance (in the give-a-quantifier and estimating-magnitude-proportionally tasks). The second part aims to answer the second research question (of study 1) by first presenting the performance results for the four number tasks (the how-many, give-a-number, non-verbal ordinal, and estimating-magnitude-numerically tasks), and then exploring the potential relationship between children’s comprehension of quantifiers and their numeracy skills. Section three presents the

results of study 2 and it is divided into two main parts. The first part provides analyses of the participants' performance in the two pragmatic experiments (enriched context and no context), followed by a comparison of children's performance in these two experiments. Then, a new procedure for analysing the ternary responses in the pragmatic task is given in the fifth section, and applied to the performance of the children on the critical (under-informative) items in the two pragmatic experiments. After the participants' performance on the pragmatic tasks is examined, in order to partially answer the third research question, in the second part, the participants' performance on the two cognitive tasks (the Corsi blocks task and the Simon task) is explored, also in part to answer the third research question but also to investigate which variables might explain variation in children's cognitive performance (as reflected in the cognitive tasks). Finally, the analyses of study 2 attempt to provide a complete answer to the third research question by exploring the potential relationship between children's pragmatic performance on the quantifiers/operators (most, some, and, or) and their performance on the two cognitive tasks.

Each section is followed by a brief summary of the main findings for the relevant task(s), and at the end of the chapter, a summary of the overall results is given.

4.2 Background characteristics

This section gives information on the number of participants in each groups, their mean age, and gender and language background (spoken language and receptive vocabulary score), for child and adult participants. In the subsection on child participants, 4.2.1, further information is given on the children's general mental ability, SES, and for the bilingual and Arabic children, language exposure and use, since those two groups had been exposed to more than one language or dialect. A summary of the findings is given at the end of the section.

4.2.1 Child participants

4.2.1.1 General measures

Table 4.1 provides descriptive information on each group of child participants. The bilingual and Arabic groups were matched for number of participants and gender,

each containing 30 participants: 17 male, 13 female. The English group included 26 participants (11 male, 15 female).

Table 4.1. Background characteristics of bilingual, Arabic-speaking, and English-speaking children

Group	N	Gender	Acquired Language	Linguistic situation	Age	NVIQ Raw	NVIQ standardised
Bilingual	30	Female 13 Male 17	English Arabic	Bilingual	M 5;6 SE 0.1 range 4;1–7;0	M 15.27 SE 0.64	M 10.8 SE 0.38
Arabic-speaking	30	Female 13 Male 17	Arabic	Monolingual or bidialectal	M 5;6 SE 0.1 range 4;7–6;9	M 15.93 SE 0.29	M 10.97 SE 0.32
English-speaking	26	Female 15 Male 11	English	Monolingual	M 5;7 SE 0.09 range 4;3–6;2	M 16.61 SE 0.53	M 11.57 SE 0.41

Comparison between groups' age

The mean age of participants was matched across the three groups (around 5;6 years old), but the age range differed slightly across groups. As table 4.1 shows, the bilingual group's age range was 4;1–7 years old, although it should also be mentioned that there was only one child aged 7 and only one aged 4;1; all other children in this group were aged between 4;4 and 6;9 years old. The English and Arabic groups had approximately similar age ranges, with the Arabic children slightly older, as shown in table 4.1. It should also be noted that in the Arabic group there were only two children aged around 4;6 years old, and only one child aged 4;3 in the English group. Thus, the majority of children in all three groups fell in the age range 5–6;3 years old.

Before exploring whether the groups differed in age to a statistically significant degree, I ran a normality test to find out if the children in each group and in the whole sample were normally distributed; this information is essential to determine which test should be applied to compare the groups' age. The Shapiro–Wilk normality test for a small sample (under 50 participants) was used first, to test the normality for each group. The results revealed that the ages of the bilingual and Arabic child groups were

normally distributed ($p > 0.05$), while the English child group was not ($p = 0.035$). For the sample as a whole, I used the Kolmogorov–Smirnov test of normality, since it works better with larger samples (more than 50); the results revealed that the whole sample was normally distributed ($p > 0.05$).

On the basis of the normality test (Kolmogorov–Smirnov), I conducted a one-way analysis of variance (ANOVA) to compare if the groups differed significantly in age. The ANOVA revealed no significant differences in age between groups ($F(2, 115) = 0.35, p > 0.05$), and post hoc multiple comparisons also showed no significant differences between groups ($p > 0.05$). To secure more confidence in the assumption of insignificance of differences in age between groups, especially the English group, which violated the assumption of normality, I ran pair-comparisons between groups using a non-parametric test (Mann–Whitney); the results confirmed the insignificance of differences between groups found by the ANOVA test.

Comparison between groups' NVIQ

Table 4.1 above has presented the NVIQ standardised and raw mean scores for all the groups. It can be seen that the three groups have relatively similar mean scores, with raw means around 15 and standardised ones around 10. In all the analyses, I used the standardised scores, which take the child's age into account.

Before conducting comparisons between groups to find any statistical differences, I first tested the normality of each group, and then of the whole sample, in the same way described for age above. The Shapiro–Wilk normality test revealed that the Arabic and English children were normally distributed ($p > 0.05$) but that the assumption of normality was violated for the bilingual children ($p = 0.001$). Next, I tested the normality of the whole sample using the Kolmogorov–Smirnov test, which showed a significant violation of normality ($p < 0.001$). Therefore, I applied a free-distribution (Mann–Whitney) test to compare the groups' NVIQ. The pair-comparisons revealed no statistically significant differences between the bilingual and Arabic children ($U = 379, Z = -.79, p = 0.43$), the bilingual and English children ($U = 285, Z = -1.74, p = 0.082$), or the Arabic and English children ($U = 321, Z = -1.15, p = 0.25$).

4.2.1.2 Child participants' socio-economic status (SES)

As this study involves an international and multi-cultural sample, the participating children's SES was assessed using a composite method, incorporating two proxy indicators: parental educational attainment and family wealth (as an indication of family income). The former indicator was assessed by asking parents about the highest level of education they had obtained, while the latter was assessed by asking them to complete a four-item questionnaire on family wealth. The two subsections below present the information related to these two indicators.

Parental education

Parents were asked to state the highest educational level they had attained; responses were classified according to UNESCO International Standard Classification of Education (ISCED) mappings for each parent's country. Then, I applied Sutherland's (2012) method of parental education classification, in which he selected the highest degree achieved by either parent as the level of parental education. Information on parental education was gathered for all children. Table 4.2 below provides information on parental level of education in each group.

Table 4.2. Parents' level of education in each child group, classified according to the UNESCO international education scale

UNESCO Scale	Bilingual children			Arabic children			English children		
	Mother (N)	Father (N)	Parental level (N)	Mother (N)	Father (N)	Parental level (N)	Mother (N)	Father (N)	Parental level (N)
1	0	1	0	0	1	0	0	0	0
2	2	1	0	0	0	0	0	0	0
3	0	0	0	3	8	2	6	7	3
4	0	0	0	0	0	0	0	0	0
5	26	20	20	26	19	26	19	16	21
6	2	8	10	1	2	2	0	2	2
Total	30	30	30	30	30	30	25*	25*	26

Note: the symbol (*) means that one of the child participants in the English group had only one parent's education level given in the questionnaire (that is, this child was raised by a single father/mother).

As can be seen, the table first shows the total number of parents classified under each UNESCO educational level in each group, and then gives the final level selected (which was the highest level between the two parents).

The classification in table 4.2 was applied using the published UNESCO education mappings (2011) for each country from which participants hailed: Libya, Iraq, Saudi Arabia, and the United Kingdom (available online as a Microsoft Excel spreadsheet, <http://uis.unesco.org/Education/ISCEDMappings/Pages/default.aspx>). The table shows that all the bilingual children's parents had completed a tertiary level of education (Levels 5 [bachelor's degree] and 6 [post-graduate degree] in the scale). It can also be seen that for the majority of Arabic children (28/30) and English children (23/26), one parent had attained the first stage of tertiary education (Level 5, considered the first 'high' level), with only two (Arabic) and three (English) children's parental level of education falling at Level 3 (medium level, equivalent to either General Certificate of Secondary Education (GCSE) or A-levels in the English group, and to secondary school qualification in the Arabic group).

Family wealth

Family wealth was assessed using the Family Affluence Scale (FAS) questionnaire (Currie et al., 1997). A composite FAS score was computed for each child based on parents' responses to the four-item questionnaire. Figure 4.1 shows the mean FAS score for each child group; it can be seen that the groups have very similar mean scores.

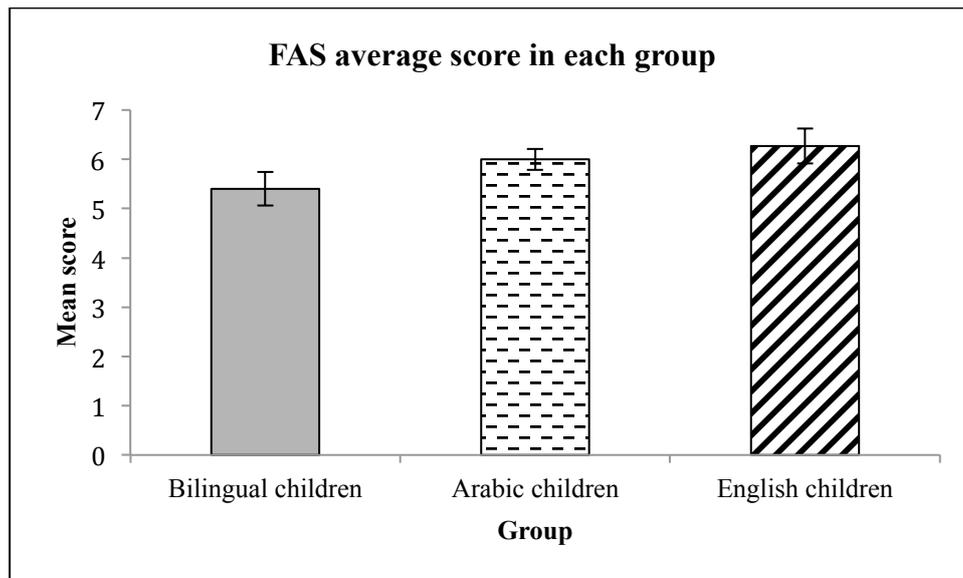


Figure 4.1. Results for family wealth using the Family Affluence Scale (FAS). Error bars represent standard error of the mean

At the individual level, following Boyce et al. (2006), FAS was grouped into a three-point ordinal scale, as low affluence (score=0, 1, 2), medium affluence (score=3, 4, 5) or high affluence (score=6, 7, 8, 9). Table 4.3 presents individual FAS within each group.

Table 4.3. Child participants' SES by the international Family Affluence Scale

FAS	Bilingual children	Arabic children	English children
Low (0, 1, 2)	(FAS 2) N=3	0	(FAS 2) N=1
Medium (3, 4, 5)	(FAS 3) N=1	(FAS 4) N=4	(FAS 3) N=2
	(FAS 4) N=7	(FAS 5) N=6	(FAS 5) N=5
	(FAS 5) N=3		
High (6, 7, 8, 9)	FAS 6) N=6	(FAS 6) N=8	FAS 6) N=5
	(FAS 7) N=6	(FAS 7) N=10	(FAS 7) N=6
	(FAS 8) N=4	(FAS 8) N=2	(FAS 8) N=5
			(FAS 9) N=2
Total	30	30	26

It can be noted that approximately half of the bilingual children can be classified as having high FAS, and the other half as medium FAS; only 3 children fall into the low FAS category. The majority of the Arabic children (around two-thirds) can be

categorised as high FAS, and the remaining third as medium FAS. The same generally applies to the English children, with (in addition) one child having a low FAS.

To investigate whether these differences between groups were statistically significant, I first categorised the results according to the criteria given in table 4.3 (using the 3-point scale: high, medium, low). The ordinal scale was coded 3, 2, 1 for high, medium, and low, respectively. Since the data were categorical, I used a non-parametric test (Mann–Whitney) to compare the groups' FAS. The pair-wise comparison did not reveal any significant difference between the bilingual and Arabic children ($U=375$, $Z=-1.29$, $p=0.19$), the bilingual and English children ($U=336$, $Z=-1.01$, $p=0.31$), or the Arabic and English children ($U=380$, $Z=-.199$, $p=0.84$).

To explore whether the two SES indicators were correlated, I conducted a bivariate correlation test. The results revealed that parental education level significantly correlated neither with the raw FAS score ($r(\text{two-tailed})=.009$, $p=0.93$) nor with the categorical FAS score ($r(\text{two-tailed})=.142$, $p=0.19$). This was expected, because the majority of the children had at least one parent with a tertiary level of education but their family wealth as measured by the FAS questionnaire varied. Thus, having a high level of education was not necessarily reflected in the FAS score and vice versa. Since the variations were more obvious in the FAS questionnaire results, the FAS score, as an indicator of SES, would be used as predictor when conducting the regression analysis.

4.2.1.3 Language measures

This subsection presents the results for two types of language assessment. First, it presents the results of a receptive vocabulary test given to all participating children. Then, it shows the results of the language use questionnaires that were completed by the bilingual and Arabic children's parents (unlike the Arabic children, who had been exposed to two dialects, the English children are excluded from this part of the analysis, as they were purely monolingual and had had no exposure to any other language variety than English).

Children's receptive vocabulary

The average values for children's raw scores in the British Picture Vocabulary Scale (BPVS) (Dunn & Dunn, 2009) and its Arabic-translated version are given in figure 4.2 below. They reveal that English children had the highest average raw vocabulary score (around 77); the Arabic children's average raw score was around 66, in the middle; and the bilingual children scored lower than these two groups in both English and Arabic (around 54 and 46, respectively). The bilinguals' average score in English was slightly higher than their score in Arabic.

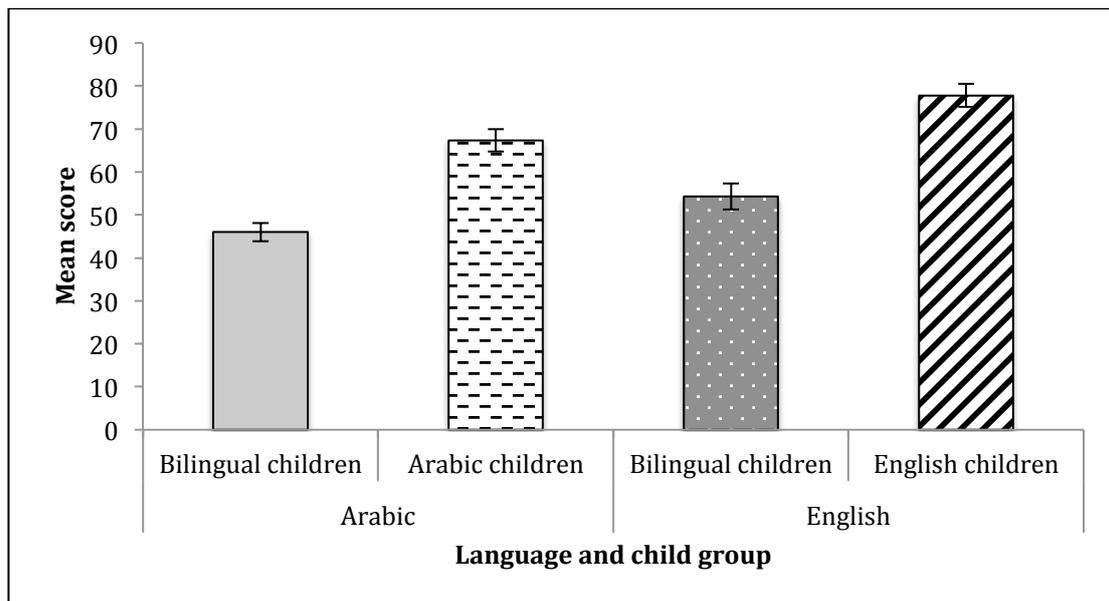


Figure 4.2. Children's average raw scores in the receptive vocabulary test. Error bars represent standard error of the mean

To explore the differences between the groups further, I examined the assumption of normality first for each group and then for the whole sample. The Shapiro–Wilk test revealed that all the groups' vocabulary scores (bilingual-English, bilingual-Arabic, Arabic children and English children) had normal distributions (all $p > 0.05$).

A Kolmogorov–Smirnov test of the normality of the whole sample revealed that it met the assumption of normality ($p > 0.05$). Thus, comparison between groups was conducted using the parametric comparative ANOVA, with group as a factor and raw score as an independent variable. The test results revealed a significant difference between the groups ($F(3, 112) = 28.04, p < 0.001$). Post hoc multiple comparisons (using the Games–Howell test, as the samples were not equal) showed strong,

significant differences between the bilingual-English scores and the English children's scores ($p < 0.001$), and also between the bilingual-Arabic scores and the Arabic children's scores ($p = 0.009$). The outcomes further revealed a significant statistical difference between the Arabic and English children ($p = 0.033$): the Arabic children had a lower vocabulary score, which might be attributed to the diglossic nature of Arabic, as this has been found to have a negative effect on children's vocabulary (Fedda & Oweini, 2012). The results showed that the difference between bilingual children's vocabulary scores in the two languages was not significant ($p = 0.119$).

Bilingual and Arabic children's language experience

The results of the questionnaire were recorded in an Excel file which incorporated a number of algorithms allowing the current study to estimate the children's language use and exposure to given languages (Arabic and English for bilingual children, Standard and Colloquial Arabic for Arabic children). The algorithms quantitatively measure (a) the amount of input/output the child has for a specific language at that time, and (b) the amount of input/output the child has had to a given language over time (the *cumulative length of exposure*). The results are given as proportions of yearly exposure to each language (percentage of input to a child) and use of these languages (percentage of output by a child). For the bilingual children, Arabic is the language that they speak at home and usually within their community (the British Arab community), so it was considered the first (home) language (or L1), while English is the language they spoke at school, and thus considered the second (target) language (or L2). Similarly, for Arabic children, their local (Colloquial) Arabic dialect, which they use at home and with friends and anyone outside school, was described as the first (home) dialect (D1), while the Standard Arabic dialect, which they use at school, was considered the second (target) dialect (or D2).

The analytic procedure used in this subsection is as follows: First, the bilingual and Arabic children's experience of the languages or dialects they have been exposed to is investigated by exploring their amount of input/output in each language or variety they speak, as given in the questionnaire completed by their parents. Then, the analyses focus on the amount, length, and onset (initial age) of exposure and explore

its potential effect on children's receptive vocabulary. This is because previous studies have found that the amount, onset, and length of input to a language are important factors in children's acquisition of a language (Unsworth et al., 2014; Thordardottir, 2011; Pearson, Fernández, Lewedeg, & Oller, 1997; Schiff & Ventry, 1976).

Results of language questionnaire

Figure 4.3 below shows the average percentages of language use for the bilingual children (Arabic v. English) and Arabic(-bidialectal) children (Standard v. Colloquial Arabic). It can be seen that bilingual children have more exposure to Arabic (around 60%) than to English (around 47%), however, they tend to use English (55%) more than Arabic (49%). Such results might indicate that not all bilingual children participating in this study can be described as balanced bilinguals; the variation in (bilingual and Arabic) children's exposure will be investigated further later in this section.

With respect to the Arabic children, the information given by their parents shows that the children have Colloquial Arabic as a dominant dialect, at 80% for both exposure and use; Standard Arabic represents only 20% of these children's language experience (used for educational purposes).

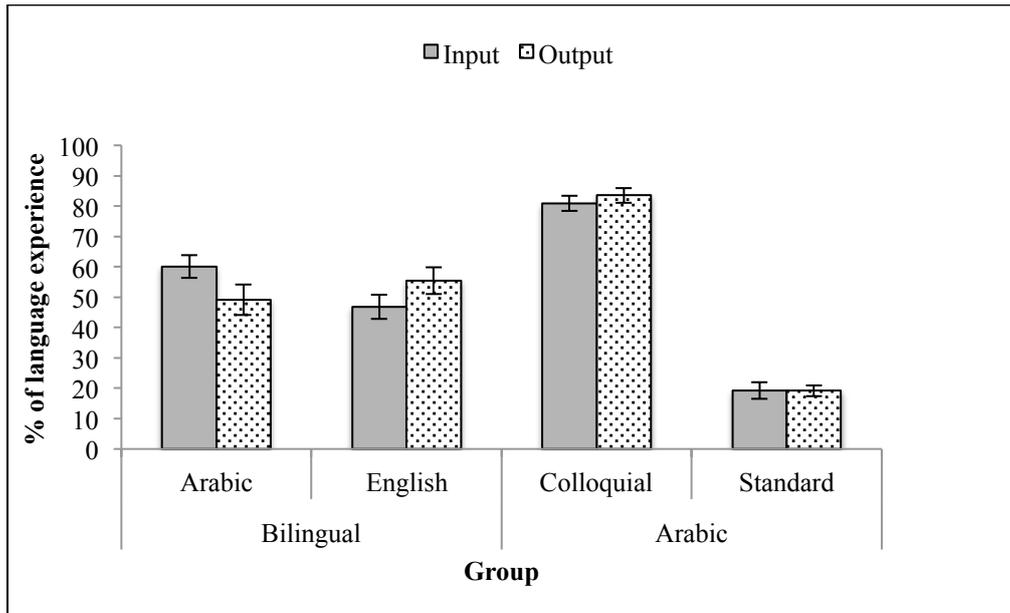


Figure 4.3. Average ratio of Arabic–English use in the bilingual child group and Colloquial–Standard Arabic in the Arabic child group. Error bars represent standard error of the mean

Critical amount of language exposure

As amount of exposure has been found to be an important quantitative factor in language acquisition in previous research, which has suggested that children may require at least 20% exposure to a language in order to produce utterances in that language spontaneously (Pearson et al., 1997; Schiff & Ventry, 1976), the individual amount of exposure for each child was explored further.

Table 4.4. Number of bilingual and Arabic children who have limited exposure to either of the two languages/dialects

Amount of exposure	Bilingual children's exposure		Arabic children's exposure	
	Arabic	English	Colloquial Arabic	Standard Arabic
Less than 20%	N=2	N=1	0	N=9
Less than 10%	0	0	0	N=10

Table 4.4 shows the number of children whose exposure to a given language or dialect was less than 20%. It can be seen that only 2 bilingual children have less than 20% exposure to Arabic, and only one bilingual child has less than this amount in

English. In contrast, the Arabic children's results reveal that approximately 9 children have less than 20% Standard Arabic input, and 10 children have less than 10%. Based on the finding in previous work that a child may require at least 20% exposure to a language in order to develop appropriate linguistic skills in that language (e.g. Pearson et al., 1997; Gutiérrez-Clellen & Kreiter, 2003), the results for the Arabic children participating in this study might indicate that the majority of those children had not reached the critical amount of exposure expected to affect their performance, especially in terms of EF abilities.

Bilingualism v. bidialectalism: Language input and output

To find out how the bilingual and Arabic groups' language experience might statistically differ, I first tested the assumption of normality for the whole sample and within each group. Then, I conducted comparisons between the two groups' amount of exposure (input) and amount of use (output) for their first and second language/dialect.

Language input

First, the two groups' exposure to their first language/dialect (Bilingual L1: Arabic; Arabic children D1: Colloquial Arabic) was checked against the assumption of normality. The test revealed that the distributions for the whole sample and within each group were normal ($p > 0.05$). Since the sample was normally distributed, I used one-way ANOVA, with group as factor and L1/D1 input as a dependent variable. The test revealed no significant differences between the groups ($F(1, 58) = .009, p = 0.92$).

Next, I explored the normality for the whole sample and within the groups for the amount of input in the second language/dialect (Bilingual L2: English; Arabic children D2: Standard Arabic). The Shapiro–Wilk test revealed that normality for the Arabic children's D2 input was violated ($p = 0.002$), whereas the bilingual children met the assumption of normality in their L2 input ($p = 0.14$). The test of normality for the sample L2/D2 input (Kolmogorov–Smirnov) showed a violation of the assumption of normality ($p = 0.003$); therefore, comparison between the groups was performed using a distribution-free test (Mann–Whitney), whose results revealed a significant difference between the two groups ($U = 241, Z = -3.1, p = 0.002$), with the Arabic

children having a significantly more limited amount of exposure to their D2 (Standard Arabic) compared to the bilingual children's exposure to their L2 (English). These results indicate that quantitatively, the bidialectal children's language experience is not similar to that of the bilinguals and thus we might describe the bidialectal children in this study as functionally monolinguals.

Language output

Similar to the above analyses, I first tested the assumption of normality for the whole sample and within each group in use of first language/dialect. As expected, the normality test (Shapiro–Wilk) revealed that the Arabic children's use of colloquial Arabic was not normally distributed ($p=0.004$), whereas the use of Arabic within the bilingual children was almost normally distributed ($p=0.083$). The sample, as expected, violated the normality hypothesis (using Kolmogorov–Smirnov; $p<0.001$). Thus, groups were compared using the Mann–Whitney test, which revealed a significant difference between the two groups' use of their L1/D1 ($U=272$, $Z=-2.63$, $p=0.009$), with the Arabic children having a significantly higher amount of output in their D1 (Colloquial Arabic) compared to the bilingual children's output in their L1 (Arabic).

After this, I explored the two groups' use of their L2/D2. Similar to the above findings, the hypothesis of normality (Shapiro–Wilk) for the Arabic children's use of Standard Arabic (D2) was violated ($p=0.001$), whereas the bilingual children's use of English (L2) was normally distributed ($p=0.17$). Thus, the sample evidently does not meet the assumption of normality (Kolmogorov–Smirnov; $p<0.001$); and indeed, the comparison between the two groups showed a significant difference ($U=112$, $Z=-5$, $p<0.001$), with the bilingual children having significantly higher amount of use of their L2 (English) compared to the Arabic children's use of their D2 (Standard Arabic).

Length of exposure

Table 4.5 provides the average age when first exposed and length of exposure of the bilingual and the Arabic children for their first and second language/dialect. Following Unsworth et al. (2014), cumulative length of exposure was calculated by first subtracting age at onset from chronological age (age at testing), then multiplying the result by the amount of exposure (input), while traditional length represents the number of years in which a child has had exposure to a particular variety, without taking into account the amount of exposure. The two formulae below schematise how length of exposure was computed.

(Traditional length of exposure=Chronological age–Age at onset)

(Cumulative length of exposure=Traditional length of exposure*Amount of input)

Table 4.5. Average age at first exposure to languages/dialects in bilingual and Arabic children

Child Group	Age at onset	Chronological age (Age at testing)	Length of exposure	
			Traditional length	Cumulative length
Bilingual-English (N=30)	M 2;2	M 5;6	M 3;5	M 1;9
	SE 0.27	SE 0.14	Se 0.32	SE 0.29
Bilingual-Arabic (N=30)	0	M 5;6	M 5;6	M 3;3
	0	SE 0.15	SE 0.15	SE.20
Arabic-Standard (N=30)	M 2.8	M 5;6	M 2;8	M 0;6
	SE 0.28	SE 0.10	SE 0.30	SE 0.12
Arabic-Colloquial (N=30)	0	M 5;6	M 5;6	M 4;5
	0	SE 0.10	SE 0.10	SE 0.16

From the table, it can be seen that the average age at onset of exposure to English for the bilingual children was slightly younger than the average age of onset of Arabic children's exposure to Standard Arabic. Since the two groups were of the same chronological age, this obviously resulted in the bilingual children's having a greater traditional length of exposure for the target language than Arabic children. Also, as the bilingual children generally had more input in English (L2) than the Arabic children in the Standard dialect (D2), as revealed in the analyses above (Language output), they also had longer cumulative exposure to the L2.

Individual variation in length of exposure

At the individual level, table 4.6 shows variation in children's age at onset of exposure to the second language/dialect. It can be noticed that approximately two-thirds of children in each group were exposed to the L2/D2 at the age of 2 or older, and only one-third from birth or before the age of 2.

Table 4.6. Individual results for first exposure to the second language/dialect in bilingual and Arabic groups

Age at onset (Year)	Bilingual children (N)	Arabic children (N)
0 (from birth)	6	3
1	4	3
2	6	7
3	7	5
4	7	8
5	0	4
Total	30	30

As the children participating in this study were relatively young and had not yet passed out of the critical period for language learning, age at onset might not have had a direct effect on their performance (Unsworth et al., 2014). However, knowing age of onset becomes rather important when it comes to calculating either traditional or cumulative length of exposure to the L2/D2, and it was therefore investigated here.

Bilingualism v. bidialectalism: Age at onset and length of exposure

To find any differences in age between groups at onset of the L2/D2, I first tested the normality of distributions for the groups and the full sample. The test (Shapiro–Wilk) revealed a violation of normality within the Arabic children ($p=0.03$) and the bilingual children ($p=0.003$), meaning that the overall sample was also not normally distributed (Kolmogorov–Smirnov; $p=0.001$). A comparison between the two groups’ age at onset revealed no significant differences ($U=346$, $Z=-1.56$, $p=0.12$).

Second, to explore differences between the two groups’ traditional length of exposure to L2/D2, I first ran a normality test (Shapiro–Wilk) which revealed that while the Arabic and bilingual children were normally distributed ($p>0.05$), the whole sample slightly violated the assumption of normality (Kolmogorov–Smirnov; $p=0.043$). The comparison between the two groups showed no significant difference with the parametric test ($t(58)=1.484$, $p=0.143$) or the non-parametric test ($U=360$, $Z=-1.33$, $p=0.18$).

After testing the normality of the distribution, I explored the differences in cumulative length of exposure to the L2/D2 between groups. The results revealed that both the bilingual children and the Arabic children violated the hypothesis of normality (Shapiro–Wilk; $p < 0.005$) and consequently that the sample was not normally distributed (Kolmogorov–Smirnov; $p < 0.001$). The difference between the two groups was very significant ($U = 164$, $Z = -4.23$, $p < 0.001$), with the bilingual children having the highest cumulative length of exposure to the L2.

I also explored the differences between the two groups' cumulative length of exposure to the L1/D1. The test of normality revealed that the bilingual and Arabic children (Shapiro–Wilk) as well as the whole sample (Kolmogorov–Smirnov) were normally distributed ($p > 0.05$); therefore, I conducted a one-way ANOVA to compare the two groups. The test revealed a very significant difference between groups ($F(1, 58) = 21.3$, $p < 0.001$), with the Arabic children having the highest cumulative length of exposure to the D1 than the bilingual to the L1.

Effect of length of exposure on receptive vocabulary

In this subsection, I looked for any effect of length of exposure to a second language/dialect on the children's receptive vocabulary score in that language/dialect. I present the results by group.

Effect of exposure on bilingual children

I remind the reader that the bilinguals were exposed to Arabic from birth, so the age of exposure onset only varies with respect to English. First, I examined the effect of exposure to English on the bilinguals' vocabulary scores in English. I ran a bivariate correlation test using vocabulary score in English, age at onset of exposure to English, language input in English, and cumulative and traditional length of exposure. The test results revealed strong positive correlations between vocabulary score in English and cumulative length of exposure ($r(\text{two-tailed}) = 0.59$, $p = 0.001$), traditional length of exposure ($r(\text{two-tailed}) = 0.58$, $p = 0.001$), amount (proportion) of input ($r(\text{two-tailed}) = 0.503$, $p = 0.005$), and chronological age ($r(\text{two-tailed}) = 0.76$, $p < 0.0001$); in addition, English vocabulary scores negatively and significantly correlated with amount of input in Arabic ($r(\text{two-tailed}) = -0.379$, $p = 0.039 < 0.05$).

A linear regression test was performed with vocabulary score as a dependent variable and each of age at onset of English and cumulative length of exposure as independent variables (I removed traditional length and age due to their strong correlation with cumulative length). The test revealed that the best predictors of children's vocabulary scores are age at onset ($t=4.53$, $p=0.025$) and cumulative length of exposure ($t=3.64$, $p<0.001$). The model was significant ($F(2, 27)=11.5$, $p<0.001$), explaining 46% of the variation in bilinguals' vocabulary in English.

Second, I explored the effect of exposure to Arabic on bilinguals' vocabulary score in Arabic. The bivariate correlation test did not reveal any significant correlation between the bilingual children's score in Arabic and either cumulative length of exposure to Arabic or language input in Arabic; nor were their scores negatively correlated with English input or onset of exposure to English ($ps>0.05$). There was however, a marginally significant correlation with chronological age ($r(\text{two-tailed})=.36$, $p=0.053$). A linear regression model also revealed an effect of age on bilingual vocabulary score in Arabic ($t=2.17$, $p=0.053$), which was not improved by adding any other predictors, and the model was only marginally significant ($F(1, 28)=4.1$, $p=0.053$), accounting for 13% of the variation in the bilinguals' vocabulary in Arabic.

Taken into consideration that the majority of the bilinguals have more input in Arabic than in English, and that only age at onset and cumulative length of exposure in English (which reflects the amount of exposure) affect the bilinguals' English vocabulary score, while amount of input in Arabic has no effect on the bilinguals' Arabic vocabulary score, such results might indicate that relying on the amount of exposure alone might not always lead to a higher level of proficiency, nor might it accurately reflect vocabulary level. In other words, it is not only the quantity of exposure that might play an essential role in vocabulary development, but also potentially the quality of exposure (e.g. reading books, educational materials, and having intensive conversations with a child).

Effect of exposure on Arabic children

To investigate the potential effect of the Arabic children's exposure to a second dialect on their vocabulary scores in Colloquial Arabic, I first ran a bivariate correlation test, which revealed no significant (positive or negative) correlation between receptive vocabulary and any of chronological age, cumulative or traditional length of exposure to Standard Arabic, or age at onset ($p > 0.05$). Despite this, I fitted a regression model (using the backward method); none of the models was statistically significant ($p > 0.05$). These results might indicate that this group was functionally monolingual in Colloquial Arabic.

4.2.2 Adult participants

Table 4.7 provides information on the adult control group, consisting of 10 Arabic-speaking adults and 11 English-speaking adults. The groups had the same average receptive vocabulary (exhibiting an appropriate ceiling effect (97%)) and approximately were the same age (18–24 years old). However, while the number of male and female participants in the Arabic group was equal (5/5), there was only one male participant in the English adult group.

Table 4.7. Background features for the Arabic and English adults

Group	N	Gender	Native language	Linguistic situation	Age	Vocabulary (raw score)
Arabic adults	10	Female 5 Male 5	Arabic	Monolingual or Bidialectal	M 22;2 SE 0.7 Range 18–24;9	M 163 (97%) SE 1
English adults	11	Female 10 Male 1	English	Monolingual	M 20;7 SE 0.6 Range 18;3–24	M 163 (97%) SE 1

Although the table describes the Arabic adults as monolingual or bidialectal and the English adults as monolinguals, all the adult participants reported that they spoke an additional language (for the Arabic adults, aside from the Arabic Standard and Colloquial dialects) but without intensive daily use or being fluent in it. However, since this research aims to find and compare the endpoints of children's language

acquisition and not to compare the effect of adults' acquiring an additional language, adults' being learners or non-fluent speakers of another language should not affect the results.

4.2.3 A summary of background measures

The analyses in this section sum up and discuss the background characteristics of the child participants, and then briefly explore basic information on the adult participants.

There were 86 child participants divided into three groups: 30 Arabic–English bilinguals, 30 Arabic-speakers, and 26 English-speaking children. The mean age for the three groups was very close, and a statistical test revealed no significant differences between the groups. The three groups had very close NVIQ scores, and no significant differences were found. With respect to SES, all the bilingual children and around 90% of the English and Arabic children had a high parental educational level, with the remaining 10% having a medium level. In terms of family wealth, the results of the FAS questionnaire revealed that around half of the bilingual children had high FAS, while the other half had medium FAS (with three children having low FAS). Within the English and Arabic children, the result showed that two-thirds had high FAS, with the remaining one-third having medium FAS (except one English child with low FAS). These two indicators of SES were not significantly correlated.

The language measures revealed significant differences between the groups. First, the receptive vocabulary test revealed that the bilingual children had the lowest vocabulary level in both Arabic and English; these differences were very significant. These results are consistent with previous studies that found a negative effect of bilingualism on children's vocabulary (Siegal et al., 2007, 2009; Antoniou et al., 2014). The results also showed that the English children had significantly larger vocabulary than the Arabic children, which might be attributed to the possible negative effect of diglossia in Arabic on children's vocabulary (Fedda & Oweini, 2012).

The analyses also considered the findings of the language questionnaire completed by the bilingual and Arabic children's parents to measure children's exposure to and use

of two languages/dialects. For the bilingual children, the L1 was Arabic and the L2 English, while for the Arabic children, Colloquial Arabic was the D1 and Standard Arabic the D2. The results of the questionnaire revealed that the bilingual children had more exposure to Arabic than English (60% v. 55%), while the Arabic children had only very limited exposure to Standard Arabic (less than 20%); this limited exposure might suggest that the Arabic children were functionally monolinguals. The analyses explored the effect of length of exposure combined with cumulative amount of input (cumulative length) on the children's vocabulary scores. The results revealed only one significant effect: of bilingual children's cumulative length of exposure to English on their vocabulary score in English. Neither bilinguals' cumulative length of exposure to Arabic nor Arabic children's cumulative exposure to standard Arabic had an effect on their vocabulary score.

Finally, the analyses gave general information on the adult control group, consisting of 10 Arabic adults and 11 English adults who were all between 18–24 years old, and showed appropriate ceiling effect in the vocabulary test (97%).

4.3 Results of Study 1: Children's comprehension of quantifiers and operators and the potential effect of numeracy

4.3.1 Semantic performance

This section explores the child and adult participants' performance on two semantic tasks. The first task (experiment 1) was a give-a-quantifier task, used as a measure of children's perception of the quantifiers 'all', 'most', 'some', and the operators 'or' and 'and' (in English and in Arabic, as appropriate). It also aimed to assess their ability to manipulate sets in relation to a given quantifier in a given context. For example, given a set consisting of 6 apples, how would a child react when asked to put 'some' of the apples in a box?

The second semantic task, the estimating-magnitude-proportionally task (experiment 2), was a production task examining children's performance on 'most' and 'some'. The aim of the task was to explore children's ability to map various proportional sets using the appropriate quantifier. The task yielded insight into children's semantic

comprehension of ‘most’ and ‘some’, allowing examination of differences in children’s production and perception of the two quantifiers.

The structure of this section is as follows. First, the children’s performance on experiment 1 is presented, followed by the adult groups’ performance on the task. Then, the results of experiment 2 are given, again for children and then adults. After that, a within-group comparison is conducted between children’s ability to correctly perceive (experiment 1) and produce (experiment 2) ‘most’ and ‘some’. The semantic-performance section concludes with a summary of the main findings on the study for the participants’ semantic comprehension of the quantifiers/operators.

4.3.1.1 Give-a-quantifier task (experiment 1)

The analyses in this section present the results for experiment 1, which aimed to assess children’s comprehension of the lexical meaning of quantifiers/operators. First, the analyses explore the child participants’ performance on the task, in three levels of analyses: a) a descriptive analysis of the children’s performance on the quantifiers/operators, b) an inferential analysis comparing the performance of the groups on each quantifier/operator to find out if they significantly differ, and c) an exploration of the types of wrong responses provided by the children over the three trials of each quantifier/operator. After this, the performance of the adult groups is presented concisely.

Children’s performance

To start with the first level of analysis, the results below compare the children’s performance in experiment 1. Following Barner et al. (2009), I used adult-like responses as criteria for judging correctness of children’s responses. This is because relying on the average numeral given by each child over the three trials for each quantifier would not provide accurate results regarding how children comprehend these quantifiers. For instance, when a child responds with (6, 1, 6), the average would be 4.3, which does indicate ‘most’ of the total for the set ($\{6\}$; see below) but does not reflect the actual comprehension of the quantifier, as the child’s score should be zero in each of the three trials.

Table 4.8 explains the criteria for each quantifier/operator for a given set size. Depending on the adult responses as well as the conventional meaning of the quantifiers, the criteria for accurate responses for the quantifier ‘all’ should refer to the whole set {6}, ‘most’ to {4, 5} (more than half and less than the whole set), and for ‘some’ {2, 3, 4, 5} (more than 1 and less than the whole set). For the disjunction ‘or’ and the conjunction ‘and’, accurate responses are considered to be {1} and {2}, respectively.

Table 4.8. Criteria for correct (adult-like) responses for each quantifier/operator within a given set size²

	All	Most	Some	Or A or B	And A and B
Set size (S)	S=6	S=6	S=6	S=3	S=3
Criteria	X=6	$X > 3 \wedge < 6$	$X > 1 \wedge < 6$	X=A	$X = A \wedge B$

Responses that met the above criteria received a score of 1, and 0 otherwise. Since the English sample had slightly fewer participants, I converted the scores to percentages so there would be no effect of unequal samples when visualising the results. Figure 4.4 shows the percentages of accurate responses given for each quantifier/operator. It can be clearly seen that the groups varied in their performance, especially on ‘most’ and ‘some’.

In more detail, the results presented in the figure reveal that all four groups were quite competent in their comprehension of ‘all’, as they all exhibit a ceiling effect, except for one bilingual child when tested in English. With ‘most’, it is clear that the bilingual children had poor semantic comprehension in both English (19%) and Arabic (26%), although scoring slightly higher in Arabic. The Arabic children had the lowest score among the groups (10%), while the English outperformed both the bilingual and the Arabic children (67%).

The results for ‘some’ showed that the bilingual children had good comprehension and almost equal performance in English (79%) and in Arabic (78%). The Arabic

² The terminology ‘semantic’ in this task is meant to encompass semantic and pragmatic meaning.

children showed weak comprehension of ‘some’, but better than for ‘most’. As can be seen, the English children were quite competent in their semantic comprehension of ‘some’ (94%) and clearly outperformed the other groups.

For the disjunction ‘or’, it can be noticed that the bilingual children have good, very similar performance in English (75%) and Arabic (71%), but still have the lowest scores compared with the Arabic and English children. The Arabic children had very good comprehension of ‘or’ (90%), and the English children exhibited a ceiling effect in their semantic performance on ‘or’, with 100% correct responses.

Finally, for the conjunction ‘and’, the results in figure 4.4 show that the bilingual children clearly scored higher when tested in English (95%) than in Arabic (70%), and that their scores in Arabic were lower than those of the Arabic children, who provided 90% correct responses. As with ‘or’, the English children showed a ceiling effect on ‘and’, with 100% correct responses.

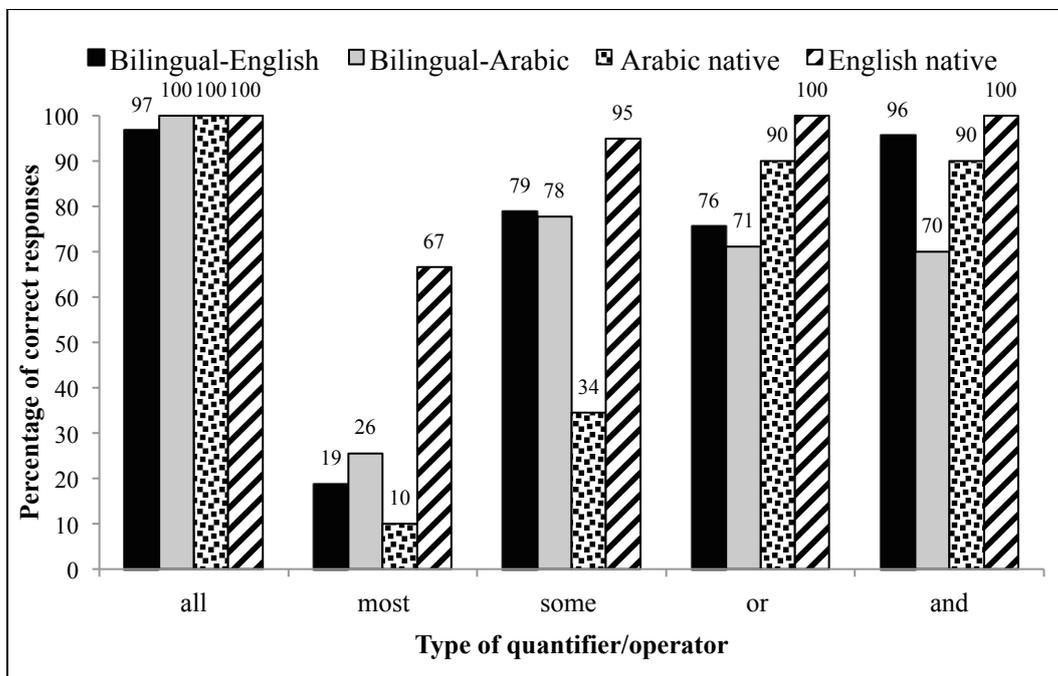


Figure 4.4. Proportions of correct (adult-like) responses given for each quantifier/operator in the bilingual (English, Arabic), Arabic, and English children’s groups (3 trials for each quantifier/operator)

Procedure of inferential analyses

The second phase of analysis aimed to find out if the differences between groups were statistically significant. I first conducted comparisons between groups on the quantifiers/operators; then, I performed separate pair-wise comparisons for each quantifier/operator between groups in order to identify differences across quantifiers/operators and across groups. However, before making any comparisons, I checked the assumption of normality for each quantifier/operator for each group (using the Shapiro–Wilk test) and for the whole sample (using the Kolmogorov–Smirnov test). The results revealed that neither the groups nor the whole sample were normally distributed for any of the quantifiers/operators ($p < 0.001$). This means that comparisons should be made using non-parametric tests. Although the analyses below report the findings of distribution-free or non-parametric tests, I conducted the comparisons using parametric tests as well (and only report the outcomes of the parametric tests when they conflict with the non-parametric results, which rarely occurred). My rationale for doing this is the claim that non-parametric statistics are less powerful than parametric ones (Larson-Hall, 2010) and also the claim that parametric tests can be still used even if the data are not normally distributed (Ghasemi & Zahedias, 2012). Given these statements from authoritative sources, in order to increase my confidence in any conclusions based on non-parametric outcomes I decided to use both kinds of tests and compare their findings.

Between-group comparisons for each quantifier/operator

I compared the groups' performance using the non-parametric alternative to the ANOVA, that is, the Kruskal–Wallis test. The results revealed significant differences between the bilingual-in-English, bilingual-in-Arabic, Arabic, and English children for 'most' ($H(3)=30.61$, $p < 0.001$), 'some' ($H(3)=32.1$, $p < 0.001$), 'or' ($H(3)=19.006$, $p < 0.001$), and 'and' ($H(3)=15.77$, $p = 0.001$). These findings were completely compatible with the ANOVA results (with group as a factor and the scores given to each quantifier/operator as dependent variables). Thus, to understand which groups differ and on which quantifiers/operators, pair-wise comparisons were conducted. The analyses below report the findings for each quantifier/operator separately.

Semantic performance on ‘most’

First, to explore differences between the bilingual group’s performance in English and Arabic, I conducted the Wilcoxon test for paired samples. The results revealed no significant difference ($Z=-.82$, $p=0.41$). Then, I used the Mann–Whitney test (for independent samples) to perform a comparison between the bilingual-in-Arabic and Arabic groups, again showing no significant difference ($U=358$, $Z=-1.65$, $p=0.1$). A comparison between the bilingual-in-English and English children, in contrast, revealed a very significant difference ($U=153$, $Z=-4.16$, $p<0.001$), with the English children scoring higher. Finally, a comparison between the Arabic and English groups again showed a very significant difference ($U=111$, $Z=-4.95$, $p<0.001$), with the English again performing better. All these results are consistent with the outcomes of the ANOVA post hoc (Games–Howell) multiple comparisons.

Semantic performance on ‘some’

A comparison between the performance of the bilingual children on ‘some’ in English and in Arabic showed no significant difference ($Z=-.52$, $p=0.6>0.05$). When the bilingual-in-Arabic performance was compared with the Arabic children’s performance, however, the Mann–Whitney test revealed a very significant difference ($U=222$, $Z=-3.61$, $p<0.001$), with bilinguals performing better. When the performance of the bilinguals-in-English was compared with that of the English children, however, the results revealed no significant difference ($U=335$, $Z=-1.35$, $p=0.18$). Finally, a comparison between the Arabic and the English children revealed a very significant difference ($U=121$, $Z=-4.93$, $p<0.001$), with the English scoring higher. All these results are consistent with the ANOVA (Games–Howell) post hoc multiple comparisons.

Semantic performance on ‘or’

First, a comparison between the bilingual children’s performance in English and Arabic revealed no significant difference ($Z=1.33$, $p=0.18$). When comparing the performance of the bilinguals-in-Arabic with that of the Arabic children, however, the test showed a significant difference ($U=303$, $Z=-2.69$, $p=0.007$), while the ANOVA post hoc (Games–Howell) showed only a marginally significance ($p=0.076$); the Arabic children scored higher. A comparison between the bilinguals-in-English and

the English children revealed another significant difference ($U=260$, $Z=-3.2$, $p=0.001$), consistent with the ANOVA results; the English children exhibited a ceiling effect (100% correct responses). Finally, a comparison between the Arabic and English children showed only a marginal significant difference ($U=338$, $Z=-1.19$, $p=0.056$), while the ANOVA post hoc revealed a completely insignificant difference ($p=0.23$).

Semantic performance on ‘and’

A comparison between the bilingual children’s performance in the two languages on ‘and’ revealed a significant statistical difference ($Z=-2.81$, $p=0.005$); the ANOVA also confirmed this ($p=0.02$); they performed better in English. A comparison between the bilinguals-in-Arabic and the Arabic children showed no significant difference ($U=371$, $Z=-1.44$, $p=0.15$), a finding compatible with the parametric comparison. A comparison between the bilinguals-in-English and the English children revealed an insignificant difference ($U=351$, $Z=-1.64$, $p=0.1$), consistent with the parametric ANOVA. Finally, a comparison between the Arabic and English children revealed a significant difference in performance ($U=299$, $Z=-2.60$, $p=0.009$), while the ANOVA post hoc (Games–Howell) comparison showed a marginal significant difference ($p=0.054$). Although the English and Arabic children showed a very good semantic comprehension of ‘and’ (100% and 90%, respectively), the significant statistical difference might be due to the English children’s exhibiting a ceiling effect while the Arabic children did not.

Understanding wrong-response variation within groups by exploring the child participants’ row results

The third level of analysis explores the performance of children who provided consistently incorrect responses (over the three trials for each quantifier/operator). Table 4.9 below provides information on the number and average age of children who showed constant incomprehension of ‘most’, ‘some’, ‘or’, and/or ‘and’ within each group. More precisely, it displays only the results of children who responded by acting either on the whole set (6/6) or just one item (1/6) for the quantifiers ‘most’ and ‘some’, or who gave 2 items when asked to give A or B, or one item when asked to give A and B.

Table 4.9. Number, average age, and performance of children who provided wrong responses consistently (in all three trials) within each group

Response	Most			Some			Or	And
	=6	=1	=1, 2, 3, 6	=6	=1	=6, 1	=2	=1
Bilingual-English	N=8	N=2	N=10	N=3	N=1	0	N=4	0
Total	20			4			4	NA
Age	M 5;6 SD 0.67			M 5;3 SD 0.17			M 5;8 SD 0.44	NA
Bilingual-Arabic	N=5	N=1	N=13	N=2	0	N=1	N=6	N=7
Total	19			3			6	7
Age	M 5;5 SD 0.79			M 5;3 SD 0.30			M 5;5 SD 0.93	M 5;5 SD 1
Arabic Children	N=13	N=5	N=5	N=11	N=4	N=1	N=2	N=1
Total	23			16			2	1
Age	M 5;5 SD 0.44			M 5;5 SD 0.40			M 5;7 SD 0.61	5;8
English Children	N=2	0	N=3	0	0	0	0	0
Total	5			0			0	0
Age	M 5;7 SD 0.29			NA			NA	NA

Before discussing the wrong responses, I remind the reader that the give-a-quantifier task aimed to explore children's comprehension of the quantifiers 'all', 'most', 'some', and the operators 'and', and 'or'. The children were presented to different sets and were asked to act upon certain instructions (e.g. put some of the carrots in the plate). For the quantifiers 'all', 'most', and 'some', each set consisted of 6 items, and for the operators 'or' and 'and', 3 items; there were three trials for each quantifier, given in a random order. For performance on 'most', starting with the lowest performers, it can be seen that the number of bilingual children who consistently gave wrong responses was approximately the same in Arabic and in English, and also that wrong responses occurred when the child acted on the whole set or on half or less of

the set items. When exploring whether it was always the same bilingual children who did not respond accurately, the row results revealed that 70% of the children had similar performance in the two languages, meaning that their semantic weakness might not be attributable to weakness in a specific language. The Arabic children's row results revealed that the majority of those who steadily gave wrong answers acted upon the whole set when asked, for example, to put most of carrots in a plate; fewer of these children acted upon only one item or on half or less than half of the set. Similarly, within the English group, only 2 children acted upon the whole set, 3 children provided consistently variably wrong answers (either acting upon the whole set, one item, or half or less of the items) and none acted upon only one item. The average age of the children who consistently responded incorrectly was approximately the same over the groups.

The row results for performance on 'some', as summarised in table 4.9, revealed that within the bilingual group few children (only 3 when tested in English and 2 in Arabic) acted steadily upon the whole set when asked to act upon 'some' of the set items. In addition, only one bilingual child responded to 'some' by acting on one item when tested in English; another gave consistently wrong answers by acting upon one or 6 of the items, across the three trials. As for which bilingual children provided wrong answers, the results revealed that half of the children who responded inaccurately in English made the same mistakes when tested in Arabic. The majority of wrong responses by Arabic children were made when the children acted upon the whole set over the three trials; only a few children acted upon only one item. Finally, none of the English children provided consistently wrong answers over the three trials of 'some'. The average age of the bilingual and Arabic children who reliably responded inappropriately was around 5;3 years old.

The investigation of children's row results for 'or' revealed that within the bilingual group, only 4 children (when tested in English) and 6 children (when tested in Arabic) responded steadily with 2 items (that is, when asked to give A or B, they gave both). When exploring whether the same children repeated the same kind of wrong responses in the two languages, the row results showed that 3 of the bilingual children who regularly gave wrong answers for 'or' in English did the same thing in Arabic. As for the Arabic children, only 2 responded consistently with 2 items, while none of

the English children did so (as they provided 100% correct responses). The average age of the bilingual and Arabic children giving consistently wrong responses for ‘or’ was around 5.7 years old.

Finally, row results for children’s performance on ‘and’ showed that the bilingual children gave steadily inappropriate answers when tested in Arabic (7 children acted upon only one item when asked to act upon A and B). Among the Arabic children, only one child responded constantly with one item when asked to act upon A and B, whereas the English children showed a ceiling effect, with 100% correct responses.

Adults’ performance

All the participating adults scored 100% on all the quantifiers/operators. With a 6-item set, the Arabic and English adults acted upon the whole set (6) when asked to put ‘all’; upon (4, 5) when asked about ‘most’; and upon (2, 3, 4) when asked about ‘some’. With a 3-item set, they gave (1) item for ‘or’ and (2) items for ‘and’. Since the adult groups had different language backgrounds (Arabic and English), I looked for any variation in responses on ‘most’ and ‘some’ by adults across groups. This would provide some evidence that Arabic-speakers in general use these quantifiers similarly to English-speakers, hopefully addressing any potential claim that the Arabic children’s poor semantic performance is due to different use of these quantifiers in Arabic. This, however, should not be taken by any means as making any claim that the study is providing evidence for complete similarity of use between the Arabic and English quantifiers, simply because making such a claim requires a larger sample.

Table 4.10. Arabic and English adults’ responses for the quantifiers ‘most’ and ‘some’ in the give-a-quantifier task

Adult group	‘most’ (set size=6)		‘some’ (set size=6)		
	Response=4	Response=5	Response=2	Response=3	Response=4
Arabic (10)	33%	67%	83%	17%	0
English (11)	55%	45%	70%	24%	6%

As table 4.10 shows, the Arabic adults slightly favoured acting upon 5 out of 6 items when asked to give ‘most’, while the English adults preferred 4 out of 6 items; none of the adults in either group gave half (3) of the items. Usage of ‘some’ was more similar across groups: the majority (around two-thirds) of adults, regardless of their native language, preferred giving 2 items when asked to give ‘some’, and the others, 3; only two responses, both in the English group, gave 4.

4.3.1.2 Estimating-magnitude-proportionally task (experiment 2)

The aim of the task was to find out how children would map various proportions using the quantifiers ‘some’ and ‘most’, across groups. They were asked to describe various scenes involving different proportions using only the quantifiers ‘most’ and ‘some’. The participants’ responses for all the proportions (2/15, 3/15, 4/15, 5/15, 6/15, 7/15, 8/15, 10/15, 11/15, 12/15, 13/15) were calculated; then, the percentages of use of the two quantifiers were computed by dividing the frequency of each quantifier (describing a proportion) by the total number of participants in each group. For example, the frequency of describing the proportion (3/15) with ‘some’ was calculated and then divided by the total number of participants (the group’s sample size).

Children’s performance

Figure 4.5 shows the percentages of use of ‘some’ and ‘most’ by the bilingual children. It can be noted that for the small proportions (2/15, 3/15, 4/15, 5/15), more than 70% of bilingual children’s responses rated the target yellow circles using the quantifier ‘some’. This percentage decreased for the fuzzy proportions (6/15, 7/15, 8/15), which represent approximately half of the set size in the scene, and dipped further with the large proportions (10/15, 11/15, 12/15, 13/15), where ‘most’ was used more than 80% of the time. Similar results were found for the English monolingual children, albeit with slightly lower percentages of ‘some’ with the proportions (10/15, 11/15, 12/15, 13/15) and lower use than the bilinguals of ‘most’ with small proportions (2/15, 3/15, 4/15, 5/15) as displayed in figure 4.6.

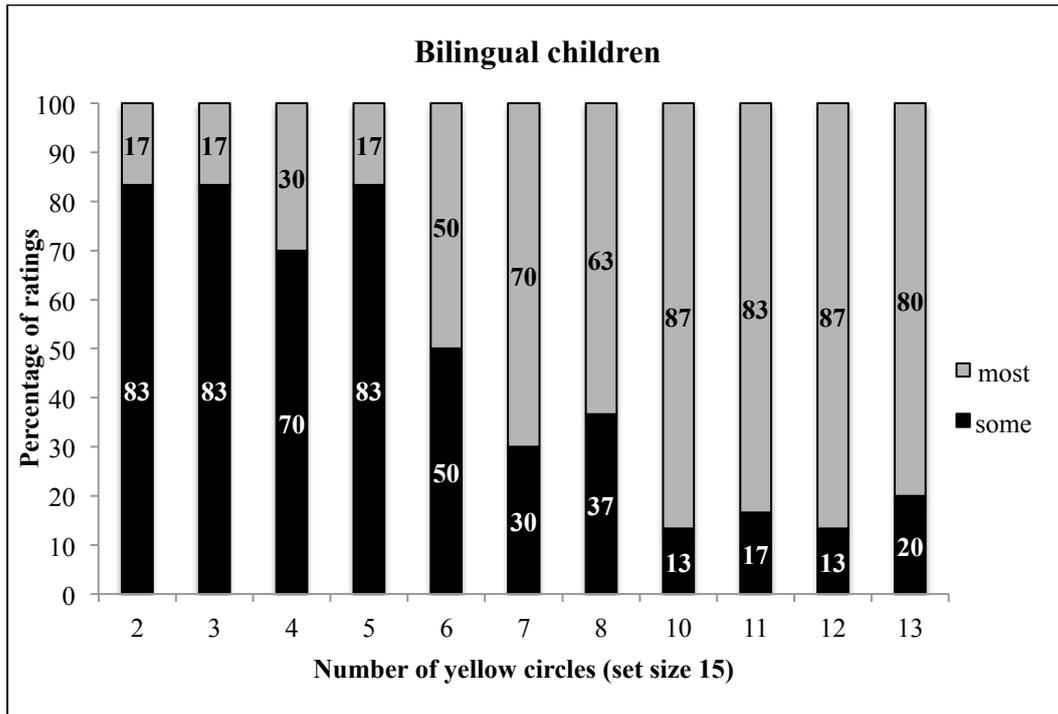


Figure 4.5. Breakdown of bilingual children’s use of the quantifiers ‘most’ and ‘some’ to describe various proportions

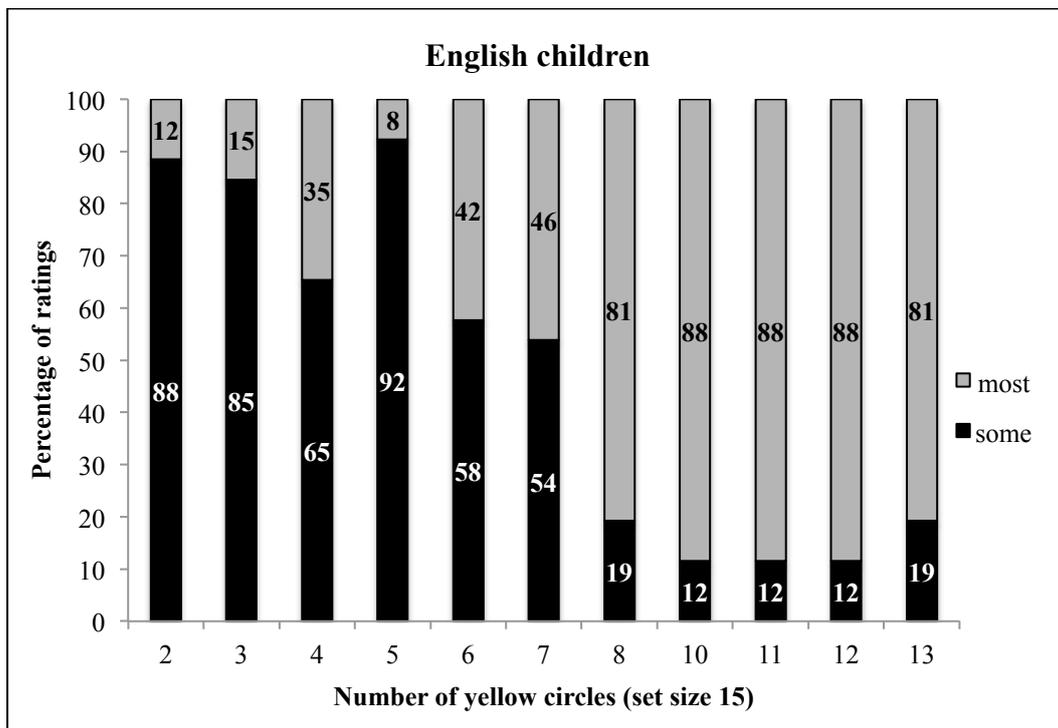


Figure 4.6. Breakdown of English children’s use of the quantifiers ‘most’ and ‘some’ to describe various proportions

The performance of Arabic monolingual children in this task did not reveal an obvious preference for either quantifier to rate a given proportion. That is, Arabic children used the two quantifiers at an approximately equal rate, as figure 4.7 shows. It can be seen, however, that there is a slightly greater tendency to use ‘most’ with larger than with smaller proportions. The inconsistent use of ‘some’ to rate various proportions might indicate incomplete comprehension of this quantifier’s lexical meaning; more precisely, it seems that those children might not be able to differentiate accurately between the positions of ‘most’ and ‘some’ in the quantifier scale, although they were able to do so accurately with numerals, which have more concrete representations.

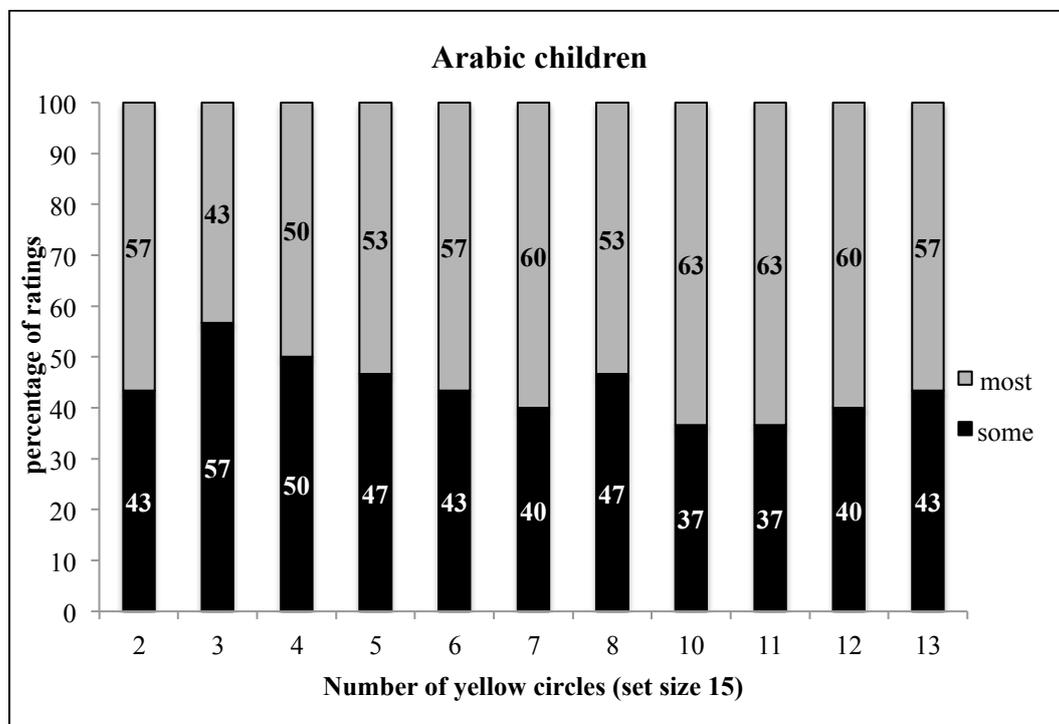


Figure 4.7. Breakdown of Arabic children’s use of the quantifiers ‘most’ and ‘some’ to describe various proportions

A bivariate correlation test revealed no significant correlation between type of response (quantifier) and group ($r(\text{two-tailed})=0.017$, $p=0.64$), but a strong positive correlation between quantifier type and proportion ($r(\text{two-tailed})=0.307$, $p<0.0001$).

Since this study is interested in investigating which proportions should clearly map to either quantifier, the fuzzy (filler) proportions (the ones that represent around half of

the total items) were removed from the following analyses. Separate pair-wise comparison of rate frequency between each two groups revealed no significant difference between bilingual and Arabic children ($U=27720$, $Z=-.823$, $p=0.41$), bilingual and English children ($U=10060$, $Z=-.760$, $p=0.45$), or English and Arabic children ($U=19686$, $Z=-.264$, $p=0.791$).

Estimating-magnitude-proportionally task results with new criteria for scoring

Although the above-discussed results do give some indication of the potential differences between groups, I decided to further explore the differences between groups, focusing only on the critical items, by applying a similar analysis to that performed on experiment 1's results. That is, I used adult-like responses as criteria for judging children's responses, scoring 0 for wrong responses and 1 for correct ones. All the adults were found to use 'some' (100%) with small critical proportions (2/15, 3/15, 4/15, 5/15) and 'most' (100%) with large critical proportions (10/15, 11/15, 12/15, 13/15); children's responses were expected to match to be scored correctly. Then, I calculated the correct responses given by each child for 'most' and 'some' separately (there were 4 trials for each quantifier). Figure 4.8 below gives the percentages of correct responses for each group after applying the new criteria. I should remind the reader that in this task the bilingual children completed the task in one language, English, only.

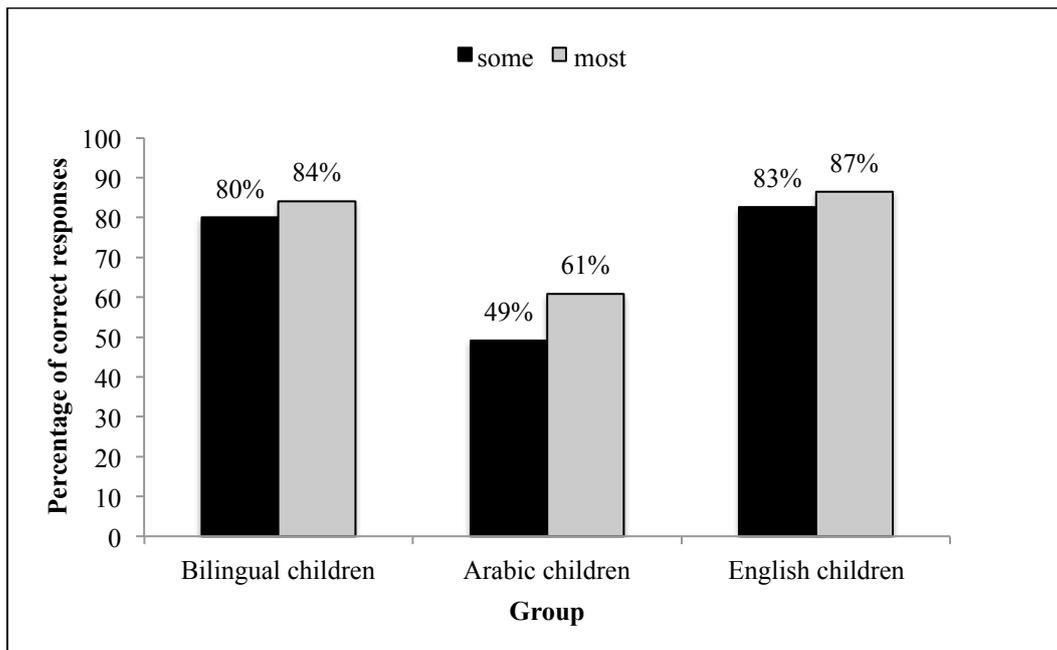


Figure 4.8. Percentages of total correct responses (adult-like choice) for ‘most’ (10/15 to 13/15) and ‘some’ (2/15 to 5/15) in the estimating-magnitude task by the child groups (four critical trials for each quantifier, fuzzy items (fillers) were excluded)

It can be seen in figure 4.8 that the performance of the bilingual and English children on ‘most’ and ‘some’ was very similar, and that although the Arabic children had the lowest score among the groups, their performance was better than in experiment 1. I will further examine the differences in performance in experiments 1 and 2 within each group in section 4.3.1.3. It was surprising to find that all the children generally, and the Arabic children specifically, performed better for ‘most’ than for ‘some’. This is because in the give-a-quantifier task (experiment 1) all the bilingual children (in Arabic and English) and the Arabic children showed poor comprehension of ‘most’, and even the English children had moderately lower performance compared with their comprehension of ‘some’ in experiment 1.

To find out if these differences were statistically significant, I conducted some comparisons using free-distribution tests, since all the groups and the whole sample violated the assumption of normality ($p < 0.05$); these results are provided in the next section. As in experiment 1, I ran the parametric ANOVA, and only report whether I found different results with this parametric test, not with the non-parametric one.

Between-group comparisons with new-scored results of experiment 2

Semantic performance on ‘some’

To explore possible differences in performance on ‘some’, I first conducted the Kruskal–Wallis test to compare the three groups. The results revealed significant differences between the groups ($Z=11.18$, $p=0.004$). Then, pair-wise comparisons were conducted between each pair of groups, revealing significant difference between the bilingual children and the Arabic children ($U=275$, $Z=-2.81$, $p=0.005$), with the bilinguals performing better, but no significant difference between the bilingual and English children ($U=387$, $Z=-.058$, $p=0.95$). Comparing the Arabic with the English children, the outcomes showed a significant difference in performance ($U=223$, $Z=2.83$, $p=0.005$); the English children performed better. All these findings are compatible with the ANOVA results.

Semantic performance on ‘most’

The same analytic procedure was applied to the results for ‘most’ to find out if the groups significantly differed in this regard. The Kruskal–Wallis test showed significant differences between the groups’ performance ($Z=12.1$, $p=0.002$), and paired-wise comparison (with the Mann–Whitney U-test) again revealed a significant difference between the bilingual children and Arabic children ($U=272$, $Z=-2.85$, $p=0.004$), with the bilinguals performing better, no significant difference between the bilingual and English children ($U=386$, $Z=-.071$, $p=0.94$), and a significant difference between the Arabic and the English children ($U=223$, $Z=2.95$, $p=0.003$); the English children performed better. As with ‘some’, all these findings were consistent with the ANOVA results.

Adults’ performance

The adult control groups completed the estimating-magnitude-proportionally task to find out which quantifier (‘most’, ‘some’) the adults would use respectively with the small (2/15, 3/15, 4/15, 5/15) and large proportions (10/15, 11/15, 12/15, 13/15). The main focus would be on critical items, as the current study is not interested in exploring behaviour on fuzzy items (representing approximately half the set) and just used them as filler. I examined performance for each group separately.

Figure 4.9 presents the results for the Arabic adult group. It can be seen that all the Arabic adults described small proportions using ‘some’ (100%) when given the option of using either ‘most’ or ‘some’, in each trial. Similarly, with large proportions they used ‘most’ (100%). Their performance on the fuzzy items only varied on the proportion 7/15, where they preferred to use ‘some’ (60%) than ‘most’. Their responses were steady on the other two fuzzy items; they used ‘some’ with 6/15 and ‘most’ with 8/15 at a rate of 100%.

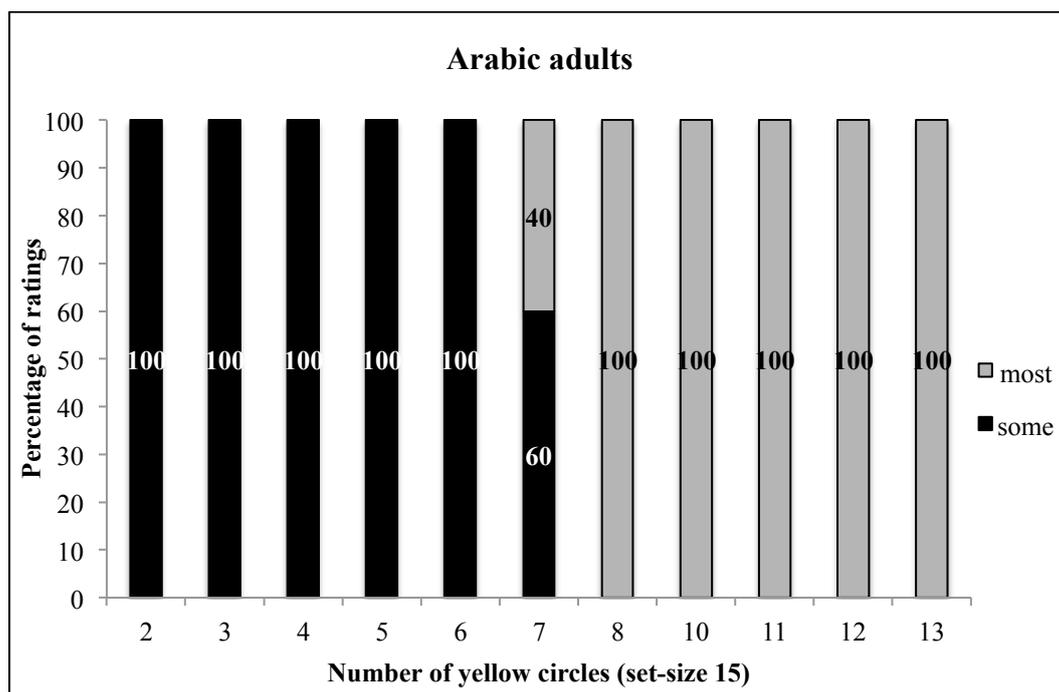


Figure 4.9. Breakdown of Arabic adults’ use of the quantifiers ‘most’ and ‘some’ to describe various proportions

The performance of the English adults on the critical items was almost exactly like the performance of the Arabic adults, as displayed in figure 4.10: they constantly used ‘most’ with the larger proportions and ‘some’ with the smaller proportions. However, their performance on the fuzzy items differed slightly from that of the Arabic adults: with the proportion 8/15, they were compatible with the Arabic adults, using ‘most’ 100% of the time, but with 6/15, one-third of the total responses still preferred ‘most’, while the remaining two-thirds used ‘some’. With 7/15, the majority of the participants (80%) used ‘some’. Although this study is not interested in exploring or even explaining the differences between the two adult groups’ performance on the

fuzzy proportions, it can be briefly noted that such variation in describing fuzzy items was also found in Yildirim et al.'s (2016) study. Not all the English adult speakers used the same quantifiers to describe proportions that represented roughly half of the set. The forced-choice method might be responsible for such variation, so it might be that Arabic adults have no preference to regarding the use of any of the quantifiers with the fuzzy items, but that these were the only allowed options.

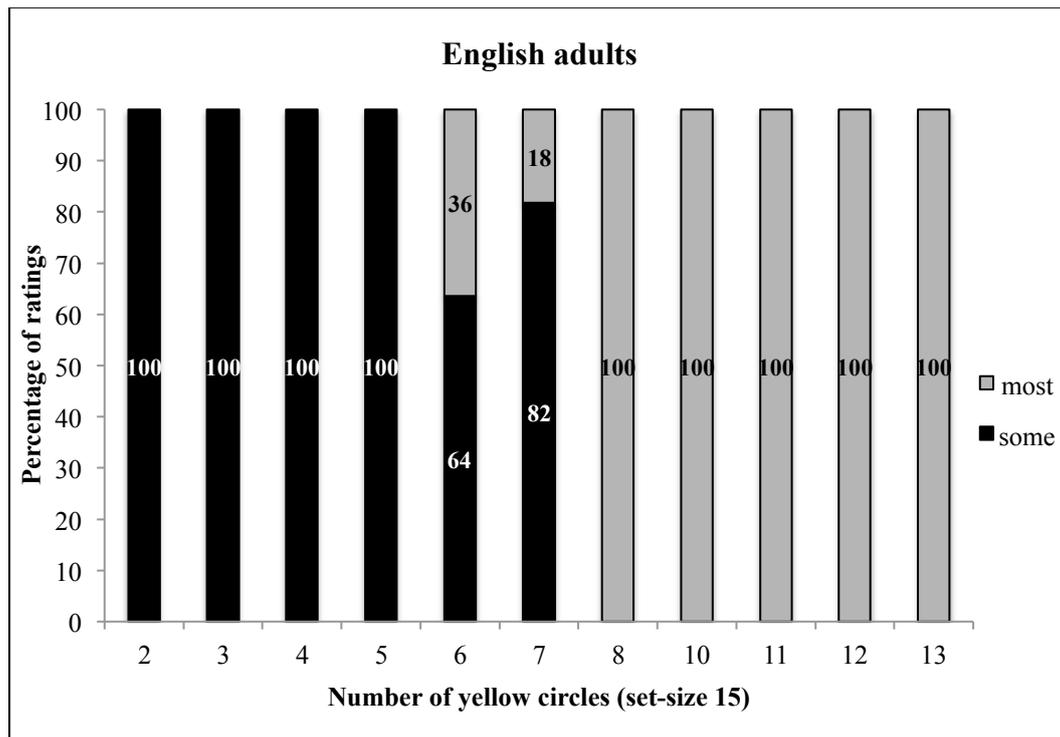


Figure 4.10. Breakdown of English adults' use of the quantifiers 'most' and 'some' to describe various proportions

4.3.1.3 Children's performance on 'some' and 'most': Perception v. production

This part of the analysis explores the differences in children's semantic comprehension of 'most' and 'some' when asked to manipulate sets according to the given quantifier (assessing perception of quantifiers, in experiment 1) to or to map proportional sets to the suitable quantifier (either 'most' or 'some', assessing production of quantifiers, in experiment 2). Figures 4.11 and 4.12 below show perceptive and productive performance for 'most' and 'some' among bilingual (in English), Arabic, and English children.

Figure 4.11 presents performance for ‘most’ in both experiments. It can be clearly seen that children in all groups were better in the production task than the perception task. That is to say, they were better at mapping various proportional sets than at manipulating sets that correctly represented ‘most’. Indeed, in bilingual and Arabic children, there was a dramatic change in performance on ‘most’ between the two experiments, and the difference was still meaningful for English children.

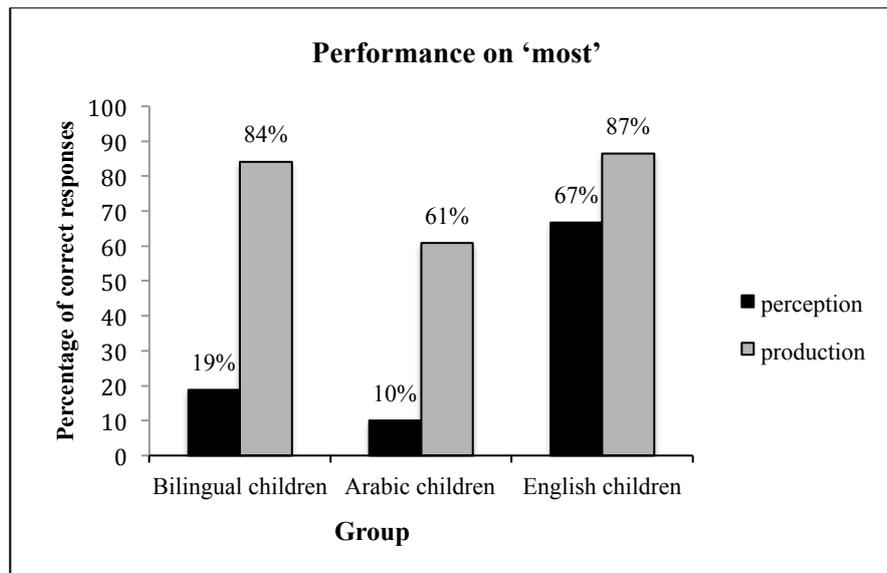


Figure 4.11. Percentages of children’s correct (adult-like) responses given for ‘most’ in experiment 1 (perception) v. experiment 2 (production)

Figure 4.12 shows performance for ‘some’ in experiments 1 and 2. Unlike the children’s performance on ‘most’ as given in figure 4.11, in general no dramatic change can be detected on ‘some’ across the two experiments. The bilingual children have exactly the same performance in the perception and the production tasks; the performance of the Arabic children slightly improved in experiment 2, while the English children performed better on ‘some’ in experiment 1.

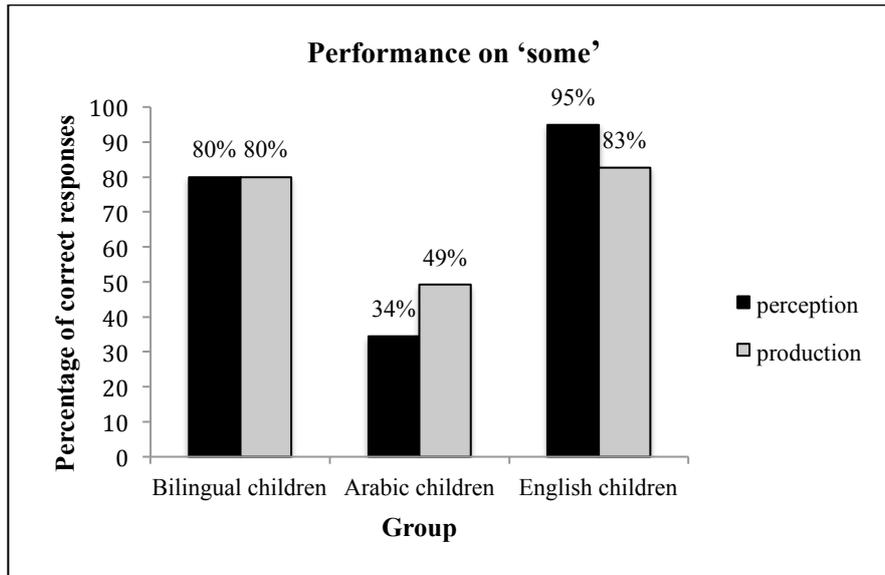


Figure 4.12. Percentages of children's correct (adult-like) responses given to 'some' in experiment 1 (perception) v. experiment 2 (production)

To find out if the above-discussed results for children's semantic performance on 'most' and 'some' are statistically significant, I compared the performance of children in each group separately using non-parametric tests, since all groups violated the assumption of normality ($p < 0.05$). To make the findings more secure, I also ran parametric tests, and only report results if the two tests give compatible results.

Bilingual children: Perception v. production

Here, I investigate the difference in bilingual children's performance on 'most' and 'some' in experiment 1 (perception) versus experiment 2 (production) using the Wilcoxon test for paired samples. The results revealed a significant difference between bilingual performance on 'most' across the two experiments ($Z = -4.46$, $p < 0.001$), with better performance in the production condition, while bilingual performance on 'some' did not significantly differ across two experiments ($Z = -.032$, $p = 0.98$). These findings are consistent with the parametric t-test (paired sample).

Arabic children: Perception v. production

The performance of the Arabic children on ‘most’ and ‘some’ was explored using the distribution-free Wilcoxon test. On ‘most’, the results showed a very significant difference ($Z=-4.47$, $p<0.001$), with better performance in the production condition, but there was no significant difference on ‘some’ ($Z=-1.35$, $p=0.17$). These results were consistent with the parametric analyses.

English children: Perception v. production

Finally, I compared the performance of the English children on ‘most’ across experiments 1 and 2, where the Wilcoxon test revealed a significant difference ($Z=-2.91$, $p=0.004$), with better performance in the production condition. When comparing the English children’s performance on ‘some’ in the two experiments, the outcomes showed a significant difference ($Z=-2.19$, $p=0.028$), unlike for the bilingual and Arabic children, the English children had better performance in the perception condition. These differences were confirmed using the outcomes of the parametric t-test for paired samples.

The relationship between perception and production

Having explored the differences between and within groups in the perception and production of ‘most’ and ‘some’, it was important to understand the potential relation between children’s performance in experiment 1 and in experiment 2, regardless of their language background. To achieve this, I first conducted a bivariate correlation test, then a regression test (a generalised linear model (GLM) with Poisson log link function). The correlation test revealed a significant positive correlation between children’s perception and production of ‘some’ ($r(\text{two-tailed})=.031$, $p=0.001$) and only a marginally significant correlation in the production and perception of ‘most’ ($r(\text{two-tailed})=.0178$, $p=0.057$).

Then I fitted a regression model (GLM with Poisson distribution) with quantifier type (most, some) and group (bilingual, Arabic, and English children) as fixed factors, and participants’ performance in the production task and age as covariates; I had the results of the perception task as a dependent variable. To understand the effects of age and performance in the production task on each quantifier score in the perception

task, I added an interaction effect between quantifier type and each of the two factors. Also, since my major aim was to explore the relationship between the two tasks and to understand the effect of age on performance, I did not include vocabulary or NVIQ in the model, since they were highly correlated with age.

The results of the model are displayed in table 4.11. It revealed a significant effect of group ($X^2(2)=43.431$, $p<0.001$), with the bilingual and the Arabic children performing significantly lower than the English children, as suggested by the model negative estimates (Bs). There was no main effect of the production task ($p>0.05$), but the model revealed a significant effect of interaction between quantifier type and the performance in the production task ($X^2(2)=28.578$, $p<0.001$). The regression analysis also showed a significant main effect of age ($X^2(1)=5.78$, $p=0.016$), and a significant effect of interaction between age and quantifier type ($X^2(2)=7.632$, $p=0.022$), although this effect was only marginally significant for the perception of ‘some’ ($X^2(1)=2.971$, $p=0.085$).

Table 4.11. GLM regression results: Exploring the relationship between children’s perception and production of the quantifiers ‘most’ and ‘some’

Parameter	B	SE	X ²	DF	P
(Intercept)	.698		11.038	1	.001
Language Group			43.431	2	.000
Bilingual children	-.463	.1403	13.043	1	.001
Arabic children	-1.24	.1937	39.203	1	.000
English children	0
Quantifier type (most)	-.072	.0611	1.397	1	.237
Quantifier Type*Production Score			28.578	2	.000
Some*Production score	.205	.038	28.578	1	.000
Most*Production score	0
Age (months)	.077	.0321	5.78	1	.016
Quantifier type*Age			7.632	2	.022
Some*Age (months)	-.016	.009	2.971	1	.085
Most*Age (months)	0

Note: Model reference levels: For group (English child group), for quantifier (‘most’)

I also conducted two other GLM models to understand the effect of performance in the production task and age on each quantifier separately; the results of these models

confirmed the above results, revealing a strong effect of the production task on the perception of ‘some’ ($X^2(1)=6.476$, $p=0.011$, $B=.125$) but only a marginal effect of age ($X^2(1)=3.416$, $p=.065$, $B=.015$); in contrast, the effect of age was significant on the perception of ‘most’ ($X^2(1)=6.245$, $p=.012$, $B=.031$), but the effect of the production task was marginal ($X^2(1)=3.437$, $p=.064$, $B=.007$).

4.3.1.4 A summary of semantic performance

The analyses in this section aim to answer the first research question, related to children’s appropriate semantic comprehension of the quantifiers by exploring the child and adult (control) groups’ performance on two semantic tasks: the give-a-quantifier task (experiment 1) and the estimating-magnitude-proportionally task. Then, I compare child performance on ‘most’ and ‘some’ across experiments 1 and 2.

The results of experiment 1 showed that the Arabic and English adults exhibited ceiling effects on all the quantifiers, with almost the same semantic performance. Children in all groups exhibited a ceiling effect in their semantic performance on ‘all’. For ‘most’, the bilingual children (in both languages) as well as the Arabic children scored very low, while the English had a significantly higher score. On ‘some’, the bilingual and English children showed good comprehension, while the Arabic children had very weak performance. Comparisons between the groups revealed no significant differences between the bilingual-in-English and English children, but a significant difference was found between the bilingual-in-Arabic and Arabic children and also between the English and Arabic children. Performance on ‘or’ and ‘and’ showed that the English and Arabic children both had very good comprehension and performed significantly better than the bilinguals (in both Arabic and English).

The outcomes of experiment 2 (the estimating-magnitude-proportionally task) showed that the two adult groups mapped large proportions with ‘most’ and small proportions with ‘some’ (100%), across languages. The children’s results showed that the bilinguals (tested only in English) and English children had a tendency to use ‘most’ with large proportions and ‘some’ with small proportions. For the Arabic children, in contrast, the results showed that the majority might not yet have the ability to map these proportions to the appropriate quantifiers. Scoring children’s performance as 0/1

based on adult-likeness of responses, the analyses revealed that all the groups, including the Arabic children, were significantly more accurate in mapping large proportions with ‘most’ rather than producing sets that represent this quantifier. With ‘some’, only the Arabic children performed significantly better in the production task. Comparisons between groups revealed significant differences only between the bilingual and the Arabic children and between the English and the Arabic children.

The analyses also explored the differences and relationships between child group performance on ‘most’ and ‘some’ in experiment 1 (a perception task) versus experiment 2 (a production task). As table 4.12 shows, there was a significant difference between tasks in performance on ‘most’, namely, that all the children performed better in experiment 2. Performance on ‘some’ did not significantly differ for the bilingual and Arabic groups, whereas the English children performed significantly better in the perception task (experiment 1) than the production task (experiment 2).

Table 4.12. A summary of within-group differences in the perception and production of ‘most’ and ‘some’ (experiment 1 v. experiment 2)

Group	Most (experiment 1 v. experiment 2)	Some (experiment 1 v. experiment 2)
Bilingual (in English)	S (production, exp. 2 +)	NS
Arabic children	S (production, exp. 2 +)	NS
English children	S (production, exp. 2 +)	S (perception, exp. 1 +)

Note: Abbreviations in the table: S (significant), NS (non-significant), exp. (experiment), (‘+’ means better performance)

The regression analysis showed a significant effect of age on children’s performance on ‘most’ in experiment 1, with only a marginally significant effect of their performance on the production task. Conversely, age seems to have only a marginal effect on children’s comprehension of ‘some’, where their performance on the production task best predicted their performance on the perception one.

4.3.2 Performance on number tasks

This section presents results for numerical tasks, and is intended to answer the second research question, related to children's acquisition of numerical system and its potential effect on the acquisition of quantifiers. This question is answered in two stages; first, by assessing children's performance on the four numerical tasks, and then by exploring the relationship between children's performance on number and quantifying tasks. Thus, I start the analysis by investigating children's ability to count various sets (the how-many task), to produce various sets when given different number-words (the give-a-number task), and to successfully differentiate several set sizes without counting (the non-verbal ordinal task). Success on these tasks demonstrates that children have acquired the cardinal principle and can differentiate various sets without counting, which is necessary before giving them the two target tasks (the (fourth) numerical task of 'estimating-magnitude-numerically' and the semantic task of 'estimating-magnitude-proportionally', as presented in section 4.3.1.2). The purpose of the two later tasks was to help explain the reasons for the weak semantic performance of the Arabic children with the quantifiers 'some' and 'most' in experiment 1, given the significant difference between their performance and that of the bilingual and English children. After exploring children's performance on the numerical tasks, I conducted regression analysis to examine the relation between the acquisition of quantifiers and that of number words. For each numerical task, the analyses start with children's results, then briefly report adults' performance. The section concludes with a summary of the main findings for the participants' performance on the number tasks.

4.3.2.1 Results for the how-many task

This task serves two goals: a) it assesses children's knowledge of the how-to-count principle and b) it evaluates children's ability to count relatively large sets {10}, {14} and map these sets to their true values in the numerical list. The latter ability is critical for success in the estimating-magnitude-numerically task. The adult groups serve as controls; their performance is reported after the child groups' results.

Children's performance

Table 4.13 shows the results for the three child groups' performance on the how-many task. It can be seen that the majority (around 90%) of bilingual and Arabic children and all of the English children were able to accurately count a set size {10}, while the bilingual and Arabic children were slightly less competent with a set size {14} (approximately 80% counted correctly). The English monolingual children exhibited a ceiling effect in the task, with 100% correct answers.

Table 4.13. Number and percentage of participants in each child group who correctly counted set sizes of {10} and {14} in the how-many task

Set size	Bilingual		Arabic bidialectal		English monolingual	
	N (30)	%	N (30)	%	N (26)	%
{10}	27	90	28	93	26	100
{14}	23	77	24	80	26	100

Breaking children's wrong responses down by type

To understand the nature of the errors the children made when counting, the individual performance (row results) of those children was explored. Table 4.14 provides information on the children's error types by age. Starting with performance on set size {10}, it can be seen that only 3 bilingual and 2 Arabic children could not accurately give the exact numeral value of this set size. These ones answered '9' or '11', and this occurred because they either skipped or duplicated one item in the row while counting the set items (in two attempts). In addition, one Arabic child answered '5'. Such errors might not be due to lack of cognitive knowledge of mapping of sets to parallel values in a numerical list or with incomplete acquisition of the how-to-count principle, but possibly instead with lack of experience or training with number words.

Table 4.14. Performance and age of children who could not complete the how-many task accurately

Bilingual				Arabic bidialectal			
ID	Age	Set {10} Answer	Set {14} answer	ID	Age	Set {10} Answer	Set {14} Answer
BC2	4;1	9	13	AMC4	5;7	10	8
BC3	4;8	10	10	AMC7	5;9	9	19
BC11	5	11	17	AMC10	5	10	15
BC17	5;3	10	15	AMC17	5	5	9
BC18	5;4	9	12	AMC29	5;9	10	13
BC19	4;6	10	18	AMC30	5;9	10	10
BC21	5;5	10	16				

Moving to set size {14}, table 4.14 shows the number of children in the bilingual and Arabic groups who could not map this set to its true value in the numerical list. The number is slightly higher than for {10}. Although some children could not give accurate values, it is noticeable that they nevertheless answered with larger numbers, reflecting the larger size of this set, except one bilingual child (BC3) who gave the same answer for both sets, and one Arabic child (AMC4) who gave a lower number for {14}. Looking at the ages of children who could not complete the task successfully, as in table 4.14, we see that the bilingual children who could not do so were slightly younger than the Arabic children who could not do so (mean ages 5 and 5;6, respectively).

Between-group differences (inferential analyses)

To find out if the differences in group performance outlined above were statistically significant, each child who could count set size {10} correctly was given a score of 1, and 0 otherwise; the same approach was taken for size {14}. These scores were added together to determine the child's overall performance. Next, I checked the assumption of normality, which was evidently violated, since the English children exhibited a ceiling effect on the task. Indeed, the Shapiro–Wilk test of normality (for each group) and Kolmogorov–Smirnov (for the sample) showed that neither any of the groups nor the whole sample were normally distributed ($ps < 0.05$). Thus, I made comparisons between groups first with non-parametric tests, then with parametric tests.

The results of the Kruskal–Wallis H-test revealed significant differences between the groups only on set size {14} ($H(2)=6.69$, $p=0.035$) and final score ($H(2)=6.67$, $p=0.036$), not set size {10} ($H(2)=2.574$, $p=0.28$). I conducted pair-wise comparisons (with the Mann–Whitney U-test) of children’s scores on each set size; the outcomes of tests with set size {10} revealed no significant difference between the bilingual and Arabic children ($U=435$, $Z=-.463$, $p=0.64$), the bilingual and English children ($U=351$, $Z=-1.64$, $p=0.1$), or the Arabic and English children ($U=364$, $Z=-1.33$, $p=0.18$). Comparisons between groups for set size {14} showed no significant difference between the bilingual and the Arabic children ($U=435$, $Z=-.311$, $p=0.76$), but there were significant differences between the bilingual and English children ($U=299$, $Z=-2.61$, $p=0.009$) and the Arabic and English children ($U=312$, $Z=-2.392$, $p=0.017$); in both cases the English children performed better, since they exhibited a ceiling effect. These findings are compatible with ANOVA results (with group as a factor and scores for each set size as dependent variables). The tests showed a significant difference only in set 14 ($F(2,83)=3.54$, $p=0.033$), not in set 10 ($F(2, 83)=1.29$, $p=0.28$). The post hoc (Games–Howell) comparisons revealed significant differences in performance on set size {14} between the bilingual and English children ($p=0.016$) and the Arabic and English children ($p=0.03$), but not between the bilingual and Arabic children. Again, the English children performed better than the other two groups, since they exhibited a ceiling effect in the task.

Adults’ performance

The two adult groups’ performance on this task showed that, as expected, all the adults were able to complete the task successfully (i.e., there were 100% correct responses).

4.3.2.2 Results for the give-a-number task

Children’s performance

This task aimed to measure children’s knowledge of the exact meanings of the numerals 1 through 6 by manipulating sets to create accurate new sets according to the target numeral in each trial. Children who could complete the task successfully by generating set sizes compatible with the given numerals were described as ‘cardinal-

principle-knowers' (CP-knowers), while children who could not do so for some numbers were described as 'subset-knowers'. Table 4.15 displays the number of children in each group (bilingual, Arabic, and English) who could correctly create set sizes matching the given numerals. All the bilingual and English children were CP-knowers, as were 25 Arabic children (out of 30); the remaining five children were subset-knowers. (That is, no child failed to give a number 100% of the time.)

Table 4.15. Number and percentage of cardinal-principle-knowers in each child group based on the results of give-a-number task

Set size	Bilingual		Arabic bidialectal		English monolingual	
	N (30)	%	N (30)	%	N (26)	%
{1}–{6}	30	100	25	83.3	26	100

Breaking children's wrong responses down by type

To understand the behaviour in particular of the Arabic children, who did not complete the give-a-number task completely successfully, I investigated their row results and at the same time compared which children were not able to complete the how-many task and the give-a-number task correctly (compared results on both tasks).

Table 4.16 displays the age of the Arabic subset-knowers and the highest numerals they could provide accurately, and compares their performance on the give-a-number task with that on the how-many task. As seen in the table, the row results show that the subset-knowers were mostly the same children who could not complete the how-many task successfully. Two children were exceptions: it can be seen that although one child (AMC7) could not complete the how-many task correctly, he was able to produce sets that accurately matched the given number. In contrast, it can also be seen that the youngest Arabic subset-knower (AMC18) successfully counted sets of 10 and 14 but failed to generate set sizes or numerals larger than 4. The highest numeral answered correctly by the Arabic subset-knowers was either 4 or 5; their mean age was 5;4 years old.

Table 4.16. Individual results for Arabic children who could not complete the how-many and/or give-a-number tasks successfully

ID	Age/Gender	Give-a-number task	How-many task	
		Highest numeral mapped successfully to correct set	Set size (10)	Set size (14)
AMC4	5;7 (M)	5	10	8
AMC7	5;9 (M)	6 (✓)	9	19
AMC10	5 (F)	5	10	15
AMC17	5 (M)	4	5	9
AMC18	4;7 (F)	4	10	14
AMC29	5;9 (M)	5	10	13
AMC30	5;9 (M)	5	10	10

Note: The symbol (✓) means the child completed the whole task successfully.

Between-group differences (inferential analyses)

Before exploring whether the Arabic children's performance differed significantly from that of the other two groups, I first checked whether the groups and the overall sample met the assumption of normality. As might be easily predicted, all the groups and the full sample violated the assumption of normality; this was revealed by the Shapiro–Wilk and Kolmogorov–Smirnov tests ($p < 0.05$). Pair-wise comparisons with non-parametric Mann–Whitney U-tests showed significant differences between the bilingual and Arabic children ($U=360$, $Z=-2.557$, $p=0.011$) and between the Arabic and English children ($U=312$, $Z=-2.388$, $p=0.017$); in both cases the Arabic children performed lower (83.3%), while the other two groups exhibited a ceiling effect (100%). The ANOVA results revealed a very significant difference between the Arabic children and the other two groups ($F(2, 83)=5.93$, $P=0.004$), and the post hoc comparisons were completely consistent with the non-parametric test results.

Adults' performance

The Arabic and English adults completed the task correctly, as expected, with 100% correct responses.

4.3.2.3 Non-verbal ordinal task

The aim of this task was to demonstrate non-verbally the availability and accuracy of children's analogue magnitudes, as this ability was taken to be critical to success in the estimating-magnitude task.

Children's performance

All the children—bilingual, Arabic, and English—exhibited ceiling effects in this task; they were all able to point out the set with bigger size (more circles) in each trial. All the child participants were able to differentiate between the two sets (appearing simultaneously on the computer screen) not only when the sets were clearly distinct (6 v. 2, 6 v. 10, 15 v. 9) but also when they were not distinct at a glance (2 v. 3, 10 v. 8, 15 v. 12).

Adults' performance

All the adult participants completed the task successfully, with 100% correct responses.

4.3.2.4 Estimating-magnitude-numerically task

After assessing children's ability to distinguish between sets which were relatively close in size, above, this task was employed as a measure of children's 'approximate number' system. That is, the task aimed to assess children's ability to map between various set sizes (magnitudes) and number words without counting.

Children's performance

Averages of numerals given by children in the three child groups for each set size (1, 2, 3, 4, 5, 6, 8, 12) were calculated. Figure 4.13 displays the averages for each child group; it can be easily seen that the numerals children provided grew linearly with the magnitudes. Since set sizes were displayed randomly, the children's results thus show that they were capable of manipulating numeral responses according to the magnitude displayed in a scene, and were able to provide large numbers when the set size increased and give small numbers when it decreased.

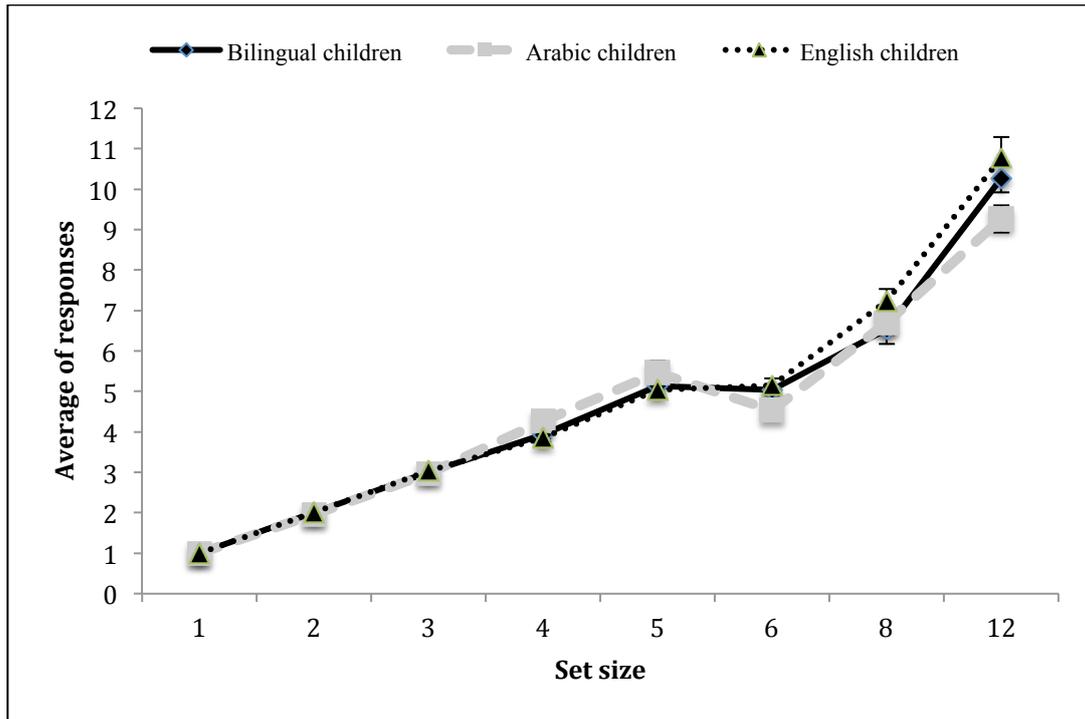


Figure 4.13. Average numeral response by set size for bilingual, Arabic, and English children in the estimating-magnitude-numerically task

A bivariate correlation test revealed no significant correlation between the answer provided (the number) and group ($r(\text{two-tailed})=0.019$, $p=0.61$) but a strong positive correlation between set size and the given number ($r(\text{two-tailed})=.917$, $p<0.001$). This finding indicates that the children performed similarly across groups in this task, and the positive correlation between set size and number provided might indicate that children in all groups were able to map various magnitudes to their corresponding number words in the numerical list.

Adults' performance

The performance of the two adult groups is displayed in figure 4.14. It can be seen that all the Arabic and English adults were able to increase their number answers as the magnitudes grew, at the same rate. The two groups had almost the same performance on the task.

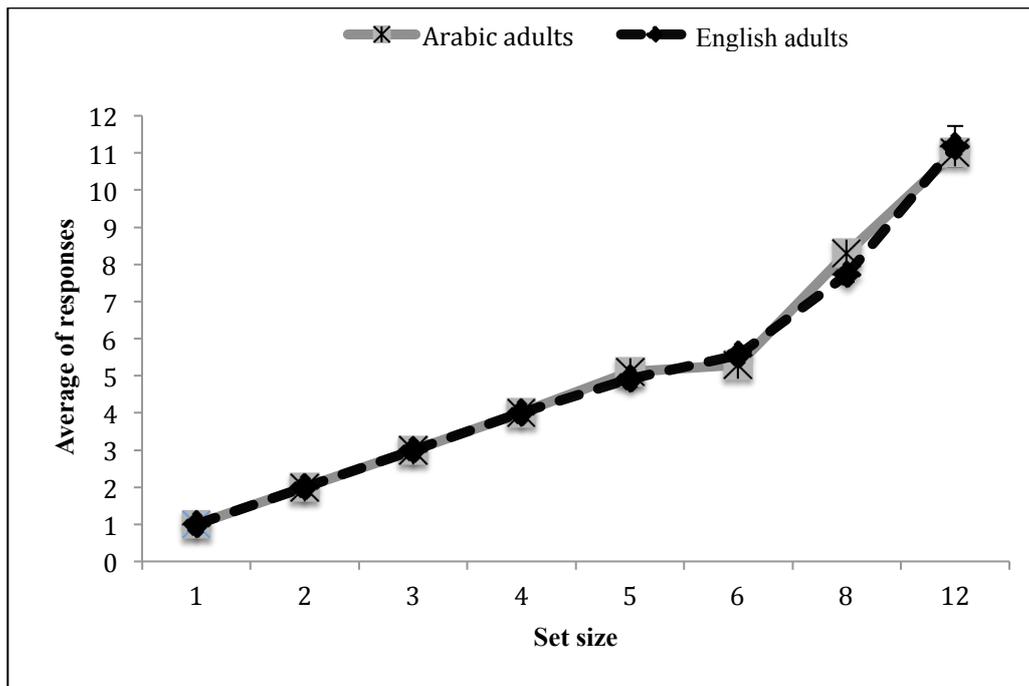


Figure 4.14. Average numeral by set size for Arabic and English adults in the estimating-magnitude-numerically task

The bivariate correlation results revealed no significant correlation between group and response (given number) ($r(\text{two-tailed})=0.007$, $p=0.83$) but a very strong correlation between set size and response ($r(\text{two-tailed})=0.98$, $p<0.001$). These results indicate that the two adult groups performed similarly on the different magnitudes.

4.3.2.5 The potential effect of pre-school learning on children's numeracy skills

I explored the effect of the amount of training with numbers children had had, as represented by their overall amount of pre-school learning. I asked the children's parents about the age at which their child had started attending nursery, playgroup, or school. Before going to the analyses, I should make it clear here that investigating children's amount of pre-school learning might not reliably predict the development of their numeracy skills (Anders et al. 2011), which are also influenced by other factors such as quality of home learning environment and quality of pre-school (Anders et al., 2011; Sammons et al., 2008). In addition, I did not ask parents to specify if the child attended a play-group, day-care or nursery. Since the educational materials might differ across these institutions, this is another reason for caution in drawing firm assumptions about the role of pre-schooling here. For example, learning in playgroup might be incidental compared with that in nursery or day-care, which

could provide a child with more advanced knowledge. Nevertheless, I think exploring children's amount of pre-school learning can help explain the weak numeracy skills of some of the Arabic children.

All the Arabic and English children's parents provided information on the age at which their child started attending pre-school. Since I decided to add this information only after collecting the bilingual children's data, I only included bilingual children whose parents had already provided this information when completing the questionnaire; as a result, one bilingual child was excluded from the analyses here.

To investigate the amount of pre-school learning, I first calculated the traditional length of pre-schooling for each child, by subtracting their age when entering pre-school from their age at testing; then I compared average age at pre-school onset and traditional length of pre-schooling of the three groups. Figure 4.15 displays the onset age and the traditional length of pre-schooling. It can be seen that the average age at onset for the bilingual and English children was around 2;6 years old, while the average age for Arabic children was around 4;5 years old. Thus, it is obvious that the Arabic children would have shorter length of pre-schooling, while the bilingual and English children would have very similar length of pre-schooling. This might indicate, leaving aside any other sources of learning, that the Arabic children have relatively limited training in numerical skills, which might be expected to be implemented in the pre-school years.

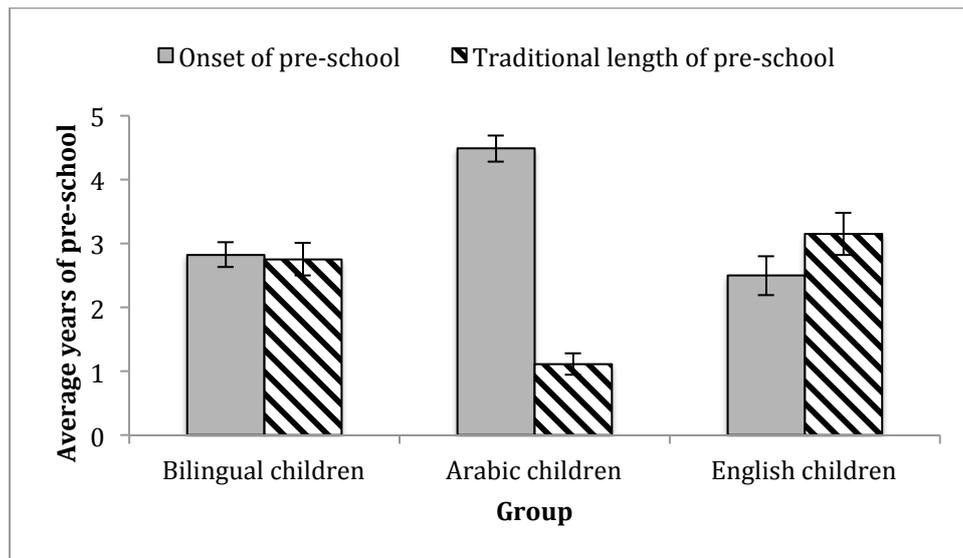


Figure 4.15. Children’s age at entering pre-school and their traditional length of pre-schooling. Error bars represent standard error of the mean

The analyses revealed that the assumptions of normality for each group and the whole sample were violated ($p < 0.05$); therefore, I made main comparisons with distribution-free tests, and checked their compatibility with the parametric ANOVA results. Comparisons with the Kruskal–Wallis H-test revealed significant differences between the groups’ age at pre-school entry ($H(2) = 32.32$, $p < 0.001$) and also their traditional length of pre-schooling ($H(2) = 31.65$, $p < 0.001$). The pair-wise comparisons revealed significant differences between the bilingual and Arabic children in onset ($U = 94.5$, $Z = -4.294$, $p < 0.001$) and length ($U = 106.5$, $Z = -4.988$, $p < 0.001$) and also between the English and Arabic children in onset ($U = 127.5$, $Z = -4.47$, $p < 0.001$) and length ($U = 110$, $Z = -4.60$, $p < 0.001$), but not between the bilingual and English children either in onset ($U = 305$, $Z = -1.25$, $p = 0.21$) or length ($U = 320$, $Z = -.961$, $p = 0.34$). These outcomes were completely consistent with the ANOVA results.

4.3.2.6 A summary of performance on number tasks

The analyses in this section explore the children’s and the adult controls’ performance on four numeral tasks: the how-many, give-a-number, non-verbal ordinal, and estimating-magnitude-numerically tasks.

The how-many task examined children’s ability to count two set sizes: {10} and {14}. The results showed that the majority (around 90%) of the bilingual and Arabic

children and all the English children were able to successfully count sets of size {10}, whereas for set size {14}, the analyses revealed that 77% of bilingual children, 80% of Arabic children, and 100% of English children completed the task correctly. The analyses also revealed that the difference between the bilingual and Arabic children was not statistically significant, while the differences between the bilingual and English children and between the Arabic and English children were significant. The adults showed a ceiling effect, with 100% correct responses.

The outcomes of the give-a-number task revealed that 100% of the bilingual and English children were able to produce sets that represented the given number in each trial (1, 2, 3, 4, 5, 6), as were around 83% of the Arabic children; for the children who did not complete the task correctly, the highest number that they could produce sets to represent was either 4 or 5. The differences between the Arabic children and the other two groups were significant. All adults completed the task successfully (100% correct responses).

Performance on the non-verbal ordinal task showed that all the children were successfully able to refer to the large set when given two sets in a context. Like the two adult groups, the children responded 100% correctly in all the trials.

The last task in this section was the estimating-magnitude-numerically task. The analysis of children's results revealed that, generally, all the child groups were able to manipulate their responses (numbers) linearly, increasing or decreasing them as needed as the magnitudes in a context were randomly changed. Statistically, the analysis revealed no significant differences among the bilingual, Arabic, and English children's performance on the task. In regard to the adults' performance, the outcomes showed that both Arabic and English adults were also able to increase their numeral responses as the magnitudes grew, and the two groups performed similarly on the task.

Finally, the analyses explored the potential effect of pre-school learning on children's numeracy skills, revealing that the Arabic children, who were the weakest group on the number tasks, had the shortest length of pre-schooling.

4.3.3. The relationship between numbers and quantifiers

The analysis in this section investigated whether there was any relationship between children's comprehension of 'most', 'some', 'or' and 'and' in experiment 1 and their performance on number tasks (the give-a-number task and the how-many task). In other words, I tried to explore what best predicts children's performance in the semantic comprehension task. Before going on to report the findings of the analysis, I should briefly justify why the estimating-magnitude tasks (numerically and proportionally) are excluded from the analysis. Although the production task might reflect, to some degree, children's ability to distinguish between various numerical and proportional scales by manipulating their response as the set-size increased, it clearly might not reflect exactly how children understand numbers and quantifiers, as the perception task does.

To examine the relationship, I first conducted a bivariate correlation to find out which of the independent variables correlate with children's scores in experiment 1: language group, receptive vocabulary, NVIQ, SES (FAS questionnaire), age, type of quantifier, and/or performance on number tasks (the give-a-number task and the how-many task). The results revealed strong positive correlations between children's semantic performance in experiment 1 and each of language group ($r(\text{two-tailed})=.149$, $p=0.001$), vocabulary score ($r(\text{two-tailed})=.234$, $p<0.001$), age ($r(\text{two-tailed})=.118$, $p=0.011$), type of quantifier ($r(\text{two-tailed})=.494$, $p<0.001$), how-many task ($r(\text{two-tailed})=.137$, $p=0.003$) and give-a-number task ($r(\text{two-tailed})=.096$, $p=0.039$). No significant correlation was found with SES or NVIQ.

Then, to determine which of these variables influenced children's performance in experiment 1, I conducted a GLM regression (with log link function for Poisson distribution) to predict the accuracy of each child's response. I first included all the variables that correlated with the dependent variable (the total score in experiment 1), and checked the goodness-of-fit. After this, I conducted two models, each with one of the variables that highly correlated with each other (age and vocabulary); the model's goodness-of-fit was better when age was included, so I removed vocabulary. Then, I added SES, which neither negatively nor positively affected goodness-of-fit; however, NVIQ improved the goodness-of-fit slightly, and having those two variables in the

model increased its goodness-of-fit, so I kept them. Finally, I was interested in identifying any possible relationship between numeracy and type of quantifier, so I added an interaction effect between the two tasks and the type of quantifier ('most' or 'some') and operator ('or' or 'and').

Thus, the final version of the model had language group (bilingual-in-English, bilingual-in-Arabic, Arabic, English) and the type of quantifier ('most', 'some', 'or', 'and') as fixed factors, and age, NVIQ, SES, the total score of each child in the give-a-number task, and the total score in the how-many task as covariates.

Table 4.17 shows the results of the regression analysis, which reveal a significant main effect of language group ($X^2(3)=62.915$, $p<0.001$); the model negative estimates (B) indicates that the English children performed better than the other three groups. There was also a significant main effect of quantifier type ($X^2(3)=20.462$, $p<0.001$); the model estimates that the children had significantly poorer performance on 'most' and 'some' when compared with 'and', but that performance on 'or' did not significantly differ from that on 'and'. The regression results also revealed no significant effect of any of the give-a-number task, the how-many task, SES, or NVIQ (all $ps<0.05$), but there was a significant effect of age ($X^2(1)=8.317$, $p<0.05$). An interaction effect between the numerical tasks and quantifier type was only found with the give-a-number task ($X^2(3)=16.77$, $p<0.005$) and was only significant with the quantifier 'some' ($X^2(3)=9.155$, $p<0.05$)³.

³ *Note:* Due to the potential effect of WM in processing quantifiers (e.g. Heim et al., 2016), I fitted another model exactly like the one in table 4.17 but with the addition of STM, since it correlated significantly with experiment 1 results ($r(\text{two-tailed})=.152$, $p=0.001$). The model revealed exactly the same results as the original one with the addition of a significant effect of STM ($X^2=8.31$, $p=0.004$, $B=.123$). Adding an interaction effect between STM and quantifier type and one between STM and language group neither revealed any significant interaction effect nor improved the model's goodness-of-fit. Ultimately, due to the structure of this thesis, specifically the fact that I included results for STM only in study 2, I avoid including it in the model at this stage.

Table 4.17. GLM regression results: Predictors for children’s semantic performance (experiment 1).

Predictor	B	SE	X²	DF	P
(Intercept)	4.36	1.5997	7.420	1	.006
Language group	.	.	62.915	3	.000
Bilingual-in-English	-.654	.1352	23.418	1	.000
Bilingual-in-Arabic	-.888	.1352	43.094	1	.000
Arabic	-.992	.1379	51.786	1	.000
English	0
Quantifier/operator type	.	.	20.462	3	.000
Most	-4.35	2.1556	4.076	1	.043
Some	-7.04	2.1556	10.666	1	.001
Or	1.65 5	2.1556	.590	1	.443
And	0
SES (FAS)	.	.	3.640	2	.162
SES (Low)	.298	.1981	2.259	1	.133
SES (Medium)	.144	.1000	2.066	1	.151
SES (High)	0
Age (month)	.019	.0064	8.317	1	.004
NVIQ	.040	.0232	2.918	1	.088
How-many task	-.292	.2742	1.137	1	.286
Give-a-number task	.094	.1707	.301	1	.583
Quantifier/operator*Give-a-number	.	.	16.772	3	.001
Most*Give-a-number task	.448	.3778	1.408	1	.235
Some*Give-a-number task	1.14	.3778	9.155	1	.002
Or*Give-a-number task	-.313	.3778	.686	1	.408
And*Give-a-number task	0
Quantifier/operator*How-many	.	.	.588	3	.899
Most*How-many task	-.035	.2341	.022	1	.881
Some*How-many task	-.135	.2341	.335	1	.563
Or*How-many task	.034	.2341	.021	1	.884
And*How-many task	0

Note: Model references levels were, for group, the English children; for quantifier/operator, the conjunction ‘and’; and for SES, High

4.4 Results of Study (2): The potential effect of bilingualism on pragmatic competence

4.4.1 Pragmatic performance: Ternary-response investigation

For better understanding of children's performance in the two pragmatic tasks, the analyses of pragmatic performance were conducted on two levels. First, I explored (using descriptive and statistical analysis) the participants' ternary responses; then, I rescored the 3-point scale as a binary response. The ternary-response analyses allow the examination of variation in pragmatic behaviour across the three groups: do the children completely reject under-informative items, or partially penalise them (with the medium strawberry)? For such analyses, non-parametric tests are employed, as the responses are categorical—coded 1, 2, 3 for small, medium, and large strawberries, respectively.

In the second type of analysis, only the results of the critical (under-informative) condition are examined, in binary fashion: 'small' and 'medium' responses are given scores of 1, and 'large' responses, 0. The rationale for this re-scoring is that both 'medium' and 'small' responses convey some pragmatic penalisation of the violation of the Gricean maxim of informativeness, regardless of the strength or the degree of this penalisation. 'Large' responses are scored 0 since they indicate complete insensitivity to the violation of informativeness.

4.4.1.1 Results for experiment 3 (enriched context)

Children's performance

This section starts with descriptive analyses of the results, then compares performance across groups to find out if the differences are statistically significant. The between-group comparisons include two levels: the first captures the differences between group performance in the three conditions (under-informative, optimal, false), and the second, the pragmatic differences for each of the quantifiers 'most', 'some', 'or', and 'and'.

Within-group variation (descriptive analyses)

Table 4.18 breaks down by proportion the responses given by the bilingual, Arabic and English children for all three experimental conditions. Starting with the bilingual children, it can be seen that, in the optimal condition, they accepted the items at a high rate only with ‘some’ or ‘and’, and that the percentages were higher overall when they were tested in English. With optimal ‘most’, the bilingual children showed a moderately high rate of acceptance (more than 60%), although they generally did better in Arabic than in English, with less frequent occurrence of complete rejection (choice of ‘small’). As for optimal ‘or’, bilingual children penalised it (with partial or complete rejection) at a high rate, as the table shows. It should be clarified here that higher rate of acceptance (‘large’ response) for optimal items and full rejection (‘small’ response) for false items is the expected and appropriate result, since optimal and false items were included as control items to explore children’s ability to accept logically and pragmatically appropriate utterances and reject logically and pragmatically inappropriate ones in the false condition.

When comparing the bilinguals’ performance in Arabic with that of the Arabic children, it can be seen that the Arabic children accepted optimal ‘some’, ‘most’ and ‘and’ at a higher rate than the bilingual children did. Although optimal ‘or’ received a lower acceptance rate than other quantifiers in the bilingual (both languages) and Arabic children, the percentage in the Arabic children group almost tripled that in the bilingual children’s groups in English and in Arabic.

With respect to the under-informative condition, table 4.18 shows that for ‘some’ and ‘most’, the bilingual children (in both English and Arabic) accepted only around half of the under-informative items, penalising the other half with partial or complete rejection (‘some’: 27.5% ‘medium’, 30% ‘small’; ‘most’: 21.7% ‘medium’, 20.8% ‘small’). Under-informative ‘or’ was less frequently penalised than other quantifiers, though penalisation was slightly higher when the bilinguals were tested in Arabic (in English, 13% ‘medium’, 22.7% ‘small’; in Arabic, 28% ‘medium’, 20% ‘small’). The results for the Arabic children showed a very weak tendency to penalise under-informative items using the quantifiers ‘some’, ‘most’, and ‘or’ (total ratio of partial or complete rejection of these did not exceed 15%).

With under-informative ‘and’, around two-thirds of the bilingual children penalised (with partial or full rejection) items in this condition, in both languages, though higher percentages were recorded for partial penalisation (in English, 43.3% ‘medium’, 30.8% ‘small; in Arabic, 49.2% ‘medium’, 25% ‘small’). The Arabic children’s behaviour with ‘and’ was completely different from their performance on ‘some’, ‘most’ and ‘or’; overall, they penalised half of the under-informative items (17.5% ‘medium’, 33% ‘small’).

In the false condition, it can be seen that the bilingual children rejected more than 88% of items with all quantifiers except false ‘and’, especially with a small-choice response (English: 55% ‘small’, 23.3% ‘medium; Arabic: 40% ‘small’, 35% ‘medium’). The Arabic children also rejected false ‘most’ and ‘some’ more than 90% of the time, but showed a lower rate of complete rejection with false ‘or’ (78% ‘small’, 11.7% ‘medium’) and false ‘and’ (60% ‘small’, 15% ‘medium’).

As for the English children, table 4.18 below shows that they accepted optimal ‘some’ and ‘most’ at a high rate (around 86%). Optimal ‘or’ was accepted only at 30%, as half of total responses partially rejected it (‘medium’ response) and the remaining ones rejected it completely (17%). Optimal ‘and’, as in all the other groups, received the highest proportion of ‘large’ ratings among the quantifiers/operators (around 90%).

In the under-informative condition, around two-thirds of the English children’s responses to ‘some’ and ‘most’ showed complete acceptance, the ‘large’ choice. Correspondingly, lower ratios of penalisation (with either ‘medium’ or ‘small’) were given to the critical items for these two quantifiers, and it can be seen that the percentage for partial rejection doubles that for complete rejection. The table also reveals that the majority of the children in this group accepted under-informative ‘or’ at a high rate (83%), partially penalising it at only 10% and completely rejecting it at a rate of only 8%. As for under-informative ‘and’, the English children penalised critical items at around 60%, at a higher rate for partial (37%) than complete (22%) rejection.

Finally, in the false condition, the English children rejected false 'some' and 'most' (with the 'small' response) at a very high rate (around 87%), while this percentage declined with 'or' (around 80%). False 'and' received a relatively low rate of complete rejection (40%) and a correspondingly higher rate of partial rejection (33%). Thus, approximately one-third of the English children's responses accepted false 'and'.

Table 4.18. Bilingual, Arabic and English children’s responses (as a percentage of each type of utterance) in experiment 3

Utterance	Type of response	Bilingual children—English				Bilingual children—Arabic				Bidialectal children—Arabic				Monolingual children—English			
		Some %	Most %	Or %	And %	Some %	Most %	Or %	And %	Some %	Most %	Or %	And %	Some %	Most %	Or %	And %
Optimal	Large	88.3	63.3	20	95	81.7	70	16.7	86.7	86.7	88.3	56.7	93.3	86.4	86.5	30.4	90.4
	Medium	3.3	13.3	40	3.3	13.3	23	31.7	8.3	6.7	6.7	23.3	1.7	9.6	5.8	51.9	4
	Small	8.3	23.3	40	1.7	5	6.7	51.7	5	6.7	5	20	5	5.8	7.7	17.3	5.8
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Under-info	Large	42.5	57.5	65.0	25.8	50	57.5	51.7	25.8	86.7	84.2	85	49.2	65.4	60.6	82.7	41.3
	Medium	27.5	21.7	13.3	43.3	20.8	21.7	28.3	49.2	5	7.5	7.5	17.5	22.1	28.8	10	36.5
	Small	30	20.8	21.7	30.8	29.17	20.8	20	25	8.3	8.3	7.5	33.3	12.5	10.6	7.7	22
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
False	Large	1.7	0	5	21.7	0	3.3	3.3	25	1.7	2	10	25	5.8	8	7.7	26.9
	Medium	0	1.7	1.7	23.3	1.7	0	8.3	35	5	6.7	11.7	15	8	9.6	13.5	32.7
	Small	98.3	98.3	93.3	55	98.3	96.7	88.3	40	93.3	91.7	78.3	60	86.5	82.7	78.8	40.4
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Note: Each quantifier is presented alongside 4 critical items (under-info=under-informative) and 4 control items (2 optimal and 2 false). The use of the term under-informative ‘and’ is meant to refer to ‘ad hoc’ scale.

Between-group variation (inferential analyses)

To find out if the groups differed statistically, I first ran a Kruskal–Wallis test (an alternative to the parametric 1-way ANOVA used to compare distribution across groups; its outcomes either refute or prove the null hypothesis) with group as an independent variable (bilingual-English, bilingual-Arabic, Arabic bidialectal, English monolingual) and response type as a dependent variable (large, medium, small). The test results revealed that the distribution of responses in experiment 3 was not the same across group categories ($p < 0.001$). To investigate if the groups differed in their rating of each quantifier, I ran the same test but with quantifier/operator as an independent variable ('most', 'some', 'and', 'or') and response type again as a dependent variable (as above). The outcomes showed that the distribution of responses in experiment 3 was not the same across quantifiers ($p = 0.021 < 0.05$). Finally, to find out if the groups differed in types of response given in each condition, I conducted a Kruskal–Wallis test with condition as an independent variable (under-informative, optimal, false) and response as dependent variable; the results showed that the distribution was not the same for all conditions ($p < 0.001$).

Since the previous tests had already demonstrated that the groups statistically and systemically differed in judging different conditions and also in rating different quantifiers/operators, it was important to break down these differences in more detail. First, to understand potential differences among the variables, and taking into account that the responses were ranked on a 3-point scale, pair-wise comparisons were conducted between each pair of groups separately (since non-parametric tests did not stand as an alternative to mixed ANOVA with post hoc pair-wise comparison; see Larson-Hall [2010]).

Between-group comparisons by quantifier/operator

To find out in which quantifier/operator the participant groups differed, separate pair-wise between-group comparisons were conducted for the quantifiers and operators across the three conditions.

Pragmatic performance on ‘most’ in experiment 3

Starting with the bilingual group, a comparison between performance in English and Arabic (using the Wilcoxon signed-rank test) for the quantifier ‘most’ revealed no significant differences between the bilinguals’ performance in the under-informative and in the false conditions ($Z=-.18$, $p=0.86$ and $Z=-1.09$, $p=0.28$, respectively), but did reveal a significant difference in the optimal condition ($Z=-2.26$, $p=0.024$). The pair-wise comparisons (using the Mann–Whitney U-test) between bilingual-in-Arabic and Arabic children revealed a significant difference in pragmatic comprehension of ‘most’ in the under-informative condition ($U=5370$, $Z=-4.238$, $p<0.001$), with the bilinguals performing better, and the optimal condition ($U=1483$, $Z=-2.352$, $p=0.019$), with the Arabic children accepting more optimal items with a ‘large’ response, but not in the false condition ($U=1685$, $Z=-1.396$, $p=0.16$). A comparison between bilingual-in-English and English children revealed no significant difference in under-informative ‘most’ ($U=5749$, $Z=-1.15$, $p=0.25$), with the bilinguals numerically performing better, but a significant difference in optimal ‘most’ ($U=1165$, $Z=2.97$, $p=0.003$), with the English children accepted more optimal items, and in false ‘most’ ($U=1314$, $Z=-2.90$, $p=0.004$), where the bilinguals rejected more false items. When comparing Arabic and English children, the test revealed a significant difference in under-informative ‘most’ ($U=4935$, $Z=-3.46$, $p=0.001$), with the English children performing better, but not in optimal ‘most’ ($U=1528$, $Z=-.32$, $p=0.75$) or false ‘most’ ($U=1438$, $Z=-1.19$, $p=0.23$).

Pragmatic performance on ‘some’ in experiment 3

The analyses revealed no significant difference between bilingual performance in English and Arabic for under-informative ‘some’ ($Z=-.583$, $p=0.56$), optimal ‘some’ ($Z=-.144$, $p=0.89$), or false ‘some’ ($Z=-.447$, $p=0.89$). The bilinguals’ performance in Arabic significantly differed from the Arabic children’s performance for under-informative ‘some’ ($U=4507$, $Z=-6.132$, $p<0.001$), with the bilinguals performing better, but not for optimal ‘some’ ($U=1720$, $Z=-.662$, $p=0.51$) or false ‘some’ ($U=1709$, $Z=-1.37$, $p=0.17$). The bilingual-in-English and English children’s results revealed significant differences for under-informative ‘some’ ($U=468$, $Z=-3.57$, $p<0.001$), with the bilinguals performing better, and false ‘some’ ($U=138$, $Z=-2.38$, $p=0.017$), where the bilinguals rejected more false items, but not for optimal ‘some’ ($U=1511$, $Z=-0.478$, $p=0.63$). Finally, the English and Arabic children significantly

differed on under-informative ‘some’ ($U=9424$, $Z=-3.71$, $p<0.001$), with the English children performing better, but not optimal ‘some’ ($U=1532$, $Z=-0.269$, $p=0.79$) or false ‘some’ ($U=1451$, $Z=-1.23$, $p=0.22$).

Pragmatic performance on ‘or’ in experiment 3

First, a comparison between the bilingual children’s performance in the two languages revealed no significant differences for under-informative ‘or’ ($Z=-1.05$, $p=0.29$), optimal ‘or’ ($Z=-1.1$, $p=0.27$), or false ‘or’ ($Z=-.347$, $p=0.73$). When comparing bilingual performance in Arabic with that of bidialectal Arabic children, significant differences were found for under-informative ‘or’ ($U=4845$, $Z=-5.34$, $p<0.001$), the bilinguals penalised more items, and optimal ‘or’ ($U=977$, $Z=-4.59$, $p<0.001$), the Arabic children accepted more optimal items, but not false ‘or’ ($U=1612$, $Z=-1.52$, $p=0.13$). Bilingual performance in English and that of the English children significantly differed in under-informative ‘or’ ($U=5070$, $Z=-3.12$, $p=0.002$), the bilinguals performed better, optimal ‘or’ ($U=1176$, $Z=-2.41$, $p=0.016$), the English accepted more optimal items, and false ‘or’ ($U=1342$, $Z=-2.15$, $p=0.032$), with similar rate of penalisation that differed only qualitatively. Finally, comparison between the Arabic and English children showed no significant difference in under-informative ‘or’ ($U=6105$, $Z=-0.44$, $p=0.66$), a marginal difference in optimal ‘or’ ($U=1255$, $Z=-1.92$, $p=0.054$), the Arabic children accepted more optimal items, and no difference in false ‘or’ ($U=1545$, $Z=-0.122$, $p=0.90$).

Pragmatic performance on ‘and’ (‘ad hoc’ scale) in experiment 3

A comparison between the bilingual children’s performance in Arabic and in English revealed a significant difference in all of under-informative ‘and’ ($Z=-.573$, $p=0.57$), optimal ‘and’ ($Z=-1.42$, $p=0.15$), and false ‘and’ ($Z=-1.48$, $p=0.14$), their performance, however, differed qualitatively rather than quantitatively. When comparing bilingual performance in Arabic with that of the Arabic children, in contrast, the comparison revealed no significant differences in under-informative ‘and’ ($U=6385$, $Z=-1.61$, $p=0.107$), optimal ‘and’ ($U=1686$, $Z=-1.15$, $p=0.25$), or false ‘and’ ($U=1530$, $Z=-1.54$, $p=0.12$). Bilingual performance in English and that of monolingual English children significantly differed for under-informative ‘and’ ($U=5167$, $Z=-2.37$, $p=0.018$), the bilinguals performed better, but not for optimal ‘and’ ($U=1486$, $Z=-.967$, $p=0.33$) or false ‘and’ ($U=1325$, $Z=-1.48$, $p=0.137$). Finally, comparison between the Arabic and English children showed no significant difference

in under-informative ‘and’ (U=6209, Z=-0.068, p=0.95), optimal ‘and’ (U=1515, Z=-.551, p=0.58), or false ‘and’ (U=1318, Z=-1.53, p=0.12).

Adults’ performance

Table 4.19 displays the percentages of the two adult groups’ ternary responses for the four quantifiers/operators (‘most’, ‘some’, ‘or’, ‘and’) across the three pragmatic conditions (optimal, under-informative, false). In the optimal condition, it can be seen that that Arabic and English adult participants rated ‘most’, ‘some’, and ‘and’ with the ‘large’ response (100%, except the English for ‘and’ at 96%). However, the two groups did not accept optimal ‘or’ at a high rate, as they did the other quantifiers and the conjunction ‘and’: only half of the Arabic adults’ responses rated ‘or’ large, while they partially penalised it at 40% and completely rejected it at 10%. The English adults’ performance on optimal ‘or’ was similar, though with a slightly higher rate of penalisation. As the table shows, only around one-third of the English adults’ responses were ‘large’ (36%), while most of the remaining two-thirds conveyed partial penalisation (59%); very few responses completely rejected optimal ‘or’.

Table 4.19. Percentages of Arabic and English adults’ ternary responses in the three conditions of experiments 3

Utterance	Type of response	Adults—Arabic				Adults—English			
		Some %	Most %	Or %	And %	Some %	Most %	Or %	And %
Optimal	Large	100	100	50	100	100	100	36.4	95.5
	Medium	0	0	40	0	0	0	59.1	4.5
	Small	0	0	10	0	0	0	4.5	0
	Total	100	100	100	100	100	100	100	100
Under-info	Large	0	0	0	0	4.5	0	9	22
	Medium	25	22.5	12.5	52.5	84.1	88.6	66	64
	Small	72	77.5	87.5	47.5	11.4	11.4	25	14
	Total	100	100	100	100	100	100	100	100
False	Large	0	0	0	5	0	0	0	0
	Medium	0	0	0	15	0	0	0	45
	Small	100	100	100	80	100	100	100	55
	Total	100	100	100	100	100	100	100	100

Note: Each quantifier is presented alongside 4 critical items (under-informative) and 4 control items (2 optimal and 2 false). The use of the term under-informative ‘and’ is meant to refer to ‘ad hoc’ scale.

With respect to the adult groups' performance in the under-informative condition, table 4.18 shows that, for 'most' and 'some', around two-thirds of the Arabic adults' responses indicated complete rejection ('small'), and the remaining third, partial rejection. The English adults, on the other hand, partially rejected under-informative 'most' and 'some' at a high rate (around 84%) and completely rejected them only at a small rate (11.4%). On under-informative 'or', the Arabic participants rejected at a very high rate (88%), and partially rejected 12% of the items. Around two-thirds of the English adults' responses on under-informative 'or' consisted of partial rejection (66%), 25% completely rejected it, and 9% did not penalise it ('large' responses). The two groups' performance on under-informative 'and' revealed that half of the Arabic adults' responses rejected it completely ('small' response) and the other half partially ('medium' response), while two-thirds of the English adults' responses consisted of partial rejection (64%) and 14%, complete rejection.

Finally, the adult groups' performance on the false condition was quite similar. All Arabic and English responses on false 'most', 'some', and 'or' were 100% complete rejection ('small' response). The two groups' performance on 'and' was slightly different: the Arabic adults rejected false 'and' ('small' response) at a high rate (80%) and partially rejected it at a rate of 15%, while the English adults' responses partially rejected half of the false items, and the other half completely.

Performance on the under-informative items in experiment 3

Since the main aim of the current study is to explore participants' (development of) sensitivity to the violation of the Gricean principle of informativeness (using quantifiers), I conducted pair-wise comparisons to investigate whether the adult groups significantly differ only in the under-informative condition. The results of the Mann–Whitney U-test revealed that the Arabic and English adults' performance statistically and significantly differed on the under-informative 'most' ($U=298$, $Z=-6.08$, $p<0.001$), the under-informative 'some' ($U=310$, $Z=-5.89$, $p<0.001$), the under-informative 'or' ($U=320$, $Z=-5.72$, $p<0.001$), and the under-informative 'and' ($U=477$, $Z=-4.14$, $p<0.001$). In general, they differed qualitatively (i.e. the type of penalisation) rather than quantitatively.

4.4.1.2 Results for experiment 4 (no context)

The analyses in this section follow the same procedure that was applied to the results of experiment 3.

Children's performance

Within-group variation (descriptive analyses)

Table 4.20 below shows the proportions of responses given for each quantifier/operator across the three conditions (*infelicitous*, *felicitous*, and *bizarre*) by the bilingual (in English and Arabic), the English and the Arabic children. Starting with the felicitous condition, when comparing the bilingual children's performance on the quantifiers/operators between the two languages, it can be noticed that the children accepted the items in this condition (with the 'large' choice) at approximately the same rate. In both languages, they tended to accept felicitous 'some' more than 'most' and the conjunction 'and' more than 'or'. With respect to the Arabic children, they accepted the felicitous items for all quantifiers/operators at a high rate (around 80%); comparing their performance on each of the quantifiers and operators, it can be seen that they accepted 'some' at a higher rate than 'most', and accepted 'and' at a higher rate than 'or'.

Moving to the infelicitous condition, it can be seen that for 'some' and 'most', around 50% of the bilingual children's responses (in English and Arabic) penalised the items in this condition, divided roughly equally between complete rejection ('small' response) and partial rejection ('medium' response). Looking at the Arabic children's performance in this condition, we see that the majority of these children accepted infelicitous items using 'some', 'most' and 'or' at a high rate (around 80%), but a slightly lower rate of acceptance was given to infelicitous 'and' (70%). As the table shows, the Arabic children penalised infelicitous 'some' (medium: 2.5%, small: 17.5%) at a clearly higher rate than 'most' (medium: 8.3%, small: 1.8%); a similar difference was found when comparing 'and' (medium: 7.5%, small: 22.5%) and 'or' (medium: 5.8%, small: 12.5%). Finally, the bilingual children's performance in the bizarre condition shows that, for all the quantifiers/operators, the children rejected bizarre items at a higher rate (more than 90% for all quantifiers except 'and' in Arabic at 86%). The Arabic children also rejected the bizarre items at a very high rate (more

than 80%). For bizarre ‘some’, ‘most’, and ‘or’, they rejected more than 90% of the items, while ‘and’ received a slightly lower rate of rejection (83%).

With respect to the English children’s performance on the felicitous condition, it can be seen that the children accepted the quantifiers/operators at a very high rate. As table 4.20 shows, they tended to accept felicitous ‘some’ slightly more than ‘most’ (87% and 71%, respectively), and the conjunction ‘and’ more than the disjunction ‘or’ (92%, and 85%, respectively). In the infelicitous condition, one-third of the English children’s responses penalised infelicitous ‘some’, ‘most’ and ‘or’, while approximately two-thirds of their responses penalised infelicitous ‘and’. They tended to partially penalise ‘some’, with the medium choice (23%), more than completely rejecting it with the small choice (14%). Similar results were found for ‘most’ and ‘or’: partial penalisation (around 18%) was seen more than complete rejection (approximately 12%). For infelicitous ‘and’, the English children chose complete rejection (31%) at a marginally higher rate than partial rejection (25%). Finally, the English children rejected the bizarre items at a very high rate, with bizarre ‘some’ and ‘most’ (96%) at a slightly higher rate than bizarre ‘or’ and ‘and’ (92%).

Table 4.20. Bilingual, Arabic and English children’s responses (as a percentage of each type of utterance) in experiment 4

Utterance	Type of response	Bilingual children—English				Bilingual children—Arabic				Bidialectal children—Arabic				Monolingual children—English			
		Some %	Most %	Or %	And %	Some %	Most %	Or %	And %	Some %	Most %	Or %	And %	Some %	Most %	Or %	And %
Felicitous	Large	73.3	60.0	66.7	91.7	70	65	68.3	96.7	90	86.7	78.3	96.7	86.5	71.2	84.6	92.3
	Medium	16.7	18.3	25	6.7	15	20	16.7	1.7	3.3	3.3	5	3.3	5.8	13.5	13.5	5.8
	Small	16.7	21.7	8.3	1.7	15	15	15	1.7	6.7	10	16.7	0	7.7	15.4	1.9	1.9
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Infelicitous	Large	52.5	53.3	70.8	54.2	56.7	51.7	63.3	57.5	80	81	81.7	70	63.5	70.2	68.3	44.2
	Medium	25	21.7	14.2	25.8	22.5	20.8	23.3	21.7	2.5	8.3	5.8	7.5	23	18.3	20.2	25
	Small	22.5	25.0	15.0	20.0	20.8	27.5	13.3	20.8	17.5	1.8	12.5	22.5	13.5	11.5	11.5	30.8
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Bizarre	Large	1.7	0	6.7	3.3	0	0	5	8.3	6.7	3.3	5	15	0	1.9	5.8	1.9
	Medium	0	1.7	1.7	1.7	5	0	5	5	1.7	3.3	0	1.7	3.8	1.9	1.9	5.8
	Small	98.3	98.3	91.7	95	95	100	90	86.7	91.7	93.3	95	83.3	96.2	96.2	92.3	92.3
	Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Note: Each quantifier is presented alongside 4 critical items (infelicitous) and 4 control items (2 felicitous and 2 bizarre). The use of the term infelicitous ‘and’ is meant to refer to ‘encyclopaedic’ scale.

Between-group variation (inferential analyses)

Exactly the same procedure used in the analyses conducted for experiment 3 will be applied in this section (to the new set of conditions). First, to explore if the groups differed statistically, the Kruskal–Wallis test was conducted, with group as an independent variable (bilingual-English, bilingual-Arabic, Arabic bidialectal, English monolingual) and response type as a dependent variable (large, medium, small). The outcomes revealed that the overall response distribution in experiment 3 varied across groups ($p < 0.001$). To investigate if and how the groups differed in the rating of each quantifier/operator, another Kruskal–Wallis test was performed, but with quantifier/operator as an independent variable (most, some, and, or) and response type as a dependent variable (large, medium, small). The test results showed a marginally significant difference in the distribution of responses in experiment 4 across categories of quantifier/operator ($p = 0.069 < 0.05$). Finally, another Kruskal–Wallis test, with condition as an independent variable (infelicitous, felicitous, bizarre) and response as a dependent variable (large, medium, small) revealed that the difference between groups in responses given in each condition was significant ($p < 0.001$).

Between-group comparisons by quantifier/operator

Pragmatic performance on ‘most’ in experiment 4

Regarding between-group differences in quantifier/operator type, the Wilcoxon signed-rank test for the paired sample bilingual-in-English v. -in-Arabic revealed no significant difference in performance on infelicitous ‘most’ ($Z = .206$, $p = 0.84$), felicitous ‘most’ ($Z = -.967$, $p = 0.33$) or bizarre ‘most’ ($Z = 1.0$, $p = 0.32$). The results of pair-wise comparison between the bilingual children in Arabic and the Arabic children, however, did reveal significant differences in performance using ‘most’ in the infelicitous condition ($U = 5097.5$, $Z = -4.676$, $p < 0.001$), the bilinguals penalised more items, the felicitous condition ($U = 1437$, $Z = -2.545$, $p = 0.011$), the Arabic children accepted more felicitous items, and the bizarre condition ($U = 1680$, $Z = -2.025$, $p = 0.043$), the bilinguals rejected all the bizarre items. There were also significant differences between the bilinguals’ performance in English and that of the English children on infelicitous ‘most’ ($U = 5059$, $Z = -2.81$, $p = 0.005$), the bilinguals penalised more items, but not felicitous ‘most’ ($U = 1384$, $Z = -1.21$, $p = 0.22$) or bizarre ‘most’

($U=1525$, $Z=-.72$, $p=0.47$). Finally, a comparison between the Arabic and English children revealed no significant difference for infelicitous ‘most’ ($U=5639$, $Z=-1.66$, $p=0.097$), only a marginal significant difference for felicitous ‘most’ ($U=1331$, $Z=-1.89$, $p=0.058$), they differed qualitatively not quantitatively, and non-significant difference for bizarre ‘most’ ($U=1516$, $Z=-.658$, $p=0.51$).

Pragmatic performance on ‘some’ in experiment 4

Comparing the performance of bilingual children on ‘some’ across the two languages the results showed no significant difference for infelicitous ‘some’ ($Z=-.737$, $p=0.46$) felicitous ‘some’ ($Z=-.82$, $p=0.41$), or bizarre ‘some’ ($Z=-.378$, $p=0.71$). A comparison between bilinguals’ performance in Arabic and that of Arabic children, in contrast, revealed significant differences for infelicitous ‘some’ ($U=5766$, $Z=-3.253$, $p=0.001$), the bilinguals penalised more infelicitous items, and felicitous ‘some’ ($U=1449$, $Z=-2.643$, $p=0.008$), the Arabic children accepted more items, but not bizarre ‘some’ ($U=1734$, $Z=-.801$, $p=0.423$). The results also revealed a marginal significant difference between bilinguals’ performance in English and that of English children on infelicitous ‘some’ ($U=5442$, $Z=-1.85$, $p=0.063$), the bilinguals penalised more items, but no significant difference on felicitous ‘some’ ($U=1365$, $Z=-1.615$, $p=0.106$) or bizarre ‘some’ ($U=1527$, $Z=-0.688$, $p=0.49$). Finally, when comparing the Arabic with the English children, a significant difference was found on infelicitous ‘some’ ($U=5439$, $Z=-2.109$, $p=0.035$), the English children penalised more items, but not felicitous ‘some’ ($U=1508$, $Z=-.546$, $p=0.59$) or bizarre ‘some’ ($U=1486$, $Z=-1.03$, $p=0.30$).

Pragmatic performance on ‘or’ in experiment 4

First, there was no statistically significant difference in the performance of bilingual children between the two languages for infelicitous ‘or’ ($Z=-.581$, $p=0.79$), felicitous ‘or’ ($Z=-.273$, $p=0.79$), or bizarre ‘or’ ($Z=-.489$, $p=0.63$). Next, when comparing the performance of bilingual children in Arabic with that of the Arabic children, a significant difference was found for infelicitous ‘or’ ($U=6034$, $Z=-2.77$, $p=0.006$), the bilinguals penalised more items, but not for either felicitous ‘or’ ($U=1681$, $Z=-.812$, $p=0.42$) or bizarre ‘or’ ($U=1714$, $Z=-.983$, $p=0.33$). The comparison of bilinguals’ performance in English and that of the English children revealed no significant

difference in infelicitous ‘or’ (U=6167, Z=-.186, p=0.85), the two groups have very similar rates of penalising infelicitous items, a significant difference in felicitous ‘or’ (U=1270, Z=-2.24, p=0.025), the English children accepted more items, and no difference in bizarre ‘or’ (U=1523, Z=-.437, p=0.66). There was also a significant statistical difference between the Arabic and English children on infelicitous ‘or’ (U=5519, Z=-1.98, p=0.048), the English children penalised more items, but not felicitous ‘or’ (U=1428, Z=-1.129, p=0.26) or bizarre ‘or’ (U=1519, Z=-.563, p=0.57).

Pragmatic performance on ‘and’ (encyclopaedic’ scale) in experiment 4

The results of the analyses showed no significant difference between the bilingual children’s performance in Arabic and English on infelicitous ‘and’ (Z=-.195, p=0.85), felicitous ‘and’ (Z=-.791, p=0.43), or bizarre ‘and’ (Z=-1.833, p=0.067). When the performance of the bilingual children in Arabic was compared with that of the Arabic children, no significant difference was found in any of infelicitous ‘and’ (U=6538, Z=-1.44, p=0.15), the bilinguals numerically penalised more instances of infelicitous ‘and’, felicitous ‘and’ (U=1799, Z=-.107, p=0.99), or bizarre ‘and’ (U=1729, Z=-.601, p=0.54). Further, when the performance of the bilingual children in English was compared with that of the English children, a marginal significant difference was found between the two groups in infelicitous ‘and’ (U=5436, Z=-1.81, p=0.071), the English children numerically penalised more infelicitous items, felicitous ‘and’ (U=1550, Z=-.118, p=0.91), or bizarre ‘and’ (U=1493, Z=-1.002, p=0.32). Finally, a comparison between the Arabic and the English children showed a significant difference in infelicitous ‘and’ (U=4839, Z=-3.27, p=0.001), the English children penalised it more, but not felicitous ‘and’ (U=1491, Z=-1.03, p=0.30) or bizarre ‘and’ (U=1407, Z=-1.56, p=0.12).

Adults’ performance

The results for the two adult groups in experiment 4 are presented in table 4.21. In the felicitous condition, the majority of the Arabic adults rated ‘most’ and ‘some’ with ‘large’ more than 90% of the time, and ‘or’ and ‘and’, 100% of the time. The English adults’, on the other hand, rated felicitous ‘some’ and ‘and’ with ‘large’ 100% of the

time, but lower percentages of 'large' ratings were found in their responses to 'most' (77%) and 'or' (81%).

The groups' performance in the infelicitous (that is, pragmatically under-informative) condition was to some extent varied, as the table shows. For 'most' and 'some', the majority of the Arabic adults completely rejected the infelicitous items (80% 'small' response), and most of the remaining responses indicated partial rejection (with 'medium' response). The English adults' responses on 'most' favoured partial rejection (61%), selected twice as much as complete rejection. Also, on 'some', the English adults partially rejected half of the critical items (55% 'medium' response), with their remaining responses divided nearly equally between complete rejection (25% 'small' response) and acceptance (20% 'large' response). Regarding the two groups' performance on 'or', it can be seen that the Arabic adults rejected two-thirds of the critical items completely, and the remaining one-third partially. In contrast, two-thirds of the English adults' responses conveyed partial rejection (68%), and most of the responses in the remaining one-third, complete rejection (27%). Finally, the adults' performance on infelicitous 'and' revealed that approximately half the Arabic adults' responses conveyed complete rejection (55%) and the other half, partial rejection (42%). Approximately half of the English adults' responses to critical 'and' conveyed partial rejection, and the remaining ones were divided between complete rejection (16%) and acceptance (29%).

Table 4.21. Breakdown of Arabic and English adults' ternary responses over the three conditions of experiment 4

Utterance	Type of response	Native adults—Arabic				Native adults—English			
		Some %	Most %	Or %	And %	Some %	Most %	Or %	And %
Felicitous	Large	95	90	100	100	100	77	81.8	100
	Medium	0	10	0	0	0	18	13.6	0
	Small	5	0	0	0	0	5	4.5	0
	Total	100	100	100	100	100	100	100	100
Infelicitous	Large	5	2.5	10	2.5	6.8	20.5	4.5	29.5
	Medium	15	17.5	20	42.5	61.4	54.5	68.2	45.5
	Small	80	80	70	55	31.8	25	27.3	15.9
	Total	100	100	100	100	100	100	100	100
Bizarre	Large	0	0	0	0	0	0	0	0
	Medium	0	0	0	0	0	0	0	0
	Small	100	100	100	100	100	100	100	100
	Total	100	100	100	100	100	100	100	100

Note: Each quantifier is presented alongside 4 critical items (infelicitous) and 4 control items (2 felicitous and 2 bizarre). The use of the term infelicitous 'and' is meant to refer to 'encyclopaedic' scale.

Finally, the two adult groups' performance in the bizarre condition, as table 4.20 shows, was exactly the same, with 100% complete rejection for all four quantifiers.

Performance on the infelicitous (under-informative) items in experiment 4

As noted above, the current study's main focus is on measuring participants' sensitivity to the violation of the Gricean principle of informativeness; the pair-wise comparisons in this section investigate differences in performance among adult groups for the under-informative condition only. The results of the Mann–Whitney U-test revealed that the Arabic and English adults' performance statistically and significantly differed on all of the infelicitous 'most' ($U=376$, $Z=-4.99$, $p<0.001$), the infelicitous 'some' ($U=474$, $Z=-4.130$, $p<0.001$), the infelicitous 'or' ($U=556$, $Z=-3.25$, $p=0.001$) and the infelicitous 'and' ($U=437$, $Z=-4.33$, $p<0.001$). It should be noted that the two groups showed high rate of penalisation, but differed qualitatively in terms of the type of response ('medium' or 'small').

4.4.1.3 Enriched context v. no context

This section applies multivariate analysis using cross-tabulation and chi-squared tests to compare the performance of each group separately over the two experiments. Each sub-subsection contains three cross-tabulation comparisons, one for each of the following pairs of conditions: under-informative versus infelicitous, optimal versus felicitous, and false versus bizarre. My reason for conducting this analysis is to find out how the children's behavioural pragmatic performance (i.e. rate of response: large, medium, small) might differ in the two context conditions (with both critical and filler items). Multivariate analysis with cross-tabulation is usually used with categorical data, and in the present study, allows me to compare one cell (in the enriched context column) to another cell (in the no context column). To explain further, let us say that we have 100 rows, each representing a participant's given response (large, medium, or small) in the parallel items in the two pragmatic experiments (e.g. felicitous v. optimal); then, the test calculates the total amount of responses in each condition, and compares number and percentage for each response type. For example, it would tell us that when 23% of responses in the enriched context were 'small', there were only 10% 'small' in the no context condition. Then, the cross-tabulation would tell us if such a difference was statistically significant.

Children's performance

The section starts with the bilingual children's performance on the two experiments in English, and then in Arabic. Next, the Arabic children's performance in the two experiments is compared, followed by the English children's performance.

Bilingual children in English

Under-informative v. infelicitous condition

A cross-tabulation comparison between the bilingual children's performance (in English) in two parallel conditions—under-informative v. infelicitous (experiment 3 v. experiment 4). The multivariate analyses (presented in Table 4.22) reveal that these children rejected around 25.8% of under-informative utterances, and rejected infelicitous utterances slightly less (20.6%). It can be seen further that the children partially rejected (with the 'medium' response) under-informative utterances

somewhat more than infelicitous ones (26.5% v. 21.7%). As these numbers imply and as cross-tabulation shows, the majority of children gave the 'large' response on these two conditions; it can be noticed, however, that the bilingual children rejected critical items slightly better when there was a context, which is to say, they accepted the under-informative items in experiment 3 less than the infelicitous ones in experiment 4 (47.7% v. 57.7%). The chi-squared test comparing the performance (rate of type) of this group in two critical (under-informative) conditions reveals a significant statistical difference ($X^2(4, N=480)=31.98, p<0.001$).

Table 4.22. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): Bilingual children in English

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	29	24	71	124
		% within experiment 3	23.4%	19.4%	57.3%	100.0%
		% within experiment 4	29.3%	23.1%	25.6%	25.8%
		% of total	6.0%	5.0%	14.8%	25.8%
	Medium	Count	29	47	51	127
		% within experiment 3	22.8%	37.0%	40.2%	100.0%
		% within experiment 4	29.3%	45.2%	18.4%	26.5%
		% of total	6.0%	9.8%	10.6%	26.5%
	Large	Count	41	33	155	229
		% within experiment 3	17.9%	14.4%	67.7%	100.0%
		% within experiment 4	41.4%	31.7%	56.0%	47.7%
		% of total	8.5%	6.9%	32.3%	47.7%
Total	Count	99	104	277	480	
	% within experiment 3	20.6%	21.7%	57.7%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	20.6%	21.7%	57.7%	100.0%	

Optimal v. felicitous condition

Table 4.23 presents the outcomes of a comparison between the bilingual children’s performance on items used as controls, to show that children could accept pragmatically and logically correct utterances. As the table shows, although the majority of responses accepted these utterances (66.7% in experiment 3 and 72.5% in experiment 4), the multivariate analyses showed some variation in responses: bilingual children penalised optimal utterances around 18.3% with ‘small’ and 15% with ‘medium’ (somewhat more than for the felicitous items in experiment 4, which were 10.4% with ‘small’ and 16.7% with medium). The chi-squared outcomes show no statistically significant difference in bilingual children’s performances in English between the experiments ($X^2(4, N=240)=3.80, p=0.433$).

Table 4.23. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): Bilingual children in English

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	5	10	29	44
		% within experiment 3	11.4%	22.7%	65.9%	100.0%
		% within experiment 4	20.0%	25.0%	16.6%	18.3%
		% of total	2.1%	4.2%	12.1%	18.3%
	Medium	Count	2	8	26	36
		% within experiment 3	5.6%	22.2%	72.2%	100.0%
		% within experiment 4	8.0%	20.0%	14.9%	15.0%
		% of total	0.8%	3.3%	10.8%	15.0%
	Large	Count	18	22	120	160
		% within experiment 3	11.2%	13.8%	75.0%	100.0%
		% within experiment 4	72.0%	55.0%	68.6%	66.7%
		% of total	7.5%	9.2%	50.0%	66.7%
Total	Count	25	40	175	240	
	% within experiment 3	10.4%	16.7%	72.9%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	10.4%	16.7%	72.9%	100.0%	

False v. bizarre condition

Table 4.24 below provides a comparison between the bilingual children’s responses in the false and bizarre conditions, both of which were used as fillers to ensure that children did not accept all items with the ‘large’ choice. The cross-tabulation analysis revealed that the children rejected these conditions (by choosing the ‘small’ strawberry), at a high rate in the two conditions (86% in experiment 3 and 96% in experiment 4). The percentage of responses that penalised the false items with ‘medium’ was higher than for the bizarre items but still low (6.7% v. 1.2%), and similarly small rates of acceptance of these conditions were found for the ‘large’ strawberry (7% in experiment 3 v. 2.9% in experiment 4). The chi-squared test, comparing responses these two conditions, reveals no significant statistical difference in bilingual children’s performance in English ($X^2(4, N=240)=4.37, p=0.359$).

Table 4.24. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): Bilingual children in English

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	199	2	6	207
		% within experiment 3	96.1%	1.0%	2.9%	100.0%
		% within experiment 4	86.5%	66.7%	85.7%	86.2%
		% of total	82.9%	0.8%	2.5%	86.2%
	Medium	Count	16	0	0	16
		% within experiment 3	100.0%	0.0%	0.0%	100.0%
		% within experiment 4	7.0%	0.0%	0.0%	6.7%
		% of total	6.7%	0.0%	0.0%	6.7%
	Large	Count	15	1	1	17
		% within experiment 3	88.2%	5.9%	5.9%	100.0%
		% within experiment 4	6.5%	33.3%	14.3%	7.1%
		% of total	6.2%	0.4%	0.4%	7.1%
Total	Count	230	3	7	240	
	% within experiment 3	95.8%	1.2%	2.9%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	95.8%	1.2%	2.9%	100.0%	

Bilingual children in Arabic

Under-informative v. infelicitous condition

The cross-tabulation analysis of the bilingual children's performance in Arabic in the two critical conditions (under-informative v. infelicitous) is presented in table 4.25. The results show that, similar to their performance in English, these children tended in Arabic to reject under-informative items (using the 'small' strawberry) at a slightly higher rate than infelicitous items (23.8% v. 20.6%). The table shows a marginally higher rate of penalising these items using the 'medium' choice than rejecting them using 'small'; however, this rate varies between the two conditions, with the higher rate recorded for the under-informative items in the enriched context condition (30%) as compared to infelicitous ones without context (22.1%). Similar to their performance in English, the bilingual children often accepted both under-informative and infelicitous utterances (with 'large'), though considerably more for the infelicitous ones (46.2% in experiment 3 v. 57.6% in experiment 4). Unlike their performance in English, a comparative chi-squared test showed no statistically

significant difference in bilingual children’s performance in Arabic between these experiments ($X^2(4, N=480)=7.95, p=0.093$).

Table 4.25. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): Bilingual children in Arabic

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	24	22	68	114
		% within experiment 3	21.1%	19.3%	59.6%	100.0%
		% within experiment 4	24.2%	20.8%	24.7%	23.8%
		% of total	5.0%	4.6%	14.2%	23.8%
	Medium	Count	36	39	69	144
		% within experiment 3	25.0%	27.1%	47.9%	100.0%
		% within experiment 4	36.4%	36.8%	25.1%	30.0%
		% of total	7.5%	8.1%	14.4%	30.0%
	Large	Count	39	45	138	222
		% within experiment 3	17.6%	20.3%	62.2%	100.0%
		% within experiment 4	39.4%	42.5%	50.2%	46.2%
		% of total	8.1%	9.4%	28.7%	46.2%
Total	Count	99	106	275	480	
	% within experiment 3	20.6%	22.1%	57.3%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	20.6%	22.1%	57.3%	100.0%	

Optimal v. felicitous

A comparison of bilingual children’s responses in the parallel conditions, optimal v. felicitous, is given in table 4.26. The outcomes revealed that, in Arabic, these children tended to accept felicitous items (75% in experiment 4) considerably more than optimal items (63.7% in experiment 3), and correspondingly, to reject (17.1% ‘small’ choice), or penalise (19.2% ‘medium’ choice) optimal items. In experiment 4, the rate of penalisation was clearly lower: only 11.7% gave responses of complete rejection (‘small’ strawberry), while 13.3% responded with partial penalisation (‘medium’ strawberry). Although these results were similar to those for bilinguals-in-English, the chi-squared outcomes this time revealed a crucial statistical difference between bilinguals’ behaviour in the optimal v. felicitous conditions in Arabic ($X^2(4, N=240)=18.098, p=0.001$).

Table 4.26. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): Bilingual children in Arabic

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	11	5	25	41
		% within experiment 3	26.8%	12.2%	61.0%	100.0%
		% within experiment 4	39.3%	15.6%	13.9%	17.1%
		% of total	4.6%	2.1%	10.4%	17.1%
	Medium	Count	1	11	34	46
		% within experiment 3	2.2%	23.9%	73.9%	100.0%
		% within experiment 4	3.6%	34.4%	18.9%	19.2%
		% of total	0.4%	4.6%	14.2%	19.2%
	Large	Count	16	16	121	153
		% within experiment 3	10.5%	10.5%	79.1%	100.0%
		% within experiment 4	57.1%	50.0%	67.2%	63.7%
		% of total	6.7%	6.7%	50.4%	63.7%
Total	Count	28	32	180	240	
	% within experiment 3	11.7%	13.3%	75.0%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	11.7%	13.3%	75.0%	100.0%	

False v. bizarre

Similar to the bilingual children’s behaviour in the two parallel conditions, false v. bizarre, the multivariate comparison in table 4.27 shows that the great majority of responses used the small strawberry (80.8% for false and 92.9% for bizarre), though the proportion of complete rejection was considerably higher for bizarre. (These are still marginally lower than the rates found in English, however.) The table also shows that higher rate of partial penalisation (with the medium strawberry choice) was given to the false items than to the bizarre items (11.2% and 3.8%, respectively). Also, children accepted false items at a higher rate (7.9%) compared to bizarre items (3.2%). A comparative chi-squared test between these responses showed no significant differences in performance ($X^2(4, N=240)=5.07, p=0.280$).

Table 4.27. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): Bilingual children in Arabic

			Experiment 4			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	182	7	5	194
		% within experiment 3	93.8%	3.6%	2.6%	100.0%
		% within experiment 4	81.6%	77.8%	62.5%	80.8%
		% of total	75.8%	2.9%	2.1%	80.8%
	Medium	Count	24	2	1	27
		% within experiment 3	88.9%	7.4%	3.7%	100.0%
		% within experiment 4	10.8%	22.2%	12.5%	11.2%
		% of total	10.0%	0.8%	0.4%	11.2%
	Large	Count	17	0	2	19
		% within experiment 3	89.5%	0.0%	10.5%	100.0%
		% within experiment 4	7.6%	0.0%	25.0%	7.9%
		% of total	7.1%	0.0%	0.8%	7.9%
Total	Count	223	9	8	240	
	% within experiment 3	92.9%	3.8%	3.3%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	92.9%	3.8%	3.3%	100.0%	

Arabic children

Under-informative v. infelicitous

The results for cross-tabulation of Arabic children's performance in the critical conditions (under-informative v. infelicitous) in experiment 3 (enriched context) versus experiment 4 (no context) are displayed in table 4.28. Notably, 14.4% and 15.8% of responses consisted of complete rejection (small strawberry) in experiments 3 and 4, respectively; tentative penalisation (medium strawberry) was less frequent, with 9.4% in experiment 3 and only 6% in experiment 4. Thus, the majority of responses completely accepted the critical utterances (large strawberry), and as the table shows, the ratios of 'large' responses were very similar in experiments 3 and 4 (76.2% and 78.1%, respectively). The cross-tabulation chi-squared test outcomes reveal a significant difference between Arabic children's performance with enriched versus no context ($X^2(4, N=480)=29.64, p<0.001$).

Table 4.28. Experiment 3*experiment 4 cross-tabulation (under-informative v. Infelicitous): Arabic children (monolingual)

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	18	3	48	69
		% within experiment 3	26.1%	4.3%	69.6%	100.0%
		% within experiment 4	23.7%	10.3%	12.8%	14.4%
		% of total	3.8%	0.6%	10.0%	14.4%
	Medium	Count	7	10	28	45
		% within experiment 3	15.6%	22.2%	62.2%	100.0%
		% within experiment 4	9.2%	34.5%	7.5%	9.4%
		% of total	1.5%	2.1%	5.8%	9.4%
	Large	Count	51	16	299	366
		% within experiment 3	13.9%	4.4%	81.7%	100.0%
		% within experiment 4	67.1%	55.2%	79.7%	76.2%
		% of total	10.6%	3.3%	62.3%	76.2%
Total	Count	76	29	375	480	
	% within experiment 3	15.8%	6.0%	78.1%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	15.8%	6.0%	78.1%	100.0%	

Optimal v. felicitous

A multivariate comparison between the responses given by Arabic children in experiments 3 and 4 for the filler items in the two parallel conditions (optimal v. felicitous) is presented in table 4.29. The comparison showed that the majority of children accepted these utterances. Of the total responses 81.2% were ‘large’ responses in experiment 3, a proportion that increased slightly in experiment 4 (87.9%). This means less complete or partial rejection of these conditions, and indeed, 9.2% of total responses used the small strawberry in experiment 3, and around 8.3% in experiment 4, whereas partial penalisation (medium choice) was approximately equal to rejection in experiment 3 (9.6%), and in experiment 4, only 3.2% of responses penalised the felicitous items. The chi-squared test, however, shows a significant difference between children’s performance in these two parallel conditions ($X^2(4, N=480)=13.13, p=0.011$).

Table 4.29. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): Arabic children (monolingual)

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	5	0	17	22
		% within experiment 3	22.7%	0.0%	77.3%	100.0%
		% within experiment 4	25.0%	0.0%	8.1%	9.2%
		% of total	2.1%	0.0%	7.1%	9.2%
	Medium	Count	2	3	18	23
		% within experiment 3	8.7%	13.0%	78.3%	100.0%
		% within experiment 4	10.0%	33.3%	8.5%	9.6%
		% of total	0.8%	1.2%	7.5%	9.6%
	Large	Count	13	6	176	195
		% within experiment 3	6.7%	3.1%	90.3%	100.0%
		% within experiment 4	65.0%	66.7%	83.4%	81.2%
		% of total	5.4%	2.5%	73.3%	81.2%
Total	Count	20	9	211	240	
	% within experiment 3	8.3%	3.8%	87.9%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	8.3%	3.8%	87.9%	100.0%	

False v. bizarre

Table 4.30 displays the results of a cross-tabulation between Arabic children's performance in the false and bizarre conditions. It can be seen that although children mostly rejected the items for these conditions, a considerable higher rate of rejection was found for bizarre items (80.8% in experiment 3 [false] versus 90.8% in experiment 4 [bizarre]). Only 1.7% of the total responses partially penalised the bizarre items in experiment 4, a proportion that was higher in experiment 3 (9.6%). The table also reveals that out of the total responses given to these items, 9.6% and 7.6% in experiments 3 and 4, respectively, completely accepted them (with the 'large' choice). The cross-tabulation chi-squared test reveals a crucial statistical difference in the Arabic children's performance in these two parallel conditions ($X^2(4, N=240)=38.65, p<0.001$).

Table 4.30. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): Arabic children (monolingual)

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	180	0	14	194
		% within experiment 3	92.8%	0.0%	7.2%	100.0%
		% within experiment 4	82.6%	0.0%	77.8%	80.8%
		% of total	75.0%	0.0%	5.8%	80.8%
	Medium	Count	17	4	2	23
		% within experiment 3	73.9%	17.4%	8.7%	100.0%
		% within experiment 4	7.8%	100.0%	11.1%	9.6%
		% of total	7.1%	1.7%	0.8%	9.6%
	Large	Count	21	0	2	23
		% within experiment 3	91.3%	0.0%	8.7%	100.0%
		% within experiment 4	9.6%	0.0%	11.1%	9.6%
		% of total	8.8%	0.0%	0.8%	9.6%
Total	Count	218	4	18	240	
	% within experiment 3	90.8%	1.7%	7.5%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	90.8%	1.7%	7.5%	100.0%	

English children

Under-informative v. infelicitous

The outcomes of a cross-tabulation between English children's total responses given to under-informative (experiment 3) versus infelicitous (experiment 4) items are given in table 4.31. The English children rejected these items at very similar rates (13.2%, 16.8% respectively). The table also shows very similar percentages for partial penalisation (medium) and acceptance (large). Out of the total responses, the English children penalised under-informative versus infelicitous utterances around 22% (with the medium choice) and accepted around 61% of the under-informative and infelicitous items in experiments 3 and 4. The chi-squared test did not reveal any significant differences in English children's performance in these two conditions ($X^2(4, N=416)=93.6, p<0.001$).

Table 4.31. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): English children (monolingual)

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	19	5	31	55
		% within experiment 3	34.5%	9.1%	56.4%	100.0%
		% within experiment 4	27.1%	5.6%	12.1%	13.2%
		% of total	4.6%	1.2%	7.5%	13.2%
	Medium	Count	13	54	34	101
		% within experiment 3	12.9%	53.5%	33.7%	100.0%
		% within experiment 4	18.6%	60.0%	13.3%	24.3%
		% of total	3.1%	13.0%	8.2%	24.3%
	Large	Count	38	31	191	260
		% within experiment 3	14.6%	11.9%	73.5%	100.0%
		% within experiment 4	54.3%	34.4%	74.6%	62.5%
		% of total	9.1%	7.5%	45.9%	62.5%
Total	Count	70	90	256	416	
	% within experiment 3	16.8%	21.6%	61.5%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	16.8%	21.6%	61.5%	100.0%	

Optimal v. felicitous

Table 4.32 provides a comparison between English children’s responses in the optimal and felicitous conditions in experiments 3 and 4, respectively. Of the total responses given to the two conditions, 73% of responses to the optimal items in experiment 3 were ‘large’ (acceptance), with a slightly higher rate for the felicitous items in experiment 4 (83%). Correspondingly, marginally higher proportions of complete and partial penalisation were found in experiment 3 (9% ‘small’ and 18% ‘medium’). The percentages of penalisation were lower in experiment 4, with 6.7% ‘small’ responses and 9.6% ‘medium’ responses. The cross-tabulation chi-squared test didn’t reveal any statistically significant difference in children’s performance ($X^2(4, N=208)=6.1, p=0.193$).

Table 4.32. Experiment 3*experiment 4 cross-tabulation (optimal v. felicitous): English children (monolingual)

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	0	2	17	19
		% within experiment 3	0.0%	10.5%	89.5%	100.0%
		% within experiment 4	0.0%	10.0%	9.8%	9.1%
		% of total	0.0%	1.0%	8.2%	9.1%
	Medium	Count	0	5	32	37
		% within experiment 3	0.0%	13.5%	86.5%	100.0%
		% within experiment 4	0.0%	25.0%	18.4%	17.8%
		% of total	0.0%	2.4%	15.4%	17.8%
	Large	Count	14	13	125	152
		% within experiment 3	9.2%	8.6%	82.2%	100.0%
		% within experiment 4	100.0%	65.0%	71.8%	73.1%
		% of total	6.7%	6.3%	60.1%	73.1%
Total	Count	14	20	174	208	
	% within experiment 3	6.7%	9.6%	83.7%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	6.7%	9.6%	83.7%	100.0%	

False v. bizarre

The results of a comparison between the English children’s performance in the false and the bizarre conditions are given in table 4.33. In both experiments, the highest rate was recorded for complete rejection (‘small’ response), with around 72% for the false condition and 94% for the bizarre condition. The percentage given to partial penalisation (‘medium’ response) out of the total responses was around 16% in experiment 3, and only 3% in experiment 4. The ratio of responses that did not penalise these two conditions at all was low in both cases but higher for false items (12%) than for bizarre ones (2%). The chi-squared results revealed a significant difference in performance between these conditions ($X^2(4, N=208)=30.9, p<0.001$).

Table 4.33. Experiment 3*experiment 4 cross-tabulation (false v. bizarre): English children (monolingual)

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	148	0	2	150
		% within experiment 3	98.7%	0.0%	1.3%	100.0%
		% within experiment 4	75.5%	0.0%	40.0%	72.1%
		% of total	71.2%	0.0%	1.0%	72.1%
	Medium	Count	25	6	2	33
		% within experiment 3	75.8%	18.2%	6.1%	100.0%
		% within experiment 4	12.8%	85.7%	40.0%	15.9%
		% of total	12.0%	2.9%	1.0%	15.9%
	Large	Count	23	1	1	25
		% within experiment 3	92.0%	4.0%	4.0%	100.0%
		% within experiment 4	11.7%	14.3%	20.0%	12.0%
		% of total	11.1%	0.5%	0.5%	12.0%
Total	Count	196	7	5	208	
	% within experiment 3	94.2%	3.4%	2.4%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	94.2%	3.4%	2.4%	100.0%	

Adults' performance

This section provides the results of multivariate analyses (cross-tabulation) comparing the performance of each of the two adult groups in the under-informative/infelicitous condition in experiment 3 and 4 (under-informative v. infelicitous) and in the parallel control conditions, optimal versus felicitous and false versus bizarre. The analyses start with the performance of the Arabic adults, followed by the English adults.

Arabic adults

Under-informative v. infelicitous

The cross-tabulation given in table 4.34 between the Arabic adults' performance on the under-informative items in experiment 3 and the infelicitous ones in experiment 4 revealed very similar performance. Of the total responses, around 71% consisted of complete rejection ('small' response) in both experiments; there were also close rates of partial rejection ('medium' response) between them (28% in experiment 3, 23%

experiment 4). The results, however, revealed that none of the Arabic adults accepted the under-informative items ('large' response) in experiment 3, while only 5% accepted the infelicitous items in experiment 4. Despite this very similar performance, the chi-squared test results revealed a significant difference, possibly due to the absence of 'large' responses in experiment 3 ($X^2(2, N=160)=71.23, p<0.001$).

Table 4.34. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): Arabic adults

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	102	7	6	115
		% within experiment 3	88.7%	6.1%	5.2%	100.0%
		% within experiment 4	89.5%	18.4%	75.0%	71.9%
		% of total	63.7%	4.4%	3.8%	71.9%
	Medium	Count	12	31	2	45
		% within experiment 3	26.7%	68.9%	4.4%	100.0%
		% within experiment 4	10.5%	81.6%	25.0%	28.1%
		% of total	7.5%	19.4%	1.3%	28.1%
	Total	Count	114	38	8	160
		% within experiment 3	71.3%	23.8%	5.0%	100.0%
% within experiment 4		100.0%	100.0%	100.0%	100.0%	
% of total		71.3%	23.8%	5.0%	100.0%	

Optimal v. felicitous

Table 4.35 displays the results of cross-tabulation between the Arabic adults' performance in the two parallel conditions optimal (experiment 3) and felicitous (experiment 4). It can be seen that performance is close between them, with the highest percentage of the total responses given to the 'large' response. The chi-squared test revealed no significant difference between the Arabic adults' performance in the two conditions ($X^2(4, N=80)=.445, p=.98$).

Table 4.35. Experiment 3*experiment 4 cross-tabulation: Arabic adults (optimal v. felicitous)

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	0	0	2	2
		% within experiment 3	0.0%	0.0%	100.0%	100.0%
		% within experiment 4	0.0%	0.0%	2.6%	2.5%
		% of total	0.0%	0.0%	2.5%	2.5%
	Medium	Count	0	0	8	8
		% within experiment 3	0.0%	0.0%	100.0%	100.0%
		% within experiment 4	0.0%	0.0%	10.4%	10.0%
		% of total	0.0%	0.0%	10.0%	10.0%
	Large	Count	1	2	67	70
		% within experiment 3	1.4%	2.9%	95.7%	100.0%
		% within experiment 4	100.0%	100.0%	87.0%	87.5%
		% of total	1.3%	2.5%	83.8%	87.5%
Total	Count	1	2	77	80	
	% within experiment 3	1.3%	2.5%	96.3%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	1.3%	2.5%	96.3%	100.0%	

False v. bizarre

The results of a comparison between the Arabic adults' performance in the two control conditions false (experiment 3) and bizarre (experiment 4) are presented in table 4.36. Notably, the Arabic adults penalised the items of these two condition with 'small' responses at a very high rate (around 95%) in the two experiments. The chi-squared test results showed no crucial difference between the Arabic adults' performance in the two conditions ($X^2(2, N=80)=.053, p=.97$).

Table 4.36. Experiment 3*experiment 4 cross-tabulation: Arabic adults (false v. bizarre)

		Experiment 4 response			Total
		Small	Medium		
Experiment 3 response	Small	Count	75	1	76
		% within experiment 3	98.7%	1.3%	100.0%
		% within experiment 4	94.9%	100.0%	95.0%
		% of total	93.8%	1.3%	95.0%
	Medium	Count	3	0	3
		% within experiment 3	100.0%	0.0%	100.0%
		% within experiment 4	3.8%	0.0%	3.8%
		% of total	3.8%	0.0%	3.8%
	Large	Count	1	0	1
		% within experiment 3	100.0%	0.0%	100.0%
		% within experiment 4	1.3%	0.0%	1.3%
		% of total	1.3%	0.0%	1.3%
Total	Count	79	1	80	
	% within experiment 3	98.8%	1.3%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	
	% of total	98.8%	1.3%	100.0%	

English adults

Under-informative v. infelicitous

The outcomes of a comparison between the English adults' performance on the under-informative and infelicitous items in experiment 3 and 4 are displayed in table 4.37. It is noticeable that, out of the total responses to the critical items in the two experiments, around 76% and 60% showed partial rejection ('medium' response) in experiments 3 and 4, respectively. The table also shows higher rates of complete rejection ('small' response) in experiment 4 (25%) than in experiment 3 (15%). The percentage of critical items accepted ('large' response) was also slightly higher in experiment 4 (15%) than in experiment 3 (9%). The chi-squared test results showed a crucial difference in the English adults' performance in experiment 3 (with context) and experiment 4 (no context) ($X^2(4, N=176)=11.26, p=0.024$).

Table 4.37. Experiment 3*experiment 4 cross-tabulation (under-informative v. infelicitous): English adults

		Experiment 4 response			Total	
		Small	Medium	Large		
Experiment 3 response	Small	Count	13	9	5	27
		% within experiment 3	48.1%	33.3%	18.5%	100.0%
		% within experiment 4	29.5%	8.6%	18.5%	15.3%
		% of total	7.4%	5.1%	2.8%	15.3%
	Medium	Count	27	87	19	133
		% within experiment 3	20.3%	65.4%	14.3%	100.0%
		% within experiment 4	61.4%	82.9%	70.4%	75.6%
		% of total	15.3%	49.4%	10.8%	75.6%
	Large	Count	4	9	3	16
		% within experiment 3	25.0%	56.3%	18.8%	100.0%
		% within experiment 4	9.1%	8.6%	11.1%	9.1%
		% of total	2.3%	5.1%	1.7%	9.1%
Total	Count	44	105	27	176	
	% within experiment 3	25.0%	59.7%	15.3%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	25.0%	59.7%	15.3%	100.0%	

Optimal v. felicitous

Table 4.38 shows the results of cross-tabulation between the English adults' performance in the two conditions: optimal (experiment 3) versus felicitous (experiment 4). It can be seen that they accepted the items of these conditions (with 'large' response) at very similar rates (83% in experiment 3 and 89.8% in experiment 4). Despite this similar performance on the two conditions, the chi-squared test result was significant ($X^2(4, N=88)=14.850$ p=0.005). This might be due to the small sample size of the adult groups, which would have made the test very sensitive to such a slight difference.

Table 4.38. Experiment 3*experiment 4 cross-tabulation: Arabic adults (optimal v. felicitous)

			Experiment 4 response			Total
			Small	Medium	Large	
Experiment 3 response	Small	Count	0	1	0	1
		% within experiment 3	0.0%	100.0%	0.0%	100.0%
		% within experiment 4	0.0%	14.3%	0.0%	1.1%
		% of total	0.0%	1.1%	0.0%	1.1%
	Medium	Count	1	2	11	14
		% within experiment 3	7.1%	14.3%	78.6%	100.0%
		% within experiment 4	50.0%	28.6%	13.9%	15.9%
		% of total	1.1%	2.3%	12.5%	15.9%
	Large	Count	1	4	68	73
		% within experiment 3	1.4%	5.5%	93.2%	100.0%
		% within experiment 4	50.0%	57.1%	86.1%	83.0%
		% of total	1.1%	4.5%	77.3%	83.0%
Total	Count	2	7	79	88	
	% within experiment 3	2.3%	8.0%	89.8%	100.0%	
	% within experiment 4	100.0%	100.0%	100.0%	100.0%	
	% of total	2.3%	8.0%	89.8%	100.0%	

False v. bizarre

The results of a multivariate comparison between English adults' performance in the two parallel conditions false (experiment 3) and bizarre (experiment 4) are displayed in table 4.39. English adults rejected all the bizarre items with 'small' response (100%) and penalised around 89% of the false items with 'small' response and 11% with 'medium' response. The chi-squared test for these two conditions could not be computed due to the constant variable in experiment 4 (100% 'small' response).

Table 4.39. Experiment 3*experiment 4 cross-tabulation: English adults (false v. bizarre)

		Experiment 4 response		Total
		Small		
Experiment 3 response	Small	Count	78	78
		% within experiment 3	100.0%	100.0%
		% within experiment 4	88.6%	88.6%
		% of total	88.6%	88.6%
	Medium	Count	10	10
		% within experiment 3	100.0%	100.0%
		% within experiment 4	11.4%	11.4%
		% of total	11.4%	11.4%
Total		Count	88	88
		% within experiment 3	100.0%	100.0%
		% within experiment 4	100.0%	100.0%
		% of total	100.0%	100.0%

4.4.1.4 A summary of pragmatic performance (based on ternary responses)

The analyses in this section explored children and adults' pragmatic performance (based on ternary responses) in experiments 3 and 4. In the enriched context (experiment 3), the results showed crucial differences in performance between the bilingual-in-Arabic and Arabic children on 'most', 'some', and 'or', with the bilinguals penalising the under-informative items (with 'small' and 'medium' responses) at a higher rate. The outcomes also revealed significant differences between the bilingual-in-English and English children on 'some', 'or', and 'and', with higher rates of penalisation (with 'small' and 'medium' responses) among the bilingual children. The comparison between the English and Arabic children showed a crucial difference between the two groups only on 'most' and 'some', with the English children performing pragmatically better by penalising the under-informative items. Finally, the two adult groups' performance significantly differed on all of the four quantifiers, though this was a result of different preferences in how to penalise the under-informative items (partially or completely) rather than different rates of penalising them overall (as opposed to accepting them).

In experiment 4, the analyses revealed significant differences between the bilinguals-in-Arabic and the Arabic children in each of 'most' and 'some', with better pragmatic performance among the bilingual children. The outcomes also showed a significant difference between the bilinguals-in-English and the English children in performance on 'most' and marginal differences on 'some' and 'and', with the bilinguals penalising the under-informative (with 'small' or 'medium' responses) at a higher rate. The comparison between the English and Arabic children showed significantly better pragmatic performance among the English children, with a high rate of penalising under-informative 'some', 'or', and 'and'. The analyses of the adult groups revealed a significant difference between the Arabic and English groups, but as in experiment 4, this might be attributed to type of penalising rather than overall rate of penalising: the Arabic adults favoured penalising the violation of informativeness with 'small' responses, while the English tended to penalise such items with 'medium' responses.

The analyses also investigated whether each group's performance differed by context: enriched context (experiment 3) versus no context (experiment 4); the results of these comparisons are displayed in table 4.40. It can be seen that in the under-informative versus infelicitous conditions, all groups (children and adults) performed significantly better pragmatically with the enriched context, except the bilingual-in-Arabic children, whose performance did not significantly differ. In the optimal versus felicitous comparison, the outcomes revealed a significant difference in the performance of only two groups: the bilingual-in-Arabic and Arabic children, who accepted optimal items at higher rates in experiment 4 than in experiment 3. For the false versus bizarre comparison, only the Arabic and English child groups differed significantly, with higher rates of rejection ('small' response) in experiment 4 than in experiment 3.

Table 4.40. A summary of the groups' performance in the parallel conditions of experiments 3 and 4 (context v. no context)

Group	Condition (experiment 3 v. experiment 4)		
	Under-informative v. infelicitous	Optimal v. felicitous	False v. bizarre
Bilingual-in-English children	S (more 'medium' and 'small' responses in exp. 4)	NS	NS
Bilingual-in-Arabic children	NS	S (more 'large' responses in exp. 4)	NS
Arabic children	S (higher penalisation; more 'medium'/'small' responses in exp. 4)	S (more 'large' responses in exp. 4)	S (more 'small' responses in exp. 4)
English children	S (higher penalisation; more 'medium'/'small' responses in exp. 4)	NS	S (more 'small' responses in exp. 4)
Arabic adults	S (slightly higher penalisation; more 'medium'/'small' responses in exp. 3)	NS	NS
English adults	S (higher penalisation; more 'medium' and 'small' responses in exp. 4)	S (though very similar performance)	The test could not be applied

Note: Abbreviations in the table: exp. (experiment), S (significant), NS (non-significant)

With regard to the adults' performance on the parallel conditions, the table shows that, for the Arabic adults, despite very similar ratios of penalisation in the under-informative versus infelicitous conditions, the chi-squared test revealed a significant difference, while there was no crucial difference in the other conditions. The results of the English adults also showed a significant difference in the under-informative versus infelicitous and also in the optimal versus felicitous conditions, again despite the very similar performance in the two experiments. The statistical test could not be completed for the false versus bizarre conditions, due to the stable performance for (complete rejection of) all the bizarre items.

4.4.2 Pragmatic performance: Another way of analysing data

After exploring, in depth, the participants' pragmatic sensitivity to the violation of Grice's principle of informativeness in section 4.4.1 as represented by their ternary judgments (large, medium, small), a different analytic procedure will be applied to the results of experiments 3 and 4. In this procedure, the participants' responses are re-scored using binary criteria; a correct response is scored as 1, and a wrong response as 0. That is, for the critical (under-informative/infelicitous) items, any response that indicates partial penalisation ('medium' response) or complete penalisation ('small'

response) is scored 1, while the ‘large’ response, which does not convey any sensitivity to the violation of informativeness, is scored as 0. The analyses in this section focus only on the under-informative/infelicitous items in experiment 3 and 4; control items (fillers) are excluded.

Before going on to examine performance on each task under binary scoring, I should justify the adoption of the new scoring system and explain why the analyses in this section focuses only on the critical items. First, a binary score should provide the study with a better understanding of children’s pragmatic behaviour in the two tasks, because it can clearly reflect the proportion of children who penalised the critical items, regardless of whether partially or completely. In addition, a binary score allows calculation of the total score for each child, which should facilitate testing of differences between groups using parametric tests. Second, the reason control items are dropped is that this study’s main aim is to investigate children’s ability to detect violation of Grice’s maxim of informativeness. Other conditions in the two experiments served as controls to find out if the participants have the ability to accept logically and pragmatically appropriate items and to reject the items which are logically and pragmatically inappropriate. Since the controls, which were intended to help determine whether participants can accept logically and pragmatically appropriate and reject inappropriate items, served no core aim, and since section 4.4.1 explored them intensively, I exclude them from the analyses here.

The structure of this section is as follows: first, it recapitulates the results of experiment 3, and then those of experiment 4. For each experiment, descriptive and then inferential analyses are given and performance on each quantifier is explored with the new scoring. After this, a comparison is made between children’s performance on the Horn lexical scale (e.g. <all, some> and an ad hoc scale (a context-specific scale, e.g. <{A}, {B}, {A and B}>) (Papafragou & Tantalou, 2004; Katsos & Cummins, 2012). This is followed by a summary of the findings using the new scoring system.

4.4.2.1 Experiment 3 (enriched context)

Before proceeding with the binary analysis of experiment 3 data, I remind the reader that the main task here is to measure participants' ability to detect the violation of informativeness by judging a fictional character's utterance describing the action in the scenario (an enriched context pragmatic task). Then, we apply the new criteria on the results: 'medium' and 'small' responses are scored as 1, 'large' as 0.

Children's performance

The scores of each child in the four trials (for each of the quantifiers/operators 'most', 'some', 'or', 'and') were calculated and then converted to percentages. Figure 4.16 shows the average percentages of responses that were correct for the bilingual children's performance in English and Arabic and the performance of the Arabic and English children. It can be seen that the bilingual children scored higher in both languages than the Arabic and English children, on all the quantifiers/operators, and further that the bilinguals performed almost equally in Arabic and English on all the quantifiers, except 'or', on which they performed slightly better in Arabic. In both languages, the bilingual children performed better on 'some' (around 55%) than 'most' (around 42%), and better on 'and' (74%) than 'or' (35% in English, 48% in Arabic). The Arabic children's average scores on the quantifier/operators were the lowest among the groups, and were basically constant across all the quantifiers and the disjunction 'or' (around 14%), except the ad hoc 'and', where their scores dramatically increased (59%). The English children's performance on 'most' and 'some' was roughly the same (around 38%), and reached its peak on 'and' (59%), but clearly declined on 'or' (17%).

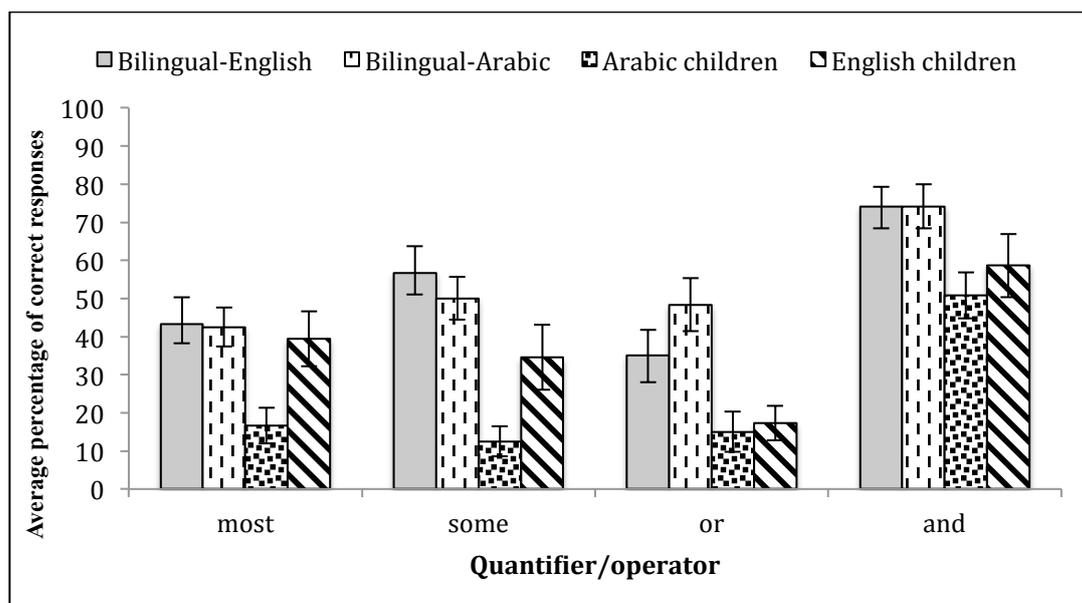


Figure 4.16. Child groups' average percentages of correct penalisation of the under-informative items in experiment 3

Between-group comparisons by quantifier/operator

Before comparing group performance for each quantifier in the under-informative condition of experiment 3, the results for all the quantifiers/operators for each group were checked against the assumption of normality. This is an essential step, as it helps determine the type of tests to be used in the analyses (parametric or non-parametric). Shapiro–Wilk tests of normality revealed that neither the bilingual children's performance in English nor that in Arabic, and neither the results of the Arabic children nor those of the English children, were normally distributed for any of all the quantifiers/operators ($p < 0.05$). Therefore, comparison between groups was basically done using distribution-free tests, following Larson-Hall (2010); although I also applied parametric tests, I only report whether their outcomes are consistent with those of the distribution-free tests or not.

Pragmatic performance on 'most' in experiment 3

The pair-wise comparison between groups on 'most' revealed no significant difference between the bilingual children's performance in English and in Arabic (Wilcoxon signed-rank test; $Z = -.137$, $p = 0.89$), nor between the bilingual-in-English and the English children (Mann–Whitney U-test; $U = 371$, $Z = -.312$, $p = 0.76$). The comparisons did reveal significant differences between the bilingual-in-Arabic and

the Arabic children ($U=226$, $Z=-3.46$, $p=0.001$), the bilinguals performed better, and between the English and Arabic children ($U=244$, $Z=-2.56$, $p=0.010$), the English children performed better. These findings are compatible with the parametric comparative ANOVA, except for the difference between the Arabic and English children, which was only marginal ($p=0.054$).

Pragmatic performance on ‘some’ in experiment 3

The pair-wise comparisons between groups showed no crucial difference between the bilingual children’s performance in Arabic and English (Wilcoxon signed-rank test; $Z=-1.016$, $p=0.31$). The comparative Mann–Whitney U-tests revealed a significant difference between the bilingual-in-Arabic and Arabic children ($U=172$, $Z=-4.62$, $p<0.001$) and between the bilingual-in-English and English children ($U=319$, $Z=-1.22$, $p=0.044$), in both cases the bilinguals performed better. There was also a marginal significant difference between the Arabic and English children ($U=289$, $Z=-1.89$, $p=0.059$), the English children performed better. The parametric ANOVA revealed a significant difference only between the bilingual-in-Arabic and the Arabic children ($p<0.001$), not between any of the other groups.

Pragmatic performance on ‘or’ in experiment 3

The investigation of differences in performance on under-informative ‘or’ between the bilingual children’s performance in Arabic and English revealed a significant difference (Wilcoxon signed-rank test; $Z=-.137$, $p=0.033$). Mann–Whitney U-tests showed significant differences between the bilingual-in-Arabic and Arabic children ($U=215$, $Z=-3.70$, $p<0.001$), the bilinguals performed better, a marginal significant between the bilingual-in-English and the English children ($U=292$, $Z=-1.71$, $p=0.086$) and there was no significant difference between the Arabic and English children ($U=334$, $Z=-1.08$, $p=0.28$). According to the parametric ANOVA outcomes, there was a significant difference only between the bilingual-in-Arabic and the Arabic children ($p=0.002$), and between the bilingual-in-English and English children ($p=0.039$) between any other groups (when comparing the bilinguals in both languages and the English v. Arabic children).

Pragmatic performance on ‘and’ in experiment 3

The comparison of bilingual children’s performance in Arabic and English showed no significant difference (Wilcoxon signed-rank test; $Z=0.00$, $p=1$). The comparative Mann–Whitney U-tests revealed a significant difference between the bilingual-in-Arabic and Arabic children ($U=271$, $Z=-.883$, $p=0.006$) but no significant difference between the bilingual-in-English and English children ($U=319$, $Z=-1.22$, $p=0.22$) or between the Arabic and English children ($U=337$, $Z=-.883$, $p=0.38$). These results are consistent with the parametric ANOVA, which showed a significant difference only between the bilingual-in-Arabic and the Arabic children ($p=0.036$)

Adults’ performance

Figure 4.17 breaks down the Arabic and English adult groups’ results. It can be seen that the Arabic adults exhibit ceiling effects on all the quantifiers/operators (indicating 100% correct responses), meaning that they penalised all the under-informative items with either ‘medium’ or ‘small’ responses. Meanwhile, the English adults’ performance on the quantifiers/operators varied slightly: they scored 100% on ‘most’ and more than 90% on ‘some’ and ‘or’, indicating that they penalised under-informative items for these quantifiers and the disjunction ‘or’ at a very high rate. Their performance on ad hoc ‘and’ showed a lower rate of penalisation, with mean score around 77%; however, the relatively big value of the standard error here reflects variation in the English adults’ pragmatic sensitivity to the violation of informativeness with an ad hoc scale.

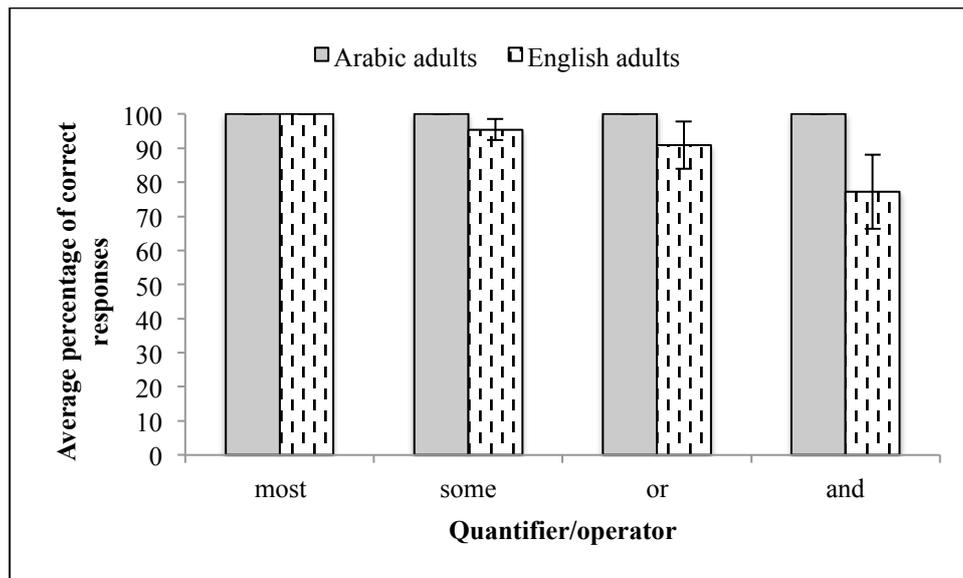


Figure 4.17. Adult groups' average percentages of correct penalisation of the under-informative items in experiment 3

The comparison between the two groups for the quantifier 'some' and the operators 'or', and 'and' using Mann–Whitney U-tests revealed no significant differences between groups in performance on 'some' ($U=45$, $Z=-1.38$, $p=0.17$) or 'or' ($U=45$, $Z=-1.38$, $p=0.17$) but a significant difference on 'and' ($U=35$, $Z=-2.054$, $p=0.040$). These results were confirmed with the parametric t-test; however, the difference between the two groups on 'and' was only marginal ($t(19)=1.98$, $p=0.064$).

4.4.2.2 Experiment 4 (no context)

Before exploring the participants' performance on this task, I remind the reader that the aim here was to measure participants' ability to detect violation of Grice's Maxim of informativeness when there is no context. That is, in this task the participant heard a number of utterances (e.g. *some elephants have trunks*) and were asked to judge them using a 3-point scale. The analyses here re-scored the 3-point scale using binary criteria for appropriate and inappropriate responses, as discussed in section 4.4.

Children's performance

The child groups' average scores on the infelicitous (meaning pragmatically under-informative) items in experiment 4 are presented in figure 4.18. It can be seen that the bilingual children performed roughly equally on 'most' when tested in English and

Arabic (around 47%); their performance on ‘some’ and ‘and’ was approximately the same, with around 47% of responses correct in English and 43% in Arabic. The Arabic children had the lowest average score among the groups, and performed almost equally on ‘most’, ‘some’, and ‘or’ (around 19% of correct responses); their performance clearly improved on that for ad hoc ‘and’ (30%). The English children performed almost equally on ‘most’ and ‘or’ (with around 30% of responses correct), and their performance slightly improved on ‘some’ (36%). It can be also seen that the English children had the highest average score on ad hoc ‘and’, with around 56% of responses correct.

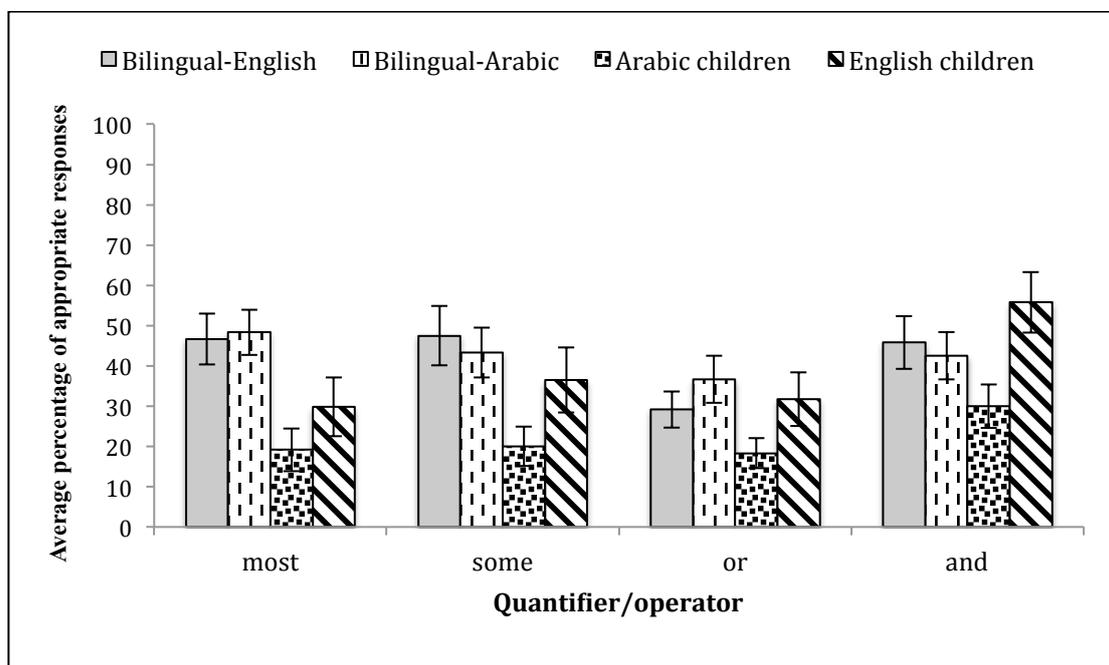


Figure 4.18. Child groups’ average percentages of correct penalisation of the under-informative items in experiment 4

Between-group comparisons (inferential analyses, experiment 4)

As with the Experiment 3 results, I first tested the assumption of normal distribution. The results of Shapiro–Wilk tests revealed that the performance of all the groups on all the quantifiers/operators was not normally distributed; therefore, distribution-free statistics were applied to explore the differences between the groups, and the results of the parametric ANOVA are only reported when they differ from the non-parametric test findings, and then only briefly.

Pragmatic performance on ‘most’ in experiment 4

The comparison between the bilingual children’s performance on ‘most’ in experiment 4 revealed no significant difference (Wilcoxon signed-rank test; $Z=-.267$, $p=0.79$). The Mann–Whitney U-tests showed a significant difference between the bilingual-in-Arabic and the Arabic children ($U=208$, $Z=-3.73$, $p<0.001$), there was a marginal significant difference between the bilingual-in-English and the English children ($U=289$, $Z=-1.71$, $p=0.087$), the bilinguals performed better than the two groups, but there was no significant difference between the Arabic and English children ($U=336$, $Z=-.978$, $p=0.33$). These results are compatible with the ANOVA outcomes, which revealed a significant difference only between the bilingual-in-Arabic and Arabic children ($p=0.002$).

Pragmatic performance on ‘some’ in experiment 4

The comparisons between the groups indicated no significant difference between the bilingual children’s performance in Arabic and English (Wilcoxon signed-rank test; $Z=-.729$, $p=0.47$). The Mann–Whitney U-tests revealed a significant difference between the bilingual-in-Arabic and Arabic children ($U=263$, $Z=-2.88$, $p=0.004$), the bilinguals performed better, but no significant difference between the bilingual-in-English and the English children ($U=323$, $Z=-1.14$, $p=0.25$) or between the Arabic and English children ($U=311$, $Z=-1.39$, $p=0.16$). These findings are consistent with the ANOVA results, which showed a significant difference only between the bilingual-in-Arabic and Arabic children ($p=0.023$).

Pragmatic performance on ‘or’ in experiment 4

The results of comparisons between the groups showed no significant difference between the bilingual children’s performance in Arabic and English (Wilcoxon signed-rank test; $Z=-1.34$, $p=0.18$). The Mann–Whitney U-tests revealed a significant difference between the bilingual-in-Arabic and Arabic children ($U=301$, $Z=-2.31$, $p=0.021$), but there was no significant difference between the bilingual-in-English and the English children ($U=383$, $Z=-.078$, $p=0.94$), nor between the Arabic and English children ($U=316$, $Z=-1.28$, $p=0.19$).

The parametric ANOVA revealed only a marginally significant difference between the bilingual-in-Arabic and the Arabic children ($p=0.053$).

Pragmatic performance on ‘and’ in experiment 4

No significant difference was found between the bilingual children’s performance on ‘and’ in Arabic and English (Wilcoxon signed-rank test; $Z=-.667$, $p=0.5$). The Mann–Whitney U-tests showed no significant differences between the bilingual-in-Arabic and Arabic children ($U=351$, $Z=-1.51$, $p=0.13$) or between the bilingual-in-English and English children ($U=328$, $Z=-1.04$, $p=0.29$), but there was a significant difference between the Arabic and English children ($U=235$, $Z=-2.60$, $p=0.009$), the English children performed better. These findings are compatible with the ANOVA results, which showed a significant difference only between the Arabic and English children ($p=0.037$).

Adults’ performance

The performance of the two adult groups under the binary scores is given in figure 4.19. It can be seen that the Arabic adults gave more correct responses than the English adults on ‘most’ (97% v. 80%) and on ‘and’ (97% v. 70%), and that the two groups scored almost equally high on ‘some’ and ‘or’ (around 94%).

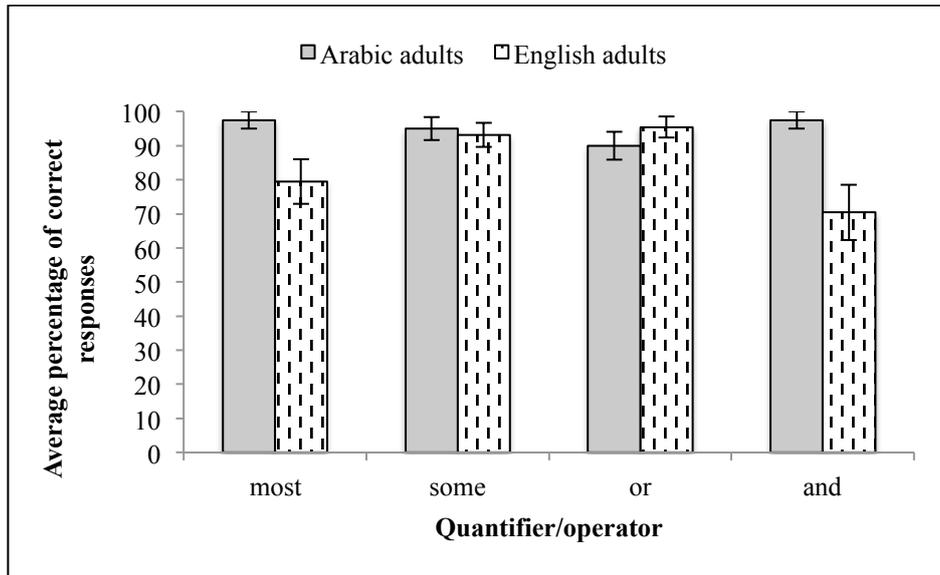


Figure 4.19. Adult groups’ average percentages of correct penalisation for the under-informative items in experiment 4

The comparison using Mann–Whitney U-tests revealed significant between-adult-group differences in performance on ‘most’ (U=25, Z=-2.48, p=0.013), and ‘and’ ((U=19, Z=-2.85, p=0.004) but no significant difference between the groups on ‘some’ (U=51, Z=-.381, p=0.7) or ‘or’ (U=43, Z=-.108, p=0.28). The parametric t-test outcomes were completely consistent with these results.

4.4.2.3 Children’s performance on different scales: Horn, ad hoc, and encyclopaedic

Since the main focus of the current study is children’s pragmatic development, the analyses in this section explore only the child participants’ performance, to find out if their pragmatic performance might significantly differ if measured with a Horn as opposed to an ad hoc or an encyclopaedic scale. Horn lexical (quantifier) scales are context free (e.g. <most, all>, <some, all>, <or, and>), whereas ad hoc scales are context dependent (e.g. <{orange}, {apple}, {orange and apple}>) and encyclopaedic scales are licensed by world knowledge (e.g. to clean your teeth, you need <{toothpaste}>, {toothbrush}, {toothpaste and toothbrush}>).

To examine the difference in children’s performance using the three scales, I first calculated each child’s correct penalisation of under-informative items in the three scalar expressions ‘most’, ‘some’, and ‘or’, dividing the result by the total number of trials for the three expressions (12); then, I computed the percentage for each child’s score in the lexical scale—for example, if a child scored 4 for each of the three scalar expressions, her/his score in the Horn lexical scale would be (‘most’=4+‘some’=4+‘or’=4)/12*100=100%). The children’s scores on the ad hoc and encyclopaedic scales were also changed to percentages in the same way, except that each child’s score was divided by 4 instead of 12.

After computing the children’s results on each scale, the assumption of normality was examined. The Shapiro–Wilk test results revealed that the assumption of normality was violated for all groups (ps<0.05); therefore, all comparisons were conducted using distribution-free tests. Each group’s performance on the two scales in experiment 3 and 4 would be investigated separately.

Comparison by scale

Performance on Horn v. ad hoc scale in experiment 3

Figure 4.20 displays the child groups' average performance in experiment 3 as measured on the Horn and ad hoc scales. In all four child groups (bilingual-in-English, bilingual-in-Arabic, Arabic, English), all the children scored higher on the ad hoc scale than on the Horn scale. For the bilingual children, the difference in performance on the two scales was approximately the same when they were tested in English and in Arabic.

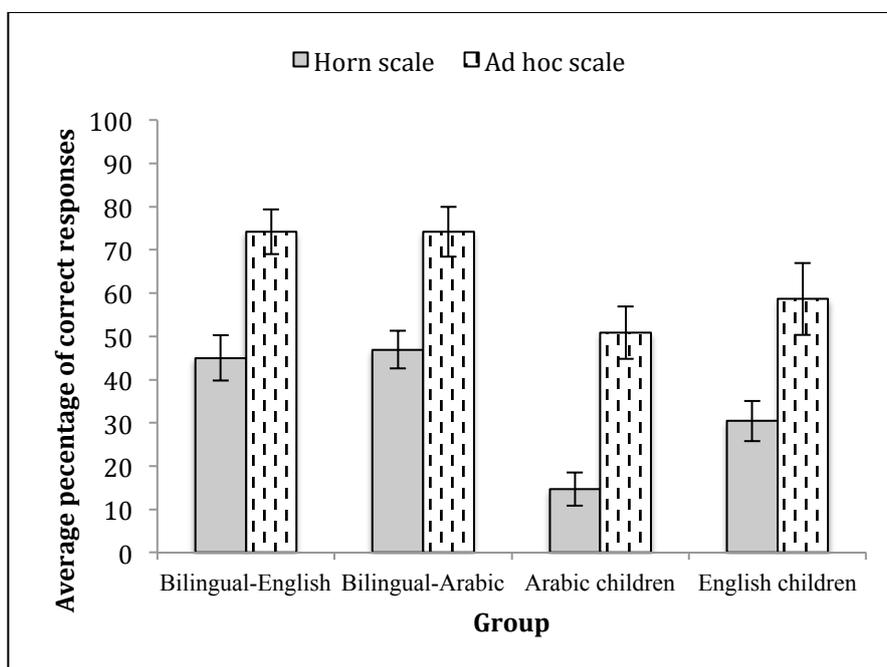


Figure 4.20. The child groups' pragmatic performance (%) on the Horn scale v. the ad hoc scale in experiment 3 (criteria for correct response: 'small' and 'medium' scored as 1 and 'large' as 0). Error bars represent standard error of the mean

To find out whether the differences were statistically significant I conducted within-group comparisons with Wilcoxon signed-rank tests for paired samples. The outcomes of the tests revealed that the bilingual children's performance in English significantly differed from scale to scale ($Z=-3.69$, $p<0.001$), and there was also a significant difference in their performance in Arabic ($Z=-3.03$, $p=0.002$). The Arabic and English children's performance on both the ad hoc and Horn scales was also significantly different ($Z=-4.006$, $p<0.001$), ($Z=-4.45$, $p=0.001$), respectively. The findings of the parametric t-tests confirmed all these significant differences.

Pragmatic performance on Horn v. encyclopaedic scale in experiment 4

The performance of each child group on the Horn and encyclopaedic scales in experiment 4 is displayed in figure 4.21. It can be seen that the bilingual children performed slightly better on the encyclopaedic scale than on the Horn scale when tested in English, but equally well on the two scales when tested in Arabic. The Arabic children scored higher on the encyclopaedic scale, while the English children's performance in the encyclopaedic scale was clearly better than on the Horn scale.

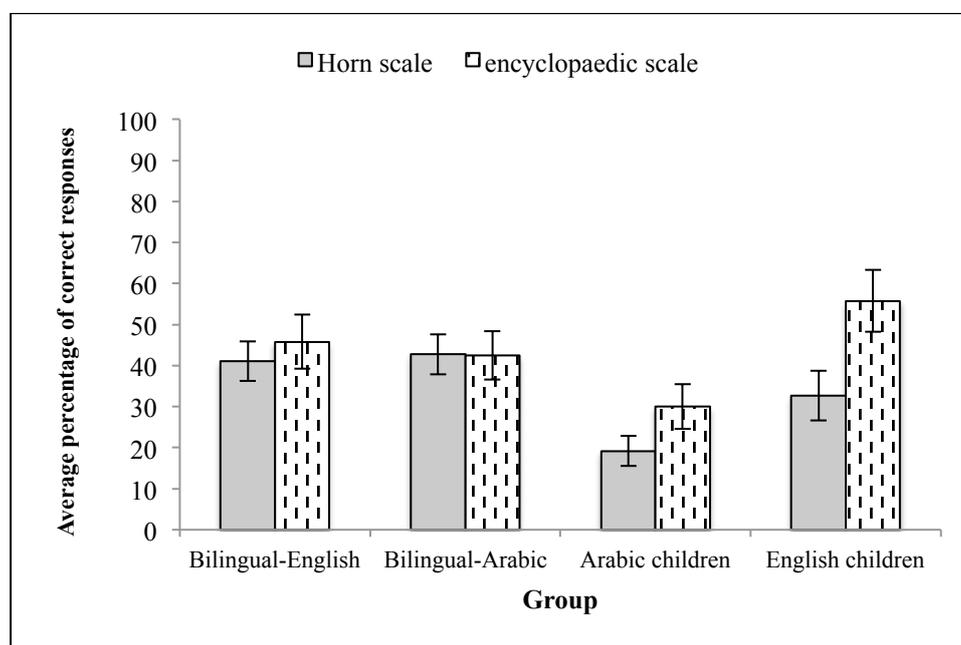


Figure 4.21. Child groups' performance (%) on the two scales (Horn v. encyclopaedic) in experiment 4 (criteria for correct response: 'small' and 'medium' scored as 1 and 'large' as 0). Error bars represent standard error of the mean

Next, I explored whether the above-discussed differences were statistically significant using the Wilcoxon signed-rank paired-sample test, which revealed no significant difference in bilingual performance between the two scales either in English ($Z=-.76$, $p=0.45$) or in Arabic ($Z=.011$, $p=0.99$). The results did, however, show a marginally significant difference in the Arabic children's performance between the two scales ($Z=-1.91$, $p=0.056$), and a significant difference in the English children's performance ($Z=-3.003$, $p=0.003$). The parametric t-test (paired-sample) also showed only a marginal difference in the Arabic children's performance on the two scales ($t(29)=-$

1.993, $p=0.056$), but a significant difference in the English children’s performance ($t(25)=-3.591$, $p=0.001$).

Comparisons of scale by context: Enriched context v. no context

Horn scale: Context (experiment 3) v. no context (experiment 4)

Figure 4.22 compares the performance of each child group in the Horn (lexical) scale in the two context conditions: enriched context (experiment 3) versus no context (experiment 4). It can be seen that each of the child groups had approximately similar performance on the Horn scale with the enriched context (experiment 3) and no context (experiment 4). The bilingual children’s performance in English as well as in Arabic on the Horn scale decreased slightly in experiment 4 from experiment 3. In contrast, the Arabic and English children’s performance without context (experiment 4) increased slightly from their performance with context (experiment 3).

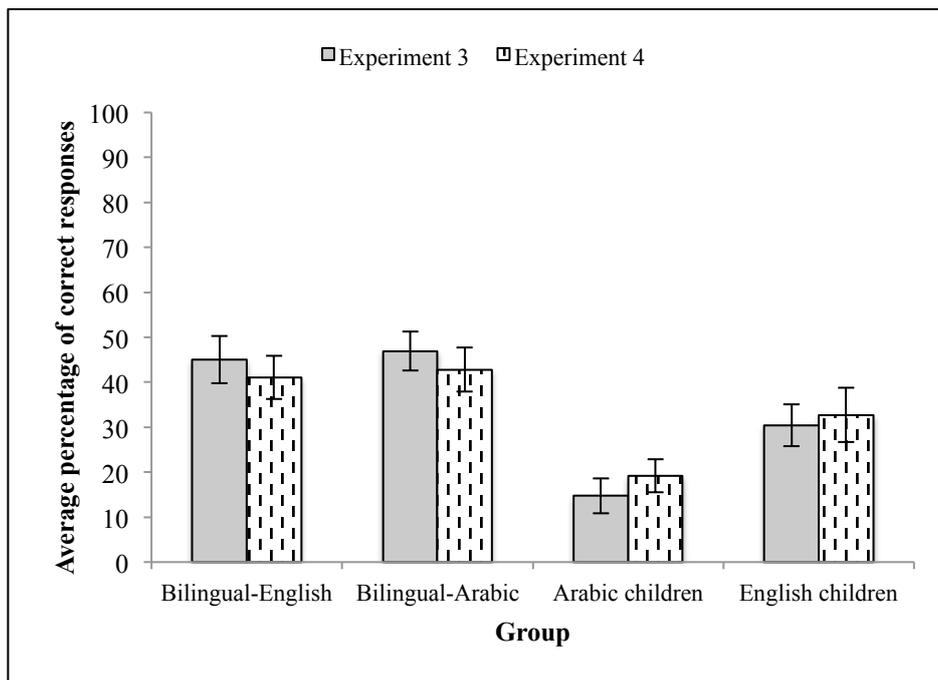


Figure 4.22. Child groups’ performance (%) on Horn lexical scale in experiment 3 v. experiment 4 (context v. no context) (criteria for correct response: ‘small’ and ‘medium’ scored as 1 and ‘large’ as 0). Error bars represent standard error of the mean

I investigated the differences within each group using the Wilcoxon signed-rank test for paired samples, and the outcomes revealed no significant difference between bilingual performance in the two context conditions in English ($Z=-.36$, $p=0.72$) or in

Arabic ($Z=-.8$, $p=0.43$). Similarly, the results showed no significant differences in the Arabic children's performance between the two context conditions ($Z=-1.36$, $p=0.18$), or in the English children's performance ($Z=-.02$, $p=0.98$). These findings are compatible with the parametric t-test (paired-sample) outcomes.

Ad hoc v. encyclopaedic scale: Context (experiment 3) v. no context (experiment 4)

Figure 4.23 presents a comparison between each group's performance on the ad hoc and encyclopaedic scales in the two context conditions. All groups scored clearly higher with enriched context (experiment 3) than with no context (experiment 4), except the English children, who performed almost equally in the two conditions. The bilingual children's performance on the two scales was very similar when tested in English and in Arabic: in both languages, the bilingual children's average score in experiment 3 was around 70%, and in experiment 4, around 43%. The Arabic children's average score on the ad hoc scale was around 50%, and clearly declined on the encyclopaedic scale (to 30%). Finally, the English children performed roughly the same in the two scales, with an average score around 56%.

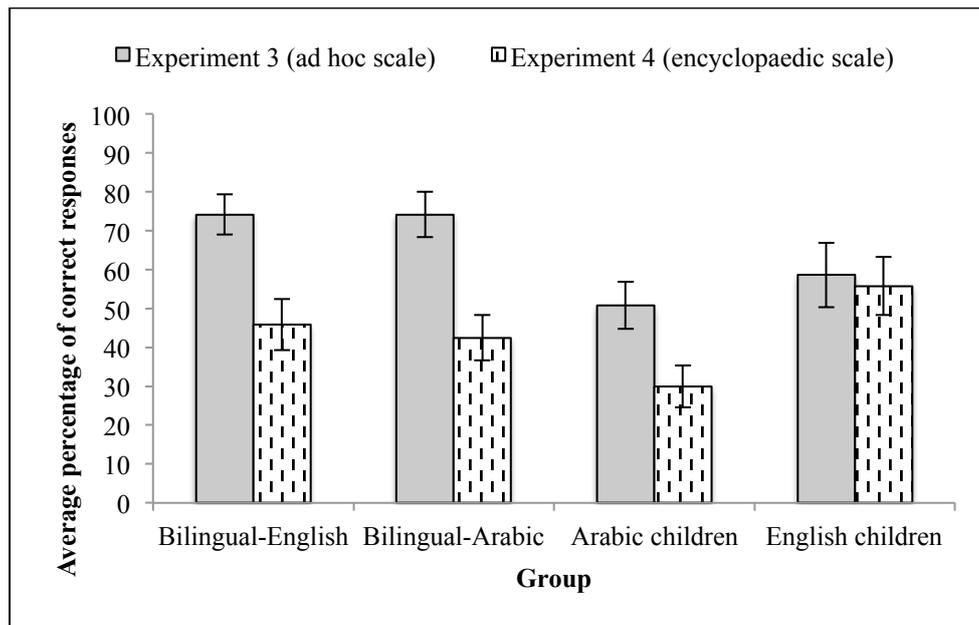


Figure 4.23. Child groups' performance (%) on the ad hoc v. the encyclopaedic scale in experiment 3 v. experiment 4 (context v. no context). (Criteria for correct response: 'small' and 'medium' scored as 1 and 'large' as 0.) Error bars represent standard error of the mean

The investigation of differences in ad hoc performance between the two context conditions revealed significant differences in the bilingual children's performance (using Wilcoxon signed-rank test) in English ($Z=-2.96$, $p=0.003$) and in Arabic ($Z=-3.357$, $p=0.001$). The Wilcoxon signed-rank tests also showed a significant difference in the Arabic children's performance between the two context conditions ($Z=-2.29$, $p=0.022$), but no significant difference in the English children's performance in the two conditions ($Z=-.144$, $p=0.88$).

These results are totally consistent with the parametric t-test (paired-sample) outcomes, which also showed a highly significant difference in bilingual performance between the two context conditions in English ($t(29)=3.66$, $p=0.001$) and in Arabic ($t(29)=4.13$, $p<0.001$), and also a significant difference in the Arabic children's performance between the two context conditions ($t(29)=5.51$, $p=0.018$); again, as with the non-parametric tests, there were no significant differences in the English children's performance in the two context conditions ($t(25)=.320$, $p=0.75$).

4.4.2.4 A summary of the new pragmatic analyses

This section re-investigated the child and adult participants' pragmatic performance on the under-informative/infelicitous items in experiments 3 and 4 (respectively) after applying new binary scoring criteria, which aggregated 'penalising' responses as correct and treated 'accepting' responses as incorrect.

In experiment 3, after applying the new scoring criteria, the bilingual participants' results showed no significant difference in performance between English and Arabic on under-informative 'most', 'some', or 'and', but there was a significant difference in performance on 'or'. When comparing the bilinguals-in-Arabic with the Arabic children, the outcomes revealed significant differences on all the quantifiers/operators, with the bilinguals scoring significantly higher. Conversely, although the bilinguals-in-English scored higher than the English children on all the quantifiers/operators, the difference between the two groups was only significant for 'some', and marginally significant for 'or'. The adult results showed that the Arabic adults exhibited a ceiling effect on all four quantifiers (100%), while the English adults showed a ceiling effect on 'most' and scored high on 'some' (95%) and 'or' (90%) but did not penalise the under-informative 'and' at quite the same rate (77%). Comparing the two adult groups' performance on 'and', the difference was significant.

In experiment 4, the child results showed no significant differences in bilinguals' performance between English and Arabic on any of the four quantifiers/operators ('most', 'some', 'or', or 'and'). Comparison between the bilinguals-in-Arabic and the Arabic children revealed statistically significant differences for 'most', 'some', and 'or', with the bilinguals scoring higher, but the groups did not significantly differ on 'and'. Comparison between the bilinguals-in-English and the English children showed no significant differences between groups, although the bilinguals scored higher on 'most', 'some', and 'or' and the English scored higher on 'and'. With respect to the adults' performance, the results revealed that the Arabic adults penalised the infelicitous 'most' and 'and' significantly more than the English adults, while the two groups performed almost equally on 'some' and 'or', with high rates of penalisation (more than 90%).

Finally, the analyses in this section also explored children’s performance on implicatures resulting from two different scales—Horn, ad hoc and encyclopaedic scales. Table 4.41 compares the results for children in each group on both the two scale; (+) is added to indicate better pragmatic performance. When comparing pragmatic performance on the Horn and ad hoc scales in experiment 3, significant differences were found in all the groups—all of them performed significantly better on the ad hoc. In experiment 4, there were no significant differences in the bilingual (in Arabic and English), a marginal significant difference in the Arabic children’s pragmatic performance on the two scales (Horn v. encyclopaedic scale), and the English children performed significantly better on the encyclopaedic scale.

Table 4.41. Summary of children’s pragmatic performance on Horn and ad hoc scales

Group	Horn v. other scales		Enriched context v. no context	
	Enriched context (experiment 3)	No context (experiment 4)	Horn scale	Other scales
Bilinguals-in-English	S (Ad hoc+)	NS	NS	S (Context+)
Bilinguals-in-Arabic	S (Ad hoc+)	NS	NS	S (Context+)
Arabic children	S (Ad hoc+)	MS	NS	S (Context+)
English children	S (Ad hoc+)	S (encyclopaedic +)	NS	NS

Note: The ‘other scales’ in the table refers to the ad hoc scale in the enriched context task and the encyclopaedic scale in the no context task. Abbreviations in the table: S (significant), MS (marginally significant) NS (non-significant)

The comparison between children’s pragmatic performance on the Horn scale in experiment 3 and that in experiment 4 revealed no significant differences in any group. On the ad hoc v. the encyclopaedic scale, the outcomes showed significant differences for the bilingual children (in Arabic and English) and the Arabic children; all performed better with the enriched context (experiment 3). The ad hoc v. encyclopaedic comparison revealed no significant difference in the English children’s pragmatic performance between experiments 3 and 4.

4.4.3 Cognitive performance

The tasks in this section were included in this study to answer, in part, the third research question, related to children’s cognitive performance. This section explores participants’ performance on two cognitive tasks: first, the Simon task, and then, the

Corsi blocks task. For each task, the study presents the results of the analysis for the child groups, explores the predictors of their cognitive performance, then briefly explores the adult groups' performance. Finally, a summary of the main findings for both cognitive tasks is given.

4.4.3.1 The Simon task

This task was conducted as a measure of cognitive conflict inhibition; more precisely, it aimed to assess children's ability to suppress interference from conflicting stimuli. The analysis first looks at the accuracy of performance of the task, then it explores the average reaction times (RTs) needed to complete the congruent and incongruent trials. After this, it investigates the differences between the groups on the Simon effect, which is the gap in RTs when correctly completing congruent trials versus incongruent trials.

Children's performance on the Simon task

In each part of the analysis presented below, I started with numerical results for the bilingual, Arabic, and English children's performance on the task, then explored whether the difference between groups was statistically significant for each.

Children's accuracy in the Simon task

The accuracy of each child group on the Simon task was next investigated. The results for the bilingual, Arabic, and English children's average accuracy in the two congruent and incongruent conditions are displayed in figure 4.24. Overall, the children had lower accuracy in the incongruent condition. Differences between the three groups were tiny in the congruent condition: the bilingual and English children had the same average accuracy (13.6 correct trials out of 14 trials), while the Arabic children had slightly lower accuracy (13.4). The differences between groups increased marginally in the incongruent condition, where the bilingual children had a slightly higher average accuracy (around 13.1 correct trials) than the other two groups (approximately 12.5 correct trials each).

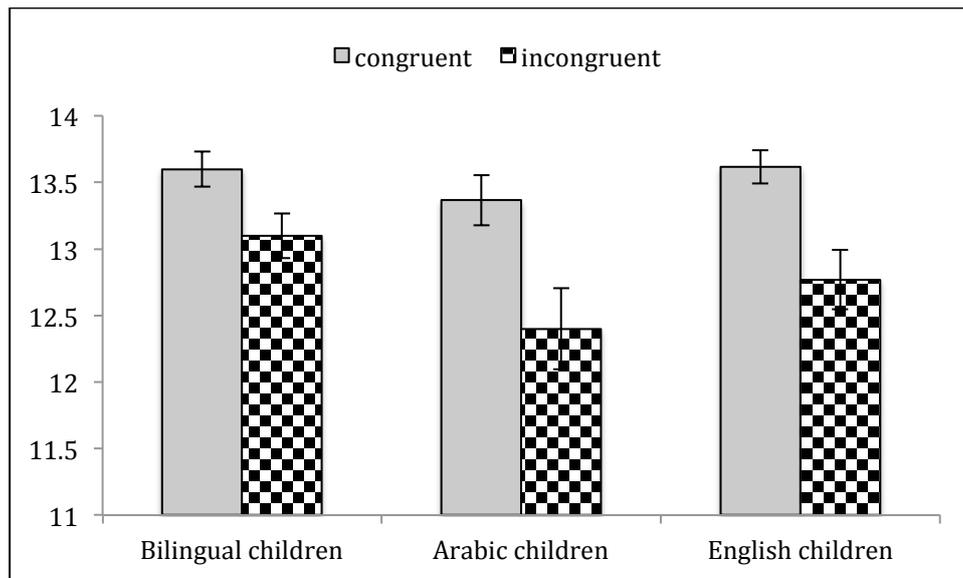


Figure 4.24. Child participants' average accuracy in the Simon task (by condition). Error bars represent standard error of the mean

Before exploring whether the differences between groups were statistically significant, I checked the assumption of normality. The Shapiro–Wilk test revealed that none of the groups' data were normally distributed in either of the two conditions, and the Kolmogorov–Smirnov test revealed that neither was the whole sample (which was expected, since the majority of children exhibited a ceiling effect in the congruent condition and had similar accuracy in the incongruent condition). Thus, I investigated the difference first with parametric tests, then with non-parametric tests. A 2 (congruency: congruent v. incongruent) x 3 (group: bilingual, Arabic, English children) mixed ANOVA was conducted, with accuracy as the dependent variable, congruency as a within-subject factor and language group as a between-group factor. The results of Mauchly's test revealed that the assumption of sphericity was violated ($W=1$, $p<0.05$); therefore, the degrees of freedom were corrected using Huynh–Feldt estimates of sphericity ($\epsilon=1>0.75$). The ANOVA results showed a significant effect of congruency ($F(1, 83)=23.12$, $p<0.001$), a marginally significant difference between the groups ($F(2, 83)=2.69$, $p=0.074$), and no significant interaction between group and congruency ($F(2, 83)=.798$, $p>0.05$). Post hoc comparisons (Games–Howell) showed no significant difference between the bilingual children and either the Arabic or the English children, or between the Arabic and English children ($ps>0.05$). The pairwise comparisons of accuracy in the congruent condition (with the Mann–Whitney U-test) revealed no significant difference between the bilingual and Arabic children

($U=389$, $Z=-1.06$, $p=0.29$), the bilingual and English children ($U=387$, $Z=-.61$, $p=0.59$) or the Arabic and English children ($U=341$, $Z=-.94$, $p=0.35$). Similar results were found when comparing accuracy in incongruent condition (bilingual v. Arabic children ($U=360$, $Z=-1.39$, $p=0.16$), bilingual v. English children ($U=332$, $Z=-1.01$, $p=0.31$), and Arabic v. English children ($U=358$, $Z=-.54$, $p=0.59$)).

Children’s performance on congruent and incongruent conditions

Mean RTs for correct responses in each condition of the Simon task (congruent and incongruent) are displayed in figure 4.25 for all the child groups. It can be seen that while the bilingual and English children had approximately the same average RT in the incongruent condition, the English were slightly faster in the congruent condition. It can be noticed also that the Arabic children were somewhat faster than the other groups in completing the task in the two congruent conditions.

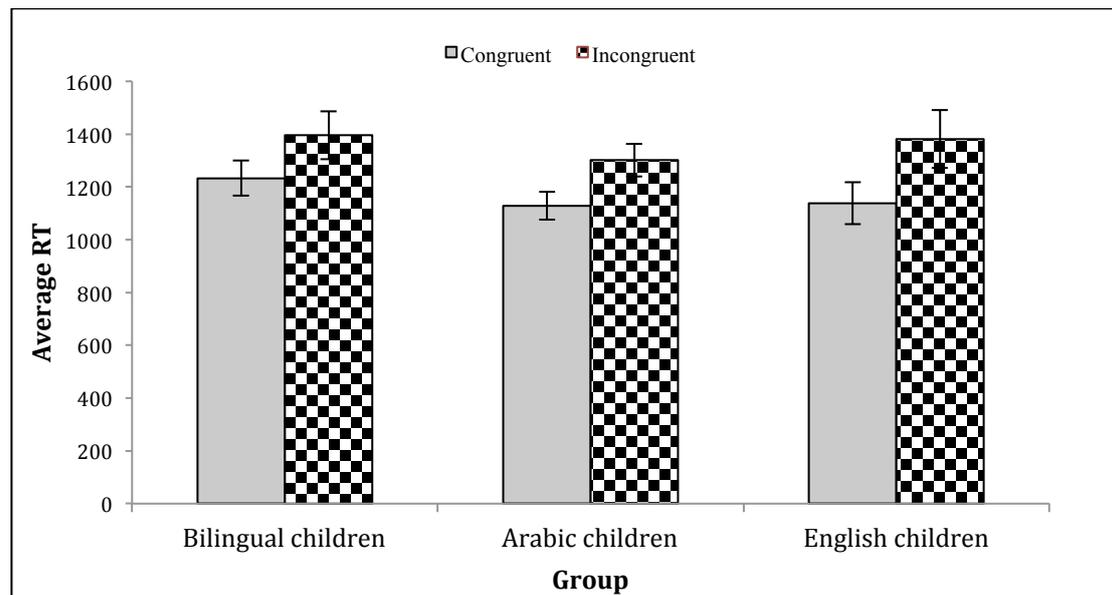


Figure 4.25. Average RT of correct responses in the congruent and incongruent trials of the Simon task for the bilingual, Arabic, and English children. Error bars represent standard error of the mean

Before proceeding to find out if the above numerical differences between groups are statistically significant, I first checked the assumption of normality, which the tests revealed was violated in all cases ($p < 0.05$). Therefore, I adopted two ways of

exploring differences between groups. First, since Simon task data have no 0-value, I was able to transform them—specifically, to attempt to make them normally distributed—using a log function. However, I checked the normality of the transformed results and found that in the incongruent condition, the whole sample and the Arabic and English children’s data still violated the normality assumption. Thus, I first applied a parametric mixed ANOVA and then investigated the differences between groups using non-parametric tests.

I conducted a 3 (group: bilingual children, Arabic children, English children) x 2 (congruency: congruent v. incongruent) ANOVA with group as between-subject factor and congruency as within-subject factor. The results of Mauchly’s test showed that the assumption of sphericity was violated ($W=1$, $p<0.001$); therefore, the degrees of freedom were corrected using Huynh–Feldt estimates of sphericity ($\epsilon=1>0.75$). The test outcomes revealed a significant effect of congruency ($F(1, 83)=60.5$, $p<0.001$) but no effect of group ($F(2, 83)=0.221$, $p=0.80$) and only a marginally significant interaction between group and congruency ($F(2, 83)=2.72$, $p=0.072$). Post hoc comparisons (Games–Howell) showed no significant difference between the groups ($ps>0.05$).

Next, I examined the differences between groups using non-parametric tests. Starting with the child groups’ performance on the congruent condition, the Mann–Whitney U-tests revealed no significant difference between the bilingual and Arabic children ($U=388$, $Z=-.971$, $p=0.36$), the bilingual and English children ($U=277$, $Z=-1.53$, $p=0.12$), or the Arabic and English children ($U=342$, $Z=-.789$, $p=0.43$).

The investigation of children’s performance in the incongruent condition also showed no significant difference between the bilinguals and either the Arabic children ($U=392$, $Z=-.86$, $p=0.39$) or the English children ($U=378$, $Z=-.197$, $p=0.84$). The Mann–Whitney test revealed no significant difference between the Arabic and the English children ($U=349$, $Z=-.674$, $p=0.50$).

Since previous work (e.g. Martin-Rhee & Bialystok, 2008) has found a bilingual advantage in the Simon task only on global RT (i.e. the time taken to complete the whole task) I explored this aspect. Average global RTs are displayed in figure 4.26

for all the child groups. It can be seen that the Arabic children completed the task faster than the other two groups, and the English children were faster than the bilinguals. To find out if these differences between the groups are statistically significant, I conducted an ANOVA test with global RT as the dependent variable. The ANOVA results revealed no significant difference between the groups ($F(2, 83)=.40, p>0.05$), and none of the post hoc comparisons were significant (all $ps>0.05$).

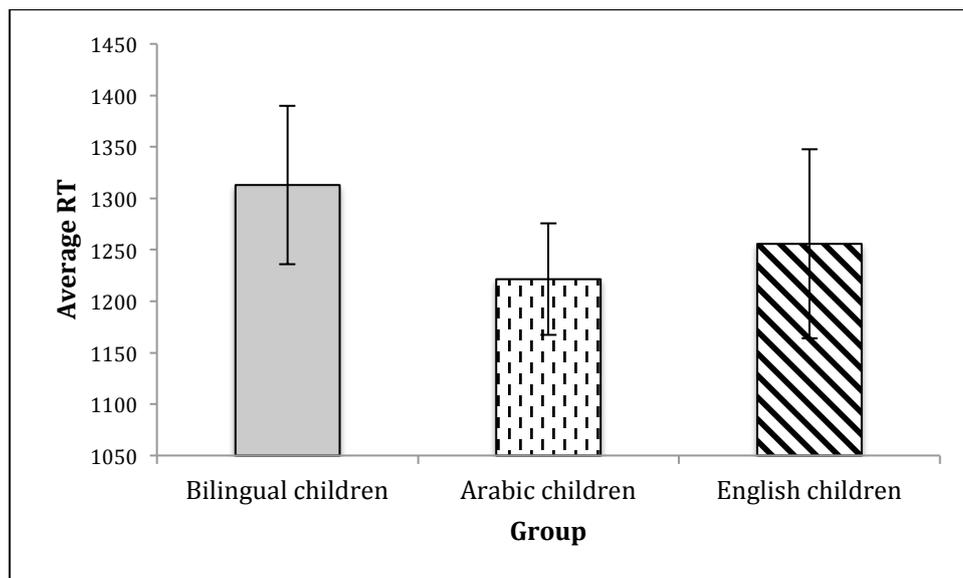


Figure 4.26. Average global RT of correct responses in the Simon task for the bilingual, Arabic, and English children. Error bars represent standard errors of the mean

Simon effect in the child groups

Although average RT of performance in each condition might provide some information on how children in each group performed on the two conditions, exploring differences in the conditions separately might more clearly reflect children’s cognitive inhibition ability, which could be measured by calculating the Simon effect (the difference between RTs of the conditions). Therefore, I computed the Simon effect for each group by subtracting the average incongruent RT for each participant from the average congruent RT for that participant. The average RT of the Simon effect for each group is shown in figure 4.27. It can be seen that the bilingual and Arabic children have roughly the same Simon effect RT, with the bilingual children’s being slightly smaller and clearly lower than that of the English children,

which might indicate better inhibitory skills for both the bilingual and Arabic children than for the English children.

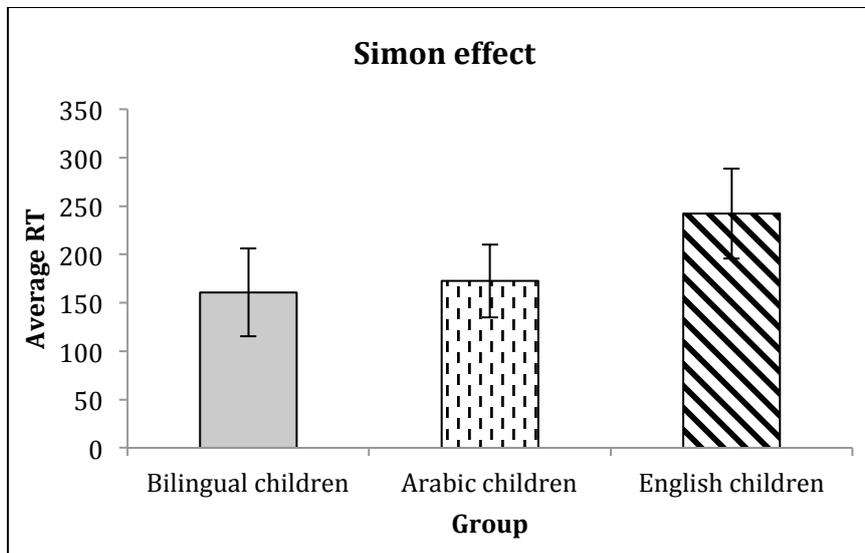


Figure 4.27. Average RT of the Simon effect for the bilingual, Arabic, and English children. Error bars represent standard error of the mean

To find out if there was any statistically significant difference between the groups, I first tested the assumption of normality. The results revealed that only the Arabic children's data were normally distributed; therefore, I transformed the overall results with a log-function to make all the groups and the whole sample normally distributed ($p > 0.05$). However, as I did with all data that violated normality in this study, I compared the groups with parametric as well as distribution-free tests. Starting with the parametric test, a 3 (group: bilingual children, Arabic children, English children) \times 1 (Simon effect) ANOVA was performed, with the Simon effect as a dependent variable. The test revealed no significant differences in the Simon effect between groups ($F(2, 83) = 1.83, p = 0.17$), nor did post hoc (Games–Howell) comparisons ($p > 0.5$). Pair-wise comparisons (with the Mann–Whitney U-test) revealed no significant difference between the bilingual and Arabic children ($U = 426, Z = -0.347, p = 0.73$), but there was a marginal difference between the bilingual and English children ($U = 278, Z = -1.84, p = 0.066$) and a significant difference between the Arabic and English children ($U = 255, Z = -2.218, p = 0.027$); however, this finding should be taken with caution, since the Arabic children had the lowest accuracy in the incongruent condition. I address this issue further in section 4.4.3.1.1.4 below.

Analysing children's performance in the Simon task using Cox proportional hazard regression

Traditional analysis of the Simon task, as found in the literature (and as conducted above), usually compares mean RTs across conditions (congruent vs. incongruent) and groups. Such an approach has been criticized for deleting or ignoring vital information, which could reveal significant outcomes, from the data (De Cat et al., under review). For example, De Cat et al. proposed that self-monitoring might have an effect in such cases, as the participant might slow down after noticing they had answered inaccurately in one trial; the possibility of such an effect, however, is usually ignored in traditional analysis. Furthermore, 'the trials form part of a time series, and there might be an effect of e.g. habituation or tiredness. Removing the trial immediately following an erroneous response results in further loss of data' (De Cat et al., under review, p. 27). More importantly, traditional analysis cannot capture accuracy and response time simultaneously, as incorrect responses are removed from data before calculating mean RTs, which could, again, lead to further loss of vital data.

These critical issues imply the need for a novel approach allowing better investigation of children's performance in the Simon task: the Cox proportional hazard (PH) model. Before doing so and reporting the children's resulting Simon task outcomes, I briefly clarify how this technique works (for a detailed description, see De Cat et al., under review). This technique takes into account both the time taken to answer correctly, and the time taken to answer incorrectly; as De Cat et al. explain, if a child A (with good inhibition ability) takes time X to respond correctly in a certain incongruent condition trial, child B (with bad inhibition ability) might be expected to take longer to respond accurately to that trial, or possibly a shorter time to respond inaccurately. The Cox PH model apprehends this by including time to an incorrect response as a *censored* observation, that is, one corresponding to 'the minimum amount of time it would have taken to produce a correct response in that trial' (De Cat et al., under review, p. 28).

In the Simon Task, the *event* in the Cox PH model is 'time to correct response' for a

trial (and is given a value of 1). The consequences of this for how data are treated are two. First, time length from stimulus performance to correct registered answer is regarded as an *uncensored*, that is, known, observation (De Cat et al., p. 29). Second, time to an incorrect answer is regarded as censored, since such a correct endpoint is not observed. Censored observations are included in the model, as they contain vital information: ‘they indicate that the amount of time to a correct response would have taken at least as long as that of the censored observation’ (De Cat et al., p. 29). Neglecting such information could lead to biased assessments.

Predictors for children’s performance in the Simon task (using Cox regression)

I fitted a Cox regression model with time to a correct response as a dependent variable; as covariates, I had congruency, bilingualism, group, age, SES (FAS), and NVIQ; item was included as a random effect. The results in table 4.42 show that children performed more poorly in the incongruent condition ($X^2(1)=52.42$, $p<0.001$); there was no significant effect of bilingualism ($X^2(1)=.054$, $p=0.59$) or group ($X^2(1)=.053$, $p=0.17$). The model revealed that age was the strongest predictor ($X^2(1)=211.3$, $p<0.001$) followed by NVIQ ($X^2(1)=75.1$, $p<0.001$), and that children with high SES performed better than those with medium or low SES ($X^2=9.88$, $p=0.002$).

Table 4.42. Coefficients of a Cox proportional hazard model fitted to the time to correct response

Covariates	B	SE	X²	DF	P
Condition (incongruent)	-.306	.042	52.415	1	.000
Bilingual (yes)	.030	.054	.298	1	.585
Group	.074	.053	1.916	1	.166
SES (FAS)	-.118	.038	9.886	1	.002
Age (month)	.046	.003	211.276	1	.000
NVIQ	.103	.012	75.100	1	.000
Item	.004	.003	2.847	1	.092

Note: Model reference levels: For condition (incongruent), for language background (bilingual), for group (English children), for SES (high FAS) (the effect of FAS was similar when the FAS raw score instead of categorical score was used)

From this model, the adjusted scores (propensity scores) were calculated for each child; the score captures the effect of all these covariates on children’s performance in

the Simon task. Since there was an adjusted value for each row (trial), the average of these values was calculated and given as a child-adjusted score in the Simon task. This final score will be used as a predictor of children’s pragmatic performance in section (4.4.4), effectively corrected for age, SES, and NVIQ.

Adults’ performance on the Simon task

Similar to the above section on children’s performance in the Simon task, the analyses of adults’ performance investigated the average accuracy of each group, then explored their RTs in the two congruency conditions (congruent and incongruent), and finally examined the difference in Simon effect between the groups.

Adults’ accuracy on the Simon task

The two adult groups’ average accuracy on the Simon task across the conditions of congruency is presented in figure 4.28; it is almost the same (around 13.9 in the congruent condition and 13.4 in the incongruent condition). The Shapiro–Wilk tests results revealed that neither the groups nor the whole sample are normally distributed ($p < 0.05$), therefore, I investigated the differences with non-parametric tests first, then parametric tests. The non-parametric Mann–Whitney U-test showed no significant difference between groups in the congruent condition ($U=54$, $Z=-.069$, $p=0.79$) or the incongruent condition ($U=51$, $Z=-.317$, $p=0.81$), a finding completely compatible with parametric t-test outcomes.

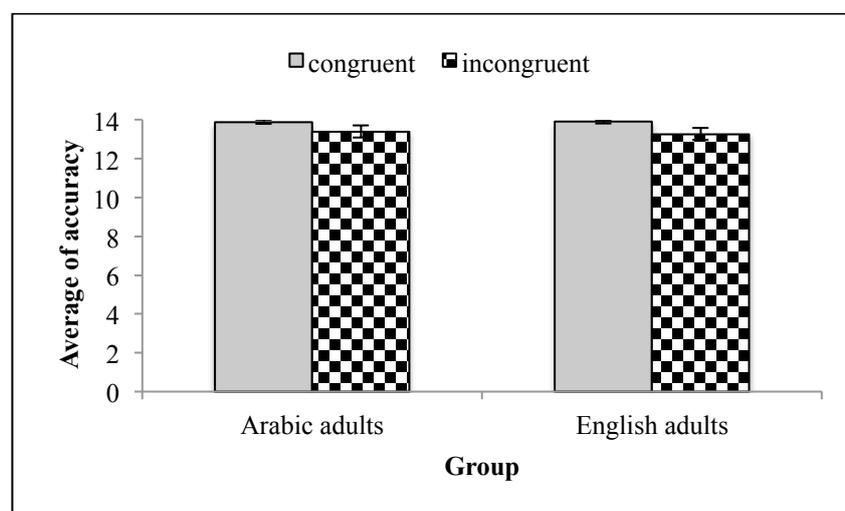


Figure 4.28. Adult participants' average accuracy on the Simon task. Error bars represent standard error of the mean

Adults' performance by congruency condition

Figure 4.29 shows the Simon task results for the two adult groups, across two conditions of congruency. The English adults completed the task slightly faster (around 60 seconds faster) than the Arabic adults did, in both conditions.

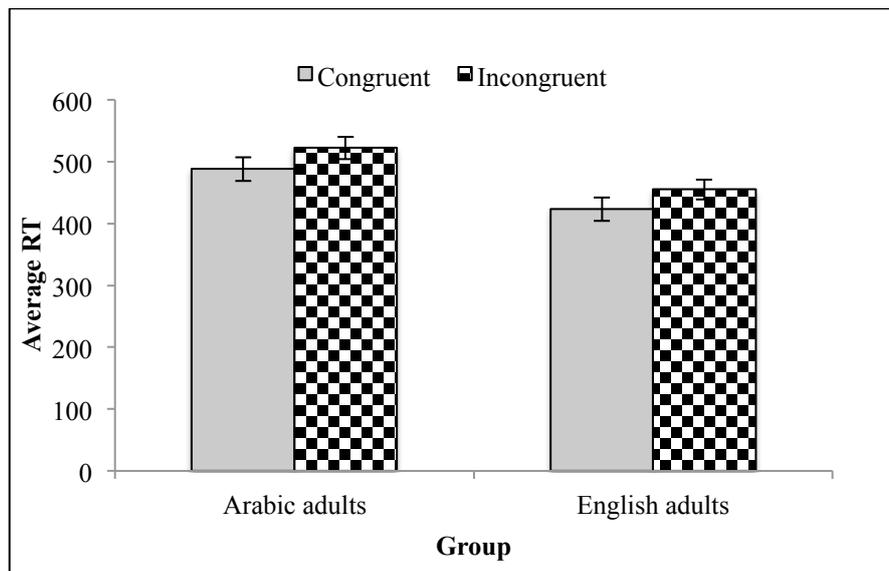


Figure 4.29. Average RT of correct responses in congruent and incongruent Simon task trials for Arabic and English adults. Error bars represent standard error of the mean

Before exploring whether the differences between the two groups might be statistically significant, I checked that the sample was normally distributed in the two conditions, using the Shapiro–Wilk test of normality for small samples. This investigation revealed that the English adults' performance in the incongruent condition was not perfectly distributed ($p=0.046$), while the Arabic adults met the assumption of normality in both conditions ($p>0.05$). The t -test results revealed significant differences between the two groups in the congruent condition ($t(19)=2.42$, $p=0.026$) and the incongruent condition ($t(19)=2.79$, $p=0.012$). These results were completely compatible with the non-parametric Mann–Whitney U -tests results.

The last aspect I investigated was the adults' average global RTs. As displayed in figure 4.30, the English adults completed the task faster. The Shapiro–Wilk test of normality revealed that the two groups and the whole sample were normally distributed ($p>0.05$), and a comparison between the groups (with T-test) showed a significant difference between groups in global RT ($t(19)=2.78$, $p<0.05$). These outcomes were totally consistent with the Mann–Whitney test results.

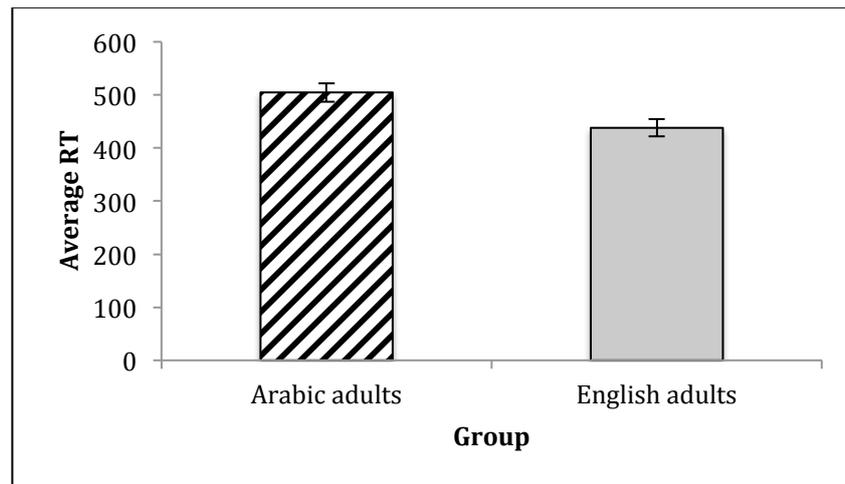


Figure 4.30. Average global RT of correct responses in the Simon task trials for Arabic and English adults. Error bars represent standard error of the mean

Simon effect in the adult groups

The average RT of the two adult groups' Simon effect (incongruent RT–congruent RT) is presented in figure 4.31. It can be seen that the two groups had almost the same Simon effect (Arabic adults 34, English adults 32). The Shapiro–Wilk test of normality revealed that both groups as well as the whole sample were normally distributed ($p>0.05$). T-test results showed no significant difference between groups in terms of the Simon effect ($t(19)=.113$, $p=0.91$). These outcomes were totally consistent with the Mann–Whitney test results.

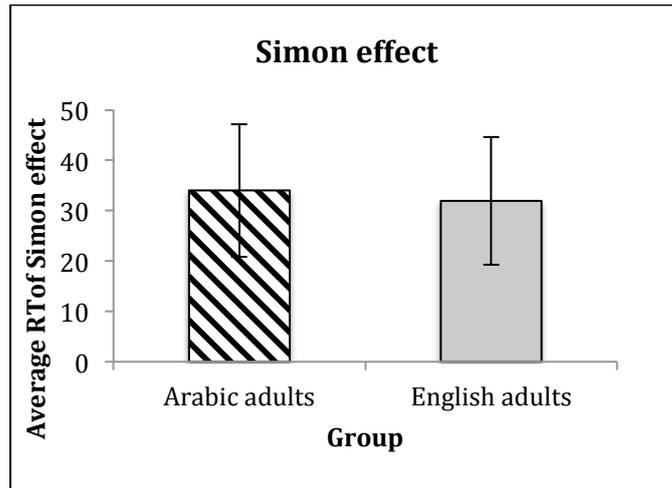


Figure 4.31. Average RT of the Simon effect for the Arabic and English adults. Error bars represent standard error of the mean

4.4.3.2 The Corsi blocks task

This task aimed to measure participants' visuo-spatial STM. The final score given to each participant represents the highest number of circles she/he could copy correctly, or her/his STM span. The analyses first explored numerically the average score of each group, then examined whether the differences were statistically significant.

Children's performance on the task

Figure 4.32 displays the average score on the Corsi block task for each child group. It can be seen that the bilingual children scored slightly higher (4.13) than the English children (3.96), while the Arabic children had the lowest STM span (3.2).

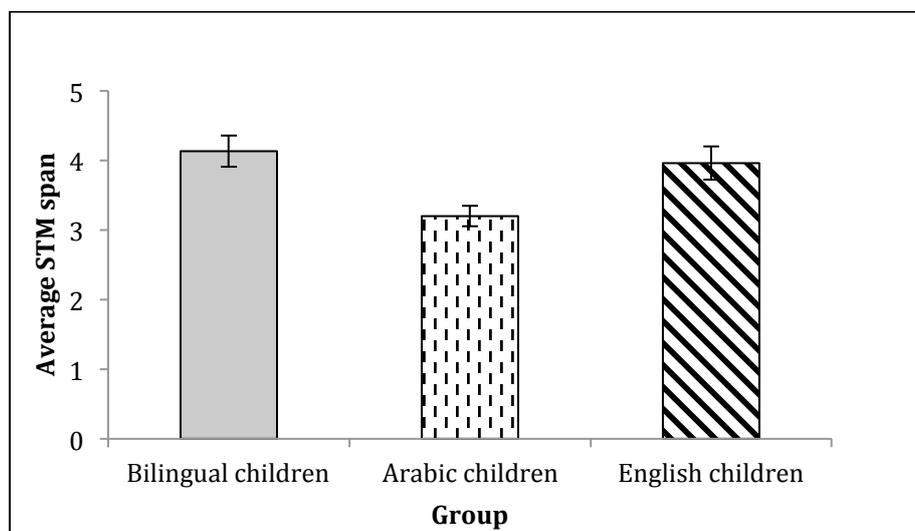


Figure 4.32. Average scores of bilingual, Arabic, and English children on the Corsi blocks task as a measure of STM span. Error bars represent standard error of the mean

To find out if such differences were statistically meaningful, I first explored the assumption of normality, then conducted comparisons between the groups. The results of Shapiro–Wilk tests revealed that none of the three child groups were normally distributed, and the Kolmogorov–Smirnov results showed that the whole sample violated the assumption of normality as well. The Kruskal–Wallis H-test revealed that the differences between groups were significant ($H(2)=10.98$, $p=0.004$), and pair-wise comparisons between groups (with the Mann–Whitney U-test) showed a significant difference between the bilingual and Arabic children ($U=245$, $Z=-3.147$, $p=0.002$) but no significant difference between the bilingual and English children ($U=359$, $Z=-.524$, $p=0.6$). The pair-wise comparison also revealed a significant difference between the Arabic and English children ($U=246$, $Z=-2.48$, $p=0.013$). These findings are completely consistent with the parametric ANOVA results.

Predictors of children’s STM span

I ran a bivariate correlation test to find any correlation between children’s performance on the Corsi blocks task and the independent variables: age, FAS, NVIQ, L2 input, group (bilingual, Arabic and English children), and bilingualism (bilingual v. not bilingual). The test results revealed significant correlations between STM and age ($r(\text{two-tailed})=.27$, $p=0.004$) and between STM and bilingualism ($r(\text{two-tailed})=.25$, $p=0.008$). There was no significant correlation with STM or any of the other independent variables ($ps>0.05$).

Next, I fitted a regression model (GLM) with STM as a dependent variable and age and interaction effects of group and bilingualism as predictors. The regression test results are presented in table 4.43. The model revealed a significant main effect of age ($X^2=4.33$, $p=0.037$) and also a significant effect of bilingualism in interaction with group ($X^2=12.976$, $p=0.002$). When we look at the effect on each group, it can be seen that there was no significant difference between the bilingual and English children ($X^2=.42$, $p=0.52$) but there was a significant difference between the English

and Arabic children ($X^2=7$, $p=0.008$), with the Arabic children having lower STM as indicated by the model negative estimate ($B=-.74$).

Table 4.43. Coefficients of a GLM regression model fitted to Corsi Blocks score

Parameter	B	SE	X ²	DF	P
(Intercept)	3.97	.206	372.32	1	.000
Age (month)	.032	.016	4.33	1	.037
Group*Bilingualism	.	.	12.98	2	.002
(Bilingual children)*Bilingualism	.181	.281	.417	1	.518
(Arabic children)*Bilingualism	-.743	.281	7.00	1	.008
(English children)*Bilingualism	0

Note: The reference level for group is 'English children'

Adults' performance on the task

The results of the Arabic and English adult groups on the Corsi blocks task are presented in figure 4.33, and show only a slight difference in average short-term memory span, with the English adults (7.27) having slightly longer span than the Arabic adults (6.8).

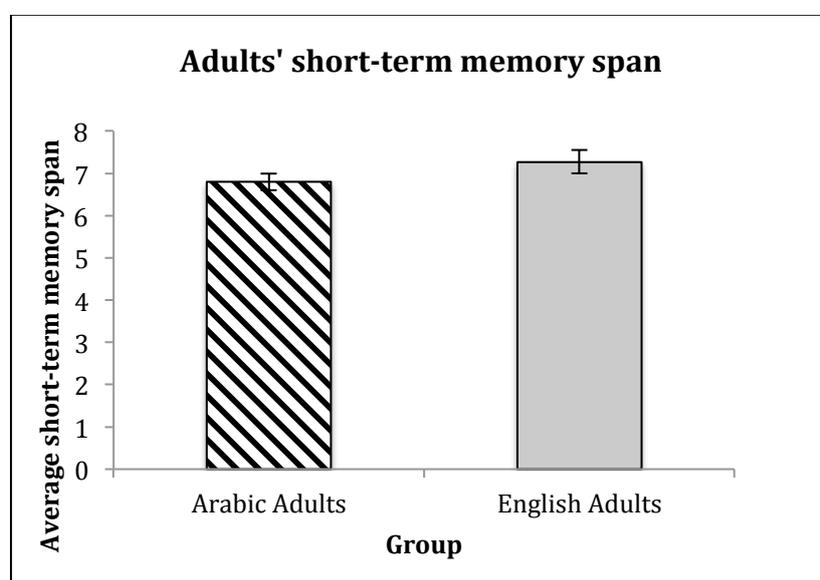


Figure 4.33. Average scores of Arabic and English adults on the Corsi block task as a measure of short-term memory span. Error bars represent standard error of the mean

The outcomes of Shapiro–Wilk tests of normality revealed that both groups as well as the whole sample violated the assumption of normality ($ps<0.05$). Therefore, I now

report the results of comparisons conducted with non-parametric tests, and then briefly report the compatibility of these results with the parametric test outcomes. The Mann–Whitney U-test showed no significant difference between the two adult groups' performance on the Corsi blocks task ($U=36.5$, $Z=-1.38$, $p=0.17$); these findings are totally consistent with the parametric t-test findings.

4.4.3.3 A summary of cognitive performance

The analyses in this section explored the child and adult participants' performance on two cognitive tasks: the Simon task (measuring cognitive inhibition ability) and the Corsi blocks task (assessing STM span).

The results of the Simon task revealed that none of the child groups—bilingual, Arabic, or English children—significantly differed in RT performance or in their accuracy (average score of correct trials) between the two congruency conditions. However, when calculating the Simon effect (incongruent RT–congruent RT), the differences between the groups became more significant: comparisons between groups (including non-parametric tests) revealed a marginal difference between the bilingual and English children and a significant difference between the Arabic and English children, but no significant difference between the bilingual and Arabic children. The Cox regression analysis, however, did not confirm these significant differences and its results revealed that children's age, NVIQ, and SES best predicted their inhibition ability. With respect to the adult groups' performance, the outcomes revealed that although the English adults were significantly faster to complete the task (as revealed by the comparison of global RT), the groups had almost the same Simon effect. The analyses showed that the two groups did not significantly differ in any of RT by congruency or accuracy.

Performance on the second cognitive task, the Corsi blocks, revealed that the bilingual children had the longest STM span and the Arabic children the shortest. The analyses showed significant statistical differences between the bilingual and Arabic children and between the Arabic and English children, but no significant difference between the bilingual and English children. The regression analysis showed that age was the best predictor of children's STM span. In regard to adult performance on the

task, the results revealed that the Arabic and English adults had very similar STM span, and the difference between the two groups was not statistically significant.

4.4.4 The relationship between children's pragmatic and cognitive performance

A bivariate correlation test revealed significant correlations between children's pragmatic binary responses and each language group (bilingual-in-English, bilingual-in-Arabic, Arabic, English) ($r(\text{two-tailed})=.171$, $p<0.001$), context condition (enriched context v. no context) ($r(\text{two-tailed})=.079$, $p=0.02$), type of quantifier ('most', 'some', 'or', 'and') ($r(\text{two-tailed})=.141$, $p<0.001$), NVIQ ($r(\text{two-tailed})=.085$ $p=0.01$), STM ($r(\text{two-tailed})=.147$, $p<0.001$); a marginally significant correlation with Simon adjusted score ($r(\text{two-tailed})=.063$ $p=0.054$); but a strong negative correlation with the Simon effect ($r(\text{two-tailed})=-.090$, $p<0.001$), and also a strong negative correlation with vocabulary ($r(\text{two-tailed})=-.130$, $p<0.001$). The test did not show correlations with age or SES (FAS score, either categorical or continuous) (all $ps>0.05$).

To explore which of the independent and experimental variables influenced children's pragmatic performance in experiments 3 and 4, I first conducted a GLM regression (with Poisson log function) based on the total score resulting from giving one point to 'small' and 'medium' responses and 0 to 'large' responses. I fitted the model by including the variables that correlated with the dependent variable (total score), and checked the goodness-of-fit. Due to the high correlation between NVIQ and vocabulary score, and also the correlation with the Simon adjusted score, I removed NVIQ from the model, and it improved slightly. Then, I added FAS score and age to the model, and the goodness-of-fit improved. Also, as I was interested to explore the effect of EF measures on the language groups, I added an interaction effect between each of the EF measures and language group, but this only added a tiny improvement to the model goodness-of-fit.

The results of the final model are given in table 4.44 below. The model has language group, context condition, and quantifier/operator type as fixed factors, and participants' age, vocabulary score, STM, and adjusted Simon score as covariates. The model was fitted based on the variables that significantly correlated with the dependent variable (added one by one to the model after checking the goodness-of-

fit); then, I added other exploratory variables that did not correlate with it and checked the goodness-of-fit again. Since I was interested in finding out which group might be significantly affected by the two cognitive measures, I added an interaction effect between each of STM and inhibition and group .

As table 4.44 shows, the regression analysis revealed a strong main effect of language group ($X^2(3)=10.87$, $p=.013$), and the model revealed that the bilinguals performed marginally better than the English children in English ($p=0.065$) and significantly better than the English children when tested in Arabic ($p=0.014$). Although the model estimate (B) suggests that the English gave more pragmatic responses than the Arabic children, the difference between the groups was not significant. The regression results also showed a significant effect of context condition ($X^2(1)=7.72$, $p=0.005$), and the model estimates ($B=.148$) indicate that children performed pragmatically better in the enriched context condition.

The model further revealed that SES (as a categorical variable) has a strong effect on children's pragmatic performance; taking children with high SES as a reference, the results showed that children with low SES had significantly poorer pragmatic performance ($X^2(1)=20.18$, $p<.001$, $B=-.622$), while children with medium SES did not significantly differ from those with high SES. It should be mentioned that the same effect was found when raw FAS (continuous variable) was included in the model; however, I used the categorical score to better understand the effect of variation in SES. Any indicative better performance of medium-SES than high-SES children should not be taken as problematic, simply because the effect of SES might be interpreted not as a linear gradient but in terms of deprivation or level of learning opportunity; that is, there might be a threshold above which results would differ little.

From table 4.44, it can be also seen that age had no significant effect ($X^2(1)=.273$, $p>0.05$) but that there was a significant negative effect of vocabulary ($X^2(1)=7.615$, $p<0.05$, $B=-.012$). This was expected, because the bilingual children had the lowest vocabulary scores but the highest pragmatic performance in terms of penalising under-informative items; this might be taken as an additional confirmation of their pragmatic advantage.

In respect to the relation between cognitive and pragmatic performance on the group level, the model revealed a main significant effect of STM ($X^2(1)=9.876$, $p=0.002$) but no effect of interaction with group ($p>0.05$). This might not surprise us not only due to the variation between groups, but also to that within groups. The regression results showed that inhibition had a significant main effect ($X^2(1)=30.906$, $p<0.001$) and that there was also a significant effect of interaction between inhibition and group ($X^2(3)=22.667$, $p<0.001$). It should be mentioned that exactly the same results were obtained when the Simon adjusted score resulting from the Cox regression was replaced with the Simon effect⁴.

⁴ I fitted additional model but with removing 'group' as explanatory factor (main effect) and including an interaction effect between STM and group and inhibition and group. The interaction effects with the two EF measure in this model became significant in all the groups.

Table 4.44. GLM Regression results: Predictors for children’s pragmatic performance on under-informative items in the two context conditions.

Predictor	B	SE	X²	DF	P
(Intercept)	.530	.2900	3.337	1	.068
Context condition (no context)	.146	.0526	7.725	1	.005
Language group	.	.	10.806	3	.013
Bilingual-in-English	.559	.3030	3.401	1	.065
Bilingual-in-Arabic	.736	.2995	6.038	1	.014
Arabic	-.175	.3835	.208	1	.648
English	0
SES (FAS)	.	.	25.691	2	.000
SES (Low)	-.622	.1385	20.182	1	.000
SES (Medium)	.099	.0587	2.865	1	.091
SES (High)	0 ^a
Age (month)	-.003	.0063	.273	1	.601
Vocabulary (%)	-.012	.0043	7.615	1	.006
Quantifier/operator Type			76.971	3	.000
Most	-.409	.0708	33.435	1	.000
Some	-.356	.0697	26.160	1	.000
Or	-.616	.0755	66.497	1	.000
And	0
STM	.156	.0495	9.876	1	.002
Inhibition	1.162	.2090	30.906	1	.000
Group*STM	.	.	6.186	3	.103
Bilinguals (English)*STM	-.111	.0676	2.695	1	.101
Bilinguals (Arabic)*STM	-.164	.0664	6.109	1	.013
Arabic children*STM	-.083	.1079	.599	1	.439
English children*STM	0
Group*inhibition	.	.	22.667	3	.000
BCE*inhibition	-.678	.2271	8.916	1	.003
BCA*inhibition	-1.05	.2252	21.605	1	.000
AC*inhibition	-1.07	.3998	7.196	1	.007
EC*inhibition	0

Note: Model reference levels: For group (English children), context (no context condition), quantifier/operator type (‘and’), and SES (High). The dependent variable is the total score of pragmatic responses (penalisation with small or mediums response). The score used for inhibition is the adjusted score from the Cox regression analysis.

To ensure that this model reflected children’s pragmatic performance accurately, I fitted another which had exactly the same exploratory factors but a categorical dependent variable (i.e. ‘large’, ‘medium’, ‘small’ strawberry). Since the 3-point scale

was the dependent variable, I fitted a GLM ordinal logistic regression to predict pragmatic sensitivity to the violation of informativeness. The model reflects exactly the same findings as the Poisson GLM regression in terms of significant effects of the exploratory variables; the results of this model are given in appendix 5.

4.5 A summary of results

This section aims to provide the reader with a brief, useful, and informative summary of the main findings of the whole chapter. I summarised the key outcomes of each section in tables. The comparisons between groups' semantic, pragmatic, and cognitive performance are based on the results of the inferential analyses, reported in sections 4.2, 4.3 and 4.4. Whenever the comparisons revealed a significant statistical difference, I noted this with an abbreviation for the group in question accompanied by a plus sign, indicating better performance of that group on that task—for example, (BCE+) means that bilingual-in-English children performed better, (EC+) that English children performed better, (AA+) that Arabic adults scored higher, etc.

4.5.1 A summary of participants' basic measures

As table 4.45 shows, the child participants were very close in average age, NVIQ, and SES, and the statistical tests revealed no significant differences between the groups. The analyses did, however, show significant differences in vocabulary score between the bilingual-in-Arabic and Arabic children, the bilingual-in-English and English children, and the Arabic and English children. The table also shows that the bilingual children had had more exposure to Arabic than to English, while the Arabic children had clearly had very limited exposure to Standard Arabic, as their dominant dialect was the Colloquial Arabic. The two adult groups had the same average vocabulary score, and the Arabic adults were slightly older than the English.

Table 4.45. A summary of participants' basic measures

Group	Age	Vocabulary		NVIQ	SES		Language exposure	
					FAS	Parents' education		
Bilingual children (30)	5.6	Arabic	English	10.8	5.4	High	Arabic	English
		46	54				60%	47%
Arabic children (30)	5.6	67		10.79	6.	28: high	Colloquial	Standard
						2: medium	81%	19%
English children (26)	5.7	78		11.58	6.27	23: high 3: medium	NA	
Arabic adults (10)	22.2	163 (97%)		NA	NA		NA	
English adults (11)	20.7	163 (97%)		NA	NA		NA	

4.5.2 A summary of semantic results

Table 4.46 gives a summary of the differences between the child and adult groups on the basis of the statistical test outcomes. It can be seen that, in experiment 1 (the give-a-quantifier ask), for the quantifier 'most' there were significant differences between the bilingual-in-English and English and between the English and Arabic children, with the English scoring higher. For 'some', the table shows significant differences between bilingual-in-Arabic and Arabic children (the bilinguals scored higher) and between English and Arabic children (the English scored higher). For the disjunction 'or', significant differences were found between the bilingual children in Arabic and in English and between the Arabic and the English children (with the two English-speaking groups scoring higher than their respective counterparts). The table shows in addition that for the conjunction 'and', there was a significant difference in bilingual performance between Arabic and English children and also between bilingual-in-English and English children, with English children scoring higher in both cases.

Table 4.46. A summary of between-group differences in performance in experiments 1 and 2

Group	Experiment 1				Experiment 2	
	Most	Some	Or	And	Most	Some
Bilingual-in-English v. Arabic	NS	NS	NS	S (EL+)	NA	NA
Arabic v. Bilingual-in-Arabic children	NS	S (BCA+)	S (AC+)	NS	S (BCE+)*	S (BCE+)*
Bilingual-in-English v. English children	S (EC+)	NS	S (EC+)	NS	NS	NS
Arabic v. English children	S (EC+)	S (EC+)	NS	S (EC+)	S (EC+)	S (EC+)
Arabic v. English adults	NS (ceiling)	NS (ceiling)	NS (ceiling)	NS (ceiling)	NS (ceiling)	NS (ceiling)

Note: (BCE+)* means that the bilingual children were tested only in English when compared with Arabic children. Abbreviations in the table: EL (in English language), BCE (bilingual-children-in-English), BCA (bilingual-children-in-Arabic), EC (English children), S (significant), NS (not significant)

On the estimating-magnitude-proportionally task (experiment 2), in which the bilingual children were tested in English only, the analyses revealed significant differences between the bilingual-in-English and Arabic children (with higher scores for the bilinguals) and also between the bilingual-in-English and English children (with better performance for the English children).

The two adult groups exhibited a ceiling effect on the two tasks.

4.5.3 Number task results

A summary of participants' performance on the number tasks is given in table 4.47. It can be noticed that in the how-many task, all the child groups performed almost similarly on set size {10}, but that on set size {14} there were significant differences between the bilingual and English children and between the English and Arabic children, with the English children exhibiting a ceiling effect, as did both adult groups also. In the give-a-number task, the bilingual and the English children as well as the two adult groups exhibited ceiling effects, while only 83% of the Arabic group completed the task correctly; the differences between the Arabic children and the bilingual and English groups were significant. As the table shows, all the child and

adult groups exhibited a ceiling effect on the non-verbal ordinal task, and, generally, all children and adults performed similarly on the estimating-magnitude-numerically task.

Table 4.47. A summary of participants' performance on the number tasks

Group	How-many task		Give-a-number task	Non-verbal ordinal task	Estimating-magnitude-numerically
	Set size (10)	Set size (14)			
Bilingual v. Arabic children	NS	NS	S (BC+)	Ceiling effect	NS
Bilingual v. English children	NS	S (EC+)	NS Ceiling effect	Ceiling effect	NS
Arabic v. English children	NS	S (EC+)	S (EC+)	Ceiling effect	NS
Arabic and English adults	Ceiling effect	Ceiling effect	Ceiling effect	Ceiling effect	NS

Note: Abbreviations in the table: BC (bilingual children), EC (English children), S (significant), NS (not significant)

When exploring the relationship between children's semantic comprehension of the logical quantifiers and operators (in the give-a-quantifier task), the results of GLM regression showed that the children's semantic performance was best predicted by their age and that there was a strong effect of the give-a-quantifier task on the comprehension of 'some'.

4.5.4 A summary of pragmatic results

Below, I summarise the findings of the analyses conducted on the ternary responses (that is, those using a 3-point scale: large, medium, small or acceptance, partial rejection, full rejection). Then, I report the precise outcomes of the different analyses applied only to the results of the two critical conditions in experiments 3 and 4, after transposing the ternary responses into binary responses by combining the two rejection or penalisation responses (since they both indicated some sensitivity to the violation of informativeness, they scored 1 point, while the 'large' response scored 0 since it implied complete insensitivity to the violation of informativeness.)

4.5.4.1 Performance with ternary responses

Table 4.48 provides a summary of comparisons between the groups' performance in experiments 3 and 4. In experiment 3, significant differences in performance were found between the bilingual-in-Arabic and Arabic children on 'most', 'some', and 'or', with the bilinguals penalising the under-informative items at a higher rate. The table also shows significant differences between the bilinguals-in English and the English children on 'some', 'or', and 'and', with higher rates of penalisation among the bilingual children. The comparison between the English and Arabic children showed significant differences between the two groups on 'most' and 'some', with the English performing pragmatically better by more often penalising the under-informative items. Finally, with regard to the adult groups, it is noticeable that Arabic and English groups significantly differed, though the difference might be attributed to the types of responses they used to penalise the under-informative items rather than the rate of penalising: as the results revealed, the Arabic group were more conservative regarding informativeness and sensitive to its violation, penalising it with the 'small' response, while the English group were more tolerant of this violation, more often penalising it only partially, with the 'medium' response.

Table 4.48. A summary of between-group comparisons of pragmatic performance in the under-informative/infelicitous condition between experiments 3 and 4 (ternary responses)

Group	Experiment 3				Experiment 4			
	Most	Some	Or	And	Most	Some	Or	And
Bilingual children (English v. Arabic)	NS	NS	NS	NS	NS	NS	NS	NS
Bilingual-in-Arabic v. Arabic children	S (BCA+)	S (BCA+)	S (BCA+)	NS	S (BCA+)	S (BCA+)	S	NS
Bilingual-in-English v. English children	NS	S (BCE+)	S (BCE+)	S (BCE+)	S (BCE+)	MS (BCE+)	NS	MS (EC+)
Arabic v. English children	S (EC+)	S (EC+)	NS	NS	NS	S (EC+)	S (EC+)	S (AC+)
Arabic v. English adults	S	S	S	S	S	S	S	S

Note: Abbreviations in the table: BCE (bilingual-children-in-English), BCA (bilingual-children-in-Arabic), EC (English children), AC (Arabic children), S (significant), MS (marginally significant), NS (not significant)

As the table shows, in experiment 4 there were significant differences between the bilingual-in-Arabic and Arabic children for ‘most’ and ‘some’, with better pragmatic performance by the bilinguals. The table also reveals a significant difference between the bilingual-in-English and the English children in ‘most’, and a marginal significant difference between the two groups in ‘some’ and ‘and’, with the bilinguals penalising the under-informative items at a higher rate. The comparison between the English and Arabic children showed that the English performed significantly better pragmatically in penalising under-informative ‘some’, ‘or’, and ‘and’. In respect to the adult groups, although the analyses revealed significant differences between groups, both Arabic and English adults penalised the under-informative items at a very similar rate, though the Arabic adults preferred the ‘small’ response and the English group, the ‘medium’.

4.5.4.2 Performance with binary responses

After re-scoring the ternary responses binary values, I conducted comparisons between groups with the re-scored results. Table 4.49 shows the outcomes. It can be seen that in experiment 3, there was a significant difference between the bilingual children’s performance on ‘or’ when tested in Arabic and in English; significant differences were also found between the bilinguals-in-Arabic and the Arabic children for all the quantifiers/operators, with the bilinguals scoring higher. The table also shows significant differences between the bilinguals-in-English and the English children for each of ‘most’ and ‘some’, with the bilinguals performing better. The comparisons between the English and Arabic children demonstrate a significant difference between the groups’ pragmatic performance on ‘most’ and a marginal significant difference on ‘some’, with the English scoring higher. In respect to the adult groups, the results revealed no significant difference except on the conjunction ‘and’, where the Arabic-speakers penalised 100% of the critical items, while the English penalised only 78%.

Table 4.49. A summary of between-group comparisons on pragmatic performance in the under-informative condition in experiments 3 and 4 (binary response)

Group	Experiment 3				Experiment 4			
	Most	Some	Or	And	Most	Some	Or	And
Bilingual children (English v. Arabic)	NS	NS	S (EL+)	NS	NS	NS	NS	NS
Bilingual-in-Arabic v. Arabic children	S (BCA+)	S (BCA+)	S (BCA+)	S (BCA+)	S (BCA+)	S (BCA+)	S (BCA+)	NS
Bilingual-in-English v. English children	S (BCE+)	S (BCE+)	NS	NS	MS (BCE+)	NS	NS	NS
Arabic v. English children	S (EC+)	MS (EC+)	NS	NS	NS	NS	NS	S (EC+)
Arabic v. English adults	NS	NS	NS	S (AA+)	S (AA+)	NS	NS	S (AA+)

Note: Abbreviations in the table: BCE (bilingual-children-in-English), BCA (bilingual-children-in-Arabic), EC (English children), AC (Arabic children), AA (Arabic adults), S (significant), MS (marginally significant), NS (not significant)

In regard to the groups' performance in experiment 4, the results in table 4.45 show a significant difference between the bilingual-in-Arabic and Arabic children in all the target items except 'and', with the bilinguals performing pragmatically better in each case. The outcomes, however, revealed only marginally significant differences between the bilingual-in-English and the English children, with the bilinguals scoring slightly higher. The comparisons between the English and Arabic children revealed significant differences only in their performance on the conjunction 'and', with the English performing better. Finally, comparison between adult groups showed significant differences on 'most' (98% v. 80%) and 'and' (98% v. 71%) but not 'some' or 'or'.

4.5.5 Cognitive task results

Table 4.50 compares the performance of the different groups on the Simon and Corsi blocks tasks. It can be seen that the results revealed no significant differences between the child groups on the Simon task for congruent RT, incongruent RT, or accuracy. The outcomes, however, did show a marginally significant difference in Simon effect between bilingual and English children and a significant difference between Arabic and English children; however, these results should be taken tentatively, as the Cox regression revealed no significant differences. The adult groups' performance on the Simon task did not differ significantly by congruency, Simon effect, or accuracy.

Performance on the STM task (the Corsi blocks) revealed significant differences between bilingual and Arabic children and between English and Arabic children, with the bilingual and English children respectively having longer STM span. There were no significant differences between the two adult groups.

Table 4.50. A summary of participants' performance on the cognitive tasks

Group	The Simon task				The Corsi blocks task
	Congruent RT	Incongruent RT	Simon effect	Accuracy	
Bilingual v. Arabic children	NS	NS	NS	NS	S (BC+)
Bilingual v. English children	NS	NS	MS (BC+)	NS	NS
Arabic v. English children	NS	NS	S (AC+)	NS	S (EC+)
Arabic v. English adults	S	S	NS	NS	NS

Note: Abbreviations in the table: BC (bilingual children), EC (English children), AC (Arabic children), S (significant), MS (marginally significant), NS (not significant)

When exploring the relationship between children's pragmatic performance and their STM and inhibition abilities, the results of GLM regression showed that STM and inhibition had strong effects on the children's pragmatic ability. The model also revealed significant effects of group (superior bilingual performance) SES, vocabulary, and context condition (better performance in the enriched context).

4.5.6 Chapter summary

To conclude, this chapter provides detailed analyses of the basic measures used to control for potential confounds (language proficiency, SES, NVIQ, and language exposure). It also gives descriptive and inferential analyses of the results of the tasks used to assess children's semantic comprehension of logical quantifiers and operators, their numeracy skills, and the potential relation between these two abilities. After this, it explores children's pragmatic ability to detect the violation of informativeness and explores the possible relationship of STM and inhibition in relation to pragmatic performance. The next chapter discusses, in detail, the implications of these results.

Chapter 5

Discussion

5.1 Introduction

This research investigated the relationship between children's comprehension of quantifiers and their numeracy skills (study 1) and the potential effect of bilingualism on children's pragmatic competence via EF ability (study 2). The results showed significant differences in both semantic and pragmatic tasks between Arabic–English bilinguals and Arabic children and English children, but not all the groups differed significantly in cognitive performance on the STM task (Corsi blocks) or the inhibitory control task (Simon task). This section discusses these results in relation to prior empirical studies, answering the following questions: (a) Do bilingual and monolingual children comprehend the quantifiers 'most' and 'some' and the operators 'or' and 'and' in a semantically appropriate (adult-like) way? (b) Does numerical system acquisition promote or (possibly) hinder acquisition of quantifiers, and to what extent? And (c) can any superior pragmatic competence in bilingual children compared to monolinguals be explained in terms of a cognitive advantage? Then, the discussion explores the implications of the children's pragmatic and cognitive performance for theories of implicature processing.

5.2 Study 1: Children's comprehension of quantifiers and operators and the potential effect of numeracy

Study 1 explored children's semantic comprehension of quantifiers and operators and the possible effect of numeracy skills on this comprehension. I discuss first the results for each semantic task (perception (experiment 1) and production (experiment 2)) and why children's performance differs across them. Next, I discuss results for the number tasks, and then for the regression analysis exploring predictors of semantic performance.

5.2.1 Performance on semantic tasks

5.2.1.1 Semantic performance on the comprehension task

The give-a-quantifier task aimed to answer the first research question, regarding semantic comprehension of the quantifiers ‘all’, ‘most’, and ‘some’, and the operators ‘or’ and ‘and’, by asking participants to create sets (quantities) representing the meaning of each quantifier or operator (e.g. for ‘some’, *put some of the apples on the plate*; for ‘or’, *give the puppet a pen or a flower*). Adults’ performance, used as a benchmark for judging the correctness of children’s responses, showed a ceiling effect, with almost exactly the same level of performance across English and Arabic adult groups. Children’s results revealed that, regardless of language background, all the children understood ‘all’, with a ceiling effect (except for one bilingual child in English) and had over 80% accuracy on ‘or’ and ‘and’ but markedly lower scores on ‘most’ and ‘some’. If we order quantifiers by children’s accuracy, we confirm Hanlon’s suggestion that that ‘cognitive development proceeds from simple to more complex forms of knowledge’ (Hanlon, 1987, p. 67). Children performed better on ‘all’ than ‘some’, better on both of them than on ‘most’, and better on ‘and’ than on ‘or’. This is because ‘all’ has a fixed meaning (it always refers to the whole set) and requires only simple cognitive operations for calculation, while ‘most’ requires more complex operations (perhaps mathematical operations, e.g. estimating half the value of the set and then creating sets of a larger number) than either ‘all’ or ‘some’ (since the latter might only require excluding 1 or the whole set). The same might apply to ‘and’ (which requires acting upon both items) compared with ‘or’ (which requires an exclusive reading: either A or B but not both).

The results were largely consistent with previous studies using similar methods. For example, in Barner et al. (2009), English-speaking children (younger than the present sample, with a mean age of 3;8 years old) were found to understand ‘all’ (around 90%) and ‘some’ (around 70%), but to have poorer comprehension of ‘most’ (around 20%). Similarly, Hanlon (1987) found that 4-to-7-year-old English-speaking children showed a ceiling effect on ‘all’ but not ‘some’ (although they understood it well at 88%).

Cognitive complexity might explain the within-group variation in semantic performance on the different quantifiers, but how can we explain the between-group differences, especially the poor performance of Arabic children on ‘most’ and ‘some’? One possibility is that ‘most’ and ‘some’ might cover different semantic realms in Arabic than in English and require more complex cognitive processes for acquisition. With little previous empirical work on Arabic quantifier acquisition, however, it is difficult to evaluate this possibility; but given the Arabic adults’ very similar semantic performance to the English adults, it seems unlikely.

Arabic children’s weak comprehension might alternatively be attributed to frequency of usage: Arabic speakers might not use these quantifiers as often as English speakers. Hanlon (1987) found a strong correlation between frequency of parental quantifier use (with data from Brown, 1973) and the performance of 4-to-7-year-old English children (Hanlon’s own participants). Although this may plausibly explain the Arabic children’s poor semantic performance in the present study, we should be cautious about making such an assumption, since we have no evidence regarding the intensity of exposure to such terms in Arabic.

A third possibility is that Arabic children’s poor semantic performance might be interpreted in terms of mathematical prerequisites (Carey, 1999). If mathematical tools are essential to gain a quantitative conception of weight, for example (which requires understanding of ratios; Carey, 1999), Arabic children’s semantic performance may have been lower because they had not yet developed the mathematical concepts required for a quantitative conception of scale. However, if intensity of exposure or availability of mathematical prerequisites affect performance, why did bilingual children perform worse than English children, since they presumably had similar exposure to the quantifier terms?

Reasons for bilingual children’s lower semantic performance might include parental use, that is, frequency of (and child’s exposure to) the quantifiers, and/or insufficient acquisition of mathematical prerequisites, both as discussed above. Although bilingual children were better than Arabic children at distinguishing ‘most’ from the whole set {6} and from value {1}, these give-a-quantifier task results taken alone, might not clarify why their performance on ‘most’ and ‘some’ was worse than that of

English children. Measuring frequency of use by relevant adults (parents in particular), was not done in this study. Thus, for better understanding of children's semantic comprehension of 'most' and 'some', further tasks (a semantic production task and number tasks) were added to explore reasons for bilingual and Arabic children's lower semantic performance. The results are discussed below.

Another possible reason for low bilingual performance might be overlap between Arabic and English quantifiers. That is, bilinguals are not likely to overlap the meanings of 'some' and 'most' in Arabic and/or English (i.e. interpreting English 'most' as Arabic 'some'), as they showed similar performance in the two languages. But does the same go for 'or' and 'and'? Although bilinguals scored similarly on 'or' across languages, their performance on 'and' significantly differed between Arabic and English (performing better in English). If this indicates overlap (i.e. interpreting 'or' as 'and' in Arabic), one reason might be phonological (due to the similar pronunciation of English [ɔ:] and Arabic [ʔw] 'or' and of [ʔw] and Arabic [wa] 'and', which might make distinguishing them difficult). In contrast, Arabic 'some' [baād] and 'most' [muādam] are clearly dissimilar. Thus, bilingual children may indeed have acquired the conceptual meaning of 'and', which is cognitively less complex than 'or', but overlapping might nevertheless affect their performance in Arabic.

5.2.1.2 Semantic performance on the production task

The production task (estimating-magnitude-proportionally task) was meant to further explore children's comprehension of quantifiers 'most' and 'some' by asking them to describe different proportions using these quantifiers. The task requires the ability to evaluate various ratios, which might also be essential to comprehend set-relational quantifiers 'most' and 'some'.

The two adult groups showed the same ceiling to their performance on the critical items; their results were taken as benchmarks for the children. Bilingual and English children were very competent at expressing different proportions with the most appropriate quantifiers under a forced-choice condition (the production task) (with responses scored 0–1 by adult-likeness). Although initially the Arabic children had no clear quantifier preferences, rescoring with 0–1 values showed slightly better

performance on ‘most’ (mapping 61% of large proportions to it but only 49% of small proportions to ‘some’).

As in the comprehension task, Arabic children’s performance on ‘some’ and ‘most’ in the production task may not be attributable to different meanings of these terms in Arabic, as the Arabic adults’ performance was similar to that of their English peers. Then, the Arabic children’s poor comprehension of the quantifiers ‘most’ and ‘some’ might be best understood in terms of mathematical prerequisites or of the (potential) rarity of these quantifiers in everyday Arabic. If so, however, how might one explain the better performance of not only the Arabic children but all the children on the second semantic task? To answer this question, I conducted further comparative analysis of the children’s performance in the two tasks, which I will discuss below.

5.2.1.3 Perception v. production

Comparison between children’s results for ‘most’ and ‘some’ on the perception task (experiment 1) and production task (experiment 2) revealed that all child groups did significantly better in the production task on ‘most’ than on ‘some’, which saw slight variation between groups. Specifically, bilingual children performed equally on ‘some’ across tasks, while English children had worse performance on ‘some’ in the production than the perception task and Arabic children had better performance on ‘some’ in the production task than the perception task. Given these results, first, what explains the gap in performance on ‘most’ in production versus perception tasks, and second, why did this gap not appear in all groups, especially in English children, with ‘some’?

As for the production–perception gap on ‘most’, let us consider similarities between the present results and those of Tillman and Barner (2015) on 4-to-7-year-old children’s production and comprehension of time duration words (*second, minute, hour*), since that study had several similarities to the present one. That is, sample age, use of both perception and production tasks, and important shared semantic characteristics between their time duration terms and the quantifiers ‘most’, ‘some’: both are abstract terms, ‘structured [in a way] that reflects some knowledge of the relative temporal magnitudes of words’ (Tillman & Barner, 2015; p. 73). Tillman and

Barner found that pre-schoolers performed better on production tasks (forced-choice method; e.g. given *jumped for a minute* vs. *jumped for an hour*, who jumped more?) than a perception task (estimating task; e.g. estimating the duration of some familiar event such as watching a movie). They suggested that pre-schoolers' knowledge of time duration words was limited to knowing their rank ordering (e.g. day>hour>minute), without fully understanding their absolute durations. They found that 6-and-7-year-old children who had been introduced explicitly (in school) to the formal definitions of duration words (e.g. 'one minute equals sixty seconds') were much better able to represent their relative durations than children who had not. This led the authors to propose that, prior to learning absolute definitions of duration words, pre-schoolers understand that they indicate lengths of time, some longer than others; thus, children often know an hour is longer than a minute but not how much longer. Applied to quantifiers, this implies that children's much better performance on the production task (especially on 'most') might indicate that they understand the relevant ordinal lexical scale before learning the exact meaning of these words.

Why did this divergence not appear in all children's performance on 'some', especially for the English children? Although English children performed well on the production task (84%), this was still slightly lower than their performance in the perception task (94%). Thus, we are not talking about poor versus good performance, but about a slight decline in performance that might indicate issue(s) relevant to how these children perceive the meaning of 'some'. To explain further, English children might be inclined to apply 'some' to large proportions (e.g. 11/15), resulting in lower performance in the production task. Furthermore, only one English child and two bilingual children, but ten Arabic children, used 'most' with small proportions (where 'some' is better). Thus, the declining performance of English children, and even the moderate performance of bilingual children, on 'some' might be due not to lack of mathematical prerequisites or inability to associate proportions to their appropriate positions (quantifiers) on a scale, but instead to the children's having fixed mental representations of 'some'—either that 'some' means {2},{3} but not more, or that it means {4},{5} or possibly more, but not less, regardless of set size in a context. Alternatively, bilingual and English children may have achieved advanced understanding/acquisition of 'some', reflected in their similar performance on the perception and production tasks, but still not of 'most', whereas Arabic children's

better production performance might indicate understanding that quantifying words indicate proportions, and that some quantifiers point to larger proportions than others, prior to learning their absolute definitions.

The question then emerges: Why do production task results and age predict children's perception performance? Although regression analysis revealed a significant positive effect of age and an interaction between quantifier perception and production, the effect of performance on production seems to significantly interact only with that on perception of 'some', while the effect of age was only marginally significant. In contrast, the effect of production of 'most' on perception was only marginally significant, while age had a clearly significant effect. Such results have a twofold implication. First, the significant effect of age on perception of 'most' but not 'some' indicates that children knew the meaning of 'some' earlier than 'most'—in line with the common claim that cognitively less complex knowledge (here, 'some') is acquired earlier than more complex knowledge ('most'). Second, the finding that only perception of 'some', not 'most', is predicted by performance on the production task, and that performance on 'most' was significantly better in the production task, means children might acquire the lexical order of these terms prior to understanding them in an adult-like way—similar to Tillman and Barner's (2015) hypothesis that children form categories for abstract terms prior to acquiring their formal meanings. I will discuss abstractness further in section 5.2.2.3.

5.2.2 Performance on number tasks

Children's numeracy skills were assessed to determine whether Arabic children's weak semantic performance on 'most' and 'some' stemmed from weak numeracy skills (possibly due to late exposure). Below, I discuss performance on tasks measuring acquisition of exact and approximate numerical systems.

5.2.2.1 Acquisition of the exact numerical system

Two tasks assessed children's acquisition of the exact numerical system. The how-many task assessed the ability to map sets to their true values, while the give-a-number task measured the ability to produce sets representing the exact meanings of number words. The how-many task results (chapter 4) showed that most child

participants had acquired the counting principle, that is, that they understood that moving (only) one item in a set should be combined with moving one item in the numeral list stored in LTM, and that the last number they count represents the set's numeral value. The English children, like adults, showed a ceiling effect in the task, with 100% correct responses, whereas with set size {10}, 3 bilingual and 2 Arabic children were not able to map the last number they counted with its true value in the numerical list; the number of children who were unsuccessful at this mapping increased slightly at set size {14} (to 7 bilingual and 6 Arabic children).

Three specific findings indicate that incorrect responses might be due not to delay in acquisition of the counting principle or incomplete cognitive ability to map sets to their parallel values in the abstract numerical list but instead to limited experience (training) with number words. First, the types of mistakes children made with set {10} might result from practical issues with counting (e.g. unintentionally jumping one item in the row) rather than cognitive incompetence regarding how counting process works (except with one Arabic child who was clearly unable to memorise the numerical list). Second, wrong responses for set {14} indicate that children had mastered counting only with small numeral values, and could not map larger values to the numeral list. The children's performance revealed that they did understand that a larger set size requires mapping to a larger number in the numerical list (since they gave relatively large numbers such as 16 and 18), but they gave inaccurate values. (However, this interpretation did not apply to one bilingual child and three Arabic children, who gave smaller or equal numbers to numeral values for set {10}.) Third, bilingual children who did not complete the task accurately were slightly younger than their Arabic counterparts, which might mean they had received less training than other children in their group; similarly, the Arabic children may have had late formal exposure to the numerical system and therefore received less training. The analysis also revealed that the Arabic children had had the least pre-schooling, so this might play a role.

The results for the give-a-number task showed that both bilingual and English children had adult-like performance, constructing sets accurately representing the meanings of {1} to {6}. Of the Arabic children, 3 were '5-knowers' and 3 children were '4-knowers' (able to produce sets only up to {5} or {4} correctly, respectively);

these were the same children who couldn't complete the how-many task correctly, except for one 4-knower, who counted {10} and {14} sets successfully but could only generate sets up to {4}. Thus, the question emerges: if these children could count set size {10} correctly, then why they could not produce sets higher than {5} or {4}?

The first possible explanation is that the give-a-number task might require more complex cognitive operations than the how-many task—producing sets might require applying some mathematical processes to given sets, e.g. {6}—while the how-many task only requires connecting each item in the set to a cardinal value and then giving the last number as a representative value for the whole set, and a child might understand both that the last number they count represents the set value and that each item should be mapped to only one numerical value even without mastering principles such as the successor function (Sarnecka & Carey, 2008). Such mathematical abilities may not have been available to the Arabic subset-knowers, although they had evidently acquired some counting principles. Another possible explanation is that the subset-knowers might have developed a good sense of numbers and understood how counting works but not acquired the exact meanings of number words, especially larger ones. Indeed, the results showed that the Arabic subset-knowers produced sets exceeding {1} and {2}, but failed with numbers higher than {4}.

Assuming that one or both of these interpretations are correct, how might one then interpret the worse performance of the bilingual children on the how-many task compared to the give-a-number task, and more importantly, why could only the Arabic children not complete the give-a-number task accurately? Bilingual children's better performance on the give-a-number task than the how-many task might reflect underlying difficulties with large numerals (more than 10), knowledge that might develop through the early years of schooling; the bilingual children who failed with large sets were relatively young (4;1–5;5), supporting this claim. Similarly, the fact that only the Arabic children could not complete the give-a-number task accurately can again be attributed to delayed exposure to the abstract numerical system, due to their having had the least pre-schooling. Of course, only 6 Arabic children (of 30) were (mere) subset-knowers, so these explanations should not be generalised to all the Arabic children.

5.2.2.2 Acquisition of the approximate numerical system

Children's competence with the approximate numerical system was assessed with the estimating-magnitude-numerically task (to assess availability of scalar variability). The results revealed that all the children showed very similar, adult-like behaviour on this task, namely, that there was a systematic change in the distribution of their responses when estimating different magnitudes. This might indicate that children acquire approximate meaning of numerals (estimation) prior to or at least while mastering counting principles. Although this finding is compatible with Huntley-Fenner's (2001) results for 5-to-7-year-old children, it contradicts Le Corre and Carrey's (2007) finding that younger children (3–4 years old) did not show scalar variability in mapping magnitudes larger than {4} to exact numerical values without counting. Le Corre and Carrey concluded that mappings between large numerals and analogue magnitudes are likely not part of the acquisition process. Although the current study is not concerned with whether mapping between numerals and magnitudes is part of acquisition or not, and instead employs this task to see if children have acquired the ability to estimate numerically different magnitudes (scalar variability), since such a cognitive skill might be essential for acquisition of quantifier terms, it might nevertheless be important to understand why the results are inconsistent with previous findings.

One possible reason is the different age ranges between Le Corre and Carrey (2007) and Huntley-Fenner (2001). Negen and Sarnecka (2010) replicated Le Corre and Carrey's estimating task (fast cards) with age-matched children and found scalar variability in pre-schoolers' responses on a number estimation task even before the children had acquired the cardinal principle of counting. The longitudinal nature of Negen and Sarnecka's study, however, might be responsible for such results—that is, instead of assuming that Negen and Sarnecka's participants developed scalar variability within the 20-week period of their study, a dramatic acquisition rate, there might have been an effect of the fortnightly repetition of the task on children's performance, a possibility it seems they did not mention.

Regardless of any age effect or of whether scalar variability is part of numerical system acquisition, my results for the estimating-magnitude-numerically task show that the child participants have acquired scalar variability. However, the robustness of

the empirical findings might be questionable, since each magnitude was estimated only once (this is one of the limitations of the current research, and should be avoided in future work).

5.2.3 The relationship between numbers and quantifiers

The discussion in this section considers the second research question, on the potential effect of mastering numeracy skills for comprehension of quantifiers ‘some’ and ‘most’. Let us first briefly recall children’s performance results on numbers and quantifiers. Starting with quantifiers, the perception task (the give-a-quantifier task) showed that the children, in general, performed better on ‘some’ than ‘most’—the Arabic children had weak comprehension of both ‘some’ (34% correct) and ‘most’ (10%), the bilingual children also had weak comprehension of ‘most’ in both languages (19% in English, 26% in Arabic) but performed better on ‘some’ (around 78% in both languages), and the English children performed better on ‘some’ (95%) than ‘most’ (67%). All groups did significantly better at describing different proportions using ‘most’ in the production (estimating-magnitudes-proportionally) task than the perception task, while on ‘some’ the English and bilingual children had approximately similar performance as on the perception and production tasks, while the Arabic children had significantly better performance on the production task.

Moving to the number tasks, most children exhibited a ceiling effect on the how-many task (of the bilinguals, 90% counting {10} did so, and 77% counting {14}); of the English children, 100% counting both {10} and {14}; and of the Arabic children, 93% counting {10} and 80% on counting {14}), and also on the give-a-number task (the bilingual and English children 100%; the Arabic children 83%); all children exhibited a ceiling effect in the non-verbal ordinal task (100% correct responses, as they were able to point to the set which had more circles), and the three groups showed scalar variability when estimating different magnitudes (although this is not necessarily reliable, since each magnitude was tested once and I relied on the average score). Regression analysis for the number and quantifier tasks showed that performance on the perception task was predicted by age, and once receptive vocabulary score was added to the model, the effect of age disappeared and vocabulary became a strong predictor (but since it was correlated strongly and

positively with children's age, it was excluded from the model). Moreover, the regression analysis revealed no significant main effect of number tasks, but did reveal a significant interaction between the give-a-number task and children's performance on 'some'.

How should we interpret these regression results? Several implications can be drawn. First, the gap between the excellent performance of the majority of children on the number tasks with the incompetent comprehension of 'some' and 'most' in the perception task (especially for 'most' in the bilingual and Arabic groups) might indicate that children generally acquire number words before understanding the adult-like meaning of quantifiers. If so, why do numbers come first? Second, as age was a strong predictor, precisely how do age and (possibly) vocabulary affect comprehension of quantifiers? Third, why does the give-a-number task have a significant effect on the perception of 'some'?

Starting with the first question, various causes might explain children's acquisition of numbers before quantifiers. First, the learning mechanisms of the two systems are extremely different; children learn the meaning of number words and the principles of counting explicitly and systematically in school, pre-school, or from parents. In such an acquisition process, children receive feedback to help them shape the right mental representations of number words and understand from an early stage how counting works. This implies not only more systematic but also more intense exposure. In contrast, children learn quantifiers implicitly by hearing and imitating adults using them in everyday conversation. This not only prevents children from correcting self-established quantifier meanings (since no feedback is given) but also might hinder their shaping mental concepts of such terms, because the same quantifiers are used even among adults to describe different quantities which might confuse children and delay understanding. For instance, in different contexts the same child's parents might use 'some' to refer to three people (say, out of six), a hundred people (e.g. a group within a crowd of 1000 people), or a single (unspecified) person (e.g. *I'm going to meet some guy*).

Second, children might acquire numbers prior to quantifiers for reasons related to the different conceptual (lexical) natures of the two systems. Numbers possess concrete

representations in the real world, and this physical concreteness might be crucial in early acquisition. For instance, the difference between {2} and {3} can easily be conveyed with concrete examples (e.g. *look, here are two horses, while over there are three horses*). In contrast, the meanings of quantifiers are fluid and depend largely on the set size and context. A young child might be able to acquire the meaning of ‘some’ if an adult explained it explicitly (e.g. *some can be used to describe more than one object and less than the whole set*). However, ‘most’ might be harder to acquire, since it is a proportional concept—we might have to explain that one needs first to estimate the half-value of a set in a context, and can then apply the meaning of ‘most’ to any value (proportion) above the half-value. Of course, intuitively, one needs basic mathematical skills to such an estimation, which small children may not have.

The reason age serves as a predictor of children’s performance on the perception task is more easily understood: as children grow, their developing cognitive ability is reflected in their performance. Let us now return to the question of how and why vocabulary predicted children’s performance on the perception task, before being removed from the model. One possibility is that the vocabulary effect is an indirect effect of age, not a pure effect of language proficiency as measured by receptive vocabulary. This is supported by two important findings. First, there was a very significant positive correlation between children’s age and their vocabulary score, and when two covariates are highly correlated, the effect on a dependent variable is likely to appear in only one, strongly predicting (in this case) children’s results on the perception task. Second, the regression analysis conducted with only age and performance on the production task found that age significantly predicted children’s performance on the perception of ‘most’ and marginally for ‘some’; such an effect might disappear with vocabulary added to the model.

Let us assume now that the effect of vocabulary is independent of age, and purely predicts children’s perception of quantifiers. Is there any good explanation for this? Specifically, is it plausible to understand this effect in terms of intensity of exposure to a language? It seems that no conclusive answer can be given here, since although bilingual children performed better than Arabic children despite their significantly limited vocabulary sources, they performed lower than the English children. Such results, although seemingly paradoxical, might be explained in various ways

pertaining to the specific characteristics of the groups. For instance, language proficiency (as measured by receptive vocabulary) might affect children's comprehension of 'most' and 'some', meaning that the English children, who had the highest vocabulary scores, would outperform the bilingual and Arabic children in the give-a-quantifier semantic task. Nevertheless, vocabulary resources are not the only factor playing a role in acquisition; intensity of exposure of quantifiers might also facilitate acquisition, explaining why the bilinguals outperform the Arabic children. That is, bilingual children might shape their comprehension of quantifiers through their daily exposure to such terms in English, and transfer their knowledge of these terms to their equivalents in Arabic.

Let us now answer the last question, why does the give-a-number task have a significant effect on perception of 'some'? The first point to be highlighted here is that only 6/30 of the Arabic children did not show a ceiling effect in the give-a-number task. Thus, finding a relationship between those children's ability to produce sets presenting numerical values (give-a-number task) and their ability to produce sets only representing the quantifier 'some' and not 'most' might be taken as a confirmation of my hypothesis that numbers are acquired first. That is to say, if we assume that the acquisition of quantifiers underpins the acquisition of numbers, then why did the interaction appear on 'some'? Similarly, if we assume that the two systems are acquired in parallel, then we should find at least similar performance in tasks on quantifiers and on numbers, but we did not.

A final possibility is that WM might play a role in children's performance. Although my research included an STM task, it was not reflected in the first study's hypothesis. However, to better illuminate the whole picture I fitted a model with STM as a predictor of children's performance in the perception task. This model showed a strong STM effect on performance, without any significant interaction with either group or quantifier type. This might be because producing sets representing the meanings of 'most' and 'some' requires an interaction between long-term memory (LTM) for word forms (the mental lexicon) and WM. Since the Arabic children had the lowest STM span, this might have affected their semantic performance. Indeed, accurate performance in the give-a-quantifier task requires a child not only to understand the lexical meaning of quantifiers (stored in LTM) but also to evaluate

magnitude in context and apply basic mathematical operations. Thus, the whole process loads heavily on WM (Heim et al., 2016; McMillan et al., 2005).

To sum up, the results of the first study revealed that all the children showed very good semantic comprehension of the logical operators ‘and’ and ‘or’ as well as the quantifier ‘all’; however, performance on ‘most’ and ‘some’ showed that the English children performed significantly better than the bilingual children only on ‘most’, and both groups performed significantly better than the Arabic children. Numeracy skills seem to play only a limited role in the semantic comprehension of quantifiers, and it may be instead that the Arabic children had limited exposure to such quantifiers, hindering them from establishing corresponding mental concepts. It is also possible that the Arabic children lacked the mathematical prerequisites and/or had limited STM ability compared to the other groups, resulting in poorer semantic performance.

The next section discusses the potential effects of bilingualism and EF abilities on children’s pragmatic ability.

5.3 Study 2: The potential effect of bilingualism on pragmatic competence

Based on the empirical evidence of a bilingual EF advantage (e.g. Bialystok, 2011; Blumenfeld & Marian, 2011; Morales, et al., 2013), the purpose of study 2 was to explore the relationship between bilingualism and the pragmatic ability to detect Gricean underinformativeness (i.e. generation of scalar implicatures). In this section, I first discuss the results for pragmatic performance by adults and children; then, I compare the performance of bilingual children with that of Arabic and English children in order to investigate a possible bilingual pragmatic advantage. After this, I discuss adults’ and children’s performance on the two EF tasks (respectively measuring STM and inhibitory ability), followed by a discussion of whether EF abilities predicted children’s pragmatic performance. Finally, I discuss the implications of these results for theories of implicature processing.

5.3.1 Pragmatic performance

Pragmatic performance was measured through two tasks designed to assess the ability to detect violations of Gricean under-informativeness in two different context conditions (enriched context v. no context). In these tasks, participants watched a short scenario (enriched-context task) or heard a statement (no-context task); in each task, there were 4 critical (under-informative) items, and 4 fillers (2 pragmatically and logically correct; 2 pragmatically and logically false). After viewing the scenario or hearing the statement, children had to reward the speaker character with a small, medium, or large strawberry depending on the extent to which the participant detected violation of informativeness. For example, children were asked to reward the character with a large strawberry if he/she described what happened in a correct way, with a medium strawberry if his/her response was not completely correct but was not totally wrong, and with a small strawberry if the response was completely wrong. Control items were used to ensure that the children were able to appropriately reject false items and accept correct ones, and thus that their responses to critical items had not been arbitrary.

In both tasks, the quantifiers ‘most’ and ‘some’ and the disjunction ‘or’ were used in the under-informative condition to reflect children’s pragmatic ability to derive inferences from implicatures generated on a lexical scale (that is, the weaker term on the scale was used when the stronger term should have been applied). For example, in the enriched context, a child would have to evaluate an utterance such as *the girl erased most of the hearts* when the girl erased all of the hearts, or *the dog picked up the banana or the orange* when the dog picked up the banana *and* the orange. Similarly, in the no-context pragmatic task, a child would have to evaluate utterances such as *some elephants have trunks* and *you clean your teeth using toothpaste or a toothbrush*. The conjunction ‘and’ in the first pragmatic task was used to measure pragmatic ability through implicatures generated on an ad hoc scale (e.g. *the girl bought the ring* when the girl bought the hat and the ring), or on an encyclopaedic scale in the no-context condition (e.g. *to clap you need to use your right hand* when the action requires the interaction of both hands, left and right). Adult and child performance in the different experimental conditions is discussed below, with reference to previous empirical results gathered using similar methods.

5.3.1.1 Children v. adults

This section compares the pragmatic performance of adults with that of children on three types of scale (lexical, ad hoc, and encyclopaedic) in two pragmatic tasks (enriched context and no context).

Pragmatic performance on the lexical scale

Let us recall the child and adult results for the critical/under-informative items in the enriched-context task. Analyses were based on ternary responses (the three-point scale), with the adult groups serving as a baseline. When implicatures were generated from a lexical scale (i.e. when the term ‘some’, ‘most’, or ‘or’ was used in a context where a stronger term in the scale should be applied), the performance of the Arabic and English adults differed qualitatively but not quantitatively. That is to say, while both groups penalised under-informative items at a high rate (Arabic adults 100%, English adults more than 90%), Arabic adults mainly used the small strawberry (complete rejection) in their responses, while English adults mainly used the medium strawberry (partial rejection). This clearly indicates that both groups were sensitive to the violation of informativeness, differing only in the degree of penalisation (English adults being more tolerant). If we compare the adults’ performance with the adult results in Katsos and Bishop (2011) (experiment 3), we find that English adults in their study penalised all under-informative items with the medium-sized strawberry as well.

Moving to the adult participants’ performance on the second pragmatic task, in which sensitivity to under-informative (pragmatically infelicitous) items was not only generated from the use of the weaker term in the lexical scale (e.g. ‘some cats have tails’ v. ‘all cats have tails’) but also required participants to draw on their LTM (more precisely their world knowledge) when evaluating the different statements. The adult participants penalised under-informative items at a similarly high rate to the enriched context task (more than 80%). These results partly replicate Noveck (2001) (where adults penalised 70% of under-informative items) and Guasti et al. (2005) (where adults penalised 50% of such items). Qualitatively, there was between-group variation in the type of response; the Arabic adults had a tendency to penalise items

with complete rejection (small strawberry), while the English more often used the medium strawberry to penalise under-informative items.

The child groups' pragmatic responses differed from those of adults in both quantity and quality. In general, in the enriched context condition, bilingual children penalised under-informative items in both languages nearly 57% of the time, the Arabic children nearly 17% and the English children nearly 40%. These results differ from those obtained from 5-to-6-year-old English children in Katsos and Bishop's (2011) ternary judgment task, where English-speaking monolingual children penalised under-informative items at an adult-like rate. In the no-context condition, bilingual children (in both languages) penalised approximately half of the items including the quantifiers 'some' and 'most', but this rate declined with items including the disjunction 'or' (to around 30%); responses were penalised almost equally with small or medium strawberries. The Arabic and English children had similar pragmatic performance on each of the under-informative items, including quantifiers 'some' and 'most' and disjunction 'or'. Arabic children penalised only around 18% of under-informative items (using the small strawberry more than the medium one), and English children, around one-third of such items (using the medium strawberry more than the small one).

I would like to answer two questions here: First, why did not all the English adults penalise the under-informative items, and more importantly, how can we understand the qualitative differences in ratings between Arabic and English adults? Second, why were children in the present study pragmatically less competent than the adults and also than the child participants in Katsos and Bishop (2011) (since experiment 3 in this study replicated their ternary judgment task)?

Before tackling the first question on adult performance, let me emphasise that the present adult sample was small (10 Arabic and 11 English adults), and potentially unreliable. In particular, it is difficult to decide which performances should be treated as outliers. Since all the Arabic adults and most English adults (10 out of 11) penalised the under-informative items, we might view penalisation as the norm. Is this then a concern? Several empirical studies found that adults did not consistently penalise under-informativeness (for enriched context, see e.g. Antoniou, Cummins &

Katsos, 2016; Guasti et al., 2005; without context see Noveck, 2001; Guasti et al., 2005). Therefore, it would not be unusual if adults did not exhibit a ceiling effect in penalising under-informative items, as the Arabic and English adult participants in this study did.

The second interesting finding for the adult groups is the qualitative differences in rating under-informative utterances; that is, the Arabic adults consistently penalised under-informative items with the small strawberry (complete rejection), while the English adults mostly used the medium strawberry (partial rejection). It is unlikely that such behaviour can be explained in terms of differences in EF abilities, since both groups penalised the critical items at a high rate. A more plausible possibility might be (partially) associated with personality traits. Although the current study did not measure personality traits, Feeney and Bonnefon (2013) found a positive relation between participants' self-rated honesty, measured on the Honesty/Integrity/Authenticity Scale (Goldberg et al., 2006) and rate of implicature detection (e.g. 'or' interpreted as 'not both'). Feeney and Bonnefon (2013, p. 7) explained that 'people who perceive themselves as honest are more likely than people who view themselves as less honest to give a scalar term its maximally informative interpretation'. Of course, even if personality traits affect adults' choice of penalty, this does not imply that participants in one group see themselves as more committed to honesty than another group; rather, they might place more importance on precision (and expect other speakers to be more precise), and might more emphatically reject any utterance that does not meet this expectation (in terms of informativeness). Alternatively, such differences might be attributed to cultural norms, where an Arabic participant, for example, might not tolerate the use of the lexical term 'some' if the speaker has no apparent reason to use it (the participant might not see it this way, and might perceive that the speaker did not give less informative information but rather misleading information), while the English hearer might instead interpret even an under-informative utterance as 'saying the truth' and simply view informativeness as a matter of gradient appropriateness or accuracy, which does not necessarily require rejecting an utterance completely. It is worth mention here that the assignment of these different positions to Arabic and English is only an arbitrary hypothesis, and the intention is not to suggest that these profiles might actually fit the two languages themselves.

Let us now compare the children's performance to that of the adults and to that of children in Katsos and Bishop (2011). If we disregard the small size of our adult sample, the children's pragmatic performance was clearly worse than the adults'. This aligns with most previous findings, which show limited pragmatic ability in young children (e.g. Noveck, 2001; Papafragou & Musolino, 2003; Guasti et al., 2005). Those studies gave several explanations for children's insensitivity to violation of informativeness and inability to derive scalar implicature: one, that children might interpret (some or all) scalar expressions logically, and thus, unlike adults, tend to accept under-informative items (Noveck, 2001); or that their limited cognitive resources (e.g. WM) compared to adults might lead to more logical interpretations given the lower cognitive effort involved (Dieussaert et al., 2011). Other scholars have suggested that 'children have less exposure to language than adults, and this limited experience may result in them being less certain about their metalinguistic judgments, and thus accepting under-informative utterances' (Katsos and Bishop, 2011, p. 77).

If the above-mentioned possibilities justify, to some degree, the pragmatic differences between the adult and child groups, how can we understand the relatively low pragmatic performance of the child participants in this study compared to those in Katsos and Bishop (2011), who showed adult-like pragmatic behaviour in penalising 89% of under-informative scalar items (in the ternary judgment task)? The current study did not include items with optimal 'all', which might have made the children in Katsos and Bishop (2011) more aware of the difference between <some, all> and thus more sensitive to the use of 'some' in the under-informative condition since it would then be easier to compare and contrast context between 'some' and 'all'. However, this explanation seems implausible because in Katsos and Bishop's binary judgment task (exactly the same as the ternary judgment task but including optimal 'all') the children could only reject 26% of under-informative items with scalar expressions.

Similarly, it is unlikely that the poor performance of children in the current study resulted from difficulty understanding the rationale of the strawberry rating scale, since they had training in its use and showed clear comprehension in the training trials. It is also unlikely that the between-group variation stems from general mental abilities as measured by the non-verbal matrix test, since the comparison between

groups revealed no significant differences. Third, although the current results are not compatible with those of Katsos and Bishop's (2011) ternary judgment task (in terms of sensitivity to the violation of informativeness), they are consistent with the widely replicated finding that children do not reach adult-like pragmatic level until around 7 years old (e.g. Guasti et al., 2005; Pouscoulous et al., 2007; Hendriks et al., 2009).

Pragmatic performance on other scales

In this section, I discuss pragmatic performance on the ad hoc and encyclopaedic scales. On the ad hoc scale (e.g. 'the dog picked up the orange' when he picked up an orange and a banana), approximately half the Arabic adults' responses indicated complete rejection and the other half, partial rejection (with zero rate of acceptance with the large strawberry). The English adults' results, in contrast, showed only 15% complete rejection (small strawberry), 64% partial rejection (medium strawberry), and the remaining ratio (22%) acceptance (large strawberry), for a total of 78% penalisation. On the encyclopaedic scale (e.g. *to clap, you need to use your right hand*), Arabic adults penalised 98% of the under-informative items (approximately equal ratios using small and medium strawberries), while English adults penalised 70% (two-thirds of which was with the medium strawberry). Thus, each adult group's performance is quite consistent across scales (100% v. 98%, 78% v. 70%), which might indicate consistent levels of sensitivity to violations of informativeness from the context-dependent condition (ad hoc condition) to that where informativeness depends on world knowledge (the encyclopaedic scale).

When comparing adults' performance on the ad hoc and encyclopaedic scales with that on the lexical scale, however, we do find changes in rate and type of penalisation: Arabic adults become more tolerant in their penalisation responses (increasing proportion of partial rejection) on the lexical scale, while English adults accept more under-informative items). One possible explanation for the Arabic adults' performance is that they might be more tolerant when the violation of informativeness results from not mentioning all the items (e.g. the ad hoc scale <orange and apple> v. <apple>; encyclopaedic scale <left and right hands> v. <right hand>), while on the lexical scale using 'some' might be seen as violating the truth condition (i.e. when the speaker can use 'all'). Arabic adults might find it misleading to use 'some' (this

hypothesis, however, lacks empirical evidence), whereas dropping some items does not involve giving misleading information. In contrast, the English adults' results might suggest that it is the semantic truth condition that directs their judgments, rather than pragmatic interpretation—that is, they might see no reason to penalise an utterance that is under-informative as long as it does not explicitly include less informative expressions (such as using 'some' instead of 'all').

Interestingly, in contrast to the adult results, in all child groups the rate of penalisation of under-informative items increased on the ad hoc scale in all groups as compared to performance on the lexical scale. Bilingual children penalised under-informative items on the ad hoc scale around 73% in both languages (more often, partial rejection with the medium strawberry); Arabic children, roughly 50% (more often complete rejection (small strawberry)); and English children, around 68% (more often partial rejection). However, children's rate of penalisation of under-informative items on the encyclopaedic scale decreased slightly among all groups as compared to the ad hoc scale except the English children, who had similar penalisation rates on encyclopaedic and ad hoc scales. In more detail, bilingual children penalised approximately 45% of under-informative encyclopaedic items (in both languages), and the English children, slightly more than half (55%) (with similar percentages of small and medium strawberries in both groups), while the Arabic children penalised only 30% of these items (using the small strawberry (complete rejection) at triple the rate of the medium one (partial rejection)).

The decrease in the rate of penalisation of under-informative items in the encyclopaedic scale compared to the ad hoc scale in the child groups was expected, because, unlike the ad hoc scale, sensitivity to under-informativeness in the encyclopaedic condition requires children to draw heavily on both STM and LTM. That is, to evaluate an utterance such as 'to clap you need to use your right hand', they need to keep the utterance active in STM while assessing its validity by connecting it to information stored in LTM, more precisely to world knowledge.

Children's performance: Horn v. ad hoc and encyclopaedic scale

Having discussed differences in children's pragmatic performance across scales, let us now consider why these differences exist. As the results revealed, children's rate of penalisation of under-informative items on the ad hoc scale was significantly greater than on the lexical scale (approximately 30% more in all child groups). This might indicate that the children (4–6 years old) had acquired the pragmatic ability to detect the violation of informativeness, but that on the lexical scale the process of detecting this violation requires lexical scale activation in order to contrast and compare the term used in a context (e.g. 'some') with the stronger (and more informative) term on the scale (e.g. 'all') that should be applied in that context. Such a step requires better *episodic buffer ability* (ability to connect STM with LTM; Baddeley & Hitch, 1974; Baddeley, 2000) to allow connection of the utterance just heard (while keeping it active in STM) with the lexical scale stored in LTM (more precisely in semantic memory). Thus, even if the child was quite competent with the semantic meaning of 'some' (as in the English child group), she would not be able to detect the violation of informativeness if she did not activate the lexical scale and compare the stronger with the weaker term. In contrast, on the ad hoc scale, such a step is not required to detect violation, since a child only needs to compare the verbal utterance with the visual context; this lower complexity facilitates implicature derivation. Since the bilingual and English children have very similar STM ability, the mechanism enabling the bilingual children to derive more scalar implicatures might be episodic buffer ability.

However, would the decline in pragmatic responses on the encyclopaedic scale refute this hypothesis, especially since the English children performed slightly better than the bilinguals? This seems unlikely, for several reasons. First, on the ad hoc scale (where episodic buffer seems to play only a limited role), the two groups had very similar quantitative performance (bilingual 73% v. English 68%), as also on the encyclopaedic scale (bilingual 45% v. English 55%); significant differences that might complicate the episodic buffer hypothesis are absent. Second, if we compare items that required world knowledge (the no-context condition) and differed in including a scalar term ('some', 'most', 'or') or not (encyclopaedic scale) we find that bilingual performance was almost the same on both scales in both languages, while Arabic children's performance was marginally significantly different across scales, ($p=0.054$) and that of English children, strongly significant ($p=0.003$), with better

pragmatic performance in utterances that did not include scalar terms. Recalling that the English children outperformed the other groups in semantic comprehension of quantifier expressions and exhibited a ceiling effect in all numerical tasks, their lower ability to detect violation of informativeness in a statement such as ‘some elephants have trunks’ is likely due neither to semantic incompetence with ‘some’ nor to incompetent numeracy. Further, it is unlikely that all child groups lack the pragmatic ability to detect violation of informativeness, since they were able to penalise under-informative statements that did not include scalar terms at a significantly higher rate than those that did. Thus, we might attribute children’s inability to penalise under-informative scalar expressions to limited ability to activate the lexical scale and compare the scalar expression to a stronger term in the scale; and this ability is based on WM, more precisely on episodic buffer ability.

The effect of context

The analysis in chapter 4 compared children’s performance within each group in the two context conditions; here I discuss only the critical (under-informative) conditions. As expected, generally speaking, children performed better when visual context was available, except the bilinguals in Arabic, where the difference was only marginal ($p=0.093$). This aligns with Guasti et al. (2005) and Feeney and Scrafton (2004). It is unlikely that the marginal difference in bilinguals’ performance in Arabic was due to having completed the same task a week earlier in English, since it has been found that the effect of training disappeared when children were tested on the same task a week later (Guasti et al., 2005). It is also unlikely that it was due to better language proficiency in Arabic, as the bilinguals’ vocabulary was better in English.

Performance on control items

Control items were included to ensure that children, in particular, are able to reject pragmatically and logically inappropriate items and accept pragmatically and logically appropriate ones. This was essential to ensure that children’s choice of acceptance of critical (under-informative) items was not arbitrary, for example, alternating acceptance and rejection, or choosing spontaneously without considering the information presented. Another purpose was to ensure that the children understand the rationale of the 3-point scale. Below I discuss the participants’ performance in a

pragmatically and logically appropriate condition in experiments 3 and 4 and then on items that were pragmatically and logically inappropriate.

Optimal v. felicitous

Starting with the optimal condition in the enriched-context task, the two adult groups accepted all the items (with large strawberry) in conditions including optimal use of the quantifiers ‘some’ and ‘most’ and the ad hoc scale with ‘and’ (except one English adult, who penalised one ad hoc item with the medium strawberry). Although the child groups did not exhibit such a ceiling effect, they accepted optimal items for ‘some’, ‘most’ and ‘and’ at more than 85% (except the bilingual children, who accepted optimal items for ‘most’ at 63% in English and 70% in Arabic). Thus, the children could seemingly differentiate between optimal and under-informative items, generally accepting the optimal ones, with the large strawberry.

For the disjunction ‘or’, Arabic adults only accepted half the items, and English adults only 40% of the items. Similar performance was found in Arabic and English children, while the bilinguals’ rate of acceptance was clearly lower (20% in English and 16% in Arabic). This might be because all optimal ‘or’ items started with the doubtful phrase *I’m not sure I saw it well; the crocodile ate the apple or the banana;* (when the crocodile ate the banana). The rationale for inserting this phrase was to create an optimal context for using the disjunction ‘or’; that is, there was no reason to use ‘or’ if the character addressing the children was sure which object. Thus, based on the assumption that ‘it may well be that the exclusive interpretation of “or” is often based on considerations of plausibility rather than implicature’ (Geurts, 2010, p. 60), I tried to experimentally justify the use of ‘or’ by adding the uncertainty phrase.

However, including this phrase may have been a methodological mistake, since it could have affected participants’ perception of the utterance in two main ways. First, it might prevent participants from reading the utterance exclusively in order to derive the implicature (‘A or B but not both’) due to the lack of a ‘Competence Assumption’—‘the assumption that the speaker knows what he is talking about’ (Geurts, 2010, p. 29)—which would strengthen the implicature. We might assume that the context (in this case, the uncertainty phrase) does not support the Competence

Assumption to generate a strong implicature (i.e. an exclusive interpretation), but it seems here that the participants were not even able to derive a weaker implicature. This is likely because inserting the doubtful phrase led participants to penalise the utterance for violating the Maxim of Quality (without even paying much attention to the validity of the content coming after the phrase). Second, if the child understood that speaker (the fictional character) had not seen what happened, his utterance might be taken to be bringing up possibilities rather than intentionally implicating some specific case, meaning that neither strong nor weak ‘exclusivity implicatures’ might be derived (Geurts, 2010, p. 61). The high rate of acceptance of optimal items (Arabic adults 100%; English adults accepted 82% and partially penalised 14%) in the no-context condition (e.g. *when babies are born they are either a girl or a boy*) suggests that the adults were able to accept items where exclusive readings seemed pragmatically and logically appropriate. One English adult who rejected such an item justify his response by referring to a birth defect, ‘atypical genitalia’, making it unclear whether a newborn is a girl or a boy; however, other adults who accepted the statement likely considered such cases to be rare, and such advanced world knowledge (of birth defects) might not be available to children.

Performance on the parallel condition in the no-context task (felicitous items) revealed that adults only exhibited a ceiling effect for some quantifiers and operators (Arabic adults accepted all felicitous items for ‘or’ and ‘and’ but not ‘most’ or ‘some’, while English adults accepted all items for ‘some’ and ‘most’ but not for the operators) but that overall acceptance rate was high in both groups (Arabic, more than 90%; English, more than 77%). Similarly, the child groups showed fairly high rates of acceptance of felicitous items (60% and above), with English and Arabic children generally accepting more felicitous items than bilingual children; and all the children recorded high rates (more than 90%) of acceptance of optimal items on the encyclopaedic scale, which require the interaction of two entities (e.g. *people hear sound from their left and right ears*). Generally speaking, children’s high acceptance rates for logically and pragmatically appropriate items in the pragmatic tasks seem to indicate that they could differentiate between the optimal and under-informative items.

False v. bizarre

First, in the enriched-context task, adults rejected all false items for ‘most’, ‘some’, and ‘or’; however, on the ad hoc scale (e.g. *the dog picked up the orange and the banana*, when he picked up the orange and the apple), Arabic adults penalised 95% of false items (15% of those with partial rejection) and the English adults, all the false items (45% with partial rejection). Similarly, child groups penalised false ‘some’, ‘most’ and ‘or’ at high rates (more than 85%, mostly complete rejection) but without reaching the ceiling as adults did. Of ad hoc false items (mismatching visual content), child groups penalised more than 75% (with high rates of complete rejection). Such high rejection rates strongly suggest that the children did not choose to accept or reject under-informative items arbitrarily but rather were able to evaluate the truth-condition value and reward the fictional character accordingly. Furthermore, these results indicate that the children understood how the three-point scale worked.

As for bizarre items in the no-context task, the two adult groups penalised all of them (with complete rejection) for quantifiers ‘some’ and ‘most’ and operators ‘or’ and ‘and’. Similarly, child groups penalised bizarre items at more than 90% for all scales; although Arabic children showed a somewhat lower percentage (around 84%) in the encyclopaedic condition, this difference was non-significant. These high bizarre item rejection rates were expected, since these items clearly made no sense, violating the truth condition (e.g. *some chairs can sing*), and had been employed to ensure that children were able to reject statements that were logically and pragmatically false. This was especially useful to ensure that the Arabic children, who often used the large strawberry, were not biased toward the use of the large strawberry as such, irrespective of the item’s linguistic or pragmatic content, but were selecting a reward based on evaluation of the given statements.

5.3.1.2 Is there a bilingual pragmatic advantage?

One of the main objectives of this study was to explore if bilingual children pragmatically outperform their monolingual (in this case, English) and the bidialectal (Arabic) peers, and to suggest possible explanations for any such outperformance. I will begin investigating this question with reference to the tasks using the three-point

scale, and then to comparisons built on the general rate of penalisation of critical (under-informative) items in the two pragmatic tasks (scored 0–1).

Regarding the ternary responses, with implicatures generated by using the weaker term in the lexical scale where the stronger should be applied, for the quantifier ‘most’ in the enriched-context condition the bilingual children performed significantly better than the Arabic children and numerically but non-significantly better than the English children, and their performance across the two languages was exactly the same. In partial contrast, in the no-context condition, bilinguals significantly outperformed both Arabic and English children (again with very similar performance in both languages). These results remain approximately the same when were re-analysed in binary terms, except that the difference between bilingual and English children becomes marginal in the no-context condition.

The results for the quantifier ‘some’ in the enriched context again saw bilingual children significantly outperforming both Arabic and English children in penalising under-informative items; similar results were found in the no-context condition, except that the difference between the bilingual and English children was only marginally significant. Bilinguals’ results were again very similar across languages. The results did not change when re-analysed using binary responses, except that the marginally significant difference between bilingual and the English children in the no-context condition disappeared (became statistically insignificant, although bilinguals still performed numerically better).

On the disjunction ‘or’, bilingual children significantly outperformed the other groups on both conditions (enriched context and no context), again with no significant difference across languages. However, re-analysing data using the 0–1 score showed that although the difference between bilingual and Arabic children remained statistically the same, that between bilingual and English children in the enriched context became marginal, and in the no-context condition, totally disappeared, with the bilinguals still numerically scoring higher.

Performance on the ad hoc scale (in the enriched context) showed a significant difference only between bilingual and English children, with bilinguals performing

better (and non-significantly better than Arabic children), again with no significant difference in their performance across languages. Re-analysed as binary responses, however, data showed no significant difference between bilingual and English children but a significant difference between bilingual and Arabic children.

On the encyclopaedic scale, ternary responses revealed no significant differences between bilingual and Arabic children or between bilingual performance across languages, but a marginally significant difference between bilingual and English children, with the English penalising more items. Re-analysed as binary responses, that marginal difference disappeared.

Generally speaking, the lexical scale shows a significant difference in pragmatic performance between bilingual and Arabic children on all words assessed ('most', 'some', 'or') regardless of context or analysis (ternary or binary), except that the difference between bilingual and English children remains significant only for under-informative 'most' in the no-context condition and for under-informative 'some' and 'or' in the enriched-context condition (with bilinguals still performing numerically better in all other conditions). The significant difference between bilingual and Arabic children also disappeared on the ad hoc and encyclopaedic scales (with better performance among bilinguals), and the significant difference between bilingual and English children disappeared on those scales under binary re-analysis (with numerically better performance for bilinguals on the ad hoc scale and for the English on the encyclopaedic scale).

In general, apart from performance on the ad hoc and encyclopaedic scales, these findings might lead us to speak about a bilingual pragmatic advantage, which we see appearing for all the scalar expressions, especially when we recall that the bilinguals had significantly lower vocabulary than the other groups. Bilingual children repeatedly outperformed Arabic children statistically and English children statistically (in three of six cases, under both types of analysis) or numerically. This outperformance is compatible with previous studies including children matching the current sample (Siegal et al., 2007, 2009, 2010). However, the question emerges: Why did this pragmatic advantage not remain on the ad hoc and encyclopaedic scales,

and only appeared when the weaker term in a scale was used in the utterance (whether supported with visual context or based on world knowledge)?

What is clear is that bilingual pragmatic outperformance on ‘most’ and ‘some’ cannot be explained by higher semantic competence, since English children outperformed the bilinguals semantically. One possible explanation is that metalinguistic advantage (from exposure to multiple languages) makes bilingual children more sensitive to under-informative utterances than their monolingual counterparts. Alternatively, or in addition, a bilingual cognitive advantage may contribute to their pragmatic outperformance—this possibility is discussed further in section 5.3.4 below.

5.3.2 Cognitive performance

Two cognitive tasks were used to assess children’s visuo-spatial STM (the Corsi blocks task) and inhibition ability (the Simon task). On the STM task, the adult results revealed very similar STM span between the two groups (non-significantly longer in the English group, 7.27 v. 6.8 average STM span). In contrast, bilingual children had slightly (non-significantly) longer STM span than English children (4.13 v. 3.96 average STM span), with the Arabic children significantly behind at 3.2). Regression analysis revealed that children’s performance on the Corsi blocks task was best predicted by age. These results do not support a bilingual STM advantage, not only because no significant effect of bilingualism appeared but might also be because the English monolinguals had significantly better STM span compared to the Arabic children (of course this in itself would not preclude a bilingual advantage; there might be other factors at play, e.g. an English advantage or an Arabic disadvantage). These results are partially consistent with Blom et al. (2014), who found a bilingual WM advantage only in 6-year-old and not in 5-year-old children; however, we still need to understand why the Arabic children had lower STM.

One possible explanation for the Arabic children’s lower STM span might involve the structure of the Arabic language. I am not referring to the morphological complexity effect that has been shown to have a negative effect on children’s verbal STM by studies in which children performed worse in recalling inflected than uninflected words (Cohen-Mimran, Adwan-Mansour, & Sapir, 2013). Instead, I am talking about

how the morphological and syntactic features of Arabic might involve less cognitive load than corresponding features in, for example, English. In Arabic, speakers do not use the verb *to be*, second pronouns (*he, she, they*) are expressed with only one letter (a suffix), and most words consisting of only two syllables and all words based on three consonant roots (e.g. *katab* ‘he wrote’, *kaatib* ‘writer’, *kitaab* ‘book’, *yaktub* ‘he writes’ are all different forms, but all depend on the three consonants *ktb* for their root meaning). Such features might result in shorter average utterance in conversational exchange, reducing the load on a child’s WM when either producing or perceiving/interpreting an utterance. However, this remains only a hypothesis, and lacks empirical support.

On the other hand, performance on the Simon task revealed that although English adults completed both congruent and incongruent trials significantly faster than Arabic adults, there were no significant differences in the Simon effect, and both groups completed all the trials correctly. However, global RT does not necessarily reflect better inhibition ability, and the small sample here makes reliable conclusions difficult.

Moving to the child groups’ performance on the Simon task, the analyses comparing RT to complete two congruency conditions revealed that the Arabic children were faster but that there were no significant differences in the congruent condition, the incongruent condition. The analyses of the Simon effect showed a marginally significant difference between the bilinguals and the English children, and between the English and Arabic children (the English had the highest Simon effect, indicating lower inhibition ability), but there was no significant difference between the bilingual and the Arabic children. These results, showing no exclusive bilingual inhibitory advantage, are consistent with some previous studies (e.g. Morton & Harper, 2007; Coderre & van Heuven, 2014) but not others that did find such a bilingual inhibitory advantage (e.g. Antoniou et al., 2014; Poarch & van Hell, 2012). Thus, although there is some evidence for a general inhibitory advantage in bilingual children, I suggest another explanation that might be relevant to Arabic–English bilinguals in particular (which might be also applied to the two adult groups if we take into account that the two groups were exposed to another European language). Coderre and van Heuven (2014) explored the effect of script similarities on bilinguals’ inhibition ability (as

measured by the Simon task). The study included three groups of adult bilinguals whose native languages' scripts had varying levels of similarity with English (German, Polish, and Arabic) and an English monolingual group. Although the study did not find significant differences between groups in Simon interference effects, the Arabic-English bilinguals showed the longest global RTs of all four groups and on this basis, Coderre and van Heuven (2014) suggest that script similarity may affect bilingual EF abilities. We should, however, be tentative with these results, since the Arabic-English bilinguals were slightly older than the other groups, and since, possibly more importantly, there is inconsistent evidence of script similarity effects on cognitive abilities (e.g. Morton & Harper, 2007).

The lower Simon effect in the Arabic children might be understood either in terms of their being less accurate in completing the task or of their having an inhibitory advantage (similar to that of the bilinguals) over the English children; in this case, we need to understand why. The first possible explanation is that bidialectalism has a similar effect to bilingualism, especially on inhibition ability, due to the need to switch between the two languages. However, we should be cautious with this explanation, since the Arabic children had only very limited exposure to Standard Arabic. Another possibility is that there might be an indirect effect of certain characteristic of Arabic (either morphological features or the right-to-left writing and reading characteristic). A third possibility is that the orthographic system of Arabic might affect inhibition ability. That is, in Arabic writing short vowels should not be written, but they still need to be pronounced (Fedda & Oweini, 2012), and so there may be a continuous need for an Arabic child, who is just starting to learn how to write words, to inhibit such vowels in order to master writing skills, and this may contribute to better inhibitory skills.

In the current study, when children's performance on the Simon task was re-analysed using Cox regression, the differences between the bilinguals and the other groups previously found to be non-significant were again non-significant. Cox regression showed a significant effect of congruency, with all the children performing better in the congruent condition—as expected, since this condition does not require inhibition of irrelevant information and thus has less cognitive cost than the incongruent

condition. The regression also revealed a strong effect of SES and significant positive effects of age and non-verbal IQ (NVIQ). How can we interpret such results?

As the regression analysis revealed, age and NVIQ predicted children's performance on the Simon task and correlated negatively with RT. These are not strange findings, since in principle, inhibition ability should be affected by age and general mental abilities (such as NVIQ). The negative effect of SES, meaning family wealth (as measured by the FAS questionnaire), on group performance would not necessarily need to be reflected in EF abilities such as inhibition, since other factors (e.g. age and NVIQ) might have more significant roles. In addition, the FAS questionnaire showed that bilinguals, in general, had lower FAS scores than English or Arabic children; however, they also had the smallest Simon effect, so we must be cautious when interpreting the negative effect of SES. Previous studies have found an effect of SES independent from that of bilingualism (De Abreu et al., 2013; Calvo & Bialystok, 2014). Note here that we are not talking about extreme SES differences; the majority of participating children in the present study had at least one parent with a high level of education (of course, this alone is not a reliable measure of SES). In addition, FAS results should be interpreted cautiously due to the multicultural sample in this research, as well as possible issues with the accuracy of the questionnaire as an indicator of real SES, which is intrinsically multifactorial and hard to measure.

5.3.3 The relationship between pragmatic and cognitive advantage

The relationship between children's pragmatic performance and their EF abilities was explored through regression analysis, as discussed in chapter 4. Regression test results revealed significant effects of language group, context condition, SES (as measured by the FAS questionnaire), and a strong negative effect of vocabulary, as well as significant main effects of STM and inhibition.

Let us start with language group and SES. The model showed a marginal significant positive effect of bilinguals-in-English and a significant effect of bilinguals-in Arabic over English children. These results are in line with Siegal et al. (2007, 2009, 2010). The model also revealed a significant effect of SES: children with low SES performed

significantly worse than those with high and medium SES (the two latter did not significantly differ). Low-SES children's poorer pragmatic performance might be attributable, for example, to resources available to higher-SES children (such as number of books and technological tools available at home), which might positively affect their general cognitive abilities (e.g. Calvo & Bialystok, 2014) and be indirectly reflected in their pragmatic performance.

Let us now discuss the effects of vocabulary and context. The bilingual children, who pragmatically outperformed the other groups, nevertheless had lower vocabulary scores (a negative effect). Although there was a strong correlation between age and vocabulary score, age did not significantly correlate with children's pragmatic performance. This was accompanied by a significant effect of context, with better performance in the enriched-context condition, where under-informative items were supported with visual context. Reasons for this might include the following. First, the interaction between verbal and visual STM might facilitate WM development (Giard & Peronnet, 1999; Prabhakaran et al., 2000). Alternatively, judging whether a scalar expression appropriately describes some of a small number of objects in context may have been easier than evaluating whether a scalar expression appropriately describes some of a countless number of exemplars of a real-world category existing in one's encyclopaedic knowledge (e.g. *some elephants have trunks*). Evaluating the latter kind of item not only requires activation of the lexical scale <all, most, some> but also accessing episodic and semantic LTM to search if the term (e.g. *some*) is appropriately applied. In addition, evaluation in the no-context condition might require a participant to apply mathematical operations while keeping the verbal utterance active in STM (McMillan et al., 2005; Heim et al., 2016), making computation more effortful. This possibility might be supported by the model results, which showed a significant main effect of STM, is in line with previous findings (e.g. Zajenkowski & Szymanik, 2013). Thus, pragmatic interpretation might require participants to hold an utterance active in STM while comparing it with either visual context or information stored in LTM.

The negative estimates and significant effect of STM should not be taken to indicate negative relations between pragmatic performance and STM, but this should be interpreted with relation to how the model was fitted (i.e. the reference level). That is

to say, since the bilinguals-in-Arabic significantly outperformed the English children while the bilingual-in-English and English groups did not differ in STM, the similar STM levels met with a significant difference in pragmatic performance among the bilinguals. It is worth mentioning here that once language group was removed from the model, the effect of STM becomes significant for all groups.

The regression model also revealed a strong significant positive effect of inhibition. This may be explained by, first, the possibility that participants inhibit the logical interpretation of a scalar expression ('some and possibly all'), which would mean that the better their inhibition ability the more pragmatic their interpretation. Indeed, the model revealed a direct negative relationship between inhibition and rate of pragmatic interpretation (that is, decrease in RT in the Simon task was met with increase in rate of rejection).

Below I discuss the implications of these results for implicature processing theories.

5.3.4 Implications for implicature processing theories

The relationship between cognitive advantage and pragmatic performance can be interpreted with reference to two theories of implicature processing: the Default (Levinson, 2000) and Relevance-theoretic (Sperber & Wilson, 1986/1995) accounts. The Default view suggests that pragmatic interpretations (GCIs) are automatic, and only require cognitive effort if the context requires them to be cancelled (PCIs). The Relevance account, in contrast, indicates that an implicature is only drawn if relevant within the context, and that construing a logical interpretation is less effortful than construing a pragmatic interpretation (Sperber & Wilson, 1986/1995; Sperber & Wilson, 2002). To explain how these two accounts might respectively explain the between-group variation in the present results, we will look at semantic and pragmatic performance with the operators 'and' (implicature generated from an ad hoc scale) and 'or' (implicature generated from a lexical scale), simply because these were the only terms with which the Arabic children were semantically competent, removing the need to worry about difference in semantic ability. Similarly, the discussion here will consider the enriched-context task only, to avoid any individual differences

resulting from varying levels of world knowledge (which is difficult to control) rather than pragmatic ability.

First, I interpret an example of pragmatic performance from the perspective of Default theory. Let us consider performance with the operator 'or'. In a scenario where two items are collected (C and D) and the character states that the animation showed the collection of *C or D*, the generated implicature would presumably be 'either C or D but not both'. If the participants tolerated or rejected as opposed to accepting this response, it was taken to mean that they were pragmatically sensitive and able to derive the implicature. Conversely, if they accepted it, we might suspect that they had made the logical (inclusive) interpretation 'either C or D, and (logically) perhaps both'. According to Default theory, this process is fast and automatic, as it relies partially on the lexical meanings of words. If Arabic and English children had been found to understand the literal meaning of 'or' as *either A or B but not both* more reliably than bilinguals, the expectation would have been that they would also be better at deriving implicatures.

This was not the case. Bilinguals, who were semantically less advanced, were still pragmatically more competent in deriving implicatures with the operator 'or'. Explaining such results in light of Default theory would preclude the argument that bilinguals had better inhibitory control and therefore were better able to withdraw the generalised implicature and derive the particularised implicature. This is because in such a lexical scale <or, and>, only generalised implicatures could be generated in experiment 3, and the context did not require their cancellation. In addition, although there was a strong effect of inhibition on pragmatic performance, there was no exclusive bilingual inhibitory advantage to justify their superior pragmatic performance. Similarly, the argument that bilinguals' higher STM ability allows them to more easily derive implicatures is difficult to accept not only in terms of the Default hypothesis but also because no exclusive bilingual advantage was found. The results revealed that only with GCIs did bilingual outperformance remain consistent (i.e. even when data were re-analysed as binary values), but differences on the ad hoc scale (PCI) were not always significant. Thus, a Default account might not explain superior pragmatic performance with GCI.

The Relevance view states that deriving an inference is an effortful cognitive process that requires contextual support. This seems promising to explain the extreme difference between bilingual and other (Arabic and English) children's pragmatic performance. Let us also use performance with the operator 'or' as an exploratory example. The results revealed that both the Arabic and the English children were very competent at understanding the literal meaning of 'or' in the give-a-quantifier task, and scored significantly higher than the bilingual children. However, comparing pragmatic performance (in the enriched-context task), it was found that although bilinguals frequently provided logical answers, they were still able to derive pragmatic implicatures significantly better than the other two groups. Thus, bilinguals, who showed the lowest Simon effect in the inhibitory control task and had longer STM spans than the Arabic and English children (but which only reached statistical significance with the Arabic children), seemed more able to engage in the costly cognitive process of deriving implicatures. Similarly, we might interpret the better pragmatic performance of English over Arabic children to be due to better STM abilities, or possibly suggest that, at least with scalar implicatures, STM ability seems to play a significant role.

For further support of this Relevance interpretation, we may consider Dieussaert et al. (2011), who suggest that children generate more logical interpretations than pragmatic inferences (as in Noveck, 2001) because '[s]ince children have a more limited cognitive capacity than adults, they will prefer the interpretation with the lowest cognitive cost—that is, the logical interpretation' (Dieussaert et al., 2011; p. 2355). Applying this explanation to the present results suggests that because the bilinguals had developed stronger cognitive capacity than the Arabic or English children, they were better able to produce the more effortful pragmatically enriched interpretations. In addition, the strong effect of inhibition and STM on children's pragmatic performance in general (i.e. regardless of language profile) might give further support. To explain this more precisely, let us recall Tomlinson et al.'s (2013) two-phase processing hypothesis. They suggest that implicature processing occurs in two steps, the first requiring decoding/retrieval of context-independent meaning, after which enrichment is implemented (or the utterance is re-interpreted with enrichment added). If we assume that this was the case in the current study, then it might be suggested that the bilinguals, who developed better inhibitory control ability, were

able to inhibit the literal interpretation that resulted in the first step and make pragmatically enriched interpretations, whereas other children (who had lower inhibition ability) did not inhibit the literal meaning and thus made more logical (rather than pragmatic) interpretations. The effect of STM can be interpreted in either or both of two ways. The first possibility is that while processing scalar implicature, children might need a good STM span to keep the utterance active while evaluating not only the truth value of a statement but also the magnitude shown in a scene and whether it accurately represents the meaning of a quantifier. However, this possibility could not explain the significant improvement in children's performance on the ad hoc scale, leading us to assume that some additional processing step might occur requiring a level of cognitive ability (possibly) only available to bilinguals. That is to say, the additional step here requires the activation of the lexical scale (stored in LTM) to compare the strongest scalar term (e.g. 'all') with the scalar expression in an utterance (e.g. 'some'). This ability to connect LTM with STM in an effective way (episodic buffer ability) enables children to derive implicatures, which might explain the effect of STM on pragmatic performance.

Further support might be given to the Relevance-theoretic interpretation if we consider the effect of context condition on children's pragmatic performance. The lower performance of all groups when there was no context (e.g. *some elephants have trunks*) might indicate that the process does not occur by default mechanism but requires the hearer to apply all available kinds of relevant information (in this case search through world knowledge) to make pragmatic interpretations. If we recall the results of Zhao et al. (2015), they found that processing statements that contradict with LTM generates more cognitive effort (as measured by higher level of neural activation), while Nieuwland et al. (2010) suggest that scalar implicature processing depends not only on contextual assumptions but also on neuropsychological factors, a claim that comes in line with the Relevance-theoretic account, which posits the presence of cognitive effort on the computation process.

Another possibility that might at least partially explain bilinguals' superior pragmatic performance is their having more efficient brain areas associated with EF. That is, bilinguals' exposure to two languages might modify their brain morphology (in e.g. grey matter) and areas associated with EF (i.e. prefrontal cortex), resulting in more

effective language processing. Thus, bilinguals might rely more on their frontal lobes, which are responsible for EF, and therefore become pragmatically more skilled than monolinguals. Direct neural evidence for this hypothesis comes from brain-imaging studies comparing the active brain areas of bilingual and monolingual adults processing language and showing greater activity in bilinguals' EF networks (Hernandez & Meschyan, 2006).

Despite the strong association between children's pragmatic ability and their performance on ToM assessments (Goetz, 2003) and the strong association between ToM and inhibition capacity (Carlson & Moses, 2001), ToM might not serve directly to explain bilinguals' strong pragmatic performance in this study. This is because deriving implicatures should not rely on the participant's ability to understand the speaker's mental state but rather on attentional control and WM capacities. Neither of the pragmatic tasks involved in this study required participants to apply mind-reading ability. Thus, the role of ToM remains for further study.

The final point to be discussed briefly in light of the two implicature processing theories is children's pragmatic performance on ad hoc scale. Let us clarify the situation with this scale using an example. Participants in all groups watched an animated scenario in which two items, A and B, were acted upon and only item B was described by an animated character. If the participants tolerated or rejected as opposed to accepting this response, it was taken to mean that they were pragmatically sensitive and able to derive the implicature that only one item was designated, not both. The results revealed that all children performed significantly better on the ad hoc scale than on the lexical scale. It seems difficult to interpret this performance in light of Default theory, however, since we are here talking only about a PCI; and if my hypothesis of the potential effect of the episodic buffer is plausible in explaining the change in children's performance as compared to on the lexical scale (as discussed in section 5.3.1.1.2.1), then Relevance theory seems to be a more appropriate explanation. That is to say, the cognitive demand for deriving an implicature with the ad hoc scale requires only comparing verbal utterances with visual context (i.e. has lower cognitive cost than scalar implicature), while the lexical scale might require additional effort (i.e. higher cognitive cost) than (e.g.) ad hoc implicatures, activating the scale in LTM, estimating the proportion/magnitude in a context, and checking if it

represents the meaning of a quantifier in an utterance). If this assumption is true, then the result is in line with Relevance theory

To sum up the discussion in this section, bilingual children outperformed their Arabic and English counterparts pragmatically despite having considerably less advanced vocabulary. Such results support the Relevance over the Default hypothesis, since the process of implicature derivation seems to be cognitively effortful.

5.4 Summary of discussion

This chapter discussed the main findings of the two studies presented in chapter 4. The first assessed children's semantic comprehension of quantifiers and the possible role of numeracy skills in semantic performance. The results of the perception (give-a-quantifier) semantic task revealed that English children outperformed bilingual children significantly on 'most' and numerically on 'some', while the Arabic children were clearly incompetent at understanding the meaning of either 'some' or 'most', which might be attributed to either limited exposure to such terms, lack of the mathematical basics essential for acquisition, or having limited WM ability. All the children showed very good comprehension of the conjunction 'and' and the disjunction 'or', which might be attributed to their being cognitively less complex than the quantifiers. The production task results showed a dramatic positive change in all children's performance on 'most', while only the Arabic children's performance on 'some' improved. The discussion suggests that children acquired the ordinal meaning (on the lexical scale) of quantifiers before acquiring their adult-like meaning.

The number task results showed that all the children had well-developed exact and approximate numerical systems, with a significant age effect on performance. The discussion suggests that such results might indicate that children acquire numbers earlier than quantifiers.

The second study, which explored the potential effect of bilingualism on pragmatic ability, revealed that bilingual children outperformed the other groups (but without reaching statistical significance, for all quantifiers/operators). No exclusive bilingual advantage was found in the two EF tasks. However, STM and inhibition had strong

effects on children's pragmatic performance. The discussion suggests that such results imply that scalar implicature requires cognitive cost, an assumption that comes in line with Relevance theory.

The next chapter summaries the major findings of the thesis and its contribution to the field, and briefly discusses its limitations and suggests some directions for future work.

Chapter 6

Conclusion

6.1 Introduction

In this chapter, I briefly summarise the major findings of the thesis, its contribution to the field, and some of the limitations resulting from methodological issues, and conclude by suggesting possible directions for future work.

6.2 Summary and main findings

The goal of the thesis was twofold. First, it aimed to explore how children with different language backgrounds comprehend quantifiers and the effect of numeracy skills on their semantic comprehension of quantifiers (study 1). Second, it aimed to investigate the effect of bilingualism on children's pragmatic ability (study 2).

In chapter 2, the literature review explained the motivation for each study. With respect to the first study, the previous literature on the acquisition of quantifiers and numbers revealed a slightly unclear picture: although a positive correlation had been identified between the acquisition of numbers and that of quantifiers, it was unclear whether the former underpins the latter or vice versa, I tried to address this issue by bringing together theories of abstract concept representation and theories on number representation, and based on this synthesis, hypothesised that numbers are acquired first. Since semantic comprehension of quantifiers might affect children's pragmatic performance (i.e. their sensitivity to the violation of informativeness on a lexical scale, investigated in study 2), it was important to assess how children understand logical quantifiers as well as operators; and due to the prior evidence of a relation between numbers and quantifiers, it was important to assess children's numeracy skills; study 1 does so. All this makes the current thesis the first (to the author's knowledge) to explain theoretically and investigate empirically how the quantificational and numerical systems are associated and to explicitly propose the role of the approximate numerical system.

With regard to the second study investigating the relationship between bilingualism and pragmatic competence, although there was strong previous evidence of bilingualism's positive effect on EF abilities (e.g. Bialystok, 2011; Blumenfeld & Marian, 2011), this effect had not emerged in all previous work; I discussed possible reasons for this with reference to the *coordination account* of Bialystok et al. (2006) and Bialystok (2011). In addition, although there is some evidence for a bilingual pragmatic advantage (e.g. Siegal et al., 2001, 2009), this evidence is still scant and did not emerge in all previous work (e.g. not in Antoniou et al. 2014). More importantly, it is still unclear how such a pragmatic advantage, if indeed it exists, can be interpreted in terms of implicature processing theories. Thus, the second study was conducted to explore this phenomenon further and explain its findings with reference to theories on implicature processing.

In chapter 3, empirical methods to robustly explore the research questions were presented in detail. Control measures included assessment of children's receptive vocabulary, general mental ability (NVIQ), SES (parents' education and the FAS questionnaire), and also bilinguals' and bidialectals' language exposure. The methods used to investigate children's semantic comprehension of the quantifiers and operators were a perception task (experiment 1) and a production task (experiment 2). To explore the possible association between quantifiers and numeracy skills, the study adopted four measures assessing children's acquisition of approximate and exact numerical systems. With respect to the potential effect of bilingualism on pragmatic abilities, two ternary judgment tasks were adopted to assess children's pragmatic ability to detect the violation of informativeness in two conditions: enriched context v. no context. The indirect effect of bilingualism on pragmatic performance was measured by assessing two EF abilities: STM and inhibition.

In chapter 4, the analyses started by exploring the background measures; the results revealed that the groups were matched in age, SES, and NVIQ, but that the bilinguals had a significantly lower vocabulary score than the other two groups. The results of the language questionnaire indicated that the Arabic children could be functionally described as monolinguals due to their limited exposure to Standard Arabic (lower than 20%), and that the bilinguals generally had greater exposure to Arabic than to English. However, the fact that the bilinguals' vocabulary score was lower in Arabic

than in English might suggest that the quantity of exposure might not be an accurate indicator of language proficiency; rather, the quality of input (e.g. the language used in educational materials, reading activities at home, or on TV and on the computer) might play a more important role.

Moving on to children's performance on the semantic perception task, the results revealed that all the children had nearly the same ceiling effect in the comprehension of the operators 'or' and 'and', whereas performance on the quantifiers 'most' and 'some' revealed that the English-speaking children had the highest performance of all the groups overall but did not significantly differ from the bilinguals on 'some'; those two groups significantly outperformed the Arabic children. In the production task, all groups' performance on 'most' was significantly better than their performance on 'most' in the perception task, but on 'some', only the Arabic children showed a significant improvement.

The results of the numeracy tasks showed that all the children had very good numeracy skills on the whole, with the majority exhibiting a ceiling effect in the numeracy tasks but only the bilingual and English children exhibiting a ceiling effect in the give-a-number task, while 6/30 of the Arabic children did not show a ceiling effect. On the how-many task, all the English children exhibited a ceiling effect, as did 90% of the bilinguals and 93% of the Arabic children. Performance on the approximate numerical system task showed that all the children were able to manipulate their responses in correspondence to a given magnitude. Regression analysis showed a strong effect of age on children's performance in the perception task, while the give-a-number task only became significant in interaction with quantifier performance for 'some'.

Performance on the two pragmatic tasks showed that the bilinguals outperformed the other two groups (although the difference did not reach statistical significance for all tested terms or experimental conditions). The bilingual advantage in general was much clearer using the lexical scale than the ad hoc or encyclopaedic scales.

Performance on EF tasks, in contrast, did not reveal a bilingual cognitive advantage. Although the bilinguals had the longest STM span, it did not significantly differ from

that of the English children; both these groups had significantly longer STM than the Arabic children. The inhibition task also revealed no exclusive bilingual advantage, although the bilinguals had the lowest reaction time on the Simon task, possibly indicating better inhibition ability. The regression analysis showed strong effects of STM and inhibition on children's pragmatic performance.

In chapter 5, the discussion attempted to explain the between-group variation in the comprehension of quantifiers, suggesting that it is due either to limited exposure to the quantifiers or to lack of mathematical prerequisites (because of limited training or limited length of pre-schooling). Another possibility that arose is that it could be WM abilities that affected the Arabic children's performance, and indeed the additional regression analysis I conducted showed a strong effect of STM on children's semantic comprehension of the quantifiers, making this study the first to attain such a finding. In addition, I suggested that the children's better performance on the production task could be due to their having acquired the ordinal position of the quantifiers before acquiring their literal (adult-like) meaning. The higher performance by all the groups on numeracy tasks than tasks measuring comprehension of quantifiers was attributed to different mechanisms in the acquisition process—for example, children are taught number words and their meanings explicitly, while they learn the meaning of quantifiers implicitly. Another mechanism facilitating acquisition might be the concrete nature of numbers compared to abstract quantifiers. The results of the give-a-number task (where only the Arabic children did not show a ceiling effect) on comprehension of 'some' seem to indicate that knowledge of numbers does play at least a partial role in children's acquisition of quantifiers; the absence of such an effect for 'most' might be either because children have acquired neither its absolute meaning nor its ordinal position or because it requires the application of relatively advanced mathematical operations that are not yet available to children at this age.

Pragmatically, the bilingual children performed better than the other groups, although the difference did not reach statistical significance for all of the four experimental quantifiers and operators (i.e. 'most', 'some', 'or', 'and'), especially on the ad hoc and encyclopaedic scales as compared to the lexical scale. I suggested that the clear improvement in the Arabic and English children's pragmatic performance on the ad hoc and encyclopaedic scales indicates that those children were indeed sensitive to the

violation of informativeness, but that, in contrast, sensitivity to informativeness on the lexical scale requires the ability to activate that scale in LTM in order to compare the stronger term in the scale (e.g. ‘all’) with the weaker scalar expression used in the utterance (e.g. ‘some’). If so, this means that the bilingual children had better ability to connect STM with LTM, that is, better episodic buffer ability (Baddeley & Hitch, 1974; Baddeley, 2000) than the others. Another possibility is that children who were pragmatically more competent had developed better inhibition ability, allowing them to inhibit the logical interpretations that might arise first (e.g. ‘some and possibly all’). This two-phase computation is in line with the results of Tomlinson et al. (2013). The results of regression analysis revealed strong main effects of STM and inhibition on children’s pragmatic performance, which might support both these possibilities.

Considering this posited association between pragmatic performance and EF abilities, this study favours Relevance theory to account for its results rather than the Default explanation, since the process of deriving implicatures seems to involve some cognitive effort (as seen in the relationship between EF and the rate of pragmatic responses). This thesis might also provide evidence that the cognitive effort involved in pragmatic processing is not limited to STM but also applies to inhibition; such results are compatible with those of Zajenkowski and Szymanik (2013), who found strong effects of STM and attention on scalar implicature processing.

6.3 Contribution and possible implications of the current research

The results of the current research can make a vital contribution to several areas of research.

6.3.1 Contribution of Study 1: The relationship between numbers and quantifiers

The first contribution is that this research theoretically supports the relationship between quantifiers and numbers, linking theories from different domains to give a clearer, novel understanding of the nature of this association. I mapped theories on the representation of abstract and number words, and proposed roles of approximate numerical system and scalar variability in acquiring the meaning of quantifier terms. I

tried to employ previous empirical findings (e.g. Barner et al., 2009) to show that it seems more plausible that children would acquire numbers first. Furthermore, I explained why and how this would be the situation, and clarified the mechanisms that might enable such early acquisition of numbers.

Another contribution of this thesis can be seen in the empirical evidence it gathers on how children acquire quantifiers. To the best of my knowledge, there is no previous work exploring both production and perception of quantifiers; this study provides evidence that the acquisition process starts by acquiring the ordinal scale, before moving to acquiring the more explicit and advanced knowledge needed to understand the quantifiers in an adult-like way. In addition, the high level of accuracy in numeracy tasks compared with tasks using ‘most’ and ‘some’ indicates that number acquisition comes earlier, which might support the claim I made in chapter 2 when mapping the mental representation of abstract concepts to that of number words.

Furthermore, the study is the first to talk about the role of not only the exact but also the approximate numerical system, and to discuss which might underpin which and why. It is also the first to empirically test the role of STM in children’s acquisition of quantifiers. Moreover, it seems that no previous work has been conducted to explore how Arabic-speaking children might acquire such terms. Given the different possible explanations regarding their poor performance on the quantifiers (in the perception task specifically), further research should be conducted to explore the possibilities that this study has generated not only in relation to Arabic children but for our understanding of the roles of numeracy and WM more widely.

6.3.2 Contribution of Study 2: The relationship between bilingualism, EF, and pragmatic competence

One of the main contributions of the second study is its effort to map theoretical pragmatic accounts of implicature processing to behavioral and neural data on scalar implicature in order to better understand the nature of the possible cognitive effort involved. I believe that this contribution is theoretically and practically relevant because, although several researchers have attempted to explore implicature processing in light of the Default and Relevance/Context-Dependent hypotheses of

computation, the nature of the cognitive effort involved is still unclear. By empirically testing children's pragmatic abilities alongside some of their EF abilities and exploring how the latter might affect the former, this study gives some insights on the nature of the cognitive effort and possibly some evidence that this effort is not limited to only one EF component, such as WM.

Although there is sizeable literature on bilingualism's effects on EF abilities, the evidence regarding pragmatic competence is still scant and inconsistent; this research might add some evidence on the bilingual pragmatic advantage. Apart from this, it seems that no previous work on children's pragmatic performance has tested bilinguals in both their languages, making the current research the first to give evidence that bilingual pragmatic advantage is neither exclusive to one language nor limited by language proficiency level.

A final contribution of this study is that it gives some evidence on the roles of STM and inhibition in pragmatic performance, which can contribute to new insights on the nature of the cognitive effort involved in pragmatic processing.

6.4 Limitations

The current project employed multiple experimental measures and aimed to control for several potential confounds, with the hope of better understanding the within- and between-group variation found. The study, however, makes no claim that the results are without limitations, and this section addresses them briefly.

The first limitation might be related to the receptive vocabulary measure, more precisely the fact that an Arabic translated version of the BPVS test originally developed for English was used. Although I was well aware of this limitation, this instrument remained the most suitable choice, since there was no available similar test in Colloquial Arabic and the age of the children did not allow the use of a Standard Arabic test; even if the children were older (and thus having a higher level of exposure to Standard Arabic), testing their vocabulary in Standard Arabic might not accurately reflect their proficiency level. Thus, we should be cautious when interpreting the results of this test. Another issue with using a translated version is the

difficulty of controlling for frequency; that is, vocabulary that might be used more frequently in English (and thus introduced earlier in the test) might differ in frequency of usage in Arabic, leading to inaccuracy in the results.

The second limitation pertains to the first study, on quantifiers and numeracy. It can be seen in the estimating-magnitude-numerically task, more precisely in the choice to test each magnitude in only one single trial. Although we expect children in this age group to have a good sense of numbers and to be able to manipulate numerical values to match a set size viewed on a computer screen, for more reliable results it would have been better if the I had tested each magnitude at least twice; this shortcoming was entirely a result of my limited experience in designing tasks at the time the study was constructed, and could be addressed easily in future.

The third limitation pertains to the second study, on children's pragmatic ability. Here, I used the uncertainty phrase 'I'm not sure I saw it well' in the optimal experimental condition for the disjunction 'or' in the enriched-context pragmatic task (experiment 3). As already discussed in chapter 5, this choice seems to preclude any exclusive reading of 'or' (see Guerts, 2010, for a further discussion). However, this limitation might not have a severe effect on the results, since it occurred only in filler items and not in the critical items (those employed to assess pragmatic ability to derive inferences). Of course, however, it should be addressed in future research.

The final limitation is related to the choice to test the bilingual children first in English and then in Arabic, that is, to the fact that the tests were not counterbalanced. It might have been more reliable if I had tested half of the sample first in Arabic and then in English and then reversed this order for the other half. In Arabic but not in English, I found no significant differences in bilingual performance on under-informative items across the two context conditions. Although previous work suggests that the effect of training disappears within a week, if the present sample had been counterbalanced, it would be clearer if the absence of any significant difference resulted from a training effect due to repeating the task, even in a different language, or if it was due to something more.

6.5 Future work

Although the current research sheds light on the relationship between quantifiers and numbers, and the potential effects of children's bilingualism and EF abilities on their pragmatic abilities, more work should be conducted in future to help us better understand these relationships. More precisely, since this research proposes that numbers seem to be acquired first, and precisely explains the mechanisms that enable children to acquire them first, this claim should be explored further in future. Also, the potential effect of early exposure to numbers might be explored in future research. Hopefully, if an adequate corpus covering various dialects of Colloquial Arabic can be developed, it will become possible to test the effect of Arabic children's intensity of exposure to quantifiers on their semantic comprehension of quantifiers. In addition, since this study showed a bilingual pragmatic advantage, it will be important for us to explore whether such an advantage is limited to scalar implicatures by testing other types of implicatures. Furthermore, researchers should remember that the absence of explicit evidence of a bilingual cognitive advantage does not mean it does not exist; rather, it might only emerge when tasks require high cognitive effort, a possibility we might test using a dual-modality task (as in Bialystok et al., 2006; Bialystok, 2011).

References

- Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta psychologica, 128*(3), 466–478.
- Abutalebi, J., Della Rosa, P. A., Gonzaga, A. K. C., Keim, R., Costa, A., & Perani, D. (2013). The role of the left putamen in multilingual language production. *Brain and language, 125*(3), 307–315.
- Abutalebi, J., Della Rosa, P.A., Green, D.W., Hernandez, M., Scifo, P., Keim, R., Cappa, S.F. & Costa, A. (2011). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cerebral Cortex, bhr287*.
- Al-Akeel, A. I. (1998). The acquisition of Arabic language comprehension by Saudi children (Doctoral dissertation, University of Newcastle).
- Alduais, A. M., Shoeib, R. M., Al Hammadi, F. S., Al Malki, K. H., & Alenezi, F. H. (2012). Measuring Pragmatic Language in Children with Developmental Dysphasia: Comparing Results of Arabic Versions of TOPL-2 and CELF-4 (PP and ORS Subtests). *International Journal of Linguistics, 4*(2), 475–494.
- Aliyu, A. A., Bello, M. U., Kasim, R., & Martin, D. (2014). Positivist and Non-Positivist Paradigm in Social Science Research: Conflicting Paradigms or Perfect Partners?. *Journal of Management and Sustainability, 4*(3), 79–95.
- Alkhamra, R. A., & Al-Jazi, A. B. (2016). Validity and reliability of the Arabic Token Test for children. *International Journal of Language & Communication Disorders, 51*(2), 183–191.
- Anders, Y., Rossbach, H. G., Weinert, S., Ebert, S., Kuger, S., Lehl, S., & von Maurice, J. (2012). Home and preschool learning environments and their relations to the development of early numeracy skills. *Early Childhood Research Quarterly, 27*(2), 231–244.

- Anderson, P. J., & Reidy, N. (2012). Assessing executive function in preschoolers. *Neuropsychology review*, 22(4), 345–360.
- Antoniou, K., & Katsos, N. (2016). The cognitive foundations of pragmatic development. In F. Salfner & U. Sauerland (eds.), *Proceedings of “Trends in Experimental Pragmatics”* (pp. 10–17). Berlin: German Research Foundation.
- Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, 99, 78–95.
- Antoniou, K., Grohmann, K., Kambanaros, M., & Katsos, N. (2016). The effect of childhood bilingualism and multilingualism on executive control. *Cognition*, 149, 18–30.
- Antoniou, K., Katsos, N., Grohmann, K., & Kambanaros, M. (2014). Is bilingualism similar to bilingualism? An investigation into children’s vocabulary and executive control skills. In W. Orman & M., J. Valteau (eds.), *Proceedings of the 38th annual Boston University Conference on Language Development* (pp. 12–24). Somerville, MA: Cascadilla Press.
- Azhar, M., Nadeem, S., Naz, F., Perveen, F., & Sameen, A. (2013). Impact of parental education and socio-economic status on academic achievements of university students. *European Journal of Psychological Research*, 1(1), 1–9.
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417–423.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, 8, 47–89.
- Ball, L. J., Lucas, E. J. & Phillips, P. (2005). Eye-movements and reasoning: Evidence for relevance effects and rationalisation processes in deontic

- selection tasks. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. Alpha, NJ: Sheridan Printing.
- Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, *56(A)*, 1053–1077.
- Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive psychology*, *58(2)*, 195–219.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism Development Disorder*, *31*, 5–17.
- Barrouillet, P. (2011). Dual-process theories of reasoning: The test of development. *Developmental Review*, *31*, 151–179.
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser (eds.). *Symbols, embodiment, and meaning* (pp. 245–283). Oxford: Oxford University Press.
- Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22(4)*, 577–660.
- Becker, J. (1989). Preschoolers' use of number words to denote one-to-one correspondence. *Child Development*, 1147–1157.
- Behrman, J. R. (1996). The impact of health and nutrition on education. *The World Bank Research Observer*, *11(1)*, 23–37.

- Bezuidenhout, A., & Cutting, J. C. (2002). Literal meaning, minimal propositions, and pragmatic processing. *Journal of Pragmatics*, *34*, 433–456.
- Bialystok, E. (2011). Coordination of executive functions in monolingual and bilingual children. *Journal of Experimental Child Psychology*, *110*(3), 461–468.
- Bialystok, E., & Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual adults. *Bilingualism: Language and Cognition*, *15*(02), 397–401.
- Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental science*, *7*(3), 325–339.
- Bialystok, E., Craik, F. I., Green, D. W., & Gollan, T. H. (2009). Bilingual minds. *Psychological Science in the Public Interest*, *10*(3), 89–129.
- Bialystok, E., Craik, F. I., & Ruocco, A. C. (2006). Dual-modality monitoring in a classification task: The effects of bilingualism and ageing. *The Quarterly Journal of Experimental Psychology*, *59*(11), 1968–1983.
- Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology and aging*, *19*(2), 290–303.
- Bialystok, E., Martin, M. M., & Viswanathan, M. (2005). Bilingualism across the lifespan: The rise and fall of inhibitory control. *International Journal of Bilingualism*, *9*(1), 103–119.
- Bick, A. S., Goelman, G., & Frost, R. (2011). Hebrew brain vs. English brain: language modulates the way it is processed. *Journal of cognitive neuroscience*, *23*(9), 2280–2290.

- Blom, E., Küntay, A. C., Messer, M., Verhagen, J., & Leseman, P. (2014). The benefits of being bilingual: Working memory in bilingual Turkish–Dutch children. *Journal of experimental child psychology*, *128*, 105–119.
- Blumenfeld, H. K., & Marian, V. (2011). Bilingualism influences inhibitory control in auditory comprehension. *Cognition*, *118*(2), 245–257.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*, 437–457.
- Boudelaa, S., Pulvermüller, F., Hauk, O., Shtyrov, Y., & Marslen-Wilson, W. (2010). Arabic morphology in the neural language system. *Journal of Cognitive Neuroscience*, *22*(5), 998–1010.
- Boyce, W., Torsheim, T., Currie, C., & Zambon, A. (2006). The family affluence scale as a measure of national wealth: validation of an adolescent self-report measure. *Social indicators research*, *78*(3), 473–487.
- Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual review of psychology*, *53*(1), 371–399.
- Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science*, *12*(3), 238–243.
- Breheny, R. E. T., Ferguson, H. J., & Katsos, N. (2012). Investigating the time-course of accessing conversational implicatures during incremental sentence interpretation. *Language and Cognitive Processes*, *28*(4), 443–467.
- Breheny, R. E. T., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. *Cognition*, *126*(3), 423–440.

- Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? *Cognition*, *100*(3), 434–463.
- Brese, F., & Mirazchiyski, P. (2010). Measuring students' family background in large-scale education studies. In *4th IEA International Research Conference in Gothenburg, Sweden, Gothenburg, Sweden*. Retrieved from http://www.iea-irc.org/fileadmin/IRC_2010_papers/TIMSS_PIRLS/Brese_Mirazchiyski.pdf.
- Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Brunetti, R., Del Gatto, C., & Delogu, F. (2014). eCorsi: implementation and testing of the Corsi block-tapping task for digital tablets. *Frontiers in psychology*, *5*, 939. [DOI: 10.3389/fpsyg.2014.00939].
- Burgaleta, M., Sanjuán, A., Ventura-Campos, N., Sebastian-Galles, N., & Ávila, C. (2016). Bilingualism at the core of the brain. Structural differences between bilinguals and monolinguals revealed by subcortical shape analysis. *NeuroImage*, *125*, 437–445.
- Butterworth, B., & Walsh, V. (2011). Neural basis of mathematical cognition. *Current biology*, *21*(16), R618–R621.
- Calvo, A., & Bialystok, E. (2014). Independent effects of bilingualism and socioeconomic status on language ability and executive functioning. *Cognition*, *130*(3), 278–288.
- Carey, S. (1999). Knowledge acquisition: Enrichment or conceptual change. In E. Margolis & S. Laurence (eds.) *Concepts: core readings*, (pp. 459–487). Hong Kong: Massachusetts Institute of Technology.
- Carey, S. (2004). Bootstrapping and the origin of concepts. *Daedalus*, *133*(1), 59–68.

- Carlson, S. M., & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental science*, *11*(2), 282–298.
- Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child development*, *72*(4), 1032–1053.
- Casasanto, D. (2010). Space for thinking. In V. Evans & P. Chilton (eds.), *Language, Cognition and Space: State of the art and new directions*, (pp. 453–478). London: Equinox Publishing.
- Chierchia, G., Crain, S., Guasti, M.T., Gualmini, A. & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In *Proceedings of the 25th Boston University conference on language development* (pp. 157–168). Somerville, MA: Cascadilla Press.
- Coderre, E. L., & van Heuven, W. J. (2014). The effect of script similarity on executive control in bilinguals. *Frontiers in psychology*, *5*, 1070. [DOI: 10.3389/fpsyg.2014.01070].
- Cohen-Mimran, R., Adwan-Mansour, J., & Sapir, S. (2013). The effect of morphological complexity on verbal working memory: results from Arabic speaking children. *Journal of psycholinguistic research*, *42*(3), 239–253.
- Coolican, H. (2004). *Research Methods and Statistics in Psychology* (4th ed.) London: Hodder Arnold.
- Cordes, S., & Gelman, R. (2005). The young numerical mind: When does it count?. In J. Campbell (ed.), *Handbook of mathematical cognition*, 127–142.
- Corsi, P. M. (1973). *Human memory and the medial temporal region of the brain* (Doctoral dissertation). Retrieved from eScholarship@McGill. (93903).
- Costa, A., Hernandez, M., & Sebastian-Galles, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, *106*(1), 59–86.

- Craik, F. I., & Bialystok, E. (2006). Cognition through the lifespan: mechanisms of change. *Trends in cognitive sciences*, *10*(3), 131–138.
- Currie, C. E., Elton, R. A., Todd, J., & Platt, S. (1997). Indicators of socioeconomic status for adolescents: the WHO Health Behaviour in School-aged Children Survey. *Health education research*, *12*(3), 385–397.
- Darby, D. (2011). *PathSpan*. Retrieved from <https://itunes.apple.com/us/app/pathspan/id521564885?mt=8>.
- De Abreu, P. M. E., Cruz-Santos, A., Tourinho, C. J., Martin, R., & Bialystok, E. (2012). Bilingualism enriches the poor: Enhanced cognitive control in low-income minority children. *Psychological science*, *23*(11), 1364–1371.
- De Cat, C., Gusnanto, A., & Serratrice, L. (under-review). Identifying a threshold for the executive function advantage in bilingual children. Retrieved from <http://ceciledecat.blogspot.co.uk/p/in-preparation.html>
- De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, *17*(5), 428–433.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load. *Experimental Psychology*, *54*(2), 128–133.
- Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, *7*(4), 145–147.
- Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, *284*(5416), 970–974.
- Dickinson, E. R., & Adelson, J. L. (2014). Exploring the Limitations of Measures of Students' Socioeconomic Status (SES). *Practical Assessment, Research & Evaluation*, *19*(1). Retrieved from <http://pareonline.net/getvn.asp?v=19&n=1>.

- Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2352–2367.
- Dunn, L. M., & Dunn, D. M. (2009). *British picture vocabulary scale* (3rd ed.). London: GL Assessment Limited.
- Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test*. Circle Pines, MN: American Guidance Service.
- Dunn, L. M., & Dunn, L. M. (2000). *Peabody Test di Vocabolario Recettivo*. Torino, Italy: Omega (Adapted by Stella, G., Pizzoli, C., & Tressoldi, P.E.).
- Dunn, L. M., L. M. Dunn, C. Whetton & J. Burley. (1997). *British Picture Vocabulary Scale* (2nd ed.). London: nferNelson.
- Ellefsen, M. R., Shapiro, L. R., & Chater, N. (2006). Asymmetrical switch costs in children. *Cognitive Development*, *21*(2), 108–130.
- Esposito, A. G., Baker-Ward, L., & Mueller, S. T. (2013). Interference suppression vs. response inhibition: An explanation for the absence of a bilingual advantage in preschoolers' Stroop task performance. *Cognitive development*, *28*(4), 354–363.
- Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology*, *87*, 223–240.
- Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Science*, *7*(10), 454–459.
- Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347.

- Fedda, O. D., & Oweini, A. (2012). The effect of diglossia on Arabic vocabulary development in Lebanese students. *Educational Research and Reviews*, 7(16), 351–361.
- Feeney, A. & Scafton, S. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, 58(2), 121–132.
- Feeney, A., & Bonnefon, J. F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of Language and Social Psychology*, 32(2), 181–190.
- Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of *some*: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, 58(2), 121–132.
- Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Cambridge, MA: Harvard University Press.
- Frye, D., Braisby, N., Lowe, J., Maroudas, C., & Nicholls, J. (1989). Young children's understanding of counting and cardinality. *Child development*, 1158–1171.
- Gangopadhyay, I., Davidson, M. M., Weismer, S. E., & Kaushanskaya, M. (2016). The role of nonverbal working memory in morphosyntactic processing by school-aged monolingual and bilingual children. *Journal of experimental child psychology*, 142, 171–194.
- Ganzeboom, H. B., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social science research*, 25(3), 201–239.

- Gelman, R., & Gallistel, C. R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 3½–7 years old on a stroop-like day-night test. *Cognition*, *53*(2), 129–153.
- Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.
- Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486–489.
- Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490.
- Goetz, P. J. (2003). The effects of bilingualism on theory of mind development. *Bilingualism Language and Cognition*, *6*(1), 1–15.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, *40*(1), 84–96.
- Goldman, M. C., Negen, J., & Sarnecka, B. W. (2014). Are bilingual children better at ignoring perceptually misleading information? A novel test. *Developmental Science*, *17*(6), 956–964.
- Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, *16*(4), 562–276.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (eds.). *Syntax and semantics, 3: Speech acts* (pp. 41–58). New York: Academic Press

- (Reprinted in Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press).
- Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole & J. L. Morgan (eds.). *Syntax and semantics, 3: Speech acts* (pp. 113–128). New York: Academic Press (Reprinted in Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press).
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). “Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116*, 42–55.
- Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A. & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes, 20*(5), 667–696.
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of qualitative research*, 105–117.
- Gutiérrez-Clellen, V. F., & Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24*(2), 267–288.
- Hanlon, C. (1988). The emergence of set-relational quantifiers in early childhood. In F. S. Kessel (ed.), *The Development of language and language researchers: Essays In honor of Roger Brown* (pp. 65–78). Hillsdale, NJ: Erlbaum.
- Hanlon, C. C. (1987). “Acquisition of set-relational quantifiers in early childhood”. *Genetic, Social and General Psychology Monographs, 113*(2), 215–264.

- Hartshorne, J. K., & Snedeker, J. (2014). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures. *Manuscript submitted for publication*.
- Heim, S., McMillan, C., Clark, R., Baehr, L., Ternes, K., Olm, C., ... Grossman, M. (2016). How the brain learns how few are “many”: An fMRI study of the flexibility of quantifier semantics. *NeuroImage*, *125*, 45–52.
- Hendriks, P., Hoeks, J., De Hoop, H., Krämer, I., Smits, E.J., Spender, J. & de Swart, H. (2009). A large-scale investigation of scalar implicature. *Semantics and pragmatics: From experiment to theory*, 30–50.
- Hernandez, A. E., & Meschyan, G. (2006). Executive function is necessary to enhance lexical processing in a less proficient L2: Evidence from fMRI during picture naming. *Bilingualism: Language and cognition*, *9*(02), 177–188.
- Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child development*, 567–582.
- Horn, L. R. (1972). *On the semantic properties of the logical operators in English* (Doctoral dissertation). ProQuest (7301702)
- Huntley-Fenner, G. (2001). Children's understanding of number is similar to adults' and rats': numerical estimation by 5–7-year-olds. *Cognition*, *78*(3), B27–B40.
- Hurewitz, F., Papafragou, A., Gleitman, L. & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language Learning and Development*, *2*(2), 77–96.
- Hyde, K. F. (2000). Recognising deductive processes in qualitative research. *Qualitative market research: An international journal*, *3*(2), 82–90.

- Iluz-Cohen, P., & Armon-Lotem, S. (2013). Language proficiency and executive control in bilingual children. *Bilingualism: Language and Cognition*, 16(04), 884–899.
- International Labour Office (1990). *International Standard Classifications of Occupations (ISCO-88)*. International Labour Office, Geneva. Retrieved from <http://www.ilo.org/public/english/bureau/stat/isco/index.htm>.
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2006). Do children need concrete instantiations to learn an abstract concept. In *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 411–416). Mahwah, NJ: Erlbaum.
- Katsos, N. & Cummins, C. (2012). Scalar implicature: theory, processing and acquisition. In: *Nouveaux Cahiers de Linguistique Française*, 30, 39–52.
- Katsos, N., & Bishop, D. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67–81.
- Katsos, N., Breheny, R., & Williams, J. (2005). Interaction of structural and contextual constraints during the on-line generation of scalar inferences. *Proceedings of GLOW 28, University of Geneva*. Utrecht: Generative Linguistics in the Old World (GLOW).
- Katsos, N., Roqueta, C. A., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification?. *Cognition*, 119(1), 43–57.
- Kormi-Nouri, R., Shojaei, R. S., Moniri, S., Gholami, A. R., Moradi, A. R., Akbari-Zardkhaneh, S., & Nilsson, L. G. (2008). The effect of childhood bilingualism on episodic and semantic memory tasks. *Scandinavian Journal of Psychology*, 49(2), 93–109.

- Kousaie, S., & Phillips, N. A. (2012a). Conflict monitoring and resolution: Are two languages better than one? Evidence from reaction time and event-related brain potentials. *Brain research, 1446*, 71–90.
- Kousaie, S., & Phillips, N. A. (2012b). Ageing and bilingualism: Absence of a “bilingual advantage” in Stroop interference in a nonimmigrant sample. *The Quarterly Journal of Experimental Psychology, 65*(2), 356–369.
- Kura, B., & Sulaiman, Y. (2012). Qualitative and quantitative approaches to the study of poverty: taming the tensions and appreciating the complementarities. *The Qualitative Report, 17*(20), 1–19.
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.
- Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*(2), 395–438.
- Le Corre, M., Van de Walle, G., Brannon, E. M., & Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive psychology, 52*(2), 130–169.
- Lee, A. S. (1991). Integrating positivist and interpretive approaches to organizational research. *Organization science, 2*(4), 342–365.
- Lemer, C., Dehaene, S., Spelke, E., & Cohen, L. (2003). Approximate quantities and exact number words: Dissociable systems. *Neuropsychologia, 41*(14), 1942–1958.
- Levinson, S. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.
- Levinson, S. (2000). *Presumptive meanings*. Cambridge, MA: The MIT Press.

- Logan, G. D. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In D. Dagenbach & T. H. Carr (eds.), *Inhibitory processes in attention, memory, and language* (pp. 189–239). San Diego: Academic Press.
- Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, *114*(1), 96–104.
- Ma, H., Hu, J., Xi, J., Shen, W., Ge, J., Geng, F., Wu, Y., Guo, J. & Yao, D. (2014). Bilingual cognitive control in language switching: An fMRI study of English-Chinese late bilinguals. *PloS one*, *9*(9), e106468.
- Mannai, H. A., & Everatt, J. (2005). Phonological processing skills as predictors of literacy amongst Arabic speaking Bahraini children. *Dyslexia*, *11*(4), 269–291.
- Martin-Rhee, M. M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism: language and cognition*, *11*(01), 81–93.
- McMillan, C., Clark, R., Moore, P., Devita, C., & Grossman, M. (2005). Neural basis for generalized quantifier comprehension. *Neuropsychologia*, *43*, 1729–1737.
- Mezzacappa, E. (2004). Alerting, orienting, and executive attention: Developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. *Child development*, *75*(5), 1373–1386.
- Miller, C. (2007). Arabic Urban Vernaculars: Development and Change. In C. Miller, E. Al-Wer, D. Caubet, & J.C.E. Watson (eds.). *Arabic in The City: Issues in Dialect Contact and Language Variation* (pp. 1–31). New York: Routledge.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their

- contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100.
- Morales, J., Calvo, A., & Bialystok, E. (2013). Working memory development in monolingual and bilingual children. *Journal of experimental child psychology*, 114(2), 187–202.
- Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental science*, 10(6), 719–726.
- Mueller, S. T. (2010). A partial implementation of the BICA cognitive decathlon using the Psychology Experiment Building Language (PEBL). *International Journal of Machine Consciousness*, 2, 273–288.
- Mueller, S. T. (2011). The Psychology Experiment Building Language (Version 0.11) [Computer software]. Retrieved from: <http://pebl.sourceforge.net>.
- National Center for Education Statistics (2004). *ECLS-K base year public-use data files and electronic codebook*. U.S. Department of Education. Retrieved from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=20>.
- Negen, J., & Sarnecka, B. W. (2010). Analogue magnitudes and knower-levels: Revisiting the variability argument. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 1252–1257.
- Neuman, L. W. (2014). *Social research methods: Qualitative and quantitative approaches* (7th ed.). Edinburgh: Pearson Education Limited.
- Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, 63, 324–346.
- Noveck, I. (2001). When children are more logical than adults: Investigations of scalar implicature. *Cognition*, 78(2), 165–188.

- Noveck, I., & Posada, A. (2003). Characterising the time course of an implicature. *Brain and Language, 85*, 203–210.
- Noveck, I., & Sperber, D. (2012). The why and how of experimental pragmatics: The case of ‘scalar implicature’. In D. Wilson & D. Sperber (eds.), *Meaning and relevance* (pp. 307–330). Cambridge: Cambridge University Press.
- Odic, D., Le Corre, M., & Halberda, J. (2015). Children’s mappings between number words and the approximate number system. *Cognition, 138*, 102–121.
- Orlikowski, W. J., & Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information systems research, 2*(1), 1–28.
- Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive psychology, 66*(2), 232–258.
- Paivio, A., Yuille, J. C. & Madigan, S. A. (1986). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology, 1–25*.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition, 86*, 253–282.
- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition, 12*(1), 71–82.
- Parker Jones, O., Green, D.W., Grogan, A., Pliatsikas, C., Filippopolitis, K., Ali, N., ... Price, C. J. (2012). Where, when and why brain activation differs for bilinguals and monolinguals during picture naming and reading aloud. *Cerebral Cortex, 22*(4), 892–902.
- Pearson Education (2011). *Versant Arabic Test*. Retrieved from <https://www.versanttest.com/products/arabic.jsp>.

- Pearson, B. Z., Fernández, S. C., Lewedeg, V., & Oller, D. K. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics*, *18*(01), 41–58.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.
- Poarch, G. J., & van Hell, J. G. (2012). Executive functions and inhibitory control in multilingual children: Evidence from second-language learners, bilinguals, and trilinguals. *Journal of experimental child psychology*, *113*(4), 535–551.
- Politzer-Ahles, S. & Gwilliams, L. (2015). Involvement of prefrontal cortex in scalar implicatures: Evidence from magnetoencephalography. *Language, Cognition and Neuroscience*, *30*(7), 853–866.
- Posner, M.I. & Petersen, S.E. (1990). The Attention system of the Human Brain. *Annual Review of Neuroscience*, *13*, 25–42.
- Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, *14*(4), 347–376.
- Prabhakaran, V., Narayanan, K., Zhao, Z., & Gabrieli, J. D. E. (2000). Integration of diverse information in working memory within the frontal lobe. *Nature Neuroscience*, *3*(1), 85–90.
- Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and vocabulary scales (Section 4: Advanced Progressive Matrices)*. London: H. K. Lewis.

- Robinson, S. J., & Brewer, G. (2016). Performance on the traditional and the touch screen, tablet versions of the Corsi Block and the Tower of Hanoi tasks. *Computers in Human Behavior, 60*, 29–34.
- Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia, 42*(8), 1029–1040.
- Sammons, P., Anders, Y., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Barreau, S. (2009). Children’s cognitive attainment and progress in English primary schools during Key Stage 2: Investigating the potential continuing influences of pre-school education. In *Frühpädagogische Förderung in Institutionen* (pp. 179–198). VS Verlag für Sozialwissenschaften. [DOI: 10.1007/978-3-531-91452-7_12].
- Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition, 108*(3), 662–674.
- Schiff, N. B., & Ventry, I. M. (1976). Communication problems in hearing children of deaf parents. *Journal of Speech and Hearing Disorders, 41*(3), 348–358.
- Schneider, S. L. (2013). The international standard classification of education 2011. *Comparative Social Research, 30*, 365–379.
- Schroeder, S. R., & Marian, V. (2016). Cognitive consequences of trilingualism. *International Journal of Bilingualism, 1*–20.
- Schroyens, W., Schaeken, W., & Handley, S. (2003). In search of counter-examples: deductive rationality in human reasoning. *Quarterly Journal of Experimental Psychology, 65*(7), 1129–1145.
- Shalan, S. (2009). Considerations for developing and adapting language and literacy assessments in Arabic-speaking countries. In E. Grigorenko (ed.),

- Multicultural Psychoeducational Assessment* (pp. 287–314). New York: Springer.
- Shaalán, S. (2010). *Investigating grammatical complexity in Gulf Arabic speaking children with specific language impairment (SLI)* (Doctoral dissertation, University College London).
- Shetreet, E., Chierchia, G., & Gaab, N. (2014). When some is not every: Dissociating scalar implicature generation and mismatch. *Human Brain Mapping, 35*, 1503–1514.
- Shneiderman, B. (1991). A taxonomy and rule base for the selection of interaction styles. *Human factors for informatics usability, 325–342*.
- Siegal, M., Iozzi, L., & Surian, L. (2009). Bilingualism and conversational understanding in young children. *Cognition, 110*(1), 115–122.
- Siegal, M., Matsuo, A., Pond, C., & Otsu, Y. (2007). Bilingualism and cognitive development: Evidence from scalar implicatures. In *Proceedings of the Eighth Tokyo Conference on Psycholinguistics* (pp. 265–280). Tokyo: Hituzi Syobo.
- Siegal, M., Surian, L., Matsuo, A., Geraci, A., Iozzi, L., Okumura, Y., & Itakura, S. (2010). Bilingualism accentuates children's conversational understanding. *PloS one, 5*(2), e9004.
- Simon, R. J. (1969). Reactions towards the source of stimulation. *Journal of Experimental Psychology, 81*(1), 174–176.
- Singh, R., Wexler, K., Astle, A., Kamawar, D. & Fox, D., (2013). Children interpret disjunction as conjunction: consequences for the theory of scalar implicature. *Carleton University, ms*.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417–453.

- Slabakova, R. (2010). Scalar implicatures in L2 acquisition. In H. Caunt-Nulton, S. Kulatilake, & I. Woo (eds.), *Proceedings of the 31st Annual Boston University Conference on Language Development* (pp. 576–584). Somerville, MA: Cascadilla Press.
- Soliman, A. M. (2014). Bilingual advantages of working memory revisited: a latent variable examination. *Learning and Individual Differences, 32*, 168–177.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition*. Oxford: Blackwell.
- Sperber, D., & Wilson, D. (2002). Pragmatics modularity and mind reading. *Mind and Language, 17*, 3–23.
- Sperber, D., & Wilson, D. (2012). Pragmatics, modularity and misreading. In D. Wilson & D. Sperber (Eds), *Meaning and relevance* (pp. 261–278). Cambridge: Cambridge University Press.
- Stuss, D.T. (2011). Functions of the frontal lobes: relation to executive functions. *Journal of the international neuropsychological Society, 17*(05), 759–765.
- Subrahmanyam, K., Greenfield, P., Kraut, R., & Gross, E. (2001). The impact of computer use on children's and adolescents' development. *Journal of Applied Developmental Psychology, 22*(1), 7–30.
- Sutherland, A. (2012). Is parental socio-economic status related to the initiation of substance abuse by young people in an English city? An event history analysis. *Social Science & Medicine, 74*(7), 1053–1061.
- Tenaw, Y. A. (2014). Investigating the Impacts of College Students Background Academic Performance. *International Journal of Computer Applications, 92*(9), 8–11.

- Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, 15(4), 426–445.
- Tillman, K. A., & Barner, D. (2015). Learning the language of time: Children's acquisition of duration words. *Cognitive psychology*, 78, 57–77.
- Tomlinson, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, 69(1), 18–35.
- Troiani, V., Peelle, J. E., Clark, R., & Grossman, M. (2009). Is it logical to count on quantifiers? Dissociable neural networks underlying numerical and logical quantifiers. *Neuropsychologia*, 47, 104–111.
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (eds.), *Organization of Memory* (pp. 381-403). London: Academic.
- Ueno, K., Nadeo, T. & Iinaga, K. (1991). *Kaiga goi hatattu kensa* (Picture Vocabulary Test). Tokyo: Nihon Bunka Kagakusha.
- UNESCO (2011). *International Standard Classification of Education* (ISCED). Retrieved from <http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx>.
- Unsworth, S. (2013). Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: The case of Dutch gender. *Bilingualism: Language and Cognition*, 16(1), 86–110.
- Unsworth, S., Argyri, F., Cornips, L., Hulk, A., Sorace, A., & Tsimpli, I. (2014). The role of age of onset and input in early child bilingualism in Greek and Dutch. *Applied Psycholinguistics*, 35(04), 765–805.

- Veenstra, A.L., Riley, J.D., Barrett, L.E., Muhonen, M.G., Zupanc, M., Romain, J.E., Lin, J.J. & Mucci, G. (2016). The impact of bilingualism on working memory in pediatric epilepsy. *Epilepsy & Behavior*, *55*, 6–10.
- Versteegh, K. (2001). *The Arabic language*. Edinburgh: Edinburgh University Press.
- Videsott, G., Della Rosa, P. A., Wiater, W., Franceschini, R., & Abutalebi, J. (2012). How does linguistic competence enhance cognitive functions in children? A study in multilingual children with different linguistic competences. *Bilingualism: Language and Cognition*, *15*(04), 884–895.
- Wagner, K., Kimura, K., Cheung, P., & Barner, D. (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive psychology*, *83*, 1–21.
- Webster, M., & Sell, J. (2007). Why do experiments?. In M. Webster & J. Sell (eds.), *Laboratory experiments in the social sciences*, (pp. 5–24). New York: Academic Press.
- Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1967/2012). *Wechsler preschool and primary scale of intelligence* (4th ed.). New York: Psychological Corporation.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.
- Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child development*, *73*(3), 768–778.
- Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193.

- Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive psychology*, *24*(2), 220–251.
- Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of memory and language*, *87*, 128–143.
- Zajenkowski, M., & Szymanik, J. (2013). MOST intelligent people are accurate and SOME fast people are intelligent: Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence*, *41*, 456–466.
- Zhao, M., Liu, T., Chen, G., & Chen, F. (2015). Are scalar implicatures automatically processed and different for each individual? A mismatch negativity (MMN) study. *Brain Research*, *1599*, 137–149.
- Zou, L., Ding, G., Abutalebi, J., Shu, H., & Peng, D. (2012). Structural plasticity of the left caudate in bimodal bilinguals. *Cortex*, *48*(9), 1197–1206.

Appendices

Appendix 1. Language exposure questionnaire

ABOUT THE CHILD

- Name _____
- Gender _____
- Place of birth _____
- Date of birth _____
- Was the child born more than 6 weeks premature? yes no
- Date of arrival in UK (if not born here) _____
- Home Language(s) of the child _____

- At what age did your child start receiving regular exposure to English?

When child was:

- 0-1 year old 4-5 years old
- 1-2 years old 5-6 years old
- 2-3 years old
- 3-4 years old

- Where did your child start receiving regular exposure to English for the first time?

- at home
- at playgroup
- at nursery
- at primary school
- somewhere else: _____

- Does your child have free school dinners

- yes
- no

P.T.O.

ABOUT THE PARENTS

● Country of origin (Mother) _____
 (Father) _____

● Date of arrival in UK (Mother) _____
 (Father) _____

● How well do you speak English?

not at all not well quite well very well

Mother

Father

● What language(s) do you speak with the child?

MOTHER		
Home Language	English	3rd language (only if there is)
<input type="checkbox"/> Always	<input type="checkbox"/> Always	<input type="checkbox"/> Always
<input type="checkbox"/> Usually	<input type="checkbox"/> Usually	<input type="checkbox"/> Usually
<input type="checkbox"/> Half the time	<input type="checkbox"/> Half the time	<input type="checkbox"/> Half the time
<input type="checkbox"/> Rarely	<input type="checkbox"/> Rarely	<input type="checkbox"/> Rarely
<input type="checkbox"/> Never	<input type="checkbox"/> Never	<input type="checkbox"/> Never

FATHER		
Home Language	English	3rd language (only if there is)
<input type="checkbox"/> Always	<input type="checkbox"/> Always	<input type="checkbox"/> Always
<input type="checkbox"/> Usually	<input type="checkbox"/> Usually	<input type="checkbox"/> Usually
<input type="checkbox"/> Half the time	<input type="checkbox"/> Half the	<input type="checkbox"/> Half the time
<input type="checkbox"/> Rarely	time	<input type="checkbox"/> Rarely
<input type="checkbox"/> Never	<input type="checkbox"/> Rarely	<input type="checkbox"/> Never
	<input type="checkbox"/> Never	

P.T.O.

- What language(s) does the child speak to you?

child to MOTHER		
Home Language	English	3rd language (only if there is)
<input type="checkbox"/> Always <input type="checkbox"/> Usually <input type="checkbox"/> Half the time <input type="checkbox"/> Rarely <input type="checkbox"/> Never	<input type="checkbox"/> Always <input type="checkbox"/> Usually <input type="checkbox"/> Half the time <input type="checkbox"/> Rarely <input type="checkbox"/> Never	<input type="checkbox"/> Always <input type="checkbox"/> Usually <input type="checkbox"/> Half the time <input type="checkbox"/> Rarely <input type="checkbox"/> Never

child to FATHER		
Home Language	English	3rd language (only if there is)
<input type="checkbox"/> Always <input type="checkbox"/> Usually <input type="checkbox"/> Half the time <input type="checkbox"/> Rarely <input type="checkbox"/> Never	<input type="checkbox"/> Always <input type="checkbox"/> Usually <input type="checkbox"/> Half the time <input type="checkbox"/> Rarely <input type="checkbox"/> Never	<input type="checkbox"/> Always <input type="checkbox"/> Usually <input type="checkbox"/> Half the time <input type="checkbox"/> Rarely <input type="checkbox"/> Never

- When mother and father are together with the child, who speaks most to the child?
 - Mother
 - Father
 - Both an equal amount

P.T.O.

OTHER HOUSEMATES

● Does your child have sisters or brothers? Yes No

● If yes, Name of sibling 1 _____ Age _____
Name of sibling 2 _____ Age _____
Name of sibling 3 _____ Age _____
Name of sibling 4 _____ Age _____
Name of sibling 5 _____ Age _____

● What language(s) do the siblings speak with the child? _____

● Besides the parents and siblings, does another adult look after your child (e.g. nanny, grandmother, aunt)?

Yes No

● If yes, what is the relation of this adult to the child? _____

● What language(s) does this adult speak to the child? _____

● What language(s) does the child speak to this adult? _____

Please fill in the information relating to this other adult in the “other” column, in the tables below!!!

P.T.O.

AVERAGE DAY

- Please describe who spends time with the child on an average day **during the week?**

Please tick the relevant boxes. If more than one person is with the child at the same time, circle the tick to show who is interacting more with the child.

	Mother	Father	Siblings	School	Other adult (specify person) _____ -
7 am – 8 am					
8 am – 9 am					
9 am – 3 pm				✓	
3 pm – 4 pm					
4 pm – 5 pm					
5 pm – 6 pm					
6 pm – 7 pm					
7 pm – bedtime					

- Please describe who spends time with the child on an average day **during the weekend?**

Please tick the relevant boxes. If more than one person is with the child at the same time, circle the tick to show who is interacting more with the child.

	Mother	Father	Siblings	Other adult (specify person) -----
7 am – 9 am				
9 am – 11 am				
11 am – 1 pm				
1 pm – 3 pm				
3 pm – 5 pm				
5 pm – 7 pm				
7 pm – bedtime				

- How many weeks per year is your child on holiday from school? ____
- How many weeks per year does the child spend in the family's country of origin? ____
- How often does your child speak English during the holidays?
 - Always
 - Usually
 - Half the time
 - Rarely
 - Never
- Please describe who spends time with the child on an average day **during the holiday?**

Please tick the relevant boxes. If more than one person is with the child at the same time, circle the tick to show who is interacting more with the child.

	Mother	Father	Siblings	Other (specify person) _____
7 am – 9 am				
9 am – 11 am				
11 am – 1 pm				
1 pm – 3 pm				
3 pm – 5 pm				
5 pm – 7 pm				
7 pm – bedtime				

P.T.O.

OTHER ACTIVITIES

- How often do you do activities with your child?

For instance: going to museums / going to the zoo / going to a film / going to the swimming pool / etc.

- Often
- Regularly
- Sometimes
- Never

- What activities does the child do each week in what language?

Please give the total NUMBER OF HOURS per week, e.g. 2 hours per week

activity	HOME LANGUAGE	
	Monday- Friday	Saturday-Sunday
Reading with an adult		
Using computer		
Watching TV		
Sports		
Playing with friends / cousins		

activity	ENGLISH	
	Monday- Friday	Saturday-Sunday
Reading with an adult		
Using computer		
Watching TV		
Sports		
Playing with friends / cousins		

Appendix 2. Item list for Experiment 3

Experiment 3: Ternary judgment pragmatic task (enriched context)					
Quantifier/operator	Scene (items in the PowerPoint slide)	Action (in the PowerPoint slide)	Utterance	Experimental condition	Utterance in Arabic
Some	The elephant likes pushing things. There are 5 buses and 5 trucks	The elephant pushed all the trucks	The elephant pushed some of the trucks	Under-info	الفيل حرّك بعض الشاحنات
	The giraffe is hungry for fruit. There are 5 apples and 5 pears on the trees	The giraffe ate all the pears	The giraffe ate some of the pears	Under-info	الزرافة أكلت بعض الكامثرات
	Mr. Tough likes lifting stuff up. There are 5 boxes and 5 stones	Mr Tough lifted all the boxes	Mr Tough lifted some of the boxes up.	Under-info	السيد عملاق رفع بعض الصناديق
	The mouse wants to pick up vegetables. There are 5 pumpkins and 5 carrots.	The mouse picked up all the carrots	The mouse picked up some of the carrots	Under-info	الفأر أخذ بعض الجزرات
	The goat likes jumping over things. There are 5 fences and 5 bushes	The goat jumped over two out of five fences	The goat jumped over some of the fences	Optimal	الماعز قفز فوق بعض الحواجز
	The boy likes carrying his things. There are 5 books and 5 shoes	The boy carried two out of five books	The boy carried some of the books	Optimal	الولد رفع بعض الكتب
	The girl likes collecting flowers. There are 5 red flowers and 5 yellow flowers	The girl collected two out of five red flowers	The girl collected some of the yellow flowers	False	البنت جمعت بعض الورد الأصفر
	The boy likes giving yummy stuff to	The boy gave the	The boy gave the elephant	False	الولد أعطى الفيل بعض

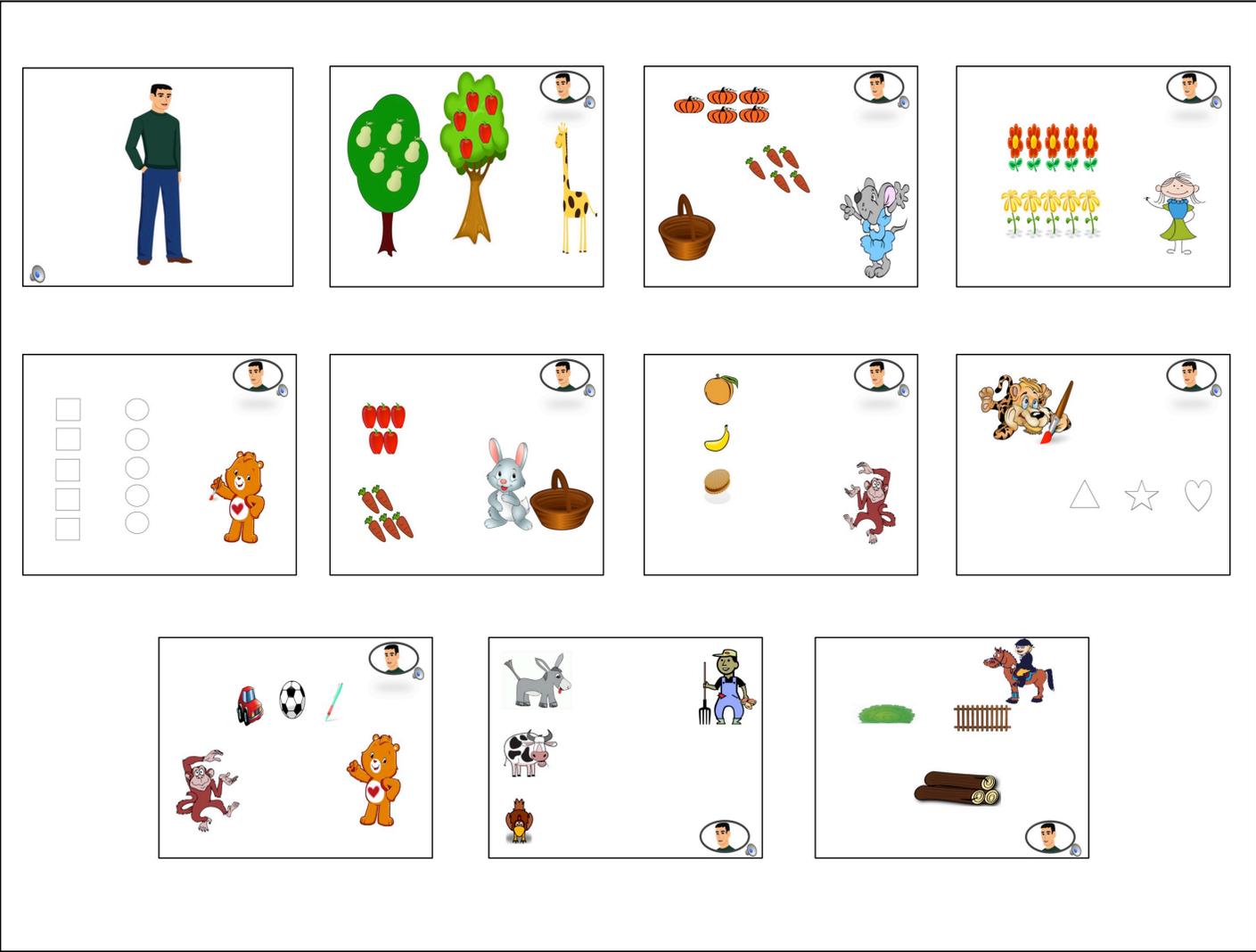
	his friend the elephant. There are 5 bananas 5 biscuits	elephant two out of five bananas	some of the biscuits		البسكويت
Most	The crocodile wants to play with toys. There are 5 cars and 5 dolls	The crocodile played with all the cars	The crocodile played with most of the cars	Under-info	التمساح لعب بمعظم السيارات
	The girl likes erasing things. There are 5 hearts and 5 stars.	The girl erased all the hearts	The girl erased most of the hearts	Under-info	البنيت مسحت معظم القلوب
	The rabbit likes putting carrots in his bag. There are 5 carrots and 5 apples	The rabbit put all the carrots in his bag	The rabbit put most of the carrots in his bag.	Under-info	الارنب وضع معظم الجزرات في السلة
	The horse like jumping over things. There are 5 fences and 5 bushes.	The horse jumped over all the fences	The horse jumped over most of the fences.	Under-info	الحصان قفز فوق معظم الحواجز
	The bear likes coloring things. There are 5 circles and 5 squares	The bear coloured four out of five circles	The bear coloured most of the circles	Optimal	الدب جمع معظم التفاحات
	The bear likes gathering yummy things; there are 5 pieces of cake and 5 apples	The bear gathered four out of five apples.	The bear gathered most of the apples.	Optimal	السلفاة لعبت بمعظم الشاحنات
	The turtle likes playing with his toys. There are 5 balls and 5 trucks	The turtle played with four out of five balls	The turtle played with most of the trucks	False	الأرنب أخذ معظم الجزرات
	The rabbit likes getting yummy things. There are 5 carrots and 5 apples	the rabbit got four out of five apples	The rabbit got most of the carrots	False	الدب لَوّن معظم الدوائر
Or	The bear likes providing presents to his friend, the monkey. There are a pen, a ball, a car,	The bear gave the monkey the ball and the pen	The bear gave the monkey the pen or the ball.	Under-info	الدب أعطى القرد القلم أو السيارة

	The boy likes watering the plants in the garden. There are a tree, a flower, and some grass	The boy watered the flower and the tree	The boy watered the tree or the flower	Under-info	الولد أسقى الشجرة أو الوردة
	The girl likes buying new things. There are a ring, a dress, a hat	The girl bought a ring and a hat.	The girl bought the hat or the ring.	Under-info	البنيت اشترت الطاقية أو الخاتم
	The horse likes jumping over several things. There are a fence, a bush, and wood.	The horse jumped over wood and a fence.	The horse jumped over the fence or the wood.	Under-info	الحصان قفز فوق الحاجر أو الخشب
	The squirrel is hungry and wants to take something tasty. There are a pizza, a hamburger, and an ice-cream	The squirrel took the pizza	(I'm not sure I saw it well), the squirrel took the pizza or the hamburger	Optimal	مو متأكدة انو شفتو بوضوح السنجاب أخذ البيتزا أو الهامبرجر
	the girl likes erasing shapes. There are a triangle, a heart and a circle	the girl erased the heart	(I'm not sure I saw it well),the girl erased the heart or the circle	Optimal	مو متأكدة انو شفتو بوضوح البنيت مسحت القلب أو الدائرة
	The man likes washing his things. There are a bicycle, a set of motorcycle, and a car	The man washed the car	The man washed the bicycle or the motorcycle	False	الرجل غسل الدراجة أو الدباب
	the farmer likes feeding his animals. There are a donkey, a cow and a chicken	the farmer fed the cow	The farmer fed the chicken or the donkey	False	الفلاح أكل الحمار أو الدجاجة
And	The monkey loves picking up yummy stuff. There are a banana, an orange and a biscuit	The monkey picked up the orange and the biscuit	The monkey picked up the biscuit	Under-info	القرود أكل البسكويت

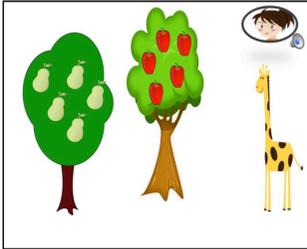
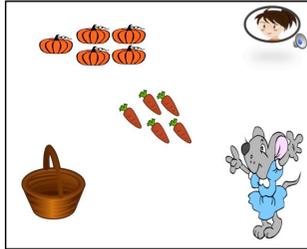
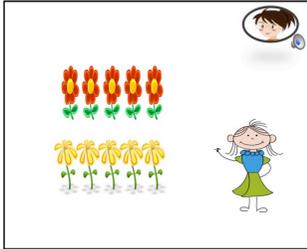
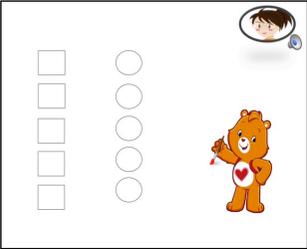
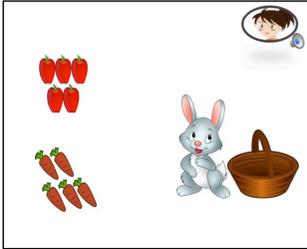
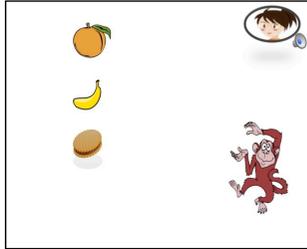
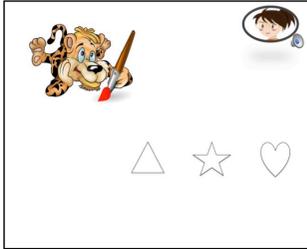
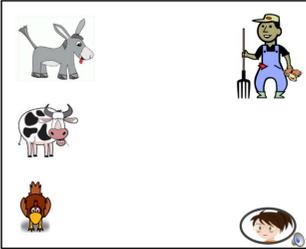
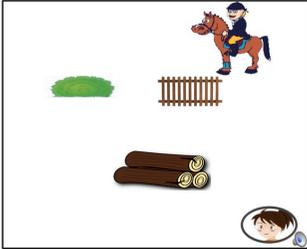
The girl likes packing beautiful clothes in her suitcase. There are a skirt, t-shirt, and a dress	The girl packed the t-shirt and the skirt	The girl packed the skirt	Under-info	البنيت اشترت التنورة
The spaceman wants to buy stuff for his new spaceship. There are a computer, a desk, a TV and	The spaceman bought the computer and the desk	The spaceman bought the desk	Under-info	رجل الفضاء اشترى المكتب
The tiger is an artist. He likes painting things. There are a star, a triangle and a heart	The tiger painted the heart and the triangle	The tiger painted the triangle	Under-info	النمر لون المثلث
The builder likes carrying things around. There are a chair, a bucket and a ladder	The builder carried the bucket and the ladder	The builder carried the bucket and the ladder	Optimal	المهندس حمل السطل والسلّم
The man likes buying new things before travelling. There are a T-shirt, a trouser, and a pair of shoes.	The man bought the T-shirt and the shoes	The man bought the T-shirt and the shoes	Optimal	الرجل اشترى التي شيرت والجزمة
The dog wants to pick up some fruit. There are an apple, an orange and a banana	The dog picked up the banana and the apple	The dog picked up the orange and the banana	False	الكلب مسك البرتقاله والموزة
The boy likes moving things. There are a box, a desk and a TV	The boy moved the TV and the desk	The boy moved the desk and the box	False	الولد حرّك المكتب والصندوق

Appendix 3. Sample stimuli used for Experiment 3

a) Sample for the task in English (Mr Kareem)



b) Sample for the task in Arabic (Ms Sara)

Appendix 4. Item list for Experiment 4

Experiment 4: Ternary judgment pragmatic task (no context)			
Quantifier/operator	Utterance	Experimental condition	Utterance in Arabic
Some	Some cats have tails	Infelicitous	بعض الكلاب لها ذيل
	Some giraffes have long necks	Infelicitous	بعض الزرافات لها رقاب طويلة
	Some televisions have screens	Infelicitous	بعض التلفزيونات لها شاشات
	Some elephants have trunks	Infelicitous	بعض الفيلة لها خرطوم
	Some people wear glasses	Felicitous	بعض الناس يلبسوا نظارات
	Some birds live in cages	Felicitous	بعض الطيور تعيش في أقفاص
	Some birds have telephones	Bizarre	بعض العصافير عندها موبايلات
	Some flowers can talk	Bizarre	بعض الورود تقدر تتكلم
Most	Most horses have four legs	Infelicitous	معظم الخيول لها أربع أرجل
	Most people have a head	Infelicitous	معظم الناس لديهم رأس
	Most fish live in water	Infelicitous	معظم الأسماك تعيش في الماء
	Most refrigerators have doors	Infelicitous	معظم الثلاجات فيها صفحات
	Most people have breakfast in the morning	Felicitous	معظم الناس يأكلو وجبة الفطور بالصباح
	Most houses have a staircase	Felicitous	معظم المنازل فيها درج
	Most chairs can walk	Bizarre	معظم الكراسي تتكلم
	Most cars can sing	Bizarre	معظم السيارات تغني
Or	When you cross the road, you have to look left <i>or</i> right before	Infelicitous	لما تقطع الشارع لازم تلتفت يمين أو يسار

	crossing		
	You clean your teeth by using toothpaste <i>or</i> a toothbrush	Infelicitous	لما تفرشي اسنانك تستعمل الفرشاة أو المعجون
	Before going out, people wear a left shoe or right shoe	Infelicitous	الناس يستطيعو التنفس من فمهم أو أنفهم
	To survive, people need to drink water <i>or</i> eat food	Infelicitous	عشان يظل الناس على قيد الحياة، لازم يشربو ماء أو يأكلو طعام
	When babies are born, they are a girl or a boy	Felicitous	في كرة القدم، فريق واحد سيفوز أو يسخر
	When writing on a paper, people use their left or right hands	Felicitous	عند الكتابة على ورقة، الناس يستخدمو يدهم اليمين أو اليسار
	To survive, people can eat books or stones	Bizarre	عشان يظل الناس أحياء، لازم يأكلو كتب أو حجار
	To write your name, you need to use a flower or a carrot	Bizarre	حتى تكتب اسمك، تحتاج تستخدم وردة أو جزرة
And	To clap, you need to use your right hand	Infelicitous	حتى تصفق، انت تحتاج تستخدم ايدك اليمين
	To ride a bike, the front wheel needs to turn around	Infelicitous	حتى تقود الدراجة، العجلة الأمامية لازم تدور
	When you walk, you use your left leg	Infelicitous	لما تمشي تحتاج انو تستخدم رجلك اليسار
	People hear sounds around them using their right ear.	Infelicitous	حتى تقص دائرة، تحتاج يكون عندك ورقة
	To make a chees sandwich, you need bread and chees	Felicitous	عشان تعمل ساندويتش جبين، تحتاج قطعة خبز وجبن
	To wash and clean your hands, you need soap and water	Felicitous	حتى تغسل وتنظف ايديك، انت تحتاج لماء وصابون
	To cut out a circle, you need a computer and a telephone	Bizarre	عشان تعمل ساندويتش تحتاج تستخدم كمبيوتر وتلفون
	For breakfast, you can eat pens and books	Bizarre	لوجبة الفطور، ممكن انو تاكل أقلام أو كتب

Appendix 5. Predictors for pragmatic performance: Ordinal logistic regression

Ordinal logistic regression results: Predictors for children's pragmatic performance in under-informative items in the two context conditions.

Predictor		B	SE	X ²	DF	P
Threshold	Large response	-.097	.3671	.070	1	.792
	Medium response	.991	.3676	7.26	1	.007
Language group		.	.	13.121	3	.004
Bilinguals (English)		.367	.3852	.908	1	.341
Bilingual (Arabic)		.843	.3865	4.76	1	.029
Arabic children		-.611	.4581	1.78	1	.182
English children		0
Context condition (no context)		.206	.0675	9.29	1	.002
SES (FAS)		.	.	47.013	2	.000
SES (Low)		-.935	.1662	31.65	1	.000
SES (Medium)		.232	.0763	9.248	1	.002
SES (High)		0
Age (month)		-.006	.0081	.513	1	.474
Vocabulary (%)		-.020	.0054	13.41	1	.000
Quantifier Type		.	.	117.836	3	.000
Most		-.690	.0927	55.37	1	.000
Some		-.571	.0918	38.77	1	.000
Or		-.994	.0962	106.66	1	.000
And		0
STM		.163	.0618	6.99	1	.008
Inhibition		1.47	.261	31.69	1	.000
Group * STM		.	.	5.667	3	.129
Bilinguals (English)* STM		-.041	.0868	.228	1	.633
Bilinguals (Arabic)* STM		-.181	.0859	4.441	1	.035
Arabic children * STM		-.001	.1276	.000	1	.996
English children * STM		0
Group * inhibition		.	.	29.462	3	.000
BCE * inhibition		-.900	.284	10.04	1	.002
BCA * inhibition		-1.48	.282	27.54	1	.000
AC* inhibition		-1.56	.467	11.12	1	.001
EC * inhibition		0

The reference levels in the model: for (dependent variable) response type (small), for group (English children), for the context (no-context condition), quantifier type 'and', for the SES (High). The dependent variable ternary responses (small, mediums, large), when using (large as a reference level, the model reveal the same results but with the estimates in reverse values (e.g. negative instead of positive). The score used for inhibition is the adjusted score resulted form the Cox regression.